UNIVERSITÀ DEGLI STUDI DIE CAMERINO

SCHOOL OF ADVANCED STUDIES
DOCTOR OF PHILOSOPHY IN INFORMATION SCIENCE AND COMPLEX SYSTEMS
XXV CICLO

SCHOOL OF SCIENCE AND TECHNOLOGY

Automatic, Format-independent Generation of Metadata for Documents Based on Semantically Enriched Context Information

BARBARA THÖNSSEN

Advisors
Prof. Dr. Emanuela Merelli
Prof Dr. Knut Hinkelmann

Wenn Herr K. einen Menschen liebte
"Was tun Sie", wurde Herr K. gefragt, "wenn Sie einen Menschen lieben?" "Ich mache einen Entwurf von ihm", sagte Herr K., "und sorge, daß er ihm ähnlich wird." "Wer? Der Entwurf?" "Nein", sagte Herr K., "Der Mensch."

Berthold Brecht

DECLARATION

I declare that this dissertation is my own work and all the sources have been quoted and acknowledge by means of complete references.

Abstract of the Dissertation

The purpose of this study was to investigate how metadata can be generated automatically for all types of documents used in an enterprise, regardless of their content. Because of the increasing number of non-textual documents, i.e. images, audio and video files, full-text indexing is not applicable and thus, the use of metadata has become more and more important for resource description and discovery. However, creating metadata manually is time consuming and error prone and moreover barely feasible for the huge amount of documents an enterprise deals with daily. Thus, an approach for automatic, format-independent metadata generation is required. To begin the documents' context was analysed. A document is considered an enterprise object, which is related to other enterprise objects such as a task the document is used in and the purpose it is created for. It was recognised that context of a document can be described formally and semantically enriched in an enterprise architecture. This enterprise architecture description can then be used for automatic metadata generation. To use the enterprise architecture description in a productive environment it was determined how its objects can be linked to enterprise components, e.g. information stored in a relational database. Finally a procedure for setting-up, conducting and utilizing the metadata generation approach in an enterprise was identified. The combination of these objectives has been called *mintApproach.* With the mintApproach system the huge annual economic loss due to the vast time wasted on information retrieval is addressed.

Research design followed the deductive approach and a mixed method strategy was employed, combining the four methods: results of a Representative Study provided a comprehensive source for the analysis of the use of document creation tools in enterprises and preferred search strategies. Qualitative interviews conducted in a survey and based on a structured questionnaire provided insights on document handling in enterprise. Action Research and prototyping was applied in two different types of organisations, a non-profit organisation (NPO) in the domain of sexual health and a small and medium-sized enterprise (SME), developing contract management software. Evolutionary 'prototyping' built an integrated part of the Action Research studies and led to the development of an executable prototype. Applying Action Research in two enterprises, with very different business and business goals, helped to avoid the common pitfalls of this method like subjectivity, lack of generality and replication.

The results of the survey and the Action Research studies endorsed the fact that for document management in enterprises and public administrations alike, a document's context is considered. Although relations between documents and other enterprise objects may be hidden, low level governance instruments like guidelines for file storage help to reveal these relations. For example relations to other enterprise objects like a product or a client are implicit in the file structure in which a document is stored. Determining the naming conventions for files is another way of implicitly stating relations between enterprise objects and documents. This explicit information is represented in a semantically enriched Enterprise Architecture description. It was found that the well-known standard for Enterprise Architecture modelling, ArchiMate, was well suited for providing the basis for core enterprise ontology. ArchiMate was refined, enhanced, and represented in RDFS-Plus, an ontology language which is machine executable but also cognitively adequate for humans.

This core ontology was enhanced by application of specific ontologies reflecting enterprise specific needs, for example for representing domain knowledge or improving contract lifecycle management. The enterprise ontology was considered a part of an enterprise repository, comprising all enterprise objects constituting an organisation despite their

representation. Thus, for automatic, format-independent metadata generation based on context, ontology-to-database-mapping was considered suitable (why not was used?)

The approach was evaluated based on an executable prototype that illustrates the scientific models and makes it easier for the evaluators to assess the underlying scientific concepts. Goal of the evaluation was to determine the appropriateness, capability and applicability of the mintApproach. The mintApproach, visualized in the MeGaWorkbench prototype, was assessed as appropriate for automatic format-independent metadata generation for business documents. Using context for metadata generation was considered promising, particularly regarding multi-media documents, respectively documents with little, meaningless or even wrong document attributes. The mintApproach was considered beneficial as it helps to meet business needs in handling the ever-increasing amount of unstructured information by reducing the amount of personnel time involved.

Aknowledgements

It is a pleasure to thank the many people who made this thesis possible.

My thanks are due first to my supervisor Professor Dr. Emanuela Merelli, for her professional guidance, contribution of ideas and the generous sharing of her copious knowledge. With her openness and warmth she has helped me through the times when encouragement and motivation were most needed. I would particularly like to thank my co-supervisor, Professor Dr. Knut Hinkelmann. With his scientific questioning spirit, he constantly challenged me but also provided clear and simple illustrations of his thoughts, sound advice, and lots of good ideas throughout my thesis-writing period.

I would like to thank my two Action Research partners, Johannes Schläpfer from AHSGA and Andreas Kyriakakis from Symfact and their teams for their time, critical discussions and valuable contributions without which this work couldn't have been done. I am grateful to everyone who agreed to the interviews and gave valuable input to solving the problem and later feedback to my results.

I am indebted to Jonas Lutz, who did the nasty Java programming and put flesh on my prototype. I also would like to thank Andreas Martin for his kind assistance with various applications, his reassurance and ability to mediate.

I am also very thankful to Antoinette Lambourne for proof-reading my thesis, helping to avoid the manifold pitfalls a foreign language entails. I wish to thank in addition Daniel Hertig, who fought for me through the perils of the text processor and organized the layout of the document.

I am grateful to FHNW's librarians for assisting me in many different ways. Silvia Schwappacher deserves special mention.

My time at Camerino was made enjoyable in large part due to the colleagues that have become good friends, first of all Dr. Luca Tesei and Dr. Barbara Re. I am grateful for time spent at Anna and Sydney Higgins' home enjoying their hospitality and their pets' company.

I would like to thank all those friends I have had to abandon during the last three years, for their understanding. I wish to thank particularly Henrike Jonsson, Christine Riderer, Monika Scherer-Grünenfelder, Eva-Maria Neisser and Sandra Hauser for helping me get through the difficult times, and for all the emotional support, distraction, and caring they provided.

I wish to thank my parents, Ernst Thönssen and Elisabeth Thurau. Long ago, each in its own way prepared the ground for the work I have now accomplished. I am thankful to my wider family, Anke, Ole, Nils and Tjark Thönssen, and Silke Kilian for their interest in my work and inspiring confidence. I want to apologize to my godchildren, Salina and Anna, for spending so much time working on dead matter rather than sharing their lively ideas. I am very grateful to their parents who put up with this, and provided me with lifts and practical help.

Lastly, and most importantly, I wish to thank my husband, Johannes, for his enduring love, for believing in me, and supporting me in many, many ways to reach my goal and complete this work.

To him I dedicate this thesis.

List of Publications

Thönssen, B. & Lutz, J., 2012. Semantically Enriched Obligation Management. An Approach for Improving the Handling of Obligations Represented in Contracts. 4th International Conference on Knowledge Management and Information Sharing.

Emmenegger, S., Lorenzini, E., Thönssen, B., 2012. Improving supply-chain-management based on semantically enriched risk descriptions. 4th International Conference on Knowledge Management and Information Sharing.

Thönssen, B., 2012. Turning Risks Into Opportunities. Electronic Government, tbp. Available at: http://www.inderscience.com/browse/index.php?journalID=72.

Brander, S., Hinkelmann, K., Hu, B., Martin, A., Riss, U. V., Thönssen, B., Witschel, H. F., 2011. Refining process models through the analysis of informal work practice. In 9th International Conference on Business Process Management. Clermont-Ferrand, France.

Thönssen, B., 2011. Forrmalizing low-level governance instruments for a more holistic approach to automatic metadata generation. In Proceedings of the 5th International Conference on Methodologies, Technologies and Tools enabling e-Government. Camerino, Italy, pp. 1-12.

Thönssen, B. & Wolff, D., 2011. A broader view on Context Models to support Business Process Agility. In S. Smolnik, F. Teuteberg, & O. Thomas, eds. Semantic Technologies for Business and Information Systems Engineering: Concepts and Applications.

Brander, S., Hinkelmann, K., Martin, A., Thönssen, B.,, 2011. Mining of Agile Business Processes. In: Hinkelmann, Knut; Thönssen, Barbara: AI for Business Agility. AAAI 2011 Spring Symposium. Technical Report SS-11-03. Published by The AAAI Press, Menlo Park, California.

Hinkelmann, K., Thönssen, B. (Chair) (2011): AI for Business Agility. Papers from the AAAI Spring Symposium, 21-23 March 2011, Stanford University, California.

Thönssen, B., 2010. An Enterprise Ontology Building the Bases for Automatic Metadata Generation. In Proceedings of the 4th International Conference on Metadata and Semantics, MTSR1200. Madrid, pp. 195-210.

Feldkamp, D., Hinkelmann, K. & Thönssen, B., 2010. The Modelling of Knowledge-Intensive Processes using Semantics. In T. Vitvar, V. Peristeras, & K. Tarabanis, eds. Semantic Technologies for E-Government. Springer.

Feldkamp, D., Hinkelmann, K., Thönssen, B., 2010. Ontologies for E-Government. In: Poli, R., Healy, M., & Kameas, A. (Eds.): Theory and Applications of Ontology: Computer Applications: Heidelberg: Springer, pp. 429-462

Hinkelmann, K., Merelli, E. & Thönssen, B., 2010. The Role of Content and Context in Enterprise Repositories. In Proceedings of the 2nd International Workshop on Advanced Enterprise Architecture and Repositories - AER 2010.

Witschel, H-F., Hu, B., Riss, U., Thönssen, B., Brun, R., Martin, A., Hinkelmann, K., 2010. A Collaborative Approach to Maturing Process-related Knowledge. In: 8th International Conference on Business Process Management (BPM 2010). New York, USA

Riss, U., Witschel, H-F., Brun, R., Thönssen, B., 2009. What is Organizational Knowledge Maturing and how can it be assessed? I-KNOW'09, 9th International Conference on Knowledge Management and Knowledge Technologies.

Feldkamp, D. et al., 2008. E-Government for Distributed Autonomous Administrations.

Brun, R., Hinkelmann, K., Telesko, R., Thönssen, B., 2008. Towards an Integrated Approach to Assess the Potential of an Enterprise to Mature Knowledge. In: WM2009, 5th Conference of Professional Knowledge Management. 2008. Jg. S. 440-449.

Typographical Conventions

The following typographical conventions are used in this thesis:

Use of Capital and Small Initial Letters
: capitals/lower case (cap/lc) is used, i.e. all words of four letters or more in headings, titles, and subtitles are capitalized

‘quotation marks’
: indicate the specific use of a term or expression, e.g. ‘semantic gap’.

Italic
: is used to highlight an issue, e.g. *harvesting* in contrary to *extracting*.

`EntityOne entityIsRelatedToEntity EntitiyTwo`
: Couries New font indicates concepts and properties of an ontology; upper and lower cases follow the W3C Working Group[1] who uses the so-called ‘camelCase’, with upper-case first letter for class names (`EntityOne`) and lower-case first letter for properties (`entityIsRelatedToEntity`); spaces between terms are removed.

`dc:creator`
: characters before colon shows abbreviations of namespaces used to qualify entities of an ontology.

<author_name>
: ankle brackets are used to indicate wildcards for the actual value.

[1] W3C Working Group Note 26 June 2007. Web Services Description Language (WSDL) Version 2.0: RDF Mapping. URL: http://www.w3.org/TR/wsdl20-rdf/ (retrieved: 18.5.2012)

CONTENT

List of Figures

List of Tables

1 Introduction

There are a large number of business-related documents of any type in companies but they are poorly managed and therefore hard to access. To improve the management of electronic documents in enterprises by taking into account the relation between the documents and its business context is the objective of this thesis.

Documents have been the storehouse of knowledge for millenniums and still are important (Eriksson 2007, p 624): "Each day, millions of people use computers as enhanced typewriters to produce documents". In recent years more and more multi-media documents (audio, video, and images) are used in enterprises. Teleconferences are recorded, workshops are videoed or drafts on flipcharts photographed and CEOs inform their staff via podcasts. Gonsalves (2005) refers to a research group saying that the number of podcast users in the United States is expected to increase nearly 15 fold over the next five years, and is expected to reach 60 Million in five years.

In 2008 the Information Overload Research Group (IORG)[2] was founded to deal with the ever increasing amount of information. Basex, an American knowledge economy research and advisory firm, set up a calculator on their web site to bring to light enterprises information overload.[3]

The situation is getting worse as corporate governance requires that all business information has to be managed adhering to law[4] - and many enterprises simply do not know what documents they have or where they are (Zöller, 2009). Finally, time becomes increasingly short for saving all the information in the given timeslot (Pickert, 2008).

Since the eighties, an ever growing number of applications for enterprise content management (ECM) have entered the market (Zöller, 2007) but many projects have failed to implement them. Le Clair & Poore (2008) from Forrester Research state that "current approaches to enterprise content management (ECM) don't work for most enterprises. Low adoption rates and frustrated users plague enterprise implementations". Even though the software has become smarter, now fulfilling the requirements they lacked in the beginning - like supporting application programming interfaces to enterprise resource planning (ERP) or company specific legacy systems, handling lots of file formats or integrating the document managing into business process execution - many companies, especially small and medium sized ones, still do not have this type of software in place (Tanner, 2009). There have been similar experiences in the public sector as Fink and Grimm (2007) showed for Germany, Austria and Switzerland. They stress that business process modelling is a pre-requisite for successful ECM implementation which most of the public administrations do not fulfil.
As with respect to the aspects addressed in the following there is no difference between enterprises, public administrations or non-profit organisations, the term 'enterprise' is used to cover all forms of organisations.

---

[2] Information Overload Research Group (IORG). URL: http://www.iorgforum.org/. (retrieved: 12.12.12) The organization was incorporated in June 2008 as a nonprofit corporation by some of the largest IT-companies (e.g. Microsoft, Intel, or IBM) together with representatives of well-known Universities like Stanford or Princeton to conduct surveys and find solutions for the problem.
[3] Basex Information Overload Calculator. URL: http://www.iocalculator.com/ (retrieved: 22.02.10)
[4] In Switzerland for example members of the management board are liable in person for document handling according to the law (Obligationenrecht Art. 957 – 963, Verordnung über die Führung und Aufbewahrung der Geschäftsbücher (GeBüV) vom 24.4.2002, Strafgesetzbuch Art. 110 plus several sector specific laws).

1.1 *Problems*

Basex claims that "According to our latest research Information Overload costs the U.S. economy a minimum of $900 billion per year in lowered employee productivity and reduced innovation. Despite its heft, this is a fairly conservative number and reflects the loss of 25% of the knowledge worker's day to the problem. The total could be as high as $1 trillion" (Spira, 2008). Comparable results have been identified for Switzerland: In Switzerland annual working time is on average 1920 hours and labour costs are 65 CHF per hour. Assuming, that 18% of the working time is used for searching, costs of CHF 22.000 per employee per year occur[5]. If it would be possible to reduce the time for searching to 10%, more than CHF 10'000 per employee per year could be saved. However, as Sieber (2009) investigated, less than 10% of the enterprises already have a solution in place that covers all document creating sources, including document formats[6].

These figures show a huge potential of cost savings by getting enterprises' or public administrations' documents under control. Although no accordant data is available for the public sector, information and knowledge management is ranked fifth out of thirteen future research themes identified by eGovRTD2020[7] (Dawes, 2008).

To address the problem metadata is needed but hard to create due to several obstacles. Manual creation is, for example, time consuming and error prone and automatic metadata generation is particularly difficult for non-textual documents.

1.1.1 Documents are not Perceived as Values

Knowledge is regarded as the primary resource of intellectual capital and is, since Drucker (1994), largely recognized as one of the most important sources of organizations' competitive advantage. Nevertheless, information objects (documents) are not perceived as values - or products as Daconta (2007) puts it. Documents are created to meet a business purpose (e.g. issuing a contract to establish a legally binding relationship between partners) but not for later retrieval. As information is not regarded as a *product* no one cares about how to make it available for the *consumers* (e.g. a manager searching a few month later for that very contract) (Daconta, 2007). Thus, little meta data (e.g. stored in Window's document properties) is available for search. Even worth, this data must not necessarily be correct, for example when the author named in a document property is not the creator of the document but of a document, which has been taken as template.

However, even if the perception of documents changes, manual metadata creation is too cost intensive with respect to human effort and time and is error prone (Albassuny, 2008). Also a huge amount of documents leads to what Liddy et al. (2002, p 401) call the "human metadata-generation bottleneck". Doctorow (2002, p 3) puts his finger on it: "People are lazy".

1.1.2 Huge Variety of Documents

Enterprise and Public administration documents vary in subject (e.g. reports, minutes, letters, applications etc.), types (data, image, sound, text etc.), and formats (file formats like pdf, doc or application specific formats). Hence full-text indexing is not sufficient (as limited to text

[5] The calculation is taken from a report published by Dr. Pascal Sieber & Partners AG and translated into English by me (Sieber, 2009).

[6] (Sieber, 2009), figure 11, p 30. The survey has been conducted with 233 representatives of Swiss enterprises and Public administrations.

[7] eGovRTD2020 is a project funded by the European Commission during 2006-07, to analyze the current status of E-Government research internationally and to develop future research themes based on a comparison of current status with visionary future scenarios.

and string comparison), metadata is needed to provide a structured description of resources (Hatala & Forth, 2003). Metadata can significantly improve resource discovery by helping search engines and people to discriminate relevant from non-relevant documents during an information retrieval operation (Greenberg, Spurgin, & Crystal, 2005). With respect to non-textual documents metadata does not only *improve* the search but in fact, it is mostly the only way *to allow for* search. As stated already by Anderson & Pérez-Carballo (2001), automatic indexing of multi-media documents like images is still usually experimental. Mipai[8], the similarity search engine, provides for example a similarity-based search, using a combination of the five (automatically created) MPEG-7 visual descriptors provided by the CoPhIR collection[9]. SAPIR[10] uses spatial information (GPS data) for similarity search. But neither the low-level features used by Mipai nor the GPS data provide any content related information. 'Squiggle'[11], a Semantic Search Engine for indexing and retrieval of multimedia content tries to meet the need. However, pre-requisite for Squiggle is that "resources [are] already annotated with keywords" (Celino et al. 2006, p 7).

Also paper documents add to the problem: Often there are huge cabinets of paper documents not accessible through any system, for example for contract managing. Scanning makes them electronically available but not searchable. With methods for optical character recognition (OCR) the image (e.g. of a contract) can be transferred into machine readable text - now searchable but still not integrated in enterprises' applications, like a Client-Relationship-Management system. If these documents are well structured or even contain interpretable codes (e.g. a barcode) metadata creation can be partly be automated but for unstructured text they have mainly to be indexed manually.

1.1.3 Commercial Products cannot get through

Despite the great variety of commercial products offered for ECM - in the European market alone there are more than fifty Document Management Systems (DMS) available (Zöller, 2006). Sieber (2009) found that less than 10% of the enterprises already have a solution in place that covers all document creating sources, respectively document formats. Despite the various surveys, there is still no clear picture of *why* enterprises do not have an Enterprise Search solution implemented[12].
For example: 45% of the 233 representatives Sieber (2009) surveyed, answered 'the amount of investment', 42% answered 'nothing', 12% 'not seeing an added value' and 6% stated having 'no need'. Whereas the importance of search solutions for electronic documents is ranked top, with 5.5 points out of a 6-point-scale (ibid.) DMS functionality *is* appreciated when it is part of the business application the user uses in daily work, e.g. a Client Relationship Management system[13] and he/she must not spend an additional effort on managing the documents.

Hence, although DMS improve document management, for example by providing metadata for search, the effort for creating them is perceived greater than the benefit.

---

[8] MiPai Similarity Search Engine.URL: http://mipai.esuli.it (retrieved: 27.5.2010)
[9] The CoPhIR (Content-based Photo Image Retrieval) Test-Collection has been developed to make significant tests on the scalability of the SAPIR project infrastructure (SAPIR: Search In Audio Visual Content Using Peer-to-peer IR) for similarity search. URL: http://cophir.isti.cnr.it (retrieved: 27.5.2010)
[10] URL: http://sapir.isti.cnr.it (retrieved: 27.5.2010)
[11] URL: http://swa.cefriel.it/Squiggl (retrieved: 12.12.12)
[12] For example: 45% of the 233 representatives answered 'the amount of investment', 42% answered 'nothing', 12% are 'not seeing an added value' and 6% stated having 'no need' ((Sieber, 2009), figure 15, p 34).
[13] As survey conducted by Bdails for example, reveals strong demand for CRM-integrated document management. URL: http://bdaily.co.uk/technology/27-04-2012/survey-reveals-demand-for-crm-integrated-document-management/ (retrieved: 18.10.2012)

As an alternative to ECM systems, tools for desktop search (e.g. Google[14], Beagle[15], Strigi[16] or Meta Tracker[17]) could be considered, supporting various file formats, by either indexing text or making use of document properties. However, none of these tools support what Forrester's Principal Analysts claim, that to meet evolving needs and to drive broad adoption of ECM, information and knowledge management (I&KM) professionals must help their enterprises understand business context, i.e. how business people and business processes use content (Le Clair & Poore, 2008).

1.1.4 View on Documents is not Process-related

A document is created, used, read and updated during the execution of business processes. The creator of a document is a participant in a process, which can be a business actor (e.g., a company, company division, or a person) or a business role (e.g., an employee, a client) that controls or is responsible for a business process (OMG, 2011a). Although documents are closely related to business processes as they provide the knowledge needed to perform the process, view on documents is not process-related. A recent study of zhaw School of Management and Law shows that in 33% of the investigated companies less than a quarter of the business processes are documented (Minonne et al. 2011, p 27)[18]. However, if business processes are not explicitely modelled constituting enterprise objects and their relations remain hidden. Hence, the notion of a document is not of an enterprise object related to other enterprise objects but as an isolated item, at most related to other documents or grouped in folders. Since relations to the aforementioned enterprise objects remain implicit they can not be used to infer (meta) data about the document, for example the tasks for which the creator of a document is responsible for.
According to Smith & Fingar (2003) the third wave of Business Process Management is not about business-process re-engineering, enterprise application integration, workflow management but is the synthesis and extension of all these technologies and techniques into a unified whole. Therefore a powerful enough representation of enterprise objects is needed.

1.1.5 Enterprise Architecture Description is not Machine-understandable

An enterprise object is considered any entity that is part of an enterprise, like a business process activity, a compliance requirement, a manufactured product or a document, regardless of its representation. As Hinkelmann et al. (2010) show, an Enterprise Architecture (EA) is often used for describing and managing information about enterprise objects and their relations. In the ISO/IEC/IEEE 42010[19] standard 'architecture' is defined as

[14]Google Desktop Search. URL: http://desktop.google.com/features.html (retrieved: 22.02.10)
[15]Beagle is an open source search tool running on UNIX or UNIX based operating systems. URL: http://beagle-project.org/Main_Page (retrieved: 22.02.10)
[16] Strigi is a daemon which uses a crawler that can index data on a hard disk. URL: http://strigi.sourceforge.net/ (retrieved: 22.02.10). Strigi runs on several operating systems, including Windows.
[17] Tracker is a search engine for UNIX or UNIX based operating systems, that allows the users to search for their files and search for content in their files, too. URL: http://projects.gnome.org/tracker/ (retrieved: 22.02.10)
[18] The aim of the empirical study was to evaluate the maturity and degree of diffusion as well as timely and medium-term trends in the German speaking countries (Germany, Austria, Switzerland). The survey was conducted between November 2010 and February 2011. A total of 219 representatives from more than 200 companies and institutions, which have BPM expertise and implement it in daily business, have participated in this survey (Minonne et al., 2011).
[19] The ISO/IEC/IEEE 42010 is an International Standard entitled, Systems and software engineering — Architecture description. The Standard was published in 2011 and is the result of a joint ISO and IEEE revision

> "fundamental concepts or properties of a system in its environment embodied in its elements, relationships, and in the principles of its design and evolution" (DSCI, 2012).

This definition implies the notion of an architecture as a *conception of a system*, and hence it may exist without being written down. If it *is* documented in an artefact, which can be a text document, graphical notations of a business process etc., this is called an architecture *description*.

> "An architecture *description* is what is written down as a concrete work product. An architecture description (AD) expresses the architecture of a system of interest. An AD could be a document, a repository or a collection of artifacts used to define and document an architecture" (DSCI, 2012).

However, even if the architecture *is* described in an enterprise, in general two important drawbacks can be identified: the architecture description is not represented in a 'machine understandable way' and it is detached from concrete components (e.g. applications and data), used on operational level, as shown by Hinkelmann et al. (2010). Thus, a lot of information available in an enterprise is used in an isolated way. Records of employees are stored in an HR managing system, data of Workflow instances is stored in a Workflow Management System, customer records in a Client Relationship Managing system and documents are stored in file system or ECM system. Since their structure and inter-relations are not 'machine understandable' represented in the enterprise architecture description, they can not be used in a broader way as shown by Hinkelmann et al. (2010). However, far too little attention has been paid to use an enterprise architecture description on operational level, for example for automatic metadata generation.
There is a need for an approach to generate metadata automatically for electronic business documents in an enterprise regardless of their format. Metadata generation should be performed un-supervised and combined with applications, used for daily business. Having business objects represented in a well-defined and machine processable way would allow for rectifying and enhancing information about documents based on its context. Therefore business objects must be represented in a unified, machine-processesable way, interlinked and linked to operational data.
Thus, an approach is needed for:

- all business-related documents regardless of their type (text, image, video, audio) (❶)
- determining the context, i.e. enterprise objectes related to documents (❷)
- representing enterprise objects in an enterprise architecture description (❸)
- using the context of the documents, represented in this enterprise architecture description, for automatic metadata generation (❹).

Figure 1 illustrates the scaffolding for automatic format-independent metadata generation.

---

of the earlier IEEE Std 1471:2000, IEEE Recommended Practice for Architectural Description of Software-Intensive Systems (DSCI, 2012).

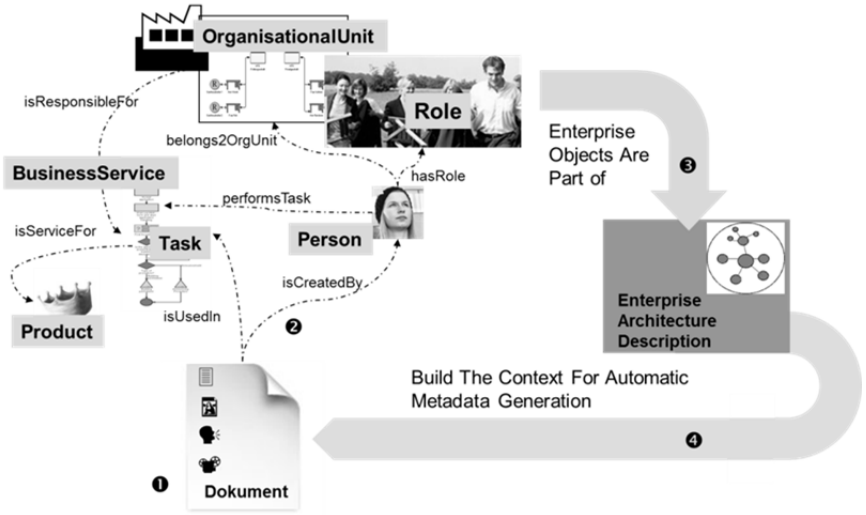


Figure 1: The Scaffolding for Automatic Metadata Generation

1.2 Thesis Statement

The thesis is that metadata for enterprise documents can be generated automatically and format-independent based on semantically enriched context information.

In light of the above presented problems and observations the thesis of this work is that metadata can be generated automatically and format-independent for all kinds of documents of an organization, based on comprehensive, formalized and semantically described enterprise objects and their relations constituting enterprise architecture. Enterprise objects, related to a document are considered the document's context. Thus, the context a document is created for or used in can be inferred to create metadata. Starting point for automatic, format-independent metadata generation is harvested metadata (e.g. from file properties). Format-dependent methods can be added, for example information extraction for text documents. This information is used to infer more metadata from context, e.g. content related metadata like associated business objects or administrative metadata like the retention period.
Based on the (automatically generated) metadata, document management can be improved, especially document lifecycle management, and manual labour can be kept low (e.g. for refining metadata elements). By regarding documents as representations of and linked to enterprise components, changes of those components can actively trigger changes of the related documents, respectively its metadata.

1.3 *Research Objectives*

Documents, created and used in enterprises, are strongly related to other enterprise objects, like a business process activity, a compliance requirement or a manufactured product. These enterprise objects can be considered as a document's *context.*
Figure 2 gives an example of some enterprise objects building a document's context. The figure immediately shows that some context information is directly related to the document whereas other must be inferred. The solid black arrows indicate context information that is directly related to the document, like the creator, the creation tool or the activity of a business process, that the document is created in. Greenberg et al. (2005) call this 'descriptive metadata'. Some of the direct context information is generally provided by the creation software when a document is stored, e.g. author's name, type of creation tool. Greenberg et al. (2005) call this 'system properties'; I will call these properties ‘document properties’. The

dotted lines indicate additional context information that can be inferred, like the organisational unit the creator belongs to, his role in the business process and the goal an activity or process has. Whereas the document properties are unconditionally withdrawable, the availability of other information, like the task for which a document is created, depends on an enterprise's description of their business processes, i.e. the description of the enterprise's architecture.

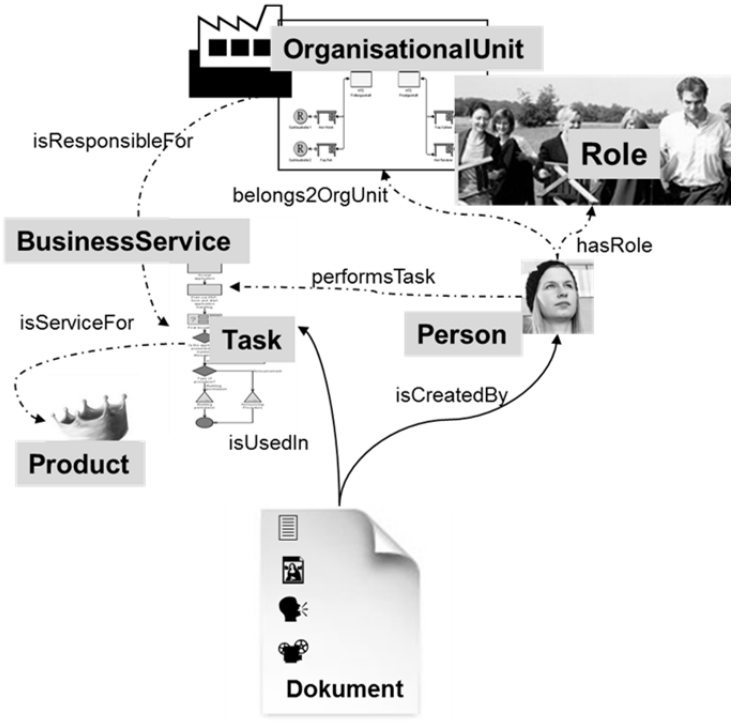


Figure 2: A Document's Business Context

The aim of this research is to provide an approach for automatic metadata generation for any type of an enterprise's document based on its business context. For that purpose four objectives have to be achieved:

(1) To generate format-independent metadata, the context of the enterprise documents is used.
(2) To use the context of a document for automatic metadata generation an Enterprise Architecture is described formally.
(3) To use an enterprise architecture description actively its objects are linked to concrete enterprise components.
(4) To establish automatic metadata generation in an enterprise, a procedure is provided for setting-up, conducting and utilizing metadata.

Figure 3 gives an overview of the research objectives leading to the thesis. In the following I describe the objectives and derive research questions related to each of it.

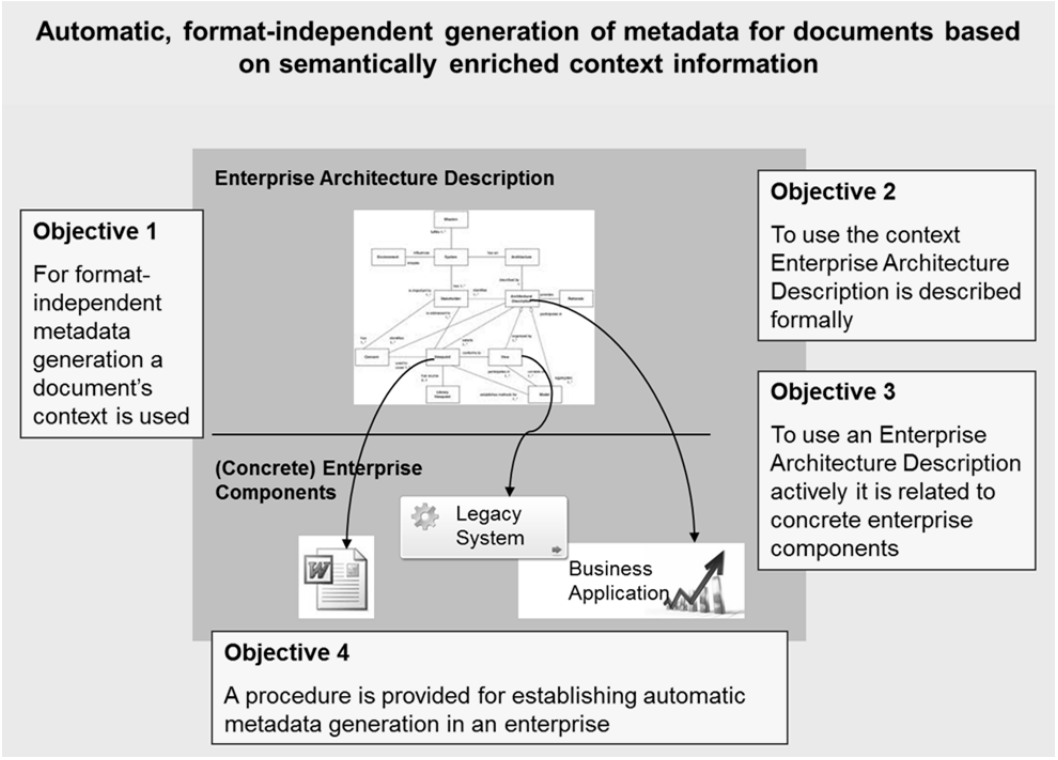


Figure 3: Research Objectives

Objective 1 To generate format-independent metadata, the context of the enterprise documents is used.

Following Winograd (2001), who defines context as an "operational term: something is context because of the way it is used in interpretation, not due to its inherent properties", enterprise objects related to a document are regarded as a document's *context.*

Thus, context of a document can be other documents, e.g. belonging to the same folder, a task in with it is created or an employee, who is the creator. In theory, every other enterprise object can become a document's context. Enterprise objects can be represented in various ways, e.g. in a database record, a video file and as a concept in an ontology. An enterprise repository is considered the entirety of explicitly represented information available in an enterprise. Making use of the whole enterprise repository then allows for relating enterprise objects regardless their representation and hence, retrieving the concrete employee's name from the ERP-System as a document's creator.

More general: related models can be regarded as the *context* of the model in focus. *Context* is considered 'everything that is not '*text'*, that is not *content* of the focused model (Hinkelmann et al. 2010). In a process model, for example, a database model, holding application data or a document created during an activity, builds its context. In contrast, for the very document the process activity is the context. Hence: what content is and what context is, are determined by their use.

Related research questions are:

- What metadata elements are important?
- What determines a document's context?

- What context entities (that is enterprise objects) can be inferred automatically to generate descriptive metadata (e.g. subject, description, coverage, or relation)?
- What harvested and extracted metadata can be used as source for metadata generation?
- What rules must be defined for metadata generation using logical inferencing?
- How to use automatically generated metadata for document management?
- How to measure quality of generated metadata?

Objective 2 To use the context of a document for automatic metadata generation an Enterprise Architecture is described formally.

As explained, the context of a document can be manifold. Usually this information is stored in different information sources. However, the various information items are not independent but related. For example, the author of a document is part of a project team, that team has a common project goal, has specific know-how and so on. These relations are important in order to derive metadata and thus have to be made explicit. Enterprise Architecture defines the structure of elements constituting an enterprise and their inter-relationship. To capitalize on enterprise architecture, it should be described in a 'machine understandable' way in an ontology. The constituent enterprise objects and their relationships are described semantically to be inferred automatically. To differentiate between ontological and non-ontological representations of enterprise objects, the latter are called enterprise components.

With this the enterprise architecture description can be exploited for automatic, format-independent metadata generation. However, enterprise architecture is not specifically modelled for metadata generation but specifically used for it.

Related research questions are:
- What enterprise objects constitute enterprise architecture?
- How to represent enterprise architecture in a way that a machine can process them?
- How to structure enterprise architecture?
- How to describe enterprise architecture in a way that it is general enough to be used ‘out of the box’ but customizable to specific companies’ needs?

Objective 3 To use an enterprise architecture description actively its objects are linked to concrete enterprise components.

Enterprise objects, like information objects, organisational units, business processes or manufactured products and their relations are part of an enterprise architecture description.

Although Enterprise Architecture Frameworks - like the one of Zachman (1987) - help to make those relations transparent, they fall short when it comes to concrete models. Hinkelmann et al. (2010) has shown that objects and relations defined in the EA models cannot be used on the implemented data level

On an operational level, enterprise objects are implemented as applications and data (generally called enterprise components):
- For the employees mentioned in the organization structure there exist records stored in the HR module of the ERP system.
- Data about clients can be stored in a Client Relationship Management System.

- Data of Workflow is stored in the Workflow Management System.
- Information is captured in documents scattered on a file server and so on.

To use an enterprise architecture description actively, e.g. for automatic metadata generation, its elements are linked to these concrete enterprise components. Semantically enriched enterprise architecture description plus the concrete enterprise components is called Enterprise Repository (ER).

Related research questions are:

- Where is the boundary between enterprise objects represented in a semantically enriched Enterprise Architecture (i.e. in an ontology) and enterprise components. For example: what employee's data must be in the enterprise architecture description to allow for reasoning and what data is stored in the ERP system?
- How to avoid redundancies?

Objective 4 To establish automatic metadata generation in an enterprise, a procedure is provided for setting-up, conducting and utilizing metadata.

To set-up, conduct and utilize automatic metadata generation with sufficient quality a procedure is needed to consider enterprises specific needs, customize domain specific parts accordingly but at the same time provide as much 'out of the box' as possible.
Related research questions are:

- What procedure model is appropriate for setting-up, conducting and utilizing metadata?
- How to incorporate and use enterprise specific knowledge, for example glossaries or filing plans?
- How to represent metadata to meet standards and enterprise specifics alike?
- How to improve quality of the metadata generation process?

The combination of the four objectives as stated above makes the approach, I call *mintApproach*[20], unique.

1.4 Course of Action

In the following the way forward to reach the defined objectives is briefly introduced. The applied research methods are described in detail in Chapter 2, page 16 ff.

In order to identify document forms and formats a representative study, conducted within the MATURE project[21] is used. The study investigated what document creation tools are used in enterprises, and thus allows for identifying the prevailing document formats.

To determine an enterprise document's context and gather requirements for metadata generation a survey is conducted with representatives of approximately 30 organisations (public administrations, enterprises of all sizes and non-profit organisations).

[20] The name is derived from Meta, which is a girl's name. In the Baltic language ‘meta’ has the meaning mint. The name has its roots in the pagan culture expressing the desires for the healing powers of the mint for a newborn. URL: http://de.wikipedia.org/wiki/Meta_(Vorname) (retrieved: 19.10.2012)
[21] MATURE is an EU funded project in which FHNW is consortium partner. An overview of the study is available on the project's web-site. URL: http://mature-ip.eu/files/Representative_Study_Info_en.pdf (retrieved: 12.12.12)

To define the scope of the enterprise architecture description and continuously verify and improve results, Action Research studies are conducted and an executable prototype is developed. Action Research Method is also applied to analyse the practical concerns of people in an enterprise and to test the procedure for setting-up, conducting and utilizing metadata.

Course of action is completed by the corresponding state of the art in research (cf. Chapter 3, 25 ff.)

1.5 Underlying Assumptions

Generated metadata is described according to the Dublin Core Metadata Initiative (DCMI) standard. Although DCMI standard has been developed to describe web resources and not enterprise documents it is the most used standard in non-library environments according to Greenberg et al. (2005).[22] Dublin Core can be refined for specific enterprise requirements. Thus, to reflect enterprise specific metadata definitions, qualified Dublin Core is used and where necessary refined or enhanced in metadata profiles.

So as not to reinvent the wheel, available tools are incorporated in the application architecture, such as tools for metadata harvesting, or ontology management. Also considered are open source tools that can be adapted or enhanced if necessary.

1.6 Delineation and Limitations

The work provided focuses on automatic metadata generation based on semantically enriched context information of an enterprise. Therefore documents regarded are those created or used within an enterprise for business purposes. Not considered are web resources (e.g. social sites like blogs, wikis or forums). Excluding those information sources allows for concentrating research on document formats, mainly used in enterprises[23] and neglected up to now, as Greenberg et al. (2005) state in the Final Report for the AMeGA (Automatic Metadata Generation Applications).

In addition to metadata harvesting, document format dependent approaches are available. For text documents metadata extraction could be performed, and text mining and analysis techniques and tools, e.g. for stemming, named entity recognition in text documents could be performed. In case of photos, video or audio file tools for feature detection can be used. Since these approaches are not format-independent they are considered complementary but out of focus of my work. Same holds true for manually created metadata, which is excluded from the mintApproach as metadata should be generated without human labour.

The work provided follows largely the recommendations for 'Automatic Metadata Generation Applications, Version 1.0' (Greenberg et al., 2005). The developed metadata generation application is described according to the 'DCMI Tools Application Profile' suggested by Greenberg & Severiens (2007) (cf. Chapter 7.3.1, p 220 ff.)

[22] As shown in table 14, page 23 of the AMeGA report, DC simple or qualified is used in almost two thirds of non-library environments.

[23] According to the MATURE representative study a significant part of documents, used in an enterprise are created with so called office software (either Microsoft products or open source) (Barnes et al., 2010). 'Endnote' for example (although investigated in the AMEGA report) is not a software usually used in an enterprise, and therefore not considered.

As quality evaluation/quality control (QC) of metadata is regarded as a separate task from the metadata creation (Greenberg et al., 2005) such functionality is not part of the automatic metadata generation. It is a separate operation that can be added.

The same holds true for metadata preservation. There is considerable work done on how to preserve documents and metadata, for example by The Preservation Metadata Implementation Strategies (PREMIS) Working Group, who developed a Data Dictionary for Preservation Metadata. The PREMIS Data Dictionary version 1.0 (PDD) was published in May 2005 a number of repositories have been built since then. In 2007 Deborah Woodyard-Robinson, commissioner of the Library of Congress, as part of the PREMIS maintenance activity, and her colleagues published a survey about PREMIS's interpretation and application (Black et al., 2007). That comprehensive report could provide a good foundation for further work on metadata prevention, but is not the topic of my work.

1.7 Rationale

Describing documents with metadata has a long tradition in libraries but is widely neglected in enterprises. Manually assigning metadata is costly and error prone and in the case of non-library organisations seldom done. As shown in the Automatic Metadata Generation Applications (AMeGA) report (about a survey with 217 participants), automatic metadata generation can help to address efficient and less costly metadata creation (Greenberg et al., 2005). However, whereas feasibility and usefulness for automatically generated 'technical metadata' (Gilliland & Baca 2008, p 9), like date, format or language scaled high (in average 2.5 out of 3), descriptive metadata like subject or description got low marks (in average 1.8 out of 3). To reach higher scales for such metadata participants were recommended to "consider context" and to "import … context-sensitive information from the authoring environment" (Greenberg et al. 2005, p 28). In addition participants indicated the importance of developing automatic methods of generating metadata for nontextual resources because of the absence of text for indexing (Greenberg et al., 2005)[24].

"The results [of the AMeGA project] demonstrate that content creation software supports metadata generation and can provide an important data source for automatic metadata generation applications" and further "Employing automatic techniques to enhance and/or refine metadata will improve the quality and overall functionality of the metadata" (Greenberg et al. 2005, p 52). Hence, harvesting and improving automatically created document properties is taken as a starting point.

Using enterprise context for automatic, format-independent metadata generation for enterprise documents has become only lately research topic. Whereas research has been done for web-resources, learning objects and most recently for multi-media documents in a personal environment, exploiting the context of enterprise documents for automatic metadata generation has been barely investigated. Mitschick (2009) works in her thesis on metadata generation for multi-media documents for personal knowledge management. She used context to enhance metadata automatically using information sources on the web, e.g. for music audio files. Brügmann (2011) investigates the management of unstructured information in an enterprise using semantic metadata. Athough he uses the context of documents for metadata

[24] Page 31, Table 21: 57.3% rate automatic metadata generation for nontextual resources very import and 38.9% somewhat important.

generation, he does not model it explicitely. Thus, in his notion of context remains arbitrary and lacks a sound foundation to build the basis for broader use.

Modlling documents' context as enterprise objects and their relations in enterprise architecture descriptions and representing it formally in an ontology supports a shared understanding (amongst different stakeholders including machines), solves ambiguity and builds a reasonable basis for automatic metadata generation. In addition enterprise objects can be related to concrete enterprise components to make active use of the enterprise architecture descriptions.

1.8 Outline of Thesis

Chapter 1 is an introduction to the topic that comprises problem identification, research objectives (RO 1-4) and related research question, thesis statement, underlying assumptions, delineation and limitations, rationale and the outline of my thesis.

The applied research methodology is explained in **Chapter 2**, introducing the methods and techniques I used. The chapter starts with the applied research design, introduces the methodology and limitations for their use.

Following this State of the Art is investigated, starting from existing approaches of 'metadata generation', continuing with the concept of 'context', followed by work on 'enterprise architecture' and 'enterprise ontologies'. The chapter ends with an overview on related work about metadata generation based on context (**Chapter 3**).

Whereas Chapter 3 gives the scientific basis of my research, in **Chapter 4** requirements gained in praxis are provided. Requirements are engineered based on the findings of a Representative study, results of a survey on document handling in enterprises and findings of the first loop of the Action Research studies with two organizations (Action Research loop 1). The chapter concludes with a specification of requirements for metadata generation that will later be part of the evaluation.

Based on the knowledge acquired from a thorough review of the associated literature and the partical requirements, general models for automatic metadata generation in an enterprise are developed. Thus, in **Chapter 5** the Enterprise Architecture Meta Model, which becomes part of the semantically enriched enterprise architecture description (seEAD), a metadata generation model and a procedure model for setting up, conducting, utilizing and maintaining metadata generation are described.
The general models, introduced in Chapter 5, are verified and improved within the two Action Research studies. Thus, for the two Action Research study partners specific application profiles are developed and implemented in demonstrators (Action Research loop 2). These application specific models are detailed in **Chapter 6**. The chapter closes with a comparison of the two specialisations.

In **Chapter 7** implementation architecture for automatic metadata generation is introduced and its components are described. An overview is given on the general approach, and the tools and techniques used are explained. After the general description the application-specific implementations of the prototype for the two Action Research partners is detailed.

Chapter 8 is about the evaluation of the results of my work. This will be done based on the evaluation criteria defined in the requirement specification, verified with both Action

Research partners and deepened by means of selective interviews in consecution of the survey on document handling in enterprises. The chapter closes with an assessment of the fulfilments of the requirements defined in Chapter 8.3.

My work concludes with **Chapter 9** giving a summary of the findings, and contributions and suggestions for further research.

Finally a glossary is provided in **Chapter 10**, as well as the bibliography in **Chapter 11**.
Details of my work and continuative material are provided in the appendix **Chapter 12**.
Figure 4 illustrates the structure of my thesis.

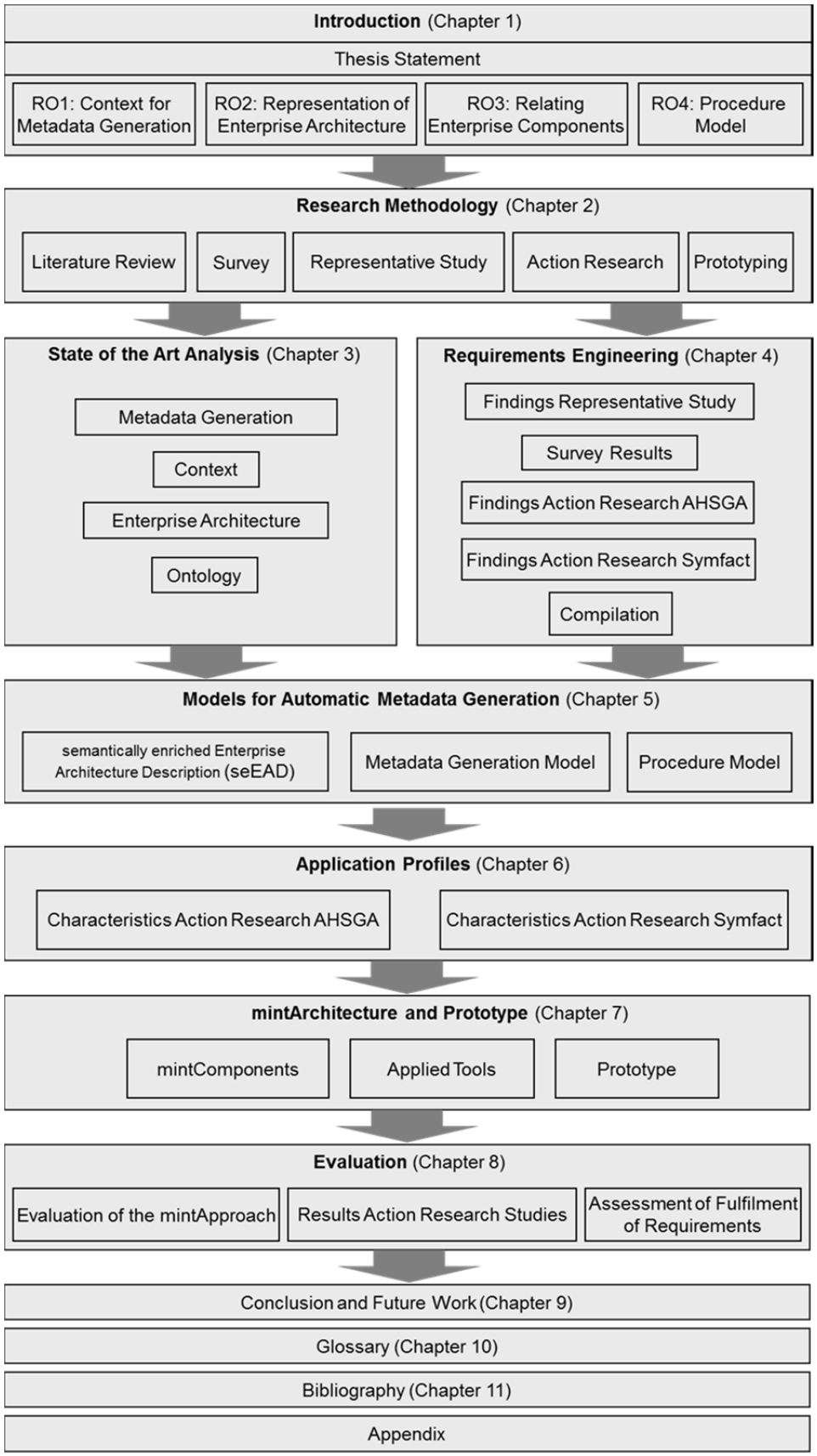


Figure 4: Outline of my Thesis

2 The Method

Following the deductive approach, I developed a theory and hypotheses about automatic format-independent metadata generation based on semantically enriched context information. Purpose of my research is explorative to clarify my understanding of the problem, test my approach and validate the results.
I employed a mixed method strategy, combining the following methods: representative study, survey, Action Research and prototyping.

2.1 *Research Design*

Figure 5 gives an overview of the research design, comprising quantitative research methods in the first phase (lighter coloured) and qualitative research methods in the second phase (darker coloured).

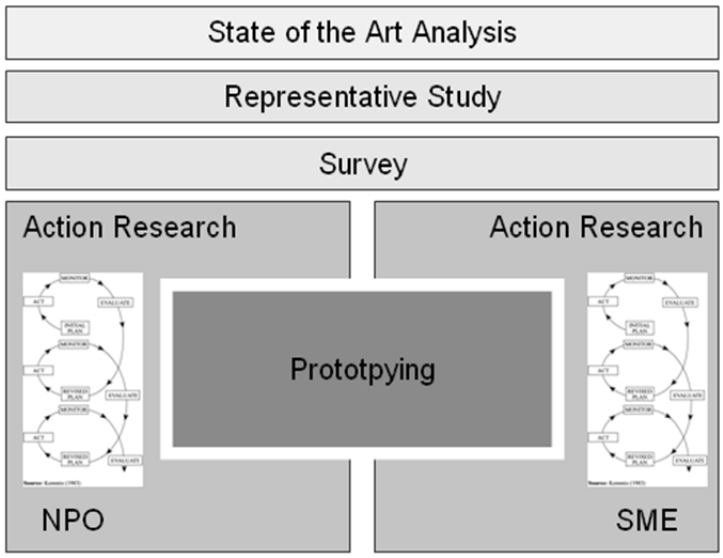


Figure 5: Research Design

After literature review, results of a comprehensive representative study conducted within the MATURE project, are analysed with respect to the use of metadata creation tools in enterprises. Although the study has been focusing on knowledge maturing activities - neglecting other purposes of document use in enterprises - the findings are important for my thesis in respect of the *types of software* used for document creation and the *context* that documents are used in.

Then, a survey is conducted about document handling in enterprises, public administrations and non-profit organisations. Findings from literature review and conclusions drawn from the representative study were translated into a questionnaire about document handling in enterprises. As employees - other than librarians - in general do not consider metadata important (Greenberg, 2004) - the questions asked are kept simple and focus on daily work behaviour. Goal of that survey was not to get a representative overview of document handling in general but to verify significance of my research topic. Hence, an explorative study, based on qualitative interviews was appropriate. The interviews provided insights into basic requirements like metadata elements that matter, and the context of enterprise documents, e.g. determined by the implemented enterprise architecture models.
As the context of documents is the most important element of the metadata generation approach introduced in my thesis, general statements, given in the interviews, are not enough.

Therefore main part of my research design is Action Research, conducted in two enterprises focusing on different aspects of automatic metadata generation. One Action Research study is carried out with AHSGA[25], a NPO with about 10 employees, the other with Symfact AG[26], a small-sized enterprise developing software for contract management. Action Research is favoured over case studies as it emphasizes the *iterative process* of diagnosing, planning and taking actions within a *specific context,* whereas case studies focus on investigating a *contemporary phenomenon in a broader context* (Saunders, Lewis, & Thornhill, 2007). Although both methods involve business users and allow proving theory in practise, Action Research is regarded more suitable as it allows for iterative testing and improving the prototype to be developed focusing on the specific goal of automatically generating metadata.

As shown in Table 1 nearly all research questions can be addressed with the Action Research method. Action Research has become increasingly important for developing information systems since the late 1990s though it has been an established research method in the social and medical sciences for more than five decades (R. L. Baskerville, 1999). Applying Action Research in two enterprises, with very different business and business goals helps to avoid the common pitfalls of this method like subjectivity, lack of generality and replicability, that (Hofstee, 2009) warns about.

As depicted in the prototyping method is used for putting theory into practise. Prototyping is embedded in Action Research as prototypes are designed, developed, implemented and evaluated within the Action Research studies. Conducting Action Research in parallel in two enterprises allows continuous testing of assumptions and drafted solutions. Prototyping comprises modelling, e.g. enterprise objects as well as defining rules (e.g. for knowledge detection and metadata inferencing from context) and implementing procedures (e.g. for automatic metadata generation). Evaluation of existing tools, e.g. for metadata harvesting, ontology creation, natural language processing is also part of prototyping. As part of Action Research, prototypes are validated in two different enterprise environments and for different business purposes. Thus, developments made to meet requirements of one enterprise can be verified in the other and terms of use can be deduced.

Table 1 gives an overview on research methods applied to objectives and research questions. Since Literature Review is assumed for every research question it is omitted in the table.

[25] AHSGA, Fachstelle für Aids- und Sexualfragen. URL: http://ahsga.ch/index.html (retrieved: 15.8.2010)
[26] Symfact Compliance Solutions. URL: http://www.symfact.com/ (retrieved: 15.8.2010)

Objectives and Research Questions (RQ)		Research Methods				
		Repre sent-ative Study	**Survey**	**Action Research (Study NPO)**	**Action Research (Study SE)**	**Proto-typing**
RQ	Objective 1	**x**	**x**	**x**	**x**	**x**
RQ1	What metadata elements are important?		x	1[27]	1	x
RQ2	What determines a document's context?	x	x	1	1	x
RQ3	What context entities (this is enterprise objects) can be inferred to automatically generate descriptive metadata (e.g. subject, description, coverage, or relation)?			1, 2	1, 2	x
RQ4	What metadata can be harvested and extracted to be used as source for metadata generation	x	x	1	1	x
RQ5	What rules must be defined for metadata generation using logical inferencing			2	2	x
RQ6	How to use automatically generated metadata for document lifecycle management?			2	2	x
RQ7	How to measure quality of generated metadata?			3	3	x
	Objective 2			**x**	**x**	**x**
RQ8	What enterprise objects constitute enterprise architecture?			1, 2	1, 2	x
RQ9	How to represent enterprise architecture in a way that a machine can process it?			1, 2	1, 2	x
RQ10	How to structure enterprise architecture?			2	2	x
RQ11	How to describe enterprise architecture in a way that it is general enough to be used 'out of the box' but customizable to specific companies' needs?			2	2	x
	Objective 3			**x**	**x**	**x**
RQ12	Where is the boundary between enterprise objects represented in a semantically enriched Enterprise Architecture (i.e. in an ontology) and enterprise components?			3	3	x
RQ13	How to avoid redundancies?			3	3	x

[27] The number indicates in which loop(s) that question is addressed. Same research question can be addressed in several loops.

Objectives and Research Questions (RQ)		Research Methods				
		Repre sent-ative Study	**Survey**	**Action Research (Study NPO)**	**Action Research (Study SE)**	**Proto-typing**
	Objective 4			**x**	**x**	**x**
RQ14	What procedure model is appropriate for setting-up, conducting and utilizing metadata?			2, 3	2, 3	
RQ15	How to incorporate and use enterprise specific knowledge, for example glossaries or filing plans?		x	1	1	x
RQ16	How to represent metadata to meet standards and enterprise specifics alike?			1	1	x
RQ17	How to improve quality of the metadata generation process?			3	3	

Table 1: Correlation Between Objectives and Research Methods

2.2 *Methodology*

To validate my thesis statement four research methods are integrated: representative study, survey, Action Research and prototyping. In the following each method is briefly explained along with research techniques used to apply the methods.

2.2.1 Representative Study

In year two of the MATURE project a representative study was conducted between 1.4.2009 and 1.4.2010. All project partners were involved in the study, and FHNW conducted 10 in Switzerland.

Objective of the study was to explore current knowledge maturing practices in medium and large sized enterprises empirically. In total, 139 interviews were conducted in this study. Although the focus was on knowledge maturing and not particularly on document management, results of the study contribute to answer two important research questions, regarding documents' forms and formats, and context.

The representative study provides data on the use of software for document creation and managing. From this I derived the prevailing document forms and formats and the context, documents are used in.

The purpose of the study for my research was to reassess practical requirements I gained from the survey and the Action Research studies on a broader basis.

2.2.2 Survey

The survey strategy, as well as the representative study introduced above, is associated with the deductive research approach. 30 structured interviews were conducted to reach a significant number of inquiries for reliable results.
Purpose of the survey was to find out people's opinion, desires and attitutes regarding document creation and storage, metadata types, creation and use, rules and regulations for document handling (e.g. naming conventions, documentation) and document retrieval strategies in their enterprise.

The survey was developed

- to evaluate document handling in daily routine by non-information specialists
- to verify the analysis of the representative study with respect to distribution and usage of document creation tools and standardization approaches
- to determine context of document handling with respect to business process management and governance instruments
- to find out search strategies and
- to identify the most important metadata elements, based on the Dublin Core Metadata Element Set (Dublin Core Metadata Initiative, 2012).

The survey was conducted in face-to-face interviews based on a standardized questionnaire (see Appendix 12.3), comprising a total of 20 questions, thereof 18 closed questions and 2 open questions. All questions were available in German and English.

Interviews were conducted in 4 countries: Switzerland, Germany, Austria and Italy. To prove that problems with document handling are not sector specific, wide coverage of different sectors and sizes of enterprises were chosen. Out of the thirty interviews eight were conducted with representatives of micro- and small sized enterprises (MiE and SE), one of a medium-sized enterprise (ME) and three of large-sized enterprises (LE). Five enterprises are classified as Non-Profit-Organisations (NPO). Eight representatives of Public administrations on the federal levels of Canton and Municipality were interviewed as detailed in Table 2:Survey Overview.

Country	total	Enterprises					Public administration		
		MiE	SE	ME	LE	NPO	State	Canton	Municipality
CH	25	7	1	1	3	5	3	4	1
DE	3	1	2						
AU	1	1							
IT	1	1							

Table 2:Survey Overview

Complementing the representative study, the survey focussed on small and micro-sized enterprises. This was due to the assumption that enterprises of these sizes can't or won't afford specific software for document management but make do with storing documents on a file server using the explorer for management.

2.2.3 Action Research

"The underlying philosophy shared by most forms of Action Research is pragmatism. As a philosophy, pragmatism concentrates on asking the right questions, and getting empirical answers to those questions" (R. L. Baskerville & Myers, 2004). 'Action Research' is research

in action, concerned with an organisation's issues, involving practitioners and performed over a certain period with iterative steps (diagnosing, planning, taking action, evaluation), aimed at linking theory to practise (Saunders et al., 2007).

According to Cronholm & Goldkuhl (2004), when Action Research is performed there is collaboration between business practise and research practise. They call it the "business change practise/empirical research practise" (ibid.), building the intersection of the two practises.

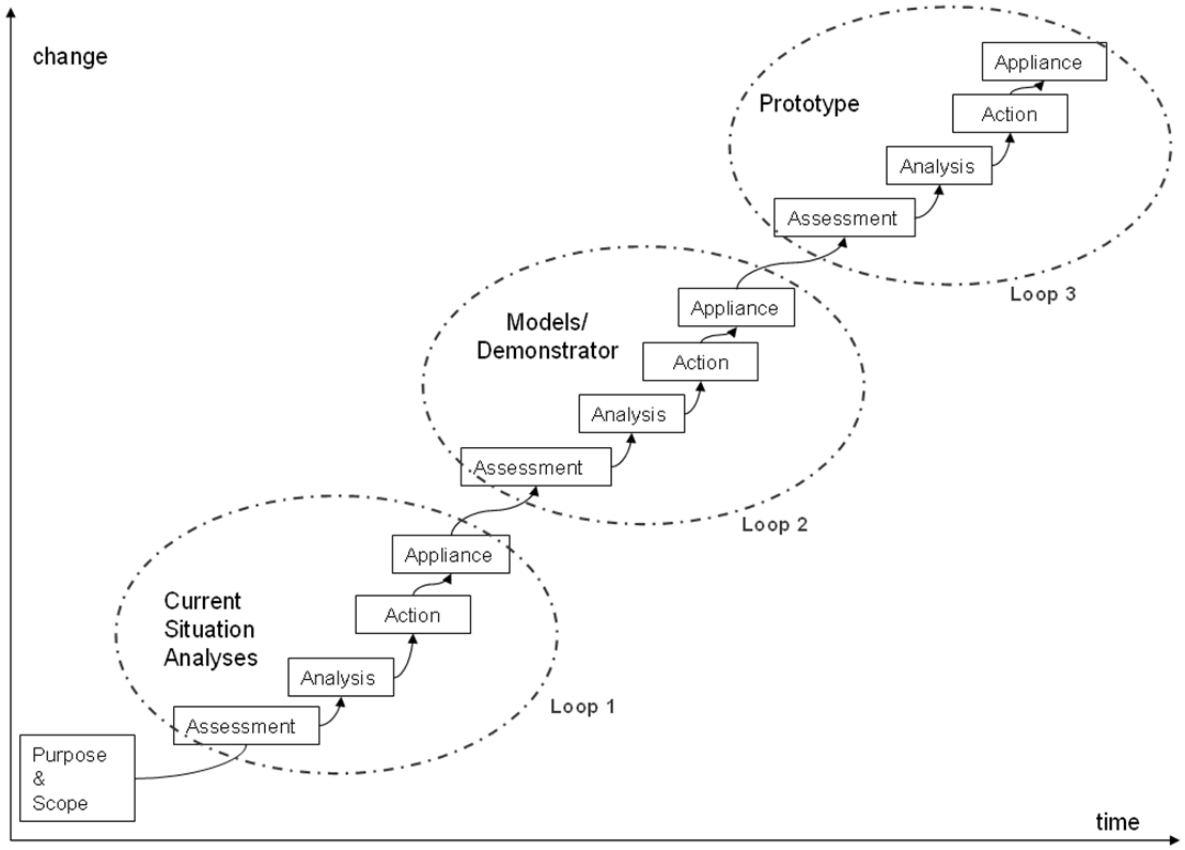


Figure 6: The Action Research Model (based on Saunders et al. 2007)

Figure 6 depicts an adapted 'Action Research Model' including the dimensions 'changes' and 'time'. In each loop of the model, scientific assumptions are verified for correctness, completeness and practical relevance (Assessment), results are evaluated (Analysis), measures are taken (Action) and introduced to business (Appliance).

However, there is no single, monolithic instrument to be used for 'Action Research' but a set of techniques that can be used. Baskerville (1999) gives an overview on "IS [Information System] Action Research Forms and Characteristics". He regards IS prototyping as a form of 'Action Research'. Following that opinion, 'Prototyping' is embedded in 'Action Research' as depicted in
Figure 5.

For empirical investigation of my research theory within its real life context, multiple sources of evidence are used.
Table 3 gives an overview of research instruments used for Action Research in my thesis.

Research Instrument	Purpose	Loop 1	Loop 2	Loop 3
In-person interviews	to get in-depth knowledge of business peoples' opinion	x	x	x
Data analysis	to get concrete examples and proof interview results	x	x	x
Paper study	to get insight on enterprise management, e.g. reading the Management Handbook of an ISO 9001 certified enterprise	x		
Observation	to get cross-check interview results	x		x
Models / Demonstrator	to test theory early in practice	x	x	x
Prototype	to evaluate software development in business			x

Table 3: Applied Research Instruments for Action Research

Loop 1
Tasks of the first iterative cycle are

1. Taking of an inventory on document handling
2. Identify enterprise specific problems with document handling
3. Collect and specify requirements
4. Compare real life situation with a priori statements on the problem
5. Review current research on the problem
6. Formulate a question or questions to be answered
7. Take action
8. Share with others (departmental meeting, publication, conference, etc.)

The cycle is described in Chapter 4.3, p 92 ff.

Loop 2
Tasks of the second iterative cycle are

1. Present and evaluate models or demonstrators
2. Identify questions to be answered
3. Identify change requests and supplementary requirements
4. Take actions to overcome or test the problem and to adapt the demonstrator to enterprise specific requirements
5. Share with others (departmental meeting, publication, conference, etc.)

The cycle is described in Chapter 6.1.6, p 179 ff and Chapter6.2.6, p 201 ff.

Loop 3
Tasks of the third iterative cycle are

1. Present and evaluate an executable prototype
2. Identify change requests and supplementary requirements
3. Identify questions to be answered

4. Plan actions to meet the requirements
5. Share with others (departmental meeting, publication, conference, etc.)

The cycle is described in Chapter 7.3.3, p 223 ff and Chapter 7.3.4, 226 ff.

2.2.4 Prototyping

Prototyping is a well-known system development methodology that comprises design, creation and test of prototypes of (parts of) systems, either to clarify requirements or to verify assumptions. Prototypes can widely differ in form and complexity, depending on their purpose and development stage. Thus, in an early stage they can simply consist of 'paper models' (e.g. visualisations using power point presentations) or 'mock-ups'/'wire-frames' (e.g. linked html-sides to simulate process flow). Later some working functions can be shown in so called demonstrators and further developed. For example, an evolutionary prototype can become a running model or part of a system. Following (Kordon & Henkel, 2003) I consider 'prototypes' to be executable, thus only the artefact developed in loop 3 of Action Research is called prototype whereas in loop 1 a 'model' and in loop 2 a 'demonstrator' is developed. Figure 7 depicts the three artefacts of prototyping.

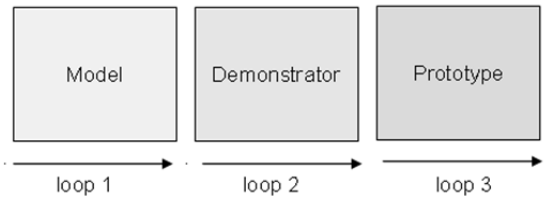


Figure 7: Prototyping Artefacts

Each of these forms of prototyping involves designing and deploying artefacts for the purpose of learning about the artefacts or their environment. The collaborative nature of prototyping and its potential for organizational change makes it useful as one form of an Action Research instrument (R. Baskerville, Pries-Heje, & Venable, 2009). As shown above in each Action Research loop a different form of prototype is used.

2.3 *Limitations*

The representative study had a goal to gather information on knowledge maturing and activities contributing to maturing were investigated. Although use and management of documents was not in the focus of the study it contributed to verify some basic aspects of my work.

Even though the survey on metadata is not representative, consistent answers for many aspects of document handling in enterprises (e.g. 80% of all representatives stated that today, they simply store their data on file servers without considering metadata) legitimates the limited sample size. Arbitrary selection of enterprises, spread over all business sectors and enterprise sizes, adds to the reliability of the results.

As Action Research is performed in two enterprises, very different in their business goals and requirements, hazards like subjectivity, lack of generalizability of results and replicability (Hofstee, 2009) can be reduced.

Using prototyping within the Action Research studies allows for evolutionary development, evaluating each artefact from a business point of view and thus, ensuring that the prototype (GMP) is applicable. However, it will remain a *prototype,* which can not be used one-to-one in productive information infrastructure.

2.4 *Summary of the Method*

In this chapter I introduced the methodology I used for my research. Based on the definition provided by Gomez-Perez et al. (2004) my methodology is composed of a representative study, exploiting it as secondary source, a survey, conducting qualitative interviews, Action Research, performing a collaboration between business practises and research practise and prototyping, applying Action Research results in an evolutionary development of the prototype. Figure 8 depicts the relationship between the methods and techniques I used in my research methodology. At the left side of the figure the chosen methods are shown and at the right side the applied techniques were listed.

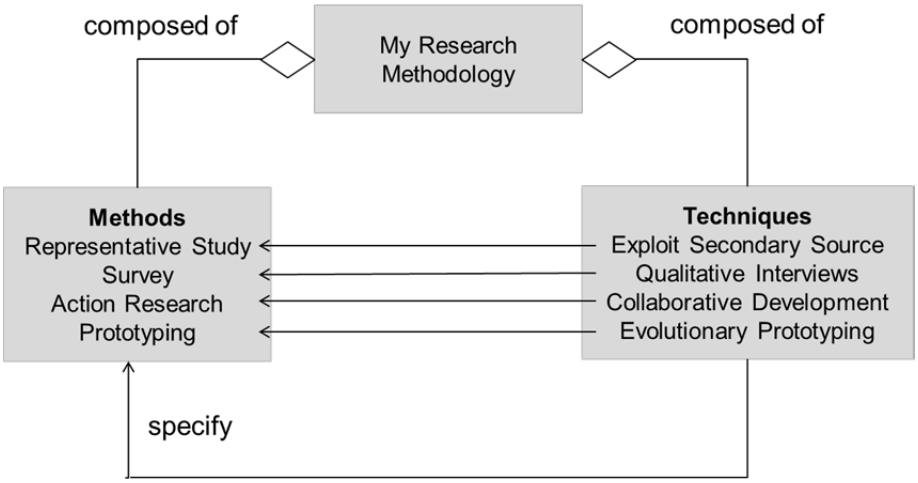


Figure 8: Graphical Representation of Methods and Techniques Used in my Research
(based on Gomez-Perez et al. 2004, p 109)

Taking the outcome of the representative study as secondary source for my research broadens the base for requirements gathering, whereas the survey adds detailed information to it. Within the two Action Research studies, in the first cycle, those general requirements are on one hand verified and on the other hand refined for enterprise specific applications. Doing two Action Research studies in parallel, focusing on different problems and solutions helps to avoid the bias of specialty that this kind of method encourages. My research methodology is completed with an evolutionary development of a prototype, which is evaluated by the Action Research partners but also assessed independently by selected representatives of the conducted survey and vendors of Information Management Systems.

3 State of the Art Analysis

Chapter 3 of my thesis provides the theory base for my work as illustrated in Figure 9.

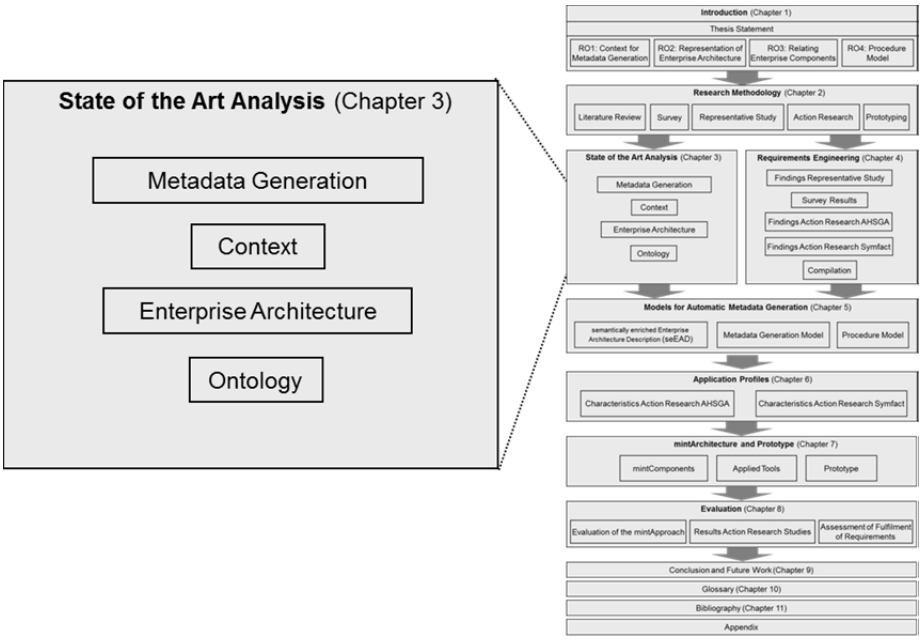


Figure 9: Position of Chapter 3 in the Overall Structure of the Thesis

To provide a sound basis for work on automatic, format-independent metadata generation, research in the following fields has been investigated:

- automatic metadata generation, as manual creation is not feasible
- context, as documents are related to (other) enterprise objects (the author, a product, a customer, etc.)
- enterprise architecture, as enterprise objects are structured and related to each other according to principles which determine its form, function, value, cost, and risk (DSCI, 2012)
- enterprise ontologies, as enterprise architecture descriptions can be represented in a machine understandable way.

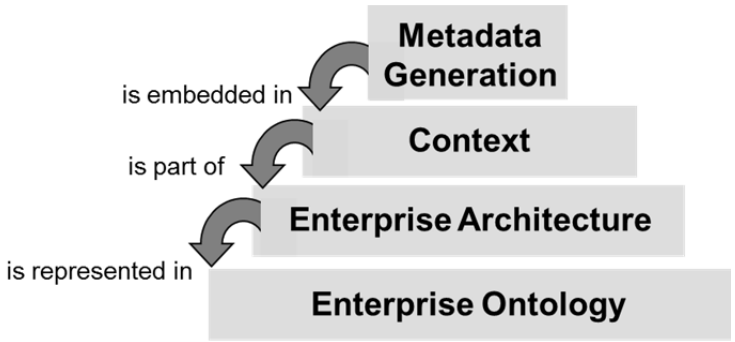


Figure 10: Metadata Generation Pyramid I

The chapter is structured according to the metadata generation pyramid provided in Figure 10. First, work related to metadata generation in general is considered. As documents, created and used in enterprises, are strongly related to other enterprise objects, like a business process activity, a compliance requirement or a manufactured product, the notion of context is

investigated. Context for automatic metadata generation can be considered a segment or a view of a more comprehensive whole that is the enterprise architecture. Therefore work in this field is reviewed. To use enterprise architecture for automatic metadata generation, enterprise architecture description can be semantically enriched and formalized in ontologies. Thus research on enterprise ontologies is analysed. The chapter concludes with a summary of related work and implications on my work.

3.1 *Metadata Generation*

In this chapter review on scientific literature about metadata generation is provided. The chapter is structured as depicted in the figure at the right. After an introduction to metadata generation in general, particular methods are discussed: metadata harvesting, metadata extraction and semantic annotations. Finally a brief overview on metadata standards relevant for my work is provided.

Structure of Chapter 3.1

Metadata is ‘data about data’, i.e. is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource (NISO, 2004). A general introduction to metadata is given by Baca et al. (2008), explaining key concepts, tools and issues concerning metadata.

Gilliland (2008) provides an overview of metadata, its types, roles, characteristics and standards. She emphazises that metadata is crucial for information management and for ensuring effective information retrieval and accountability in record keeping, and stresses, that metadata is particularly important with the increasing use of electronic commerce and e-government services.

The act of producing metadata, is called *metadata generation* and often used synonymously to *metadata creation*. It can be done by humans or machines (Greenberg & Severiens, 2007). Manual creation of metadata is costly, time consuming, and error prone. Thus, time and effort required to manually create metadata for multiple resources, as well as inconsistencies and idiosyncrasies in interpretation, are major obstacles to a widespread adoption of management systems for unstructured information (Hatala and Forth, 2003). Given the massive number of digital resources requiring metadata it is unrealistic to depend on traditional humanly generated metadata approaches (Greenberg et al, 2006). In the last decade, much research on automatic metadata generation has been done, especially in the field of libraries, but not limited to it, and several tools or add-ons to exiting systems have been developed (amongst others by Liddy et al. (2002), Hatala and Forth (2003), Patton et al. (2004) and Phipps et al. (2005).

According to Greenberg (2004), there are two different automatic metadata generation methods: extraction and harvesting. Whereas *extraction* refers to metadata that is extracted from a resource's content (e.g. a term or phrase of a text document), *harvesting* means that metadata is collected from already existing tags, e.g. found in the 'header' of an HTML-resource or added by the operating system to a document. However, in literature one finds both terms used interchangeable.

To emphasize the particular step of inferring additional metadata from context *semantic annotating* is introduced as a third method that complements the other two. The World Wide Web Coalition (W3C) defines semantic annotation as "additional information that identifies

or defines a concept in a semantic model in order to describe part of that document" (Farrell & Lausen, 2007).

3.1.1 Metadata Harvesting

"Harvesting, […] occurs when metadata is automatically collected from META tags found in the 'header' source code of an HTML resource or encoding from another resource format (e.g., Microsoft WORD documents)" (Greenberg 2004, p 63). The glossary of the Dublin Core Metadata Initiative (2007) defines *metadata harvesting* as: "Automatically gathering metadata that is already associated with a resource, and which has been produced via automatic or manual means".
The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) provides an application-independent interoperability framework based on metadata harvesting (OAI, 2008). The OAIMH protocol was designed for supporting interoperability through metadata harvesting. A tutorial on the OAI metdata harvesting protocol is provided by Warner (2001).
Tennant (2004) describes problems when harvesting bibliographic records with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Beisler & Willis (2009) compare transformation and presentation of Dublin Core metadata extracted by the two metadata harvesters OAIster and the Online Computer Library Center's (OCLC) WorldCat from CONTENTdm.

Often *harvesting* is used for metadata *extraction* when talking about web resources, for example by Ciravegna et al. (2004) in their book chapter about "Learning to Harvest Information for the Semantic Web". They describe the Armadillo system that "annotates by extracting information from different sources and integrating the retrieved knowledge into a repository."

Woodley (2008) discusses metadata harvesting of web resources in the light of uniting various sources. The harvested metadata records are to be unified and aggregated in one repository. To reach the goal Woodley (2008) describes crosswalks and metadata mapping techniques. Refer to Chapter 3.1.3 about semantic annotions of web resources.

A comprehensive survey is provided by Albassuny (2008), in which Automatic Metadata Generations Applications (AMeGAs) for web resoureces were investigated. The results show that Meta Tag Generators (MTGs), which generate metadata data from web pages were prevailing, representing almost 90% of the tools. Only three out of 205 investigated AMeGAs are harvesting tools (collecting metadata found in the header source code of an HTML-document).

Research on metadata harvesting in the sense of Greenberg (2004) is not very large and most belongs to the digital library domain. Hillmann et al. (2004) explore options for augmenting harvested metadata of resources of the National Science Digital Library[28] in order to improve metadata quality. Hillmann et al. (2004) identify four problems encountered with metadata harvesting: missing data, incorrect data, confusing data and insufficient data. With their approach Hillmann et al. (2004, p 2) wanted to improve the metadata quality automatically for example by removing harvested metadata with no information value like "unknown" or "n/a". To augment services they retrieve additional metadata from trusted web-sites, e.g. provided

[28] The NSDL is the nation's online portal for education and research on learning in Science, Technology, Engineering, and Mathematics. URL: http://nsdl.org/ (retrieved: 29.11.2012)

by the Eisenhower Noationa Clearinghouse (ENC)[29] for adding information on audience and education level of a resource.

In a research project, which was part of the Master programme Business Information Systems at FHNW, supervised by me, Johner (2011) investigated harvesting tools for documents used in enterprises, e.g. MS word documents. Since the three harvesting tools Albassuny (2008) investigated are for web resources and only semi-automatic (they harvest meta tags and provide them to the user for verification), they weren't considered in the evaluation. Johner (2011) assessed seven harvesting tools, selected on the basis of evaluation criteria derived from Albassuny (2008) and (Ares Casal, Dieste Tubío, García Vázquez, López Fernández, & Rodríguez Yáñez, 1998). The evaluation showed that there is no *one best tool* that fits all requirements. Which tool is most appropriate depends on requirements for document formats, frequency of changes of document formats, technical skills, budget constraints etc. Refer to Chapter 12.1, p 288 for details on the evaluation.

3.1.2 Metadata Extraction

As my thesis is about a generic approach applicable to all kinds of enterprise documents, *metadata extraction* is only partially interesting as it deals solely with text documents. *Metadata extraction* is the process of automatically pulling (extracting) metadata from a resource's content; i.e. resource content is mined to produce structured ('labelled') metadata for object representation (Greenberg et al., 2005).

Metadata Extraction in this sense is closely related to information extraction, a research field for more than two decades. A survey of information extraction research is provided by Sarawagi (2008, p 2) who investigated the field along various dimensions "derived from the nature of the extraction task, the techniques used for extraction, the variety of input resources exploited, and the type of output produced".

Due to the long period of research and the interests of very different research directions manifold approaches for information extraction are available. The development of information extraction approaches have been largely influenced by the Message Understanding Conferences (MUC)[30]; the MUC proceedings provide a reference source to understanding the evolution of information extraction and state of the art (Appelt, 1999).

With the advent of the internet, its ever-growing information source and increase of e-commerce information extraction (IE) has become one of the most active research areas in database and information system research (Weikum & Theobald, 2010). Besides academic research the demand of commercial software developers for identification and extracting business relevant information from textual documents advanced the development (Balke, 2012).

In view of the plethora of approaches for information extraction and its limitation to textual documents I waive the presentation of the various approaches and refer to Appelt & Israel (1999), Appelt (1999) and Tang et al. (2008) for a concise description.

[29] The Eisenhower National Clearinghouse (ENC) now goENC.com, is a contributor in building a "core collection of digital resources in K-12 education". URL: http://www.goenc.com/ (retrieved: 29.11.2012)
[30] "The 'Message Understanding Conference (MUC)' was supported by SAIC during the 1990's. SAIC sponsored the MUC website for a short time after MUC-7. NIST was asked to archive the MUC information on their web site to preserve this resource." URL: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/ (retrieved: 29.11.2012)

However, to complement format-independent metadata generation based on context it is relevant *what* can be extracted from text. These are above all 'named entities'. Named Entity Recognition (NER) aims at finding real-world objects in texts and classifying these objects into predefined categories such as names of persons, organisations, locations, temporal expressions, products, etc. (Balke, 2012). NER can be reached in different ways, ranging from simple look-ups in gazetteers (Cunningham et al., 2011) to statistical models (e.g. hidden Markov models, maximum entropy Morkov models, or conditional random fields), which take into consideration how entities are embedded in phrases, sentences or segments (Balke, 2012).

The business data provider Thomson Reuters provides the OpenCalais Web Service framework[31] which allows submitting unstructured text and getting entities, facts, and events extracted from the text back. "The set of recognized entities ranges from persons, organizations, and locations via structural content like email ad- dresses, medical conditions, or phone numbers, to events like anniversaries, product recalls, or bankruptcies" (Balke, 2012). (Rizzo & Troncy, 2011) developed evaluation framework and assessed five popular extractors (including OpenCalais). The intial requirement for NER is considered met (Balke, 2012) – despite the still remaining challenges like word sense disambiguation – and a comprehensive survey is provided by (Nadeau & Sekine, 2007).

However, the basic problem of determining what *general* entities might be of interest to some users and how can extracted entities be classified with respect to suitable concepts remains (Balke, 2012). Balke (2012) discusses the use of taxonomies and ontologies for this purpose, for example to identify 'is-a' relationships based on lexicosyntactic patterns in text.

The YAGO-NAGA approach (Kasneci, Ramanath, Suchanek, & Weikum, 2009) extracts information from infoboxes and category system of Wikipedia based on rules and reconciles the resulting facts with WordNet's taxonomical class system. The output is assesses against the YAGO knowledge base. With this approach potential inconsistencies can be identified, for example "that a person's birth place is unique and that certain cities are located in Europe so that an American-born person cannot be born in such a city" (Kasneci et al. 2009, p 43). Also KYLIN, a prototype introduced by Wu & Weld (2007), uses Wikipedia as intial data source to derive new facts and use an ontology to improve its quality.

Even though Balke (2012) provides a good overview of recent research, considering the diversity and complexity of information extraction problems and solutions still no approach is prevailing.

A specific type of information extraction – applicable to multi-media documents – is called 'feature extraction'. A feature is a characteristic of a multi-media document, e.g. color, shape or texture of an image. These features are often called 'low-level features' to indicate that they have little meaning with respect to the depicted object. The resulting 'semantic gap' between the information need, e.g. searching for a particular product in an organization's multimedia collection and the retrieval possibilities, e.g. searching for a prevailing color yellow. To close this gap some research has been done in the field of semantic annotations as detailed below.

[31] Thomson Reuters OpenCalais is a service and open API for tagging people, places, facts and events in text. URL: http://www.opencalais.com/blogs/kristathomas/introduction-opencalais (retrieved: 30.11.2012)

As shown above comprehensive research has been done on information extraction from text and particularly approaches for NER can be considered to complement format-independent metadata generation if need be. To what extent feature extraction from multi-media documents can contribute is not that clear, especially as the value of low level features isn't obvious for business purpose.

3.1.3 Semantic Annotations

The difference of semantic annotations to harvesting or extracting metadata from documents is "the process of tying semantic models and natural language together […] which may be characterised as the dynamic creation of inter relationships between ontologies […] and documents" (Malik et al. 2010, p 16). In case of metadata generation for contracts this could mean that not only a name is identified in a contract but also the 'semantics' of that name, e.g. the person that 'owns' the name, the role that person plays in the company, the address of the department she is working for and so on. The example shows why the boundary between information extraction and semantic annotation is blurred: whereas the role a person plays in a certain *document* can be extracted based on rules (e.g. to populate the metadata element "contributor"), the role that very person plays in a *company* must be inferred from a knowledge base, here from the enterprise ontology. The gained knowledge can then be used to infer additional metadata, for example the address of the contractor person and to improve quality of generated metadata, e.g. by checking if the contractor person has the competency - according to her - role to sign a contract. If so, likeliness of being the contractor is increased.
Semantic Annotation enriches the unstructured or semi-structured data with a context that is further linked to the structured knowledge of a domain[32]. Uren et al. (2006) give a comprehensive overview and comparison of tools and frameworks for semantic annotations but approaches handling non-web resources are rare.

3.1.3.1 Semantic Annotations for Web Pages

Also web pages are not considered in the mintApproach research on web-resources is investigated due to its importance and the potential to apply the results on other document forms. As annotating web pages manually is an expensive, laborious and error-prone procedure, much research effort has been put into developing annotation tools that automatically or semi-automatically (with human interaction) create metadata for web-sites. Also several tools for manual, semi-automatic and fully automatic semantic annotations have been developed as showed in a recent overview given by Kiyavitskaya et al. (2009).

Already in 1998 a framework for using ontologies to annotate web documents and to provide query access and inference service that deal with the semantics of the presented information has been developed within the Ontobroker project (Fensel, Decker, Erdmann, & Studer, 1998). In their approach, Fensel et al. (1998) base on the HTML representation of a web page and add – system supported but manually – ontological tags, i.e. an instance relationship between the page and a class of the ontology. Therefore, they extend the HTML anchor tag by adding a few keywords that let the parser locate the annotated information (Decker, Erdmann, Fensel, & Studer, n.d.) This approach has the advantage that the pages remain readable by standard browsers with all factual ontological information stored in the HTML page itself. Ontobroker has been originally developed by the Institute of Applied Informatics and Formal

[32] Definition of Semantic Annotation by Ontotext. URL: http://www.ontotext.com/kim/semantic-annotation (retrieved: 22.12.2011)

Description Methods (AIFB)[33] at the University of Karlsruhe, and is now commercialized by the company Ontoprise[34].

Also the approach of Stuckenschmidt and Harmelen (2001) takes advantage of structured information like X/HTML-tags and relations between sides, e.g. via links. To augment syntactic information presented in web resource ontologies are used as a semantic foundation for semi-automatic metadata creation. They aim for classifying web pages on the basis of their structure and relate web pages to a pre-existing ontology in a way that the formal semantics of the ontology can be used for consistency checking and filtering of web pages.

SHOE (Simple HTML Ontology Extensions), another early system for adding semantic annotations to web pages, is a knowledge representation language allowing for the design and use of ontologies directly on the World Wide Web (Heflin, Hendler, & Luke, 1999). SHOE Knowledge Annotator allows users to mark up pages manually or integrate 'mark-up scripts' if documents are created, for example from databases (Dill et al., 2003).
Both approaches, Ontobroker and SHOE, are not suitable for my approach because of its restriction to web-sites and necessity of human interaction.

CREAM (Creating RElational, Annotation-based Metadata), introduced by Handschuh et al. (2001) is a framework for an annotation environment that allows to manually associate instances with ontological concepts or to semi-automatically extract information using wrappers. An implementation of CREAM is "Ont-O-Mat, a component-based, ontology-driven markup tool" (Handschuh et al., 2001). Ont-O-Mat, and its commercial version OntoAnnotate, is suited for highly structured web documents (Kiyavitskaya et al., 2009) and therefore has the same limitations as Ontobroker and SHOE.
In addition, the principal technology for those early approaches has been 'wrapping' and "the consensus view is that they require significant training before they are productive" (Dill et al. 2003, p 117). That could become difficult if not many training documents are available or if they are too heterogeneous.

Despite the advantage of having structural information, like HTML/XHTML tags, links etc. that can be exploited for metadata generation, the "semantic annotation task tackles the same class of text analysis problems that Artificial Intelligence- (AI)-based NLP attacked in the early 80s" Kiyavitskaya et al. (2009, p 1473). Thus linguistic analysis is often part of the annotating, resp. extraction task.

“SemTag”, introduced by Dill et al. (2003) performs automated semantic tagging of large corpora. "SemTag annotates text with terms from the TAP ontology, using corpus statistics to improve the quality of annotations. The TAP ontology contains lexical and taxonomic information about a wide range of named entities, as for instance, locations, movies, authors, musicians, autos, and others. SemTag detects the occurrence of these entities in web pages and disambiguates them using a Taxonomy Based Disambiguation (TBD) algorithm" (Kiyavitskaya et al., 2009, p 1472). Although Kiyavitskaya et al. (2009) speak about large text corpora SemTag has been evaluated only for web pages.

The KIM platform comprises services and infrastructure for semantic annotation, indexing and retrieval for web pages, too (Malik et al., 2010). Like SemTag, KIM focuses on assigning

[33]Ontology Based Access to Distributed and Semi-Structured Information. URL: http://www.aifb.kit.edu/web/OntoBroker/en (retrieved: 11.2.2012)
[34] Ontoprise. URL: http://www.ontoprise.de/en/ (retrieved: 11.2.2012)

links from the entities in the text to their semantic descriptions, provided by an ontology. The analysis is based on GATE[35] (the General Architecture for Text Engineering). "KIM recognizes occurrences of named entities from the KIMO ontology that, apart from containing named entity classes and their properties, is pre-populated with a large number of instances. The generated annotations are linked to the entity type and to the exact individual in the knowledge base" (Kiyavitskaya et al. 2009, p 1473).
The SemTag and KIM approaches are related to the approach I will pursue in my Action Research study with Symfact with respect to using an ontology for metadata generation but which differs with respect to the procedure: where those tools use the ontology to find occurrences of concepts' instances in documents I use ontologies to infer additional metadata or verify already generated metadata.

After discussing research on semantic annotation for web-sites I will now introduce the existing approaches for non web pages.

3.1.3.2 Semantic Annotations for non Web Pages

One approach is CERNO, a framework and tool for supporting semantic annotation of textual documents in the legal domain. In order to support regulation-compliant systems organisations have to analyse legal texts and elicit the requirements (Kiyavitskaya, Krausová, & Zannone, 2008).
Dealing with the problem of organisations of how to align information systems requirements with regulations Breaux & Antón (2008) extended CERNO to automate the extraction of rights and obligations from regulations. They start from manually marking obligations, associated constraints and condition keywords including natural language conjunctions in text. For that they base on the method of Breaux & Antón (2007) and Breaux et al. (2006).

As mentioned above, semantic annotations for non web pages have been researched in particular for multi-media documents. Athanasiadis et al. (2005) implemented a tool that links MPEG-7 visual descriptors to high-level, domain-specific concepts. For the representation of knowledge Athanasiadis et al. (2005) extended and enriched current general-purpose ontologies to include low-level visual features.

In his thesis Lux (2006) developed an architecture for implementing a search engine for semantic metadata. Considering the increasing amount and importance of multi-media documents (Lew, Sebe, Djeraba, & Jain, 2006), Lux focuses on images (stills) also described with the MPEG-7 Semantic Description Scheme. He considers the nature of metadata ('classical' and 'semantic'), the various kinds of representation (e.g. XML, RDF) and storage possibilities (e.g. embedded in HTML code as for web pages) or externally, as it is the case for the MPEG-7 descriptions (Lux, 2006). However, Lux' interest is on improving search strategies for multimedia documents and thus he neglects the problem of poor semantics of the MPEG standard. He refers to the work of Sabol et al. (2005) with respect to research on closing the semantic gap[36] and claims, that Sabol et al. do not consider semantic annotations (Lux 2006, p 27). The opposite is the case: Sabol et al. (2005, p 350) suggest the use of

[35] The General Architecture for Text Engineering (GATE) is open source free software, already 15 years old and used by commercial companies, research labs and Universities worldwide. URL: http://gate.ac.uk/ (retrieved: 9.6.2011)
[36] The difference between user's need for searching multi-media documents, e.g. "Maud in her new dress" and metadata in general available for these documents forms (like number of pixels, colour, resolution, etc.) are called the "*Semantic Gap* between the low-level feature representing and high-level semantics" in non-textual documents (H. H. Wang, Dzulkifli, & Ismail, 2010).

"ontology based inference mechanisms for resolving ambiguities and enriching extracted concepts". An example is given in which an ontologically described meeting scenario "serves as semantic backbone" for mapping low level features to concepts of the ontology, as "detected persons and their current actions in the meeting are mapped to the corresponding concepts". Sabol et al. (2005) introduce a service oriented system architecture for semantic extraction and retrieval of multimedia data, called MISTRAL[37].

Improving multimedia retrieval by relating MPEG-7 to ontologies is a large research area of its own and several interesting approaches have been provided. As an example (Rahman, Hossain, Kiringa, & El Saddik, 2006) is called. The authors present an ontology called Semantic Content Description Ontology (SCDO) to unify different MPEG-7 semantic descriptions of multimedia contents. They introduce a framework that allows combining SCDO with domain- specific ontologies that are used in different description schemes.

Another approach has been followed by Celino et al. (2006) introducing "Squiggle", a framework which supports the building of a domain-aware semantic search engine for multi-media documents. Although it focuses on the search, and only "a little semantics" at indexing time is provided (Celino et al., 2006) it is an interesting approach using ontologies for automatic tagging of non-textual documents. "Squiggle" has been implemented for music files, automatically inferring authors, song titles and music genres from publicly available meta-databases, and then representing the retrieved information in an ontology.

Analysis of research on the various methods of metadata generation has shown that for format-independent metadata generation approaches metadata harvesting and semantic annotations are more valuable as metadata extraction (it deals solely with text documents) and feature extraction (it deals with low-level features like colour or shape of multi-media files). Although the general approach of Semantic Annotation to enrich unstructured or semi-structured data with structured knowledge of a domain provides valuable input, the vast majority of work is done on web-resources. Major difference of using context for automatic format-independent metadata generation is that these approaches do not use context to characterize the *situation* of an entity (the enterprise object) but the *meaning* of a *term*.

After giving an overview on research into various methods for (semi) automatic metadata generation, the next chapter is about metadata standards. McCray & Gallagher (2001) provide ten principles for building digital libraries; one is "adopt and adhere to standards" to ensure consistency, interoperability and automation. These requirements are also important for my approach.

3.1.4 Metadata Standards

Metadata standards have been generally developed in response to the needs of specific resource types, domains or subjects. Therefore, many standards are available but none is specific for enterprise documents and only a few are relevant for my thesis, namely metadata for general purposes. An overview on current standards is given by Zeng & Qin (2008).

[37] "Mistral is a project founded by the Forschung, Innovation, Technologie - Informationstechnologie (FIT-IT) research programme in the programme line of 'semantic systems and services'. It has started January 2005 and has a planed duration of 2 years." Project overview: URL: http://mistral-project.tugraz.at/overview (retrieved: 12.2.2012)

Using a standard to describe an enterprise's documents is of advantage for the following reasons:

- Comprehensibility
  A standard can be regarded as a set of terms, definitions and guidelines created by bringing together the experience and expertise of all interested parties (experts, regulators, users etc.) to make life simpler and to increase the reliability and the effectiveness of services[38].
- Availability
  Besides the standard itself related specifications and tools provide a comprehensive basis of work one can benefit from, e.g. guidelines to ensure quality, templates one can build on (instead of reinventing the wheel) or software that can be customized (instead of starting from scratch, e.g. for metadata harvesting[39]).
- Interoperability
  Standardization of structure and content of information supports interoperability amongst humans, between humans and machines, and between machines only.
  This is important to make information accessible beyond the scope of a certain project: making information exchangeable with other/new stakeholders (for example in case of networking or merging with other companies) and making an application portable.

Because of the large variety of existing metadata standards, developed in response to the needs of specific resource types, domains or subjects, I concentrate on metadata for general purposes. Therefore in the following sections the three standards Dublin Core (DC), Metadata Object Description Schema (MODS) and Metadata for Learning Resources (MLR) are introduced.

3.1.4.1 Dublin Core (DC)

The Dublin Core Metadata Initiative is an open organization engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. The name “Dublin” is due to its origin at a 1995 invitational workshop in Dublin, Ohio; “core” because its elements are broad and generic, usable for describing a wide range of resources[40]. Goal of Dublin Core is to provide a minimal set of 15 descriptive elements "to describe any kind of resource - including various collections of documents and non-electronic forms of media such as a museum or library archive"[41]

DCME No	DCME	Explanation
1	Contributor	An entity responsible for making contributions to the resource; Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity
2	Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.

[38] Based on a definition of The British Standards Institution, 2009. URL: http://www.bsigroup.com/en/Standards-and-Publications/About-standards/What-is-a-standard/ (retrieved: 7.11.2010)
[39] For example, the Dublin Core Metadata Initiative (DCMI) Tools Community maintains a compilation of tools and software related to the standard. URL: http://dublincore.org/tools/ (retrieved: 7.11.2010)
[40] Dublin Core Metadata Initiative. About us. URL: http://dublincore.org/about-us/ (retrieved: 5.11.2010)
[41] DCMI Frequently Asked Questions (FAQ). URL: http://dublincore.org/resources/faq/ (retrieved: 5.11.2010)

DCME No	DCME	Explanation
3	Creator	An entity primarily responsible for making the resource; Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity
4	Date	A point or period of time associated with an event in the lifecycle of the resource.
5	Description	An account of the resource; description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource.
6	Format	The file format, physical medium, or dimensions of the resource.
7	Identifier	An unambiguous reference to the resource within a given context; Recommended best practice is to identify the resource by means of a string conforming to a formal identification system
8	Language	A language of the resource; Recommended best practice is to use a controlled vocabulary such as RFC 3066
9	Publisher	An entity responsible for making the resource available; Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity
10	Relation	A related resource; Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.
11	Rights	Information about rights held in and over the resource; Typically, rights information includes a statement about various property rights associated with the resource, including intellectual property rights.
12	Source	The resource from which the described resource is derived; The described resource may be derived from the related resource in whole or in part. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.
13	Subject	The topic of the resource; typically, the topic will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary. To describe the spatial or temporal topic of the resource, use the Coverage element.
14	Title	A name given to the resource.
15	Type	The nature or genre of the resource; Recommended best practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMITYPE]. To describe the file format, physical medium, or dimensions of the resource, use the Format element.

Table 4: Dublin Core Metadata Element Set, Version 1.1[42]

Table 4 provides an overview on the Dublin Core Metadata Element Set. The fifteen-element "Dublin Core" achieved wide dissemination as part of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and has been ratified as IETF RFC 5013, ANSI/NISO Standard Z39.85-2007, and ISO Standard 15836:2009.) [43]

However, those 15 elements, called 'simple' Dublin Core Metadata Element Set are not precise enough with two respects. Firstly, they are ambiguous: the element 'Date' for example can be the creation date of a resource, the modification date, the archiving date and so on.

[42] Dublin Core Metadata Element Set. URL: http://dublincore.org/documents/dces/ (retrieved: 24.10.2010)
[43] Metadata Basics. URL: http://dublincore.org/metadata-basics/ (retrieved: 5.11.2010)

Secondly, they leave room for interpretation. To address these flaws DCMI has developed two broad classes of qualifiers, namely *Element Refinement* and *Encoding Scheme*.

- Element Refinement: These qualifiers make the meaning of an element narrower or more specific. A refined element shares the meaning of the unqualified element, but with a more restricted scope. A client that does not understand a specific element refinement term should be able to ignore the qualifier and treat the metadata value as if it were an unqualified (broader) element. The definitions of element refinement terms for qualifiers must be publicly available" (Kokkelin and Schwänzl 2001). The element date for example has been 'qualified', that means redefined to *dateAccepted*, *dateCopyrighted*, *dateSubmitted*, etc.
- Encoding Scheme: These qualifiers identify schemes that aid in the interpretation of an element value. These schemes include controlled vocabularies and formal notations or parsing rules. A value expressed using an encoding scheme will thus be a token selected from a controlled vocabulary (e.g., a term from a classification system or set of subject headings) or a string formatted in accordance with a formal notation (e.g. 2000-01-01 as the standard expression of a date). If an encoding scheme is not understood by a client or agent, the value may still be useful to a human reader. The definitive description of an encoding scheme for qualifiers must be clearly identified and available for public use." (Kokkelin and Schwänzl 2001).

An overview on refinements and encoding schemes for the DCMES elements is provided by DCMI[44]. Please note that all terms are optional, can be used manifold and in any order. But it is not possible to create a new Dublin Core element whose meaning goes beyond the scope of the original elements in the Dublin Core Metadata Element Set (DCMES). For example: it is not allowed to introduce a metadata element for a creator's birthday. As this metadata does not describe the resource but the author of a resource it infringes the one-to-one principle[45]. However, a DCMES element specified with an element qualifier is, effectively, a *new* element or property (with a more specialized meaning than its parent element) and if has not been approved by the Dublin Core community, should be handled with caution and with respect to the interoperability the standard provides.

As Dublin Core can be encoded in terms of the Resource Description Framework (RDF), recommended by the W3C, the standard is well suited to describe knowledge artefacts in a way that is machine processable. Dublin Core provides an RDF/XML structure for unambiguous expression of DCMES, as well as the possibility of straightforward addition of more detailed descriptions from the communities concerned[46]. As D. G. Campbell (2002, p 105) shows "the Dublin Core is expressly committed to fostering the development of metadata description across multiple domains, and to facilitating the interoperability necessary for cross-domain resource discovery."

To adapt automatic metadata generation to enterprise specific needs the possibility of making refinements is also of major importance. As 'refinement' is a basic concept of qualified DC, elements can be made more specific if necessary or if that is not sufficient another scheme can

[44] DCMI provides a convenience copy on their web site. URL: http://dublincore.org/documents/dcq-rdf-xml/#sec6. The normative reference is http://dublincore.org/documents/dcq-rdf-xml/#DCQual (retrieved: 7.7.09)
[45] Dublin Core Wiki. URL: http://wiki.dublincore.org/index.php/Glossary/One-to-One_Principle (retrieved: 4.5.2012)
[46] Please refer to http://www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/ (retrieved: 12.12.12) for more information.

be added[47]. Furthermore, to adapt Dublin Core to company specific needs the Dublin Core Metadata Initiative provides a framework for designing a Dublin Core Application Profile (DCAP). "A DCAP defines metadata records which meet specific application needs while providing semantic interoperability with other applications on the basis of globally defined vocabularies and models".[48]

3.1.4.2 Metadata Object Description Schema (MODS)

The Metadata Object Description Schema (MODS) has been developed in 2002-2003 by the Network Development and MARC Standards Office at the Library of Congress in collaboration with metadata experts. MODS can be regarded a transformation of the MARC format. MARC stands for MAchine-Readable Cataloging, a family of standards used in the field of library science. As metadata "in its broadest sense also includes traditional cataloguing data stored in computer systems" (Zeng & Qin, 2008), MODS has been developed with the goal to make the MARC format readable, processable and retrievable by new technologies and tools. Zeng & Qin (2008) give an overview on the transformation process of MARC formats to MODS and on their relationship
The current version of MODS is version 3.4, comprising 2 root elements and the 20 top level elements depicted in
Table 5.

titleInfo	note
name	subject
typeOfResource	classification
genre	relatedItem
originInfo	identifier
language	location
physicalDescription	accessCondition
abstract	part
tableOfContents	extension
targetAudience	recordInfo

Table 5: MODS Top Level Elements[49]

Same as for DC, guidelines and description schemas are available for MODS[50]. The standard is maintained by the Network Development and MARC Standards Office of the Library of Congress. "MODS is an XML schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications."[51]

[47] The DC has proven to be able to interact with other vocabularies and community defined schemes; please refer to http://dublincore.org/documents/dcq-rdf-xml/#sec6 for further information.
[48] Guidelines for Dublin Core Application Profiles. URL: http://dublincore.org/documents/profile-guidelines/ (retrieved: 24.10.2010)
[49] Outline of Elements and Attributes in MODS Version 3.4 (listed in order, read down each column). URL: http://www.loc.gov/standards/mods/mods-outline.html#titleInfo (retrieved: 6.11.2010)
[50] MODS User Guidelines (Version 3). URL: http://www.loc.gov/standards/mods/userguide/generalapp.html (retrieved: 6.11.2010)
[51] MODS Introduction and Implementation.
URL:http://www.loc.gov/standards/mods/userguide/introduction.html (retrieved: 6.11.2010)

Dublin Core as well as MODS are general purpose metadata standards. Whereas DC is simple and therefore easy to implement, MODS is more specific but also more complicated to use in applications, especially for non-library professionals. "DC has been adopted by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH); MODS is the major descriptive format of the Metadata Encoding and Transmission Standards" (METS) (Zeng & Qin, 2008).

3.1.4.3 Metadata for Learning Resources (MLR)

Although called "Metadata for Learning Resources" the MLR standard, the International Organization for Standardization (ISO) is currently working on, is *not* restricted to learning resources but intended to cover digital resources in general. The first part of the standard[52] provides the "General Framework for Metadata and Application Profiles that is completely interoperable and compatible with the Dublin Core (DC) metadata standard" (Stracke, 2010). The MLR standard will comprise the following parts:

- Part 1: Framework
- Part 2: Core Elements
- Part 3: Core Application Profile
- Part 4: Elements Elements
- Part 5: Educational Elements
- Part 6: Availability, Distribution, and Property Elements[53]

Unfortunately, hardly any information on the MLR is published. Not even a complete list of metadata elements defined is openly available. This might be due to the fact that the standard is not fully developed yet but also that ISO standards in general are not freely available on the web. To my knowledge, the latest publicly available publication on MLR is still the one from Currier (2008).

3.1.4.4 Summary of Review on Metadata Standards

In addition to the current standards for metadata for general purpose, introduced above, Zeng & Qin (2008) present an overview on purpose specific metadata standards for Cultural Objects and Visual Resources (e.g CDWA[54]), Educational Resources (e.g. LOM[55]), Archival and Preservation (EAD[56]), Rights Management (e.g. ODRL[57]) and Multi-media Objects (e.g. MPEG[58]).

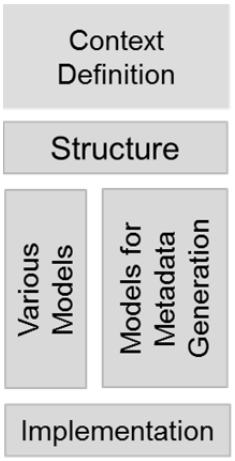
The decision which metadata standard should be chosen, if a standard should be modified, or if no standard at all is taken but a proprietary schema is to be developed etc., is enterprise

[52] ISO/IEC 19788-1:2010 Information Technology - Learning, Education, and Training - Metadata for Learning Resources. Part 1, the framework, is about to be published.
[53] ISO/IEC JTC1 SC36 WG4 (MD) - Overview. URL: http://www.cen-ltso.net/main.aspx?put=991 (retrieved: 5.11.2010)
[54] Categories for the Description of Works of Art (CDWA).. URL: http://standards.jisc.ac.uk/catalogue/CDWA.phtml (retrieved: 30.11.2012)
[55] Learning Object Metadata (LOM). URL: http://ltsc.ieee.org/wg12/20020612-Final-LOM-Draft.html (retrieved: 30.11.2012)
[56] Encoded Archival Description (EAD). URL: http://www.loc.gov/ead/ (retrieved: 30.11.2012)
[57] Open Digital Rights Language (ODRL). URL:. http://odrl.net/1.0/ODRL-10-HTML/ODRL-10.html (retrieved: 30.11.2012)
[58] The Moving Picture Experts Group (MPEG) Web-Site. URL: http://mpeg.chiariglione.org/ (retrieved: 11.2.2012); many links to further publications are provided at the web-site URL: http://mpeg.chiariglione.org/mpeg_books.php (retrieved: 11.2.2012)

specifc as it depends for example on the types of resources metadata is to be generated for, and the strategy the enterprise follows.
From the standards for metdata for general purpose Dublin Core is prevailand and also many crosswalks, i.e. mapping from another standard to Dublin Core are available (Godby, Young, & Childress, 2012).

3.2 *Context*

Documents managed in an organisation (enterprise or public administration) can be considered business relevant, i.e. they are created, read, updated, deleted or archived within business processes, from business actors for business purposes. Business builds the context of the documents I consider in my thesis.
As it is agreed that giving an universally valid definition of context is difficult (e.g. Mena et al. 2007, Baldauf et al. 2007) because context definitions vary depending on its use and the nature of system context is used in, Chapter 3.2.1 starts with the clarification of the different definitions of context followed by an introduction to the structure of context. After these various models for context are discussed first, then related research on context models for metadata generation is provided. The chapter closes with approaches for implementing context awareness.



Structure of Chapter 3.2

3.2.1 Definition of Context

According to Dey & Abowd (1999) context is any information, which can be used to characterize the situation of an entity. "An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves" (ibid.)
Although that definition is very general, it is the most widely accepted (Zimmermann et al. 2007) and reflects the fact that context is considered in very different disciplines like linguistics, philosophy, cognitive psychology, artificial intelligence and more recently in ubiquitous and mobile computing (amongst others Schmidt 2002).
Dey & Abowd (1999) give an overview on early context definitions, like the ones from

- Schilit & Theimer (1994), who refer to context as location, identities of nearby people and objects, and changes to those objects, and
- Brown (1998) considers context as those elements of a user's environment the system is able to 'understand'.

More recently Bazire & Brézillon (2005) collected and analysed a corpus of 150 definitions, but Mena et al. (2007) identified so called 'invariants' in the definitions:

- context relates always to an entity
- context is used to solve a problem
- context depends on the scope of use (domain) and time
- context is evolutionary, i.e. it changes over time.

With respect to the objective of using context for automatic, format-independen metadata generation, the definition Winograd (2001) gave is of particular interest as he inferred its use in the linguistic sense. He introduced context as an "operational term: something is context because of the way it is used in interpretation, not due to its inherent properties" and concludes: "Features of the world become context through their use" (Winograd 2001, p 405).

Following Winograd (2001), Hinkelmann et al. (2010) consider "*context* as everything that is not *text*, that is not *content* of the focused model" and "what content is and what context is, is determined by its use" (original emphasis by the authors). For my thesis I take up on this definition, regarding context as those entities of an enterprise architecture that are needed for the automatic generation of metadata for documents, managed in an enterprise, e.g. business processes, activities, organisational units etc. However, these entities are *not* specifically *modelled for* the purpose of metadata generation but exist independently - as parts of an enterprise architecture - and become context only through its use.

3.2.2 Structure of Context

At the end of the Eighties, beginning of the Nineties research started on formal theory of context by J. McCarthy (1993) and Giunchiglia (1993). "The goal was to explain the properties of the context and the contextual reasoning in a systematic way" (Mena et al. 2007). Find a comprehensive overview on approaches focussing on context for example Bouquet et al. (2003).
Context often is distinguished into dimensions, starting with Lenat (1998) who categorized context as a point in a twelve dimensional space. Zimmermann et al. (2007) reduce the categories to five (individuality, activity, location, time, relation) whereas Leppänen (2005) distinguishes seven domains (purpose, actor, action, object, facility, location, time). To deal with web service Maamar et al. (2006) see three perspectives: participation (to properly specify a web-service), execution (to monitor computing resources a web-service needs), preferences (to specify user preferences) and eventually history as a fourth perspective could be added (to draw a benefit from former situations). Mena et al. (2007) educe six dimensions from their literature study of context: a physical, a cognitive, a social, a cultural, an emotional and a linguistic dimension whereas Saidani & Nurcan (2007) again distinguish between four kinds of context: Location, time, resource and organization related context. Prekop & Burnett (2003) differ between *external context elements*, considered for example in approaches for mobile and ubiquitous computing and *internal context elements* to support cognitive activities (like process modelling or metadata generation).

Dey & Abowd (1999, p 308) differentiate between primary and secondary context. "Location, identity, time, and activity are the primary context types for characterizing the situation of a particular entity". This primary context can be used to "[…] find secondary context (e.g., the email address) for that same entity as well as primary context for other related entities (e.g., other people in the same location)" (ibid.)

All of these works define extra context models for specific or general purpose(s). Contrary to that a semantically enriched enterprise architecture description can be used to answer the questions of Dey & Abowd (1999) about the "who's, where's, when's and what's" context-aware applications should look at. The resulting primary context types are then used to find secondary context types and related entities or vice versa. For example: primary context of the creator of a document is the role this person has in an organisation; secondary context is the job description of the assigned role. Please refer to Chapter 5.2.3 on how I applied this notion of context types in my approach.

3.2.3 Models for Context

In general there are two types of context models: those specifically built to manage use-oriented knowledge for a certain purpose and non-dedicated models.

Some well-known examples of specific models are built for

- improving the understanding of natural language utterance (Guha 1991), (McCarthy 1993), (Lenat 1998)
- supporting ubiquitous computing (Linnhoff-Popien & Strang, 2004)
- detection of context for mobile information systems (Schulz 2003)
- increasing agile business process management (Saidani et al. 2007).

The downside of such purpose-specific modelling approaches is considered by Linnhoff-Popien & Strang (2004, p 1): "While early models mainly addressed the modelling of context with respect to one application or an application class, generic context models are of interest since many applications can benefit from these." As already mentioned above, approaches for generic context models are provided for example by Winograd (2001), Hinkelmann et al. (2010).

How to represent context - either to serve a specific purpose or not - is another topic of research Linnhoff-Popien & Strang (2004) address. They give an overview on context modelling approaches, e.g. the Key-Value-Models, Mark-up Scheme Models, Graphical Models like UML, Object Oriented Models, Logic Based Models and Ontology Based Models and evaluate the models with respect to the requirements they defined. Linnhoff-Popien & Strang (2004) conclude that ontologies are the most promising assets for context modelling (for ubiquitous computing environments). The advantage of using an ontology to model context has long been identified (Guha 1991), and is now agreed widely. Schulz et al. (2003) for example stress the understandability of ontologies for modelling whereas Linnhoff-Popien & Strang (2004) emphasize the capabilities of ontologies for knowledge sharing, logic inferencing and knowledge reuse. Context ontologies are introduced amongst others by Chen et al. (2004) , Schulz et al. (2003), Wang et al. (2004) and Jonsson (2007).

Even though the benefits of using ontology based context models is without controversy, two main obstacles have to be overcome:

- picking the (right) vocabulary to represent the knowledge of the domain (Guha 1991), and
- finding the right granularity level to model context (Lenat 1998).

Again it appears to be of advantage to have a context model that is standard practice, like an enterprise architecture description. Kang et al. (2010, p 1461) explicate "the description of enterprise architecture includes components definition, characteristic and constraints" and the Object Management Group (OMG, 2008) for example, provide the Semantics of Business Vocabulary and Business Rules (SBVR) supporting a coherent description of enterprise architecture to enable communication among stakeholders (Jonkers, Lankhorst, Buuren, Hoppenbrouwers, & Bonsangue, 2004).

There are several approaches to deal with the 'granularity problem'. According to (Guha 1991) contexts are rich objects in a domain that cannot be completely described. To deal with incompleteness, expressiveness and simplicity must be weighed up, i.e. a model should be expressive enough to serve the intended purpose but simple enough to be represented in an inferable model and maintainable by business users not technicians or knowledge experts. Giunchiglia (1993, p 354) argued that "reasoning is usually performed on a subset of the global knowledge base" and Bouquet et al. (2003) differentiate between the 'divide-and-conquer context type' (in which context represents partitions of a global model of the world) and the 'compose-and-conquer context type' (in which context is regarded as a local theory of the world in a network of relations with other local theories).

However, these concerns can be regarded as *not* context specific but as general problems of building an ontology (cf. Gruber et al. 1993, Uschold & King 1995) and thus will be investigated in more detail in section 3.4.3.

Besides the questions of *what* to model, the question of *how* to model context has been investigated. Mena et al. (2007) provide a method to explicitly model context in a 'context engineering' process because developing context is often hidden in the engineering process. Saidani & Nurcan (2007) claim that context modelling should be an integrated part of business process modelling to be most beneficial. A practice oriented, concise procedure for building an ontology and inferencing rules is given by Feldkamp et al. (2010). The model has its foundation in procedures for IT project management adapted to the task of modelling semantically described processes in the E-Government domain.

3.2.4 Implementation of Context Awareness

To use context the model must be implemented and integrated into the existing IT environment. The applicability of context models to existing IT infrastructure is important and can be reached for example using a service framework such as Web Services (Linnhoff-Popien & Strang 2004). Baldauf et al. (2007) present architecture principles of context-aware systems and provide a conceptual design framework.

Winograd (2001) differentiates three types of implementations to reach context-awareness: "context widgets" (proposed by Dey & Abowd 1999), an infrastructure-centred distributed services model J. I. Hong & Landay (2001) and the blackboard architecture [of Winograd], also proposed by Korpiää (2005). Using rules to model context-aware behaviour is suggested by Beer et al. (2003).

Context awareness is mainly discussed in relation to ubiquitous or pervasive computing to provide adequate service for the users, applications and services and to automatically adapt to their changing contexts (Hong et al. 2009). An interesting implementation for example is presented by Kröse et al. (2008) in the application area of ambient assisted living. Kröse et al. (2008) introduce a system that is able to monitor activities of people in a domestic environment. Such kinds of context awareness exceed by far what is needed for automatic, format-independent metadata generation. Current literature review on context-aware systems is provided by Hong et al. (2009).

3.2.5 Using Context for Metadata Generation

There is little research specifically on how context can be used for automatic metadata generation. One may argue that using ontologies to create semantic annotation is kind of exploiting the context – as shown mainly of web resources – but these approaches do not use context to characterize the *situation* of an entity (Dey & Abowd, 1999) but the *meaning* of a term. Stuckenschmidt and Van Harmelen (2001) for example use context to restrict the interpretation of terms and their relations to other terms. In the SemTag approach context is simply regarded as "ten words to either side as a 'window' of context" around an identified object (Dill et al. 2003, p 119), an approach known as word-space-model (Peirsman, Deyne, Heylen, & Geeraerts, 2008). Handschuh et al. (2003) regard context as information that supports semantic disambiguation.

Davis et al. (2004) introduce a method for generating metadata for photos using spatial, temporal, and social context. They developed the MMM-Prototype (Mobile Media

Metadata")[59] which gathers metadata from the context of capture, the 'when' (the date and time of image capture) and the 'where' (the location of the camera when the image was captured) and suggests additional metadata based on a database of similar annotated images. However, to my knowledge all approaches in the multimedia domain primarily consider the *physical* context (temporal and spatial information about an object) but not the *business* context, i.e. concern, reason and purpose why for example an image has been taken.

A sort of business context is taken into account in the domain of learning object. Cardinaels et al. (2005) and Ochoa et al. (2005) worked on frameworks for automatic indexing of learning objects, making use of relations between objects in Learning Management Systems. Recently Margaritopoulos et al. (2008) took on research on automatic metadata creation for learning object using the pre-existing metadata of related resources as context. As the approaches are specifically designed for creating metadata for learning objects and limited to the context provided within the environment of Learning Management Systems results cannot be (re-)used in a general way.

The notion of context of Mei and Zhai is even broader than the one of Dey & Abowd (1999), since "any metadata entry of a document can indicate a context […] For example, the source of a news article, the author's age group, occupation, and location of a weblog article, and the citation frequency of a research paper" (Mei & Zhai 2006, p 649). Mei & Zhai (2006) provide an approach for contextual text-mining. They act on the assumption that topics covered in a document are usually related to the context of the document and thus, try to extract themes from a text collection with the help of context information (e.g., time and location). To model context the authors introduce so called context features, which can be any metadata of a document. These context features are used for topic mining and clustering of documents. Although an interesting work on documents' context it is of limited relevance for automatic metadata generation based on context since here context is not used to create metadata but is built on metadata to support text-mining.

In the ACTIVE Project[60] technology has been developed to overcome information overload in enterprise (Simperl et al., 2010). Active Knowledge Workspace (AKWS) is a tool that tailors information to a user's context, and manages users' informal processes. The AKWS allows (or requires) a user to define their working contexts and to associate information objects with those contexts. User defined context can be augmented by automatically learned contexts. This is achieved by observing users' behaviour and semiautomatic identification of relevant patterns (Simperl et al., 2010). Text mining is used to handle textual data (e.g. emails) connected to the users (e.g. email senders and receivers) and augmented with background knowledge (e.g. user's position in the organization), modelled in an ontology. Implementation is based on Markov models to identify "frequent sequences of lower-level actions that potentially represent more complex user tasks" (Simperl et al. 2010, p 42). Although there are similarities to my work, ACTIVE focuses on shared context to improve collaboration. Metadata (in ACTIVE called tags) is created manually or semi-automaticly based on file contents and tags assigned to similar files, where similarity is based on the information retrieval concept of cosine similarity. Hence, multimedia documents in ACTIVE are not considered and context is not used for automatic metadata generation.

[59] The prototype has been used within the SIMS 202: Phone Project, conducted at UC Berkeley, School of information management & systems. URL: http://www2.sims.berkeley.edu/academics/courses/is202/f03/phone_project/index.html (retrieved: 16.2.2012)
[60] ACTIVE was a European FP7 project which ended 28/02/2011. ACTIVE has worked on increasing the productivity of knowledge workers. URL: http://active-project.eu (retrieved: 6.5.2012)

Closer to my work is research done by Mitschick and Meissner (2008) and (Mitschick, 2009) introducing an implementation of an ontology-based management system for personal multimedia documents. As Mitschick and Meissner (2008) focus on multimedia documents starting point is the extraction of document properties, plus additional information, for example EXIF[61] metadata, that is provided by digital cameras like aperture, shutter, focal length, flash usage, etc. Regardless of the question for what use these information might be, the extracted metadata is related to context information about the resource. What is considered context remains vague, described as data about document's usage provided by personal information management applications, like e-mail clients (Mitschick & Meissner, 2008). Another type of context information is retrieved from the web, for example by making use of linked open data[62]. To enhance metadata for music audio files information from sources like MusicBrainz[63] is retrieved. Extracted and augmented metadata is stored in a RDF store. Although there are some similarities to my work with respect to the general approach of using context information to automate metadata generation, representing meta knowledge (meta data, context information and their interrelations) in an ontology and providing access to the modelled information for miscellaneous front end applications there are some major differences. Above all goal of the approach of Mitschick and Meissner (2008) and (Mitschick, 2009) is an ontology-based document *management* system for *private use*. Therefore their approach focuses a) on how document's lifecycle is better supported (e.g. providing help for detecting duplicates) and b) how personal documentary practices can be improved. On the contrary my work targets *business* use and shows a) how documents (regardless of their format) are related to enterprise objects (e.g. customer, product, etc.) and how these relations can be exploited for better business *object* management (e.g. obligations *represented* in a contract). The different objectives result in completely different ontological models. Whereas in my approach I want to formally represent enterprise objects for operational use (e.g. obligation management), Mitschick & Meissner (2008) want to improve document life-cycle management. Thus, seEAD is an enterprise ontology whereas the core ontology Mitschick & Meissner (2008) use is an adaptation of the ABC ontology. The ABC model has been developed "to provide a common conceptual model to facilitate interoperability between metadata ontologies from different domains" (Lagoze & Hunter, 2001).

Closest to my work is research done by (Brügmann, 2011) who introduces an approach for managing unstructured information using semantic metadata. He aims to support the enterprise wide lifecycle management of electronic documents "by utilizing data about other data and information (metadata)". The 'ConSense' approach integrates different types of available document-related metadata: metadata extracted from local file system, extracted from email conversations and deduced from desktop activities (Brügmann, 2011). From this metadata semantic relations amongst documents and between documents and business-domain specific entities are deducted and combinded into a semantic knowledge base. The

[61] "Exchangeable image file format (Exif) is a standard that specifies the formats for images, sound, and ancillary tags used by digital cameras (including smartphones), scanners and other systems handling image and sound files recorded by digital cameras." Wikipedia. URL: http://en.wikipedia.org/wiki/Exchangeable_image_file_format (retrieved: 17.2.2012)

[62] "The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web. Key technologies that support Linked Data are URIs (a generic means to identify entities or concepts in the world), HTTP (a simple yet universal mechanism for retrieving resources, or descriptions of resources), and RDF (a generic graph-based data model with which to structure and link data that describes things in the world)." Linked Data. URL: http://linkeddata.org/faq (retrieved: 17.2.2912)

[63] "MusicBrainz is an open music encyclopedia that collects music metadata and makes it available to the public." URL:http://musicbrainz.org/ (retrieved: 16.2.2012)

ConSense approach of Brügmann (2011) allows for extracting metadata from documents and detecting and recording the context of user's activities related to the document's handling. To represent concepts specific to the enterprise domain and their relations the ConSense approach uses a combination of existing controlled vocabularies, like Dublin Core (transforming the orginal RDF Schema notation into OWL-DL), Friend-Of-A-Friend (FOAF), and newly created ontologies as for example the ConSense Document ontology and the ConSense Process ontology. Furthermore Brügmann (2011) uses rules to determine the relations amongst documents and between documents and enterprise entities. Brügmann (2011) introduces seven heuristic rules to recognize the relationships between documents. Therefore he determines similarities in file and folder names based on string comparison algorithms. Another rule set is used to infer similarties between emails, e.g. based on matching sender, recipient and subject. The third set of rules is applied to the context in which the user accesses and modifies documents. To capture the context of user's actions a client-side sensor plugin is considered. Most interesting with respect to my work is what Brügmann (2011, p. 76) suggests for linking "external entities" like product, project, process, role etc., to documents. Unfortunately Brügmann (2011) only briefly describes two potential sources: Personal Information Management (PIM) applications like Microsoft Outlook, and Central Enterprise Applications. Regarding PIM he vaguely suggests to retrieve the relevant metadata, for example from contact reports stored in a PIM application, and insert it into the semantic knowledge base. His suggestion for information stored in Central Enterprise Applications remains even more shallow concluding in the remark that more than 100 free converters exist to transform data from relational databases into RDF (Brügmann 2011, p 77). Although Brügmann (2011) provides an interesting approach for managing unstructured information in an enterprise, implemented in a comprehensive prototype, his work remains in some parts superficial. For example with respect to the aforementioned over-simplified suggestion for ontology-to-database mappings but also regarding the ontology models. Hence, no justification is given why the newly created ontologies are necessary, why existing ontologies are not (re-)used, why OWL-DL is chosen as ontology representation language and which strategy is chosen for ontology-to-database-mapping.

Deriving context information from communication details is part of the approach provided by Brügmann (2011). Although Emails clearly are business documents they aren't considered in my work. All of those documents are already managed by an Email management system like Microsoft's Outlook™ and various aspects like personal information management (e.g. by (Whittaker, Bellotti, & Gzisda, 2006), (Whittaker, 2011), (H. Zhang, 2011)), information creation (Lichtenstein & Parker, 2003) or semantic task management (Riss, Jurisch, & Kaufman, 2009) have already been a research topic.

3.3 *Enterprise Architecture (EA)*

To support the notion of context as a non-dedicated model which can be used for various purposes, for example for business and IT aligment (Woitsch, Karagiannis, Plexousakis, & Hinkelmann, 2009), according to Kang et al. (2010) it is favourable to consider standard practice. In the business domain this can be work related to enterprise architecture, including Enterprise Architecture Frameworks (EAF) and Enterprise Architecture Descriptions (EAD).

The chapter about enterprise architecture is structured as follows: first the notion of EA is discussed, then enterprise architecture descriptions (EAD) are considered. After that Enterprise Architecture Language (ADL) is investigated. Then research on Enterprise Architecture Frameworks (EAF) is discussed and the two well-known examples, Zachmann (Zachman, 2012) and ArchiMate (The Open Group, 2012), are detailed. Finally research on relating EAD to concrete enterprise objects is investigated. The figure at the right depicts the chapter's structure.

Structure of Chapter 3.3

3.3.1 Notion of Enterprise Architecture (EA)

Referred to ISO 15704, an enterprise is one or more organizations sharing a definite mission, goals and objectives to offer an output such as a product or a service (quoted after Chen et al. 2008). As the authors point out, that definition also covers networks of enterprises, e.g. partners in a supply chain but also partners in virtual enterprises.
Management of enterprise is about "The organization and co-ordination of the activities of an enterprise in accordance with certain policies and in achievement of defined objectives."[64] Chen et al. (2008) consider enterprise architecture as the foundation of enterprise systems engineering with the goal to support stakeholders of an enterprise to manage system engineering and changes.

Basically, there are two different notions of enterprise architecture. One perception is as a high level abstraction (of reality) with the purpose of reducing complexity and increasing stakeholder's understanding and communication (amongst others by Chen et al. 2008, Dietz 2006 and Zachman 2012). According to Dietz (2006) the most dominant problem, stated in scientific as well as in popular science on enterprise management, is complexity and how it can be managed. He claims that because of the complexity of enterprises a conceptual model is needed that "only shows the essence of the operation of an enterprise" and therefore "the model abstracts from all realization and implementation" (Dietz, 2006, p 8).

The other, more recent notion of enterprise architecture focuses on integration for example by Woitsch et al. (2009), Hinkelmann et al. (2010) and Valtonen et al. (2011).

The ISO/IEC/IEEE 42010 standard provides a conceptual model of Architecture Description I use to structure my investigation on Enterprise Architecture. The ISO/IEC/IEEE 42010 emphazises that a system-of-interest (which could be an enterprise, a system of systems, a product line, a service, a subsystem, or software) has an architecture *even if that architecture is not written down* (DSCI, 2012). Hence, work on enterprise architecture clearly

[64] Definition provided by the Business Directory. URL: http://www.businessdictionary.com/definition/management.html (retrieved: 16.5.2011)

distinguishes between enterprise *architecture* and enterprise architecture *descriptions*. ISO/IEC/IEEE 42010 standard provides a conceptual model of the terms and concepts belonging to Architecture Description, partially depicted in Figure 11. The framed rectangles indicate the concepts most important for automatic metadata generation based on context.

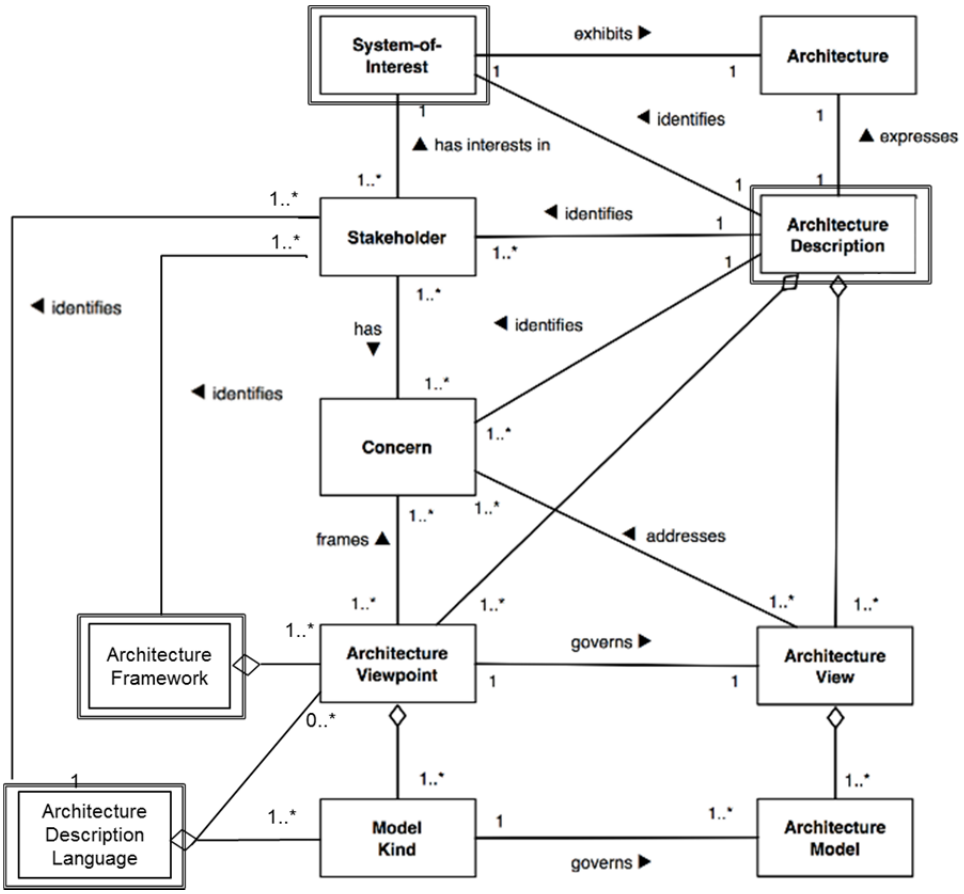


Figure 11: Conceptual Model for EAD (excerpt from ISO/IEC/IEEE 42010 provided by DSCI 2012)[65]

In the following research on the highlighted aspects is investigated, starting with work on enterprise architecture descriptions. Refer to (DSCI, 2012) for details on the other concepts.

3.3.2 Enterprise Architecture Description (EAD)

According to ISO/IEC/IEEE 42010 an "Architecture Description (AD) is an artifact that expresses an Architecture. Architects and other system stakeholders use Architecture Descriptions to understand, analyze and compare Architectures, and often as 'blueprints' for planning and construction" (DSCI, 2012).

Since architecture description can be any artefact that documents an enterprise's architecture it is often described in enterprise handbooks focusing on supporting day-to-day business. Please refer to Chapter 4.2.1.12 for details on the survey's result. Molina et al. (2005, p 309) point out "that Enterprise Modelling largely remains a concept or is even completely ignored by most SMEs (Small and Medium Enterprise)".

[65] The conceptual model is presented in the Standard using UML class diagrams to represent classes of entities and their relationships. URL: http://www.iso-architecture.org/ieee-1471/cm/ (retrieved: 12.10.2012)

Though Chen et al. (2008) refer to work of the Open Group for practical principles, there is still no clear definition of what a good enterprise architecture description is. Johnson et al. (2007) explain this by showing the various purposes an enterprise model can be used for.
Kang et al. (2010), criticize EA (descriptions) for the lack of detailed models of the components, the vagueness of the modelled relationships between the components and the lack of a model for implementation. To overcome these drawbacks Kang et al. (2010) take the Federal Enterprise Architecture that is based on Zachmann's EAF, and use the structure of WordNet to describe terms and SBVR to structure the relationships. However, the model is very detailed and seems difficult to extend into a full blown ontological representation of the EA (in addition there is no description of how the ontologies are formally represented).

Also for Hepp & Roman (2007) expressiveness and formality are missing from enterprise architecture models and, because of this there is only a limited degree of automation. Therefore the authors introduce "an approach of increasing the level of automation of BPM [Business Process Management] by representing the various spheres of an enterprise using ontology languages and Semantic Web Services frameworks" (Hepp & Roman, 2007, p 2).

Allemang et al. (2005) built the Federal Enterprise Architecture Reference Model Ontology (FEA-RMO). Even though they provide a lot of insights about dealing with problems, modelling an EA as ontology it is limited with respect to general use and content. Kang et al. critize "Although FEA-Reference Model Ontology (FEA-RMO) […] is proposed in order to share meanings of FEA reference models, it is nothing but the model which describes FEA reference models with Web Ontology Language (OWL). It is only for FEA reference models and is short of concrete method to share common meanings of Enterprise Architecture components" (Kang et al. 2010, p 1457).

Goudos et al. (2006) based on the Governance Enterprise Architecture (GEA) to address the problem of matching a citizen's needs with available public services. Based on GEA they created an ontology represented in OWL-DL and created with Protégé (with the OWL plug-in). The approach lacks some important points: the GEA is not implemented by a PA[66] and it does not consider various knowledge levels. This Hinkelmann et al. (2010) consider the crucial point, as an EA is not necessarily used throughout a company as monolithic construct. That is especially true with respect to the increasing number of virtual enterprises.

As the investigated research showed, enterprise architecture description can not be regarded without considering the 'language' it is expressed with. According to the ISO/IEC/IEEE 42010 standard an Architecture Description Language (ADL) is any form of expression for use in Architecture Descriptions (DSCI, 2012). As shown in Figure 11: An ADL might include a single Model Kind, a single viewpoint or multiple viewpoints. Examples of ADLs: Rapide, SysML, ArchiMate, ACME, xADL.

3.3.3 Architecture Description Language (ADL)

Schelp & Winter (2009) provide an overview of research on enterprise architecture language. They identified thirty-three distinct enterprise architecture research approaches, documented in nearly 100 publications. For their work on language communities in enterprise architecture research Schelp & Winter (2009) singled out seven approaches with results either achieved by a scientific community and documented in several publications or developed by practitioners

[66] According to (Liimatainen, Hoffman, & Jukka, 2007) the adoption of xGEA (an extension of GEA, a model for cross-Government Enterprise Architecture) is only starting in the UK.

basing on defined terminology provided in Enterprise Architecture Frameworks, e.g. TOGAF (The Open Group, 2009a). All analysed approaches provide EA description languages, the Telematica Institute of The Netherlands for example the well-known ArchiMate EA modelling language (The Open Group, 2012).
However, none of the research communities considers representational languages like RDF, OWL or WSML, developed by the semantic web community, to express the description more formally. Only recently, in the plugIT project OWL has been used for enterprise modelling in order to support business and IT alignment (Kondylakis et al. 2010).

This reflects the fact that enterprise architecture modelling (and description) and ontology modelling originally stem from two different application domains and only recently started to be merged. According to Dietz & Hoogervorst (2008, p 572) "the terms 'Enterprise Ontology' and *Enterprise Architecture* [now] belong to the standard vocabulary of those professionals who are concerned with (re) designing and (re) engineering enterprises". Whereas the term *ontology* emerged in the context of Artificial Intelligence and the Word Wide Web, particularly of the Semantic Web (Dietz, 2006), *enterprise architecture* became generally known as a management topic in the end of the 1980ies, for example through the Zachmann's EAF (Matthes, 2011).

Dietz (2006, p8) stresses the need of a conceptual model because of the complexity of enterprises but his notion of an ontology is of a shared understanding of a domain (Uschold & Grüninger, 1996), completely independent of any ICT implementation. Thus, the author regards an ontology as a white-box (WB) model, "that captures the construction and the operation of system while abstracting from implementation details" (Dietz, 2006, p 65). He also defines five quality criteria for evaluating enterprise ontologies: coherence, comprehensiveness, consistency, conciseness and essence, abbreviated as C_4E. To build an enterprise ontology Dietz (2006) introduces the DEMO[67] methodology. Although Dietz (2006) provides a comprehensive approach to enterprise engineering and gives copious theoretical background of ontology creation and representing, his work does not rely on standards. Instead he suggests a propriety notation for describing enterprise architecture and chooses the DEMO method instead of procedure or architecture models provided in EAFs (e.g. TOGAF).

There is broad consensus that using semantic technologies, i.e. an ontology, is an appropriate approach to represent enterprise architecture knowledge and Fox, Barbuceanu & Gruninger (1996, p 123) give the reason for it as follows: "As information systems play a more active role in the management and operations of an enterprise […] departing from their traditional role as simple repositories of data, information systems must now […] not only answer queries with what is explicitly represented in their Enterprise Model, but must be able to answer queries with what is implied by the model."

Therefore literature review of enterprise ontology is provided in a separate, sub-sequent chapter.

3.3.4 Enterprise Architecture Frameworks (EAF)

According to the ISO/IEC/IEEE 42010 standard “An architecture framework establishes a common practice for creating, interpreting, analyzing and using architecture descriptions

[67] DEMO is an acronym for Design and Engineering Methodology for Organizations, provided by the Enterprise Engineering Institute. URL: http://www.demo.nl/methodology (retrieved: 12.5.2011)

within a particular domain of application or stakeholder community" (DSCI, 2012). A "Framework is a logical structure for classifying and organizing complex information" (FEAF, 1999, C-6).
There is a huge variety of EAF and Matthes (2011) points out, to date more than 50 frameworks are available. In his compendium he gives a detailed description of 34 EAF, based on clearly structured and well defined criteria. His compendium is the first complete and comprehensive overview of Enterprise Architecture Frameworks not only assessing them but also giving the history of a framework, its varying versions and inheritance (if a framework builds the base for or depends on another one).
Ohren (2005) provides an approach for characterising Enterprise Architecture Frameworks using an ontology and defines thirteen criteria she applied to six EAF[68], namely DoDAF (Departement of Defense, 2007), FEAF (FEAF, 1999), GERAM (IFIP IFAC Task Force, 1998), TEAF (Departement of Treasury, n.d.), TOGAF (The Open Group, 2009a) and Zachmann (Zachman, 2008). As the ontology can be easily extended (Ohren, 2005) it would be interesting to research whether the criteria of Matthes (2011) could be incorporated and if questions like "what is the best EAF for my purposes" could be answered with it.

Two Master's theses, carried out by Martin (2010) and Brun (2010), the EAF TOGAF (The Open Group, 2009a), Zachmann (Zachman, 2008), ARIS (Scheer, 2000) plus BPMS (Dimitris Karagiannis, 1995), Best Practice Enterprise Architecture (Hanschke, 2009) and the enterprise architecture framework developed in the Plug-IT project (Wache et al. 2010) have been considered as sources for standard practice. Inter alia that work shows that many EAFs support structuring enterprise architecture according to various perspectives and aspects. Often, the intersection of perspective and aspect is represented by a specific model kind, for example a process model or an organizational model.

The ISO/IEC/IEEE 42010 standard provides the concepts of Architecture Viewpoints and Architecture Views (cf. Figure 11). An architecture viewpoint is a way of looking at a system. An architecture view expresses the architecture of the system of interest from the perspective of a chosen viewpoint. Hence, an Architecture View is defined by one or more perspectives and aspects represented by one or more Architecture Models.

From the Enterprise Architecture Frameworks Matthes (2011) evaluated, I consider ArchiMate, and Zachmann most valuable for automatic metadata generation based on context. The EAFs will be described in more detail in two subsequent sections as the former provides a language for describing enterprise architecture and the latter was very influential on later EAF work.

3.3.4.1 Zachmann Framework

The framework has been developed by John A. Zachmann and described first in 1987[69]. Today an extended version, co-authored with Johan F. Sowa, is available (Sowa & Zachman, 1992). Although all relevant information is subscribable and freely available on Zachmann's web-site the most famous publication is the 'EAF matrix' as depicted in Figure 12.
According to Matthes (2011) the approximate market share amounts between 22% and 25%.

[68] Since some of the frameworks are not accessible anymore in the version Ohren (2005) assessed, the latest version of the standards is cited instead.
[69] An overview on the development of the framework over the years can be found on the internet. URL: http://test.zachmaninternational.com/index.php/ea-articles/100 (retrieved: 12.8.2011)

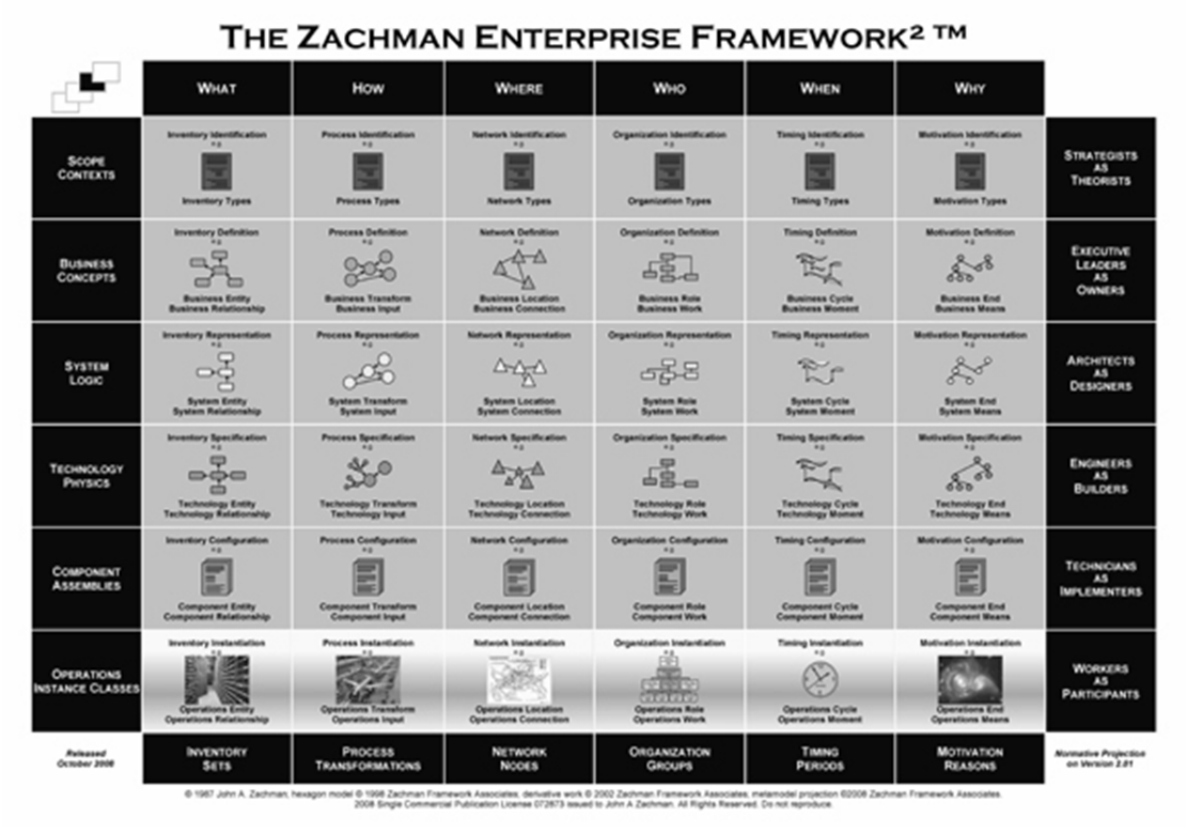


Figure 12: Zachmann's EAF Matrix[70]

The framework provides different views and representations of an enterprise and a classification scheme for organising the several aspects and perspectives of the enterprise. As shown in Figure 12 these views are arranged into a two dimensional matrix. I regard rows as different perspectives of the role a stakeholder may take (named planner, owner, designer, builder and subcontractor), and columns as representations of the various aspects that should be considered. They are "different abstractions from or different ways to describe the real world" (Sowa & Zachman 1992, p 592). The aspects (rows) are named based on the fundamentals of communication. The interrogatives What (data), How (function), When (time), Who (people), Where (network), and Why (motivation) build the basis for the concise description of complex ideas (Zachman, 2008).

Intersections of perspectives and aspects can be represented in models of various model types, like a data model or a process model. Those model types can in turn be represented in various languages. A process for example can be modeled in the BPMN (OMG, 2011a) or with a mark-up language for Event Driven Process Chains (Mendling & Nüttgens, 2002).

Despite the comprehensive version of Zachmann's ebook (according to Matthes (2011) it comprises more than 500 pages), providing "a framework description and use for analysing enterprises and exhaustive descriptions of each of the 30 Cells"[71] and the widespread use of the Zachmann Framework, there is some criticism. Schönherr (2004) sees the disadvantage of

[70] Current Framework2 Elaboration for Zachmann. URL: http://www.zachmanframeworkassociates.com/index.php/home/26-articles/100-framework2-for-zachman (retrieved: 1.11.2012)
[71] The Zachman eBook. URL: http://zachmanframeworkassociates.com/index.php/home-article/25#maincol (retrieved: 18.6.2011)

the framework in the lack of consideration of existing infrastructure. He points out that this drawback is a significant problem when running an integration project. Lankhorst (2009) complains about arbitrarily labeled roles, especially planner and subcontractor, and a low level of detail of relations between some cells. Kang et al. (2010) too, argue that the framework has some drawbacks in modelling detailed components and further criticise that the relations between the individual components are not elaborated in detail. Same criticism is made by Z. Chen & Pooley (2009a) stating that there is no clearly defined dependency between the cells.

3.3.4.2 ArchiMate

ArchiMate as a standard modelling language for describing enterprise architecture and is "complemented by some considerations regarding language extension mechanisms, analysis, and methodological support" (The Open Group 2012, p 2).
According to Matthes (2011) ArchiMate has been developed by Dutch co-operation between government, industry and education. Since 2008 ArchiMate has served as an open standard by the Open Group and in 2009 ArchiMate (1.0) became a technical standard (The Open Group, 2009b): "ArchiMate is an open and independent modelling language for enterprise architecture, supported by different tool vendors and consulting firms" and the "ArchiMate Forum is open to all organizations that apply enterprise architecture in practice or support its use and development" [72]. In 2012 ArchiMate 2.0 (The Open Group, 2012) an upwards-compatible evolution from ArchiMate 1.0 was published, which added new features and integrated usage feedback.

The ArchiMate language aims to support enterprise architects in describing, analyzing and visualizing the relationships among business domains in an unambiguous way. It is thought of as complementary to TOGAF (also an Open Group standard) in the way that the structure of the ArchiMate language neatly corresponds with the three main architectures as addressed in the TOGAF ADM, namely Business Architecture, Information Systems Architecture and Technology Architecture.
ArchiMate provides a graphical representation of its language elements based on UML class diagram but customized and limited to a small set of modelling constructs in the interest of simplicity of learning and use. That notation is widely accepted and thus supported by several tools, amongst others by the Enterprise Architect[73] and Agilian[74].

[72] The ArchiMate Forum. URL: http://www.opengroup.org/archimate/ (retrieved: 18.6.2011)
[73] Enterprise Architecture is UML team-based modeling environment with visual tools for business modeling, systems engineering and enterprise architecture. URL: http://www.sparxsystems.com.au/ (retrieved: 18.6.2011)
[74] Agilian is a modeling environment supporting business process modeling, enterprise architecture and software development. URL: http://www.visual-paradigm.com (retrieved: 17.1.2012)

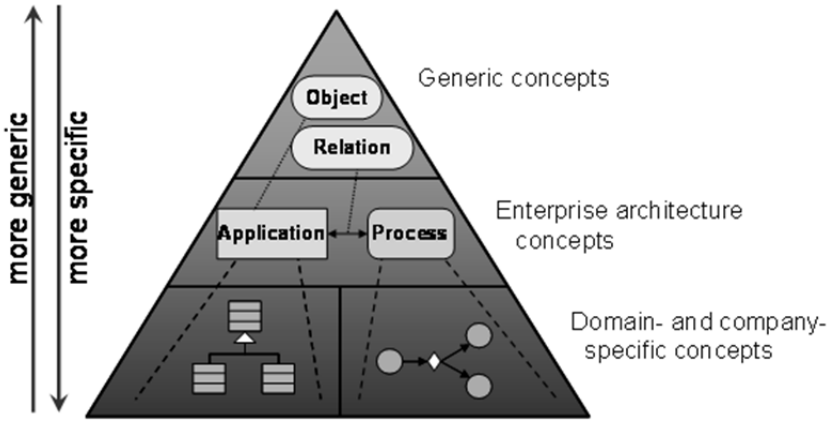


Figure 13: ArchiMate Design Approach (The Open Group 2012, p 3)

Figure 13 depicts ArchiMate's notion that enterprise concepts can be described at different levels of specialization, starting with most general concepts on top of the triangle. At the base of the triangle, domain and company specific concepts are shown. That notion correlates with a common approach for structuring an enterprise ontology into a set of upper or top level concepts, a set of domain concepts and a set of application concept (see Chapter 3.4.1).

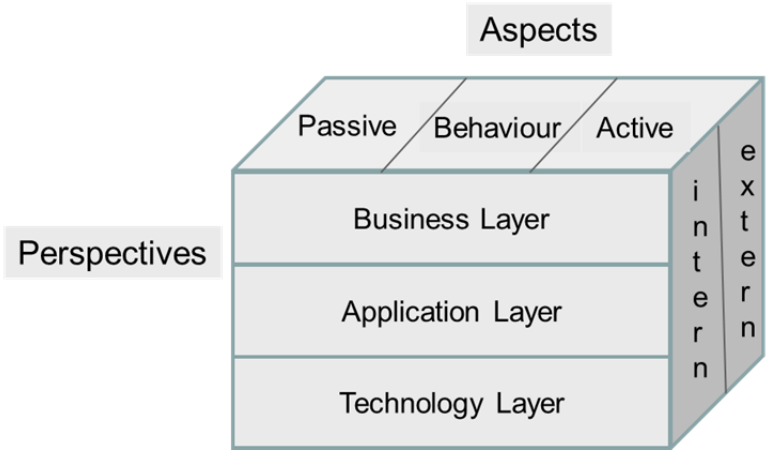


Figure 14: ArchiMate Structure (own presentation)

Figure 14 depicts the general structure of ArchiMate represented in a cube. ArchiMate consist of three aspects ('passive', 'behaviour', 'active'[75]) and three perspectives ('business', 'application' and 'technology', called 'layers'); in addition two views ('external' and 'internal') are considered (The Open Group 2012, p 7).
Starting from five core concepts for each aspect on a – not explicitly specified – 'Top Layer', concepts on each lower layer are specialized together with their relations. Like the concepts, relations are kept generic: 'composition', 'aggregation', 'association', and 'specialization' (taken from UML 2.0) plus 'access', 'assignment', 'composition', 'flow', 'triggering' and 'used by' (The Open Group 2012, p 162).
Since the ArchiMate core language contains only the basic concepts and relationships that serve general enterprise architecture modelling purposes, the language is designed to allow

[75] "These *three aspects* – active structure, behavior, and passive structure – have been inspired by natural language, where a sentence has a subject (active structure), a verb (behavior), and an object (passive structure)." (The Open Group 2009a, p 9)

for extensions and specialisations (The Open Group, 2012). The possibility of extending and refining ArchiMate is particularly important because of the heterogeneity of models present in an enterprise, with varying degrees of maturity and the difficulty to determine their interrelations and the need to do so (Lankhorst, 2004). For automatic metadata generation based on context it is important as for example, the document model ArchiMate provides can be detailed with respect to specific documents types and extended by attributes defined by Dublin Core.

In contrary to Schekkerman (2004) I'm convinced that a greater level of detail forms not an obstacle to engaging the stakeholders in the enterprise architecture development but an argument as it allows for turning 'passive descriptions' into 'active support'. In other words, using the enterprise architecture description in computer programs, as for automatic metadata generation, makes for greater complexity.

3.3.5 Objects of an Enterprise Architecture

As depicted in Figure 12 an enterprise architecture description identifies and documents a system-of-interest. The system consists of the *concrete* objects, e.g. an employee, a document, a product. If an EAD shall be used on operational level – as intended for automatic metadata generation – concepts in the EAD must be related to these objects.
To keep the focus on automatic metadata generation based on context, research is restricted to literature about relating ICT representations of concrete objects, e.g. database records *about* the employee and digital instances *of* a document, to concepts of enterprise architecture descriptions – which in fact build the metadata for the objects. Referring to EAF, the structural element aspect of data is considered (in Zachmann's EAF depicted in the column "What").

Fox & Gruninger (1998, p 110) identified a so called 'correspondence problem' "that the legacy systems that support enterprise functions were created independently and, consequently, do not share the same enterprise models." This means that a database model of an ERP system which is used for the calculation of salaries is completely independent from an enterprise architecture model describing the various database management systems of the enterprise.

Hepp & Roman (2007) use the term *semantic bottleneck* in business process management (a part of enterprise architecture) to explain the gap between the business perspective on operations and the actual execution of operations; the former described in the business perspective, the latter in the technology layer of an enterprise architecture description. Thus the authors introduce Semantic Business Process Managing (SBPM), an approach "to represent both the business perspective and the systems perspective of enterprises using a set of ontologies, and to use machine reasoning for carrying out or supporting the translation tasks between the two spheres" (Hepp & Roman, 2007, p 6). For SBPM semantic web service frameworks, ontology infrastructure and the ARIS EAF and tools are combined. Although Hepp & Roman (2007) provide requirements for representing Semantic Business Process Management and ask competency questions to define scope and a set of ontologies, content of the ontologies is rather small. In addition the ontologies are not yet formalized, which is future work planned to be done in WSML[76].

[76] The WebSevice Modeling Language (WSML) has been developed within the projects DIP, Knowledge Web, InfraWebs, SEKT, ASG and Esperanto, funded by the European Commission. URL: http://www.wsmo.org/TR/d28/d28.3/v0.1/20061218/d28.3v0.1_20061218.pdf (retrieved: 6.5.2011)

In general research on mapping enterprise architecture descriptions, formalized in an ontology, to 'real life' business applications and data is limited. Already in 2003 Maedche et al. (2003) introduced an architecture for implementing an ontology-based knowledge management system (OKMS) dealing with the problem that "a large body of information in an enterprise typically already exists outside the knowledge management system – for example, in other applications such as groupware, databases, and file systems" and needs to be integrated (Maedche et al. 2003, p 2). Although their research was done within the EU-funded "Ontologging" project[77] focusing on distributed ontology-based knowledge management applications and improvement on traditional knowledge management systems using ontologies, their focus was on ontology mapping and evolution. Thus, the authors provide solutions (wrappers) "that lift the content of the different relevant, existing applications to the ontology level", for example for Lotus Notes (documents) (Maedche et al. 2003, p 3). Regrettably, the authors only briefly describe a "virtual mapping engine" that maps concepts and relations to entities of a relational database but do not explain how the mapping is done nor what problems they faced with the approach.

More recent work has been done by Umar & Zordan (2009). Their approach interconnects repositories, in which knowledge about different aspects of an enterprise is stored, on the basis of an ontology (Umar & Zordan, 2009). The authors rely on the MIT Process Handbook (PH) Project that provides a huge online knowledgebase with entries for more than 5000 business activities and tools to manage that knowledge. Umar & Zordan (2009) refer to the Artificial Intelligence Applications Institute at the University of Edinburgh that developed an Enterprise Ontology, a collection of terms and definitions relevant to business enterprises. However, both web sites have not been updated since 2003 and publications about the Edinburgh Enterprise Ontology are dated in the late nineties, e.g. by Uschold et al. (1997).

In 2008 the Object Management Group (OMG, 2008) provided the Semantics of Business Vocabulary and Business Rules (SBVR). Even though some of the approaches are very comprehensive, like the MIT PH or SBVR, they do not provide guidelines, best practices or recommendations on how to link and use ontology concepts to operational data (and business applications). Whereas commercial approaches to ontology based repositories can be found with two large software vendors: Umar & Zordan (2009, p 354) pointed out that SAP and IBM use business support ontology based repositories. The SAP "tool helps the users to analyze their business needs and map them to enterprise application packages from SAP […] IBM's eBusiness Framework shows how to map eBusiness to IBM's application suite".
Yet, linking between ontology concepts and business entities is considered good for knowledge management purposes but not for operational use and thus, "mappings between business to enterprise applications and then to the IT infrastructure do not typically exist" (Umar & Zordan, 2009, p354). Even though the authors bridge the gap, providing a comprehensive approach for enterprise ontology modelling and Business to Business (B2B) integration, used in more than 40 real-life business scenarios, the level of integration is on business *applications* and not on business *data*.

To the best of my knowledge currently there is only one scientific group dealing with the problem of integrating application data and ontologies in an enterprise, which is the Dipartimento di Informatica e Sistemistica 'A.Ruberti' at 'Sapienza' University of Rome. They developed the QuOnto system (a plug-in for Protégé) within a project for data and service

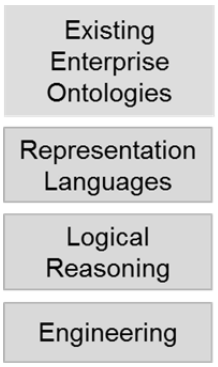
[77] The project web site (www.ontolog- ging.com) was not accessible at retrieval time; some information on the project can be found at 'The Centre for Advanced Learning Technologies' (CALT). URL: http://www.calt.insead.fr/Project/OntoLogging/ (retrieved: 29.10.10)

integration[78]. On the project's web site several papers about the topic are available, for example the one published by Poggi et al. (2008, p 133) about ontology-based data access.

As I will draw upon research findings on how non-ontological data can be technically linked to an ontology to build an enterprise repository, refer to Chapter 5.1.5 for more details on this topic.

3.4 *Enterprise Ontologies*

In that nearly every research paper starts with at least a brief explanation and history of ontologies it is not necessary for me to cover this ground. Therefore the interested reader is referred to some well-known authors like Uschold et al. (1997), Fensel (2004) and Dietz (2006).



Structure of Chapter 3.4

Considering an ontology an explicit specification of a shared conceptualization (Gruber et al. 1993, Studer et al. 1998), in essence means, providing "a view on how the world or a specific domain is structured as agreed upon by the members of a community" (Buitelaar & Cimiano 2008, p 45). Thus, considering enterprise ontology the regarded domain is the domain of enterprise and members of the community are the enterprise's stakeholders.

The chapter on enterprise ontologies is structured as follows: In the first place well-known existing enterprise ontologies are investigated, after that various representational languages for ontologies are discussed, followed by a section on logical reasoning. The chapter closes with work on ontology engineering.

3.4.1 Existing Enterprise Ontologies

Due to the various notions of ontologies (Dietz, 2006, p 9) a huge variety of already existing ontologies have been developed over the last two decades. They largely differ in purpose, coverage, complexity, level of formalization and degree of formality (Fox & Grüninger, 1998).

Dietz (2006) differentiates between world (e.g. world of travelling or world of dining) and system ontology, like enterprise ontology. He considers an enterprise ontology a system ontology, with the goal to understand the structure and operation of a whole system – an enterprise – which includes the notion of world ontology (Dietz, 2006, p 10).

Another approach distinguishes between domain and upper ontologies (Mascardi, Cordì, & Rosso, 2007).
Pinto & Martins (2004, p 442) provide the following definitions:

- an upper-level ontology defines the very general concepts that are are highly reusable across several domains and applications
- a domain ontology defines the concepts from a given domain.

In their report Mascardi et al. (2007) analyse seven upper ontologies (BFO[79], CYC[80], DOLCE[81], GFO[82], PROTON[83], Sowa's ontology[84], and SUMO[85]) based on standard software engineering criteria.

[78] QuOnto. Querying Ontologies. URL: http://www.dis.uniroma1.it/quonto/?q=node/30 (retrieved: 5.11.2010)

Within the domain of enterprises Bertolazzi et al. (2001) analyzed and compared existing ontologies, namely the TorontoVirtual Enterprise (TOVE) and The (Edinburgh) Enterprise Ontology (in the following called "TheEO"[86]), with their own proposal for a Core Enterprise Ontology (CEO). Leppänen (2005) introduced a context-based enterprise ontology and refers in his contribution in addition to TOVE and TheEO to the REA Enterprise Information System. Most recently Thönssen & Wolff (2010) introduced a ContextOntology for more effectively dealing with change in enterprises.

In order to stay focused literature review is restricted to well-known existing *enterprise* ontologies and detailed in the following sections.

3.4.1.1 Toronto Virtual Enterprise (TOVE)

The goal of the TOVE Enterprise Modeling project is to create an ontology that is a 'common sense model' of an enterprise (Fox, Barbuceanu, Grüninger, & Lin, 1996). With common sense the authors mean an enterprise model that is able to deduce information based on rather shallow knowledge of the domain. Actually, TOVE consists of a set of ontologies for modeling enterprises and public administrations alike[87], as depicted in Figure 15. Entities in TOVE are structured into taxonomies and the definitions of objects, attributes and relations are specified in first-order logic; its semantics is implemented in a set of Prolog axioms (Fox & Gruninger 1998).

---

[79] Basic Formal Ontology (BFO). BFO was developed initially by Barry Smith and Pierre Grenon and comprehensive material as well as publications are provided at the web-site. URL: http://www.ifomis.org/bfo/ (retrieved: 10.6.2011).

[80] CYC was developed within the Cyc project (founded in 1984 by D. Lenat). CYC is considered "an expert system with a domain that spans all everyday objects and actions" (Lenat, 1995, p 33). More information can be found on Cycorp web-site, which was founded to further develop, commercialize, and apply the Cyc technology. http://www.cyc.com/ (retrieved: 10.6.2011).

[81] a Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE). URL: http://www.loa-cnr.it/DOLCE.html (retrieved: 10.6.2011). DOLCE has been developed within the European funded IST-project WonderWeb: Ontology Infrastructure for the Semantic Web (WONDERWEB). URL: http://wonderweb.man.ac.uk/ (retrieved: 10.6.2011). DOLCE is introduced in the WonderWeb Deliverable D17 Preliminary Report (Masolo et al., 2003).

[82] General Formal Ontology (GFO). CEO is considered a top-level ontology for conceptual modelling developed by Onto-Med. Although regarded as an upper ontology GFO is applied in the field of biomedical science. More information on ontologies in medicine and life sciences foundations can be found on the Onto-Med web-site. URL: http://www.onto-med.de/ontologies/gfo/index.jsp (retrieved: 10.6.2011).

[83] PROTo Ontology (PROTON) has been developed in the scope of the SEKT Project. "PROTON is a development of the KIMO ontology, which had been created and used in the scope of the KIM platform for semantic annotation, indexing, and retrieval." URL: http://proton.semanticweb.org/ (retrieved: 10.6.2011).

[84] The ontology is based on the book Knowledge Representation by John F. Sowa, available via his web-site. "The basic categories and distinctions have been derived from a variety of sources in logic, linguistics, philosophy, and artificial intelligence." URL: http://www.jfsowa.com/ontology/ (retrieved: 10.6.2011).

[85] Suggested Upper Merged Ontology (SUMO) " and its domain ontologies form the largest formal public ontology in existence today. They are being used for research and applications in search, linguistics and reasoning." URL: http://www.ontologyportal.org/ (retrieved: 10.6.2011). A download of SOMA in owl-format is available on the web-site.

[86] 'TheEO' is an abbreviation I introduce for better reading and to avoid confusion with the acronym EO, I use as a general abbreviation for enterprise ontology.

[87] At the project's web-site for each ontology a link to a paper that defines the ontology is provided, as well as links to additional papers. URL: http://www.eil.utoronto.ca/enterprise-modelling/tove/ (retrieved: 11.6.2011).

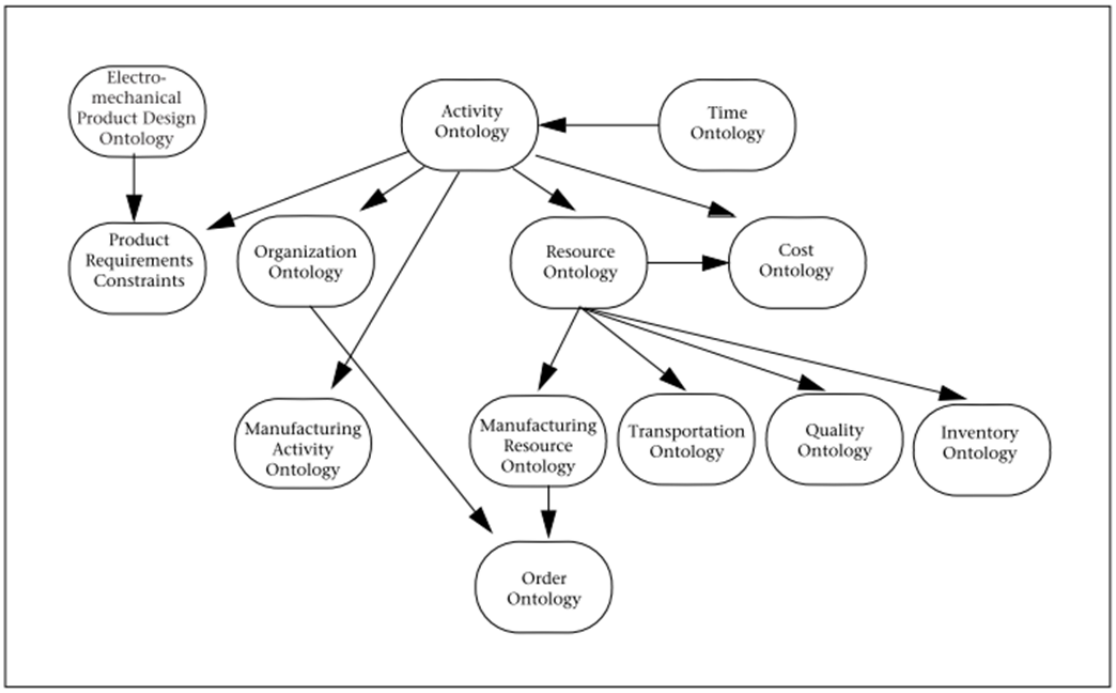


Figure 15: Toronto Virtual Enterprise Ontologies (Fox & Grüninger, 1998)

Fundamental for the TOVE enterprise model are the ontologies of time and action, used to represent the behaviour of an organisation. „An important component of representing behaviour is the ability to temporally project, that is, to determine the possible set of future states given a current state“ (Fox, Barbuceanu, Gruninger, et al. 1996, p 6 ff.) However, no explanation is given why behavior is modeled in the ontology instead of leaving it to operational systems like Workflow-Management-Systems, respectively what the advantage is of doing so. The lack of a theoretical basis for the content in general has been observed recently by O'Leary (2010).
Born et al. (2008) criticize the inconsistency of granularity of the ontologies making it inoperable to use. Filipowska et al. (2009, p 2) repeat the conclusion, stating that „the granularity of developed ontologies may be perceived inconsistent and this hampers their potential application“.
To my knowledge TOVE has not been developed further in recent years. Last update of TOVE web-site was in 2002 and all up-to-date research papers basically compare new approaches with TOVE.

3.4.1.2 The Enterprise Ontology (TheEO)

TheEO was developed as part of the Enterprise Project, an initiative of UK's government to promote the use of knowledge-based systems in enterprise modelling[88].
Although TheEO has been developed largely from scratch it was inspired and influenced by other projects and efforts, for example by TOVE (Uschold et al., 1997). TheEO is thought of as "*one* [emphasis by the author] set of terms and definitions which adequately and accurately covers the relevant concepts in the enterprise modelling domain" (Uschold et al. 1997, p 2) and can be extended to meet particular requirements.

[88] Detailed information on the project and TheEO can be found at the project's web-site. URL: http://www.aiai.ed.ac.uk/project/enterprise/ (retrieved: 13.6.2011).

A meta-ontology has been used to develop the Enterprise Ontology itself, consisting of the concepts “entity”, “relationship”, “state-of-affairs” and “role”. The enterprise meta-ontology has been kept as small as possible; TheEO consists of something more than one hundred concepts. Definitions for the sections as well as a complete list of the concepts including explanations and relations between concepts can be found in Uschold et al. (1997). Figure 16 depicts the conceptual model of TheEO comprising the meta-ontology plus the five sections as defined by Uschold et al. (1997, p 10).

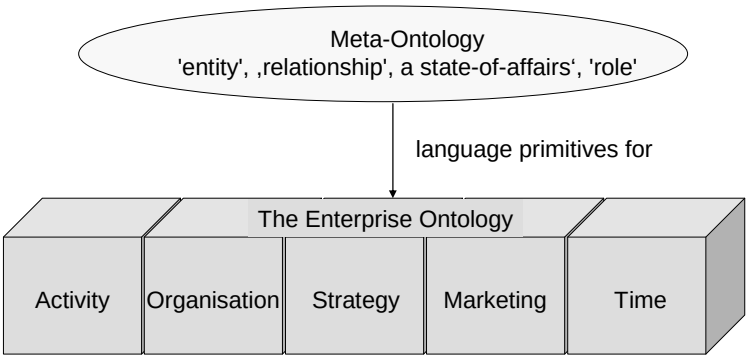


Figure 16: TheEO Conceptual Model (own presentation)

TheEO is represented both in informal way (text version) and in the formal language of Ontolingua[89], giving "the full expressive power of first-order logic" (Uschold et al. 1997, p 35). Besides the content of TheEO the authors provide guidelines for the transformation process from natural language definitions into formal ones. The existence of TheEO in a carefully created natural language glossary (together with the Ontolingua form) makes the ontology readable to non-technical readers which enterprise stakeholders mostly are. Starting ontology creation with natural language definitions is a common approach proposed by others, too, like Gómez-Pérez et al. (2004) and Feldkamp et al. (2010).

3.4.1.3 Context-Based Enterprise Ontology (CbEO)

Although the Context-Based Enterprise Ontology is not as well-known as the before mentioned TOVE and TheEO it is worth considering as it is based on contextual approach. According to Leppänen (2005, p 17) "A thing gets its meaning through the relationships it has with the other things in that context" and thus he proposes "that the semantic and pragmatic interoperability of applications in enterprises should be advanced by the more explicit use of context and other contextual concepts in enterprise ontologies". An approach known from research on Semantic Networks in the 1970ies and 80ies, for example by (Woods, 1975).
The Context-Based Enterprise Ontology is regarded a top-level ontology according to the definition given by Guarino (1998), providing a "conceptualization of the structure and behavior of the enterprise through considering things as contexts, and/or as parts thereof" (Leppänen, 2005, p 23).

[89] Ontolingua is a representation language for ontologies based on the Knowledge Interchange Format (KIF) (Uschold et al., 1997).

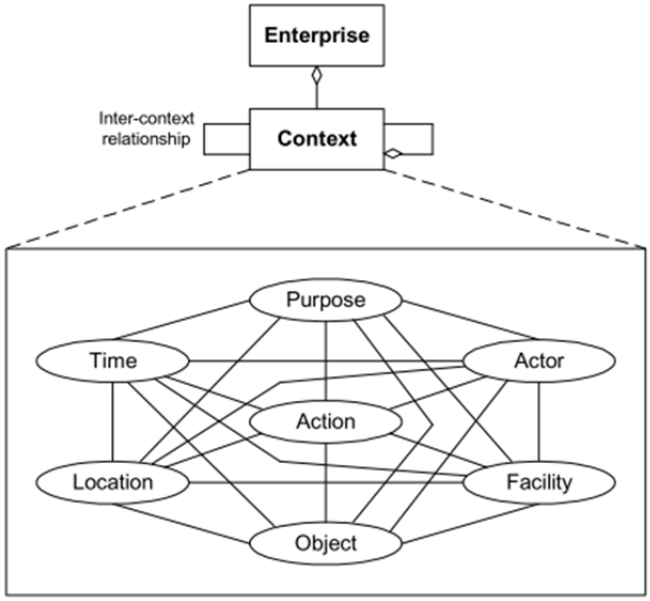


Figure 17: Overall Structure of the Context-Based Enterprise Ontology (Leppänen 2005, p 18)

The set of contextual concepts – Leppänen calls them domains – as depicted in Figure 17: Overall Structure of the Context-Based Enterprise Ontology **(Leppänen 2005, p 18)**, is based on relevant theories (like case grammar and activity theory). Its structure and relations follow the 'seven S's scheme': "For Some purpose, Somebody does Something for Someone, with Some means, Sometimes and Somewhere" (Leppänen 2006, p 274). The ontology can be enhanced to meet special needs of the enterprise, but still maintaining connections of the specialized things to their contexts (Born et al., 2008).

The concepts in the Context-Based Enterprise Ontology are defined in English and represented in UML-based class diagrams. Although concepts and their relations are strictly defined the ontology is considered a light-weight ontology in the sense that it is not formalized[90]. Using UML to represent the ontology has the advantage of being easy to understand but the drawback of not being expressive enough. To overcome that hindrance it has to be translated into more formal language, like RDF but integration of UML into ontological tools is still in a development stage (Gómez-Pérez, Fernández-López, & Corcho, 2005). According to Leppänen elaboration of the ontology towards a more formalized form was intended but has been put back because of research on other issues[91].

3.4.1.4 Core Enterprise Ontology (CEO)

Bertolazzi et al. (2001) propose a methodological frame to construct enterprise ontologies. The authors criticise existing approaches that either concentrate on knowledge representation, i.e. on representation languages for ontologies, or on selecting the right concepts from a knowledge domain. Instead Bertolazzi et al. (2001, p 104) introduce a Core Enterprise Ontology (CEO), comprising a categorisation of the enterprise concepts and a first proposal of an upper ontology. CEO then builds the starting point for a specific enterprise ontology "proceeding top-down in the refinement and decomposition hierarchies".

[90] As stated by Mauro Leppänen in a personal e-mail to me dated from 8.2.2010.
[91] ditto

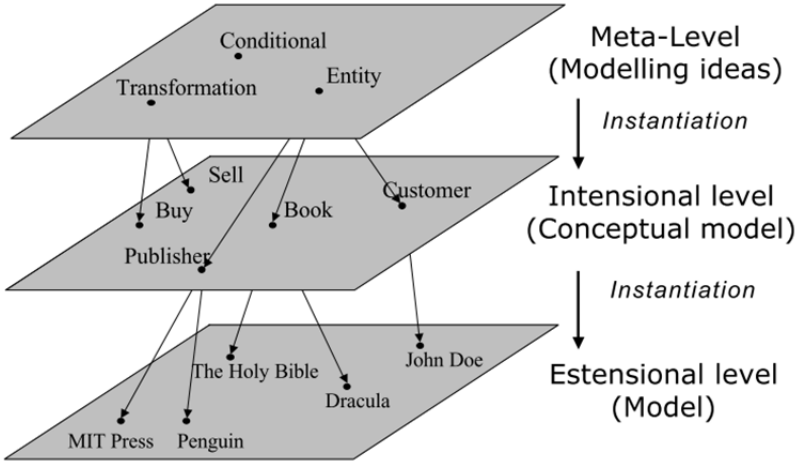


Figure 18: The Three Levels of CEO (Bertolazzi et al. 2001, p 105)

Figure 18 presents the three levels of CEO. Bertolazzi et al. (2001) draw upon knowledge modelling methods providing a three level architecture. The authors compare the content of the layers with database architecture: the lower level holds information about individuals (i.e. the database entities), the middle layer consists of the concepts (i.e. the database schema) and the top level "corresponds to the definition of the schema of the database" (Bertolazzi et al. 2001, p 105). Unfortunately, besides the analogy to database modelling, the problem of 'an instance of an instance' (Brunner et al., 2007), when relating individuals of the ontology to entities of business databases, e.g. an ERP system, is not considered. In other words: If an enterprise ontology is to be used in a business application, for example to infer similarities between customers who bought the same book some information about the book will be stored in the ontology but also information – possibly the same – is stored in the ERP system (styling, price, stock, delivery conditions etc.). In this case an instance of the book is stored in the ontology *and* a record (an instance) of the *same* book is stored in the EPR's database. Since my approach has to deal with that problem too (metadata about documents are stored in the ontology but the documents and potentially other metadata are stored in business information systems like contract management systems), it would have been interesting if Bertolazzi et al. (2001) had provided an approach from a business point of view.
However, goal of the work of Bertolazzi et al. (2001) is to identify a set of general concepts, that have been compared to well-known proposals of methodologies and languages (IDEF5[92], PIF[93] and BEM[94]) for and developments of enterprise ontologies (TOVE, TheEO, MIT[95]). Although Bertolazzi et al. (2001, p 108) claim that they have carefully analysed the before

[92] IDEF5 is the Ontology Description Capture Method, one of several Integrated DEFinition Methods (IDEF) developed and maintained by Knowledge Based Systems, Inc. URL: http://www.idef.com/IDEF5.htm (retrieved: 23.12.2011)
[93] "The PIF Project has been merged with the PSL (Process Specification Language) Project at NIST. The PIF CORE and its extensions have been incorporated into the PSL CORE and its extensions." URL: http://ccs.mit.edu/pif/ (retrieved: 23.12.2011). "The Process Specification Language (PSL) defines a neutral representation for manufacturing processes that supports automated reasoning." URL: http://www.mel.nist.gov/psl/ (retrieved: 23.12.2011)
[94] The Business Engineering Model (BEM) is part of the Open Information Model OIM. "The Meta Data Coalition (MDC) Open Information Model (OIM) is a vendor-neutral and technology-independent specification of core metadata types found in operational, data warehousing, and knowledge management environments." URL: http://xml.coverpages.org/mdc-oim.html (retrieved: 23.12.2011)
[95] MIT Process Handbook. URL: http://process.mit.edu/Directory.asp?ID=970203154850AB5013 (retrieved: 23.12.2011)

mentioned work and "integrated the extracted (most general) concepts with other business concepts that we considered important, and missing in the mentioned proposals", the selection remains arbitrary. For example: in CEO a customer is defined as an actor that receives goods or services in exchange for a consideration (e.g. a payment). This modelling approach carries the risk that an individual or customer may be also an individual or vendor. In that case the same individual would be modelled twice. Of course the two individuals could be linked – e.g. the built-in OWL property owl:sameAs allows for expressing equivalences[96] - but the specifications of actor might as well be regarded as roles. That would solve the problem in a way suggested by the ArchiMate standard[97] I intend to use.
Even though Bertolazzi et al. (2001) consider their generic upper ontology as a starting point and they intended to continue their work, to my knowledge there is no newer version. Whereas the modeling framework of CEO – starting with a core ontology which is gradually refined – is pragmatic approach also discussed by Uschold & Grüninger (1996), Gomez-Perez et al. (2003) and Cardoso (2010).

3.4.1.5 Resource-Event-Agent (REA)

Geerts & McCarthy (2002), propose an enterprise domain ontology based on previous achievements of Geerts, McCarthy and Sowa[98]. In their notion "an enterprise ontology is the conceptualization of the common economic phenomena of a business enterprise unaffected by application-specific demands" (Geerts & McCarthy 2000, p 8). Their work started from an existing conceptual accounting framework based on economic and accounting theory, The Resource-Event-Agent (REA) model.
REA was developed in 1982 and initially was intended as a generalized accounting framework to represent economic resources, events, and agents plus their relationships (McCarthy 1982). Initially REA focused on accounting but has been extended for enterprise's value and for workflow and task specification (Geerts & McCarthy 2002). Geerts & McCarthy (2000, p 8) differentiate between "physical objects describe actual phenomena, while abstractions are information structures that are used to characterize the corresponding physical categories". Thus, the REA ontology comprises conceptualizations of actual economic phenomena (current and future), called operational infrastructure, and conceptualisations of the abstract phenomena, called knowledge infrastructure. REA was originally intended to model the semantics of accounting databases. For its implementation the language Prolog has been used, "to record explicitly both data (operational) and a knowledge infrastructure that consisted of a conceptual schema, a set of declarative primitives, and a taxonomy of shareable and reusable accounting concepts" (Geerts & McCarthy 2000, p 20). As REA has been developed gradually and its extensions are manifold multiple conceptualizations in various formats exist. Thus, Gailly & Poels (2007) propose a new, uniformed and unified representation of REA in UML. The major change Gailly & Poels (2007) carried out was incorporating business domain axioms in the class diagram, "instead of describing them separately (and informally) as in the 'old' REA". Besides UML, OWL DL is proposed for representing REA in order to use the ontology at run-time. In addition SWRL[99] is suggested to formalize and execute business policies as semantic rules.

[96] OWL Web Ontology Language Reference. URL: http://www.w3.org/TR/owl-ref/ (retrieved: 23.12.2011)
[97] "A business role is defined as a named specific behavior of a business actor participating in a particular context" (The Open Group 2009a, p 15)
[98] An overview on work preparing the ground is given in (Geerts & McCarthy 2002)
[99] SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission 21 May 2004. URL: http://www.w3.org/Submission/SWRL/ (retrieved: 25.12.2011)

G. Zhang et al. (2010) also studied the representation of REA in OLW DL. Obviously unaware of the research by Gailly & Poels they add new concepts for an additional strategic layer to the REA infrastructure but completely disregard previous formalization approaches.

Although REA is acknowledged in sciences and economy[100], has been continuously further developed and has been - as shown - recently transformed into standard representation languages like UML and OWL it is of limited use for enterprise modelling. REA, as well as for example e-BMO[101] or e^3-value ontology[102], focuses on *business* modelling, the creation and transfer of economic value (Gailly & Poels, 2007) but not on enterprise *architecture*, i.e. on organizational structure, activities or management. However, it could be interesting to relate an enterprise ontology to REA if dynamic behaviour is to be modelled as for example to better support supply chain management as in the APPRIS[103] project.

3.4.2 Representation Languages for Enterprise Ontologies

Fensel (2004) regards ontology a "silver bullet for knowledge management" and a huge variety of different types of ontologies exist "ranging from simple word lists to comprehensive ontologies with the expressive power of full first-order logic" (De Bruijn 2003, p ii). Thus, expressiveness of ontologies can vary extremely and choosing the most suitable ontology language is a difficult task.
According to Bechhofer et al. (2002, p 7) ontology languages can be classified into
(1) vocabularies defined in natural language,
(2) object-based knowledge representation languages such as frames and UML, and
(3) languages based on predicates expressed in logic such as Description Logics.
Despite the consent about using an ontology for describing an enterprise architecture no agreement has been achieved yet on the appropriate representation language. Extensive research on ontology engineering in general has been done, amongst others by Gruber et al. (1993), Studer et al. (1998), De Bruijn (2003), Sure et al. (2009) and in particular on enterprise ontologies, for example by Uschold & Grüninger (1996), Fox & Grüninger (1998), Leppänen (2005) and Dietz (2006). As result a list of generally accepted requirements for ontology engineering can be derived, comprising criteria like 'clarity', 'coherence', 'comprehensiveness' or 'consistency'. Fox & Gruninger (1998, p 109) regard an enterprise model as "a *computational* representation of the structure, activities, processes, information, resources, people, behaviour, goals, and constraints of a business, government, or other enterprise". That means the enterprise ontology should be represented in a way that enables machines to *process* it.

However, all requirements are basically domain independent and if any, then the requirement for "usability with existing platforms" (De Bruijn 2003, p 18) can be interpreted as an enterprise specific one.
Whereas knowledge representation has been a major topic for long time within the Artificial Intelligence community (Bechhofer et al., 2002), the Enterprise Architecture community has as yet given little attention to executable ontology languages. Most of the approaches in the enterprise architecture domain consider ontology languages of class (1) or (2), as for example (Dietz, 2006), (Lankhorst, 2009) and the ArchiMate standard – the latter two based on UML the first one on a propriety notation.

[100] Examples of acknowledgements are given by (Gailly & Poels, 2007)
[101] e-Business Model Ontology (Osterwalder & Pigneur, 2002)
[102] e^3-value methodology (Akkermans & Gordijn, 2003)
[103] Better support of Supply Chain Management is the goal of the Advanced Procurement Performance and Risk Indicator System (APPRIS), national funded research project.

Even Kang et al. (2010) who proposes an ontology-based enterprise architecture stressing the importance of strictly defined semantics (based upon WordNet and SBVR) do not provide any suggestions for implementation in a language of class (3).

Although Hinkelmann et al. (2010) introduce two systems for modeling (enterprise) ontologies (Athene and Resourcesome[104]), they simply mention OWL-DL – an ontology language that can be regarded as class (3) - but do not justify that decision. To my knowledge no enterprise specific requirements have been defined yet for enterprise ontology languages of class (3). Thus it is indeed difficult to decide on the most appropriate language for representing an enterprise architecture description.

For a domain independent overview on early web-based ontology languages and their foundations, for example Ontolingua (which has been used to represent TheEO as before mentioned) please refer to De Bruijn (2003). Another overview on early web-based ontology languages is provided by Gomez-Perez et al. (2003). A more recent introduction is given by Allemang & Hendler (2008). Corcho & Gomez-Perez (2000) provide a framework to compare expressiveness and reasoning ability of ontology languages. Unfortunately – because of the early date of publication – OWL (W3C OWL Working Group, 2004) and its sub-languages were not available for consideration. Su & Ilebrekke (2006) provide a framework for quality evaluation of ontology models and languages. Whereas the evaluation criteria are very useful – I will rely on it for evaluation of my prototype (cf. Chapter 8.2.1 and Chapter 8.2.2) – the comparison itself is outdated. Figure 19 gives a clear illustration of the various ontology languages – with the exception of OWL.

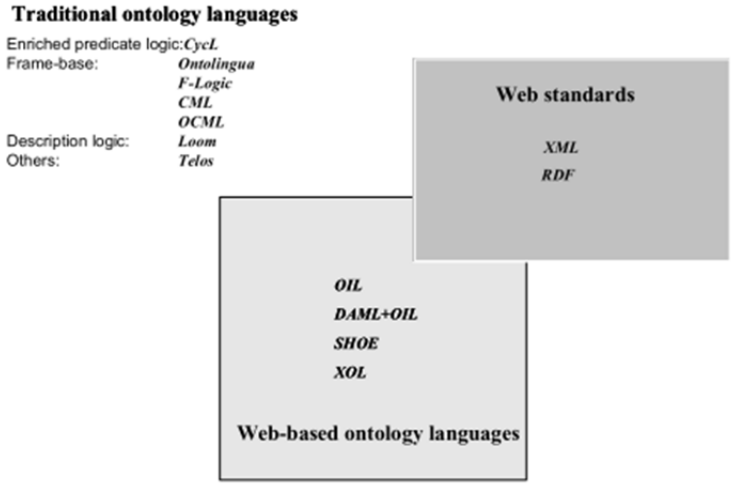


Figure 19: Classification of Ontology Languages (Su & Ilebrekke, 2006)

An easy to understand description of the development of OWL (1) and its improvement to OWL 2[105] is provided by Yu (2011). Grau et al. (2008) also give a clear and comprehensive description of OWL 1 its insufficiency and its requirements for further development to OWL 1.1 which finally became OWL 2.

Whereas OWL 1 provides the sub-languages 'OWL Lite', 'OWL DL' and 'OWL Full' to address the different requirements, OWL 2 provides three further sub-languages: 'OWL EL',

[104] ATHENE is an ontology modelling tool developed at FHNW; Resourcesome is a tool developed at UNICAM
[105] OWL 2 is the latest standard of the W3C (W3C OWL Working Group, 2009)

'OWL QL' and 'OWL PL' (W3C OWL Working Group, 2009). Krötzsch et al. (2010, p 114) even claim that "the new version of OWL is the first that adequately addresses the trade-off between logical expressivity and scalability that is inherent to formal knowledge representation by specifying additional light-weight language profiles".

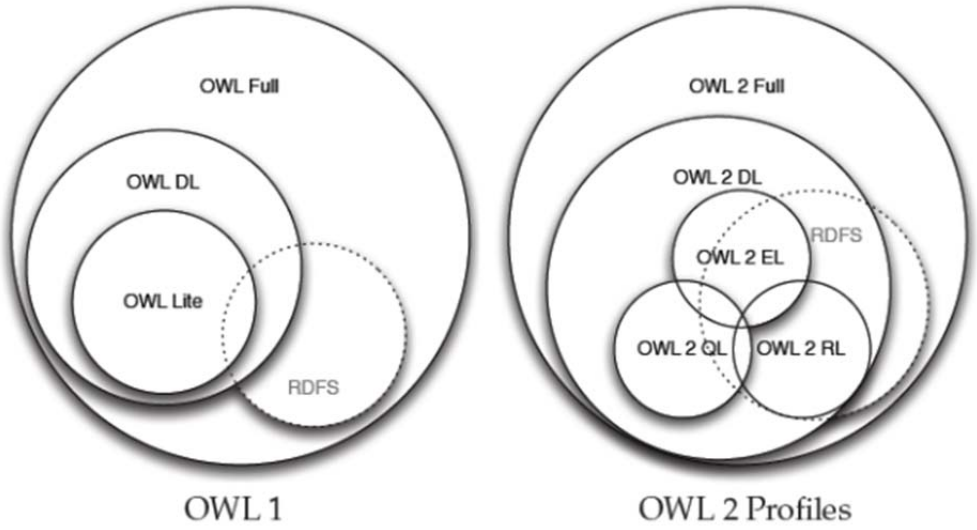


Figure 20: OWL 1 & 2 (OpenStructs TechWiki[106])

Figure 20 gives an overview on the OWL profiles. Each of them provide different expressive power and target different application scenarios (Grau et al., 2008), and therefore addresses Su & Ilebrekke's (2006, p 776) statement: "Which language to choose is dependent of the problem domain and modeling requirement, like how much reasoning support is needed".

3.4.3 Implementing Logical Reasoning

As I want to use the enterprise ontology operationally, i.e. used by machines, an important issue is the tradeoff between expressiveness and the efficiency of the reasoning process (Yu, 2011). I will address this issue in the following sub-section. Another challenge for the operational use of the enterprise ontology is the distributed information. As already mentioned in an enterprise most likely not all information will be stored in the ontology but remain in business applications, i.e. in non-ontological data stores. Research on combing both ways of storage is investigated in the second sub-section.

3.4.3.1 Reasoning on Ontologies

According to Corcho & Gomez-Perez (2000) there is a clear distinction between knowledge representation and reasoning for all languages. It is common knowledge, that OWL Full is the most expressive of the web ontology languages but not. On the other hand RDF and RDF Schema were developed in the early days of the Semantic Web and many reasoners have been developed since (Zhang 2005) - but these languages were quickly considered too limited in expressive power (Grau et al. 2008)[107].

[106] OpenStructs TechWiki. URL: http://techwiki.openstructs.org/index.php/Metamodeling_in_Domain_Ontologies (retrieved: 25.12.2011)

[107] In their work on the development of the web language Horrocks et al. (2003, p 3) give several examples, e.g. "OWL classes can be specified as logical combinations (intersections, unions, or complements) of other classes, or as enumerations of specified objects, going beyond the capabilities of RDFS".

Wang et al. (2004) differentiate between 'ontology reasoning' and 'user-defined reasoning'. OWL-Lite ontology reasoning is used to infer implicit context from low-level, explicit context information. In the approach of Wang et al. (2004) explicit context information is provided by sensors (e.g. a person is in a bathroom); the respective implicit information is that the person is in her home building. First-order logic is used to define application specific, user-defined rules, for example if a person is in her bedroom, and the light is low and the curtain is drawn then the person is sleeping.

A comprehensive introduction to reasoning with rules and ontologies is provided by Eiter et al. (2006). The authors give a slightly simplified figure of the 'Semantic-Web layer cake' emphasizing that it is not yet completely clear where and how to fit in rules and therefore depict rules and ontologies side by side.

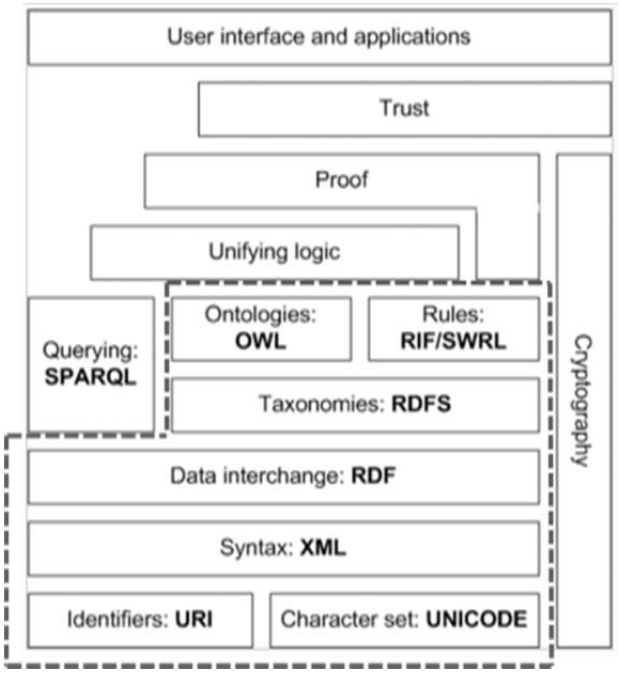


Figure 21: Semantic Web Layer Cake[108]

Figure 21 complements the illustration of Eiter et al. as SPARQL[109] is added to the considered layers, a W3C recommendation to query RDF data. The illustration also defines rules more precisely by naming RIF[110] and SWRL[111].
Eiter et al. (2006, p 123) present a number of approaches for combining rules and ontologies but "the quest for the Holy Grail of an ideally suited formalism (which might not exist) is still on-going". Eiter et al. (2008) complete previous work by giving an overview of existing languages and systems implementations, of their features and of the theoretical approaches they build upon.

[108] The figure is taken from Semantic Web Stack provided by Wikipedia. The dashed boarder indicates the elements presented by Eiter et al. (2006, p 95). URL: http://en.wikipedia.org/wiki/Semantic_Web_Stack (retrieved: 8.1.2012)
[109] SPARQL Query Language for RDF. W3C Recommendation 15 January, 2008. URL: http://www.w3.org/TR/rdf-sparql-query/ (retrieved: 8.1.2012)
[110] Rule Interchange Format (RIF) is work in progress of W3C Working Group (Note 22 June 2010). It aims to create a standard for exchanging rules among rules systems but not to develop a single rule language. URL: http://www.w3.org/TR/rif-overview/ (retrieved: 8.1.2012)
[111] Semantic Web Rule Language (SWRL). W3C Member Submission (21 May 2004). URL: http://www.w3.org/Submission/SWRL/ (retrieved: 8.1.2012)

Xian Yi Cheng et al. (2011) describe the principles of semantic reasoning about OWL and briefly introduce some reasoning machines. Xian Yi Cheng et al. (2011) give an example of practical reasoning about an OWL DL ontology but the ontology that is used is tiny and does not in the least reflect the complexity of a 'working ontology'. OWL 2 improvements are not considered at all.

Shi et al. (2011) also compare ontology reasoning systems but focus on the specific issue of scalability with respect to custom rules (i.e. user defined rules). In particular, the authors analyse SPARQL as the standard query language for RDF and SWRL, a web rule language to express custom rules. They analyse the scalability of Jena[112], Pellet[113], KAON2[114], Oracle 11g[115], and OWLIM[116] focussing on the issue of supporting user defined rules. Others, for example Sirin et al. (2006), Lee et al. (2008) and more recently Meditskos & Bassiliades (2010) study these systems in terms of complexity of the ontology and the number of its entities.

An overview on implemented reasoners can be found at the OWL working group web-site[117].

In 2011 a submission request was made to W3C for the SPARQL Inferencing Notation (SPIN). "Based on RDF, RDF Schema and SPARQL, SPIN implements a rule and constraints language […] that is easy to implement and understand. SPIN uses an RDF based syntax for SPARQL that makes it possible […] to link RDFS/OWL classes with rules and constraints that formalize the semantics of the instances of those classes. […] SPIN also provides a vocabulary to describe executable SPARQL functions that can be used to extend the range of available functions to a SPARQL processor in a platform independent way. Finally, SPIN includes a mechanism to formalize and share SPARQL query templates. These templates make it possible to build higher level modelling languages from reusable building blocks. "[118]

In their contribution Fürber and Hepp (2010) give a good example of the power of the language, showing how SPARQL and SPIN can be utilized to identify data quality problems in Semantic Web. Spohr et al. (2012) investigate in their paper the use of SPIN to formalize accounting regulations, specifically the representation of accounting regulations as rule constraints. They showed how "SPIN enhanced formalization enables inferencing of financial statement facts associated with financial reporting concepts and sophisticated consistency checks, which evaluate the correctness of reported financial data with respect to the calculation requirements imposed by accounting regulation". Spohr et al. (2012) as well as Fürber and Hepp (2010) showed the appropriateness of SPIN for formal rule represention – though not saying that another rule language like SWRL couldn't also be used. The TopSPIN

[112] Jena is a Java framework for building Semantic Web applications. Jena provides a collection of tools and Java libraries for developing applications, tools and servers. Apache Jena. URL: .http://incubator.apache.org/jena/ (retrieved: 10.1.2012); see also (Meditskos & Bassiliades, 2010) proving an approach for combining Jena and Pellet.
[113] Pellet is an OWL 2 Reasoner for Java. URL: http://clarkparsia.com/pellet/ (retrieved: 10.1.2012)
[114] KAON2 is an infrastructure for managing OWL-DL, SWRL, and F-Logic ontologies. URL: http://kaon2.semanticweb.org (retrieved: 10.1.2012); comprehensive literature on KAON is provided on the web-site.
[115] Oracle Database 11g Release 2. URL: http://www.oracle.com/technetwork/database/enterprise-edition/overview/index.html (retrieved: 10.1.2012)
[116] OWLIM is a family of RDF database management systems. URL: http://www.ontotext.com/owlim (retrieved: 10.1.2012); comprehensive literature is provided on the web-site
[117] W3C Implementations of Reasoners. URL: http://www.w3.org/2007/OWL/wiki/Implementations (retrieved: 15.5.2012)
[118] SPARQL Inferencing Notation (SPIN). Submission request to W3C 22 February, 2011. URL: http://www.w3.org/Submission/2011/02/ (retrieved: 30.7.2012)

reasoner has a built-in functionality provided by TopBraid[119], an ontology modelling environment of TopQuadrant, one of the W3C members who submitted SPIN.

3.4.3.2 Conjunctive Reasoning

In order to use an enterprise ontology (seEAD) operationally, for example for contract management, the capability to reason about an ontology and data stored in a *'non-ontological way'*, e.g. in a relational database, is important (an example is given in Chapter 3.4.1.4). Kontchakov et al. (2010) point out that novel ways of using ontologies go beyond "the 'classical' reasoning tasks such as satisfability and subsumption" and ontology-based data access will be of particular interest (Kontchakov et al. 2010, p 247). In particular they investigated how integration can be supported by storing background knowledge on database and related information systems in an ontology. The authors give an overview on the main approaches for accessing Relational Databases (RDB) – but surprisingly do not consider KAON2 although it allows for reasoning on ABox[120] axioms stored in a RDBMS (Motik & Sattler, 2006). Kontchakov et al. (2010) also show that extending reasoning on RDBMS was the grounds for developing a new family of Description Logics, DL-Lite, and subsequently OWL 2 and the OWL 2 profile QL.

Relating databases and ontologies has been investigated from the very beginning of the semantic web, focussing on the most representative and widespread technologies, namely relational databases (RDB) and ontologies (Spanos, Stavrou, & Mitrou, 2011). Approaches to combine both technologies have become known as 'database to ontology mapping problem', or more generally characterized as 'object-relational impedance mismatch problem' (Spanos et al., 2011). The problem that is to be solved lies in the structural difference of the relational and object-oriented models. It has been studied from different points of view for various kinds of reason (Auer, Feigenbaum, Miranker, Fogarolli, & Sequeda, 2010), like semantic annotation of dynamic web pages, heterogeneous database integration, mass generation of Semantic Web data or ontology learning (Spanos et al., 2011). In the beginning of the Semantic Web mainly for finding a way of efficient ontology storage (Beckett & Grant, 2003). Drawing upon the mature techniques of RDB management, systems optimized for persistent storage, maintenance and querying of ontologies have been developed, known as 'triple stores'[121] (Spanos et al., 2011). Starting from the other side, much research has been done on transforming relational data into an ontological representation (amongst others by Maedche & Staab 2001, Volz et al. 2004 and more recently Būmans & Cerans 2010). Specific interest here was on database integration – or more broadly: information system integration – and ontology population and learning. Because of the multifaceted issues involved in mapping RDB and RDF the W3C launched the RDB2RDF incubator group (Sahoo et al., 2009). The group had the two objectives: to examine and classify existing approaches and to examine and classify the approaches in associating OWL classes to SQL queries (ibid.) Sahoo et al. (2009) distinguish between 'Automatic Mapping Generation' and 'Domain Semantics-driven Mapping Generation'. Whereas the first method directly maps RDB and

[119] TopBraid Composer is available as Free Edition, Standard Edition and Maestro Edition. URL: http://topquadrant.com/products/TB_Composer.html (retrieved: 30.7.2012)

[120] "A DL knowledge base is typically partitioned into a terminological (or schema) part, called a TBox, and an assertional (or data) part, called an ABox" (Motik & Sattler 2006, p 227). Compared to database concepts, ABox equals the database instances, TBox the database schema.

[121] A triple store is as database specifically tailored to store RDF statements, also called triples. A list of triple store implementation can be found on Wikipedia. URL: http://en.wikipedia.org/wiki/Triplestore (retrieved: 26.1.2012); An overview on large triple stores is provided by the W3C. URL: http://www.w3.org/wiki/LargeTripleStores (retrieved: 26.1.2012)

RDF schemas, the latter considers "domain semantics that is often implicit or not captured at all in the RDB schema" (Sahoo et al. 2009, p 5). A very helpful and easy to read introduction on how RDF and SPARQL can be used to improve access to relational databases is provided in ("RDF and SPARQL: Using Semantic Web Technology to Integrate the World's Data," 2007); another introduction on database technologies for RDF is given by (Das & Srinivasan, 2009).
To stay focused I refer to the W3C RDB2RDF Working Group who provide many publications on automatic mapping, for example a strategy for direct mapping relational data to RDF (Arenas, Bertails, Prud'hommeaux, & Sequeda, 2011), and a language specification (R2RML) to express customized mappings from relational databases to RDF datasets (Das, Sundara, & Cyganiak, 2011).

(Hert, Reif, & Gall, 2001) developed a framework to compare the state-of-the-art RDB-to-RDF mapping languages and report the findings in their paper. A comprehensive list of direct mapping implementations has also recently been published by the W3C Working Group (2011). Latest research on 'Domain Semantics-driven Mapping Generation' has been done for example by Vavliakis et al. (2010) introducing RDOTE[122]. RDOTE provides a graphical user interface for mapping multiple relational databases into different ontology schemata and integrate them into a single ontology file.

Spanos et al. (2011) base their comprehensive and contemporary survey on the various approaches on a clearly defined, reasonable classification schema. The first major division is on whether a new ontology is to be created or an existing ontology (and an existing database) is to be mapped. (Ghawi & Cullot, 2007) took a similar starting point but remain more general in their classification of mapping approaches.

Although the vast majority of approaches for database to ontology mapping are for creating ontology (cf. Spanos et al. 2011), mapping existing ontologies to existing databases has been also topic of research for more than a decade. Wache et al. (2001) differentiate between single, multiple and hybrid mapping. For single ontology to database mapping one global ontology is considered to which all information systems are related. Multiple ontology to database mapping is an approach in which each database is related to its own ontology and relations between the ontologies are created to express the relationships amongst them. In a hybrid approach single and multiple mapping is mixed as there exist mappings between a database and its (local) ontology plus mappings between the local ontologies and a global one.

Since there is a plethora of approaches addressing database to ontology mapping and some excellent papers on the subject are available, for example of the abovementioned (Barrasa et al. 2004, Kontchakov et al. 2010 and Spanos et al. 2011), as well as a collection of links is provided by the W3C[123], I restrict myself to a few works which were in particular mentioned in the literature.

In 2004 Barrasa et al. introduce R2O, a language to describe mappings between relational database schemas and ontologies implemented in RDF(S) or OWL. Their considerations are based on the assumption that database and ontology models are likely to be different and both

[122] RDOTE is available under the GNU/GPL license. URL: http://sourceforge.net/projects/rdote/ (retrieved: 4.2.2012). A brief introduction on the mapping process is given in a short video at YouTube. URL: http://www.youtube.com/watch?v=pk7izhFeuf0 (retrieved: 4.2.2012)
[123] W3C: RdfAndSql. URL: http://www.w3.org/wiki/RdfAndSql (retrieved: 12.12.12) and Benchmarking RDB-to-RDF Tools. URL: http://www.w3.org/wiki/RdfStoreBenchmarking (retrieved: 5.2.2012);

pre-exist and are not created specifically for this purpose. R2O is a database independent high level language for defining mappings that are to be executed by tools, middleware APIs, etc. (Barrasa et al., 2004). R2O has been used in the context of the ESPERONTO project, in particular for the Fund Finder application[124], for migrating relational database content about funding opportunities to the Semantic Web. The ontology was populated with instances extracted from the DB using R2O and ODEMapster (Barrasa et al., 2004). A case study of database-to-ontology mapping with the Fund Finder application and the ODEMapster mapping processor is given by Barrasa et al. (2003).

Konstantinou et al. (2006) developed the tool VisAVis for mapping the relational database contents to the TBox of the ontology which does not contain an ABox, but instead references to the dataset in the database. The approach of Konstantinou et al. (2006) is to enhance the initial ontology by references to datasets. "These references will be under the form of class properties in the ontology, all as- signed as value a string containing the actual SQL query that returns the dataset" (Konstantinou et al. 2006, p 1054). For semantic query execution they introduce a layer of interoperability that allows for checking if a mapping property exists; if so the query is redirected to the database. This hybrid approach leaves the ontology and the mapped database untouched but adds semantics to existing systems (Konstantinou et al., 2006). VisAVis comes with a graphical user Interface for entering a query; implementation is done as a plug-in on top of the Protégé[125] and connections to the database are supported via JDBC. The ontology is represented in OWL-DL and semantic web query language RDQL[126] has been preferred to SPARQL.

D2RQ is a declarative language to describe mappings between relational databases and RDF-S/OWL ontologies. The D2RQ platform[127] has been jointly developed at the Freie Universität and Technische Universität Berlin. It provides the D2RQ Engine, a plug-in for the Jena and Sesame Semantic Web toolkits, the D2R Server, an HTTP server and links to literature about the topic. Work on D2RQ already started in 2004 (Bizer & Seaborne, 2004) and has been continuously improved since then. Spanos et al. (2011) consider D2RQ one of the most prominent tools in the field of relational database to ontology mapping. D2RQ supports both ways of data access: ETL and OBDA. D2RQ works with several RDBSs like ORACLE and Microsoft SQL Server and MySQL. The approach is detailed in a comprehensive manual (Bizer, Cyganiak, Garbers, Maresch, & Becker, 2009). TopBraid Composer uses D2RQ and provides an interface to arbitrary relational databases, so that the databases are treated as a (read-only) triple store. The approach is based on mapping files that declare how tables in the database are mapped to instances of an ontology. Lessons learned about D2RQ, used with TopBraid Composer, have been reported by Bizer and Cyganiak (2007) in their position paper to the W3C Workshop on RDF Access to Relational Databases[128]. Unfortunately D2RQ is not available in TopBraid Composer Free Edition.

[124] Unfortunately the provided internet link (URL: http://www.esperonto.net/fundfinder) is not available anymore.
[125] The Protege plugin can be found at URL: http://www.cn.ntua.gr/~nkons/VisAVisTab.jar; (retrieved: 12.12.12) the source code at URL: http://www.cn.ntua.gr/~nkons/source.zip (retrieved: 1.2.2012)
[126] RDQL is the query language of Jena, the framework upon which Protégé is built. RDQL and the Web API have been submitted to the W3C for standardization in October 2003 (J. J. Carroll et al., 2003)
[127] The D2RQ Plattform - Treating Non-RDF Databases as Virtual RDF Graphs. URL: http://www4.wiwiss.fu-berlin.de/bizer/d2rq/ (retrieved: 4.2.2012)
[128] W3C Workshop on RDF Access to Relational Databases 25. – 26 October, 2007 — Cambridge, MA, USA. URL: http://www.w3.org/2007/03/RdfRDB/ (retrieved: 31.7.2012)

(Vavliakis, Symeonidis, Karagiannis, & Mitkas, 2011) introduce Iconomy, an integrated framework for manipulating and querying data residing both in ontologies and relational databases. "It provides advanced options on the creation and synchronization of an ontology to and from a relational database, the automatic creation of queries, and data viewing/editing" (Vavliakis et al. 2011, p 3846). Iconomy supports direct transformation of relational database entries to its respective ontology instances via a simple and friendly graphical user interface. Iconomy enables on-demand incorporation of the well-established triplestores (Jena and Sesame) and the Pellet reasoner.

Another major development is OpenLink Virtuoso, "a multi-purpose and multi-protocol (Hybrid) Data Server from OpenLink Software[129] that includes SQL Object-Relational, RDF, XML, and Free Text data management, alongside Web Application (HTTP, SOAP, WebDAV), SyncML, and Discussion Server functionality, in a single server" (Erling & Mikhailov 2009, p 7). Comprehensive information about Virtuoso is provided in the wiki of the W3C (2009). According to Spanos et al. (2011) functionality of Virtuoso is similar to D2RQ's.

(Calvanese et al., 2011) introduce MASTRO, a Java tool for ontology-based data access (OBDA). In combination with rewriting queries for RDBMS access MASTRO manages OBDA systems through semantic mappings. The ontology is also specified in DL-Lite but specifically tailored to ontology-based data access. The semantic mapping associates SQL queries about the external data to the elements of the ontology. The authors verified the approach successfully on real-life data but admit there remain several open issues, for example the exponential blow-up in rewritten queries.

Because of the manifold approaches for database to ontology mapping and complexity of the topic a sound evaluation goes beyond the scope of my thesis. A feasibility study of relational database to ontology mapping in enterprises has been conducted within a master thesis of (Akabuilo, 2012), co-supervised by me. Akabuilo elaborates on how direct mapping can be done for relational data to ontological data, represented in OWL 2 QL. For proof of concept he implemented a prototype based on R2RML technique for customized mappings from relational databases to RDF datasets. However, Akabuilo (2012, p 79) also showed that "practical implementation of OWL 2 QL is not possible because the profile depends on toolset implementation. Practitioners could only leverage the OWL 2 QL if it is implemented in a tools set. At the moment the only tool that supports QWL 2 QL implementation is QuOnto from Università di Roma".

3.4.4 Ontology Engineering

According to De Bruijn (2003) many ontology engineering methodologies do exist and some general design principles can be identified. All of the before mentioned enterprise ontologies come with methods for their development and a good and condensed overview is provided by Gomez-Perez et al. (2003). The authors introduce the various methodologies and methods for building ontologies and present some well-known approaches like the Cyc method, Uschold and King's method, the METHONTOLOGY etc. However, there is still no "explicit and totally documented conceptual model upon which the ontology is built" (Gómez-Pérez, Fernández, & de Vicente, 1996).

[129] OpenLink Virtuoso Universal Server: Documentation – Contents. URL: http://docs.openlinksw.com/virtuoso/contents.html (retrieved: 5.2.2012)

Arguing that ontology development is part of an IT project, Feldkamp et al. (2010) provided a comprehensive approach for ontology creation as an element of an ontology's lifecycle. Gomez-Perez et al. (2003) give a similar view on the developing process but add upstream management activities, like scheduling, control or quality assurance and downstream support activities, e.g. configuration management.

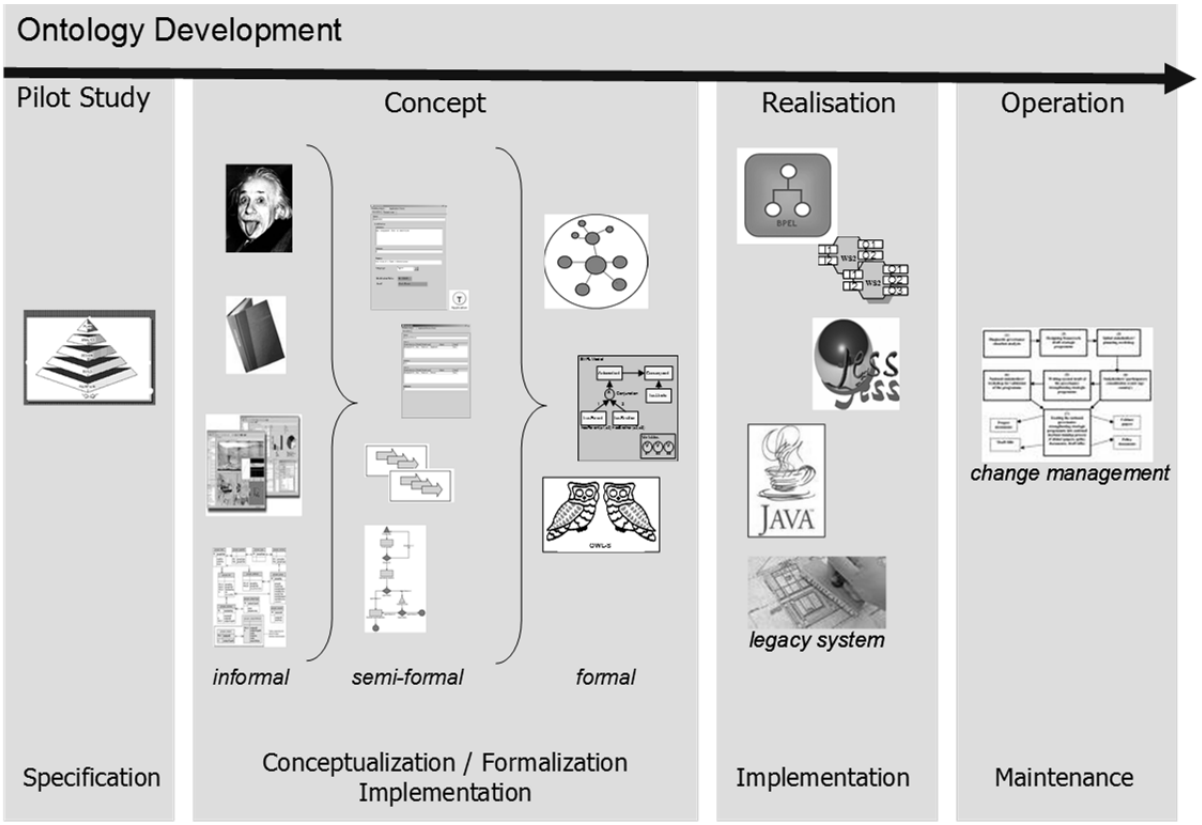


Figure 22: Ontology Lifecycle (Feldkamp et al. 2010)

Figure 22 depicts the ontology lifecycle as introduced by Feldkamp et al. (2010). In the first phase, the pilot study, the purpose and goal of the project should be investigated with respect to the business strategy. This phase is called 'specification' in other publications it is called 'capture motivating scenarios' in TOVE (Fox & Grüninger, 1997) or 'identify purpose and scope' by Uschold & King (1995) or 'requirement specification' in METHONTOLOGY (Fernandez, Gomez-Perez, & Juristo, 1997).

Within the concept phase ontology development is performed comprising three levels of formalisation: informal (knowledge is captured in natural language), semi-formal (knowledge is represented in a semi-formal way e.g. in structured templates) and formal (knowledge is strictly formalized in OWL, SWRL and OWL-S) (Martin et al. 2004). This approach of increasing formality is also suggested by Fox & Grüninger (1997) and Gómez-Pérez et al. (1996).

The realisation phase then, is about implementation of ontologies, making ontologies accessible out of program code (e.g. legacy systems or web services). Finally, maintenance is performed in the phase operations. I will adapt that approach for my procedure model and apply it to the development of the enterprise ontology. Thus, specification and conceptualization is performed within Action Research loop 1, formalization (semi and fully formal) is done within loop 2 and implementation is done within loop 3.

Whereas I draw upon the method for ontology development introduced by Feldkamp et al. (2010) I don't consider the suggested technique appropriate. As ontology engineering is time-consuming (knowledge acquisition, consensus building, formalization, etc.), enterprise can barely afford building an ontology from scratch. Thus, most of the approaches recommend *not* starting all over again but to draw upon existing work.
Fox & Gruninger (1998, p 110) suggest a so called "General Enterprise Model" (GEM), that is an "object library that defines the classes of objects that are generic across a type of enterprise, such as manufacturing or banking, and can be used (that is, instantiated) in defining a specific enterprise". The authors see the benefits of that approach in not having to start from scratch (but using the predefined object library) to ensure quality with respect to completeness and to improve a shared understanding of an enterprise model.

Bertolazzi et al. (2001, p 104) also believe that it is useful to start "with a few, well established, general concepts that will guide business experts in defining their enterprise ontology" and thus propose a Core Enterprise Ontology (CEO) independent of the specific domain to serve that purpose (cf. 3.4.1.4). As detailed in Chapter 5.1.2 I adapt that approach for the development of the enterprise ontology.

Another common approach in ontology engineering is answering competency questions. Fox et al. (1996, p 134) consider competency questions as "benchmarks in the sense that the ontology is necessary and sufficient to represent the tasks specified by the competency questions and their solution." This technique has been used to build the TOVE ontology and since then been assumed amongst others by Uschold & Grüninger (1996), De Bruijn (2003) and Abramowicz et al. (2007).
In general, competency questions are formulated in natural language to determine the scope and evaluate appropriateness of an ontology. Questions and answers lead to the required concepts and their properties, relations and axioms of the ontology. In a further step then, competency questions are (re-)written in a formal way, Grüninger & Fox (1995) used first-order logic to specify terminology and axioms.
I agree with Gomez-Perez et al. (2004, p 119) appreciating formal methodology as it "takes advantage of the robustness of classic logic and can be used as a guide to transform informal scenarios in computable models."
For that reason I also applied that technique in my Action Research Studies and a description can be found in Chapter 6.1.1 and 6.2.1, respectively.

Relating an enterprise ontology to an Enterprise Architecture Framework (EAF) has been suggested by Kang et al. (2010), Hinkelmann et al. (2010) and Thönssen (2010) to increase quality for example with respect to completeness. To the best of my knowledge despite the before mentioned work the particular aspect of relating an enterprise ontology to one or more Enterprise Architecture Framework(s) has not been researched yet. Some cognate subjects have been addressed by Vernadat (2010) and Chen & Pooley (2009b). Vernadat (2010) uses the European Interoperability Framework (EIF) as a foundational baseline to discuss technical, semantic and organizational aspects of enterprise interoperability and networking. Chen & Pooley (2009b) have a look at Enterprise Architecture Frameworks from a requirement engineering perspective. They suggest an extension of the Zachman framework using an ontology.

3.5 *Concluding the State of the Art Analysis*

In the previous sections literature on research with strong links to my work has been investigated. Figure 23 depicts the completed metadata generation pyramid I used to structure the chapter.

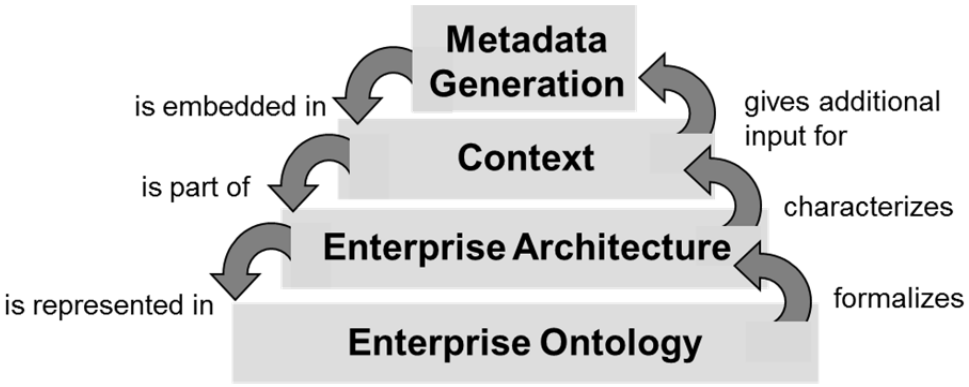


Figure 23: Metadata Generation Pyramid II

I will start my conclusion with work on enterprise ontologies. As shown several enterprise ontologies have been developed, some of them well-known and often referred to. However, each approach I analysed has some drawbacks I want to avoid in the mintApproach:

- inconsistency of granularity (TOVE)
- no 'development community' (TOVE, TheEO, CEO) as no recent further developments of the content can be recognized nor is the representation in a language that is of recent interest (unlike for example UML or RDF)
- not formalized (CbEO, CEO)
- not focussed on enterprise modelling (REA).

Analysis of research on representation languages for ontologies has shown that there is no 'silver bullet for formalization'. Even if a computational level of formalization is given, from several modelling language – or dialects – can be chosen. To my knowledge no research has been done yet on a comprehensive criteria catalogue to determine the appropriate ontology language from a business point of view, neither has an evaluation from this perspective been done.

Various methods for logical reasoning on ontologies and conjunctive on ontologies and relational databases have been investigated. I will follow the approach of Wang et al. (2004) using ontology reasoning to infer general context information for metadata generation (q.v. 5.1.4) and rules to reason on application specific metadata (q.v. 6.1.3 and 6.2.4). There are manifold approaches for database to ontology mapping but also no 'silver bullet'. As database-to-ontology mapping is not the focus of my research I will draw on methods I analysed within literature research and which are investigated in a current research project (APPRIS[130]) I am involved in, namely VisAVis, Virtuoso and D2RQ.

Ontology Engineering is a broad research field of its own and many different aspects have been investigated but no model has been commonly agreed on. For automatic metadata generation based on context I consider research of Feldkamp et al. (2010), who regard

[130] Advanced Procurement Performance and Risk Indicator System (APPRIS) is national funded research project about an integrated early warning system for supply risk & opportunities using semantic technologies. KTI-Nr. 12102.1 PFES-ES

ontology engineering part of an IT project, appropriate for the overall development procedure. To determine the content of seEAD I will draw upon research on (re)using a core ontology, amongst others by Fox & Gruninger (1998).
Asking competency questions – as suggested by Fox et al. (1996) and others – is an approach valuable for my Action Research Studies in order to validate and if necessary to enhance seEAD for enterprise and applications specific needs.

Regarding enterprise ontology as a formal representation of enterprise architecture description has become a research topic in the past two years. The few approaches considering Enterprise Architecture Frameworks (Kang et al. 2010, Hinkelmann et al. 2010, Thönssen & Wolff 2010) for their work refer to Zachmann's Framework, though the Zachmann Framework does not provide languages to describe the content of an enterprise architecture. Thus, the Zachmann Framework is very well suited to ensure quality of an enterprise ontology with respect to completeness by considering the perspectives stakeholders may take and the various aspects (who, what, etc.) that should be covered. To describe the content of an enterprise architecture the ArchiMate standard is more suitable as it specifies enterprise entities and their relations, and provides a language for representing it. Therefore I draw on the ArchiMate enterprise architecture framework for enterprise ontology development, particularly for the development of core concepts and relations and use Zachmann's Framework for quality control. To my knowledge this has not been done before. Refer to Suárez-Figueroa et al. (2011) for the latest research on methodologies, languages, and tools for building ontologies.

Prevalent in research on enterprise architecture descriptions is the question of how to reduce complexity (amongst others by Dietz 2006) but little research has been done on how to address the complexity. One approach has been provided by Hinkelmann et al. (2010) relating enterprise entities, *described* in an enterprise architecture to entities *used* in business applications, e.g. an employee's record that is stored in an ERP system, the document, that is stored on file server or the customer's preferences which are stored in a Client Relationship Management system. In Hinkelmann et al. (2010) and Thönssen & Wolff (2010) we call this total of entities managed in an enterprise 'an enterprise repository'. For my approach I take on this notion.

There are plenty of opinions on what context is and how it can be structured and modelled. Following Dey & Abowd (1999) who regard context as any information,which can be used to characterize the situation of an entity, I consider context as a view on enterprise architecture. This can be called a generic context model that can be used by various applications as called for by Linnhoff-Popien & Strang (2004). Although research has been done on using context mainly for metadata generation, to my knowledge no comprehensive approach has been introduced so far to exploit the *organisational* context of documents in order to address *all* document formats an enterprise has to deal with.
However, analysis of literature in this field has provided proof that my approach to automatic metadata generation based on context information is a widely accepted method for web-resources. Major difference to my work is that these approaches do not use context to characterize the *situation* of an entity (the enterprise object) but the *meaning* of a *term*.
Other approaches using context for metadata generation are system dependend, i.e. the document management or creation system provides the context. This is for example the case in the work of Ochoa et al. (2005) on metadata creation for learning objects.
There are some approaches based on the same notion of context as I have most of them deal with multimedia media document primarily considering the *physical* context (temporal and

spatial information about an object) but not the *business* context, i.e. concern, reason and purpose why for example an image has been taken.

To my knowledge only the approach of (Brügmann, 2011) considers *organisational* context for automatic metadata generation for enterprise documents of all formats (text, image, audio- and video-files). Although the context of a document is considered as source for semantic metadata, it misses a sound foundation und thus remains arbitrary. Furthermore, several context models are used in parallel which leads to (unresolved) conflicts and does not provide a comprehensive view as in a consistent enterprise architecture description.

Thus, although representing context information formally in an ontology is considered very valuable for enhancing documents description, the "design of appropriate ontologies and operational conceptualizations for context elements is [yet] a major area for new research" (Winograd 2001, p 407). This is particularly true for a context model based on a semantically enriched enterprise architecture description.

4 Requirements Engineering

Chapter 4 of my thesis provides the practical requirements for my work, complementing the review on scientific literature presented in the previous chapter, as illustrated in Figure 24.

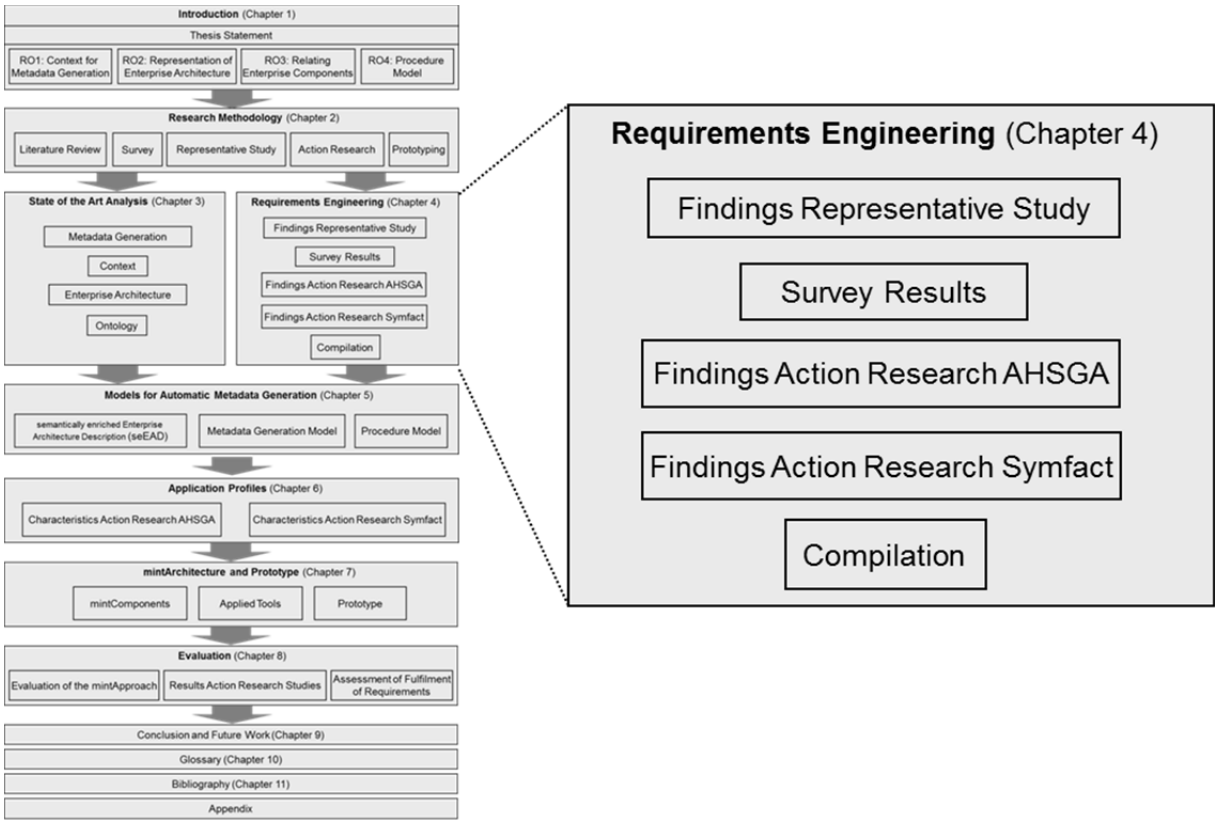


Figure 24: Position of Chapter 4 in the Overall Structure of the Thesis

Each of the chosen research methods, introduced in Chapter 2, provides requirements for my research. The chapter is organized as follows. First, conclusions of the representative study in European enterprises are presented. After that results of the survey on document handling in enterprises are provided. Then the first loop of the two Action Research studies is detailed. The chapter concludes with a summary of all requirements for automatic metadata generation and their criteria for measurement.

4.1 *Outcome of the Representative Study*

After literature review, results of a comprehensive representative study conducted within the MATURE project, are analysed with respect to the use of metadata creation tools in enterprises. The representative study about knowledge maturing in enterprises has been conducted in year two of the MATURE project. The study was conducted between 1.4.2009 and 1.4.2010 by all project partners. FHNW conducted 10 in Switzerland, I personally did 3 out of that 10.

In total, 139 interviews were conducted. A total of 128 interviews met the defined criteria. Of the 128 interviews in the sample, 43 were with representatives of medium-sized organisations and 85 were with interviewees representing large organisations. 43 organisations had their main area of business classified as industry, whereas 77 organisations were classified as service.

The size of an organisation was measured according to recommendations for innovation surveys. (OECD and EUROSTAT, 2005):

- Small 10 to 49
- Medium 50 to 49
- Large 250 and more.

Due to the assumption that medium and large companies have more systematic activities concerning knowledge maturing, potentially with designated roles that can provide a reflected perception of the importance, support and success of activities concerning knowledge maturing in their organisations, small sized organisations were not part of the study.

Out of the 128 interviews, two had to be omitted because the interviewees terminated the interviews during interviewing. Thus, 126 cases remained in the sample. These had a maximum amount of 7.6% of missing data per case concerning the closed questions that were quantitatively analysed. As less than 10% of missing data per case can generally be ignored, all 126 cases were part of the quantitative evaluation (Barnes et al., 2010).

The study's objective was to investigate the development of knowledge within and across companies and to develop supporting tools. As often digital resources are part of a knowledge maturing activity, the MATURE representative study provides insight in the use of documents in enterprises. Please refer to Chapter 12.1 in the appendix for more details.

With respect to my research questions (RQ1 and RQ4), the most important findings are the types of document creation software used in enterprises and objects determine the context of document creation and use. Table 6 provides an overview of software identified in the study. To achieve comparability coded types of software tools have been introduced. Three types of software were used by far the most on a general basis: intranet-based services (intranet.generic), mail programs (PIM.mail) and office software (office.generic). These types are highly ranked (within the top 5) in the knowledge maturing process.

Although it was generally aimed at being as specific as possible, greater generalisations had to be made for two specific codes: all answers related to portal solutions were coded as “intranet.generic” and all answers related to office tools were coded as “office.generic”. This leads to codes that may be a subset of other codes: If the answer “MS Office” was provided, this would result in the code “office.generic”. If the answer “MS Word” was provided, this would result in the code “office.word_processing”. As MS Word is part of the MS Office suite, the code “office.generic” would also be true. This approach leads to a constellation of codes, where more specified codes could also be counted towards “*.generic” codes. Please find a detailed description of the codes in the appendix 12.2.1.

The first column of Table 6 lists the codes, ordered by overall number. Columns I to V relate to knowledge maturing phases[131]. The figures outside the brackets represent the number of code occurrences within the phase, whereas the figures in brackets represent the rank of the code within the phase.

[131] The knowledge maturing process is structured into six phases: Ia. Expressing ideas (investigation), Ib. Appropriating ideas (individuation), II. Distributing in communities (community interaction), III. Formalising (in-form-ation), IV. Ad-hoc training (instruction) and V. Standardising (institutionalisation). MATURE project web-site. URL: http://mature-ip.eu/knowledge-maturing (retrieved: 3.4.2011)
Please refer to (Maier & Schmidt, 2007) for more information on the characterization of knowledge maturing or access the project’s web-site for more publications (URL: http://mature-ip.eu/publications) (retrieved: 12.12.12)

Code	# total	Ia expressing ideas	Ib appropriating ideas	II distribution in communities	III formalisa-tion	IV ad hoc trai-ning	V stand-ardisation	Unrelated to phases
intranet.generic	198	28 (2)	25 (3)	37 (2)	27 (3)	28 (1)	45 (1)	3 (1)
PIM.mail	177	28 (2)	29 (1)	48 (1)	19 (4)	14 (5)	19 (4)	0 (-)
office.generic	173	37 (1)	29 (1)	15 (4)	37 (1)	28 (1)	24 (2)	0 (-)
office.word_processing	107	10 (7)	10 (5)	17 (3)	28 (2)	13 (6)	24 (2)	0 (-)
internet.generic	83	18 (4)	23 (4)	3 (19)	4 (17)	2 (22)	5 (17)	1 (4)
office.presentation	82	9 (8)	8 (8)	13 (6)	16 (5)	22 (3)	13 (6)	0 (-)
intranet.wcms.wiki	69	11 (5)	10 (5)	13 (6)	10 (10)	8 (9)	8 (11)	1 (4)
Filebrowser	67	8 (10)	9 (7)	10 (9)	13 (6)	9 (8)	15 (5)	1 (4)
collaboration_tool. instantmessenger	63	5 (12)	3 (16)	9 (12)	1 (28)	2 (22)	2 (23)	1 (4)
office.spreadsheet	57	11 (5)	9 (7)	10 (9)	11 (8)	8 (9)	8 (11)	0 (-)
PIM.generic	57	8 (10)	7 (9)	13 (6)	5 (13)	6 (12)	6 (15)	0 (-)
project_management_tool. generic	53	6 (11)	5 (10)	1 (29)	13 (6)	10 (7)	13 (6)	0 (-)
intranet.social_software	46	9 (8)	4 (12)	15 (4)	4 (17)	4 (14)	6 (15)	0 (-)
DMS.generic	41	4 (14)	3 (16)	3 (19)	11 (8)	6 (12)	10 (10)	3 (1)
custom.generic	39	4 (14)	5 (10)	5 (15)	5 (13)	7 (11)	11 (9)	2 (3)
elearning_tool	32	2 (20)	2 (22)	3 (19)	2 (21)	15 (4)	7 (13)	1 (4)
collaboration_tool.confere ncing. desktop	30	2 (20)	2 (22)	10 (9)	1 (28)	3 (18)	1 (31)	0 (-)
ERPgeneric	27	3 (18)	3 (16)	5 (15)	5 (13)	4 (14)	7 (13)	0 (-)
intranet.wcms	27	3 (18)	4 (12)	6 (14)	5 (13)	2 (22)	3 (21)	0 (-)
desktoppublishing.pdf	25	0 (-)	1 (29)	2 (23)	4 (17)	4 (14)	13 (6)	0 (-)
modeling_tool. design_and_engineering	23	2 (20)	4 (12)	1 (29)	8 (11)	3 (18)	5 (17)	0 (-)
modeling_tool.enterprise	22	2 (20)	2 (22)	1 (29)	6 (12)	3 (18)	5 (17)	1 (4)
modeling_tool.mind_maps	21	4 (14)	1 (29)	2 (23)	0 (-)	0 (-)	0 (-)	0 (-)
modeling_tool.generic	18	1 (28)	1 (29)	0 (-)	4 (17)	0 (-)	5 (17)	1 (4)
informally.not_existent	18	0 (-)	0 (-)	0 (-)	0 (-)	0 (-)	0 (-)	0 (-)
internet.social_software	16	1 (28)	2 (22)	2 (23)	0 (-)	0 (-)	1 (31)	0 (-)
collaboration_tool.confere ncing. video	14	1 (28)	0 (-)	8 (13)	2 (21)	3 (18)	0 (-)	0 (-)
ERPfinance	13	4 (14)	4 (12)	1 (29)	1 (28)	0 (-)	2 (23)	1 (4)
suggestion_system.generic	13	5 (12)	3 (16)	2 (23)	2 (21)	0 (-)	0 (-)	1 (4)

Table 6: Types of Software (Barnes et al. 2010, p 50)

Findings of the representative study:

1. Firstly the table clearly gives evidence for the wide use of "office documents" in enterprises (indicated by the shadowed lines in Table 6) and secondly, it shows moderate use of document management systems ("DMS.generic", indicated by the bolder boundaries of the line in Table 6). With respect to my research question (RQ4: What metadata can be harvested and extracted to be used as source for metadata generation?) it means that document properties, created by MS Office tools and the underlying operating system, can build the source for automatic metadata generation.
2. With respect to my second research question (RQ2: What determines documents' context?) the representative study does not provide such a clear answer but shows that a "filebrowser" (like the Windows Explorer) is used often to navigate through file systems on own desktop or on network share and is ranked highest in the "standardization phase" of knowledge maturing. This indicates that structure and naming of directories is somehow used to standardize the management of documents and hence provide context information about the documents. Answers to a correlating question of the Survey on document handling corroborates this hypothesis (please refer to Chapter 4.2.1.7).

The highly ranked internet tools ("internet.generic") and personal email systems ("PIM.mail") are not considered further as they are out of scope of my thesis and a lot of research has already been done in these fields.
Although tools for modeling enterprise architectures (modeling.tool.enterprise) are ranked low for knowledge maturing, this does not imply that they are of no value for metadata generation. But today, as shown by Hinkelmann et al. (2010) they cannot be operationalized and thus their use is limited and not regarded as eligible to support knowledge maturing.

Please refer to Chapter 4.4 for the requirements derived from the findings.

4.2 *Results of the Survey on Document Handling in Enterprises*

As the focus of the representative study has been on knowledge maturing and excluded small- and micro-sized enterprises a more comprehensive investigation into document handling in enterprises was needed to back my hypotheses. Therefore, I created and conducted a survey that was guided by the following four research questions:

- What metadata elements are important? (RQ1)
- What determines documents' context? (RQ2)
- What metadata can be harvested and extracted to be used as source for metadata generation? (RQ4)
- How to incorporate and use enterprise specific knowledge, for example glossaries or filing plans? (RQ16).

The survey was carried out over a period of eight months, from April until December 2010. A total of 30 face-to-face interviews were conducted with business managers of micro-sized and small-sized enterprises (MiE and SE) and non-profit organizations (NPOs), with heads of organisational units of medium-sized and large-sized enterprises (ME and LE), and representatives of Swiss Public administrations on all federal levels, state (PA-CH), canton (PA-C) and municipality (PA-M).

The size of an organisation was measured according to the EUROSTAT Commission Recommendation (96/280/EC) of 3 April 1996 concerning the definition of small and medium-sized enterprises (SMEs):

- Micro-enterprises: less than 10 employees
- Small enterprises: less than 50 employees

- Medium-sized enterprises: less than 250 employees (Schmiemann, 2006).

Enterprises with 250 and more employees are considered large as in the MATURE representative study.
Contrary to the representative study, organisations designated as "micro", having less than 10 employees and "small", having 10 to 49 employees were part of the study. This was due to the assumption that micro-sized and small enterprises can't or won't afford enterprise search software or document management software but make do with storing documents on a file server using the explorer for management.

None of the interviewees is an expert in information management or is assigned to a specific task of information management in her/his organisation. Participants were recruited via personal contacts (e.g. business partners either of University of Applied Sciences, Univeristy of Camerino or of Action Research partners, participants in courses of Applied Sciences) as interest was not in a particular branch or industry sector and no specific competencies were expected. Participants were selected based on the following selection criteria:

- for universality the survey should not be restricted to one country, and each size of enterprise (from micro- to large-sized), as well as each federation level of Swiss Public administrations should be represented.
- for verification of the NGO Action Research Partner's requirements, at least 10% of the participants should work in that segment.
- for verification that document handling in public and private sector is alike, at least 25% of the participants should work in Public administrations.

As interviews are conducted in face-to-face meetings a soft factor for selection was the willingness to spend at least one hour for an interview plus additional time for subsequent questioning in the evaluation phase, if appropriate.

Figure 25 shows the distribution of the participants by country (25 Switzerland, 3 Germany, 1 Italy and 1 Austria). Figure 26 gives an overview on the distribution of participants by organisations (7 MiE, 1 SE, 1 ME, 3 LE, 5 NPO, 8 Swiss Public administrations thereof 3 on state level, 4 on canton level and 1 on municipality level).

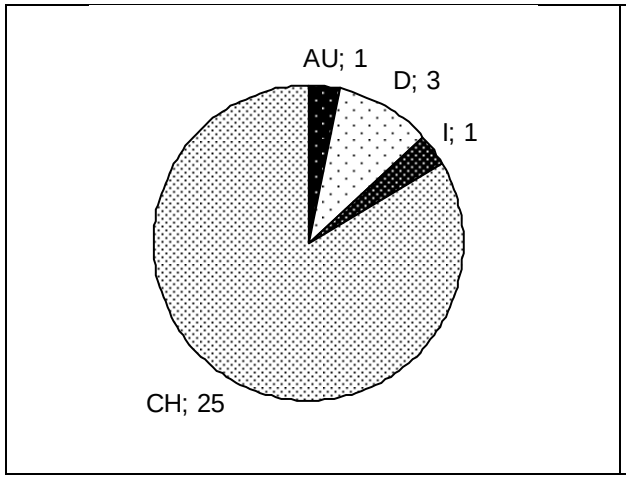


Figure 25: Distribution of Interviewees by Country

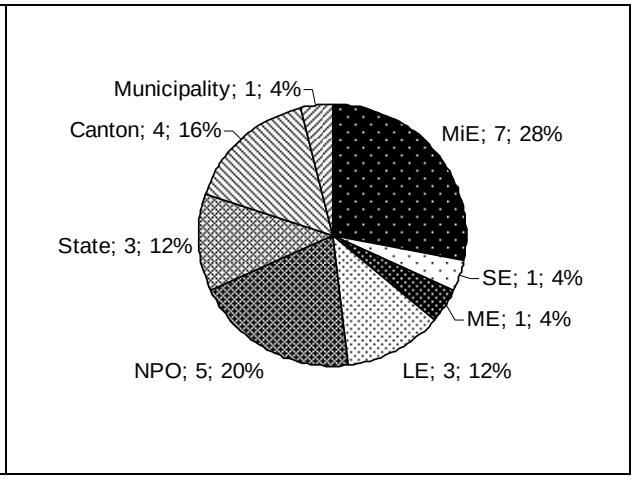


Figure 26: Distribution of Interviewees by Organisation

All interviews were based on a standardized questionnaire (see Appendix 12.3), comprising a total of 20 questions, thereof 18 closed questions and 2 open questions. All questions were available in German and English.

The survey was developed

- to evaluate document handling in daily routine by non-information specialists
- to verify the analysis of the representative study with respect to distribution and usage of document creation tools and standardization approaches
- to determine context of document handling with respect to business process management and governance instruments
- to find out search strategies and
- to identify the most important metadata elements (based on the Dublin Core Metadata Element Set)[132].

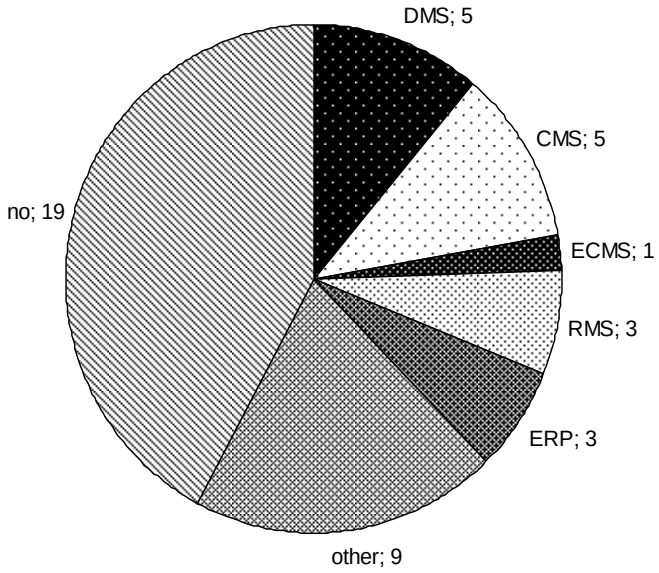
4.2.1 In-depth Report

In the following details on the results of the survey on document handling in enterprises are provided. The results have been published in Thönssen (2011).

4.2.1.1 Tools for Document Handling

Survey results show that more than 60% of the organizations do not use any tool to handle their documents (Figure 27). As expected, many are micro-sized enterprises (7 out of these 19) and only one is a large-sized enterprise. The tools used to manage documents show large variance: ranging from Document Management Systems (DMS), Content Management Systems (CMS), Records Management Systems (RMS) to Enterprise Resource Planning Systems (ERP). 9 out of the 30 representatives answered that the business software they use provides document management functions; 4 out of that 9 organisations are micro-sized enterprises. Organisations, like NPO, for which business software is barely available, lack of document management software: only one of the NPO is about to implement a business application with document management functionality but has none yet.

[132] Dublin Core Metadata Element Set, Version 1.1. URL: http://dublincore.org/documents/dces/ (retrieved: 14.12.2010)

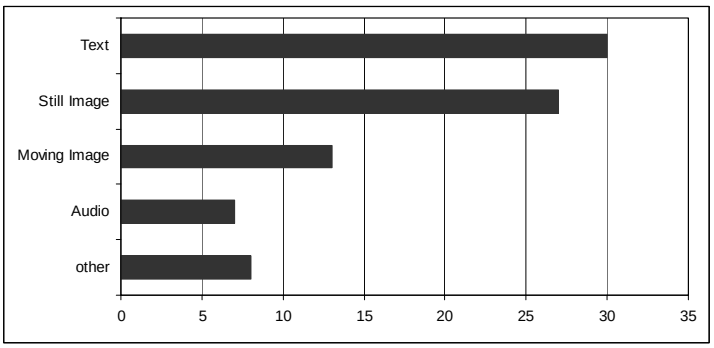


Labeling of the table:
tool type; absolute number, percentage of use.
Participants checked all that applied.

Figure 27: Usage of Tools for Document Handling (Question 1)

4.2.1.2 Document Formats

In all organisations textual documents are used (Figure 28) in daily routine. In 27 organisations "still images" (photos, diagrams etc.) are business relevant and 13 are dealing with "moving images", i.e. video, podcasts, DVD etc. (DCMI Usage Board, 2012). 7 organisations use files of audio format (audio) and 8 organisations reported the use of "other" document formats. Those documents show a great variety like CAD graphics, maps or specific formats generated by business software, XML, RDF, OWL, books, patents, construction plans, law (e.g. court decisions) or software source files.

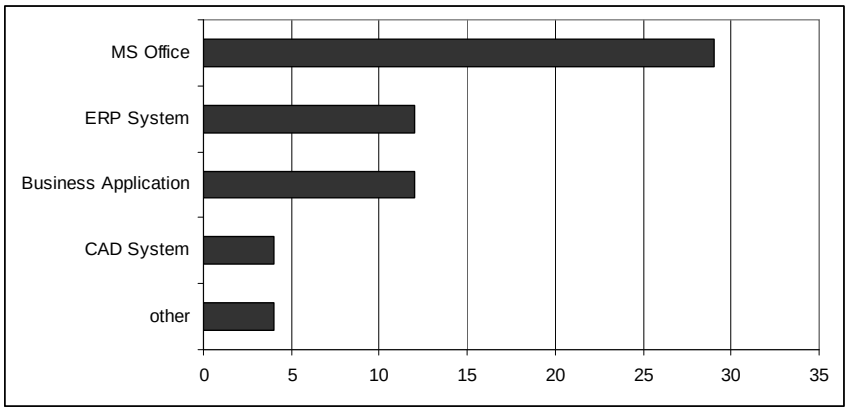


Participants checked all that applied

Figure 28:Document Formats (Question 2)

4.2.1.3 Document Creation Software

Nearly all participants (29) use MS-Office tools for document creation, and then most often MS-Word. One participant uses a business application only (Figure 29). Other systems used for document creation are production systems (e.g. creating reports), scientific systems (e.g. laboratory systems producing test logs), image processing systems or software compiler.



Participants checked all that applied

Figure 29: Document Creation Software (Question 3)

4.2.1.4 Use of Templates

Questions 4 and 5 are about the use of templates. 28 participants answered that they use templates in their daily routine of document creation; only two answered in the negative (1 MiE and 1 PA-C).
Up to 10 different templates are used by 12 organisations, between 11 and 30 templates by 8 organisations and more than 30 by 7 organisations (Figure 30).

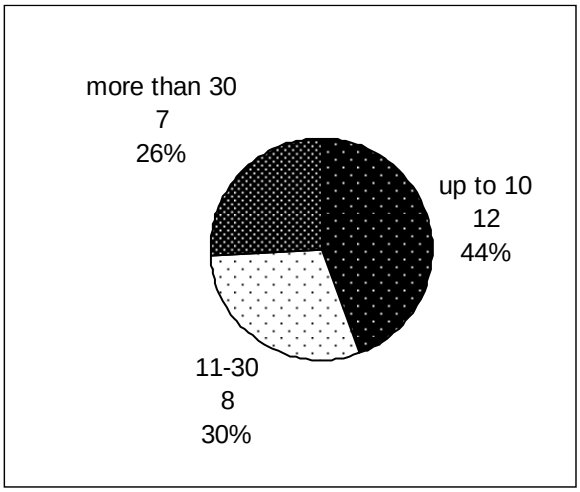


Figure 30: Use of Templates (Question 5)

4.2.1.5 Metadata Attributes

Document creation software (e.g. MS-Word) automatically adds attributes to a document, like creation date or file size. Question 6 is about knowledge on additional attributes users can add to describe the documents (e.g. for search). Equal number of answers was for “yes, I know” and “no, I don’t' know” (11 nominations each). One participant answered that the functionality did not work as expected and 6 reported that they have heard about it but never tried (Figure 31). Answers to question 6 are cross checked with answers to question 12, about what attributes/terms are used to search for documents. None of the participants answering “yes, I know” to question 6 reported on any user specific attribute she uses for search. Thus, the positive answer to question 5 should be regarded as similar to the, “heard about but never tried” answer, which means that almost nobody uses that functionality to manually add metadata to a document.

For the participant who uses a business application only, the question was not applicable.

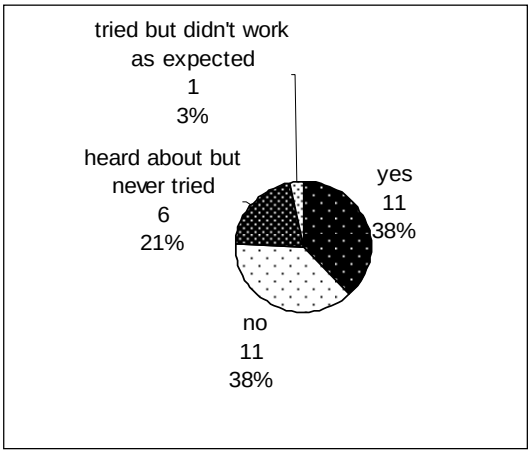


Figure 31: Adding User Specific Document Properties (Question 6)

4.2.1.6 Document Storage

Two thirds of the participants answered that they store business documents on a server accessible for all employees (Figure 32). Almost half of the 13 participants reported that they store their documents on a personal computer. Most of these 6 participants are solely responsible for a business domain and thus nobody else needs access to the documents. For security reasons documents are stored additionally on external hard disks (other storage). The question was similarly answered by interviewees exclusively working in a business domain, storing documents on a server but in a personal directory.
Those participants using a tool for document management store most but not all documents within the application. In one public administration for example only final documents are stored in a system; work in progress is stored on server accessible for the organisational unit.

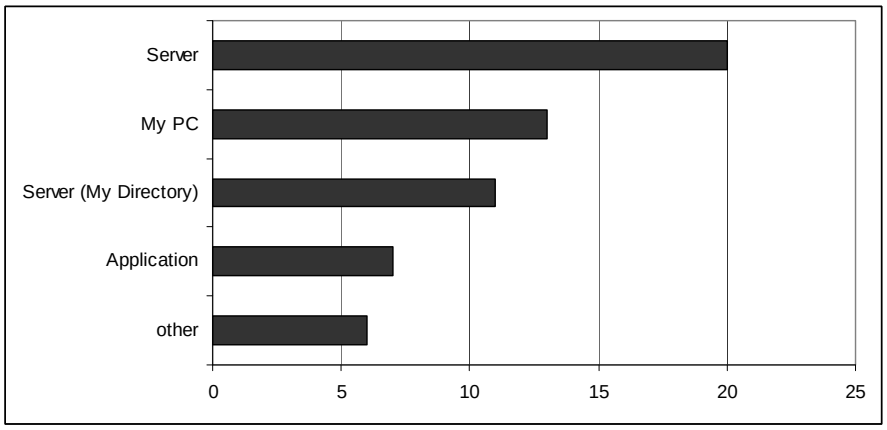


Figure 32: Document Storage (Question 7)

4.2.1.7 Directory Structure

About 90% of the organisations at least partially define a directory structure for document storage (Figure 33). 2 NPOs tried but failed to get it accepted and 1 MiE answered "no" without an explanation. The answer "yes" mainly was understood as defining a directory structure for the three upper levels; 'partially' was chosen if only the top level was defined or

if a given structure could be enhanced (but not changed) by individuals. “No” was answered if participants were completely free in defining their directory structure.

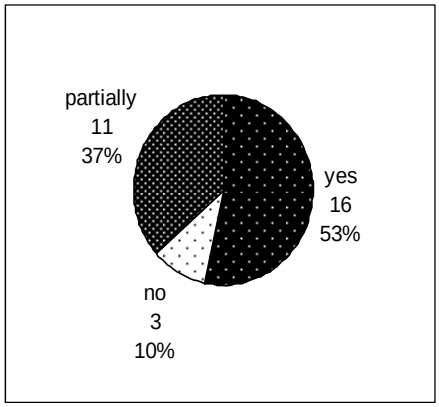
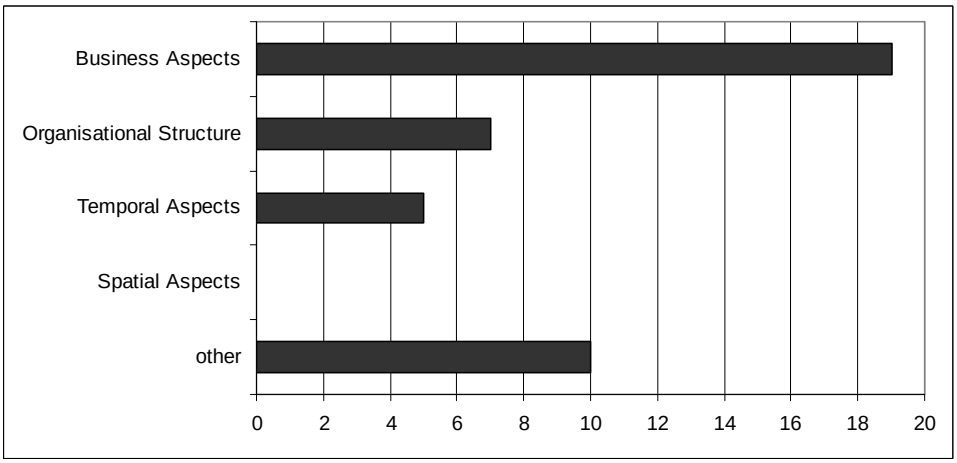


Figure 33: Standardization of Directory Structure (Question 8)

In 19 out of 30 organisations criteria for structuring a file system are business aspects like project, customer or product (Figure 34). The relatively high number of participants (10) reporting “other” criteria for a directory structure correlates with answers to question 7: users storing business data on personal computers or personal partitions have their own - not totally rational criteria. Criteria indicated for “other” were “financial aspects”, “lectures” or “fields of work”.
None of the organisations considered spatial aspects for structuring file storage.



Participants checked all that applied

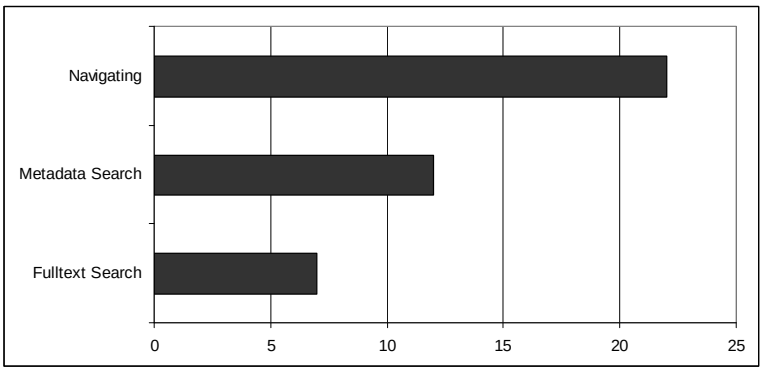
Figure 34: Criteria for Directory Structure (Question 9)

To the related question 10, if the directory structure correlates with a filing structure (filing plan) 10 participants answered “yes” whereas 16 answered “no”. 4 interviewees didn't know the answer.

4.2.1.8 Searching

About ¾ of all participants reported “navigating” as the preferred method for search (Figure 35). Reasons mentioned for this are simplicity ("I know what to look for") and format independence ("I can't search for a file name of an image as they often are just numbers"). 12 participants use functionality which the operating system provides and search for document

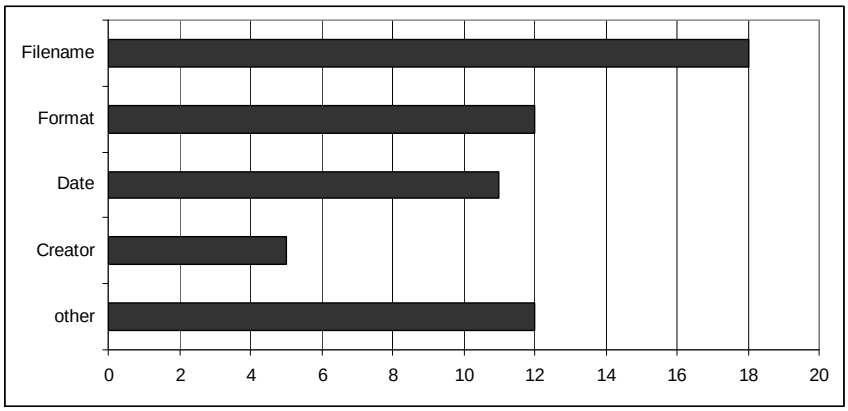
properties like title or creation date (metadata search). All interviewees performing full-text search use functionality as an application or one the operating system provides. None uses a desktop search tool.



Participants checked all that applied

Figure 35: Search Method (Question 11)

More than half of the participants (18, 60%) search for the filename (Figure 36). Those interviewees searching for "other" document properties (12 times) use metadata provided by a tool they use for document handling, e.g. a business application with the respective functionality or the full-text search function provided by the operating system. Equally often chosen was the document property "format" (12 times), followed by "date" (11 times). Only 5 interviewees indicated using "creator" for search. Reasons given for that are bad quality (the creator often isn't the author of the document but author of a document that was taken as "template"), and that searching is mainly performed in electronic storage (i.e. in directories) dedicated exclusively to themselves (cf. answers to question 7).



Participants checked all that applied

Figure 36: Document Properties Used for Search (Question 12)

4.2.1.9 Use of Dublin Core Metadata Elements

Participants were asked to indicate which of the fifteen simple Dublin Core Metadata Elements they would like to use for search. In addition they could choose "other" for specific metadata they could think of (Figure 37). Dublin Core has been chosen because of its use in non-library environments, as shown in the AMEGA report (Greenberg et al., 2005): About a quarter of the participants indicated use of simple or qualified Dublin Core (participants checked all that applied). However, none of the participants of the survey on document handling was familiar with the standard.

Nearly 80% of the participants chose “subject” (23 times) and “description” (23 times), metadata they cannot search for at present. Nearly equally often “relation” was selected as metadata for search (22 times). All interviewees that chose the “relation” metadata stressed its importance and would like to have a graphical representation of the related documents in a way as provided for example by the quintura[133] search engine for images.

“Title” (18) and “format” (12 times) were selected as often as “filename” and “format” in question 12. Participants were little interested in administrative metadata like “rights” (5 times), “language” (4 times), “identifier” (3 times) and “publisher” (3 times).

“Other” metadata (selected 8 times) shows a huge variety: one interviewee wants to search for context (like field of work, business area), one wants to search for key words (representing the content) and industry-sector specific terms, one wants to search for a certain version of a document (or the latest respectively), another one wants to search for 'mood' in audio and images.

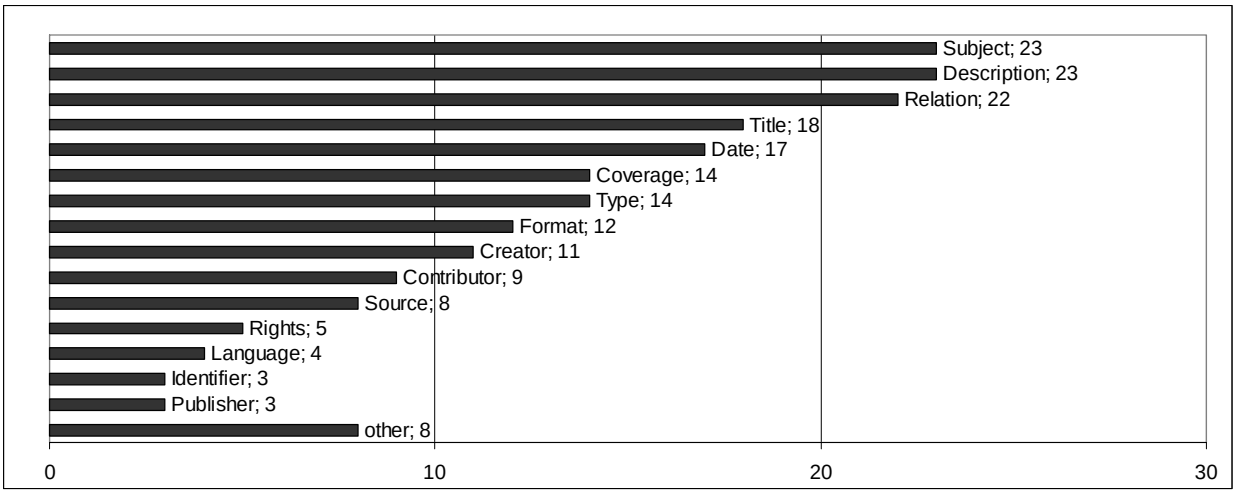


Figure 37: Dublin Core Element Set (Question 13)

4.2.1.10 Naming Conventions

Question 14 is about naming conventions for file names. 14 participants answered “yes”, in their organisations they have standards for file names and 15 participants answered “no”, they don’t have one. 1 participant couldn't answer the question as file names were created by the business application he uses.

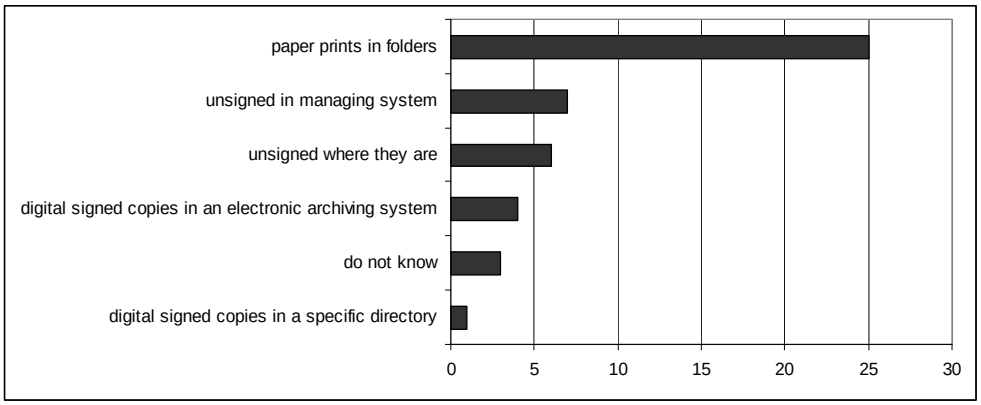
4.2.1.11 Legally Binding Documents

In question 15 interviewees were asked if they know which of the electronic documents they deal with are legally binding (e.g. a project offer you sent out via mail). 14 participants answered “yes”, 16 answered “no”. Reason for the rather high number of negative answers is that many participants were not aware that as well as paper documents electronic documents can be legally binding (e.g. emails).

25 Participants store legally binding documents as paper copies in folders, often kept in centrally organized file cabinets (Figure 38). 20% of the participants store legally binding

[133] quintura. Source: http://www.quintura.com/ (retrieved: 13.1.2011)

documents “unsigned where they are”. However, all participants but one, additionally store these documents either as paper form (5 times) or in a management system (1 time in an ERP system). 3 participants do not know where legally binding documents are stored because another organisational unit is responsible for that issue. Only one interviewee stated that digital signed copies in his ME are stored in a specific directory on a server.

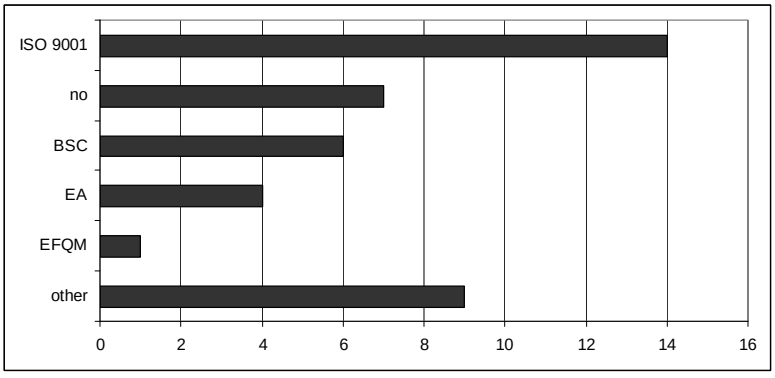


Participants checked all that applied

Figure 38: Storage of Legally Binding Documents (question 16)

4.2.1.12 Governance Instruments

Question 17 the use of governance instruments is based on considerations of (Lankhorst, 2009) about enterprise architecture at work, asking for well-known standards in enterprise management. Almost half of the participants (14) indicated the ISO 9001standard of the International Standardization Organisation (ISO) for quality management (Figure 39), although they are not certified. Balanced Score Card (BSC) was mentioned by 6 participants, the use of an Enterprise Architecture (EA) by 4 participants and the Excellence Model of the European Foundation for Quality Management (EFQM) was indicated by one interviewee. “No” use of a governance instrument was answered by seven participants (3 MiE, 1 ME, 1 LE, 1 PA-C, 1 PA-M). Nine interviewees indicated the use of “other” governance instruments, like descriptions of operating procedures, operating manuals, internal guidelines and principles, individually developed management handbooks.



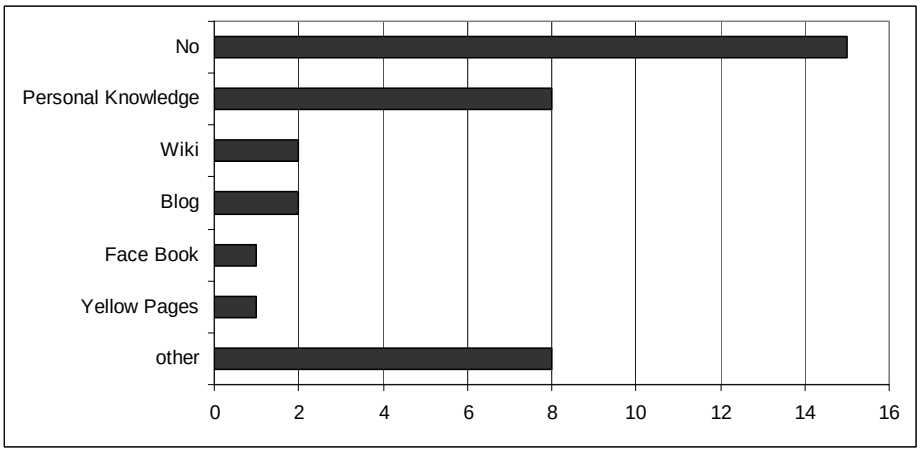
Participants checked all that applied

Figure 39: Usage of Governance Instruments (Question 17)

4.2.1.13 Skills and Experience Management

15 Participants (50%) do not use any tool for skills or experience management in their organisation (Figure 40). 8 participants rely on personal knowledge of their colleagues (2 MiE, 4 NPO, 1 PA-CH, 1 PA-C). An enterprise Wiki or an enterprise blog is used in 2 organisations, an enterprise "face book like application" is used in 1 organisation as are "yellow pages". The use of "other" tools for skills or experience management was stated by 8 interviewees.

Some participants developed in-house applications for skills management. For example one umbrella organisation of an NGO started a web site to provide knowledge (e.g. living facilities for seniors) accessible for all (not only the related NGOs) to increase transparency. One PA-C facilitates a web-site with "wiki and blog" functions accessible for members of the Public administration. One MiE stores 'lessons learned reports' (MS Word files). One MiE uses the FAQ function of MS Sharepoint.



Participants checked all that applied

Figure 40: Skills and Experience Management (Question 18)

4.2.1.14 Advantages and Disadvantages of Document Handling

At the end of the interview two open questions asked participants about advantages (question 19) and disadvantages (question 20) of document handling as it is in their organisation. No correlation between enterprise size or type and liking or not liking could be identified.

Participants reported they like:

+ navigating the file system as they are familiar with the structure
+ document handling as-is as it is simple and easy to use
+ the possibility to easily publish documents out of a system on the internet or intranet
+ navigating the file system as it is easy to understand and simple to use
+ if no additional effort for document handling is needed
+ if document management is deeply integrated in business or production software
+ if document management is customizable
+ that a Wiki provides convenient access to documents linked to an article
+ systems that automatically store documents in electronic archiving system
+ that a systems allow to link context knowledge (e.g. about a customer) to documents
+ the freedom to organize their documents as they want to
+ easy central storage of document that everybody can see everything at any time
+ quick access to electronic documents (rather than searching in paper cabinets).

Participants reported both, aspects they don't like and features they would like to have: :

→ qualified search for dates
→ versioning control (retrieve most recent per default)
→ change tracking (historiography)
→ enterprise wide document management instead of dealing with 'document islands'
→ one place (directory or system) where all documents belonging to the same project are stored
→ ,intelligent search for content (but no full text retrieval)
→ automatically generated metadata for content
→ centrally published documents accessible via the web
→ search for document components not only for complete documents
→ search for key words (of domain knowledge)
→ relations between documents
→ relations between in-house documents and documents published on the web (e.g. relate own decisions on a court case to court decisions already published on the web)
→ relations between documents and e-mails, or attachments and documents
→ relations between documents and business processes
→ a positive trade-off between standardisation and freedom of document management
→ the gap closed between information need and information retrieval, "what I find is what I want not what I get"
→ simple to use tool
→ graphical search and representation tool, e.g. in form of mind-maps
→ possibility to search for DC elements
→ traceability of document management (who changed what and when)
→ guarantee that all documents are found
→ data (e.g. of a CRM system) linked to documents
→ no redundantly stored documents.

4.2.2 Findings of the Survey on Document Handling

In the following findings of the survey on document handling relevant to answer my research questions – as defined in Chapter 2.1 – are provided.

1. Beside text, still images are by far the most used document formats in organisations. Thus automatic metadata generation for images is of importance.
2. MS Office is almost always used for document creation thus harvesting document properties of those documents is useful in all enterprise (answer to RQ4)[134]
3. Additional user-defined document properties cannot be expected for metadata harvesting (answer to RQ4).
4. Low-level governance instruments, like templates for document creation, definitions for directory structure and naming conventions are widely used in enterprises. Having those low-level governance instruments formally represented in the enterprise ontology they can build documents' context (answer to RQ2).
5. Although no user-defined document properties are added manually, having content-related metadata, like subject, description and relation would be highly appreciated (answer to RQ1). However, exactly for these metadata manual processes were considered more appropriate than automatic as they require greater intellectual discretion (Greenberg et al., 2005). Thus, in the AMEGA report a more holistic creation was suggested like considering context approach for automatic metadata.

[134] in general multi-media documents are not created in-house but imported or downloaded from the internet

6. Naming conventions are defined by many of the organisations and thus can be used for inferring metadata like creation date or creator. This is especially useful if this information is compared with the respective harvested document properties to discover faults (answer to RQ15).
7. Written information about skills and experiences cannot be expected and thus cannot be used for context definitions (anwer to RQ2).
8. As no additional effort for metadata creation is wanted, automatic metadata generation must be performed transparently in the background.
9. Document management functionality is used when deeply integrated in business applications. Therefore automatically generated metadata should have a format that can easily be imported into existing tools.

Please refer to Chapter Requirements for Automatic Metadata Generation in Enterprises for the requirements derived from the findings.

4.3 *Requirements of Action Research and First Models (Loop 1)*

Action Research studies are conducted within two enterprises: the non-profit organisation (NPO) AHSGA, a cantonal Swiss institute for the prevention of AIDS and for sexual health, located in St. Gallen, Switzerland. The NPO AHSGA (Aids-Hilfe St. Gallen und Appenzell) acts locally in the cantons St. Gallen and Appenzell Innerrhoden and Ausserrrhoden[135].

The other enterprise, Symfact AG, is a software development and consulting company, specializing in contract management software, located in Sugiez-Bern, Switzerland. Symfact is small-sized but acts globally. The company wants to expand its business operations offering not only consultancy for their contract management tool but automating metadata generation and an approach for active document lifecycle management.

Because of the very different business goals of the two enterprises, requirements for metadata generation differ significantly, too. The NPO AHSGA struggles with many different document formats (a wide range of video/DVD and images besides various text document formats) but has no specific requirements for metadata elements. Symfact mainly deals with text documents, respectively images of text documents (especially contracts) but wants to populate sophisticated and enterprise specific metadata elements and improve active lifecycle management for contracts, respectively of the reported obligations.

In this chapter enterprise specific requirements of the two Action Research partners are specified and answers to the research questions addressed within the first loop of the Action Research study are provided.

4.3.1 Action Research Study With AHSGA

AHSGA wants to increase employee productivity by decreasing time for searching documents. As manual metadata creation is labour intensive and error prone and information retrieval techniques are not applicable for many types of documents (e.g. images or audio files) a solution is desired that allows for generating metadata automatically for all kinds of documents used in the enterprise.

[135] Fachstelle für Aids und Sexualfragen. Homepage. URL: http://ahsga.ch/ (retrieved: 3.4.2011)

The AHSGA Action Research team involved in the study consists of five full-time employees: the business manager, three pedagogues in human sexual behaviour and the secretary[136]. Research is conducted in three loops (as introduced in Chapter 2.2.3). First loop of the study focuses on as-is analysis and was executed between July 2010 and March 2011. Within this loop three meetings took place: first on July 23rd, 2010, second on December 30th, 2010 and third on February, 28th, 2011.

4.3.1.1 Results of the First Loop of Action Research AHSGA

AHSGA has to maintain a large number of documents of all forms, i.e. text, image, audio and video and many formats (e.g. doc, pdf, jpg, png). All documents are stored on file server, accessable by all employees. The directory structure is partly preset but not actively managed.

In the following the results of the first iterative cycle of tha Action Research study are presented.

1. Inventory on document handling

Table 7 gives an overview on document handling as it is performed at AHSGA.

[136] Details on the team can be found at AHSGA homepage. URL: http://ahsga.ch/000001985b0d93482/index.html (retrieved: 14.12.2010)

Aspect	Occurrence
File Organisation	Documents are stored in a file system of the operating system (MS Windows XP) and organized according to business aspects. Paper registration matches with directory structure. Problems: - redundant storage of files - no versioning
Naming Conventions	For directory structure: First level: organisational unit (e.g. professional services) Second Level: business aspects (e.g. projects, customers, suppliers)[137] For file names: Every file name starts with an employee identification number (e.g. 1) followed by the year (two digit, e.g. 10). The actual name of the document is not restricted Example (complete path): F:\TEXT\AH_GS\OEFF_ARB\Jubiläum2010**110** ProtokolldererstenSitzung.doc
Document Types[138]	Text Image o still image o moving image Sound Physical Object (paper document)
Document Formats[139]	doc, pdf, ppt, xls, jpg, gif, png, bmp, mp3, mp4
Document Layout	All text documents must have the "AHSGA footer" with full path information, creator and creation date Example: F:\TEXT\AH_GS\OEFF_ARB\Jubiläum2010\110 ProtokolldererstenSitzung.doc Erstellt von Johannes Ernst Schläpfer, Erstelldatum 01.06.10

[137] Actually there are two levels above: the name of the disk drive (e.g. "F") and the directory indicating business issues, called "TEXT". As they do not change they are omitted for better reading.
[138] Classification according to the DCMI Type Vocabulary. URL: http://dublincore.org/documents/dcmi-type-vocabulary/ (retrieved: 20.11.2010)
[139] Following Dublin Core reference for term 'Format': URL: http://www.iana.org/assignments/media-types/ (retrieved: 20.11.2010)

Business Governance Instruments	Quality Management Manual Mission Statement *(Leitbild)* Service Level Agreements *(Leistungsvereinbarungen)* Activity Report *(Tätigkeitsbericht)* Organisational Structure *(Organisationskonzept)* Job Descriptions *(Stellenbeschreibungen)* Annual Activity Plan *(Jahresplan)* Auxiliary Material (e.g. forms, regulations, check lists) Naming Conventions Document Templates
Software / Tools	MS Office tools Image Editing tool (Coral Photo Paint) ERP-Database Information and Talks Recording System (ITRS), an in-house development of the Swiss institute for the prevention of AIDS in Bern; based on MS-Access
Access Rights	No restriction
Backup / Archiving	Daily by external supporter
File Maintenance	Performed manually if running out of storage space periodically, once a year if a certain parameter is met (e.g. a project ends, a contract is signed etc.)

Table 7: Overview on AHSGA's Document Handling as-is

A total 187 electronic documents of all of the above listed file formats were selected by the Action Research partner as data source for the Action Research study. The chosen documents were original documents and document templates used within the enterprise over the last five years, either created by AHSGA's employees or imported (copy) from external sources like the internet. The documents were stored in a total of 25 directories at the enterprise's file server.

2. Identification of enterprise-specific problems with document handling
   The most important problem AHSGA deals with is the huge number of images about sexual health, representing 'everyday life' (e.g. a kissing couple, a naked girl etc.) Whereas in production plants, for example in a factory manufacturing bicycles, images represent concrete products or parts of it (e.g. a frame or saddle) and therefore feature analysis for content identification might be possible, for AHSGA's images it is not. The represented topics are too general and can be considered from various viewpoints (e.g. from the perspective of sex education or the perspective of prevention). Even worse is the often menaningless document properties since in many cases the multi-media documents are downloaded from the internet or taken over from other AIDS agencies.

   Another problem AHSGA faces is the increasing production of moving images on the topic of sexual health. Many TV movies (documentations, feature films, talk shows) are recorded but there isn't enough time to register them and manually add metadata in AHSGA's Information DB.

The third problem AHSGA faces is managing paper-based documents. Newspaper and journal articles are collected and stored on paper base in suspension files, ordered by keywords. The list of keywords is available electronically and stored in a directory but not linked to the paper documents. As this problem will be solved in the short term by the national Swiss Aids Federation (they will provide press reviews online) it is not addressed any further in the Action Research work.

3. Collection and specification of requirements
   Business goal in general is to increase productive labour by decreasing time for searching, especially for images and improving support for (moving) image administration. The following requirements were specified:
   - metadata generation must be applied to documents of all formats (text, image, audio, video)
   - metadata generation must be performed automatically without user intervention
   - metadata generation must be performed even if only little information about the document is available (e.g. none or meaningless document properties of multi-media documents imported from the web)
   - documents must be stored in the enterprise's filesystem as hitherto
   - no additional system for document management is wanted but the existing Information- and Task Reporting system should be used
   - documents should be retrieved based on information recorded about tasks which are reported in AHSGA's Information- and Task Reporting system
   - related documents should be identified.

4. Comparison of real life situation with a priori statements on the problem
   The situation discovered at AHSGA mirrors the anticipated problems and survey results: an increasing number of non-textual documents has to be managed but is too laborious to be done manually. In case of AHSGA the problem is even worse as the images they deal with often do not provide any useful document property and are of general content (e.g. a kissing couple).

   AHSGA's business governance is based on the ISO 9001 standard although the NGO is not officially certified. Instead of using full-blown Enterprise Architecture (EA) AHSGA relies on low-level governance instruments. For document management they use instruments like document templates, naming conventions for file names and specifications for file organisation. In addition, business processes, production procedures, business areas and working fields are documented in a management handbook.

   As file search functionality, provided by the operating system, is barely used by AHSGA, providing a separate retrieval interface - even it is very simple and easy to use - seems inappropriate. Instead the already existing information and time recording system (further called AHSGA-ITRS) should be used.

   Whereas generating metadata automatically based on context information seems to be an appropriate approach and feasible for the AHSGA, it seems to be difficult to define general rules for reducing metadata candidates, e.g. by probability reasoning in order to get real metadata. That means inferring context for metadata generation may deliver too many metadata candidates which cannot be reduced without manual interaction.

5. Review of current research on the problem
   Metadata generation for images handled by AHSGA is particularly difficult as they often depict daily life scenes: a kissing couple, a naked girl, two aged men holding hands etc. Thus, low-level features as colour, shape or texture will be of very limited use for discovering images' content. The only way to narrow 'the semantic gap' (Enser & Sandom, 2003) is by inferring the image's context. This has been researched for example for personal knowledge management, amongst others, by (Mitschick, 2009) and (Carvalho, 2008) but not for professional use in enterprise or public administration respectively. In AHSGA, as in many organisations, structure and naming of a file system is defined by low-level governance instruments. Thus this information, if formally represented in the enterprise ontology, can be inferred for context-related metadata (Thönssen 2011).

6. Questions to be answered
   Results of the first loop of Action Research with AHSGA lead to the following questions:
   - How to represent AHSGA low-level governance instruments in semantically enriched Enterprise Architecture Description (seEAD)?
   - How to represent the document model in seEAD (according to the ArchiMate standard)?
   - How to match recorded information in AHSGA's ITRS, e.g. consultancy performed by an employee for a customer, with automatically generated metadata candidates?

7. Actions to be taken
   First loop of Action Research with AHSGA results in the following actions:
   - Enhance/adapt enterprise ontology to AHSGA's specifics
     - Represent AHSGA low-level governance instruments in seEAD
     - Represent AHSGA's document in seEAD
   - Run through example for an AHSGA document
   - Check on availability for 'Linked Data' to get information about movies (e.g. "Linked Movie Data Base"[140], by Oktie Hassanzadeh and Mariano P. Consens or the "'DBpedia Knowledge Base"[141]) or "Nanoo,TV"[142] by Werft22 for information on TV productions
   - Check on the use of AHSGA's ITRS for evaluating metadata candidates for documents
   - Create model to visualize and discuss approach with Action Research team.

8. Dissemination of results
   To share results of the first loop of Action Research with others the model (first version of prototype) and general approach was presented and discussed with the Action Research team on February, 14th, 2011.
   Part of the results have been published in Thönssen (2011).

[140] Linked Movie Data Base (Linked MDB). URL: http://wiki.linkedmdb.org/Main/About (retrieved: 21.11.2010)
[141] DBpedia Knowledge Base. URL: http://dbpedia.org (retrieved: 21.11.2010)
[142] Nanoo.TV. URL: http://www.nanoo.tv (retrieved: 21.11.2010)

4.3.1.2 Research Questions Addressed Within the First Loop of AR With AHSGA

The Research questions worked on within the first loop of Action Research with AHSGA have been answered as follows:

1. Important metadata are those concerning the content of a document, like subject and relation (answer to RQ1). Subject of a document can be a client, a service, a product etc., i.e. related enterprise objects.

2. With respect to RQ2 about determination of documents' context, Action Research with AHSGA proves results of the survey. In general it can be said that low-level governance instruments, like descriptions of organization units, business functions and services or product descriptions for intangible and tangible products determine documents' context.

   According to Dey & Abowd (1999) context characterizes the situation of a particular entity, like location, time or activity. This is called *primary context* and does not "only answer the questions of who, what, when, and where, but also acts as indices into other sources of contextual information" (ibid.) For AHSGA's documents primary context is considered the place (location) where the document is stored (directory of their file system) and business entities determine its structure, like business actor, i.e. organisation unit, business service or a related resource.
   Figure 41 provides a cartoon of a document's context (here for an image of a kissing couple) that is stored in the AHSGA's file system. In the cartoon, primary context is represented by solid arrows from the document to the five entities `Node` (where), `BusinessActor` (who), `Product` (what), `BusinessService` and `BusinessBehaviourElement` (how). The rounded rectangles in the figure indicate that the depicted concept (e.g. `BusinessActor`) actually is a super concept for instance of `OrganisationalUnit, LegalEntity` or `Person.`

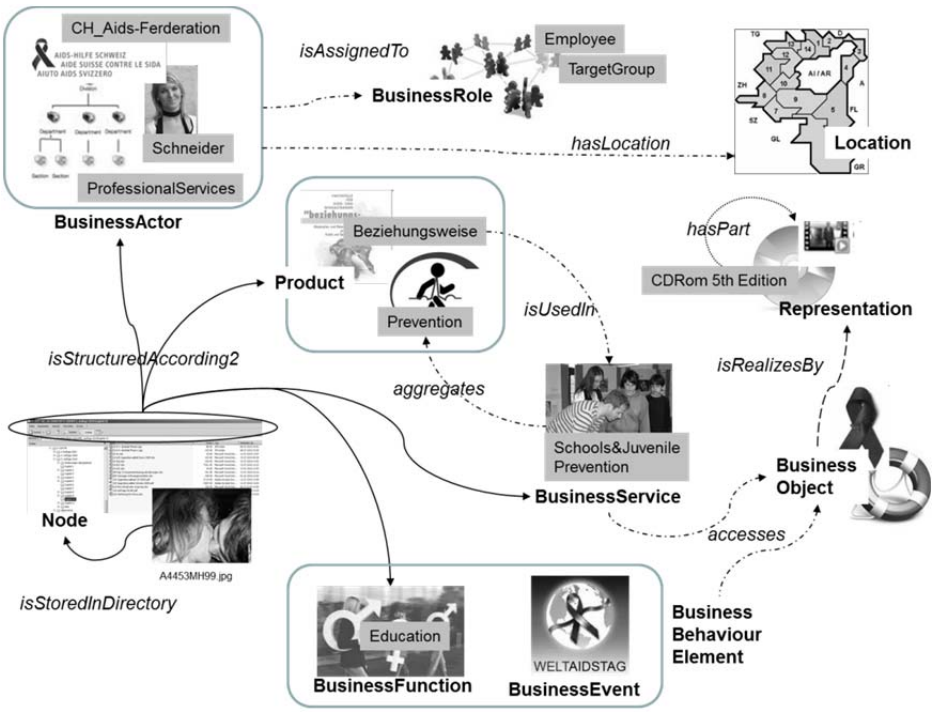


Figure 41: AHSGA Documents' Context

As described by Dey & Abowd (1999), primary context elements are used to find secondary context, namely `BusinessRole`, `Location` and `BusinessObject` with reference to `BusinessService`. Secondary context is represented by dashed arrows from the primary context elements introduced above to the five secondary context elements. As AHSGA has business processes not explicitly modelled documents cannot be related directly to an activity or process.

Whereas Dey & Abowd (1999) introduce a *two-tiered system* (all context elements that are not on primary level are considered to be on secondary level), I suggest a multi-tier context model. In Figure 41 tertiary context level is depicted as broader-dashed arrows), for example from `BusinessObject` (secondary context) to `Representation`, and a fourth from `Representation` to its parts `Document`, which is illustrated as a movie pictogram in the cartoon.

3. All types of AHSGA's documents provide document properties that can be harvested (answer to RQ4). But often, especially in case of an image, only the path of the directory that an image is stored in is of value. In the example depicted in Figure 41 the path name that is harvested for the image "A4453MH99.jpg" is: F:\TEXT\PROFSERVICE\DIRECTPREV\JUVENILE_Homes\AUBODEN.[143] To use that information for metadata generation the underlying structure can be exploited as described above.

4. RQ3 is about context entities that can be inferred for metadata generation. As aforementioned a multi-tier context model is suggested although it is assumed that with growing distance to the document the certainty of generated metadata might decrease because of the growing number of inferred possibilities.

5. In Hinkelmann et al. (2010) and Thönssen & Wolff (2010) we introduced Enterprise Architecture Frameworks as good guidance to determine constituents of an enterprise architecture. Within loop 1 of the Action Research study this approach has been challenged as AHSGA do not follow an architecture framework but simply defined low-level governance instruments. These instruments are considered organisation specific objects which constitute AHSGA's enterprise architecture (answer to RQ8).

6. The low-level governance instruments—templates for document creation, descriptions of organization units, business functions and tasks or product descriptions for services and tangible products, naming conventions for files and directories and pre-defined directory structures—can be formally represented in the enterprise ontology (answer to RQ9).

7. The concept `Document`, considered a sub-concept of `Representation` in the enterprise ontology (according to ArchiMate), is described based on Dublin Core (DC). To indicate whether a property is a simple, qualified or non-standard qualified DC element the respective element type is indicated in parenthesis after a property name. For example: `subject(dc)`, indicating a simple DC element, `created(dcq)`, indicating a qualified DC element, and `subjectRole(dceo)`, indicating a non-standard qualified DC element (answer to RQ16). Qualifications of metadata elements are

[143] Path name is a translation of "F:\TEXT\AH_GS\DIREKTEP\HEIME_JUG\Auboden "

modelled as sub-properties to allow for applying the Dublin Core Dumb-down Principle.[144]

8. The concept `subject` is special as it indicates the use of SKOS, a "standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web" (W3C, 2010). That approach is chosen to represent domain knowledge, for example about “sexual health”, as needed by AHSGA (answer to RQ15).

Requirements derived from the findings described above are detailed in Chapter 0.

Results of the second loop of the Action Research study with AHSGA are provided within Chapter 6.1.6, p 148 ff. and results of the third loop are described in Chapter 8.2.1, p 240 ff.

4.3.2 Action Research Study With Symfact AG

The Action Research study with Symfact is also conducted in three loops (as introduced in Chapter 2.2.3). First loop of the study focuses on as-is analysis and was executed between July 2010 and June 2011. The study took place within the DokLife[145] project.

Symfact AG develops and sells software, amongst others for Compliance and Contract Lifecycle Management (CLM). The CLM-system is used by more than 150 customers in 20 different countries. Symfact’s configurable, XML-based CLM-system supports companies of all business sectors in managing their contracts. However, today all contracts must be imported and annotated manually. Thus, within the DokLife project a prototype for automatic metadata creation for contracts has been developed.

Contrary to AHSGA documents are of pdf format and mainly of text type (annexes to contracts could be images) and metadata to be generated is pre-defined by the attributes describing contracts in the CLM-system. Focus in this research study is on improving contract lifecycle management with respect to event-based obligation and contract retention management based on semantically enriched context information.

4.3.2.1 Results of the First Loop of Action Research With Symfact

Results of the first iterative cycle with Action Research partner Symfact are as follows.

1. Inventory on document handling

   Table 8 gives an overview on document handling as it is. Unlike AHSGA the documents considered are not managed by Symfact but by companies implementing Symfact's CLM-system.

[144] "The qualification of Dublin Core properties is guided by a rule known colloquially as the Dumb-Down Principle. According to this rule, a client should be able to ignore any qualifier and use the value as if it were unqualified. While this may result in some loss of specificity, the remaining element value (minus the qualifier) must continue to be generally correct and useful for discovery. Qualification is therefore supposed only to refine, not extend the semantic scope of a property." URL: http://dublincore.org/documents/usageguide/ (retrieved: 5.3.2011)

[145] DokLife research project was funded by the Commission for Technology and Innovation (CTI). Project KTI Nr. 10902.2 PFES-ES des Bundesamtes für Berufsbildung und Technologie. URL: http://www.doklife.ch/ (retrieved: 2.12.2012)

Aspect	Occurrence
File Organisation	Documents are either stored electronically or in paper form in folders. For metadata generation documents are scanned and/or transformed into pdf-files.
Document Types	Text PhysicalObject (paper document)
Document Formats	pdf and paper
Document Layout	Layout is very heterogeneous as contracts of many different companies are to be processed
Metadata Elements	Metadata elements are pre-defined by the CLM-system; the whole metadata set comprises on average 30 elements[146]
Domain/Context Knowledge	Symfact Contract Glossary
Software / Tools	Symfact Contract Lifecycle Managing System

Table 8: Overview on Document Handling With Symfact's CML-System as-is

In contrary to AHSGA in Symfacts case no original contract documents could be used due to data protection concerns. Therefore test data was created to represent the original data.

2. Identification of enterprise-specific problems with document handling
   As Symfact does not deal with documents on their own but wants to provide a solution that is applicable for 'all' of their customers, the variety of contracts (with respect to layout, content, wording etc.) is enormous. Because of the so called 'freedom of contract' most of the contracts are hybrids containing arbitrary elements of legally defined types but freely defined by the contract parties. Those contracts are called innominate contracts.
   Input format of all documents to be processed is PDF/A, a file format for the long-term archiving of electronic documents. It is based on the PDF Reference Version 1.4 from Adobe Systems Inc. (implemented in Adobe Acrobat 5 and latest versions) and is defined by ISO 19005-1:2005. [147]

3. Collection and specification of requirements
   Goal of the DokLife project was to automatically create as many metadata values for the given metadata set as possible in good quality. Requirements are specified in an internal requirement specification (Thönssen et al. 2011). Relevant for the Action Research study are the two enhancements: verification of automatically extracted metadata and improving obligation management. That is, automatically extracted metadata (Named Entities like names of contract partners) should be verified, for example to identify the role of the underwriter in the company. Contract management should be improved by actively supporting the handling of obligations.

   Generated metadata values must be presented in XML-format to be imported in the CLM-System of Symfact.

[146] The CLM-system allows for great flexibility. I.e., the definition of a metadata set is application dependent and part of a customizing process

[147] ISO 19005-1:2005. Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1). URL: http://www.iso.org/iso/catalogue_detail?csnumber=38920 (retrieved: 12.2.2011)

4. Comparison of real life situation with a priori statements on the problem
   The analysis showed that document properties do not provide useful information as many contracts are scanned documents or converted from original formats like "doc". Thus, the document properties "creator" or "creation date" that are automatically created by the operating system are related to the converted file and not to the original documents. This means they are not meaningful for a metadata seed and thus, metadata harvesting is not applicable. Instead metadata extraction is to be performed.

   The automation of metadata generation for contracts is difficult because of the following issues:
   - Due to the prevalent innominate contracts automatic classification is almost impossible.
   - Information extraction is extremely difficult because of the huge variety of the contracts, regarding layout, content and language (how facts are expressed).
   - Because of the 'freedom of contract' the assumption that content-related metadata can be derived from context information has only partially proved true. It is not possible to conclude from 'general knowledge' to specific one, expressed in a contract as for every rule there is an exception. For instance: in general a contract is signed by two parties. However, a non-disclosure agreement could be signed only by one and a master contract could be signed by more than two parties.
   - Instead of content-related metadata, administrative metadata comes into focus, as managing documents according to the law is especially important for contracts. As the retention period of a document depends not only on its type (contract) but also on the business sector it is produced in, that information can be used to infer the respective metadata. For example: a leasing agreement in Switzerland may not only comply with Swiss Code of Obligations[148] but also with industry-sector-specific regulations.
   - Today, active contract lifecycle management is limited to time related triggers, for example a date when an obligation is due. By relating a document to the enterprise object(s) it represents changes of these objects can be monitored. For example: in case of bankruptcy of a contract partner affected obligations can identified and the related contract could be suggested for review.
   - For Symfact describing documents with Dublin Core metadata elements is not sufficient. Therefore additional metadata specific for contracts must be defined.

     To sum up: In place of content-related metadata mainly administrative metadata is generated automatically in Symfact's case. Extracted metadata is to be verified, e.g. if the person who signed the contract is eligible to do so (as he/she is authorized to sign in the contractee's company). Instead of using context only for metadata *generation* focus has shifted to using context for *active contract lifecycle management*.

5. Review of current research on the problem
   Approaches for automatically creating semantically enriched metadata for documents in the legal domain are rare. One is CERNO, a framework and tool for supporting semantic annotation of textual documents in the legal domain. In order to support regulation-compliant systems organizations have to analyze legal texts and elicit the requirements (Kiyavitskaya et al., 2009). The authors identify a number of issues to be addressed that

[148] In Switzerland Art. 957 ff. Obligationenrecht (OR) is about book-keeping and retention obligations. SR 220 Bundesgesetz betreffend die Ergänzung des Schweizerischen Zivilgesetzbuches (Fünfter Teil: Obligationenrecht). URL: http://www.admin.ch/ch/d/sr/220/a957.html (retrieved: 12.3.2011)

apply to DokLife, too. Namely finding the relevant pieces of information in legal documents or determining the true meaning of the law because "legal texts are frequently affected by ambiguities and lacunae" (Kiyavitskaya et al., 2009). The approach of Kiyavitskaya et al. (2009) aims for semi-automatic annotations of text resources with 'light-weight' tools and techniques. 'Semi-automatic', because a human is involved in the annotation process. 'Light-weight', first because the semantic model is represented in UML class diagrams and not in terms of an expressive description logic (like OWL), second because source code analysis tools and techniques are used instead of Natural Language Processing (NLP) techniques.

Dealing with the problem of organizations of how to align information systems requirements with regulations (Breaux & Antón 2008) extended CERNO to automate the extraction of rights and obligations from regulations. They start from manually marking obligations, associated constraints and condition keywords, including natural language conjunctions in text. However, regulations and law differ from contracts in many ways, e.g. a regulation does not have persons named or a specific product described.

Palmirani & Brighi (2003) also deal with regulations and introduced an approach to convert legislative text into XML documents. For this structural elements of the text are extracted, semantically annotated by the user and outputted in an XML document. As the approach is carried out for *Italian* legal documents it cannot be applied to DokLife - besides the difference between legislative texts and contracts. Very interesting is the development of the xmLegesEditor, a "Legislative drafting environment developed at ITTIG/CNR for supporting the adoption of Italian Legislative National XML Standards (NIR)" (Agnoloni, Francesconi, & Spinosa, 2007). The xmLegesEditor does not only allow the writing of new legal documents according to the standard but can be extended to the xmLeges Suite. The xmLeges Suite provides modules like a structural parser that transforms a legacy normative document in plain text into XML-NIR format, or a classifier to detect semantics in a normative document. Although the approach is promising it is developed for writing/editing regulations (not contracts) and relying on Italian law.
To my knowledge only two approaches specifically deal with automatic metadata generation for contracts. Both are implemented in commercial products, one by Mumboe[149], the other by openSource[150], both Symfact competitors. Both companies provide the contract management software as a service, that is a contract is uploaded into the system and stored on the provider's web-site. Once a document is imported into the system, specific details such as effective dates, expiration dates, parties etc. can be extracted automatically.[151] Both approaches are patented (Guerra Currie et al. 2007, Zernik 2008).

However, neither of the tools seems to model semantics explicitly let alone considers the context of contracts. Therefore improving metadata extraction by exploiting context information and improving active contract lifecycle management would bring Symfact a significant competitive advantage.

6. Question to be answered

[149] Mumboe. URL: http://mumboe.com/index.php (retrieved: 12.2.2011)
[150] openSource. URL: http://www.opensourceinc.com/drupal/ (retrieved: 12.2.2011)
[151] Detailed information on Mumboe's extraction process is provided in an online tutorial. URL: http://mumboe.com/features/features_auto_extract.php (retrieved: 12.2.2011)

Results of the first loop of Action Research with Symfact lead to the following questions:

- How to represent the contract model, according to the ArchiMate standard in seEAD?
- What context elements are relevant for contracts?
- What background knowledge must be considered to improve CLM with respect to event triggered obligations and records management?

7. Actions to be taken
   First loop of Action Research with Symfact results in the following actions:
   - Enhance/adapt enterprise ontology to Symfact's specifics
     - Evaluate ArchiMate representation of contract model for DokLife
     - Represent Symfact's document model for contracts in seEAD
     - Define competency question to evaluate seEAD
   - Run through example for one contract type, e.g. Non-Disclosure Agreement (NDA)
   - Create model to visualize and discuss approach with Action Research team.

8. Dissemination of results
   Results of the first loop of Action Research with Symfact have been presented and discussed within a DokLife project meeting on June, 6th, 2011.
   The general approach has been presented to a broader audience at Symfact's User Conference in Sugiez, Switzerland from September 28- 29, 2011.

Within the first loop of Action Research with Symfact, answers to the following research questions have been worked out.

4.3.2.2 Research Questions Addressed Within the First Loop of AR With Symfact

Research questions worked on within the first loop of Action Research with Symfact have been answered as follows:

1. Although the metadata element set has been pre-defined by the contract attributes Symfact's CLM-system manages, actually most of these attributes do not describe the contract *document* but terms and conditions contract partners agreed on. Hence I differ between metadata of the contract document itself (answer to RQ1) and properties, *recorded in a contract document* but determine the contract in real terms. That differentiation has been made by the ArchiMate standard, too. Refer to Chapter 6.2.2, p 187 for the respective context model.

2. Properties determine the contract in real terms are considered the contract document's context, e.g. the product the contract is about, the contract partner the contract is concluded with and so on (answer to RQ2). Figure 42 gives an overview on context elements of a contract document.

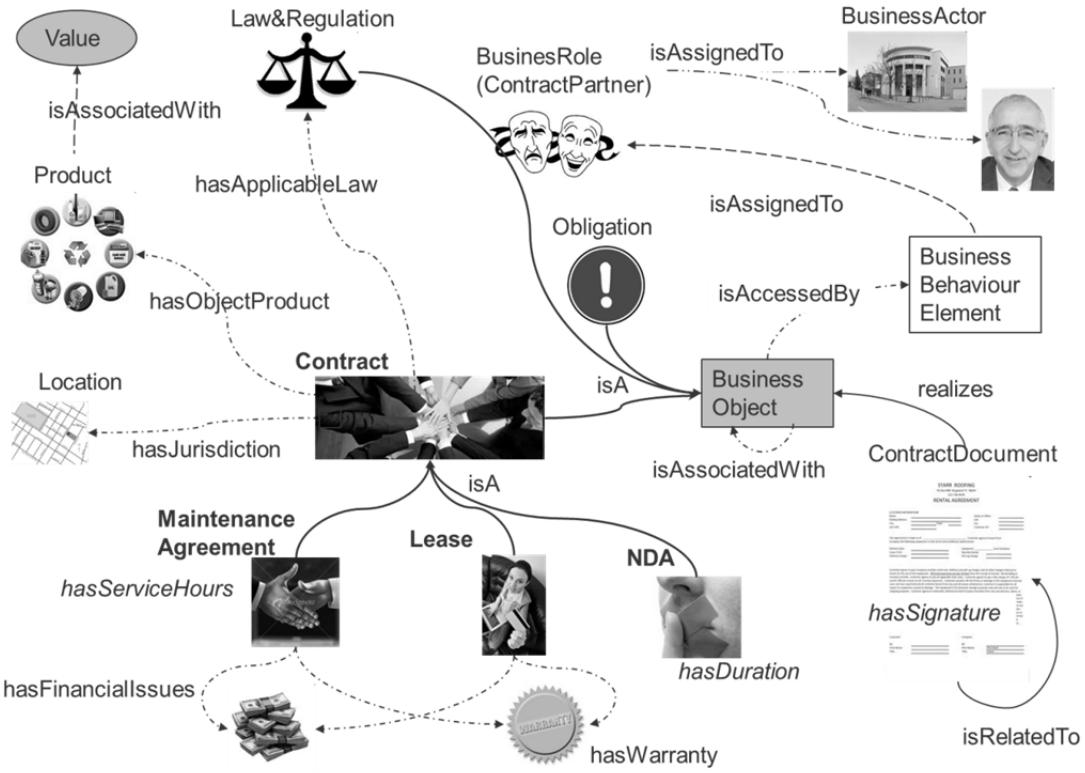


Figure 42: Contract Documents' Context

Primary context of a contract document is the `BusinessObject`, respectively its specialisations, e.g. `Contract` or `Obligation.`
Secondary context (again depicted as narrow-dashed arrows) are for example `Product, Location` or `Law&Regulation.` In Figure 42 tertiary context level is depicted as broader-dashed arrows and shown for the elements `BusinessRole` and `Value.` A fourth context level is depicted from `BusinessRole` (tertiary context) the assigned `BusinessActor`, which can be for example a legal entity or a natural person.

3. As contract documents are legally binding electronic storage must comply with special provisions of law for records management. In Switzerland for example a contract is considered *business correspondence* and therefore retention period is determined by law inter alia by Civil Law, especially the Swiss Code of Obligations. Thus a retention period can be inferred automatically (answer to RQ3). For contracts retention period is at least 10 years after the contract is expired. Of course Swiss Code of Obligations is not the only regulation controlling management of legally binding documents as there are trade-specific rules, like banking law. Modelling archiving obligations in seEAD would improve automatic records management with respect to automatically compute retention period. For example, based on a document's type, the industry sector it belongs to and the underpinned and law and regulations, retention period could be calculated automatically and if applicable be used for automated document lifecycle management. This means that within the DokLife project seEAD is primarily used for generating administrative metadata.
   However, descriptive metadata will be used to verify (automatically extracted) metadata and actively manage a contract's lifecycle. By inferring for example affected options in case of a certain business event (e.g. force majeure event, disaster, bankruptcy, injunction, etc.) a warning might be issued automatically.

4. As almost all of the documents considered in this Action Research study with Symfact are scanned documents of PDF/A format, harvesting document properties is of limited use. Therefore source for metadata generation is extracted information, for example named entities of categories such as the underwriter of a contract (e.g. the employee representing the contract partner), legal entities (e.g. the contract partners) or locations (e.g. city of jurisdiction or country of applicable law) (answer to RQ4).

5. To define enterprise objects constituting an enterprise architecture for contract management (RQ8) I draw upon the ArchiMate standard.

6. To represent enterprise architecture formally it can be modelled in an enterprise ontology (RQ9). Please refer to Chapter 6.2.2 for details.

7. As already described for AHSGA the concept "subject" refers to a SKOS thesaurus to represent legal domain knowledge (answer to RQ15). In order not to start from scratch several online legal dictionaries have been investigated, amongst others "The free legal dictionary by Farlex"[152], "JurWordNet"[153] and "EuroVoc"[154]. In addition Symfact's internal contract glossary has been considered but is too 'ad hoc' for scientific use.

8. To meet standards and enterprise specifics alike metadata for contracts are described with Dublin Core metadata elements, extended by ArchiMate descriptions (RQ16). For example, the contract *document* is described with (refined) Dublin Core elements, e.g. `dateSigned(myDC)` a refinement of `date(DC)`. The *contract* a contract document represents is described by the ArchiMate *realization* relation: `contractDocumentRealizesContract`.

Requirements derived from the findings described above are detailed in Chapter 0.

Results of the second loop of the Action Research study with Symfact are provided within Chapter 6.1.6, p 179 ff. and results of the third loop are described in Chapter 8.2.2, p 243 ff.

[152] 'The free legal dictionary by Farlex. "The main source of TheFreeDictionary's legal dictionary is West's Encyclopedia of American Law, Edition 2, which contains more than 4,000 entries detailing terms, concepts, events, movements, cases, and individuals significant to United States law. The legal dictionary also incorporates The People's Law Dictionary, by renowned authorities Gerald and Kathleen Hill. It includes definitions, context, and usage for more than 3,000 terms. Regarded by scholars, jurists, leading attorneys and reviewers as one of the most practical works of its kind, The People's Law Dictionary is a comprehensive source of meanings and use for thousands of today's most common legal terms. It has gained widespread praise for its scope and clarity." URL: http://legal-dictionary.thefreedictionary.com/ (retrieved: 18.3.2011).

[153] 'JurWordNet' is a multilingual legal wordnet development made within the LOIS (Lexical Ontologies for Legal Information Society) Project co-ordinated by the ITTIG and co-financed by the European Union under the eContent Program (2003-2005). ITTIG is the Institute of Legal Information Theory and Techniques (ITTIG) belonging to the Italian National Research Council. URL: http://www.ittig.cnr.it/IndexEng.htm (retrieved: 18.3.2011)

[154] EuroVoc is a multilingual, multidisciplinary thesaurus provided by the Publications Office of the European Union. "The aim of the thesaurus is to provide information management and dissemination services with a coherent indexing tool to effectively manage their documentary resources and to enable users to carry out documentary searches using controlled vocabulary." URL: http://publications.europa.eu/eurovoc/index_en.htm (retrieved: 5.9.2011)

4.4 *Requirements for Automatic Metadata Generation in Enterprises*

In the following the core requirements for automatic metadata generation are specified based on analysis of literature review, representative study, survey on document handling and Action Research studies.

Table 9 is structured according to five aspects: quality (Q), context (C), metadata (MD), rules (R), and implementation (I). The first column of the table contains the aspect's shortcut followed by a consecutive number for each requirement, second column describes briefly the requirement, third column gives criteria for measurement, a remark can be provided in column four and the last column indicates the source the requirement is derived from.

Requ. No	Requirement	Criteria for Measurement	Remark	Source
Q1	implement and evaluate automatic metadata generation within a real use case	implementation and evaluation is done within two Action Research studies based on real data and documents	"Research in this area [automatic metadata generation] is important, although examination is generally limited to selected experimental domains" (Greenberg et al., 2006).	(Greenberg et al., 2006), AHSGA, Symfact
Q2	build seEAD in the right balance of expressiveness and decidability	seEAD is expressive enough to model the required knowledge but remains decidable		(Bechhofer et al. 2002)
Q3	use standards to build seEAD	seEAD is based on evaluated and approved standards		(Bechhofer et al. 2002)
Q4	ensure quality of seEAD by sticking to essence	only those enterprise objects are represented in in the ontology and related to enterprise component that are required	in contrary to the definition of the Enterprise Engineering Institute and (Dietz, 2006) the ontology is not regarded independent from its implementation	(Enterprise Engineering Institute, n.d.)
C5	provide stakeholder specific views on seEAD	depending on the stakeholder (e.g. AHSGA or Symfact) other context is used for metadata generation or document lifecycle management	another research question raised by Greenberg et al. was about "the different contextual needs of metadata (e.g., which metadata elements are important for which functions and which classes of users)?"	(Greenberg et al., 2006), AHSGA, Symfact

Requ. No	Requirement	Criteria for Measurement	Remark	Source
C6	use context of documents for active support of document life-cycle-management	based on context, dependencies or implications of change on documents are identified, e.g. if a product changes what specifications are affected	although for implemented prototype will be specific for contract management the solution can easily be generalised as it is based on the generic notion of documents as representations of business objects (The Open Group, 2009b)	Symfact
C7	adapt and enhance seEAD based on enterprise specific governance instruments	seEAD reflects content of enterprise specific governance instruments like a management handbook		Survey, AHSGA
C8	represent governance instruments formally	low-level governance instruments are modelled in an enterprise ontology		Survey, AHSGA
C9	relate directory structure to seEAD	storage location of a document is parsed and analysed for information about the document (context) (Soules & Ganger, 2005)	criteria for directory structure are business entities like product, customer or business function	Representative Study, Survey, AHSGA
MD10	provide context for metadata generation	business objects - documents represent – and their relations are formally modelled in the semantically enriched Enterprise Architecture Description (seEAD)	participants in the AMEGA project suggest "taking a more holistic approach to metadata creation, highlighting the need for information systems to consider context" (Greenberg et al., 2005)	(Greenberg et al., 2005), AHSGA, Symfact
MD11	automatically generate metadata regardless of a document's type	content related metadata is automatically generated for all kinds of documents (text, image, sound)	"Research and experimentation on automatic indexing of language text has been under way for several decades, and there has been much progress and growing use of automatic indexing of language texts for retrieval. But the automatic indexing of image text has barely begun" (Anderson & Pérez-Carballo, 2001)	(Anderson & Pérez-Carballo, 2001), AHSGA, Symfact

Requ. No	Requirement	Criteria for Measurement	Remark	Source
MD12	harvest document properties of the following file formats doc, pdf, ppt, xls, jpg, gif, png, mp3, mp4	all document properties of the specified formats are harvested		Representative Study, Survey, AHSGA
MD13	determine retention period of a document based on qualitative instead of formal criteria	metadata (dceo:archiveDate) is generated automatically based on a document's context (the business object a document represents, the branch of trade the enterprise is in, the law that have been obeyed, etc.)	(Bock, 2005, p.2) stresses the fact, that today information lifecycle management is based on a few operating system attributes although "business value of ILM [..] depends on our capability to categorize content--accurately and succinctly".	(Bock, 2005), Symfact
MD14	specify rules for inferring context to automatically generate metadata	generic rules are specified (like "for all primary context elements of a metadata seed all n-ary? context elements are inferred as metadata candidates")		(Dey & Abowd, 1999), AHSGA
MD15	derive rules for analysing file names based on low-level governance instruments (e.g. naming conventions)	rules are defined reflecting naming conventions, e.g. 'if an employee's number in a file name = 1 then 'creator is <employee name>'		Survey, AHSGA
R16	provide interface between seEAD and existing platforms	seEAD can be used by target systems, like AHSGA's Time Recording System, or Symfact's Contract Lifecycle Management System		(De Bruijn 2003)
I17	keep metadata generation solution independent from upstream and downstream function	the solution is independent from harvesting or extraction tools (upstream function) and from Information- or Document-Management-Systems (downstream function)		AHSGA, Symfact
I18	enable machine-processing of seEAD	the semantically enriched enterprise architecture can be used by humans and machines alike		(Bechhofer et al. 2002)

Requ. No	Requirement	Criteria for Measurement	Remark	Source
I19	create metadata with as little user interaction as possible	automatic metadata generation is performed in the background and no extra effort from the user is required	"people are lazy"	(Doctorow, 2002), Survey, AHSGA

Table 9: Requirements for Automatic Metadata Generation

4.5 *Summary of Requirements Engineering*

In addition to requirements taken from literature research, in Chapter 4 field requirements were analysed and specified for the mintApproach.

First source for field requirements analysis was the representative study conducted in the MATURE project. From it I derived requirements for the file formats to be supported and indicators for the enterprise objects constituting enterprise documents' context. From the survey I gained requirements from non-information specialists for improving document handling in daily routine. Furthermore I could derive requirements for modelling the context of documents from the governance instruments used in the enterprise. Finally I could identify requirements for the types of metadata elements practitioners are most interested in.
I complemented requirements engineering within two Action Research studies. In AHSGA's case requirements are foremost determined by the variety of document formats (text, images and increasingly audio and video) and the need to find documents related to task AHSGA employees report in their Information- and Task Reporting system.
Symfact's requirements concern mainly administrative metadata and possibilities to verify metadata. Most important in Symfact's case are the requirements defined for improving contract lifecycle management. Instead of using context only for metadata generation focus has broadened to using context for active contract lifecycle management.

5 Models for Context-Based Metadata Generation

Chapter 5 of my thesis provides the general, enterprise and application independent, conceptual models, derived from the requirements presented in the previous chapter, as illustrated in Figure 43.

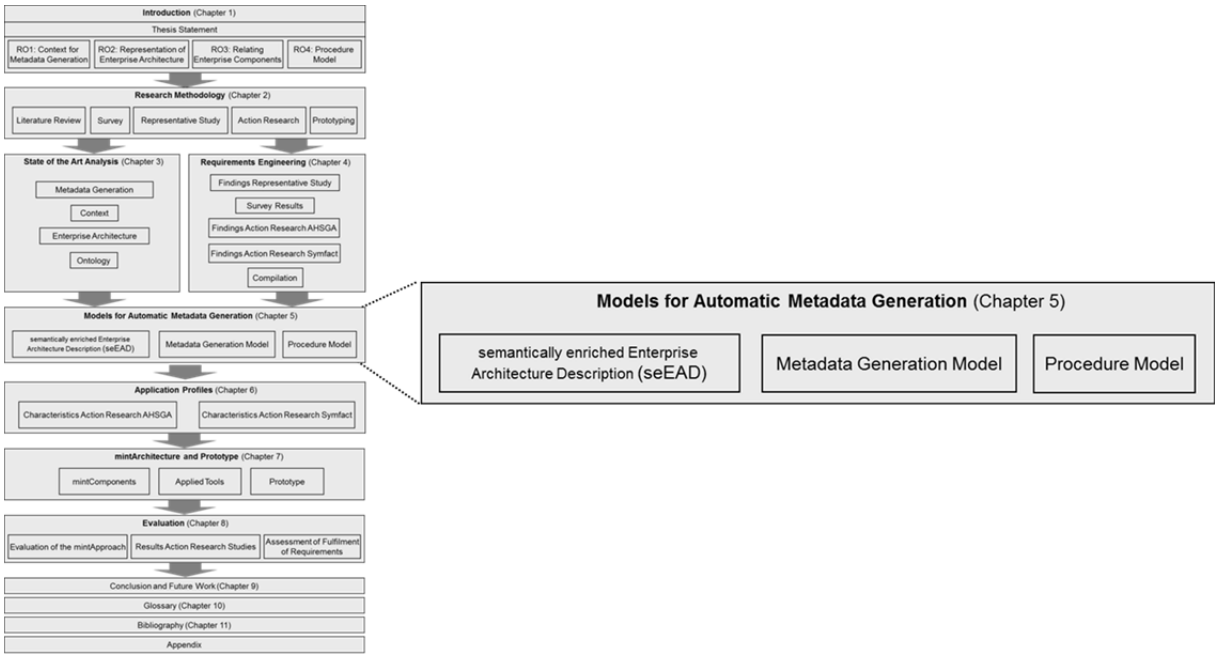


Figure 43: Position of Chapter 5 in the Overall Structure of the Thesis

Automatic metadata generation based on context comprises several conceptual models, which are enterprise-independent but customizable to enterprise specific needs. In Chapter 5 these general models – I call mintModels – are introduced. In Chapter 6 examples of enterprise-specific customizations of the models are provided for the two Action Research partners. In Chapter 7 prototypical implementations of the enterprise-specific mintModels are described as proof of concept.

According to Moody (2005, p 243) "Conceptual modelling is the process of formally documenting a problem domain for the purpose of understanding and communication among stakeholders" and "Conceptual models are central to IS analysis and design, and are used to define user requirements and as a basis for developing information systems to meet these requirements".
The notion of model as used in my thesis meets the definition given in the Stanford Encyclopedia of Philosophy (Frigg & Hartmann, 2012) stating that a model is considered a representation of a selected part of the world (the 'target system'). Depending on the nature of the target, a model is either a model of phenomena or a model of data. Likewise, a model can represent a theory in the sense that it interprets the laws and axioms of that theory. The models I developed are representations in both senses at the same time as the two notions are not mutually exclusive (Frigg & Hartmann, 2012).

In this chapter the general models of the mintApproach are explained: the models that constitute the context for metadata generation (mintContext), the models for metadata generation and the model for implementing and customizing the general models in a specific enterprise. Figure 44 depicts the general models of the mintApproach introduced in this chapter. Chapter 5 is structured accordingly: first the conceptual model to determine the

context for automatic metadata generation is provided, then the metadata generation model is detailed and after that the procedure for implementing and customizing of the mintApproach is explained. The chapter concludes with the achieved findings.

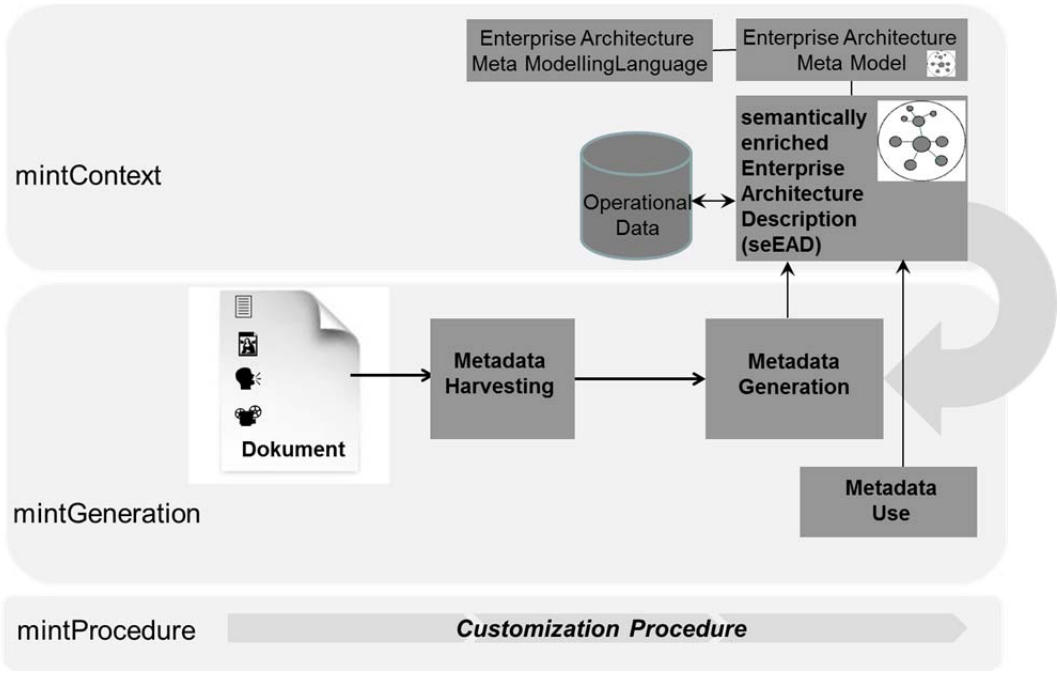


Figure 44: General Models of the mintApproach

5.1 *The Context Model*

Most important for the mintApproach is the context, which can be inferred for metadata generation. Context, used for automatic metadata generation for business documents, is composed of enterprise objects. Enterprise objects – like a person, a role, a function, a task etc. – are defined and related to each other according to an enterprise's conception of its business, its Enterprise Architecture. Enterprise Objects and their relations are defined in an enterprise architecture description. Since the main goal of the mintApproach is to automatically generate metadata based on a document's context a sound foundation is a condition as well as a machine processable representation.

Following Kang et al. (2010) and Ettema & Dietz (2009) I consider an enterprise architecture description well suited to determine context, needed for automatic metadata generation. However, as already shown in Chapter 3, existing Architecture Description Languages are not formalized enough for a machine processing representation.

Because of the consensus that representing enterprise architecture description in an ontology is appropriate (amongst others Dietz & Hoogervorst 2008, Ettema & Dietz 2009 and Hinkelmann et al. 2010), I also decided to formally represent the enterprise architecture description in an ontology.

As also shown in Chapter 3, existing well-known enterprise ontologies have some drawbacks with respect to their use for automatic metadata generation based on context. An enterprise architecture description, represented in an ontology to be used for automatic metadata generation has to cover all of the below listed characteristics:

- is machine processable but cognitive adequate for humans

- is formalized in a way that allows for reasoning
- has an appropriate a level of granularity which allows for operational use (i.e. can be used in business applications)
- is part of an enterprise repository (i.e. is related to information not stored in the ontology)
- is based on standards (e.g. all and foremost ArchiMate, Dublin Core and the W3C standards for ontology languages).

This justifies the creation of a new enterprise ontology since it is not "yet another point of view on enterprises" as Dietz (2006, p 6) puts it, but "assist[s] in coping with current and future problems related to enterprises", namely the management of unstructured information, i.e. documents, both text and multi-media.

However, ontology development is a labour intensive task, although much research has been done on automatic ontology learning (amongst others Maedche & Staab 2001, Dellschaft & Staab 2008). Since expert knowledge is needed for ontology development, costs increase if it is not available within the enterprises. Thus, it is not feasible to create an enterprise ontology form scratch all over again in an enterprise that wants to improve their document handling by automatic metadata generation. Neither is it possible to create an enterprise ontology that can be used out-of-the-box in every enterprise. Furthermore, creating an ontology which only serves one purpose would not justify the effort – same as shown for dedicated context models (cf. Chapter 3.2.3, p 40) and only partially exploit the potential an enterprise ontology has.

5.1.1 Ontology Design Rationale

To decrease ontology development costs and better exploit the potential of a semantically enriched enterprise architecture description I decided to create a meta model that can be used to model the enterprise specific semantically enriched Enterprise Architecture according to the Meta Object Facility specification (OMG, 2011b) and thus, can serve as 'General Enterprise Model' or 'Core Enterprise Ontology' as suggested by Fox & Gruninger (1998) and Bertolazzi et al. (2001), respectively. This meta model I call 'ArchiMEO'.
Figure 45 shows the Conceptual Model for Architecture Description (AD) defined by the ISO/IEC/IEEE 42010 (DSCI 2012) applied to the enterprise architecture domain. For my approach I extended the model by the two concepts 'Architecture Meta Model' and 'Meta Modelling Language'. Examples for the ISO/IEC/IEEE 42010 concepts related to my work are expressed in italic.

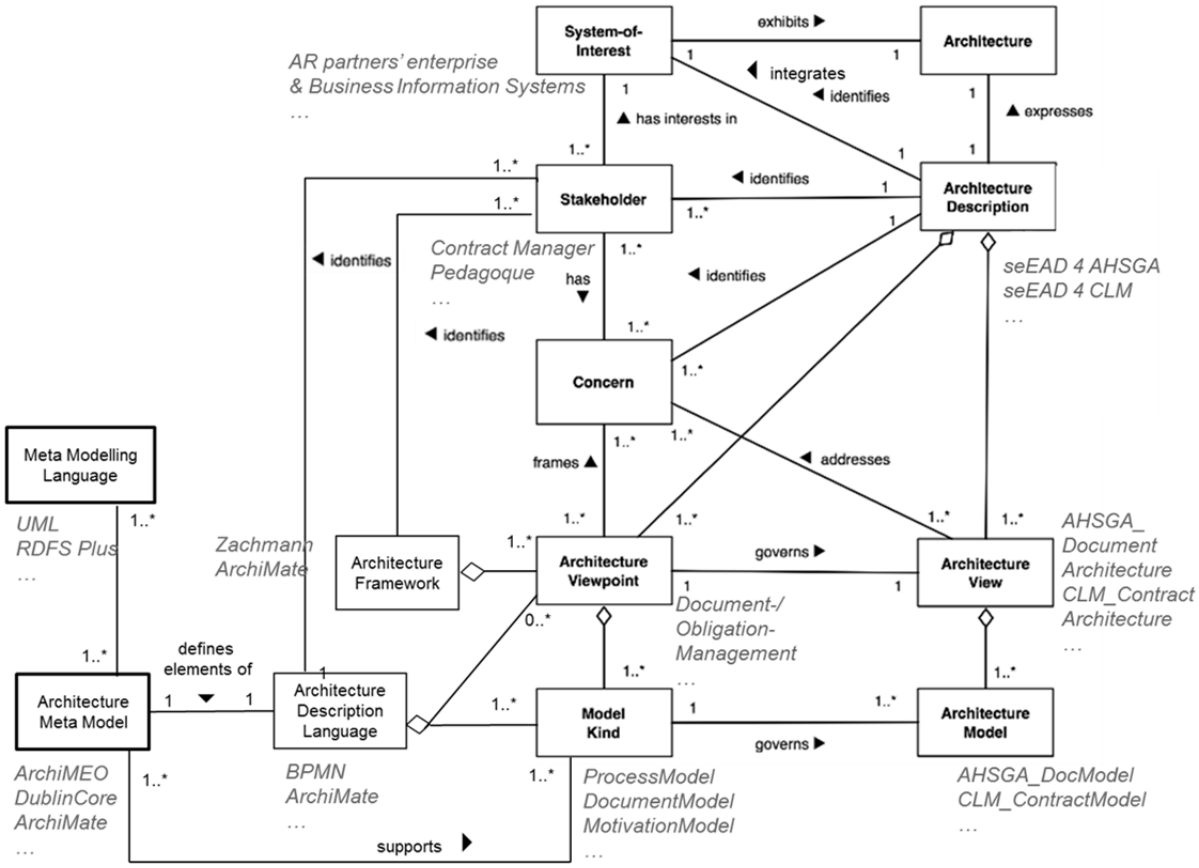


Figure 45: Extended Conceptual Model for EAD (based ISO/IEC/IEEE 42010 provided by DSCI 2012)

At the top left of the figure the System-of-interest is depicted. Besides the notion of the Action Research partners' enterprises, system of interest is although the partners' Business Information Systems. Thus, to express the operational use of an enterprise architecture description I add the relation '◂ integrates' between Architecture Description and System-of-interest. 'seEAD 4 AHSGA' and 'seEAD 4 CLM' indicate the two realizations of enterprise architecture description I did in my work. The Architecture Descriptions identify the Stakeholders, e.g. a contract manager and a pedagogue and their concerns. Architecture Descriptions aggregate 1 ore more Architecture Viewpoints and Architecture Views.

According to ISO/IEC/IEEE 42010 an Architecture Viewpoint represents the interest of one ore more stakeholders. In AHSGA's case the management of documents and in Symfact's case the management of obligations. An Architecture Viewpoint governs an Architecture View. "An Architecture View in an Architecture Description expresses the Architecture of the System-of-Interest from the perspective of one or more Stakeholders to address specific Concerns" (DSCI 2012). Examples are given in the figure for 'AHSGA_Document_ Architecture' and 'CLM_Contract_Architecture'. Within an Architecture View, *context* expresses parts of the view from the prespective of the model of interest (the context of an AHSGA document and the context of contract document).
An Architecture View consists of one or more Architecture Models, each governd by its Model Kind. For example, the AHSGA_DocModel is governed by a DocumentModel. "A Model Kind defines the conventions for a type of Architecture Model" (DSCI 2012) and is expressed in an Architecture Description Language; the DocumentModel for example in the ArchiMate Language.

On the lower left side of the figure the two concepts I introduced are depicted: The Architecture Meta Model and the Meta Model Language. The Architecture Meta Model supports the Model Kind by providing blueprints, constraints, refinements, encoding schemes etc. ArchiMate is regarded a Meta Model. ArchiMEO is the Meta Model I introduced and which is explained in detail in the following sections of this chapter. A Meta Model is expressed in a Meta Model Language, e.g. UML, RDFS, OWL and the ArchiMate Notation. A Model Kind can be expressed in the same language as its Meta Model.

Figure 46 shows that the semantically enriched Enterprise Architecture (seEAD) in turn is part of the Enterprise Model Ontology (EMO) which adds meta information derived from Enterprise Architecture Frameworks like Zachman's. The database icons depicted in the figure indicate the mapping of concepts of the semantically enriched enterprise architecture description to operational data, as proposed by Hinkelmann et al. (2010) in order to build an enterprise repository.

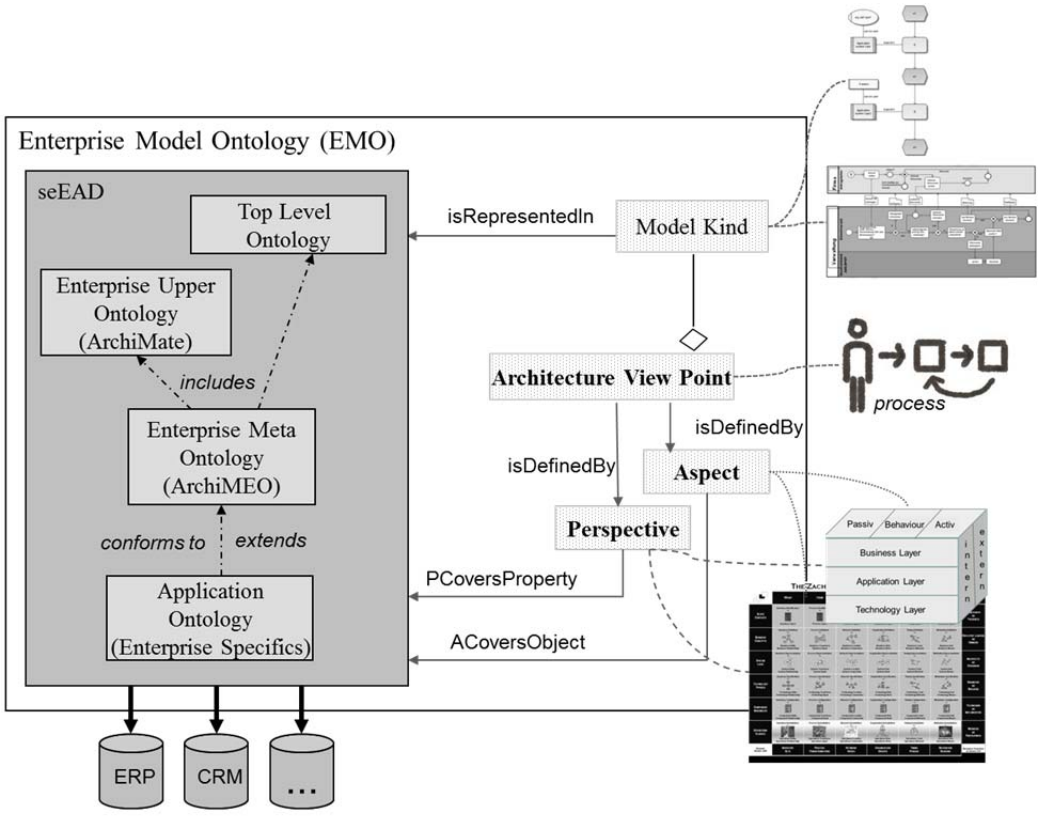


Figure 46: Structure and Representation of Enterprise Objects

Figure 46 also shows how Enterprise Architecture Frameworks can be related to an enterprise ontology, by enhancing the enterprise ontology for the concepts `ModelKind` and `ArchitectureViewPoint` (as defined in the ISO/IEC/IEEE 42010 standard plus `Aspect` and `Perspective`. `ModelKind` (e.g. a process model, a document model) describes a certain section of the enterprise ontology, e.g. all entities needed to model business processes, expressed by the `isRepresentedIn`-relation. `ModelKind`s are aggregated in `ArchitectureViewPoint`s that frames the concerns a stakeholder has. An `ArchitectureViewPoint` is defined by `Perspective` and `Aspect`, which represent the represent items of the Enterprise Architecture Frameworks for example the cells in Zachmann's EA matrix. `Perspective` is related to the enterprise ontology via the

property `PCoversProperty`; `Aspect` is related to the enterprise ontology via the property `ACoversObject`.

With this approach quality of the enterprise ontology can be improved as completeness can be ensured. Furthermore, it provides the basis for creating specific views on an enterprise architecture description: for specific viewpoints, model kinds and architecture description languages (like BPMN, not depicted in the figure).

5.1.2 Modelling Approach for seEAD

As shown in Figure 46 seEAD, logically consists of several ontologies, and all have been developed in an iterative process. A first version has been elaborated in two Master's theses, supervised by me, for supporting business process management (Brun, 2010) and metadata generation (A. Martin, 2010). Focus of the development was on meeting business needs, i.e. that structure and content of the ontology was mainly driven by the business partners in order to create an ontology which is proven by practitioners.

In order not to reinvent the wheel existing enterprise ontologies have been considered for reuse (cf. Chapter 3.4.1, p 56ff). As most of the enterprise ontologies describe the most general and reusable concepts (Pease, Niles, & Li, 2002) and their relations, for this work, the following enterprise ontologies have been analysed: TOVE (Fox et al. 1996), The Enterprise Ontology (Uschold et al., 1997), a Context Based Enterprise Ontology (Leppänen 2005; 2007) and the Core Enterprise Ontology, introduced by Bertolazzi et al. (2001) The existing enterprise ontologies have been used for input and guidelines for the development of seEAD.

However, as stated above, to satisfy business needs already existing ontologies were not sufficient (cf. Chapter 3.4.1, p 56 ff) and thus, based on existing approaches, a new ontology has been created. In order to avoid propriety development but to improve common understanding and facilitate re-use I decided to develop an Enterprise Architecture Meta Model (ArchiMEO), based on the ArchiMate standard. ArchiMEO can then be used to create enterprise specific architecture descriptions. Hence, I adapted the approach of (re)using a core ontology introduced by Fox & Gruninger (1998) and started with the transformation of the enterprise architecture elements and their relations defined in ArchiMate to build the Enterprise Architecture Meta Model. To do so several refinements were necessary as described in the following sections.

Since ArchiMate does not provide general concepts, like location or time, I introduced a Top Level Ontology (TOL) following Bertolazzi et al. (2001), who propose business independent ontologies. The Enterprise Upper Ontology is the ontological representation of ArchiMate, based on the language scope of ArchiMate 1.0 Specification[155] (The Open Group, 2009b). The Top Level Ontology and the ontological representation of ArchiMate together build the ArchiMEO ontology. The name (ArchiMEO) is chosen to indicate its foundation in ArchiMate plus its adaptation and enhancements to serve as a meta model (Meta Enterprise Ontology). Conform to the Meta Object Facility specification (OMG, 2011b), seEAD is expressed in the language of its meta model (Bézivin, 2004), allowing for enterprise specific enhancements needed for operational use in a certain enterprise.

[155] Not considered are the two optional extensions of the ArchiMate language for motivation and implementation and migration. Both extensions are not necessary for automatic metadata generation based on context but provide additional information, for example on the reason lying behind the architecture of an enterprise (The Open Group 2012, p 8 ff)

A semantically enriched Enterprise Architecture Description (seEAD) is then logically structured into four parts:

- a Top-level Ontology (TOL), comprising generic concepts of the world like `time`, `location` and `event`
- an enterprise upper ontology, comprising the ArchiMate concepts represented in an executable ontology language
- an meta enterprise ontology (ArchiMEO), adapting and enhancing the ArchiMate standard by additional concepts and relations, for example to describe business processes and business representations (e.g. documents),
- an application specific ontology, comprising specific concepts of a certain enterprise (one for AHSAG and one for Symfact).

To determine the enterprise specifics I followed the approach of Uschold & Gruninger (1996), asking competency questions to derive domain and application specific concepts and relations (cf. Chapter 6.1 and 6.2 for details on the enhancements related to the two Action Research studies). However, to keep the effort for modelling the enterprise specifics at bay I also followed Chen et al. (2008) who propose the 'principle of fitness-for-purpose' for developing enterprise architecture descriptions. Hence, seEAD for AHSGA and seEAD for Symfact have been developed only so far to fit the purpose of automatic metadata generation and obligation management, respectively. For my thesis and prototyping, physically there is a single seEAD for both Action Research Partners. In the real world there would be one seEAD for each enterprise. When talking about seEAD in the following *a* semantically enriched Enterprise Architecture is meant, i.e the one I created within my thesis is *representative* of an enterprise ontology including application specifics, whereas ArchiMEO is perceived as *the* enterprise *meta* ontology that can be used in any enterprise *as-is* to build the semantically enriched Enterprise Architecture of the enterprise.

I consider the current version of ArchiMEO a “first tentative set” after having carefully analysed the ArchiMate standard and integrated relevant results of previous work, that will vary in time as indicated by Bertolazzi et al. (2001). ArchiMEO is already used successfully in other projects, for example in the APPRIS project for supply risk management (Emmenegger, Laurenzi, & Thönssen, 2012). Its will also be used in the newly started SEEK!sem project[156] to improve knowledge management in an enterprise.
Refer to Chapter 5.3.4, p 143 for a suggestion on how further development could be conducted.

5.1.3 Content of ArchiMEO as Meta Model for seEAD

ArchiMEO comprises the top-level concepts `NCO, Event, Location, Time` and the upper-level concept `Enterprise Object`, which is the root concept for the enhanced ArchiMate concepts, as depicted in Figure 47.

[156] The SEEK!sem research project is funded by the Commission for Technology and Innovation (CTI). Project KTI Nr. 14604.1 PFES-ES des Bundesamtes für Berufsbildung und Technologie.

Figure 47: Root Concepts of ArchiMEO

The top-level concepts `Time` and `Location` are self-explaining; the top-level concept `Event` has been introduced to model non-enterprise related facts like natural disaster or man made events that might affect an enterprise. The top-level concept `NCO` (Non-Categorised Object) is used for concepts and their instances that are used in the mintApproach (e.g. to determine relevant regulations for archiving) but are not part of the enterprise architecture description itself.

Most important is the upper-level concept of ArchiMEO, the `Enterprise Object` that will be detailed in the following. Figure 48 depicts the `Enterprise Object` concept with its six sub-concepts representing the ArchiMate Core Concepts - as illustrated in the ArchiMate 1.0 Specification (The Open Group 2009a, p 21) - namely: `Interface`, `BehaviourElement`, `Service`, `Relation`, `StructureElement` and `Object`.

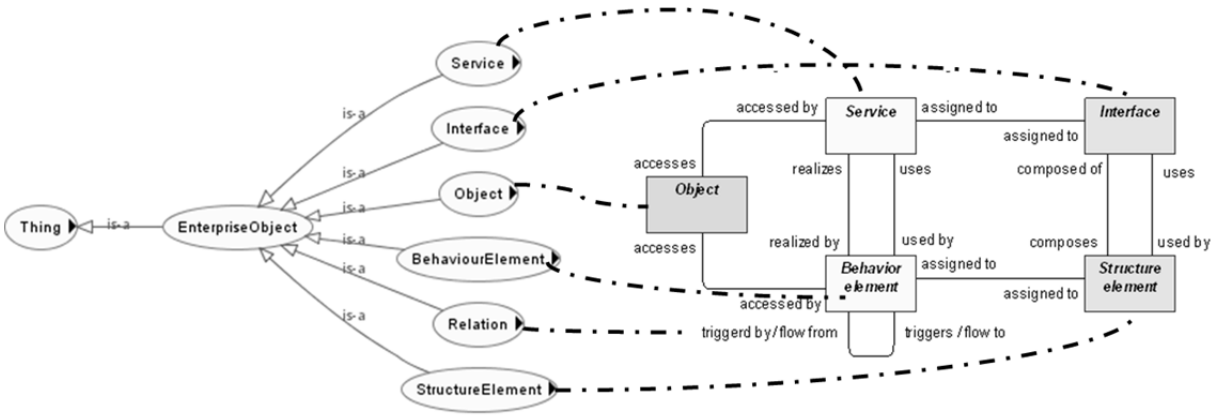


Figure 48: Fitting ArchiMate into ArchiMEO

All concepts of the sub-sequent layers are considered specifications of these super concepts. The little black arrows in the circles representing the concepts indicate that a concept is further specialized.

Figure 49 gives an example of sub-concepts of the concepts `Service`, `Relation` and `Interface`.

Figure 49: Extract of Sub-Concepts Representing ArchiMate Concepts

Note that in ArchiMate these super-sub-concept relations are not explicitly expressed but stipulated for example by the chosen noun compounds to name the concepts, e.g. `Infrastructure`**`Service`**, `Business`**`Service`** and `Application`**`Service`** *are* `Services`.

In order to fomally represent ArchiMate various weaknesses have emerged and were addressed during modelling. Namely:

- No XML-representation of the language is available. Therefore all concepts and relations have been (re-)built manually.
- There is no comprehensive diagram of all super-concepts over all levels available. Thus the relation between the core concepts (e.g. ‘object’) and concepts of the corresponding ‘Business Layer’ (e.g. ‘value’, ‘product’, ‘contract’, ‘meaning’, ‘business object’ and ‘representation’) are not described explicitly. Because of the description a super-sub-concept relation was assumed and modelled accordingly.
- Furthermore, when extending the concepts for example on the ‘Business Layer’ many concepts are clearly not of ‘active’ nor of ‘behaviour structure’ but of ‘passive structure’ and therefore all of them necessarily became sub-concepts of ‘object’ (as it is the only passive core concept).
- Because of the restriction to the domain of “information-intensive organizations, which is the main focus of the language” (The Open Group 2012, p 4), requirements of modeling manufacturing operations cannot be mapped.
  For example: a product is considered to aggregate business services. That assumption holds true for intangible products but not for tangible goods. Therefore, to describe tangible products the model was enhanced. Figure 50 depicts the enhanced model in ArchiMEO.

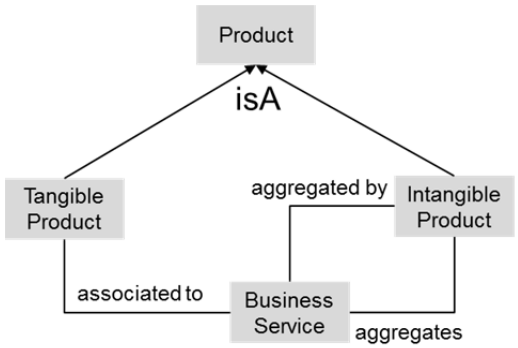


Figure 50: ArchiMate Product Business Service Relation

- On the technical layer a specialization relation is depicted, namely 'System Software' and 'Device' are specializations of 'Node' (Figure 51). As 'System Software' is considered to belong to the 'Business Perspective' and thus it must at least also be a specialization of the generic concept 'Behaviour Element'. However, *that* relationship is not documented in the 'Overview of Relationship' section (The Open Group 2012, p 163) but instead 'System Software' is regarded as specialisation of 'Device' and of itself, which is logically not correct.

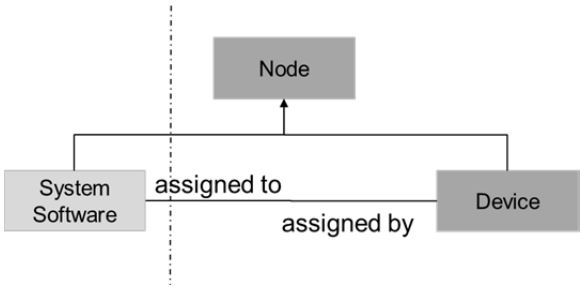


Figure 51: ArchiMate Inconsistent Inheritage

- In the ArchiMate 1.0 specification ArchiMate is considered "a tool for high-level enterprise architecture modeling" and thus, it intentionally does not cover the lower levels of detail of architecture. Instead 'links' to other standards are suggested for describing lower levels of detail (The Open Group 2009a, p 105).
  In the recent specification this suggestion is detailed by providing mechanisms for extending ArchiMate (The Open Group 2012, p 115 ff). However, integrating those other standards like Dublin Core (DC), Business Process Model and Notation (BPMN) but also the Business Motivation Model (BMM) is not a trivial undertaking. Trying to avoid extensions of ArchiMate's core concepts has led to new concepts being described as specializations of existing (core) concepts whenever possible. Figure 52 depicts an excerpt of the Dublin Core Hedgehog Model (Dublin Core Metadata Initiative, 2002) as integrated in ArchiMEO.

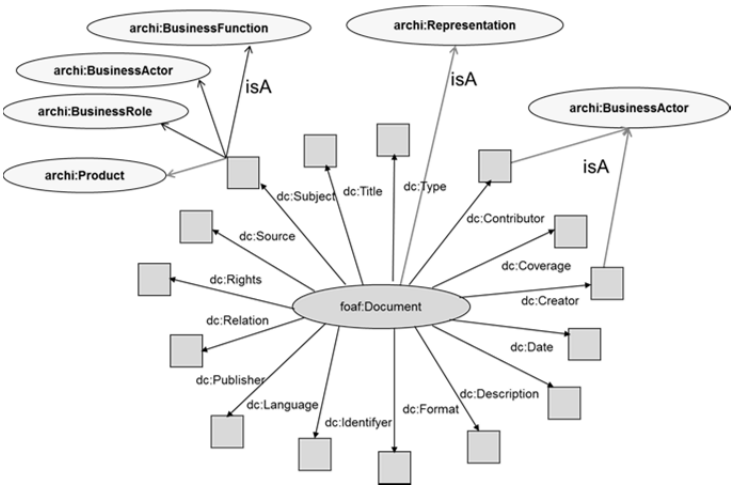


Figure 52: Adapted Hedgehog Model

Whereas ArchiMEO, is meant to be (re)used in other research projects (as for example in the aforementioned APPRIS project), the application specific ontologies are most likely to be substituted or at least supplemented in each project.
Hence, seEAD is considered to be the total concepts and relations needed to semantically describe a particular enterprise and thus, always comprises ArchiMEO plus one or more Application Ontology. At the time of writing seEAD (with ArchiMEO and the Application Ontologies for Symfact and AHSGA) consists of over 7000 RDF triples, i.e. statements about entities in the form of subject-predicate-object expressions, build on about 300 concepts, 270 object properties and 140 data properties.

5.1.4 Architecture Meta Modelling Language

As shown in Chapter 3.4.2, the Semantic Web provides a number of modelling languages that differ in their level of expressivity, but just because one model is more expressive than another doesn't make it more appropriate (Allemang & Hendler 2008). Because there is no 'right' language to formally represent seEAD but "The choice of the language to use in a system or analysis will ultimately depend on what types of facts and conclusions are most important for the application" (Brachmann & Levesque 2004, p 43), I opted for a pragmatic approach. I considered what modelling requirements were to be met, i.e. how detailed constraints between classes, instances and properties should be expressed, how much reasoning is needed and how good the language supported by tools was.
Starting with the most basic requirement, that I wanted to be able to "say anything about anything" (Klyne & Carroll 2002) in a flexible but standardized way led inevitably to RDF. With RDF, which is called data model by Allemang and Hendler (2008, p 75), every relationship between any two data elements can be expressed explicitly, unambiguously and with great flexibility.
Clearly, RDF is not sufficient as it is "dumb data" as Allemang and Hendler (2008, p 79) put it, as only little information about the data model (i.e. the graphs) can be expressed (e.g. rdf:type, rdf:subject). The RDF Schema Language (RDFS) provides a way to *describe* the data, precisely: the data *sets*. This is important as it is the key to inferencing, i.e. from some given data, other related data can be determined as it would have been stated. Table 10 gives examples how seEAD is queried by taking advantage of RDFS language constructs.

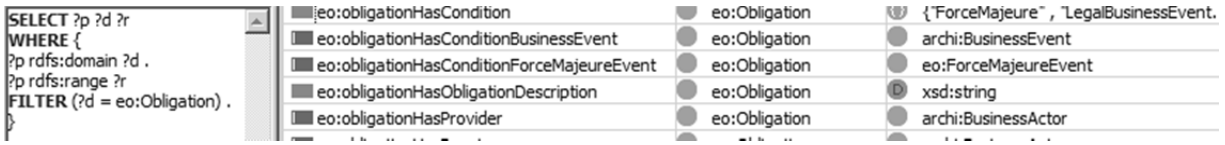
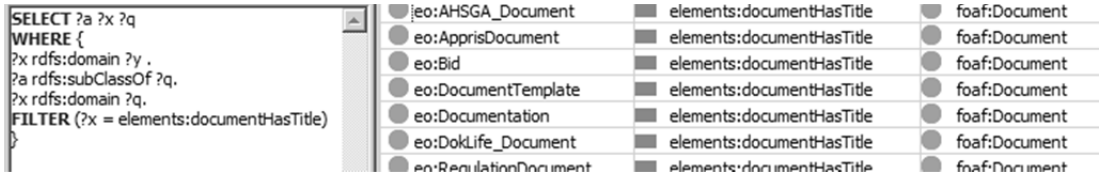
RDFS Inference Patterns (based on Allemang & Hendler 2008, p 93 ff.)		Example
Type Propagation	rdfs:subClassOf	Selects all documents of type subClassOf of `Document`[157]
		
Relationship Propagation	rdfs:subPropertyOf	Selects all documents with subPropertyOf `documentIsAssociatedToContext`
		
	rdfs:domain rdfs:range	Selects range and domain for properties with the domain `Obligation`
		
Combination of Domain (and Range) with Type Propagation	rdfs:subClassOf rdfs:domain rdfs:range	Selects domain of properties that is inherit from the super class `Document;` filter on `documentHasTitle` property
		

Table 10: Language Examples for seEAD

These few patterns RDFS provides can be combined to create more powerful patterns, which make it possible "to simulate certain logical combinations of sets and properties" (Allemang & Hendler 2008, p 122), like Set Intersection and Set Union (two ways of using the rdfs:subClassOf pattern) and Property Intersection and Property Union (two ways of using the rdfs:subProperty pattern).

All patterns RDFS supports are used in seEAD for automatic, format-independent metadata generation and hence, RDFS is necessary. The question is if it is sufficient. Taking a closer look at object properties shows that a more precise description of a relationship would be helpful, for example to express if a property can have a data value as object (`owl:DatatypeProperty`) or a resource (`owl:ObjectProperty`). Although the inverse of a property can be expressed with RDFS, the `owl:inverseOf` property OWL

[157] All print screens showing examples of queries and rules are taken from TopBraid Composer Free Edition

provides, allows for making the specifics of the relation explicit. Same goes with other property types like the `owl:SymmetricProperty` and the `owl:TransitiveProperty`. As seEAD is meant to be (re-)used in further projects intergration of newly created or already existing application ontologies is foreseen. Hence, equivalent classes must be expected. Again a more precise way than RDFS is supported by the OWL constructs `owl:equivalentClass` and `owl:equivalentProperty`. Another two constructs OWL offers are the `owl:FunctionalProperty` and `owl:InverseFunctionalProperty` that allows to infer that two instances are the same if they share the same (`Inverse)FunctionalProperty`.
The aforementioned subset of OWL language constructs is called *RDFS-Plus* by Allemang & Hendler (2008, p 123) because they "see a trend among vendors of Semantic Web tools and Web applications designers for determining a subset of OWL that is at the same time useful and can be implemented quickly". Almost the same subset has been published by members of the W3C working group, which is called RDFS 3.0[158]. The subset can be expressed entirely in RDFS, distinguished by the namespace `owl`. Figure 53 positions RDFS-Plus in relation to OWL.

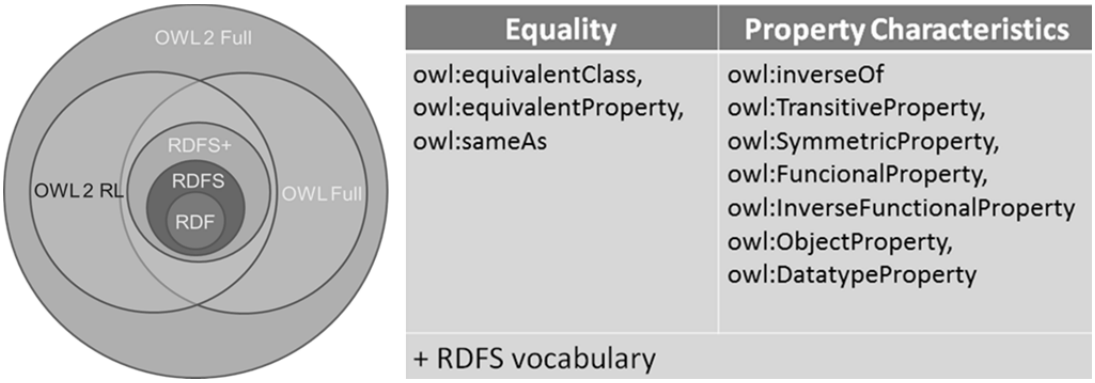


Figure 53: Positioning RDFS-Plus in Relation to OWL (illustrations taken from Bao 2008)

RDFS-Plus can be regarded an answer to OWL, which is a powerful language but hard to apply in real-world semantic technology projects due to its open-world assumption (Knublauch, 2009). Knublauch (2009) refers to the video lecture, taken at 7th International Semantic Web Conference (ISWC), in which Decker et al. (2008) point out that "OWL cannot even be used to check whether an instance of a class meets the cardinality restrictions". Tim Finin[159], a member of the W3C Web Ontology Working Group that standardized the OWL Semantic Web language, argues at the same panel, that OWL is great for some but not for all modelling tasks and that instead the appropriate model language (always) depends on the use case.
As RDFS-Plus' language constructs were sufficient to model seEAD within the two Action Research studies I refrained from investigating the use of more elaborate OWL language constructs. What is more, starting with RDFS-Plus does not hinder the later use of other OWL vocabulary if in other projects more sophisticated expressivity is necessary.

[158] W3C OWL Working Group: RDFS 3.0. URL: http://www.w3.org/2007/OWL/wiki/Fragments (retrieved: 2.8.2012)
[159] Tim Finin is professor of Computer Science and Electrical Engineering at the University of Maryland, Baltimore County (UMBC) and helds a fulltime position at MIT AI Lab. URL: http://www.csee.umbc.edu/~finin/home/ (retrieved: 3.8.2012)

Another advantage of choosing RDFS-Plus as a good tool support is shown in the W3C Wiki that lists a collection of tools for developing Semantic Web applications[160].

5.1.5 seEAD as Part of an Enterprise Repository

"The Semantic Web will not replace the Web as it is known today. Instead, it will be an addition, an upgrade of the existing content in an efficient way that will lead to its integration into a fully exploitable world-wide source of knowledge" (Konstantinou et al., 2006). Similarly I might say that semantic technologies will not replace existing data stores in an enterprise but will lead to its integration into a fully exploitable enterprise-wide source of knowledge. Figure 54 gives an example of enterprise objects that belong to an enterprise repository, structured according to the EAF of Zachman (2008). The cartoon emphasizes the notion of an enterprise repository that comprises all (physical) objects of an enterprise, their various representations and the relations between them. Thus, seEAD is considered a representation of an enterprise architecture but at the same time an enterprise object itself.

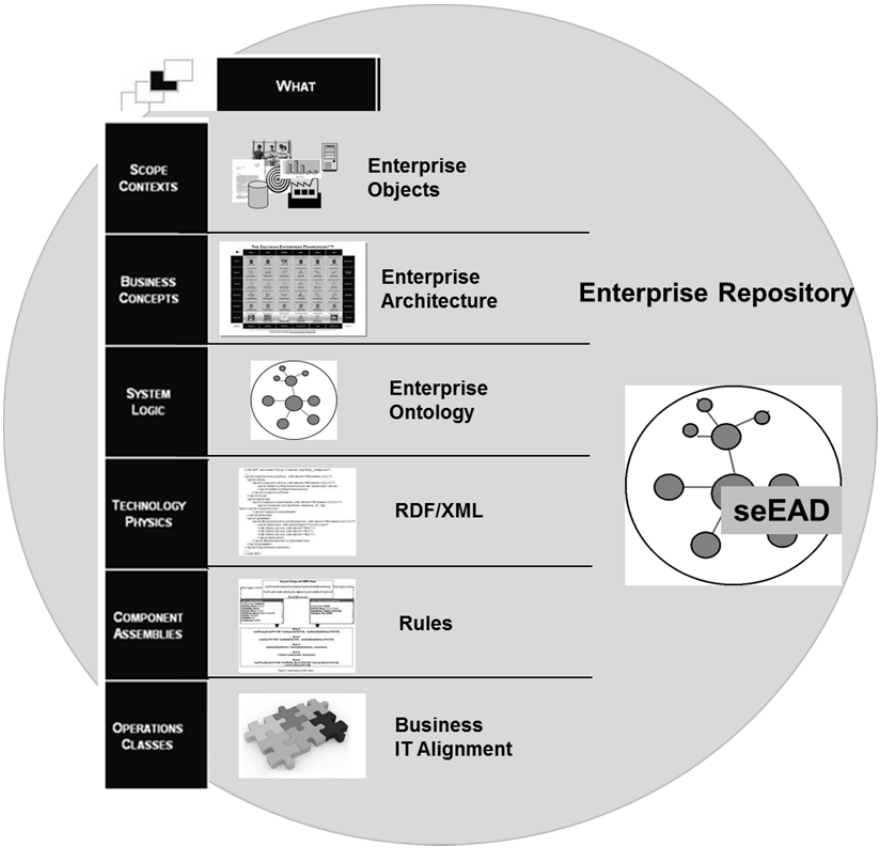


Figure 54: Example of Enterprise Objects Belonging to an Enterprise Repository

However, defining a holistic enterprise repository model is beyond the scope of my thesis. For my work the following enterprise objects are considered: people, documents, infrastructure (file storage) and software, namely AHSGA's Information and Task Recording System (ITRS) and Symfact's Contract Lifecycle Management (CLM) system. In both information management systems data is stored in a relational database (MySQL and MS SQL, respectively). From this it follows that already existing relational data stores are to be related

[160] W3C Wiki: Semantic Web Development Tools. URL: http://www.w3.org/2001/sw/wiki/Tools (retrieved: 3.8.2012)

with seEAD, since parts of the ontology are enterprise independent and thus to be considered already existing. Figure 55 depicts the strategies I have chosen from for my approach.

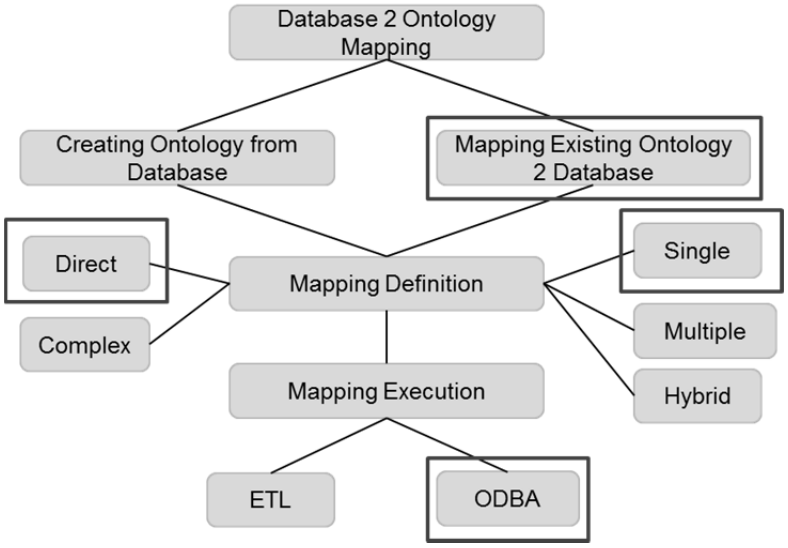


Figure 55: Strategies for Database-to-Ontology Mapping Approaches (based on Ghawi & Cullot 2007 and Spanos et al. 2011)

After deciding on the data sources to be mapped, I considered 'single mapping' (Wache et al. 2001) since one global ontology (seEAD) is used to which all information systems are related. To build an enterprise repository based on seEAD the single approach seems appropriate: each information system is mapped to the enterprise ontology. Notabene, "The global ontology can also be a combination of several specialized ontologies" (Wache et al. 2001, p 109). Moreover, that decision does not hinder a later extension to hybrid mapping. Regarding the type of mapping, in general it differentiates between direct and complex mapping. Direct mapping expresses 'one-to-one' relations between a table and a concept in the ontology, i.e. every record of the table will correspond to an instance of an ontology concept (Barrasa et al., 2004). That notion is based on the definition given by Berners-Lee (1998):

- "a record is an RDF node;
- the field (column) name is RDF propertyType;
- and the record field (table cell) is a value".

Complex mapping, or 'Domain Semantics-driven Mapping Generation' as Sahoo et al. (2009) put it, allows for modeling more sophisticated dependencies as for example if more than one database field forms a property or, if constraints (already existing in the database or not) should be expressed in the ontology. Refer to Barrasa et al. (2004) for an excellent overview of the various mapping possibilities. Even if one can argue that in case of creating an enterprise repository direct mapping is not sufficient, here too, I stick to the 'principle of fitness-for-purpose'. Since for both of my Action Research studies direct mapping is sufficient and ontology-to-database mapping is not the focus of my thesis I decided on this type of mapping.

Finally I chose a method for executing the mapping. 'Extract, Transform, Load' (ETL), a term borrowed from data warehousing techniques (Vassiliadis 2009), indicates that the mapping result is *materialized* (Spanos et al., 2011). This was not necessary for what I intended and thus 'Ontology-Based Database Access' (ODBA) was chosen. Considering the main advantage of an ODBA architecture that allows one to pose semantic queries directly against a database, without the need to replicate its contents in RDF (Spanos et al., 2011).

Figure 56 depicts the level of overlapping of ontological and relational representations of enterprise objects considered in my approach.

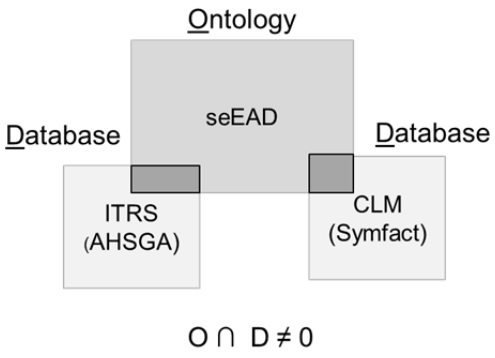


Figure 56: Partial Intersection (based on Barrasa et al. 2004)

The figure illustrates that overlapping between information stored in seEAD and stored in the application systems is kept to a minimum in order to avoid redundancy, as claimed for data and information engineering (Studer, Abecker, & Decker, 1999).

5.1.6 Summary of the Semantically Enriched Enterprise Architecture Description Model

In previous chapters the foundation for automatic format-independent metadata generation was introduced. The mintContext models comprise Enterprise Architecture Meta Model (ArchiMEO), which includes a Top Level Ontology (TOL) and an ontological representation of the ArchiMate standard. ArchiMEO also refines standard and Top Level Ontology (TOL) and refines and enhances it, for example by other standards like Dublin Core. ArchiMEO is considered the Architecture Meta Model which is used as core ontology to build the enterprise specific Architecture Descriptions according to the MOF specification.

RDFS-Plus has been chosen as meta modelling language for ArchiMEO und hence, for seEAD. RDFS-Plus is based on the RDFS standard but has overcome drawbacks by adding some language constructs from OWL. RDFS and OWL are both W3C standards. Finally it was shown how seEAD can be embedded in an enterprise repository and a strategy for database-to-ontology mapping was outlined.

In Chapter 5.1 the foundation, the mintContext, for automatic, format-independet metadata generation was detailed and in the next chapter the generation process will be explained.

5.2 *The Metadata Generation Model*

In this chapter the second foundation of the mintApproach is introduced, namely the mintGeneration model. Since metadata is to be generated for all types of documents (text, image, audio and video), content is considered. Instead, document properties, which are available for all documents build the input for the mintGeneration. These properties can be harvested, unified and processed to build the seeds for automatic metadata generation. As depicted in Figure 57, metadata harvesting can be complemented by format-depending methods like information extraction if requested and applicable.

For the mintGeneration the documents must be represented as enterprise objects in seEAD. Thus, for each document for which metadata is to be generated an instance in seEAD must be created with its properties populated from the seeds. After this metadata can be inferred from the document's context.

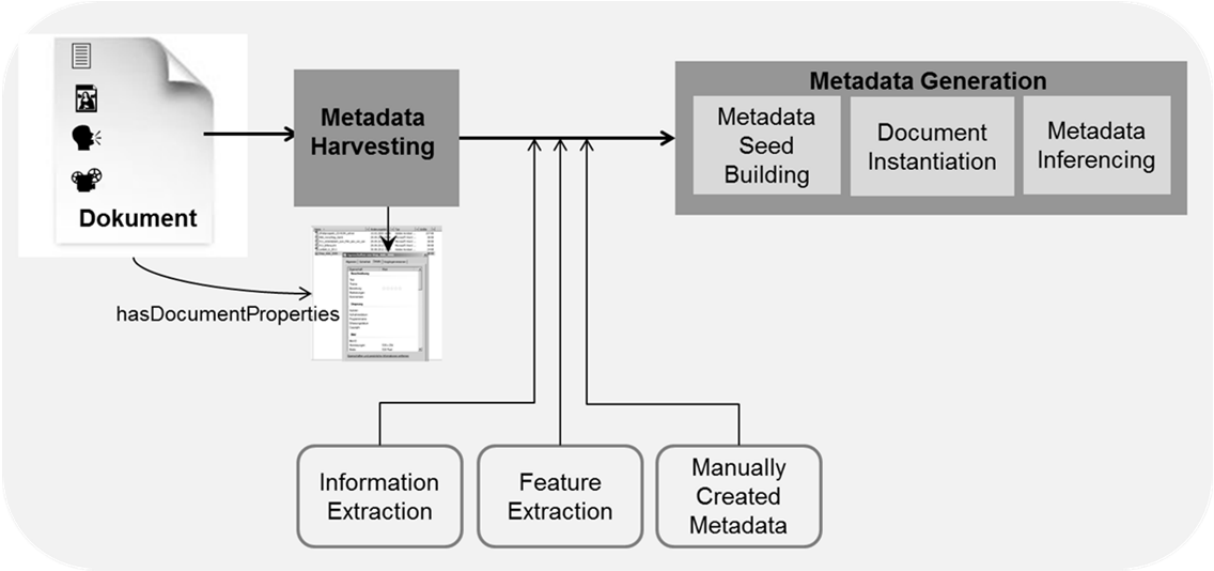


Figure 57: mintGeneration

Since the mintApproach also comprises a procedure model for introducing and customizing automatic metadata generation to enterprise specific needs (cf. Chapter 5.3) in the following diagrams are provided, which can serve as blueprints.
The diagrams also specify details for the prototypical implementation of the mintGeneration as detailed in Chapter 7.

5.2.1 Use Cases for Metadata Generation

Figure 58 provides an overview on the main use cases described in the subsequent sections. All actors refer to software agents. Human interaction for setting up and customizing the metadata generation process are not depicted here.

As metadata generation is regarded as a background process, performance is not an issue. Also not within the scope of my work is the implementation of document management systems but only the aspects of metadata generation. As in general enterprises aim to use already existing systems, document management must be performed in these. In the case of my Action Research studies AHSGA uses their Information and Task Reporting System (ITRS) while Symfact uses their Contract Lifecycle Management (CLM) system. However, these systems are part of the Action Research partner's enterprise repository and some of the automatically generated metadata will be stored in those systems, too. Thus mapping of relational and ontological representation of metadata is also considered.

The following use case specification is represented in UML 2.0 Use Cases and described according to Pohl & Rupp (2009).

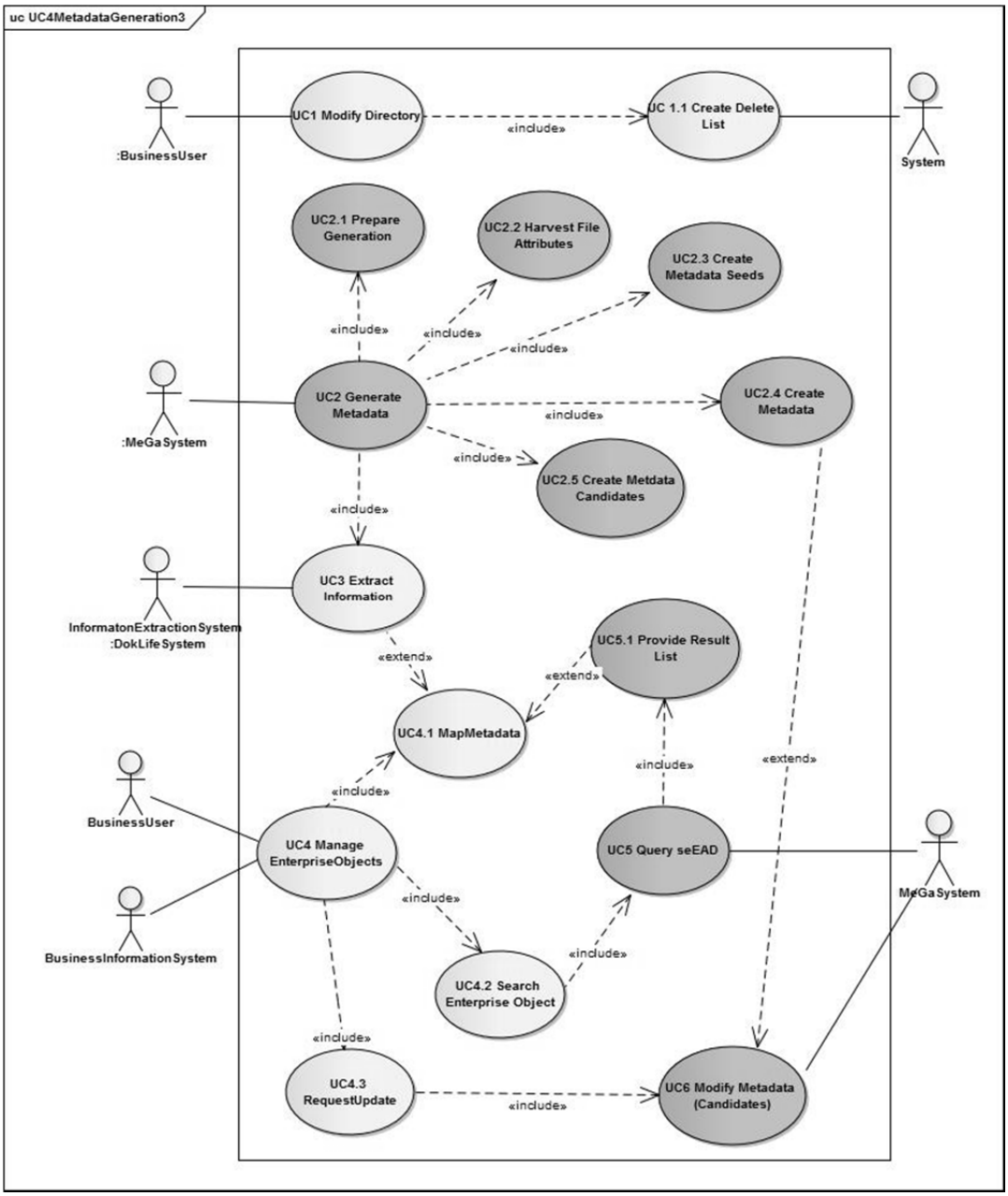


Figure 58: Metadata Generation Use Case Diagram

Figure 58 depicts the use cases relevant for automatic format-independent metadata generation.

- UC1 and UC1.1 are about modifications a user or system makes to the directory on which the harvester will operate.
- UC2 and its inclusions are about generating metadata (candidates or true metadata).
- UC3 is about information extraction.
- UC4 is about managing enterprise objects with respect to map relational and ontological representations of metadata (UC4.1), searching for enterprise objects – either business objects like contracts or their representation like the contract document (UC4.2) and updating metadata (UC4.3).
- UC5 is about transferring the request from a third party system into an executable format to query seEAD and UC5.1 is about transforming the query results into a format a business user can understand and respectively a third party system can import.

- UC6 is about modifying metadata or metadata candidates.

UC2, UC5 and UC6 and their inclusions are implemented in the Metadata Generation Prototype, depicted as grey bubbles and performed by the MeGaSystem actor. Other actors are System, InformationExtractionSystem, BusinessUser and BusinessInformationSystem.

All use cases depicted in Figure 58 are detailed according to the schema, provided in Table 11 and listed in the Appendix 12.4.

Section	Content
Identifier	Unique identifier of the use case.
Name	Unique name of the use case
Description	Brief description of the use case.
Triggering event	Description of the event that triggers the execution of the use case.
Actors	List of actors involved in the use case.
Pre-condition	A list of the necessary prerequisites before starting the use case.
Post-condition	A list of the situation in which the system is in after the execution of the use case.
Result	Description of the result of the use case.
Main scenario	Description of the main scenario of the use case.
Alternative scenarios (optional)	Description of the alternative scenario of the use case including alternative triggering events, pre- and post- conditions.
Exceptional scenarios (optional)	Description of the exceptional scenario of the use case including exceptional triggering events, pre- and post- conditions.

Table 11: Use Case Template

5.2.2 Activity Diagrams for Metadata Generation and Use

For the use cases implemented within the Metadata Generation Prototype (MeGaWorkbench) activity diagrams for metadata generation and use are provided in the following.

5.2.2.1 AD1 Metadata Generation Preparation

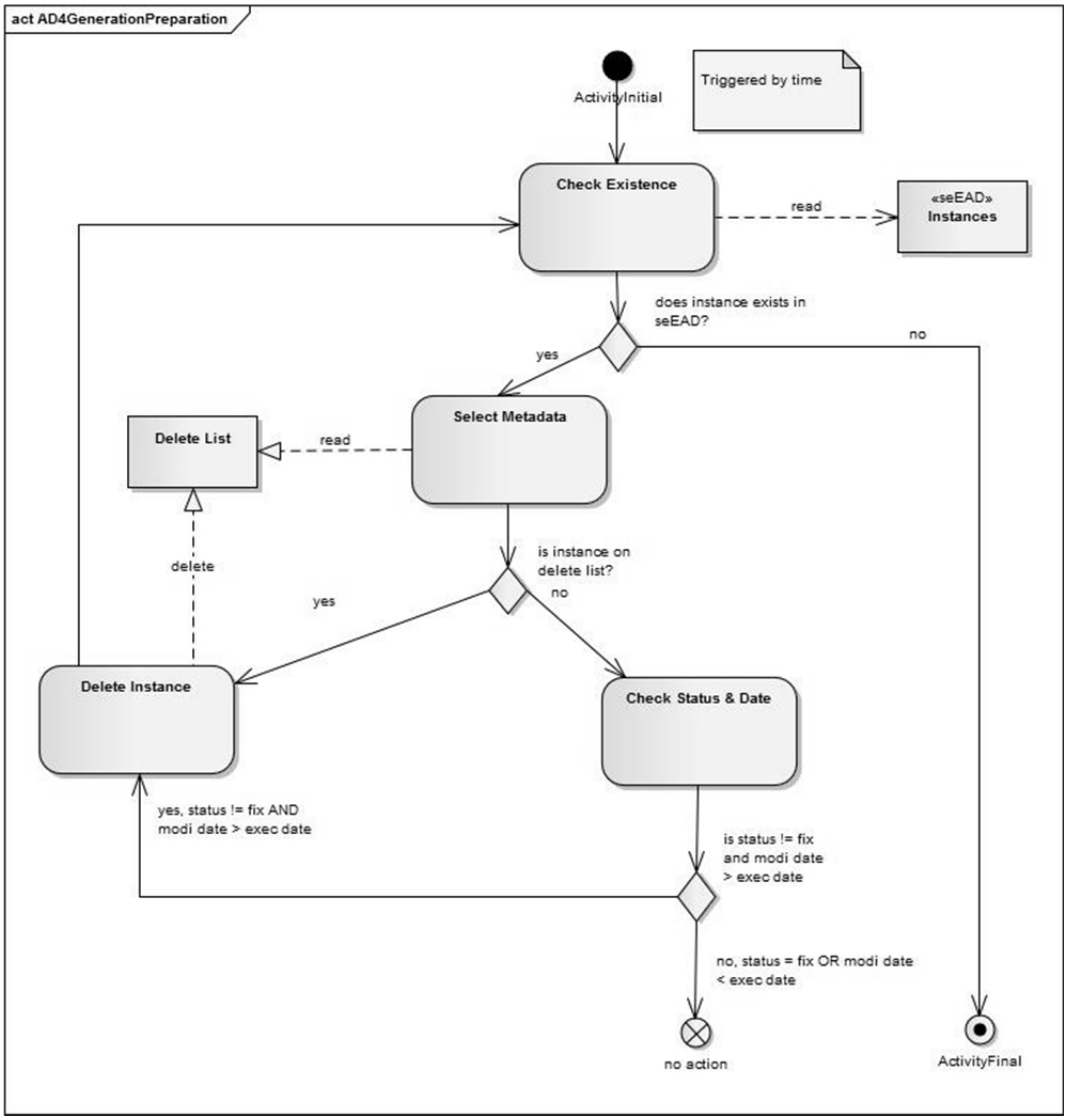


Figure 59: Activity Diagram for Metadata Generation Preparation

The activity diagram for metadata generation preparation, as depicted in Figure 59 depicts the activities that are performed in UC1 plus UC1.1. The first activity (Check Existence) verifies it for documents, which are arranged for metadata generation, already metadata ('true' metadata or candidates) in seEAD exist. If not no action is required. If metadata does exist the status of the metadata is set (e.g. 'fix') and date and time of the metadata generation run is selected. Next it is checked if for that document an entry exists on the delete list. If not it is checked if the document has been updated after the metadata has been generated or, if the status of metadata generation is 'fix'. If document status is fix or timestamp of last modification is older than timestamp of last execution of metadata generation no action is required for that instance. Otherwise the metadata instances in seEAD and the according entry in the delete list are deleted.

5.2.2.2 AD2 Metadata Creation

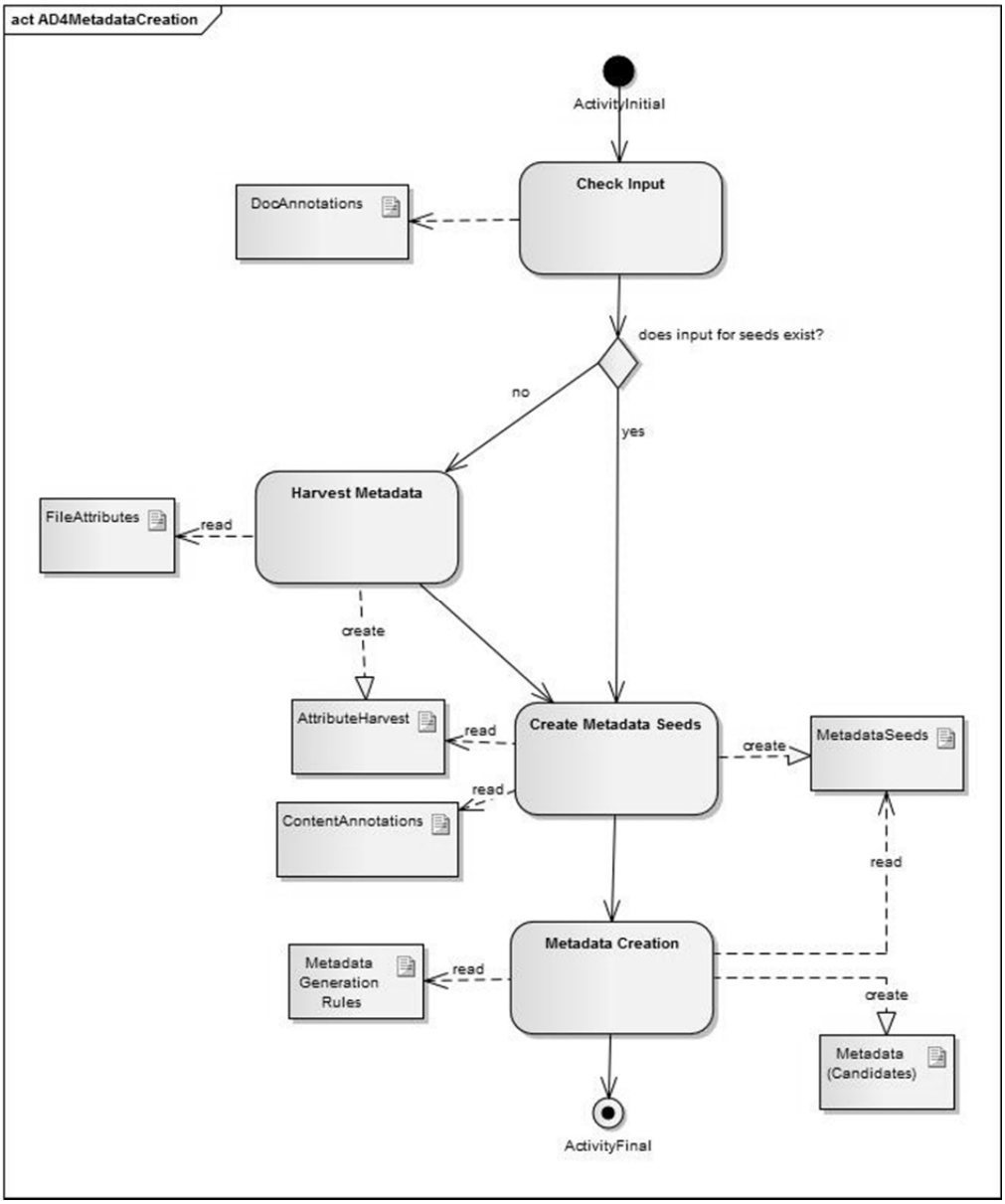


Figure 60: Activity Diagram for Metadata Creation

Figure 60 depicts the four steps of metadata creation. First it is checked, if input for seed creation already exists, i.e. an XML-file with annotations, like the result of an information extraction activity. If not, the harvesting activity is executed that harvests the document properties of documents stored in a repository, i.e. in a specific directory. Result of that activity is an XML-file with the harvested document properties. Next activity is to create metadata seeds from either harvested or extracted or otherwise created input in XML-format.

Creating metadata seeds basically means creating instances of the document concept in seEAD based on the document's document properties or on extracted information if available (refer to Chapter 5.2.3.2 for details). These metadata are called seeds because they build the basis for further metadata creation.

For metadata creation the metadata generation rules are used. Rules define what context, i.e. what enterprise objects, represented in seEAD, are used and how metadata is created. As many of these rules are application specific they refer to the respective chapters for details (Chapter 6.1.3 for AHSGA and Chapter 6.2.4 for Symfact). However, some general rules and an example of data sources and sinks used and created for metadata generation is provided in Chapter 5.2.3.

A distinction between metadata and metadata candidates is made because of the Dublin Core 'one-to-one principle' that requires the creation of one metadata description for one and only one resource (Powell, Nilsson, Maeve, Johnston, & Baker, 2007). Thus, the description of a document is strictly separated from information about the author/creator or – more general – from the context of the resource.

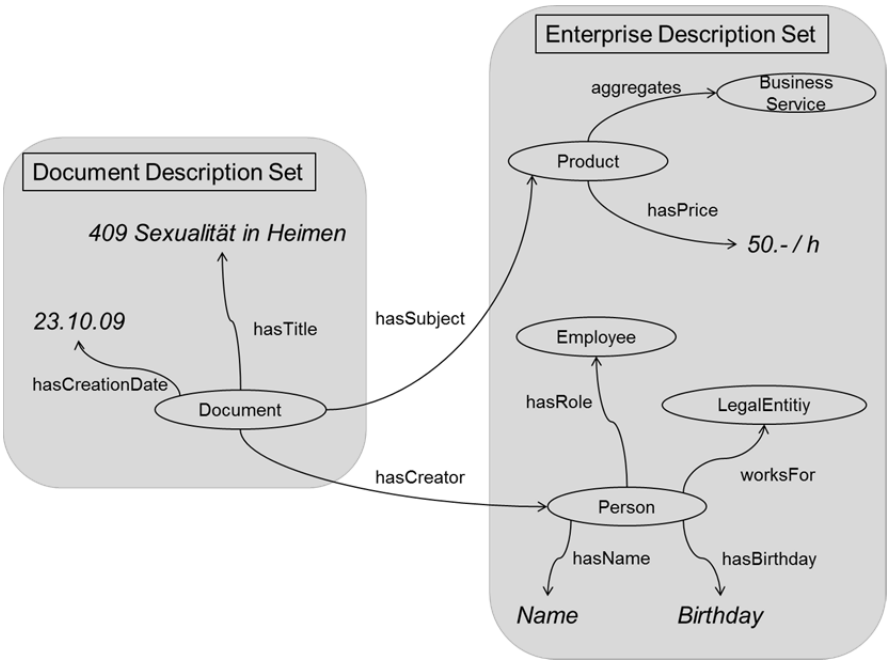


Figure 61: One-to-one Principle (based on Zeng & Qin 2008, p 153)

Figure 61 sketches the one-to-one principle adapted to the mintGeneration: on the left hand side of the figure the description set for the document is depicted, on the right hand side the 'description set' for enterprise objects (i.e. seEAD) is shown.
Therefore, all automatically generated metadata that belong to the document description set are considered 'true' metadata, all others are considered candidates.

5.2.2.3 AD3 Search Enterprise Object

Figure 62 depicts the activities involved in searching seEAD for an enterprise object. That can be for example a business actor (an employee, a contract partner, a client, etc.), a business event (a bankruptcy, a talk, etc.) or representation of a business object (a contract, a presentation, etc.) The request is issued by a third party system and translated into the query language of the ontology.

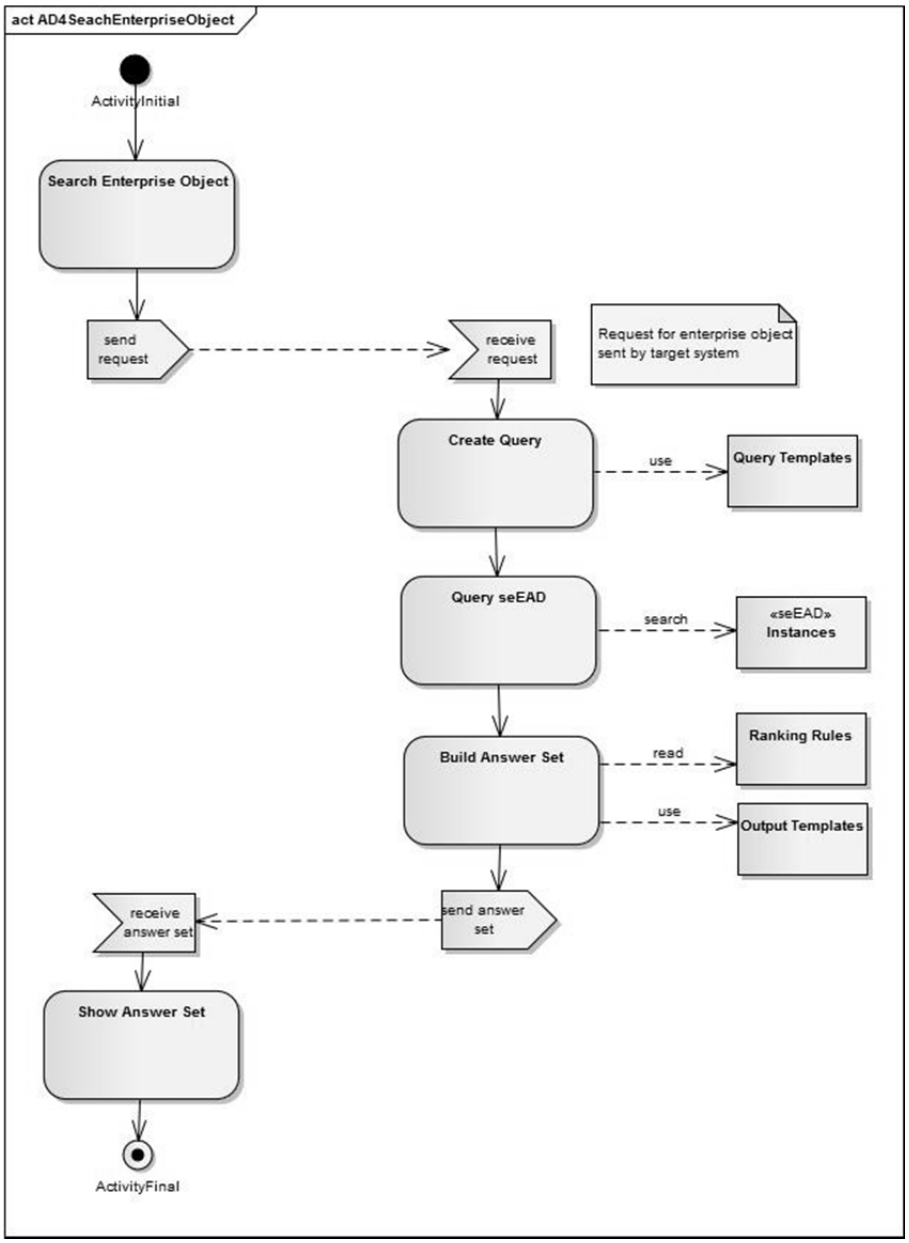


Figure 62: Activity Diagram for Searching seEAD

Ranking of the query results is based on applications specific rules. As RDF triples are not convenient to read for business users, the results are presented also using predefined templates.

5.2.2.4 AD4 Modify Metadata

Figure 63 illustrates the activities for modifying metadata. Again the request for modification is issued by a third party system. The modification can be delete, update or create metadata. As metadata creation is regarded an extension the respective rectangle in the figure is colored in a darker grey.

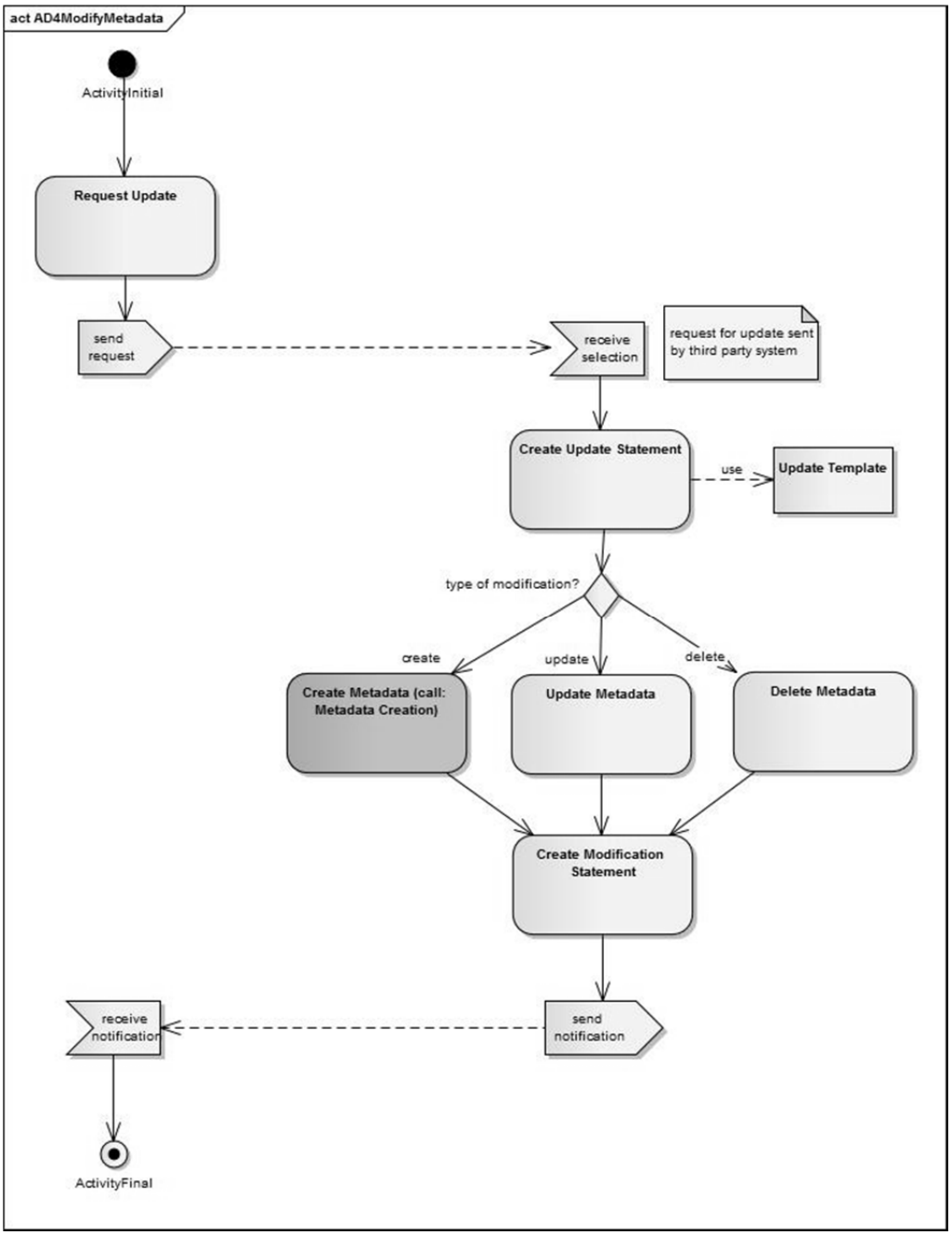


Figure 63: Activity Diagram for Metadata Modification

As for searching seEAD templates are used to transfer the request for modification into executable operations on seEAD.

The provided activity diagrams capture the dynamic behaviour of the automatic metadata generation. In the following sub-sections of Chapter 5.2 data sources and sinks for the mintGeneration are detailed.

5.2.3 Data Sources and Sinks

The artifacts, shown in the activity diagram for metadata creation (cf. Chapter 5.2.2.2) are detailed in the following. Therefor one must distinguish between the general approach and specific characteristics of the enterprise in which automatic metadata generation is processed. Figure 64 depicts the generic sources and sinks of the automatic metadata generation. As already mentioned, metadata harvesting can be performed for all types of documents and hence is shown in the figure. Input provided by other sources like information extraction or manually created metadata is not considered here since its use would be similar. Harvested document properties (or otherwise created metadata) are the source for building metadata seeds; metadata seeds are used to build instances of the documents with the respective properties; from the documents' context metadata is inferred for generating metadata and metadata candidates (cf. Chapter 5.2.2.2 for an explanation of the difference). Metadata (candidates) can be exported in an XML-file to be stored in an enterprise's operative system.

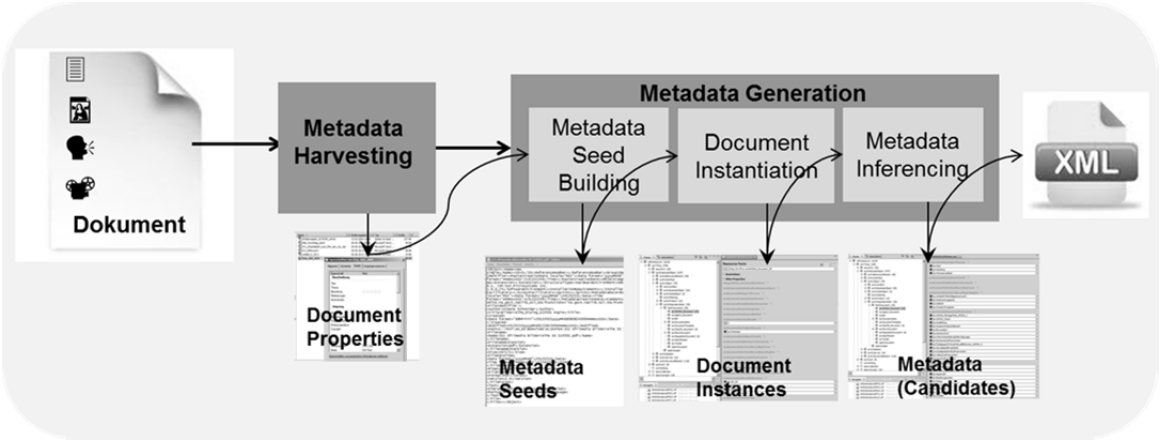


Figure 64: General Data Sources and Sinks for Automatic Metadata Generation

In the following data sources (e.g. document properties) and the corresponding data sinks (e.g. file harvest) are detailed.

5.2.3.1 Document Properties – File Harvest

The ‚FileAttributes' artifact is an input in the activity diagram for metadata seeds creation (cf. Chapter 5.2.2.2).

Table 12 provides an overview on document properties of popular document formats, like DOC or PDF that build the sources for harvesting. For each document format the names of the attributes are provided that are to be harvested.

In the table only those attributes are listed that are in general relevant for business purpose, for example attributes for authorship but not file size (of course, in case of a media concern this property may be important, too, and then is also to be harvested). The far right column provides the harvested attributes (data sink) expressed in DC metadata terms. As shown, some document properties are mapped 1:1 to the corresponding DC metadata term whereas others will be consolidated into one DC metadata, namely 'author' and 'writer' which are labeled uniquely as 'contributor'; and 'topic' and 'keywords' are labeled uniquely as 'subject'. Omitted in the table are the document templates the MS Office tools provide. Hence, attributes for '.doc' are listed but not for '.dot' since the attributes are the same for both.

DOC	PDF	JPEG	MP3 (audio)	MP4 (video)	PNG	GIF	XLS	PPT	All	Dublin Core
name	name	name	name	name	name	name	name	name	(file)name	**alternative**
data type	data type	data type	data type	data type	data type	data type	data type	data type	format	**mime-type**
location	location	location	location	location	location	location	location	location	location	**location**
created	created	created	created	created	created	created	created	created	created	**created**
modified	modified	modified	modified	modified	modified	modified	modified	modified	modified	**modified**
title	title	title	title	title	title	title	title	title	title	**title**
	writer								writer	
author		author					author	author	author	**contributor**
topic	topic	topic					topic	topic	topic	
	keywords								keywords	**subject**
			publisher	publisher					publisher	**publisher**

Table 12: Metadata Harvest

As recognizable in Table 12 with this step also a crosswalk from the proprietary document properties to the standard Dublin Core metadata is achieved.

5.2.3.2 Attribute Harvest – Metadata Seeds

Metadata seeds are instances of the concept `Document` and its properties in seEAD created on the basis of the harvested document properties (AttributeHarvest artifact) or content annotations (ContentAnnotations artifact) or a mix of both.

Figure 65 shows a cartoon of the population of seEAD. The illustration of seEAD entities closely follows the Protégé notation depicting classes with a circle, data properties with a green square (lighter grey in the figure) and object properties with a blue square (darker grey in the figure).

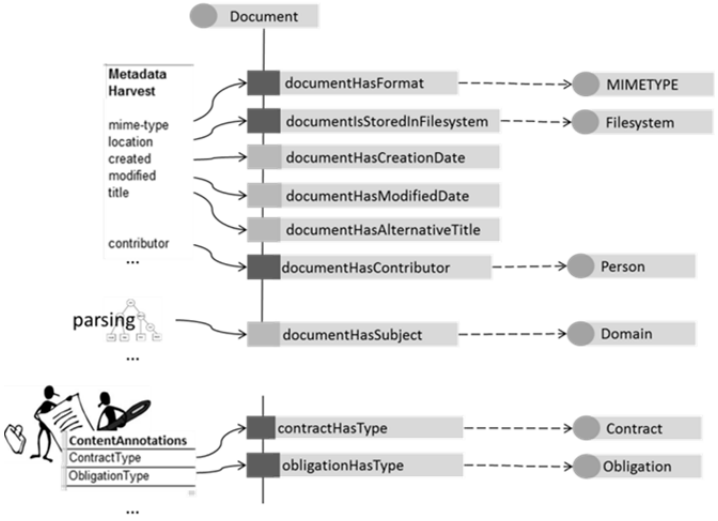


Figure 65: Data Source and Sink for Creating Metadata Seeds

In addition to the population of seEAD document instances with harvested document properties further seeds can be created. In Figure 65 an example is given for a metadata element `documentHasSubject.` In AHSGA's case for example additional metadata seeds are the result of string manipulation like the extraction of nouns from the title that becomes seeds for the `documentHasSubject` property. In Symfact's case content annotations like contract type and obligation type become additional metadata seeds.

Which harvested information can be used in which way for metadata seed creation must be defined when setting up the procedure for metadata generation. Refer to the respective

chapters on examples of specific adaptations for AHSGA (cf. Chapter 4.3.1.1) and Symfact (cf. Chapter4.3.2.1).

5.2.3.3 Metadata Seeds - Metadata

The 'MetdataSeeds' artifact is an input of the 'Metadata Creation' activity in the activity diagram for metadata creation (cf. Chapter 5.2.2.2). Main goal of the activity is to create additional metadata and metadata candidates for a document based on its metadata seeds.

Although the context of a document is enterprise specific the approach for metadata generation can be generalized:

1. All instances of a document's primary context elements are compared with metadata seeds. In case of a match the respective metadata (candidates) are created.

2. For comparison not only string matching is performed but the seed's annotation is considered. For example: if a seed (e.g. creator = 'ABC') matches with more than one instance in seEAD (e.g. 'ABC' is the instance of a `Product` and of a `LegalEntitiy`) then – as general rule – only the business actor would be considered as a match as a product cannot create a document.

3. If all metadata is generated based on primary context elements, secondary context elements are inferred for metadata candidates. Also for metadata candidates general rules for generation can be defined: First instances of all sub-concepts are inferred, e.g. specific youth centers (depth-first-traversal). After that, next neighbors i.e. instances of all object properties of a primary context element are inferred.

4. For all metadata candidates an instance of the inferred context element is created in the following way
   `Document documentIsAssociatedTo2Context` <*`ContextElement`*>.
   The angle brackets indicate that the range of the object property `documentIsAssociatedTo2Context` can be an instance of any context element that has been inferred; the number '2' in front of the property name indicates that a metadata candidate has been inferred from secondary context elements. This might be of use for later document retrieval. The association relation between a document (a specification of `representation`) and enterprise objects is according to the ArchiMate standard.[161]

5. The same approach for metadata candidate creation is used inferring tertiary, etc. context elements, i.e. for all metadata not directly related to a document (cf. Figure 61, p 132for details on the underlying Dublin Core one-to-one principle).

For automatically generating metadata from context the reasoning method of forward chaining is applied, starting from the metadata seeds and ending with the inferred metadata (candidates), stored as new properties of the document instance in seEAD.

[161] Appendix B (Overview of Relationships) of the ArchiMate 1.0 Specification states an association relationship between representation and any other enterprise object. URL: http://www.opengroup.org/archimate/doc/ts_archimate/ (retrieved: 13.10.2011)

After all metadata (candidates) have been generated and stored in seEAD, relational representations can be created and stored in a business information system too, for example in Symfact's Contract Lifecycle Management system or AHSGA's Information and Task Recording System.

5.2.4 Summary of mintGeneration Model

The mintGeneration Model consists of various parts. As a blueprint use cases and activity diagrams have been provided for metadata generation and use in an enterprise.

Three kinds of metadata are generated:

- metadata *seeds*, derived from harvested document properties or otherwise created metadata, represented as enterprise objects in a seEAD
- *true* metadata, inferred from a document's context
- metadata *candidates*, inferred from a document's context but not directly belonging to a document's description set (governed by DC's one-to-one principle).

Generic sources and sinks for the metadata creation activity have been described. Table 13 depicts input and output used and created during the activity.

Input Data	Output Data
document properties	file harvest
file harvest	metadata seeds
metadata seeds	metadata metadata candidates

Table 13: Metadata Input and Output

The reasoning method of forward chaining is applied for automatic metadata generation, starting from the metadata seeds and ending with the inferred metadata (candidates), stored as new properties of the document instance in seEAD and/or exported into a relational representation of an enterprise's business information system.

Also the mintApproach provides re-usable models, like ArchiMEO and the use case and activity diagrams, it becomes clear that the models must be customized to enterprise specific needs. After illustrating the general approach for automatic metadata generation in Chapter 5.2 in the next chapter the procedure for setting up and customizing metadata generation in an enterprise is detailed.

The approach was verified within the two Action Research studies as described in Chapter 6.

5.3 *The Procedure Model*

As shown in the previous chapters, automatic metadata generation is not a 'plug-and-play' approach but the general models introduced above must be customized to enterprise specific needs. Therefore, a procedure model is provided to guide and support the customization of metadata generation in enterprise, depicted in 139Figure 66.

The suggested procedure is based on the approach we provided in Feldkamp et al. (2010). The model has its foundation in procedures for IT project management (amongst many others Kuster et al. 2005). Although the depicted cartoon resembles what is called the 'waterfall

method'[162], in which one phase starts after the previous has ended, this is not strictly followed. Instead, an iterative process is suggested as a beneficial alteration to the waterfall approach Charvat (2003).

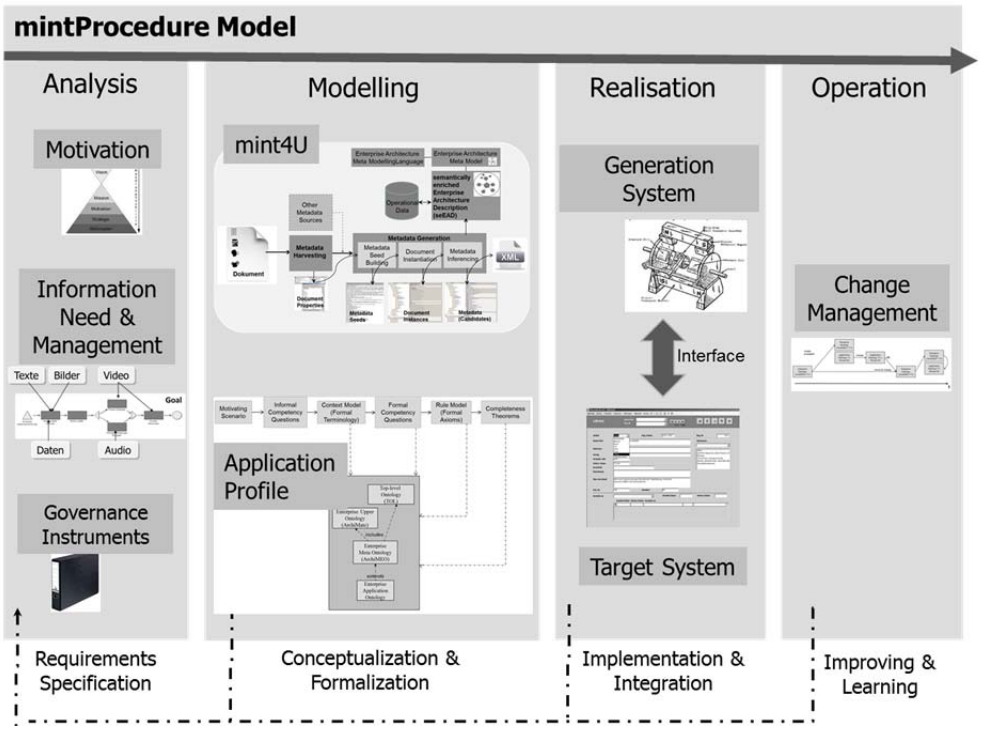


Figure 66: Procedure Model for Metadata Generation (based on Feldkamp et al. 2010)

The procedure model of the mintApproach consists of four phases, (1) analysis, (2) modelling, (3) realisation and (4) operation. In the following the phases are described putting the focus on specifics of the mintApproach. As many project management guidelines and tools for the (modified) Waterfall model are available[163] organisational aspects are not considered here.

5.3.1 Analysis

In preparation of automatic format independent metadata generation, the business need is to be analysed and the enterprise's motivation for automatic metadata generation should be described, e.g. in a motivating scenario as suggested by Uschold & Grüninger (1996).
To capture the motivating scenario no specific method is needed. However, good practise is to create UML diagrams, e.g. Use Case diagrams, for documentation (van Lamsweerde, 2009) in addition to natural language text. The diagrams and paraphrasing tables developed within the mintApproach can serve as templates.
After that the actual information need is for the documents to be identified. To do so (Lagerström, Saat, Franke, Aier, & Ekstedt, 2009) suggest the combination of two approaches: the stakeholder-oriented approach and the causality-oriented approach.

[162] "The waterfall model is a sequential design process, often used in software development processes, in which progress is seen as flowing steadily downwards (like a waterfall) through the phases of Conception, Initiation, Analysis, Design, Construction, Testing, Production/Implementation and Maintenance." Wikipedia. URL: http://en.wikipedia.org/wiki/Waterfall_model (retrieved: 14.9.2011)
[163] In Switzerland for example comprehensive material and tool support is provided by the Swiss Administration for HERMES, a waterfall modell designed for project in the public sector. URL: http://www.hermes.admin.ch/dienstleistungen/hilfsmittel/studbro (retrieved: 5.8.2012)

Lagerström et al. (2009, p 384) understand a stakeholder "as a role within an organization that may benefit from the information provided by the enterprise architecture". The causality-based approach regards specific goals at various levels of concretization (Lagerström et al., 2009). Drawing upon the approach of Lagerström et al. (2009) I suggest basing them on the stakeholder types defined in the ArchiMate standard (The Open Group 2009a, p 70) and deriving the enterprise specific roles. While Lagerström et al. (2009) proposes a method of breaking down high-level goals gradually into fine-grade ones I suggest identifying the business processes that are performed to reach enterprise goals. This is an approach pursued for example in business process-oriented knowledge management initiatives in order to provide "process participants with the information needed to successfully perform their tasks/activities as defined in process models" (Holz, Maus, Bernardi, & Rostanin, 2005).
Figure 67 provides a cartoon to illustrate the method. On the left hand side of the figure a process model is depicted surrounded by information needed to perform process activities and the process as a whole.

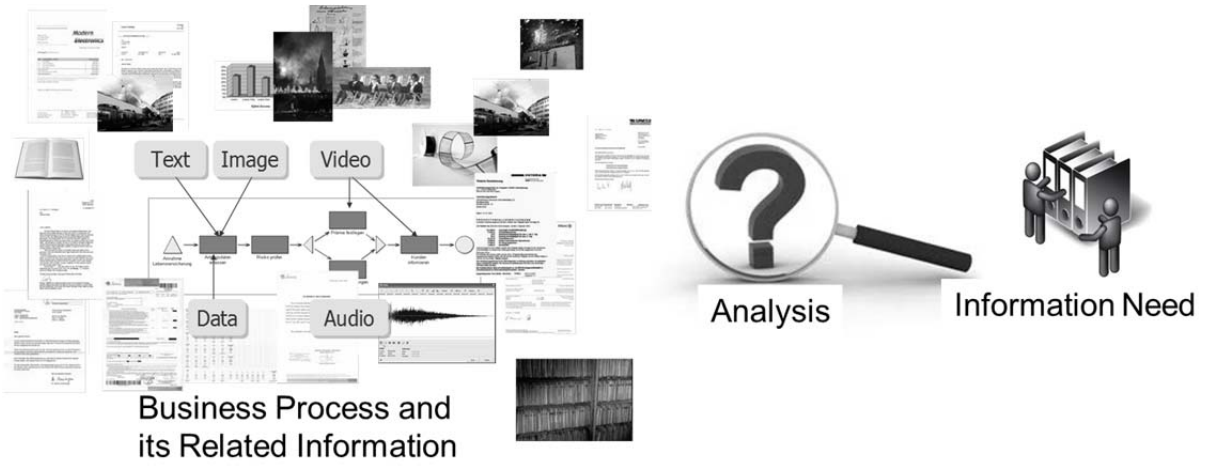


Figure 67: Analysis of the Information Need

Once stakeholder specific information need is determined, analysis of document creation software has to be made to identify documents' type and format. After that the documents' document properties are to be studied to define the base of operations for automatic metadata generation. Then the templates, used for document creation, are to be checked. Templates can provide information about the purpose of the documents created from it. Finally the systems used for business information management, e.g. for document management, must be analysed to identify potential overlapping of information represented in seEAD and in non-ontological representations.

As the corner stone of the mintApproach is a document's context, information governance instruments of an enterprise must be studied in order to customize seEAD to enterprise specific requirements. Governance instruments can range from a simple collection of templates and naming conventions (as in AHSGA's case) to full-blown enterprise architecture descriptions.

The following deliverables should be achieved in this phase:

- description of the motivation scenario
- description of (UML) Use Cases
- determination of stakeholders and their information needs
- determination of documents and document templates, including type and format

- definition of the document properties to be harvested
- determination of the target system (i.e. the system used in the enterprise for business information management) and definition of the metadata to be stored there.

5.3.2 Modelling

Based on the use cases, activity diagrams are created passing over analysis results to design models. For each artefact used in the activity diagrams sources and sinks must be defined. One deliverable of that phase is an enterprise specific description set profile. Again, the general models I've created and introduced in the previous sections may serve as templates.
Depending on the activity diagrams the mintGeneration models are defined, i.e. which components will be used for metadata generation in which way. For AHSGA for example the metadata harvesting component is used, whereas in Symfact's case metadata seeds are built upon extracted information.

Another important step in this phase is the creation of the enterprise specific Architecture Description based on the meta enterprise model ArchiMEO. As emphazised by Stuckenschmidt (2011) and others, re-use of an ontology is a huge asset. That is modelling effort can be reduced, consistency of a domain's concepts can be ensured, etc. One of the problems Stuckenschmidt (2011) identified when reusing an ontology is that most likely it does not meet (fully) requirements of a new application. Since ArchiMEO is an Enterprise Architecture Meta Model providing general, basic concepts and relations based on a standard, it provides a sound basis. Thus, as suggested by Uschold & Gruninger (1996) competency questions can be used to verify the meta model related to a "motivating scenario". Uschold & Gruninger (1996, p 29) claim: "By specifying the relationship between the informal competency questions and the motivating scenario we give an informal justification for the new or extended ontology in terms of these questions". Examples of competency questions are given in the respective sections for AHSGA's and Symfact's motivation scenario (cf. Chapter 6.1 and 6.2). From the informal competency questions the terminology is extracted that will be formally represented in the ontology (A. Gomez-Perez et al., 2004).

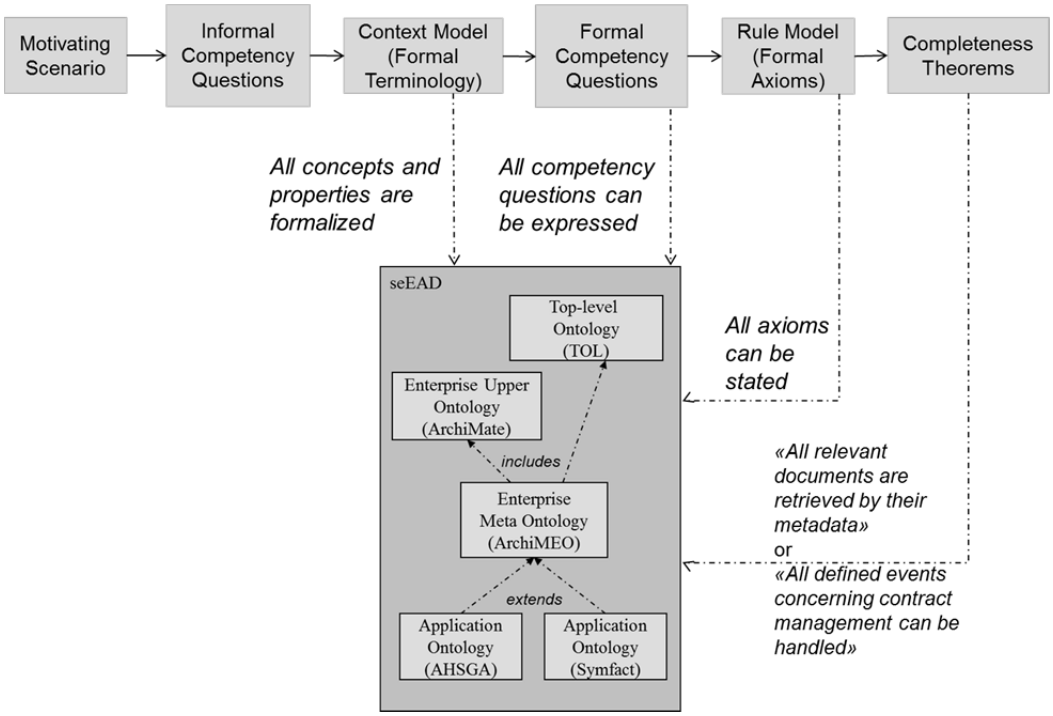


Figure 68: Verification and Enhancement of seEAD (based on Asuncion Gomez-Perez et al. 2004, p 120)

Figure 68 depicts the complete process of using and verifying ArchiMEO and developing application specific enhancements to create seEAD based on Gruninger and Fox's methodology, quoted from Asuncion Gomez-Perez et al. (2004, p 120). After the formal terminology has been defined, ArchiMEO – building the core ontology – is checked. To facilitate check-up it is suggested to represent concepts and properties semi-formally (for example in the ArchiMate notation that is well understandable by business users) and formal (for example in RDFS Plus for machine processability).[164] After that informal competency questions are transformed into formal ones, for example expressed in SPARQL Query Language[165]. The methodology of Fox & Gruninger (1997) proposes axioms "to specify the definitions of terms in the ontology or constraints in their interpretation". In my approach formal axioms are either modelled in seEAD (e.g. subclass-of relations) or expressed as rules (SPIN[166]), applied to seEAD. Finally the conditions under which the solutions are complete must be defined. However, following Asuncion Gomez-Perez et al. (2004, p 123) "a formal formulation for completeness" is not needed in my approach since seEAD is complete if all metadata values can be generated that are needed to satisfy an enterprise's information need. Since information need cannot be generalized I only can state that seEAD is complete if the MeGaSystem generates all metadata values needed to retrieve the relevant documents required by the business, for example by the Action Research Partners.

The following deliverables should be achieved in this phase:

- (UML) activity diagrams
- set of informal competency questions
- definition of formal terminology (context model): domain and enterprise specific enhancements of ArchiMEO to build seEAD (concepts, data and object properties that describe the enterprise architecture of the specific enterprise)
- set of formal competency questions
- rules for enterprise and application specific automatic metadata generation
- description set profile (to describe application specifics)

5.3.3 Realization

Alter (2002) provides four approaches for building information systems, amongst it prototyping. When using a prototype development and implementation are blended. Users can try out the prototype during successive iterations and thus gain experience with the solution and gradually decide how the final system should operate (Alter, 2002). A gap analysis can be performed to determine the implementation adaptations for the final system.
Figure 69 sketches the evaluation process. As within this thesis an evolutionary prototype for automatic metadata generation, called MeGaWorkbench, has been developed (cf. Chapter 7.3), it can be used as Alter (2002) suggests.

[164] An UML plug-in is freely available under the open source Mozilla Public License. The plug-in provides an import and export mechanism between the Protégé knowledge model and the object-oriented modeling language UML. It allows for exchanging ontologies and UML class diagrams, so that Protege can be used in conjunction with software engineering tools (Holger Knublauch) like the Enterprise Architect of SPRARX Systems (wich in turn supports the ArchiMate notation). URL: http://protege.cim3.net/cgi-bin/wiki.pl?UMLBackend (retrieved: 30.10.2011).

[165] SPARQL Query Language is a W3C Recommendation (15 January 2008). URL: http://www.w3.org/TR/rdf-sparql-query/ (retrieved: 30.10.2011)

[166] The SPIN Modeling Vocabulary is W3C Member Submission (22 February 2011) providing a light-weight collection of RDF properties and classes to specify rules and logical constraints as enhancements to SPARQL. URL: http://spinrdf.org/spin.html#spin-constraint-construct (retrieved: 30.6.2012)

The MeGaWorkbench can be modified until an appropriate level of maturity is reached or canceled if modifications run out of proportion of costs and benefits. If the required maturity is reached the prototype has become an application for automatic, format-independent metadata generation and its operation can activated.

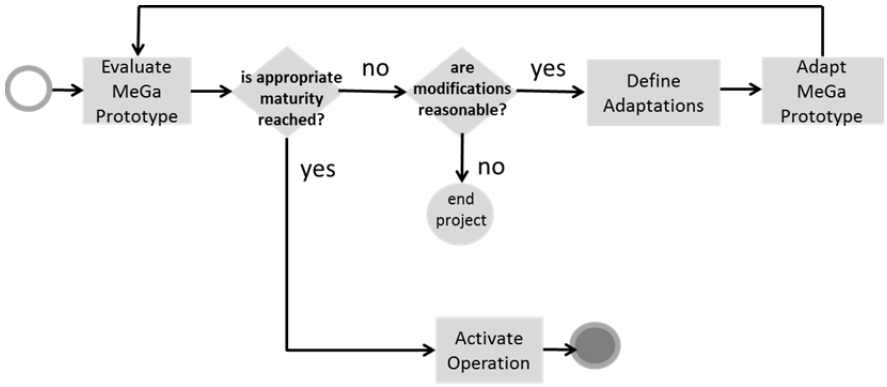


Figure 69: Prototype Approach for Realizing a Metadata Generation System (based on Alter 2002, p 490)

The following deliverables should be achieved in this phase:

- prototype is reviewed and results are documented
- if necessary metadata harvester and/or metadata extractor is adapted to the documents to be analysed
- ArchiMEO is used to build the core of the enterprise specific semantically enhanced enterprise architecture description and one or more application ontologies are created or linked
- rules are modified and enterprise and application specific rules are created
- if applicable ontology and RDBMS(s) entities are mapped
- middleware services are adapted or additional ones are created
- interfaces to enterprise applications (e.g. the ITRS and CLM system) are made ready, and
- functional interaction of all components is tested and approved.

5.3.4 Operation

The operation and maintenance phase starts after the system has been approved by the users, respectively by the commissioner(s) responsible for the development. Regardless of the type of realization this phase comprises two activities: 1) ongoing operation and support and 2) maintenance, including further development of the running system. With respect to these activities the Metadata Generation system is not different from others.

However, as the non-application specific ontologies of seEAD are to be used in other projects and hence are further developed over time the most important question is how to keep the different versions synchonized.

As emphazised by De Leenheer & Mens (2008, p 131) “there is still little understanding of, and support for, the evolutionary aspects of ontologies”. Whereas methods for ontology engineering have been researched for years, maintenance of ontologies, especially in collaborative settings, is much less explored (De Leenheer & Mens, 2008). To solve the problem, once more one can draw on system engineering techniques and tools for versioning, merging and evolving collaboratively developed software artefacts. In their chapter De Leenheer & Mens (2008) provide an overview of models and mechanisms that can be used to support ontology evolution.

Based on the approach introduced by De Leenheer & Mens (2008) I suggest a change procedure for the interorganisational use of seEAD as depicted in Figure 70. Consider an

owner of the Enterprise Architecture Meta Model ArchiMEO (for example FHNW) and distribution of the meta model as open source[167]. Assume each implementation of seEAD starts with a copy of the meta model, in the figure depicted by the square labled 'ArchiMEO V3'. As detailed by Maedche et al. (2003) if reuse is based on a copy, problems arise when the reused ontology changes, as these changes must be reproduced on all copies. Hence, "each ontology should be a closed, consistent, and a self-contained entity, but open to extensions in other ontologies" (Maedche, et al. 2003, p 290).
As ArchiMEO is complemented by application specific ontologies – illustrate as 'Application Ontology' (AO) – to build the enterprise specific seEAD, for each project (A, B, C) two squares are depicted: one for the copy of ArchiMEO (ArchiMEO') and one for the AO. Whereas ArchiMEO is considered 'closed', the AOs are open for changes. seEAD might be adapted to enterprise requirements when set up and by degrees over time (indicated in the figure by sub-version numbers, e.g. V1.1). If an enterprise wants to participate in further development of ArchiMEO they can submit change requests, depicted as 'CR' in Figure 70. "The process starts with capturing changes either from explicit requirements or from the result of change" (Maedche, et al. 2003, p 292). De Leenheer & Mens (2008) emphazises that argumentation and negotiation methodologies, complemented by support for context dependency management are most important for proper support of distributed ontology development and change management.
Figure 70 illustrates the procedure for three projects: Project A starts with a copy of ArchiMEO and creates its Application Ontology dependently but does not contribute to the evolvement of ArchiMEO. Projects B and C link existing AOs and report changes requests for ArchiMEO. Based on the change requests a new version of ArchiMEO is developed (ArchiMEO V4) and made ready for propagation.

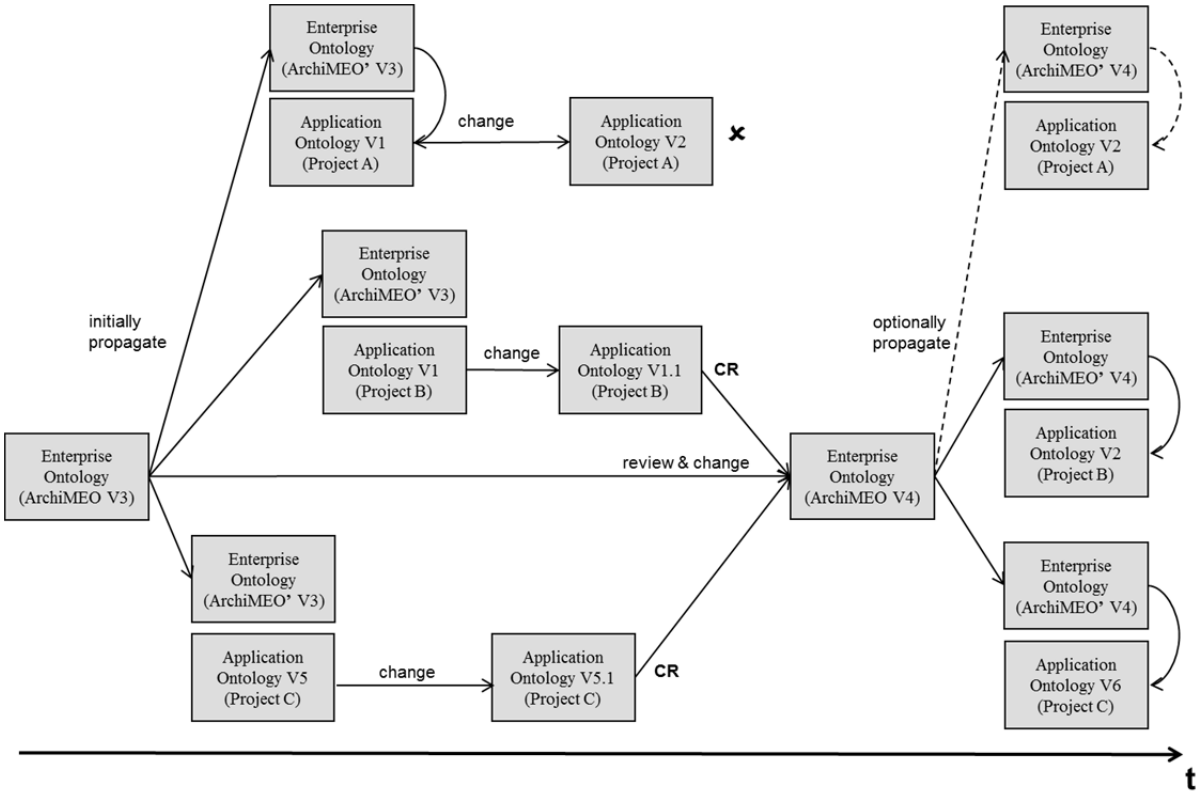


Figure 70: seEAD Change Process (based on De Leenheer & Mens 2008)

[167] Open Source Initiative. GNU General Public License Versions. URL: http://opensource.org/licenses/gpl-license.php/ (retrieved: 10.8.2012)

The approach can be considered a combination of the approaches suggested by (Maedche, et al. 2003) and De Leenheer & Mens (2008), as the copies of ArchiMEO remain unchanged but collaborative development is still performed based on the change requests resulting from applications specific adaptions. However, due to the complexity of the subject it could only be touched briefly here but as ArchiMEO is already used in various projects at FHNW the aforementioned procedure is set up and currently beeing tested.

The following deliverables should be achieved in this phase:
- a decision on participation in collaborative development of ArchiMEO has been taken
- if collaboration is wanted change process has been negotiated and set up
- changes are reported as defined.

5.3.5 Summary of the mintProcedure Model

The mintProcedure model for automatic, format-independent metadata generation has its foundation in procedures for IT project management. For the mintApproach a modified waterfall method is chosen that does not strictly require completing a development phase before starting with the next but allows for an iterative process. In this sense I recommend four phases (analysis, modelling, realization and operation), focussing in each on mintApproach specific aspects. The first three phases of the procedure model were applied to my Action Research studies.

5.4 *mintApproach Findings I*

For automatic metadata generation three kinds of models are needed:
- the mintContext Model to determine the enterprise objects related to a document
- the mintGeneration Model to specify the dynamic behaviour of the automatic metadata generation, its sources and sinks, and
- the mintProcedure Model to identify enterprise specific customization needs.

A document used in an enterprise is not created as an end in itself but contributes to reach a business goal. Thus, it is related to other enterprise objects like the employee who created the document, the task the document is used in, the template from which the document is created from, the product, the document describes and so on. How enterprise objects are related to each other is partially general, e.g. that a document is used in a task and that an employee is responsible for a task. That way, for a person who created a document, her responsibilities can be identified and from this the tasks can be inferred the document might be used in. Exploiting this knowledge for generating metadata for business documents automatically is the goal of the mintApproach.

If the context of a document shall be inferred automatically it must be represented in a machine processable way. There is broad consensus that using ontologies is an appropriate way to represent enterprise architecture knowledge. In addition enterprise architecture modelling (and description) and ontology modelling recently started to be merged. However, it turned out that neither existing Architecture Description Languages – like ArchiMate – nor existing enterprise ontologies – like TOVE – meet all requirements for modelling the mintContext. Thus, a semantically enriched Enterprise Architecture Description (seEAD) has been developed that represents the context of enterprise documents ontologically and allows for inferring metadata automatically and un-supervised.

To decrease ontology development effort and better exploit the potential of a semantically enriched enterprise architecture description I created a Enterprise Architecture Meta Model (ArchiMEO) that can be used to model the enterprise specific semantically enriched Enterprise Architecture according to the Meta Object Facility specification (OMG, 2011b). Thus, it can serve as 'General Enterprise Model' or 'Core Enterprise Ontology' as suggested by Fox & Gruninger (1998) and Bertolazzi et al. (2001), respectively. For a sound foundation ArchiMEO is based on the ArchiMate standard (The Open Group, 2009b), enhanced by other standards, e.g. Dublin Core.

Analysis of research on representation languages for ontologies has shown that there is no 'silver bullet for formalization'. Even if a computational level of formalization is given, there are several modelling languages – or dialects – that could be selected. For automatic metadata generation based on context I relied on W3C standards for ontology representation (RDFS & OWL), on the W3C recommendation for ontology managing (SPARQL), and on the W3C submission for rule formalization (SPIN). For the mintApproach I consider RDFS-Plus appropriate.

Since seEAD is regarded part of an enterprise *repository*, comprising all entities constituting an organisation, it has been analysed to show how existing datastores can be linked with seEAD. As strategies for database to ontology mapping 'direct and single mapping' were suggested for linking seEAD to non-ontological datastores and ODBA (ontology-based database access) for querying.

Although the mintContext and mintGeneration Model provide a sound basis for automatic metadata generation it became apparent that they are to be customized to enterprise specific needs. While in general, introduction of the mintApproach in an enterprise can be regarded the same as for any other business information system, it is specific with respect to context modelling. In the mintProcedure modelling the context is not simply defining APIs to other information systems but to expressing the centrepiece of an enterprise explicitely and representing it formally in a machine understandable way. The more precisely an enterprise's architecture is already described, the easier it can be transformed into the mintContext model and the better is the quality of the automatically generated metadata. If for example business processes are already comprehensively described, including resources, roles and organisational models, rich context information is available for documents. If on the other hand side only low-level governance instruments are used in an enterprise little context information can be inferred. However, as detailed in Chapter 6 also in this case automatic metadata generation based on context is possible. Proof of concept is given in Chapter 7.3.

6 Application Profiles

Chapter 6 of my thesis provides the enterprise and application dependent, conceptual models, derived from the requirements presented in Chapter 3 and Chapter 4, as illustrated in Figure 71.

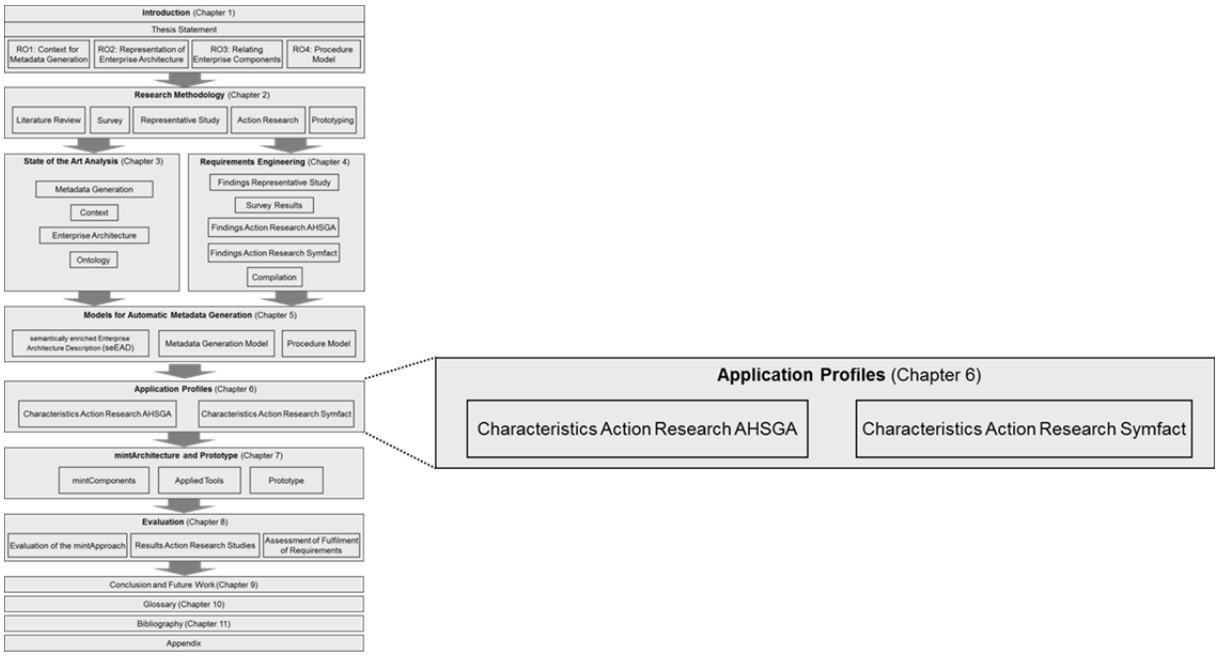


Figure 71: Position of Chapter 6 in the Overall Structure of the Thesis

Whereas Chapter 5 presents the general models for automatic, format-independent metadata generation, Chapter 6 introduces characteristics of the Action Research partners' implementations. Since Action Research is research in action, the general models are applied to enterprise specific requirements involving practitioners within Loop 2 of my Action Research study. Purpose of this loop is to get in-depth knowledge of the business' peoples' opinion on the models and to reassess theory in practice (cf. Chapter 2.2.3).

In this chapter the Action Research partners' specific models are described based on the guidelines Dublin Core provides (Coyle & Baker, 2009). The Dublin Core Application Profile specifies and describes the metadata used in a particular application. To accomplish this, a profile each for AHSGA and Symfact is created. Figure 72 depicts the mintApproach with Action Research partner's characteristics. Besides the aforementioned enterprise specific creation of the seEAD and customization of metadata generation procedure, rules are added to the figure. The enterprise specific rules define the context of the respective enterprise and what and how enterprise objects related to a document are inferred for automatic metadata generation.

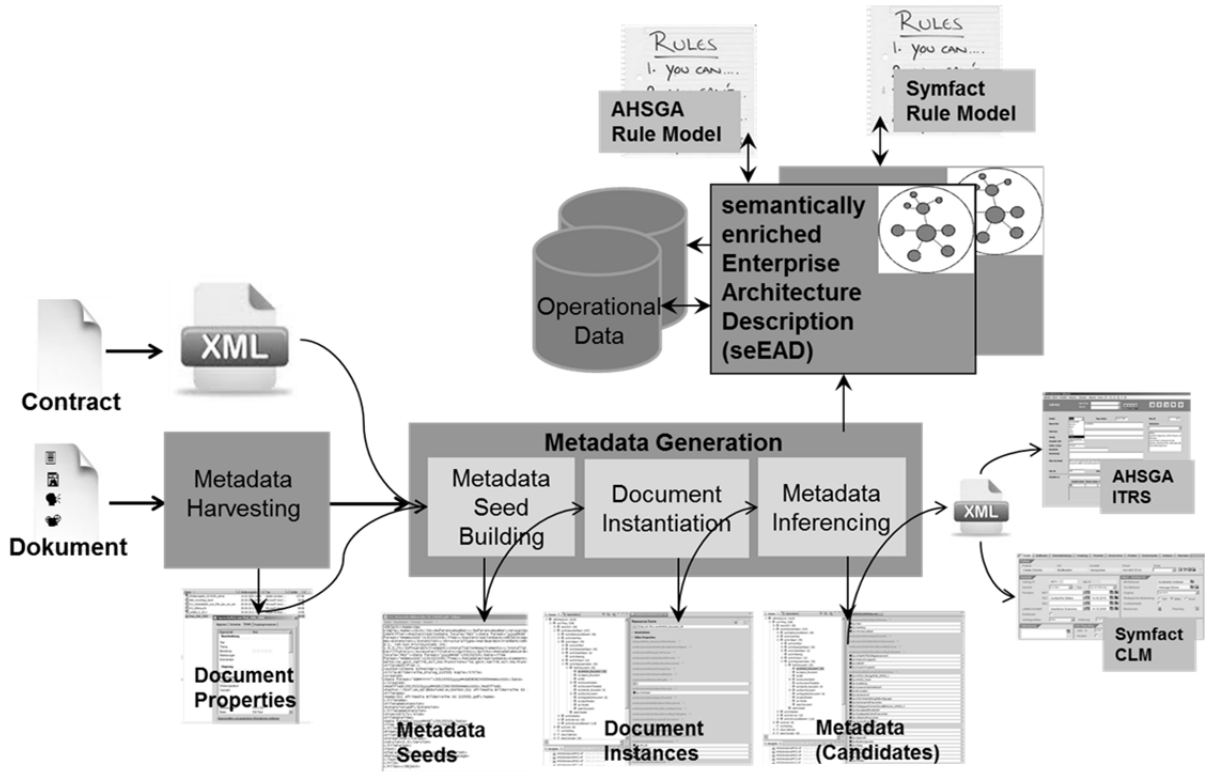


Figure 72: Customized mintApproach for the Action Research Partners

The mintProcedure Model introduced in Chapter 5.3 was applied to each research partner and thus, each profile

- briefly recapitulates what the research partner wants to accomplish with automatic metadata generation; for this the 'motivation scenario' (Uschold & Gruninger 1996) is sketched and informal competency questions are derived
- characterizes the things the metadata describes and its relationships
- provides the formal representations of the competency questions and
- defines the formal axioms for automatic metadata generation, i.e. the inferencing rules.

To address the completeness theorem (cf. Chapter 5.3.2), i.e. defining when the partner's needs are met, each profile

- enumerates the metadata elements, defines their properties and shows the formal representation in seEAD (Description Set Profile)
- defines the enterprise specific sources and sinks for automatic metadata generation
- sketches the use of the generated metadata.

Each profile ends with results of the second loop of the Action Research study with the partner.

The chapter is structured accordingly and ends with a comparison of both applications.

6.1 *AHSGA Application Profile*

In order to answer the two questions of Dietz (2006, p 6) - "Why and how would enterprise ontology assist in coping with current and future problems related to enterprises?" and "Why would this approach be more appropriate and more effective than some other one?" I will elaborate further questions.

Following Uschold & Gruninger (1996) competency questions are used to evaluate ArchiMEO related to AHSGA's motivating scenario in order to create the enterprise and application specific seEAD. As detailed in in Action Research Loop1 (cf. Chapter 4.3.1) AHSGA wants to increase employee productivity by decreasing time for searching documents. As manual metadata creation is labour intensive and error prone and information retrieval techniques are not applicable for many types of documents (e.g. images or audio files) a solution is desired that allows the generating of metadata automatically for all kinds of documents used in the enterprise. Furthermore, documents should be handled as usual (i.e. simply stored on a file server) and no additional tool for document management should be implemented. Instead, the existing Information- and Time Recording system (ITRS) will be used for document retrieval. Therefore generated metadata elements are used to identify the documents related to a reported task and the user then can select the relevant one(s) from a hit list (refer to Chapter 6.1.6 for details).

6.1.1 AHSGA Informal Competency Questions

In the following, informal competency questions are phrased to determine whether the proposed seEAD is required and adequate. The questions have been drawn from knowledge I gained from my work with the Action Research Partners and was reviewed by them within Action Research Loop 2.

1) Given a document stored in a directory at the enterprise's file server (AH_GL, AH_GS, AHS, BAG, etc.)
   AND
   some constraints (regarding authorship or directory structure)
   which secondary context[168] elements can be determined?

2) Given the primary context elements of a document (business actor, business service, product, etc.)
   AND
   some constraints (regarding some business functions or business events)
   which secondary context elements can be determined?

3) Given any context elements of a document (primary, secondary, etc.)
   AND
   some constraints (regarding authorship or directory structure)
   which metadata candidates can be determined?

4) Given the business object a documents represents (project, advice, invoice, offer, etc.)
   AND
   some constraints (regarding some business functions or business services)
   which documents can be determined?

5) Given the business behavior element related to a business object (business event, business function)
   AND
   some constraints (regarding time or location)
   which documents can be identified?

[168] Required by Dublin Core's one-to-one principle and according to Dey & Abowd (1999) I differenciate between various levels of context (cf. Chapter 4.3.1)

6) Given the contributor of a document (a person, an organizational unit, a legal entity)
   AND
   some constraints (regarding business objects or products)
   which documents can be identified?

To stratify the set of competency questions I refer to (Asuncion Gomez-Perez, Fernandez-Lopez, et al. 2004, p 121 f.), claiming that an ontology is well-designed if "competency questions can be split off into more specific (or atomic) competency questions, and the answer to a question can be used to answer more complex questions".

To illustrate the procedure two of the above phrased competency questions are elaborated as follows:

1)
a) Given a document stored in a directory at the enterprise's file server (AH_GL, AH_GS, AHS, BAG, etc.)
   AND
   some constraints for authorship (has role manager, expert, etc.)
   which primary context elements can be determined?

b) Given a document stored in a directory at the enterprise's file server (AH_GL, AH_GS, AHS, BAG, etc.)
   AND
   some constraints for the directory structure (related to business actor, business service, etc.)
   which primary context elements can be determined?

2)
a) Given the primary context elements of a document
   AND
   some constraints regarding authorship
   which metadata candidates can be determined?
b) Given the primary context elements of a document
   AND
   some constraints regarding directory structure (AH_GL, AH_GS, AHS, BAG, etc.)
   which metadata candidates can be determined?

After splitting off the competency questions into more specific questions, it is used as a basis to derive the inherent assumptions, constraints and the necessary data (A. Gomez-Perez et al., 2004). Figure 73 gives an example of the approach for competency question 1b.

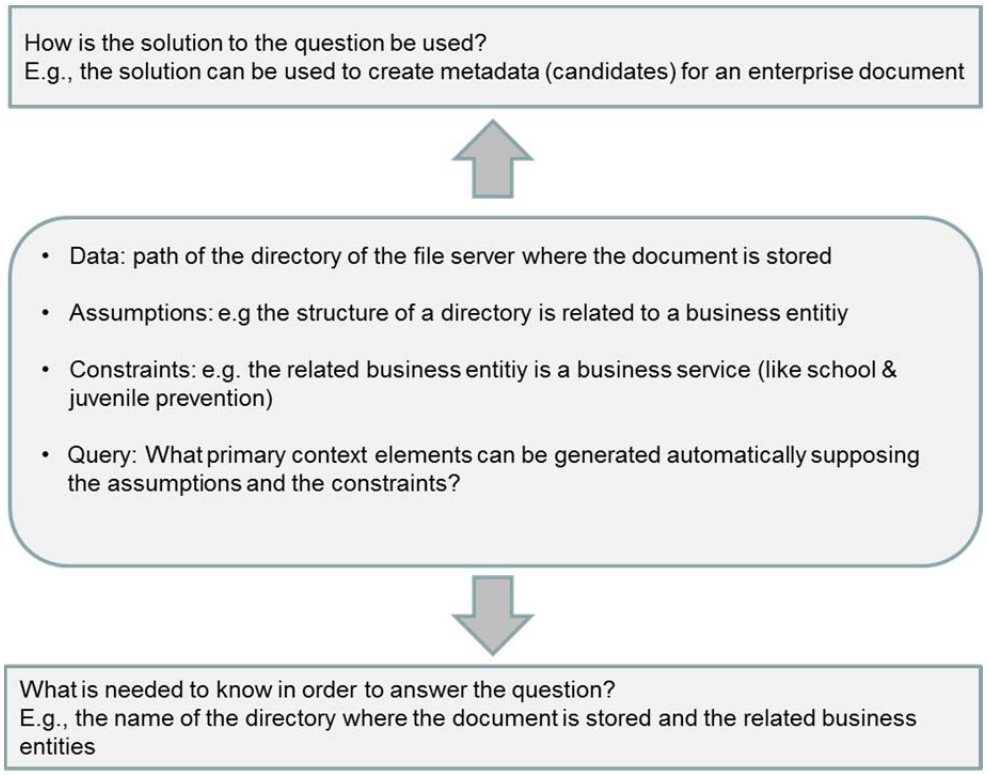


Figure 73: Decomposition of the Competency Question 1b (based on Gomez-Perez et al. 2003)

First is determined how the answer to the question will be used, here the answer is for the creation of metadata (candidates) of a document. After that the inherit data is derived, the underlying assumption is stated (here that the storage location of a document is related to a business entity) and constraints are defined (here that the related business entity is a business service). Finally the competency question is generalized to a query supposing the assumptions and the constraints as shown in Figure 73.

Today, none of the previously elaborated competency questions can be answered with AHSGA's system of document management or existing methods of document indexing. For providing answers the context of the documents must be known and formally represented in an enterprise architecture description as implemented in seEAD.

After working on competency questions Gómez-Pérez et al. (2005) suggest specifying the terminology. That is done in the following way for the metadata that is to be generated for AHSGA and the data recorded in their ITRS.

6.1.2 AHSGA Context Model

As the AHSGA Context Model is based on ArchiMEO, and thus on ArchiMate, definitions of the formal terminology haven't been made from scratch. Instead existing concepts were reviewed to see if they meet AHSGA's requirements (cf. Chapter 4.3.1) and enhanced where necessary. Thus, for example, a `Document` in seEAD is considered a specialization of a `Representation`, which realizes a `BusinessObject`. Table 14 gives an example of objects, which are instances, of AHSGA's domain and the corresponding concepts as exists in seEAD. All references to ArchiMate in this chapter are based on the ArchiMate 1.0 Specification (The Open Group, 2009b).

Instances	Concepts	Relations to ArchiMate concepts
AHSGA_Document1	`Document`	`Document` isA sub-concept of `Representation`
AH_GL, AH_GS, AHS, BAG, etc.	`Directory`	`Directory` isA sub-concept of `Node`
C:\TEXT\AH_GS\DIREKTEP\MSM etc.	`Filesystem`	`Filesystem` isA sub-concept of `Node`
Simone Schneider, Johannes Schlaepfer, etc.	`Person`	`Person` isA sub-concept of `BusinessActor`
Manager, Expert, Padagogue in human sexual behavior, etc.	`Employee`	`Employee` isA sub-concept of `BusinessRole`
General Information and Advise, Expert Services and Information, Public Relations, etc.	`BusinessService`	already exists

Table 14: Instances Derived From AHSGA's Competency Question 1

After all objects were derived from the competency questions and checked how they can be represented in seEAD their properties were determined. Table 15 gives an example of some object properties derived from AHSGA's Competency Question 1 (the list is not conclusive).

Instances	Property	Remark
AHSGA_Document1	`documentHasCreator`	property is an enhancement to ArchiMate for Dublin Core Elements
	`documentHasDateCreated`	property is a refinement of the Dublin Core Element 'date'
	`documentHasSeedEmpNo`	property is a AHSGA-specific, non-Dublin Core Element
AH_GL, AH_GS, AHS, BAG, etc.	`directoryHasStructureElement`	the property is a refinement of the ArchiMate relation 'Node Association EnterpriseObject[169],
	`rdf:label`	e.g. `AHSGA_Geschäftsstelle {@de}, AHSGA_ProfServices {@en}`
	`enterpriseObjectHasName`	e.g. AH GS
C:\TEXT\AH_GS\DIREKTEP\MSM etc.	`filesystemIsStructuredInDirectory`	the property is a refinement of the ArchiMate relation 'Node Association Node'
Simone Schneider, Johannes Schlaepfer, etc.	`personHasFamilyName`	the property is a foaf property[170]
	`personHasNameInitials`	property is AHSGA-specific
	`personWorksForLegalEntity`	the property is a refinement of the ArchiMate relation 'Business Actor Association Business Actor'

[169] EnterpriseObject as super-concept in seEAD stands for all concepts defined in ArchiMate as explained in Chapter 5.1.3 (Content of ArchiMEO as Meta Model for seEAD)
[170] FOAF Vocabulary Specification 0.98. URL: http://xmlns.com/foaf/spec/#term_Person (retrieved: 25.7.2012)

Instances	Property	Remark
Manager, Expert, Padagogue in human sexual behavior, etc.	`employeeHasEmployeNo`	property is AHSGA-specific
	`rdf:label`	e.g. `PadagogueInHumanSexualBehaviour {@en}, Sexualpaedagoge {@de}`
	`employeeIsResponsibleForProduct`	the property is a refinement of the ArchiMate relation 'Business Actor Association Product'
General Information And Advise, Expert Services and Information, Public Relations, etc.	`businessSerivceIsAssociatedToBusinessFunction`	the property conforms to the ArchiMate relation 'Business Service Association Business Function'

Table 15: Object Properties Derived From AHSGA's Competency Question 1

Figure 74 gives an overview on context elements relevant for metadata generation for AHSGA. As in the structure of the directory, where a document is stored is relevant not only with respect to the business layer but also to the technology layer these elements are depicted, too. Most important here is the relation between the `node` (`AHSGA_Directory`) and the structuring elements `Product`, `BusinessService`, `BusinessBehaviourElement`, and `BusinessActor` (indicated by the solid black lines).

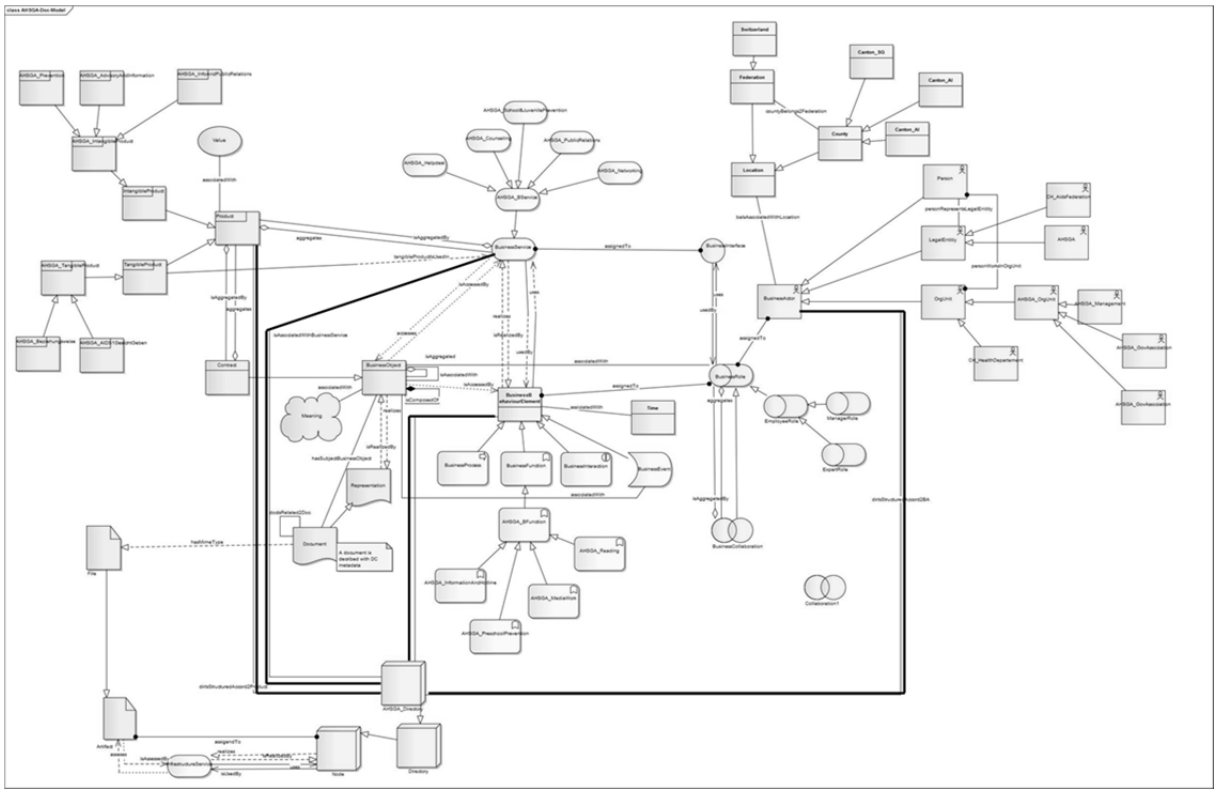


Figure 74: AHSGA Context Model Overview

All concepts and properties (data propterties and relations between concepts) are formally represented in seEAD.

6.1.3 AHSGA Formal Competency Questions

After AHSGA's terminology has been specified formal competency questions were written using the query language SPARQL. "The definition of a formal semantics for SPARQL has played a key role in the standardization process of this query language. Although taken one by one the features of SPARQL are intuitive and simple to describe and understand, it turns out that the combination of them makes SPARQL into a complex language" (Arenas, Gutierrez, & Pérez, 2010). SPARQL is a language designed to query data in the form of sets of triples, namely RDF graphs, which become a W3C Recommendation in 2008. Angles and Gutierrez (2008) consider its expressive power as potent as "relational algebra under bag semantics". SPARQL's expressiveness and resemblance to SQL allowed phrasing competency questions formally but in a language still understandable by non-experts, like my Action Research partners. Table 16 gives an example of the competency questions (1a and 1b) rewritten in SPARQL. Below each query statement an excerpt of the result set is listed.

QNo	Informal Question	SPARQL Query

1a	Given a document stored in a directory at the enterprise's file server (AH_GL, **AH_GS**, AHS, BAG, etc.)
	AND
	some constraints for authorship (has **role** manager, expert, etc.)
	which **primary** context elements can be determined?
		SELECT DISTINCT ?doc ?y ?x
		WHERE {
		?dir rdfs:label "AHSGA_ProfServices"@en .
		?dir eo:directoryContainsDocument ?doc .
		?doc elements:documentHasCreator ?creator .
		?role eo:businessRoleIsperformedByPerson ?creator.
		?role rdfs:label "PadagogueInHumanSexualBehaviour" @en .
		?doc ?y ?x.
		?y rdf:type owl:ObjectProperty .
		FILTER (?y **!=** eo:documentIsAssociatedTo2Context) .
		FILTER (?y **!=** eo:documentIsAssociatedTo3Context) .
		}
		ORDER BY ?x

Extract of Result Set

doc	y	x
eo:AHSGA_Document_1	elements:documentIsAssociatedWithDocument	eo:AHSGA_Document_10
eo:AHSGA_Document_9	extenteo:documentHasSubjectProduct	eo:AHSGA_InfUOeffArbeit
eo:AHSGA_Document_1	extenteo:documentHasSubjectProduct	eo:AHSGA_Praevention
eo:AHSGA_Document_14	extenteo:documentHasSubjectProduct	eo:AHSGA_Praevention
eo:AHSGA_Document_2	extenteo:documentHasSubjectProduct	eo:AHSGA_Praevention
eo:AHSGA_Document_1	eo:documentIsStoredInDirectory	eo:AH_GS
eo:AHSGA_Document_14	eo:documentIsStoredInDirectory	eo:AH_GS
eo:AHSGA_Document_2	eo:documentIsStoredInDirectory	eo:AH_GS
eo:AHSGA_Document_9	eo:documentIsStoredInDirectory	eo:AH_GS
eo:AHSGA_Document_9	eo:documentHasSubjectDomain	eo:AIDS
eo:AHSGA_Document_1	eo:documentIsArchivedAccordingToRegDoc	eo:Archivgesetz
eo:AHSGA_Document_1	eo:documentIsArchivedAccordingToRegDoc	eo:Artikel957962Obligationenrecht
eo:AHSGA_Document_1	extenteo:documentHasSubjectBusinessActor	eo:AusbildungAuboden
eo:AHSGA_Document_1	eo:documentIsArchivedAccordingToRegDoc	eo:Datenschutzgesetz
eo:AHSGA_Document_1	eo:documentIsStoredInDirectory	eo:Dir_Auboden

QNo	Informal Question	SPARQL Query

1b	Given a document is stored in a directory at the enterprise's file server (AH_GL, AH_GS, AHS, BAG, etc.)
	AND
	some constraints for the directory structure (related to business actor, business service, etc.) which primary context elements can be determined?
		SELECT DISTINCT ?doc ?y ?x
		WHERE {
		?dir eo:directoryContainsDocument ?doc .
		?dir eo:directoryHasStructureElementBusinessActor ?BA .
		?doc ?y ?x.
		?y rdf:type owl:ObjectProperty .
		FILTER (?y != eo:documentIsAssociatedTo2Context) .
		FILTER (?y != eo:documentIsAssociatedTo3Context) .
		}
		ORDER BY ?x

Extract of Result Set

doc	y	x
eo:AHSGA_Document_10	elements:documentIsAssociatedWithDocument	eo:AHSGA_Document_1
eo:AHSGA_Document_1	elements:documentIsAssociatedWithDocument	eo:AHSGA_Document_10
eo:AHSGA_Document_8	extenteo:documentHasSubjectProduct	eo:AHSGA_Fachberatung
eo:AHSGA_Document_3	extenteo:documentHasSubjectProduct	eo:AHSGA_InfUOeffArbeit
eo:AHSGA_Document_7	extenteo:documentHasSubjectProduct	eo:AHSGA_InfUOeffArbeit
eo:AHSGA_Document_9	extenteo:documentHasSubjectProduct	eo:AHSGA_InfUOeffArbeit
eo:AHSGA_Document_1	extenteo:documentHasSubjectProduct	eo:AHSGA_Praevention
eo:AHSGA_Document_14	extenteo:documentHasSubjectProduct	eo:AHSGA_Praevention
eo:AHSGA_Document_2	extenteo:documentHasSubjectProduct	eo:AHSGA_Praevention
eo:AHSGA_Document_5	eo:documentIsStoredInDirectory	eo:AH_GL

Table 16: AHSGA's Competency Question 1 Rewritten in SPARQL

6.1.4 AHSGA Rule Model

In the following the rules, which complete the formal axioms modeled as concepts and properties in seEAD, are presented. To express the rules the SPARQL Inferencing Notation (SPIN) is chosen. As already stated with SPIN, rules are expressed in SPARQL, in fact, SPIN is also referred to as *SPARQL Rules*[171]. SPIN also provides meta-modeling capabilities that allow users to define their own SPARQL functions and query templates and includes a library of common functions. More information on SPIN rules used in my approach is provided in Chapter 7.2.3,

Table 17 lists the SPIN rules defined for AHSGA. Rules AHSGA_IR_1 to AHSGA_IR_20 infer metadata for AHSGA's documents and rules AHSGA_IR_21 to AHSGA_IR_23 generate SKOSConcepts if they don't exist.

[171] SPIN. SPARQL Inferencing Notation. URL: http://spinrdf.org/ (retrieved: 15.8.2012)

SPIN Rule	Remark

CONSTRUCT {
?this extenteo:documentHasDocumentSeedingDate ?edate .
}
WHERE {
?this a eo:AHSGA_Document .
NOT EXISTS {
?this extenteo:documentHasDocumentSeedingDate ?sdate .
} .
FILTER (!bound(?sdate)) .
BIND (afn:now() **AS** ?sdate) .
BIND (spif:dateFormat(?sdate, "dd.MM.yyyy") **AS** ?ddate) .
BIND (spif:parseDate(?ddate, "dd.MM.yyyy") **AS** ?edate) .
}
	AHSGA_IR_1
	Creates the seeding date of a document: for all documents that do not already have a seeding date the current date is taken as seeding date and transformed into an appropriate format.

CONSTRUCT {
?x elements:documentHasCreator ?creator.
}
WHERE {
?x eo:documentHasSeedEmpNo ?eno .
?y eo:employeeHasEmployeeNo ?bno.
?y eo:businessRoleIsperformedByPerson ?creator .
LET (?yes :=fn:matches(?eno, ?bno)) .
Filter (?yes =true) .
NOT EXISTS { ?x elements:documentHasCreator ?creator }
}
	AHSGA_IR_2
	Infers the creator of a document (metadata set) based on the match between employee seed no and employee no for all instances where a creator does not already exist

CONSTRUCT {
?this extenteo:documentHasSubjectBusinessActor ?BA .
}
WHERE {
?this eo:documentIsStoredInAHSGA_Directory ?AHSGA_DIR .
?AHSGA_DIR eo:dAHSGA_DirectoryHasStructureElementBusinessActor ?BA .
?BA eo:businessActorHasAssignedBusinessRole ?BR .
}
	AHSGA_IR_3
	Infers the business actor related to a AHSGA_Directory

SPIN Rule	Remark

CONSTRUCT {
?this extenteo:documentHasSubjectBusinessService ?BS .
}
WHERE {
?this eo:documentIsStoredInAHSGA_Directory ?AHSGA_Dir
?dAHSGA_IR eo:AHSGA_DirectoryHasStructureElementBusinessService ?BS .
}
	AHSGA_IR_4
	Infers the business service related to a AHSGA_Directory

CONSTRUCT {
?this extenteo:documentHasSubjectProduct ?BP .
}
WHERE {
?this eo:documentIsStoredInAHSGA_Directory ?AHSGA_Dir
?dAHSGA_IR eo:AHSGA_DirectoryHasStructureElementProduct ?BP .
}
	AHSGA_IR_5
	Infers the product related to a AHSGA_Directory

CONSTRUCT {
?this extenteo:documentHasSubjectBusinessRole ?BR .
}
WHERE {
?this eo:documentIsStoredInAHSGA_Directory ?AHSGA_Dir
?dAHSGA_IR eo:AHSGA_DirectoryHasStructureElementBusinessRole ?BR .
}
	AHSGA_IR_6
	Infers the business role related to a AHSGA_Directory

CONSTRUCT {
?this extenteo:documentHasSubjectBusinessFunction?BF .
}
WHERE {
?this eo:documentIsStoredInAHSGA_Directory ?AHSGA_Dir
?dAHSGA_IR eo:AHSGA_DirectoryHasStructureElementBusinessFunction ?BF .
}
	AHSGA_IR_7
	Infers the business function related to AHSGA_Directory

SPIN Rule	Remark

CONSTRUCT {
?this eo:documentIsAssociatedTo2Context ?2context
.
}
WHERE {
?this elements:documentHasCreator ?creator .
?creator a eo:Person .
?creator eo:personPerformsBusinessRole ?2context .
}
	AHSGA_IR_8
	Infers a business role related to the creator

CONSTRUCT {
?s eo:documentIsAssociatedTo2Context ?2context .
}
WHERE {
?s extenteo:documentHasSubjectBusinesService ?BS
.
?BS
archi:businesServiceIsAssociatedToBusinessFunction
?2context .
NOT EXISTS {
?s extenteo:documentHasSubjectBusinessFunction
?2context .
} .
}
	AHSGA_IR_9
	Infers a business function related to a business service

CONSTRUCT {
?s eo:documentIsAssociatedTo2Context ?2context .
?s eo:documentIsAssociatedTo3Context ?3context .
}
WHERE {
?s extenteo:documentHasSubjectBusinessActor ?BA .
?BA eo:businessActorHasAssignedBusinessRole
?BR .
?BR rdf:type ahsga:AHSGA_Client .
?BA eo:businessActorIsSituatedInLocation ?2context
.
?2context top:cityIsLocatedInPartOfCountry
?3context .
}
	AHSGA_IR_10
	Infers the location of a business actor and the canton of the location

SPIN Rule	Remark

CONSTRUCT {
?this eo:documentIsAssociatedTo3Context ?3context
.
}
WHERE {
?this eo:documentIsAssociatedTo2Context ?BS .
?BS archi:businesServiceIsAssociatedToBusinessFunction ?3context .
NOT EXISTS {
?this extenteo:documentHasSubjectBusinessFunction ?3context .
} .
}
	AHSGA_IR_11
	Infers the business function aggregated by business services associated to 2Context

CONSTRUCT {
?this eo:documentIsAssociatedTo3Context ?3context
.
}
WHERE {
?this elements:documentHasCreator ?creator .
?creator eo:personPerformsBusinessRole ?role .
?role eo:employeeIsResponsibleForProduct ?3context
.
NOT EXISTS {
?this extenteo:documentHasSubjectProduct ?3context .
} .
}
	AHSGA_IR_12
	Infers a product for which a business role is responsible related to a business actor

CONSTRUCT {
?this extenteo:documentHasSubjectBusinessObject ?BO .
}
WHERE {
?this extenteo:documentHasSubjectProduct ?p .
?p rdfs:label "AdvisoryAndInformation"@en .
?BO eo:businessObjectIsAssociatedTo ?p .
}
	AHSGA_IR_13
	Infers business objects related to a specific product

CONSTRUCT {
?this elements:documentHasType ?dcmitype .
}
WHERE {
?this elements:documentHasFormat ?mimetype .
?mimetype eo:MIMETypeIsRelated2DCMITYPE ?dcmitype .
}
	AHSGA_IR_14
	Infers document type for a harvested document format

SPIN Rule	Remark

CONSTRUCT {
?this extenteo:documentHasArchiveDate ?year .
}
WHERE {
?this a eo:AHSGA_Document .
?this ahsga:AHSGA_DocumentHasStatus "closed" .
OPTIONAL {
?this extenteo:documentHasArchiveDate ?xdate .
} .
FILTER (!bound(?xdate)) .
BIND (afn:now() **AS** ?sdate) .
BIND (spif:dateFormat(?sdate, "dd.MM.yyyy") **AS** ?ddate) .
BIND (fn:substring(?ddate, 7) **AS** ?year) .
}
	AHSGA_IR 15
	Creates archiving date for documents based on closing date

CONSTRUCT {
?this eo:documentIsArchivedAccordingToRegDoc ?lawdoc .
}
WHERE {
?this extenteo:documentHasArchiveDate ?adate .
?this elements:documentHasCreator ?creator .
?creator eo:personWorksForLegalEntity ?lentity .
?lentity eo:businessActorIsSituatedInLocation ?location .
?location eo:locationHasLawAndRegulation ?law .
?law a eo:GeneralArchivingLaw .
?law eo:lawAndRegulationIsExpressedInRegulationDocument ?lawdoc .
	AHSGA_IR 16
	Infers law that determines retention of documents based on a document creator's status and location of the enterprise she is working for

SPIN Rule	Remark

CONSTRUCT {
?this eo:documentIsArchivedAccordingToRegDoc ?lawdoc2 .
}
WHERE {
?this extenteo:documentHasArchiveDate ?adate .
?this elements:documentHasCreator ?creator .
?this extenteo:documentHasSubjectBusinessActor ?BA .
?BA eo:businessActorHasAssignedBusinessRole ?BR .
?BR a ahsga:AHSGA_Client .
?creator eo:personWorksForLegalEntity ?lentity .
?lentity eo:businessActorIsSituatedInLocation ?location .
?location eo:locationHasLawAndRegulation ?law .
?law a eo:DataProtectionAct .
?law eo:lawAndRegulationIsExpressedInRegulationDocument ?lawdoc2 .
}
	AHSGA_IR 17
	Infers law that determines retention of documents based on a document creator's status, location of the enterprise she is working for and the involvement of client

CONSTRUCT { ?doc1 elements:documentIsAssociatedWithDocument ?doc2}
WHERE {
?doc1 ahsga:AHSGA_DocumentHasStatus "selected" .
?doc1 extenteo:documentHasReportedDate ?rdate1.
?doc2 ahsga:AHSGA_DocumentHasStatus "selected" .
?doc1 extenteo:documentHasReportedDate ?rdate2.
FILTER (?doc1 != ?doc2) .
FILTER (?rdate1 = ?rdate2)
}
	AHSGA_IR 18
	Infers relations between documents based on status (selected) and similar time (date, hour, minutes)

CONSTRUCT
{
?t extenteo:documentHasSubjectBusinessActor ?BA
}
WHERE
{
?t eo:documentHasSubjectDomain ?domain .
?subloc rdfs:subClassOf top:PhysicalLocation .
?domain rdf:type ?subloc .
?BA eo:businessActorIsSituatedInLocation ?domain .
}
	AHSGA_IR 19
	Infers Business Actors related to the document from a document's subject domain

SPIN Rule	Remark

CONSTRUCT {
  ?this eo:documentHasArchivingTime ?maxperiod .
}
WHERE {
  {
  SELECT ?this ((**MAX**(?period)) **AS** ?maxperiod)
  WHERE {
  ?this eo:documentIsArchivedAccordingToLaw ?law .
  ?law eo:lawDefinesArchivingPeriod ?period .
  }
  GROUP BY ?this
  } .
}
	AHSGA_IR_20
	Infers the total archiving period for a document
	Rule is also valid for contract documents (cf. Symfact_IR_6)

CONSTRUCT {
  ?domain a skos:Concept .
}
WHERE {
  ?document eo:documentHasSubjectDomain ?domain .
  NOT EXISTS {
  ?domain a skos:Concept .
  } .
}
	AHSGA_IR 21
	Creates an instance of a skos:concept if it does not yet exist

CONSTRUCT {
  ?this rdfs:label ?label .
}
WHERE {
  ?this a skos:Concept .
  NOT EXISTS {
  ?this rdfs:label ?label .
  } .
  BIND (str(**?this**) **AS** ?testString) .
  BIND (fn:substring(?testString, 19) **AS** ?prelabel) .
  BIND (strlang(?prelabel, "de") **AS** ?label) .
}
	AHSGA_IR 22
	Creates labels with language “de” for newly created concepts

SPIN Rule	Remark

CONSTRUCT {
?this skos:broader ?t .
}
WHERE {
?this a skos:Concept .
?this rdfs:label ?label1 .
FILTER (lang(?label1) = "de") .
BIND (str(?label1) **AS** ?shortLabel1) .
BIND (fn:lower-case(?shortLabel1) **AS** ?llabel1) .
?t a skos:Concept .
?t rdfs:label ?label2 .
FILTER (lang(?label2) = "de") .
BIND (str(?label2) **AS** ?shortLabel2) .
BIND (fn:lower-case(?shortLabel2) **AS** ?llabel2) .
BIND (fn:ends-with(?llabel1, ?llabel2) **AS** ?yes) .
FILTER (?yes = true) .
BIND (fn:string-length(?llabel1) **AS** ?llength1) .
BIND (fn:string-length(?llabel2) **AS** ?llength2) .
FILTER (?llength1 > ?llength2) .
NOT EXISTS {
?this skos:broader ?t .
} .
}
	AHSGA_IR 23
	Creates a broader term relation for newly added concepts based on the linguistic method that the most right element is the specified element and thus the broader term

Table 17: SPIN Rules for AHSGA

6.1.5 AHSGA Description Set Profile

As shown in Chapter 3.1.4 no specific standard is available for metadata describing enterprise documents but still there are still good reasons to describe an enterprise's documents based on a standard. One reason might be to make information accessible beyond the scope of a certain project. For making information exchangeable with other/new stakeholders as for example in case of the AHSGA with other regional institutes for AIDS, use of a standard is clearly beneficially.

As Greenberg et al. (2005) showed in the AMEGA report, organizations are using a variety of different metadata standards, but Dublin Core (simple or qualified) is prevailing in non-library environments[172]. Therefore Dublin Core is chosen for describing enterprise objects in my thesis and, adapted to the specific for purpose of an enterprise as for AHSGA.

Dublin Core (DC) has been chosen for AHSGA for the following reasons:

- DC can be customized according to enterprises' specific needs by designing an application profile[173]

[172] [172] Page 23, Table 14: DC simple has been mentioned 33 times (42.3%), DC qualified 32 times (41.0%) and DC application profile 17 (21.8.%).

[173] The Dublin Core Metadata Initiative has addressed the problem of adapting metadata schemas to specific needy by providing a framework for designing a Dublin Core Application Profile (DCAP). "A DCAP defines metadata records which meet specific application needs while providing semantic interoperability with other

- DC is machine processable (can be represented in terms of RDF/XML)[174]
- DC can be combined with / integrated into knowledge representations of an ontology (Lux 2006)
- DC provides a comprehensive set of specifications, guidelines and recommendations[175]
- DC is supported by many tools[176]
- DC has gateways to many other standards, for example to MODS[177].

The below listed metadata elements (Table 18) have been defined by the AHSGA team based on the Dublin Core standard. They can be considered as the target elements AHSGA expects from automatic metadata generation based on semantically enriched context information.

DCMI No	Dublin Core Metadata Element	Property in seEAD	Refinement	Syntax Encoding Scheme	Value Encoding Scheme (range in seEAD)
1	Contributor	dc:documentHasContributor			archi:BusinessActor
3	Creator	dc:documentHasCreator			archi:BusinessActor
4	Date	dc:documentHasDate	dcq:documentHasCreationDate dcq:documentHasModifiedDate	http://www.w3.org/TR/NOTE-datetime	
6	Format	dc:documentHasFormat			iana:MIMETYPE [178]
7	Identifier	dc:documentHasIdentifier		unique number	
10	Relation	dc:documentHasRelationToDocument			foaf:Document [179]
13	Subject	dc:documentHasSubject			emo:AHSGA_SKOS
13	Subject	dc:documentHassubject	dceo:documentHasSubjectEnterpriseObject		eo:EnterpriseObject
14	Title	dc:documentHastitle			

applications on the basis of globally defined vocabularies and models". URL: http://dublincore.org/documents/profile-guidelines/index.shtml (retrieved: 6.11.2010)

[174] Expressing Dublin Core metadata using the Resource Description Framework (RDF). URL: http://dublincore.org/documents/dc-rdf/index.shtml (retrieved: 6.11.2010)

[175] DCMI Specifications. URL: http://dublincore.org/specifications/ (retrieved: 6.11.2010)

[176] Tools and Software URL: http://dublincore.org/tools/index.shtml (retrieved: 6.11.2010)

[177] Dublin Core Metadata Element Set Mapping to MODS Version 3. URL: http://www.loc.gov/standards/mods/dcsimple-mods.html (retrieved: 6.11.2010)

[178] MIME Media Types. URL: http://www.iana.org/assignments/media-types/index.html (retrieved: 15.8.2011)

[179] FOAF Vocabulary Specification 0.98. URL: http://xmlns.com/foaf/spec/#term_Document (retrieved: 15.8.2011)

DCMI No	Dublin Core Metadata Element	Property in seEAD	Refinement	Syntax Encoding Scheme	Value Encoding Scheme (range in seEAD)
15	Type	dc:documentHasType			dcmi:DCMITYPE
	Non DC Metadata Elements				
	Storage	dceo:documentHasStorage			archi:node
	Status	ahsga:AHSGA_DocumentHasStatus			eo:documentStatus
	Compulsory Archiving	eo:documentIsArchivedAccordingToLaw			eo:LawAndRegulation
	Regulation Document	eo:documentIsArchivedAccordingToRegDoc			eo:RegulationDocument

Table 18: AHSGA Metadata Element Set

In the following the enterprise specific characteristics of automatic metadata generation for AHSGA are detailed, based on the general approach introduced in Chapter 5.2.3. First document properties are harvested to build the basis for automatic, format-independent metadata generation (cf. Chapter 5.2.3.1). From the harvest the metadata seeds are derived (cf. Chapter 5.2.3.2) and from the metadata seeds metadata (candidates) are inferred (cf. 5.2.3.3).

6.1.5.1 AHSGA Document Properties – File Harvest

Table 19 provides an overview on document properties provided by various document creation software systems or the operating system. Since AHSGA works with the German versions of the software Table 19 provides the document properties in German as used in the prototype. Again, the far right column provides the harvested attributes (data sink) expressed in DC metadata terms.

DOC	PDF	JPEG	MP3 (audio)	MP4 (video)	PNG	GIF	XLS	PPT	Alle	**Dublin Core**
Name	Name	Name	Name	Name	Name	Name	Name	Name	(Datei)Name	**alternative**
Dateityp	Dateityp	Dateityp	Dateityp	Dateityp	Dateityp	Dateityp	Dateityp	Dateityp	Dateityp	**mime-type**
Ort	Ort	Ort	Ort	Ort	Ort	Ort	Ort	Ort	Ort	**location**
Erstellt	Erstellt	Erstellt	Erstellt	Erstellt	Erstellt	Erstellt	Erstellt	Erstellt	Erstellt	**created**
Geändert	Geändert	Geändert	Geändert	Geändert	Geändert	Geändert	Geändert	Geändert	Geändert	**modified**
Titel	Titel	Titel	Titel	Titel	Titel	Titel	Titel	Titel	Titel	**title**
	Verfasser								Verfasser	
Autoren		Autoren					Autoren	Autoren	Autoren	**contributor**
Thema	Thema	Thema					Thema	Thema	Thema	
	Stichwörter								Stichwörter	**subject**
			Herausgeber	Herausgeber					Herausgeber	**publisher**

Table 19: AHSGA's Data Source and Sink for Metadata Harvesting

The provided list contains all document properties relevant to meet AHSGA's requirements. However, if in future more or other document properties should be harvested or other document creation software is used, the list must be adapted and the adapters of Metadata Extractor harvester must be modified accordingly or newly created.

6.1.5.2 AHSGA File Harvest – Metadata Seeds

Metadata seeds are built from harvested document properties. In other words: for each document that has been harvested an instance of the `Document` concept in seEAD is created and each harvested document property becomes a property of the instance. Table 20 gives an example of the data source (the metadata harvest) and sink of metadata seeds for AHSGA. Column three indicates the property type: lighter grey is a data property and darker grey an object property.

Example	Metadata Harvest		Metadata Seed	
Homosexualitaet in Heimen	Name (Dateiname)		`documentHasTitle`	
pdf	Mime-Type (Dateityp)		`documentHasFormat`	`MIME-Type`
C:\TEXT\AH_GS\DIREKTEP\HEIME JUG\Auboden	Location (Ort)		`documentIsStoredInFilesystem`	`Filesystem`
1.8.2011	Created (erstellt)		`documentHasCreationDate`	
1.12.2012	Modified (geändert)		`documentHasModifiedDate`	
Arbeit in Heimen 409	Title (Titel)		`documentHasAlternativeTitle`	
CJaeggi	Contributor (Verfasser, Autoren, Besitzer)		`documentHasContributor`	`BusinessActor`
Jugendsexualitaet	Subject (Stichwörter)		`documentHasSubjectFromFA`	
not applicable in pdf files	Publisher (Herausgeber)		`documentHasPublisher`	`BusinessActor`
	Derived Metadata Harvest			
4	EmployeeNo		`documentHasSeedEmpNo`	
Homosexualitaet ; Heim	Subject		`documentHasSubject`	`SKOSConcept`
AH_GS; DIREKTEP; HEIME JUG; Auboden	Directory		`documentIsStoredInDirectory`	`Directory`

Table 20: AHSGA's Data Source and Sink for Metadata Seeds

In addition to harvested document properties further seeds are created. In Table 20 an example is given for three metadata elements: `documentHasSeedEmpNo`, `documentHasSubject`, and `documentIsStoredInDirectory`.

Which harvested information is used in which way for metadata seed creation has been defined with AHSGA. In the example depicted in Table 20 metadata seed creation is based on AHSGA low-level governance instruments, namely rules for file name and path name conventions.

- AHSGA has defined the convention file names should start with the employee number. To build the `documentHasSeedEmpNo` metadata seed, string operation is performed on the filename: The first digit of the filename is separated and stored as employee number (the seed for 'creator'); if the file name does not start with a number nothing is extracted. Table 20 gives an example for the file name '409 Homosexualität in Heimen.doc'. When it is parsed '4' is separated as the number of the employee who created the file[180].
- Nouns in filenames usually refer to the subject of the document. In the example the terms 'Homosexualität' and 'Heim' are extracted from the file name and stored as seeds for `documentHasSubject`.
- To separate the directories the harvested path name is parsed and its parts are separated. In the depicted example the path name 'C:\TEXT\AH_GS\DIREKTEP\HEIME JUG\Auboden' is segmented and partition (C:) and top-level directory (TEXT) are omitted as they have no meaning for metadata creation. For the remaining elements values of four metadata seeds are generated:
  `documentIsStoredInDirectory` AH_GS,
  `documentIsStoredInDirectory` DIREKTEP,
  `documentIsStoredInDirectory` HEIME JUG, and
  `documentIsStoredInDirectory` Auboden,
  whereas `directory` is sub-concept of `node`.

6.1.5.3 AHSGA Metadata Seeds – Metadata

Table 21 depicts an example of primary context elements that are inferred from metadata seeds for additional metadata for AHSGA's documents. For better reading namespaces are omitted in the table.

Metadata Seed	Value	Primary Context	Instance	Example / Remark
`Creator`	*4*	`Person`	*SimoneSchneider*	'4' is value of the data property AHSGAEmployeeHasEmployeeNo; Simone Schneider is an AHSGA employee with the employee no 4.
Inferred Metadata	`Document DocumentHasCreator Person`		*SimoneSchneider*	if desired the respective metadata seed can be removed

[180] '4' is the employee ID, '09' is the year in which the resource has been created

Contributor	*GJaeggi*	-	-	no match in seEAD can be found; actually the person is a former employee of AHSGA
No metadata (candidate) is inferred				
Directory	*AH_GS*	OrganisationalUnit	*AHSGA_ProfServices (german: AHSGA_Geschaeftsstelle*[181]*)*	'AH_GS' is an instance of the concept eo:Directory which is sub-concept of Node
Inferred Metadata	Document DocumentHasSubjectBusinessActor OrganisationalUnit		*AHSGA_ProfServices (german: AHSGA_Geschaeftsstelle*	Directory is related to sub-concepts of archi:BusinessActor
Directory	*DIREKTEP*	IntangibleProduct	*Prevention (german: Prävention)*	'DIREKTEP' is an instance of the concept eo:Directory which is sub-concept of Node
Inferred Metadata	Document DocumentHasSubjectProduct Product		*Prevention (german: Prävention)*	Directory is related to sub-concepts of archi:Product
Directory	*HEIME JUG*	BusinessRole	*JuvenileHome (german: Jugendheim)*	'HEIME_JUG' is an instance of the concept eo:Directory which is sub-concept of Node
Inferred Metadata	Document DocumentHasSubjectBusinessRole BusinessRole		*JuvenileHome (german: Jugendheim)*	Directory is related to sub-concepts of BusinessRole
Directory	*HEIME JUG*	BusinessService	*School and Juvenile Prevention (german: Prävention Schule und Jugend)*	'HEIME_JUG' is an instance of the concept Directory which is sub-concept of Node
Inferred Metadata	Document DocumentHasSubjectBusinessService BusinessService		*School and Juvenile Prevention (german: Prävention Schule und Jugend)*	Directory is related to sub-concepts of BusinessService
Directory	*AUBODEN*	LegalEntity	*Educational Institution Auboden(german: Ausbildungsstätte Auboden)*	'AUBODEN' is an instance of the concept Directory which is sub-concept of Node

[181] Since Action Research is performed in a German speaking organisation the prototype must deal with German terms. Thus, German translation is put in brackets where necessary.

Inferred Metadata	Document DocumentHasSubjectBusinessActor LegalEntitiy		*Educational Institution Auboden(german: Ausbildungsstätte Auboden)*	Directory is related to sub-concepts of BusinessActor
MIMEType	*pdf*	DCMIType	*Text (german: Text)*	
Inferred Metadata	Document DocumentHasSType DCMIType		*Text (german: Text)*	
Subject	*Homoexualität*	SKOSConcept	*Homosexuality (german: Homosexualität)*	If the subject not already exists, a new instance of the concept SKOSConcept is created
Inferred Metadata	Document DocumentHasSubjectDomain Domain		*Homosexuality (german: Homosexualitaet)*	AHSGA_Document is related to SKOSConcept; i.e. the seed is reified by the rdfs:range SKOSConcept

Table 21: AHSGA's Data Source and Sink for Creating Metadata Based on Primary Context Elements

Following the general approach for metadata generation all instances of a document's primary context elements (in the example Person, OrganisationalUnit, IntangibleProduct, BusinessRole, LegalEntity and SKOSConcept) are inferred for the respective metadata seeds.

6.1.5.4 AHSGA Metadata – Metadata Candidates

After all metadata are generated based on primary context elements, secondary context elements are inferred for metadata candidates. However, not all primary context elements are considered but only those that might be relevant for later retrieval with AHSGA's ITRS. For example: the context of Person is inferred because a creator's BusinessRole can be used to broaden a query in order to find documents created by people with a similar role. In contrast, the context of OrganisationalUnit is not considered because it isn't stored in ITRS and thus cannot be searched for. Which context elements are relevant and which not have been defined with AHSGA within loop 2 of the Action Research. Table 22 provides for the aformentioned example the primary context elements and the inferred secondary context elements.

Primary Context	Instance	Secondary Context	Instance	Remark
Person	*SimoneSchneider*	ExpertRole (is a BusinessRole)	*PadagogueInHumanSexualBehaviour (german: Sexualpädagoge)*	
BusinessService	*School and Juvenile Prevention*	Business Function	*Education (german: Ausbildung)*	At this point in the metadata

Primary Context	Instance	Secondary Context	Instance	Remark
	(german: Prävention Schule und Jugend)		*ExtracurricularDomain (german: AusserschulischerBereich*	generation process all business functions related to the business service are created as instances of level two context concepts
			(german: InfoUWeiterbildungInBerufsgruppe)	
			SecondarySchoolPrevention(german: KantonsschulPraevention)	
			SexualPadagogicsIndividualWork(german: SexualpaedEinzelarbeit)	
			SocialEducationalInstitutes(german: SozialpaedInstitutPraevention)	
			ElementarySchoolPrevention(german: VolksschulPraevention)	
			PreschoolPrevention(german: VorschulPraeventi on)	
LegalEntiti y	*EducationalInstitution Auboden (german: Ausbildungsstätte Auboden)*	*Location*	*Brunnadern*	The institution is located in the Swiss City of Brunnadern

Primary Context	Instance	Secondary Context	Instance	Remark
Subject	*Homosexuality (german: Homosexualitaet)*	BroaderTerm	*Sexuality (german: Sexualitaet)*	'sexuality' is an instance of AHSGA's SKOS domain knowledge; 'sexuality' is considered a broader term to 'homosexuality' based on knowledge about German Determinativ-komposita [182]
Inferred Metadata	Document documentIsAssociatedTo2Context <*ContextElement*>		<*ContextElement*>	For all metadata candidates an instance of the inferred context element is created <*ContextElement*>.

Table 22: AHSGA's Data Source and Sink for Creating Metadata Candidates Based on Secondary Context Elements

Table 23 depicts the metadata candidates inferred for AHSGA's document from tertiary context information. For all inferred context elements the Document documentIsAssociatedTo3Context <*ContextElement*> instance is created.

Secondary Context	Instance	Tertiary Context	Instance	Remark
Location	*Brunnadern*	PartOfCountry	*Appenzell*	The Swiss City of Auboden belongs to the Canton Appenzell

[182] In German the modifier in noun phrases (german: Determinativkomposita) come before the head, e.g. the left term *homo* specifies the right term *sexualitaet* in the noun compound *Homosexualitaet*. Thus, the word position can be used to determine the meaning of a noun. Refer to Donalies (2005) on the segmentation of Determinativkomposita.

Secondary Context	Instance	Tertiary Context	Instance	Remark
Business Role	*PadagogueInHumanSexualBehaviour (german: Sexualpädagoge)*	Product	*EroticGame (german: Erotikspiel)*	The creator's business role is responsible for products; here for the tangible product of a game. The role is also responsible for the intangible product of 'AHSGA_prevention' but as this has been already generated as primary context element it is omitted here.
Inferred Metadata	Document documentIsAssociatedTo3Context <*ContextElement*>		<*ContextElement*>	Similar to level two context elements for all metadata candidates an instance of the inferred context element is created <*ContextElement*>

Table 23: AHSGA's Data Source and Sink for Creating Metadata Candidates Based on Tertiary Context Elements

6.1.5.5 AHSGA ITRS Data

AHSGA defined the Information and Time Recording System (ITRS) as target system for managing AHSGA's documents. As depicted in

Figure 75 with a black rectangle around the tab called 'Dokumente' (documents) import of documents is already envisaged. Currently this functionality is not used as documents must be retrieved manually. That is, when executing the function the explorer is opened and the user can search for the documents he/she wants to attach to a task.

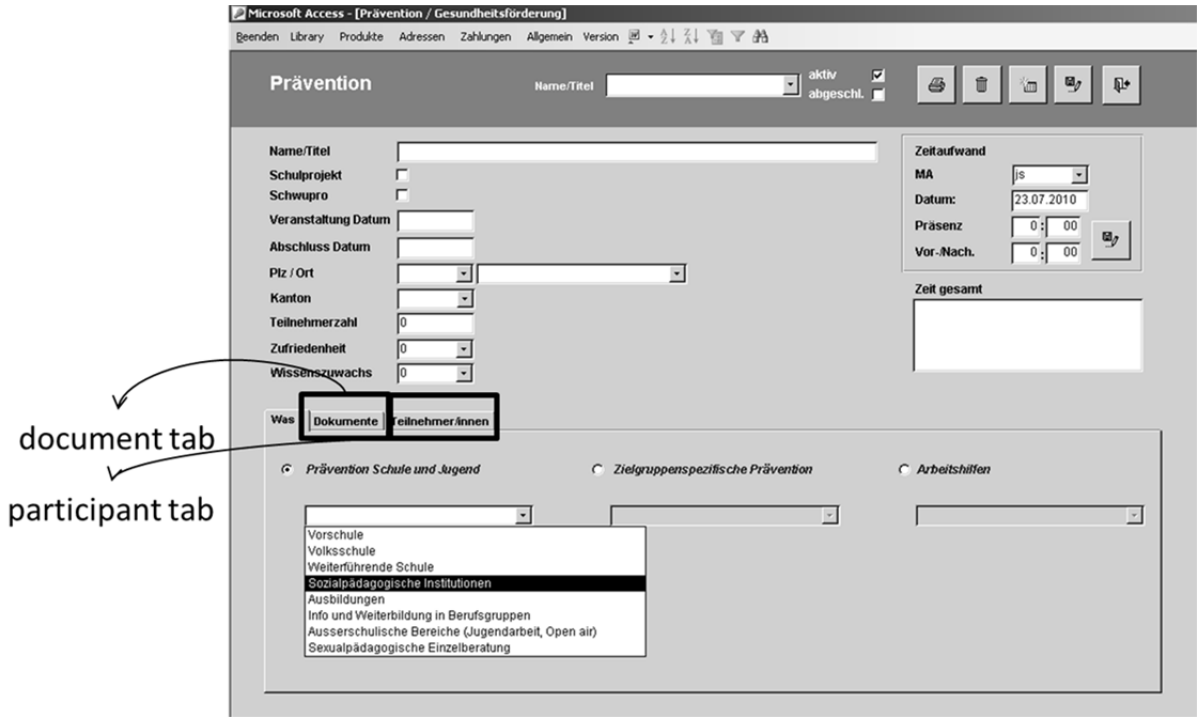


Figure 75: ITRS Print Screen as-is

On the left hand side of the print screen the data fields are shown that are to be filled when reporting on a task. This data will be used to query seEAD for documents that might be relevant for the task. If, for instance, an employee reports time on consultancy she has performed for a client in the ITRS, this information is used to query seEAD for documents having metadata (candidates) related to that client (e.g. `documentHasSubjectBusinessActor`). From a result list the user can select the documents she wants to add to the task. After that the documents will be imported automatically and stored under the 'Document' tab.

In the following an overview is given of the information recorded in ITRS. Table 24 lists the data that has to be reported for every task. On the right hand side of the table the affected axioms in seEAD are listed.

- For each task the name of the employee, the date of the entry and the canton for whom the task is done is reported
- All reported tasks are related to one of AHSGA's intangible products or to a non-billable business service:
  - Advisory and Information (Fachberatung u. Information)[183]
  - Prevention (Prävention, shown in
  - Figure 75)
  - Information & Public Relations (Informations- u. Oeffentlichkeitsarbeit)
  - Non-billable Services (Nicht verrechenbare Leistung).

[183] For easier mapping between database labels and seEAD concept labels German translations are given in brackets.

ITRS data element	ITRS Datenelement [184]	Axiom in seEAD	Remark

Name Initials of Employee	MA (MitarbeiterIn)	`Document documentHasCreator Person`
		
		optional
		
		`Document documentHasContributor Person`
			seEAD can be queried for creator and contributor of a document;
			To query metadata for name initials they must be inferred from the `BusinessActor` concept:
			`Person personHasNameInitials`

Date	Datum	`Document documentHasCreationDate`
		
		optional
		
		`Document documentHasModifiedDate`
			seEAD can be queried for creation and modification date of a document;

Canton	Kanton	`Document documentIsAssociatedToContext PartOfCountry`	`documentIsAssociatedToContext` is super-property to `documentIsAssociatedTo2Context`, `documentIsAssociatedTo3Context`
Intangible Product[185]	Produkt	`Document documentHasSubjectProduct IntangibleProduct`	
Business Service[186]	Leistung	`Document documentHasSubjectBusinesService BusinessService`	
Business Function[187]	Funktion	`Document documentHasSubjectBusinessFunction BusinessFunction`	

Table 24: Data Recorded in AHSGA's ITRS

[184] For convenience the original German labels of the data elements are listed, too
[185] Find the complete list of intangible products in Chapter 12.5.
[186] Find the complete list of services in Chapter 12.5.
[187] Find the complete list of business functions in Chapter 12.5.

In addition to general information, recorded for every task, product specific data can be entered.
Figure 75 illustrates the specific information for the intangible product 'Prevention'

A complete list of addition information is provided in Table 25.

Product	ITRS data element	ITRS Datenelement	seEAD axiom	Remark
Advisory and Information	Subject Matter	Beratungsinhalt	`Document documentHasSubjectBusinessObject BusinessObject`	In ITRS subject matter is represented as pull down list; subject matter is represented as `BusinessObject` in seEAD that is – according to ArchiMate – associated with `Product`, here: `AdvisoryAndInformation`

Prevention / Information & Public Relations	Name/Titl e	Name/Titel	`Document documentIsAssociatedTo2Context BusinessActor`
			
			optional
			
			`Document documentIsAssociatedTo2Context location`
				free text that can be parsed to identify matches of parts of the file name with client names or locations

	Schwupro	Schwupro (Schwulen-projekt)	`Document documentHasSubjectBusinessFunction BusinessFunction`
			
			optional
			
			`Document documentIsAssociatedTo2Context BusinessFunction`
			
			optional
			
			`Document documentIsAssociatedTo3Context BusinessFunction`
				tick box (if ticked it can be used to find a business object 'gay project'); 'Schwupro' is represented as `BusinessFunction` in seEAD that is – according to ArchiMate – associated with `BusinessService`, here: `TargetGroupSpecificPrevention` (german: Zielgruppenspezifische Prävention)

Product	ITRS data element	ITRS Datenelement	seEAD axiom	Remark
	Event Date	Veranstaltungs-datum	`Document documentHasCreationDate` < EventDate optional `Document documentHasModifiedDate` < EventDate	event date can be used to restrict time period of eligible documents; e.g. documents created after the event date will not be considered
	Closing Date	Abschlussdatum	`AHSGA_Document documentHasArchiveDate`	closing date can be used to set the date by when a document should be archived
	Postal Code	Postleitzahl	`City hasPostalCode`	not considered in the prototype
	City	Ort	`Document documentIsAssociatedTo2Context location`	information of the city can be used to confine documents related to business actors; e.g. given a canton all documents are eligible related to a business actor located in the respective canton; by adding the city only those business actors located in the very city remain

Table 25: ITRS Product Specific Data Elements

In addition to task reporting ITRS will be used for document search. In this case data entered in an ITRS panel is also transformed into query terms which seEAD is queried with but instead of importing retrieved documents they are displayed. Refer to Chapter 7.3 on details related to the prototype.

Figure 76 depicts the general approach: information of a task is recorded in ITRS: the symbols a7, b4, c3, d6, e1 illustrate the values of the data elements entered for task as explained above. The data elements correspond with metadata stored in seEAD and the values are used to query seEAD for matching metadata candidates of documents. In the example shown in Figure 76 the documents with DocID 1, 2, and 3 are retrieved, because at least one metadata value matches with one data value. The result list is presented to the user to choose the most appropriate (ones) – either for import or for display.

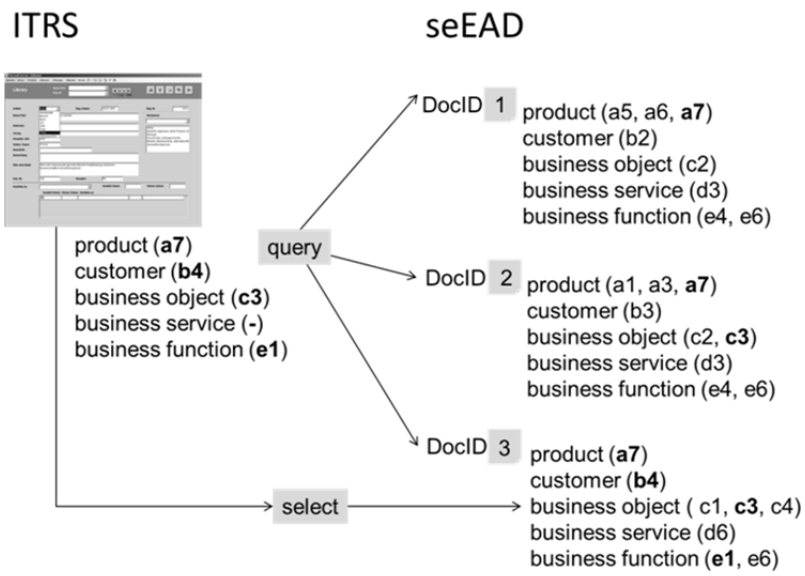


Figure 76: Matching ITR Information With Metadata Candidates

6.1.5.6 AHSG Metadata Update

The final step of automatic metadata generation for AHSGA's documents is to update metadata (candidates) of the documents that were selected for import into ITRS.
Metadata of these documents are automaticly updated. First of all, a document status is added (e.g. 'selected') and, if applicable, metadata candidates that match with query terms are transformed into true metadata. For example the metadata candidate

```
Document documentIsAssociatedTo2Context BusinessFunction
```

is transformed into the metadata

```
Document documentHasSubjectBusinessFunction BusinessFunction
```

because it is now assigned as subject, which is a DC metadata element. Remember: due the Dublin Core one-to-one principle meaning of metadata must not go beyond the scope of the original element, that is, a person might be added as creator of a document but her responsibility for a product must not. This is why such kind of inferred metadata is classified as context of level 2 to n. However, if a document is selected these metadata candidates can become true metadata as they are of direct relevance to the document and thus, their 'status' is changed into 'true' metadata, i.e. they now become subject of a document. If required the candidate can be removed.
Table 26: Update Statements for Selected Documents
lists update statements that could be executed after a document has been selected188.

[188] Due to the handover of arguments (?arg1, ?arg2) the listed statements are not executable as SPIN rules.

Update Statement	Remark

CONSTRUCT {
?arg1 ahsga:AHSGA_DocumentHasStatus "selected" .
?arg1 extenteo:documentHasReportedDate ?edate. }
WHERE {
?arg1 a eo:AHSGA_Document .
BIND (afn:now() AS ?sdate) .
BIND (spif:dateFormat(?sdate, "dd.MM.yyyy.hh:mm") AS ?ddate) .
BIND (spif:parseDate(?ddate, "dd.MM.yyyy.hh:mm") AS ?edate) .
}
	Creates AHSGA_document properties (document status and reported date) for selected documents;
	variable ?arg1 is used to indicate the handover of selected documents

CONSTRUCT
{ ?arg1 ahsga:AHSGA_DocumentHasStatus "closed" }
WHERE {
?arg1 a eo:AHSGA_Document .
NOT EXISTS {?arg1 ahsga:AHSGA_DocumentHasStatus "closed" }
}
	Creates a AHSGA_document property (document status) for selected documents if a task has been closed;
	variable?arg1 is used to indicate the handover of selected documents

CONSTRUCT
{?arg1 extenteo:documentHasSubjectBusinessActor ?BA }
WHERE {
?arg1 a eo:AHSGA_Document .
?arg1 ahsga:AHSGA_DocumentHasStatus "selected" .
?arg1 eo:documentIsAssociatedToContext ?BA.
?BA eo:businessActorHasAssignedBusinessRole ?BR .
?BR a ahsga:AHSGA_Client .
?arg2 a eo:LegalEntity
FILTER (?BA = ?arg2)
}
	Creates a AHSGA_document property (document has subject) for selected documents if a string parsed from the reported “name/title” matches with an instance of the eo:enterpriseObjectHasName property of a business function and this business function has been associated to the selected document(s) as a context element (level 1 or level 2);
	variable?arg2 is used to indicate the handover of parsed strings

Update Statement	Remark

CONSTRUCT
{?arg1
extenteo:documentHasSubjectBusinessFunction
?BF }
WHERE {
?arg1 a eo:AHSGA_Document .
?arg1 ahsga:AHSGA_DocumentHasStatus
"selected" .
?arg1 eo:documentIsAssociatedTo3Context ?BF
.
?BF a ahsga:AHSGA_BusinessFunction .
?arg2 a ahsga:AHSGA_BusinessFunction .
FILTER (?BF = ?arg2) .
}
	Creates a AHSGA_document property (document has subject) for selected documents if a string parsed from the reported "name/title" matches with an instance of the eo:enterpriseObjectHasName property of a business function and this business function has been associated to the selected document(s) as a context element (level 1 or level 2);
	variable?arg2 is used to indicate the handover of parsed strings

CONSTRUCT
{?arg1
extenteo:documentHasSubjectBusinessFunction
?BF }
WHERE {
?arg1 a eo:AHSGA_Document .
?arg1 ahsga:AHSGA_DocumentHasStatus
"selected" .
?arg1 eo:documentIsAssociatedToContext ?BF.
?BF a ahsga:AHSGA_BusinessFunction .
?Sch rdfs:label "Schwupro"@de .
FILTER (?Sch = ?BF)
}
	Creates a AHSGA_document property (document has subject) for selected documents if the particular business function 'Schwupro' is reported, and this business function has been associated to the selected document(s) as a context element (level 1 or level 2)

Table 26: Update Statements for Selected Documents

More Updates were triggered by SPIN rules (e.g. AHSGA_IR15); cf. Chapter 6.1.4 for details.

6.1.6 Summary of AHSGA Action Research Loop 2

As shown in the previous sections the chosen modeling procedure has been applied successfully in the Action Research study with AHSGA. Additionally, the actions defined in loop 1 have been performed. In agreement with the Action Research partner it was decided to go without the integration of external information on movies and TV production but to stay focused on business documents. Therefore the use of AHSGA's Task and Information Recording System has been investigated in detail as explained above.
Action Research Loop 2 focuses on creating a model to visualize and discuss improvements of the approach with Action Research team. Thus, a demonstrator has been developed that could be presented to the Action Research partner for evaluation, to capture questions to be answered, change requests and supplementary requirements. For the demonstrator not all components are automated yet and their execution is performed 'unlinked', i.e. each component is started manually, and the output result is taken as input for the next component.

6.1.6.1 Results of the Second Loop of Action Research With AHSGA

The second loop of the study was executed between March 2011 and July 2012. Within this loop four meetings took place: first on July 29th, 2011, second on December 26th, 2011, third on April 5th, 2012, and the fourth one on June, 18th, 2012.
In the following, results of the second iterative cycle are provided as specified within the Action Research method (cf. Chapter 2.2.3).

1. Presentation and evaluation of the demonstrator
   As defined in the prototyping method (cf. Chapter 2.2.4) within loop 2 a demonstrator is developed. Figure 77 depicts the components used to build AHSGA's demonstrator. Boxes with the icon of hands indicate that for the demonstrator those functions are performed manually. Hatched arrows indicate that the components are performed unlinked. Below the boxes the tools are shown, which were used to perform the component.

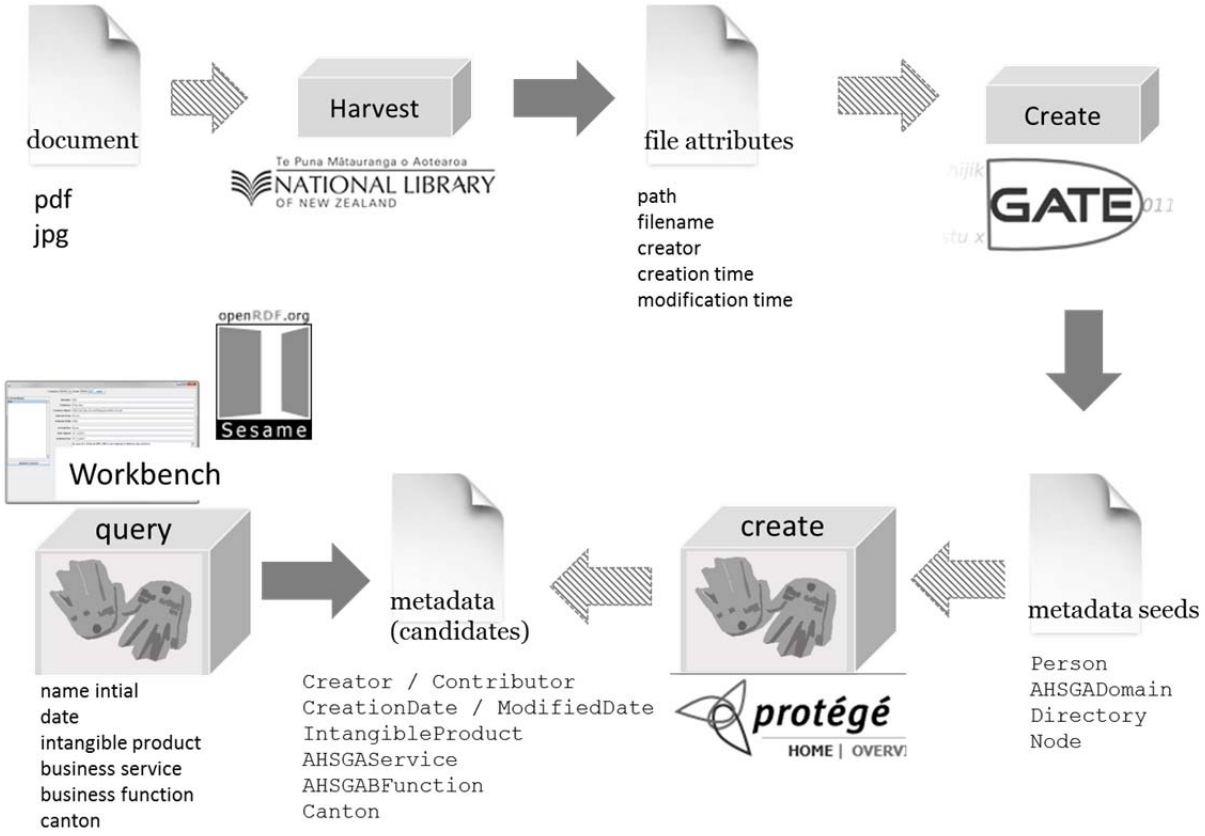


Figure 77: AHSGA Demonstrator Components

The demonstrator has been set up for two document formats (pdf and jpg). First, document properties are harvested using the Metadata Extraction Tool provided by the National Library of New Zealand (NLNZ). Harvested results are stored in xml-format. After that the xml-file is imported into GATE[189] to create additional metadata seeds from document properties.

[189] General Architecture for Text Engineering (GATE) is an open source free software, developed since 15 years and providing a comprehensive set of functions and plugins. URL: http://gate.ac.uk/overview.html (retrieved: 27.8.2011)

For the demonstrator metadata seeds derived from the document properties 'path' and 'filename' are processed. As described in section 6.1.5.2, simple text analysis is performed, to split for example the pathname (e.g. C:\TEXT\AH_GS\DIREKTEP\HEIME JUG\Auboden) into its segments or to separate the employee ID from the filename (e.g. the digit '4' in 409_Homosexualitaet in Heimen.pdf[190]). To do so some other modules of the GATE family are used, namely ANNIE English Tokeniser, ANNIE POS Tagger, GATE Morphological Analyser and RegEx Sentence Splitter (Cunningham et al., 2011). The parts build the values of the metadata seeds 'documentHasSeedEmpNo', 'documentHasSubject', and 'documentIsStoredInDirectory'.
Based on the metadata seeds the respective instances in the ontology are created. After creating metadata seeds metadata candidates are inferred. As indicated in Figure 77 for the demonstrator that step is performed manually.
As usual in software development the new functionality has not been implemented into the production system (ITRS) but instead a simple workbench has been developed that simulated recording data in AHSGA's Information and Task Reporting System and querying seEAD. The workbench is called Metadata Generation (MeGa) Workbench and data is entered via a graphical user interface as depicted in Figure 78. Contrary to ITRS all data is entered on one panel and for the demonstrator product specific data (as shown in Figure 75 for the product 'Prevention') has been omitted. As the Action Research partner's business language is German values of data elements are in German191. On the left hand side of the figure the data fields are depicted that must be entered for every task reported in the ITRS (cf. Table 24). Content of these fields is used to query seEAD. In the lower part of Figure 78 the retrieved documents are displayed. On the right hand side the metadata for a retrieved and selected document is shown. In the demonstrator these metadata cannot be modified, in a real implementation this should be possible.

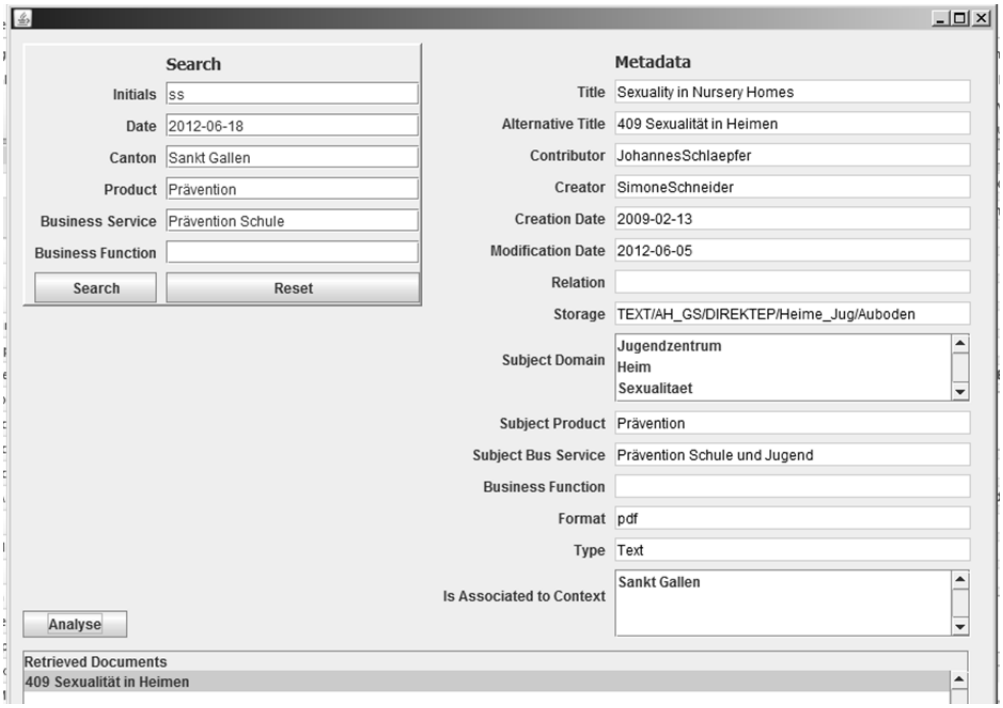


Figure 78: GUI of the MeGaWorkbench for AHSGA

[190] '4' is the employee ID, '09' is the year, the resource has been created
[191] For all instances related to AHSGA the following naming convention is chosen: instance names – if not automatically generated as for 'document' – are given in German language; labels are defined in German and English.

Figure 78 illustrates the metadata of the example document that matches with the ITRS elements used for retrieval (cf section 6.1.5.3). Creator of the Document is '`Simone Schneider`' (derived from the seed '`documentHasSeedEmpNo`'), intangible product is '`Prevention`' (derived from information about the directory), and business services is '`School and Juvenile Prevention`' (derived from information about the directy). Associated context in this example is the canton (`Saint Gallen`), which has been inferred from the information about the directory, which is named after a location (`Auboden`).

2. Captured questions to be answered
   As required by the Action Research method (cf. Chapter 2.2.3) questions that arose during demonstrator development are to be answered in collaboration with the Action Research partner. In AHSGA's case the questions were related to retrieval and to rules for metadata generation. Since the retrieval questions were about usability, ranking of results, etc. but do not contribute to my research I waive discussing them here.
   In the following the relevant aspects are briefly summarized.
   - Since many documents like images were downloaded from the internet or imported from an image library, the file names are often simply a number of characters and numbers (e.g. P1010435.jpg) which does not have a meaning for AHSGA and therefore does not allow for harvesting metadata candidates. In this case the author is either anonymous or without relevance for metadata generation as there is no relation to an enterprise object. Thus, in these cases neither a creator (inferred from the employee no given in the file name) nor a contributor (extracted from the author document property) can be determined. Nevertheless, these documents are of relevance and should be retrieved. As with my approach this is possible, e.g. via infered metadata like product, canton, these documents should appear on the hit list with low priority.
   - All business objects needed for AHSGA's metadata generation are currently represented in seEAD.
   - The correlations between low level governance instruments and generated metadata are correct and sufficient.

3. Captured change requests and supplementary requirements
   All nouns, extracted from a document's title are considered terms of the domain and stored as metadata element `documentHasSubjectDomain`. However, the filename could also imply information about the customer group (e.g. juvenile homes), the customer (Juvenile Home Auboden) and the customer's location. As with the demonstrator these specific metadata are not generated the inferrencing rules should be enhanced.

4. Actions to overcome or test the problem and to adapt the demonstrator to enterprise specific requirement
   According to Action Research method (cf. Chapter 2.2.3) the following aspects should be considered in order to improve the final prototype that will be evaluated in loop 3:
   - The demonstrator showed that AHSGA's document storage does not always comply to their directives. For example, images were stored in an undefined 'picture folder' instead of in the required directory. AHSGA will clean-up their documents' storage and eventually move documents to other folders according to their governance guidelines.

- Adapters provided out of the box by NLNZ are almost sufficient with respect to supported file formats and extracted attributes. No new adapters are needed but existing have been slightly adjusted. For all file formats considered by AHSAG (doc, ppt, xls, pdf, jpg, gif, png, mp3, mp4) the provided adapters have been modified in order to harvest the modification date attribute. All adapters are used for the prototype, which is evaluated in Action Research loop 3 (cf. Chapter 0).
- For operational use of my approach the application programming interface (API) of ITRS is to be checked for possibilities to document management, i.e. searching seEAD for documents via the ITRS GUI and importing selected document into the ITRS.

5. Share with others (departmental meeting, publication, conference, etc.)
   To share results of the second loop of Action Research with others the demonstrator (second version of prototype) was presented and discussed with the Action Research team on June, 18th, 2012.

Findings of the second loop of Action Research with AHSGA:

- As depicted in the overview of AHSGA's context model (Figure 74), content of AHSGA's business governance instruments can be completely and sufficiently represented in seEAD.
- The ArchiMate standard is appropriate for representing AHSGA's document model but is enhanced by the Dublin Core standard. Furthermore, application specific refinements are made, e.g. `eo:documentHasDocumentSeedingDate` is a refinement of `elements:documentHasDate`.
- Information kept in the ITRS has been analysed and compared to the information stored in seEAD in order to match the data entered in ITRS with the metatdata for retrieval. Correlations are provided in Chapter 6.1.5.5.

6.1.6.2 Research Questions Addressed Within the Second Loop of AR With AHSGA

As proposed in my research design (cf. Chapter 2.1, Table 1) within the second loop of Action Research with AHSGA several research questions were addressed.

1. Context entities that can be inferred to automatically generate descriptive metadata, (answer to RQ3) are the following: `Product`, `BusinessService`, `BusinessBehaviourElement`, `BusinessRole` and `BusinessActor` (cf. Chapter 6.1.2).

2. To determine what rules must be defined for metadata generation using logical inference (RQ5) three constraints – defined by AHSGA – must be met:
   - no additional tool for document management is allowed and thus primarily metadata that can be retrieved via the ITRS interface matters
   - rules must mirror enterprise's governance guidelines
   - rules must effectuate law for document archiving.

   The approach is explained by the following example: AHSGA's guideline says that every file name should start with a one digit representing the author of the document (e.g. '4'). The digit is extracted from the filename to build a metadata seed (`documentHasSeedEmpNo`). From this seed the creator can be inferred (e.g. 'Simone Schneider') and the acronym of the person's name (e.g. 'ss'). This is important as the

acronym is a data element of ITRS, which is used to query seEAD for the creator or contributor of documents. Although possible, the organizational unit the person belongs to could be inferred accordingly, this is not done as it is not relevant for retrieval (cf. Chapter 6.1.5.4).
In order to manage documents compliant to law rules that infer the applicable principles and appropriate retention period are defined. This approach allows for restricting the automatic generation of metadata to what is needed.

3. The documents attached to a task which is reported in ITRS must be considered business relevant and thus, must be archived according to the law – as already said. In addition to law generally binding for business relevant documents, specific regulations must be obeyed depending on the business sector. Health care and finance for example have particular requirements with respect to information security. Since in my approach the context of documents is known, a customer related to a document can be inferred. If the customer is not an individual but a legal entity the industry sector can also be inferred. Hence additional requirements for archiving can be determined. Automatically generated metadata support document lifecycle management as it allows the identification of the documents that must be archived and the law with which archiving must comply (answer to RQ6).

4. RQ8, RQ10 and RQ11 are concerned with enterprise architecture. As shown, the ArchiMate standard identifies the enterprise objects which constitute an enterprise architecture (answer to RQ8) and provides also its structure as shown in Figure 14 (answer to RQ10). For AHSGA the modifications of ArchiMate to become ArchiMEO (cf. Chapter 5.1.3) were sufficient, i.e. no changes on the core ontology have been necessary. AHSGA characteristics could be represented in the enterprise specific part of seEAD.
For example: the data properties `eo:documentHasDocumentSeedingDate` and `ahsga:personHasNameInitials` have been added to address application specifics. This enhancement does not affect ArchiMEO as no changes were to be made but simply extended the enterprise specific part of AHSGA's semantically enhanced enterprise architecture description.
Thus, answer to RQ11 is given, as ArchiMEO – based on ArchiMate – is general enough to be used 'out of the box' but customizable to specific companies's needs by enhancing the core ontology to an enterprise specific ontology, i.e. to seEAD.

5. First proof of concept that representing enterprise architecture in an ontology is possible and appropriate is given by AHSGA's demonstrator (answer to RQ9). It has been shown, that context information, e.g. information about a customer's location, represented in seEAD can be inferred automatically to generate metdata candidates.

6. Phases 1 and 2 of the procedure model I introduced in Chapter 5.3 have been successfully applied in the Action Research study with AHSGA. As detailed in the previous sections all required deliverables have been created and validated against the demonstrator. Thus, the first two phases of the procedure model have been proved appropriate for setting-up, conducting and utilizing metadata (answer to RQ14).

Results of the third loop of the Action Research study with AHSGA are provided within Chapter 8.2.1, p 240.

6.2 *Symfact Application Profile*

Also for Symfact the two questions asked by Dietz (2006, p 6) - "Why and how would enterprise ontology assist in coping with current and future problems related to enterprises?" and "Why would this approach be more appropriate and more effective than some other one?" – have been answered. Following again the mintProcedure Model introduced in Chapter 5.3 first the motivation scenario is briefly recapitulated, after that informal competency questions for Symfact were phrased, too, to determine whether the proposed seEAD is required and appropriate or whether the competency questions could be already be answered by an existing approach. After that the formal terminology is expressed in Symfact's context model, then the informal competency questions are transformed into formal ones and formal axioms are defined. Finally the completeness theorem is verified, i.e. if generated metadata satisfy the enterprises' information need. This has been done within Action Research Loop 2.

The motivating scenario for Symfact is as follows: contract lifecycle management needs to be improved with respect to records management and obligation management. Up to now Symfact's system supports only time related triggers for obligations, e.g. the periodic delivery of reports. In case of unforeseeable events, like force majeure events (natural disasters, war, riots, etc.) or business events (outsourcing, merger, bankruptcy, etc.) affected obligations cannot be identified automatically or in some cases at all, nor can an update or delete of the respective contracts be enforced.

6.2.1 Symfact Informal Competency Questions

In the following, informal competency questions are phrased to determine whether the proposed seEAD is required and adequate. The questions have been drawn from knowledge I gained from my work within the DokLife project.

1) Given the role of a business partner (contractor, supplier, customer, etc.)
   AND
   some constraints regarding a force majeure event of type act of God (natural disasters like earthquakes, hurricanes, floods, etc.) or men made (wars, riots or other major upheaval events)
   which business actors are affected ?

2) Given a business actor that is a certain type of corporate body (legal entity, organizational unit, person)
   AND
   some constraints regarding business role (supplier, consumer, client, etc.) and business relationship (contract, maintenance, advising, etc.)
   which business objects (contracts, obligations, etc.) are affected?

3) Given a business actor is affected by a business event (merger, bankruptcy, injunction, etc.)
   AND
   some constraints regarding business relationships or business objects (effectivity, type, etc. of a contract)
   which obligations are due and in which contract documents are they reported?

4) Given a legal entity is affected by a business event (merger, bankruptcy, injunction, etc.)
AND
some constraints regarding business role (business relationship, business contract partner)
what obligations are due?

5) Given a legal entity is affected by a business event (merger, bankruptcy, injunction, etc.)
AND
some constraints regarding business actors location (within or outside of Europe) or business roles (contract type, up-to-dateness, etc.)
what obligations are due?

6) Given a business object (contract, obligation, etc.) is affected by an event (force majeure or business event)
AND
some constraints regarding business actors (contractee, contractor) and representations (regulation document, contract document, etc.)
what document realizes the business object?

7) Given a business actor (legal entity, person etc.) is affected by an event (force majeure or business event)
AND
some constraints regarding obligations (force majeure, business event, etc.) and the role of the business partner (collaborative development, consultancy, etc.)
which documents realizes the obligations?

8) Given a business relationship (supply, advice, etc.)
AND
some constraints regarding events (force majeure or business events)
what type of obligation is due (to report, to notify, to prove, etc.)?

9) Given the occurrence of some business events (disaster, bankruptcy, injunction, etc.) or force majeure events (act of god or men made) and affected business objects
AND
some constraints regarding their representations (documents with a certain status e.g. signed, workInProgress, etc.)
which documents must be updated?

10) Given a document represents a business object (contract, report, etc.) and is associated to a business actor (legal entity, organisational unit, person)
AND
some constraints regarding location (Switzerland, Japan, etc.) or industry sector (bank, trading, etc.)
what law and regulations determine the retention period for that document?

None of those questions can be answered today with Symfact's existing Contract Management System. For providing answers the context of contracts must be known, as modeled in seEAD.

6.2.2 Symfact Context Model

Also in Symfact's case the Context Model is based on ArchiMEO, and thus on ArchiMate. Again, the definition of the formal terminology hasn't been done from scratch; instead existing concepts were enhanced were appropriate and reviewed by the Action Research Partner within Loop 2 of the study.

Table 27 gives an example of objects, which are instances of Symfact's domain and the corresponding concepts that exist in seEAD. All references to ArchiMate in this chapter are based on the ArchiMate 1.0 Specification (The Open Group, 2009b).

Instances	Concepts	Remark
DokLife_Document_35	`Document`	`Document` isA sub-concept of `Representation` (an ArchiMate concept)
contractor, supplier, customer, etc.	`BusinessRole`	as exists
earthquakes, hurricanes, floods, etc. wars, riots or other major upheaval events, etc.	`ForceMajeureEvent`	`ForceMajeureEvent` is a sub-concept of `Event`(a top-concept that has been introduced to seEAD and that is not an ArchiMate concept)
merger, bankruptcy, injunction, etc.	`FinancialBusinessEvent`	`FinancialBusinessEvent` is a sub-concept of `BusinessEvent` (an ArchiMate concept)
GiveMeFive, Symfact, DontWorryInsurance, etc.	`LegalActor`	`LegalActor` isA sub-concept of `BusinessActor` (an ArchiMate concept)
business consultancy, collaborative development	`BusinessCollaboration`	as exists; `BusinessCollaboration` is a sub-concept of `BusinessRole` (an ArchiMate concept)
contract obligation	`Contract`	as exists; `Contract` isA sub-concept of `BusinessObject` (an ArchiMate concept)
	`Obligation`	`Obligation` isA sub-concept of `BusinessObject` (an ArchiMate concept)
Japan, Switzerland, Zuerich, etc.	`Location`	`Location` is a top-concept that has been introduced to seEAD and that is not an ArchiMate concept

Table 27: Excerpt of Instances Derived From Symfacts's Competency Questions

After all objects were derived form the competency questions and checked how they can be represented in seEAD their properties were determined. Table 28 gives an example of some object properties derived from Symfact's Competency Questions.

Instances	Property	Remark
DokLife_Document_35	documentHasContributor	property is an enhancement to ArchiMate for Dublin Core Elements
	documentHasSubjectContract	property is a refinement of the Dublin Core Element 'subject'
	documentHasSubjectObligation	property is a refinement of the Dublin Core Element 'subject'
contractor, supplier, customer, etc.	businessRoleIsAssignedToBusinessActor	the property conforms to the ArchiMate association between Business Role and Business Actor
earthquakes, hurricanes, floods, etc. wars, riots or other major upheaval events, etc.	forceMajeureEventIsOfType	ActOfGod and ManMade are possible types
GiveMeFive, Symfact, DontWorryInsurance, etc	legalEntityBelongsToLegalEntity	the property is a foaf property[192]
	businessActorIsSituatedInLocation	property is AHSGA-specific
	legalEntityDescribedByGICS	the property conforms to the ArchiMate association between Business Actor and Business Actor
contract	agreesUponObligation	the property conforms to the ArchiMate aggregation between Contract and Business Object
	contractIsRealizedByDocument	the property conforms to the ArchiMate relation 'Contract Realization Representation'
	contractHasContractee	the property conforms to the ArchiMate association between Contract and Business Actor
obligation	obligationIsPerformedByBusinessActor	the property conforms to the ArchiMate association between Business Object and Business Actor
	obligationHasObligationDescription	e.g. "GiveMeFive reports damage by ForceMajeurEvent"
	obligationHasCondition	e.g. "ForceMajeureEvent"

Table 28: Extract of Object Properties Derived From Symfact's Competency Questions

[192]FOAF Vocabulary Specification 0.98. URL: http://xmlns.com/foaf/spec/#term_Person (retrieved: 25.7.2012)

Figure 79 gives an overview on context elements relevant for metadata generation for Symfact. As already shown for AHSGA, a `Document` in seEAD is considered a specialization of a `Representation` (2), which realizes a `BusinessObject` (3); a `ContractDocument` again is a specification of a document (1). `Contract` (4), `Obligation` and (5) `Product` (6) are specifications of `BusinessObject` and according to ArchiMate interlinked via `association` relations, respectively specifications of it. A `LegalEntitiy (7)` is a specification of `BusinessActor.` The relation of `Contract` and `LegalEntitiy` again is a specification of `association`. A `BusinessActor` has a `BusinessRole` (8) assigned; that can be a single role, like `ContractPartner` or a collaborative role like `BusinessRelationship` As the location is important for a contract (e.g. as place of jurisdiction) but also for the contract *document* (to archive it compliant to law and regulations), the top level concept `Location` and its specifications (9) are required. Although not needed for metadata *generation* but for contract management the `BusinessEvent` (10) and its specifications are depicted in the figure.

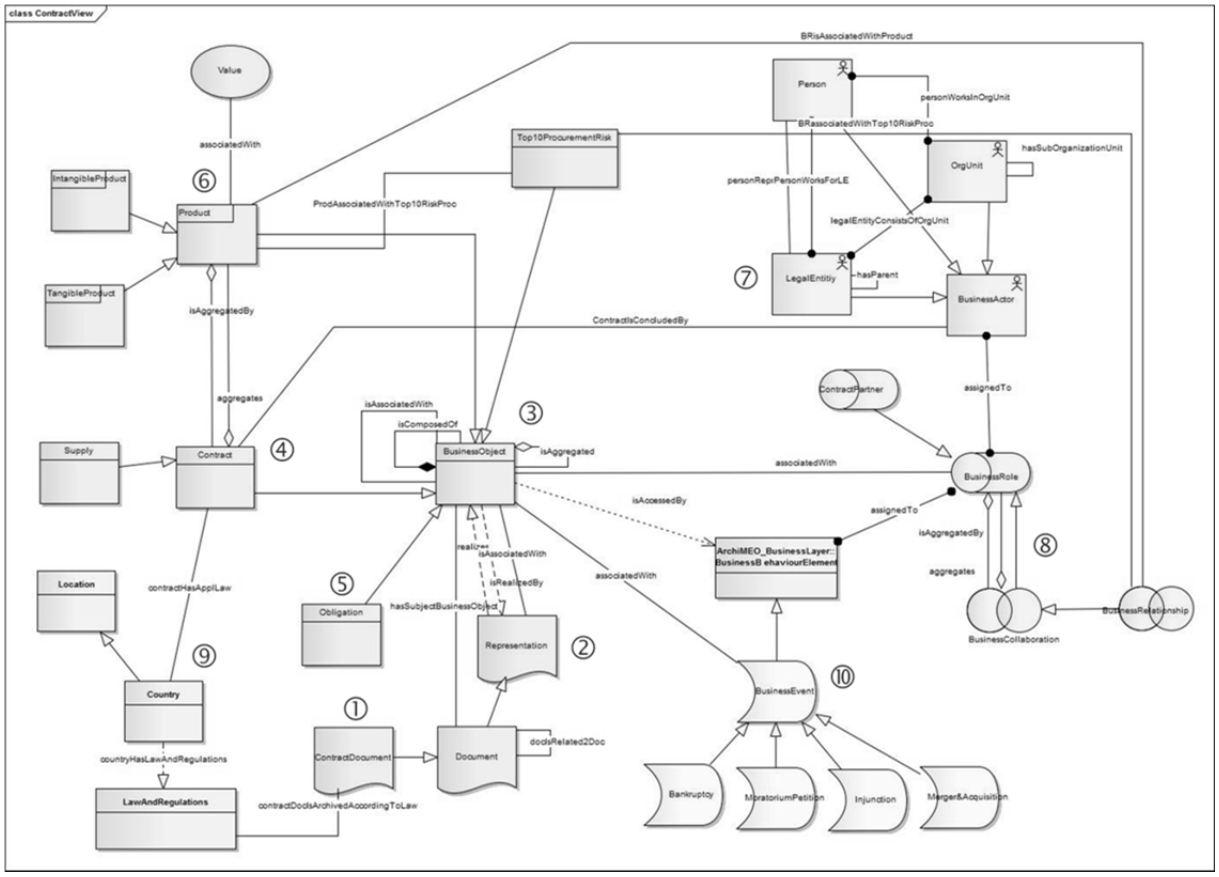


Figure 79: Symfact Context Model Overview

Same as for AHSGA, all concepts and properties (data propterties and relations between concepts) are formally represented in seEAD.

6.2.3 Symfact Formal Competency Questions

After Symfact's terminology has been specified competency questions were expressed formally.

Table 29 gives an example of two more complex competency questions (7 and 10) rewritten in SPARQL. Below each query statement an excerpt of the result set is listed.

QNo	Informal Question	SPARQL Query

7	Given a **business actor** (legal entity, person etc.) is affected by an **event** (force majeure or business event)
	AND
	some constraints regarding **obligations** (force majeure, business event, etc.) and the role of the business partner (collaborative development, consultancy, etc.)
	what documents realizes the obligations?
		SELECT ?BusPartner ?obligation ?description ?contract ?doc
		WHERE {
		?event a eo:ForceMajeureEvent .
		?event rdfs:label "**Earthquake**"@en .
		?poC a top:PartOfCountry .
		?poC rdfs:label "Fukushima"@en .
		?BusPartner eo:businessActorHasAssignedBusinessRole ?role .
		?role rdfs:label "**Collaborative Development**"@en .
		?BusPartner eo:businessActorIsSituatedInLocation ?busLoc .
		?busLoc top:cityIsLocatedInPartOfCountry ?poC2 .
		FILTER (?poC = ?poC2)
		?obligation eo:obligationIsPerformedByBusinessActor ?BusPartner .
		?obligation eo:obligationHasCondition "ForceMajeure" .
		?obligation eo:obligationHasObligationDescription ?description .
		?contract eo:contractAgreesUponObligation ?obligation .
		?contract **eo:contractIsRealizedByDocument** ?doc
		}

Result		

[BusPartner]	obligation	description	contract	doc
eo:GiveMeFive	eo:Obligation_DokLife_6	In Case of a Force Majeure Event liabilities ar...	eo:DokLife_Contract_35	eo:DokLife_Document_35

10	Given a document represents a business object (contract, report, etc.) and is associated to a **business actor** (legal entity, organisational unit, person)
	AND
	some constraints regarding location (Switzerland, Japan, etc.) or **industry sector** (bank, trading, etc.)
	what **law and regulations** determine the retention period for that document?
		SELECT DISTINCT ?code ?law ?regdoc ?source
		WHERE {
		?contract a archi:Contract .
		?contract eo:contractHasContractor ?contractor.
		?contractor rdfs:label "**Symfact PLC**"@en .
		?conractor eo:legalEntityDescribedByGICS ?code .
		BIND (fn:substring(?code, 1, 2) **AS** ?icode) .
		?law **eo:lawAndRegulationIsAssociatedToGIGS**_Code ?icode.
		?law **eo:lawAndRegulationIsExpressedInRegulationDocument** ?regdocx.
		?regdocx rdfs:label ?regdoc.
		FILTER (lang(?regdoc)="en") .
		?regdocx elements:documentHasIdentifyer ?source
		}

Result		

[code]	law	regdoc	source
eo:GICS_Code_45103010	eo:SwissProductLiabilityLaw	Product Liability 221.112.944	http://www.admin.ch/ch/d/sr/2/221.112.944.de.pdf

Table 29: Symfact's Competency Questions (7 and 10) Rewritten in SPARQL

6.2.4 Symfact Rule Model

In the following the rules, which complete the formal axioms modeled as concepts and properties in seEAD, are expressed in SPIN.
Table 30 lists the SPIN rules; Symfact_IR_1 to Symfact_IR_6 infer metadata for contract documents, rule Symfact_IR_7 constructs a relation to a monitoring service, rule Symfact_IR_8 generate an instance of legal entity if none exits and rules Symfact_IR_9 to Symfact_IR_11 generate relations between legal entities, i.e. contract partners. The rules are listed completely.

SPIN Rule	Remark

CONSTRUCT {
?s extenteo:contractDocumentHasSubjectContract ?contract .
?s eo:contractDocumentHasSubjectObligation ?obligation .
}
WHERE {
?s eo:contractDocumentRealizesContract ?contract .
?s eo:contractDocumentRealizesObligation ?obligation .
}
	Symfact_IR_1
	Creates refined Dublin Core elements for subject

CONSTRUCT {
?this eo:documentIsArchivedAccordingToLaw ?law .
?this eo:documentIsArchivedAccordingToRegDoc ?lawdoc .
}
WHERE {
?this eo:contractDocumentRealizesContract ?contract .
?contract eo:contractHasApplicableLaw ?location .
?location eo:locationHasLawAndRegulation ?law .
?law a eo:GeneralArchivingLaw .
?law eo:lawAndRegulationIsExpressedInRegulationDocument ?lawdoc .
}
	Symfact_IR_2
	Determines law generally considered for contract archiving based on the location indicated in the applicable law metadata

SPIN Rule	Remark

CONSTRUCT {
?this eo:documentIsArchivedAccordingToLaw ?law .
}
WHERE {
?contract eo:contractHasContractor ?le .
?le eo:legalEntityDescribedByGICS ?gics .
?gics eo:GICSHasCode ?code .
BIND (fn:substring(?code, 1, 2) **AS** ?icode) .
?y eo:GICSHasCode ?icode .
?law eo:lawAndRegulationIsAssociatedToGIGS_Code ?y .
?contract eo:contractIsRealizedByDocument **?this** .
}
	Symfact_IR_3
	Determines law considered specifically for contract archiving based on the industry sector, represented as GICS-Code[193], the contractor belongs to

CONSTRUCT {
?this extenteo:documentIsStoredInFilesystem ?fsystem .
}
WHERE {
?this a eo:DokLife_Document .
?fsystem a eo:Symfact_Filesystem .
}
	Symfact_IR_4
	Specifies the location of contract document; in a productive environment were Symfact's CLM system is connected with seEAD this rule is obsolete and instead an unique contract document identifier would be used

CONSTRUCT {
?this extenteo:documentHasArchiveDate ?year .
}
WHERE {
?this a eo:DokLife_Document .
?this eo:contractDocumentHasStatus "expired" .
OPTIONAL {
?s extenteo:documentHasArchiveDate ?xdate .
} .
FILTER (!bound(?xdate)) .
BIND (afn:now() **AS** ?sdate) .
BIND (spif:dateFormat(?sdate, "dd.MM.yyyy") **AS** ?ddate) .
BIND (fn:substring(?ddate, 7) **AS** ?year) .
}
	Symfact_IR_5
	Creates the year archiving must start after a contract has expired

[193] Global Industry Classification Standard (GICS) is a collaboration between Standard & Poor's andMorgan Stanley Capital International. GICS codes are 8-digit codes that correspond to various businessor industrial activities, such as Oil & Gas Drilling or Wireless Telecommunication Services. URL: http://be.ncue.edu.tw/compustat/manual/globdata/Part3d.pdf (retrieved: 22.7.2012).

SPIN Rule	Remark

CONSTRUCT {
?this eo:documentHasArchivingTime ?maxperiod .
}
WHERE {
{
SELECT ?this ((**MAX**(?period)) **AS** ?maxperiod)
WHERE {
?this eo:documentIsArchivedAccordingToLaw ?law .
?law eo:lawDefinesArchivingPeriod ?period .
}
GROUP BY ?this
} .
}
	Symfact_IR_6
	Infers the total archiving period for a document
	Rule is also valid for AHSGA's documents (cf. AHSGA_IR_20)

CONSTRUCT {
?this eo:objectManagedByService ?task
}
WHERE {
?this eo:contractHasStatus "invalid" .
?task rdfs:label "Track Obligation"@en .
}
	Symfact_IR_7
	Creates a relation between a contract and a service that monitors contracts which are not valid any more

CONSTRUCT {
?lentity a eo:LegalEntity .
}
WHERE {
?contract a eo:DokLife_Document .
?contract elements:documentHasContributor ?lentity .
NOT EXISTS {
?lentity a eo:LegalEntity .
} .
}
	Symfact_IR_8
	Creates an instance of a legal entity for contract partners mentioned in a contract if none exists

CONSTRUCT {
?contract eo:contractEstablishesBusinessCollaboration ?BR
}
WHERE {
?contract eo:contractHasContractType "CDA" .
?BR eo:roleHasCharacteristics "collaboration" .
}
	Symfact_IR_9
	Creates a relation between business partners for a specific type of contract expressed in a contract document

SPIN Rule	Remark

CONSTRUCT {
?contract
eo:contractEstablishesBusinessCollaboration ?BR
}
WHERE {
?contract eo:contractHasContractType "NDA" .
?BR eo:roleHasCharacteristics "consultancy" .
}
	Symfact_IR_10
	Creates a relation between business partners for a specific type of contract expressed in a contract document

CONSTRUCT {
?this eo:documentIsUsedInTask ?task .
}
WHERE {
?this eo:contractDocumentHasStatus "affected"
.
?task a bpmn:ServiceTask .
?task rdfs:label "Monitoring Task"@en .
}
	Symfact_IR_11
	Creates a relation between business partners for a specific type of contract expressed in a contract document

Table 30: SPIN Rules for Symfact

6.2.5 Symfact Description Set Profile

In Symfact's case documents are managed with the company's Contract Lifecycle Management (CLM) system. To avoid redundancy in seEAD only this metadata is considered relevant either for records management (e.g. `CompulsoryArchiving`) or for contract and obligation management that a contract document *represents*. Table 31 provides the according minimal subset of metadata elements for contract documents stored in seEAD.

DCME No	DCME	Property	Refinement	Syntax Encoding Scheme	Value Encoding Scheme
1	Contributor	`dc:documentHasContributor`			`archi:BusinessActor`
4	Date	`dc:documentHasDate`	`dceo:documentHasArchiveDate`	http://www.w3.org/TR/NOTE-datetime	
13	Subject	`dc:documentHasSubject`	`dceo:documentHasSubjectContract`		`eo:contract`
13	Subject	`dc:documentHasSubject`	`dceo:documentHasSubjectObligation`		`eo:obligation`

DCME No	DCME	Property	Refinement	Syntax Encoding Scheme	Value Encoding Scheme
	Non DC Metadata Elements				
	Storage	dceo:documentHasStorage			archi:node
	Status	eo:ContractocumentHasSatus			
	Compulsory Archiving	eo:documentIsArchivedAccordingToLaw			eo:LawAndRegulation
	Regulation Document	eo:documentIsArchivedAccordingToRegDoc			eo:RegulationDocument

Table 31: Symfact Metadata Element Set

In the following the enterprise specific characteristics of automatic metadata generation for Symfact are detailed, based on the general approach introduced in Chapter 5.2.3.

6.2.5.1 Symfact Content Annotations – Metadata Seeds

In the Symfact case I found the happy situation that a system exists which extracts information from text documents, developed within the DokLife project. The metadata generation starts with the subset of extracted – and already annotated information – that is relevant for further metadata creation. The subset has been drawn from knowledge I gained within the DokLife project and has been reviewed by the Action Research Partner within Loop 2 of the study. Table 32 provides the metadata seeds created on content annotations. Column three indicates the property type: lighter grey is a data property and darker grey an object property.
Note that only the first four annotations depicted in Table 32 are metadata seeds of the *contract document* itself. All other annotations become instances of the *business objects* they represent, which in this case are either contract or obligation. These instances are considered seeds of primary context elements of the contract document. For better readability namespaces are omitted.

Example	Content Annotation		Metadata Seed	
Symfact	InternalParty		documentHasContributor	LegalEntity
GiveMeFive	ExternalParty		documentHasContributor	LegalEntity
Symfact_Software	ContractObject		documentRealizesContract	Contract

Example	Content Annotation		Metadata Seed	
Obligation_DokLife3 ; Obligation_DokLife4 ; Obligation_DokLife5 ; Ogligation_DokLife6	Obligation		`documentRealizesObligation`	`Obligation`
			Seeds of Primary Context Elements (Contract)	
Collaborative Development Agreement (CDA)	ContractType [194]		`contractHasType`	
Switzerland	ApplicableLaw		`contractHasApplicableLaw`	`Country`
12.05.2012	ContractBegin		`contractHasBegin`	
31.12.2015	ContractEnd		`contractHasEnd`	
Symfact	InternalParty		`contractHasContractor`	`LegalEntity`
GiveMeFive	ExternalParty		`contractHasContractee`	`LegalEntity`
Obligation_DokLife3 ; Obligation_DokLife4 ; Obligation_DokLife5 ; Ogligation_DokLife6	Obligation		`contractAgreesUponObligation`	`Obligation`
			Seeds of Primary Context Elements (Obligation)	
Report[195]	ObligationType		`obligationHasType`	
GiveMeFive reports damage by ForceMajeurEvent	ObligationDescription		`obligationHasObligationDescription`	
ForceMajeure	ObligationCondition		`obligationHasCondition`	
GiveMeFive	ObligationResponsibility		`obligationIsPerformedByBusinessActor`	`BusinessActor`

Table 32: Symfact's Data Source and Sink for Creating Metadata Seeds

[194] Find the complete list of contract types in Chapter 12.6
[195] Example is provided for *Obligation_DokLife6*

6.2.5.2 Symfact Metadata Seeds – Metadata

After the metadata seeds were created based on the content annotations metadata were inferred using the rules of 6.2.4 for Symfact's documents. Table 33 depicts an example of a metadata elements inferred from the metadata seeds. As mentioned above, most seeds already constitute primary context elements.

Metadata Seed / Primary Context	Instance	Secondary Context	Instance	Remark
Contract	*DokLife_Contract_35*			From Symfact specific Annotations refinements of Standard Dublin Core Metadata are derived
Inferred Metadata	Document documentHasSubjectContract Contract		*DokLife_Contract_35*	
Obligation	*Obligation_DokLife3;*			
	Obligation_DokLife4;			
	Obligation_DokLife5			
	Obligation_DokLife6			
Inferred Metadata	Document documentHasSubjectObligation Obligation		*Obligation_DokLife3;*	
			Obligation_DokLife4;	
			Obligation_DokLife5	
			Obligation_DokLife6	
Contract	*DokLife_Contract_35*	LawAndRegulation	*SwissCodeOfObligation*	relevant for contract archiving in first phase is the location annotated in 'ApplicableLaw'
			SwissArchivingLaw	
Inferred Metadata	ContractDocument contractDocIsArchivedAccordingToLaw		*SwissCodeOfObligation*	
			SwissArchivingLaw	
LegalEntity	*Symfact*	GICS_Code	*GICS_Code_45103010*	Global Industry Classification Standard (GICS)
LegalEntity	*GiveMeFive*	GICS_Code	*GICS_Code_40301030*	

Metadata Seed / Primary Context	Instance	Secondary Context	Instance	Remark
Inferred Metadata	`ContractDocument contractDocIsArchivedAccordingToLaw`		*`ProductLiability`*	Specific regulations depend on industry sector a legal entity belongs to; here: Symfact as software developer must consider law about product liability

Table 33: Symfact's Data Source and Sink for Creating Metadata Based on Primary and Secondary Context Elements

6.2.5.3 Symfact Metadata – Metadata Additions

In Symfact's case focus is on improving *contract* management and thus additional metadata are inferred for the business objects a contract document *represents*, i.e. contract and obligation.
Although these metadata do not serve as metadata candidates for contract *documents* but augment document's context elements, the creation procedure remains the same. These metadata are called metadata additions. Table 34 gives an example of the inferred metadata for business entities related to contract documents.

Primary Context	Instance	Secondary Context	Instance	Remark
`LegalEntity`	*`Symfact`*	`BusinessCollaboration`	*`CollaborativeDevelopment`*	According to ArchiMate `BusinessCollaboration` is a `BusinessRole`; `BusinessRole` is created based on `ContractType`
	`DontWorryInsurance`			
Inferred Metadata	`LegalEntity businessActorHasAssignedBusinessRole`		*`CollaborativeDevelopment`*	The metadata is created for both of the legal entities

Table 34: Symfact's Data Source and Sink for Creating Metadata Additions Based on Tertiary Context Elements

6.2.5.4 Symfact CLM Data

As today Symfact's Contract Lifecycle Management system does not deal with obligations triggered by events the CLM data elements listed in Table 35 were introduced within the Action Research study.

CLM data element	Example	Axiom in seEAD	Remark
Name of Business Partner		`Contract contractHasContractee LegalEntity`	It is assumed, that business partner is always the contractee;
Business Relations		`LegalEntity legalEntityBelongsToLegalEntity LegalEntity`	seEAD can be queried for business relations a business partner has
Business Role		`LegalEntity businessActorHasAssignedBusinessRole BusinessCollaboration`	seEAD can be queried for a role a business partner has

Event Date	*11. März 2011*	`Contract contractHasBegin`
		
		`Contract contractHasEnd`
			It is checked whether a contract is valid (its begin is before/at the same date as the event date and its end is after the event date)

Event	*Earthquake*	`Event isA Event`
		
		optional
		
		`Event isA BusinessEvent`	

Event Location	*Iwaki*	`LegalEntity businessActorIsSituatedInLocation Location`	In case of a non-business event seEAD is queried for a business partner situated in the area an event takes places; depending on the requirements the query can be broadened to a country or narrowed to a place

Table 35: CLM Sample Data

Event data is transformed into query terms seEAD is retrieved with. Refer to Chapter 8.2.2 on details related to the Symfact Prototype.

6.2.5.5 Symfact Metadata Update

To improve contract lifecycle management metadata might be updated in case of an event. Table 36 gives an example of possible changes.

Concept	Instance	Metadata	Instance	Example / Remark
`ContractDocument`	*DokLife_Contract_35* *DokLife_Contract_123*	`contractDocumentHasStatus`	*`affected`*	In case a force majeure event, e.g. earthquake, has been inferred document status is set to ‘affected’

Inferred Metadata	`ContractDocument contractDocumentHasStatus`		*affected*	
`ContractDocument`	*DokLife_Contract_999*	`contractDocumentHasStatus`	*expired*	In case business event, e.g. bankruptcy, has been inferred document status is set to 'expired'
Inferred Metadata	`ContractDocument contractDocumentHasStatus`		*expired*	
`ContractDocument`	*DokLife_Contract_999*	`DocumentHasArchiveDate`	*2012*	If document status is 'expired'start of retention period is inferred (which is the end of the calendar year)
Inferred Metadata	`ContractDocument documentHasArchiveDate`		*2012*	

Table 36: Symfact Metadata Updates

Table 37 lists update statements that could be executed after a document has been selected[196].

Update Statement	Remark

CONSTRUCT {
?arg1 eo:contractDocumentHasStatus "affected" .
?contract eo:contractHasStatus "affected" }
WHERE {
?arg1 a eo:DokLife_Document .
?arg2 rdfs:subClassOf eo:Event .
?arg1 eo:contractDocumentRealizesObligation ?obligation .
?arg1 eo:contractDocumentRealizesContract ?contract .
?obligation eo:obligationHasCondition "ForceMajeure" .
}
	Creates a DokLife_document property (document status) and a DokLife_contract property (contract status) for documents affected by an event;
	?arg1 variable is used to indicate the handover of affected documents; ?arg2 variable indicates the type of event

[196] Due to the handover of arguments (?arg1, ?arg2) the listed statements are not executable as SPIN rules.

Update Statement	Remark

CONSTRUCT {
?arg1 eo:contractDocumentHasStatus "expired" .
?contract eo:contractHasStatus "invalid" }
WHERE {
?arg1 a eo:DokLife_Document .
?arg2 a eo:FinancialBusinessEvent .
NOT EXISTS { ?arg1 eo:contractDocumentHasStatus "expired" }
?arg1 eo:contractDocumentRealizesObligation ?obligation .
?arg1 eo:contractDocumentRealizesContract ?contract .
?obligation eo:obligationHasCondition "FinancialBusinessEvent" .
}
	Creates a DokLife_document property (document status) and a DokLife_contract property (contract status) for documents affected by a business event;
	?arg1 variable is used to indicate the handover of affected documents; ?arg2 variable indicates the type of business event

Table 37: Update Statements for Affected Contract Documents

More Updates were triggered by SPIN rules (e.g. Symfact_IR7); cf. Chapter 6.2.4 for details.

6.2.6 Summary of Symfact Action Research Loop 2

A demonstrator has been developed for AHSGA and also for Symfact. The actions defined in loop 1 have been performed and main results were presented in the previous chapters. While the general procedure for metadata generation is similar to the approach for AHSGA there are some important differences:

- In contrary to AHSGA's documents, contract documents do not contain any document properties that are useful to harvest. Thus, instead of harvesting information extraction is performed. This component has been developed within the DokLife project.
- Metadata seeds are derived from a subset of content annotations created within the DokLife project.
- As the generated metadata elements are well defined, e.g. the archiving date clearly depends on facts; it is metadata that is generated not metadata *candidates*.

Action Research Loop 2 focuses on creating a model to visualize and discuss improvements of the Contract Lifecycle Management addressed within the DokLife project. For the demonstrator not all components are automated yet and their execution is performed 'unlinked', i.e. each component is started manually, and the output result is taken as input for the next component.

6.2.6.1 Results of the Second Loop of Action Research With Symfact

Second loop of the study was executed between July 2011 and June 2012. Within this loop four meetings took place: first on September 19th, 2011, second on November 17th, 2011, third on April 3rd, 2012 and the fourth one on June, 14th, 2012.

In the following, results of the second iterative cycle are provided as specified within the Action Research method (cf. Chapter 2.2.3).

1. Presentation and evaluation of the MeGaSystem demonstrator
   As defined in the prototyping method (cf. Chapter 2.2.4) within loop 2 a demonstrator is developed. Figure 80 depicts the components used to build Symfact's demonstrator. Boxes with the icon of hands indicate that for the demonstrator those functions are performed manually. Hatched arrows indicate that the components are performed unlinked.

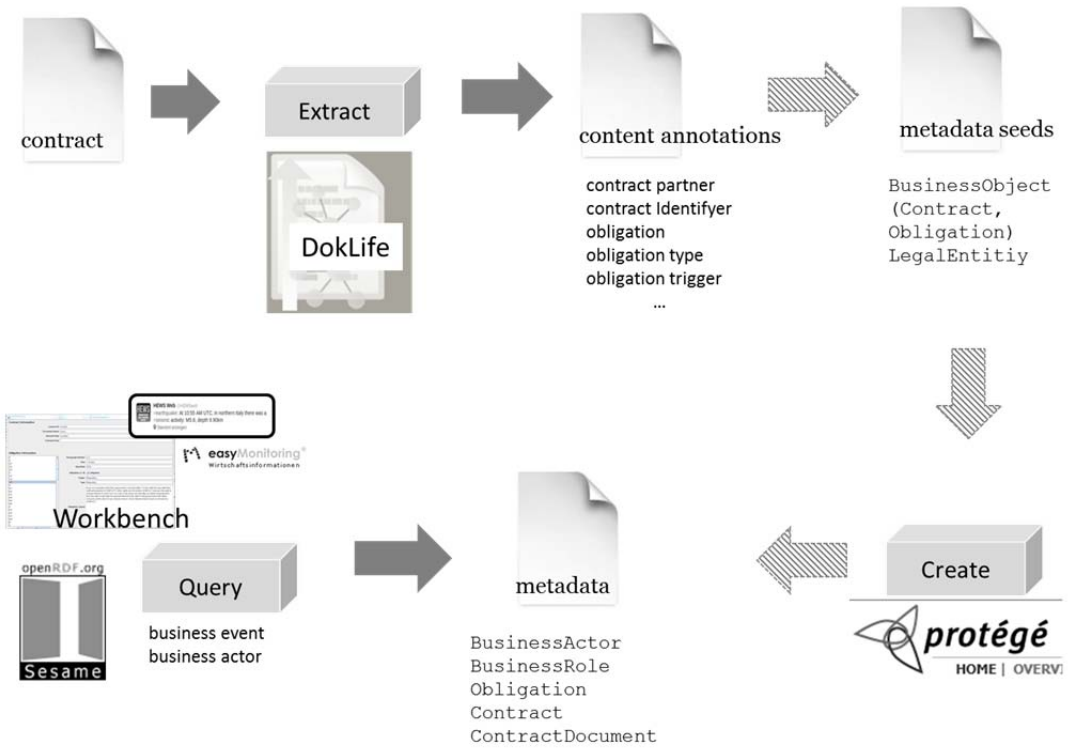


Figure 80: Symfact Demonstrator Components

The demonstrator focuses on the generation and use of additional metadata built on extracted and annotated metadata for contracts. For the demonstrator two use cases have been considered: a) a force majeure event happens and b) a financial incident occurs. In both cases the affected business partners, their business roles and business relations and if any, obligations and their representation in contracts are to be identified. In the demonstrator the query terms that represent this information are entered manually. In a production system, i.e. integrated in Symfact's CLM, this process could be automated for example based on a twitter service[197] and financial information service[198].
The demonstrator provides a GUI (icon depicted in the lower left corner of Figure 80 with the label 'workbench') to query seEAD for business relationships affected by an event. In case a) the earthquake, which happened 2011 in Fukushima has been taken as an example, in case b) bankruptcy of a company has been assumed.
In both cases inferencing is performed: In case a) to determine if a given event is a force majeure event and if it took place in a region where a business partner is located. In case b) to also identify related companies that might be affected by the financial event, e.g. if

[197] Humanitarian Early Warning Service (HWES) offers a twitter service about natural hazards, that can be parsed, for example to extract the type of hazard (earthquake, tornado, etc.), geographic coordinates, gravity of incident. URL: http://www.hewsweb.org/hp/ (retrieved: 12.12.12)
[198] In Switzerland Easymonitoring AG offers services for monitoring business partners with respect to financial incidents. URL: http://www.easymonitoring.ch (retrieved: 12.12.12)

a company which is assumed to be bankrupt is not a *direct* business partner but a partner of a partner like a supplier of a supplier.

2. Captured questions to be answered
   As required by the Action Research method (cf. Chapter 2.2.3) questions that arose during demonstrator development are to be answered in collaboration with the Action Research partner.
   - As contract management is done with Symfact's CLM system it is of great importance to relate seEAD and CLM. In order to achieve this, the ontological representation of a contract document must be mapped with the according entity in the CLM. As detailed in Chapter 5.1.5 this mapping could be done by direct mapping of the database entities to the respective concepts in seEAD.
   - To avoid manual creation of non-enterprise specific information, e.g. locations and geographic coordinates, relevant for example to determine if an event affects a business partner's construction plant, linked open data could be used. As mentioned in Chapter 3.2.5 many information collections are already available on the internet – often free of charge – like GeoNames[199] - that could be downloaded and integrated into seEAD. As GeoNames is not available in RDF yet, for geographic data alternatively Telegraphics Linked Open Data[200] could be checked on in order to further reduce effort.

3. Captured change requests and supplementary requirements
   Neither changes nor supplementary requirements have been requested. To recall Symfact's specifics: in place of *content-related* metadata *administrative* metadata is generated automatically for contract documents and, instead of using context only for generating metadata for documents it is used for generating metadata for business objects, too.

4. Actions to overcome or test the problem and to adapt the demonstrator to enterprise specific requirement
   In the main the demonstrator meets the Action Research Partner's expections and no major adaptions were required. Although possibilities of automating contract monitoring are of importance with respect to the production system – e.g. how twitter message from the Humanitarian Early Warning Service could be parsed and content elements transformed into query terms – this is no concern of my thesis and therefore not pursued.

5. Share with others (departmental meeting, publication, conference, etc.)
   To share results of the second loop of Action Research with others the demonstrator (second version of prototype) was presented and discussed with the Action Research team on June 14th, 2012.
   To broaden the audience the demonstrator has been presented to the consortium of the APPRIS project on May 30th, 2012.
   Part of the results have been published in (Thönssen & Lutz 2012) and were presented at the 4th Conference on Knowledge Management and Information Sharing (KMIS2012) in October 2012 in Barcelona, Spain.

[199] GeoNames is geographical database that covers all countries and contains over eight million placenames. Information is available for download in tab-delimited text free of charge. URL: http://www.geonames.org/ (retrieved: 17.8.2012)
[200] Telegraphis Linked Open Data provides data on countries, continents, capitals, and currencies collected from GeoNames and Wikipedia data. URL: http://telegraphis.net/data/ (retrieved: 17.8.2012)

Findings of the second loop of Action Research with Symfact:
Questions from loop 1 have been answered as follows:

- To represent the contract model in seEAD the ArchiMate standard is well suited. It clearly differentiates between business objects (contract and obligations) and their representation (contract documents). Furthermore background knowledge like relations between business partners, roles a business partner has and so on could be modelled appropriately. Refer to Symfact's context model (Figure 74) for details.
- To improve Contract Lifecycle Management with respect to managing obligations triggered by events (either force majeure or business events), events must be represented in seEAD. ArchiMEO already includes the top-level concept `Event` in addition to the ArchiMate concept `BusinessEvent`. Also the top-level concept `Location` is already available in ArchiMEO, which is needed to identify business actors having offices or production plants in a region an event took place. As business events like bankruptcy can have an impact on business eventhough a business partner is not directly affected, relations between business partners should be represented, which is already covered by the ArchiMate standard. Another concept needed and already available in ArchiMEO is `eo:LawAndRegulation`. To support compliant records management relevant law and regulations should be represented. The corresponding concepts and relations were modelled in ArchiMEO based on the results achieved in the OntoGov project[201].

6.2.6.2 Research Questions Addressed Within the Second Loop of AR With Symfact

As proposed in my research design (cf. Chapter 2.1, Table 1) within the second loop of Action Research with Symfact several research questions were addressed.

1. Context entities that can be inferred to automatically generate metadata, (answer to RQ3) are the following `LawAndRegulation`, `BusinessActor`, `BusinessRole`, `BusinessRelationship` and `BusinessObject` (cf. Chapter 6.2.2).

2. In order to answer RQ5 about the rules that are to be defined for metadata generation (it is important ot remember) I call to mind that in Symfact's case administrative metadata is in focus. The requirement is to improve compliant archiving of the contract documents and to better support contract lifecycle management, particularly in case of event triggered obligations. To determine the inferencing rules for metadata generation in Symfact's case events that might trigger obligations have been analysed. Two major types have been identified: a) specific events regarding a company's business (e.g. a service has not been delivered as contracted in a Service Level Agreement) and b) enterprise independent events, like force majeure events (natural disasters, war, riots, etc.) or business events (outsourcing, merger, bankruptcy, etc.) In my thesis I create a solution applicable to enterprises of all business domains. If for example a hurricane ravages an area it is for all companies of vital interest whether a business partner is affected. Refer to sections 6.2.4 for details on rules for metadata generation for Symfact.

[201] Ontology-enabled e-Gov Service Configuration (OntoGov) was a STREP project funded by the European Union (IST PROJECT 507237) from 01.01.2004 to 30.6.2006. FHNW has been partner in the project consortium and substantially contribute to the ontology model created in WP4: OntoGov Ontology Management System (B. Thönssen, Stojanovic, & Pariente, 2005).
http://cordis.europa.eu/fetch?ACTION=D&CALLER=PROJ_IST&QM_EP_RCN_A=71252 (retrieved: 24.7.2012)

3. In Symfact's case automatically generated metadata is used to improve records management with respect to determine start of retention period (inferred from the date a contract is closed) which law and regulation records management must comply with. This can be inferred from (1) the location of a company , e.g. if the company is based in Switzerland the general archiving law in Switzerland is relevant; and (2) the industry sector a company is doing business in, e.g. if a company does business in the manufacturing sector law regarding product liability must be obied. This answers RQ6.

4. As for AHSGA also for Symfact RQ8, RQ10 and RQ11 concerning enterprise architecture can be answered by referring to the ArchiMate standard, seEAD is built on. The enterprise objects which constitute an enterprise architecture are defined by ArchiMate (answer to RQ8) and the standard also provides an appropriate structure (answer to RQ10). ArchiMate, respectively the enhancements that ArchiMEO provides, have proved appropriate for Symfact, too. No enhancements to the core ontology have been necessary as enterprise characteristics could be represented in the enterprise specific part of seEAD.
   For example: the properties `sym:contractAgreementHasOption` and `sym:contractAgreesUponObligation`, which have been added extended the enterprise specific part of Symfact's semantically enhanced enterprise architecture description but does not affect ArchiMEO. Thus, also in Symfact's case ArchiMEO – based on ArchiMate – is general enough to be used 'out of the box' but customizable to specific companies's needs by enhancing the core ontology to an enterprise specific ontology, i.e. to seEAD (answer to RQ11).

5. As shown in Chapter 6.2.3 also in Symfact's case RDF Plus has been proven the appropriate ontology language for representing enterprise architecture in a way that is machine processable (answer to RQ9).

6. Phases 1 and 2 of the procedure model I introduced in Chapter 5.2 have been successfully applied in the Action Research study with Symfact, too. As detailed in the previous sections all required deliverables have been created and validated against the demonstrator. It could be proved that the first two phases of the procedure model have been appropriate for setting-up, conducting and utilizing metadata for Symfact (answer to RQ14).

Results of the third loop of the Action Research study with Symfact are provided within Chapter 7.3.4.

6.3 *Summary of Application Profiles*

As shown in the previous sections of this chapter the mintApproach for automatic metadata generation can be adapted to enterprise specific models, verified in the second loops of my Action Research studies with AHSGA and Symfact. I applied the Generic models, i.e. the Context Model (cf. Chapter 5.1), the Metadata Generation Model (cf. Chapter 5.2) and the Procedure Model (detailed in Chapter 5.3) in both Action Research studies and showed that the mintApproach builds a sound basis for enterprise specific adaptations. Results have been demonstrated, reviewed and discussed with the Action Research Partners and disseminated to a broader audience.
The differences between the two Action Research studies and the resulting diverging requirements have been particularly helpful in order to keep the generalizability of the mintApproach.

Figure 81 shows the differences and similarities between the studies: In the centre seEAD is depicted as it is the pivotal point of both applications, comprising the Enterprise Architecture Meta Ontology ArchiMEO as core ontology, and the application ontologies of AHSGA and Symfact. On the left hand side of the figure input for metadata generation is depicted. Metadata, whether based on harvested or extracted information, is generated with seEAD. On the right hand side of the figure use of metadata is sketched which can be for example for document retrieval related to reported tasks, document lifecycle management and obligation management.

Although in both Action Research studies generation of and search based on metadata is needed, the partner's focus is different as different business needs are to be met. AHSGA's main interest is to increase employee productivity by decreasing time for searching documents, related to a task. The documents are increasingly multi-media documents like videos presented in sexual health education or images presented in an exibition, and therefore automatic generation of metadata is of great importance. Symfact is most interested in improving contract management by identifying business-relevant events automatically and performing compliant records management.

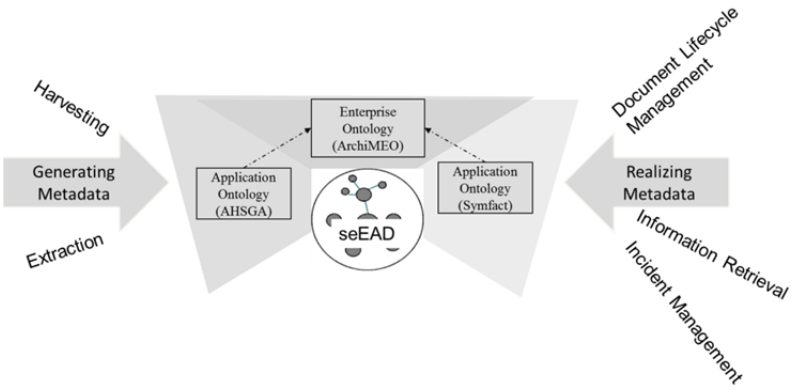


Figure 81: Metadata Generation and Realization

6.4 *mintApproach Findings II*

Within Action Research loop 2 I could verify the scientific assumptions of the mintApproach with respect to correctness, appropriateness and practical relevance.

It could be shown that it is possible to model the context for automatic metadata generation based on the Enterprise Architecture Meta Model (ArchiMEO), complemented by enterprise specifics. It proved that the two enterprise specific, semantically enriched Enterprise Architecture Descriptions (seEAD) provide a sound basis of well-defined enterprise objects and relations.

However, it became clear that a full-blown enterprise architecture description can not be expected in an enterprise. Thus, instead low-level governance instruments like templates, naming conventions and guidelines for structuring electronic filing e.g. written down in organisational handbooks, can serve as basis for the development of a seEAD. Ordering systems, particularly in the form of file plans or business classifications schemes can also build a valuable source for enterprise specific enhancements of seEAD since they are considered as a core element of electronic records management systems (Spree, 2009). In particular process-oriented file plans can provide the missing link between filed documents and business processes (Steinbrecher & Müll-Schnurr, 2010) to be represented in seEAD.

If file plans bear no relation to business processes templates, documents are created from, can be considered. Templates, e.g. '.dot files' for MS Word documents support users in routine tasks like creating minutes, writing the quarterly report and filling out a form. Hence, enterprise objects related to templates, e.g. Business Function, Business Service, Product or Customer Group, can provide context information for the documents created from it. Metadata seed for this additional context information can be derived from the documents' document property 'template', which provides the file name of the used template.

It turned out that simply harvesting document properties for metadata seed creation is not sufficient. Firstly, names of document properties vary depending on the document creation software. For example the creator of a document may be expressed as 'author', 'creator', 'publisher', etc. Hence, naming must be homogenized, which is done in the mintApproach by transforming document properties with the same meaning to the same Dublin Core metadata element. Secondly, document properties, created by document creation software or operating systems are unreliable as they might be wrong (e.g. the author of a document is not the creator but the creator of another document the current one is a copy of), meaningless (e.g. a randomly generated file name of an image) or even completely missing. With the mintApproach the weakness is alleviated in various ways. For example, the harvested document properties for author build the metadata seed 'contributor'. Now rules can be applied to derive a document's 'creator', e.g. based on naming conventions for file names as in AHSGA's case or inferred from background knowledge defined in seEAD for example about responsibilities of employees. Checking on background knowledge can be applied also to verify extracted information, for example to check whether the signatory of a contract is authorized to sign.
Since the mintApproach takes all harvested (and homogenized) document properties as metadata seeds missing or meaningless information can compensated. If for example a file name is meaningless still information about the document's storage is available and can be exploited for metadata generation.

It became apparent that SKOS is well suited to structure domain knowledge, e.g. about sexual health, and to support the search for related documents. Hence, it is possible to broaden the search for documents based on `skos:semanticRelations`. If for example the `skos:narrower` property is used for relating terms (e.g. 'sexuality' to 'bi-sexuality' and 'trans-sexuality') search can be expanded from a document having the broader term as subject (here: 'sexuality') to documents having the narrower terms in their subject. Besides this obvious use SKOS can be used to handle changes. With the `skos:mappingRelations` organisational changes can be expressed as well as replacements of employees. Thus for example, search for documents performed by a new employee may retrieve documents created by her processor.

Finally it could also been shown that low-level governance instruments can build a starting point for defining the rules for automatic metadata generation. Complemented by the methodology of Uschold & Grüninger (1996) for ontology development, rules can be developed from statements expressed in natural language to machine processable inferencing rules. The ontology representation language RDFS Plus and the SPIN SPARQL Inferencing Notation proved appropriate for this.

7 mintArchitecture and Prototype

Chapter 7 of my thesis introduces the prototype, created for proof of concept of the theoretical models and used for illustration within the evaluationphase, as illustrated in Figure 82.

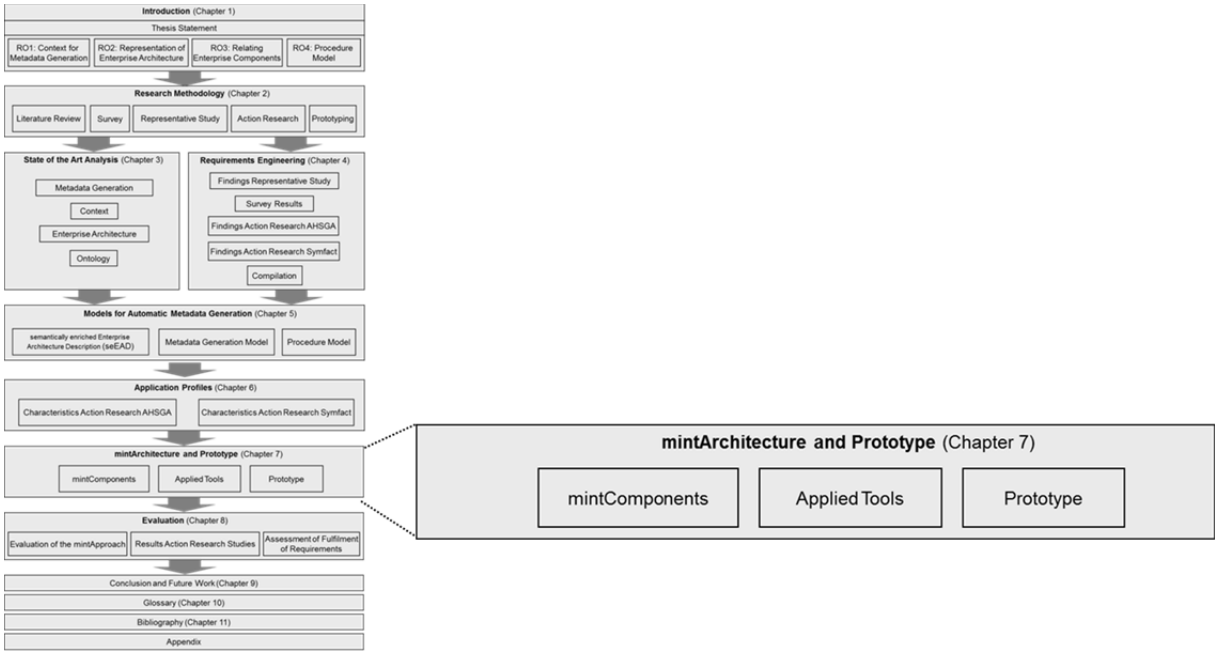


Figure 82: Position of Chapter 7 in the Overall Structure of the Thesis

The chapter describes how automatic, format-independent metadata generation is implemented in the MeGaWorkbench prototype. Whereas in Chapter 5 and Chapter 6 conceptual models for automatic, format-independent metadata generation have been introduced, Chapter 7 provides the description of a concrete model, what Dietz (2006, p 64) called an imitation of a concrete system.

The chapter starts with a “top level” view on the Metadata Generation Architecture. After that an overview on the application components and dependencies between them is given. Components – by UML definition –represent logical components (e.g., metadata harvesting, metadata creation), and physical components (e.g. NLNZ, GATE). In section 7.2 the applied toos are described and in section 7.3 the MeGaWorkbench prototype is described. Chapter 7 closes with a summary on implementation and prototyping.

To introduce implementation and prototyping of automatic, format-independent metadata generation the Metadata Generation Architecture is provided. The purpose of the Metadata Generation Architecture is to establish the metadata related services between metadata builders and metadata users such that any business information system can use the metadata generation system. Figure 83 depicts the model for the metadata generation architecture according to the Service oriented architecture Modelling Language (SoaML) Specification provided by the OMG (2012).

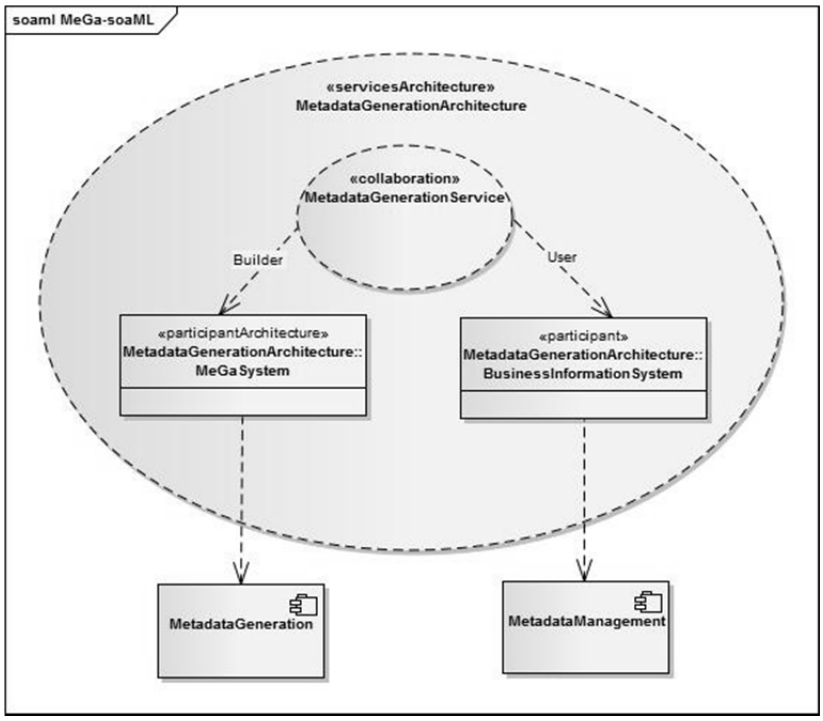


Figure 83: Metadata Generation Architecture

As defined for SOA the Metadata Generation Architecture depicts the participant roles providing and consuming services to fulfill a purpose. The two roles participants play in the Metadata Generation Architecture are the 'Builder' and the 'User'.
Figure 83 presents a "top level" view of how the independent participants work together for the purpose of improving document management. The diagram shows two composite application components, one for MetadataGeneration and one for MetadataManagement.

7.1 *mintComponents*

Figure 84 depicts the components of the MeGaSystem. The components realized in the prototype are depicted unshaded and for better readability only the interfaces relevant for the prototype are shown. At the upper left hand side the (graphical) User Interface (UI) for the prototype is depicted, called MeGaWorkbench. It is linked to two components: MetadataGeneration and MetadataManagement. For the prototype in the MetadataGeneration component the MetadataHarvesting Services and MetadataCreation Services are realized. The MetadataHarvesting Service consists of a Java component and the National Library of New-Zealand's (NLNZ) harvester. The General Architecture for Text Engineering (GATE) component is used for enhancing harvested file attributes (e.g. parsing nouns from a document's title). The MetadataCreation Services consist of another Java component and the TopBraid component with the SPIN rule and SPARQL query parts.
The "ball-and-socket" connection indicates assembly interfaces between components; connections between the external interface (i.e. the MeGaWorkbench) and internal components which realize the behaviour are depicted as delegation connectors.

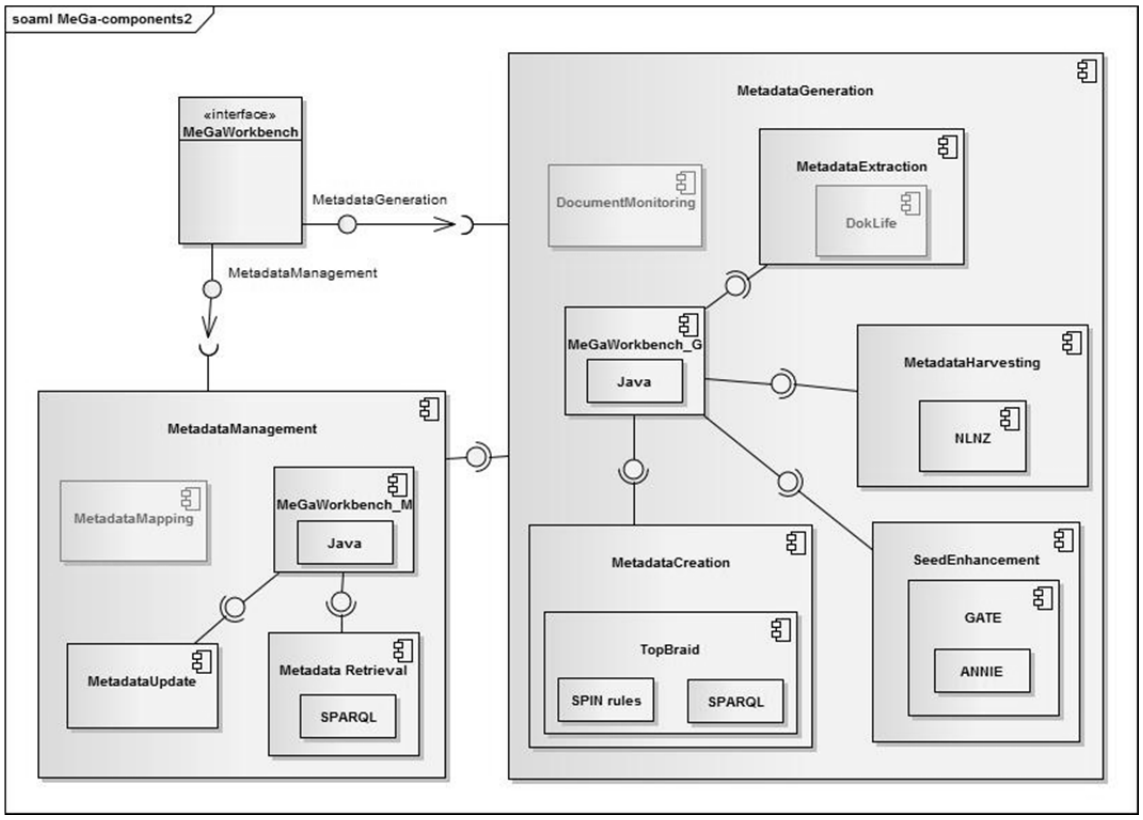


Figure 84: MeGaSystem Components

In the following the components, implemented in the MeGaWorkbench prototype are briefly characterized. For details reference to the respective use cases in Chapter 5.2.1 is provided. The utilized tools are named there but described in Chapter 7.2. Use of components is widely independently, i.e. depending on whether and how the approach is used with an enterprise's application the appropriate component(s) can be applied.

7.1.1 MetadataGeneration Component

The following components – depicted in Figure 84 – are implemented in the MegaWorkbench prototype: DocumentMonitoring, MetadataHarvesting, MetadataExtraction, SeedEnhancement, MetadataCreation and MeGaWorkbench_G.

The MetadataExtraction component includes the DokLife component as described by (Thönssen & Lutz 2012). For a contract document as a whole as well as for its segments metadata are automatically extracted and annotated, e.g. contract partner, contract beginning, contract end, applicable law. For information extraction GATE is used and some JAVA web-services. In addition to metadata for the whole contract particular metadata is created for single paragraphs, e.g. obligation type, trigger, dates and conditions. For that regular expressions are used to extract due dates, conditions and triggers (cf. UC3 Chapter 12.4.9). Results are stored in an XML-file. For the MeGa Workbench prototype this XML-file is the basis for metadata seed creation.

The MetadataHarvesting component includes the harvester of the National Library of New Zealand (NLNZ). The component readouts file attributes of documents stored in pre-defined storage locations (cf. UC2.1 Chapter 12.4.4) and stores the harvest in an XML-file (cf. UC2.2

Chapter 12.4.5). The harvested filename and the XML is further processed with GATE, i.e. it is segmented in its parts and nouns are identified). The results are stored as additional seeds.

To the MetadataCreation component creates metadata inferred from metadata seeds (cf. UC2.4 Chapter 12.4.7), as well as metadata candidates (UC2.5 Chapter 12.4.8). Therefore the SPIN Library provided by TopBraid is used. SPIN rules are detailed in Chapters 6.1.4 and 6.2.4.

The MeGaWorkbench_G Component builds the glue between the various components and creates the metadata seeds (UC2.3 Chapter 12.4.6), i.e. the instances of classes and properties in seEAD, on the basis of harvested file and content annotations. It is a Java program that acts as a pipe, i.e. processing of components is arranged in a way that the output of each component is the input of the next.

7.1.2 MetadataManagement Component

Figure 84 depicts the components of MetadataManagement implemented in the MeGaWorkBench prototype: MetadataMapping, MetadataRetrieval, MetadataUpdate and another MeGaWorkbench component.

The MetadataMapping component maps instances of seEAD to attribute values represented in a relational database (cf. UC4.1 Chapter 12.4.11). This can also be done with TopBraid Composer (standard edition).

MetadataRetrieval is a component that allows for querying the enterprise repository, i.e. seEAD and if mapping has been done the respective non-ontological data stores. For operational use search is executed via the user interface of a third party system for example Symfact's CLM system (cf. UC4.2 Chapter 12.4.12). To query the ontology templates are used in the form of pre-defined SPARQL queries (cf. UC5 Chapter 12.4.14). The query results, the RDFS triples, are rehashed to a more user friendly result list (cf. UC51 Chapter 12.4.15) as common in information retrieval.

The MetadataUpdate component allows for updating metadata based on user interaction. If for example an AHSGA user selects two documents from the result list to add to the reported task execution of SPIN rules are triggered to determine relations between these documents. The component operates an interface to TopBraid Composer for this. Related use cases are UC4.3 Chapter 12.4.13 and UC6 Chapter 12.4.16.

Also within the MetadataManagement component the MeGaWorkbench_M component serves as the glue between the included components and adds the functionality not provided out-of-the-box. As for example the execution of SPIN rules and functions are subject to certain restrictions workarounds have been implemented in Java. Furthermore the MeGaSpine_M component fills the terms entered by the user into the query templates, e.g. transforms them into variables of the SPARQL queries.

7.1.3 Graphical User Interface

The MeGaWorkbench graphical user interface (implemented in Java 7) is a substitute of the interface an enterprise's application provides. It allows a user to interact with the components described above and hence to generate, retrieve and update metadata.

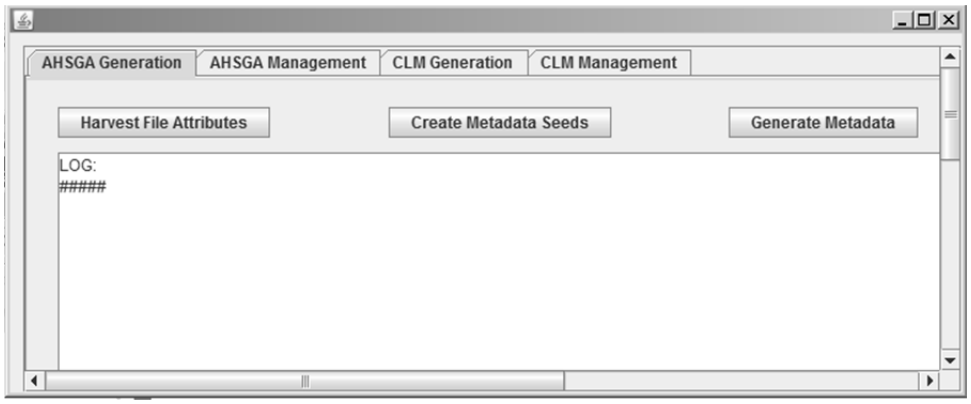


Figure 85: MeGaWorkbench GUI

Figure 85 depicts a print screen of the MeGaWorkbench after starting the prototype. It consists of four tabs, two substituting each Action Research partner's application and allowing for metadata generation and retrieval. As the print screen shows AHSGA metadata generation comprises three parts: harvesting of file attributes, creating metadata seeds and generating metadata. Similarly metadata generation is performed for Symfact except for metadata harvesting. Instead data seeds are created form extracted information. In productive use these parts would be performed transparently to the user. They made explicit in the prototype to make the process of automatic metadata generation testable.

Whereas the GUI for metadata generation is similar for both Action Research partners, the screens for retrieval differ completely. In AHSGA's case the MeGaWorkbench GUI substitutes the screens for task reporting AHSGA's ITRS by one entry screen as depicted in Figure 95. On the left hand side product independent entry data is depicted whereas on the right hand side data, specific for the product 'AHSGA Prevention' is shown (refer to 6.1.5.5 for details on AHSGA's ITRS data).

Functionality of the MeGaWorkbench prototype is described in detail in Chapter 7.3.

7.2 *Tools*

Tool selection has been driven by two aspects: 1) to avoid reinventing the wheel if ever possible existing software is used; 2) to make my results publicly available all components are available free of charge, either as open source software or as free edition. The latter implies that the newly developed Java components are also open source.
Whereas GATE for natural language processing and TopBraid for ontology development and management are used in many research projects, no tool for metadata harvestings is prevailing. Thus an evaluation of harvesting tools has been performed within a student project (Johner, 2011) supervised by me. For the mintApproach the Metadata Extraction Tool, developed by the National Library of New Zealand, was considered most appropriate, as it is open source, adaptable to the needed document properties and extensible if need be for other than the default file formats. Table 38 gives an overview on the software used in the metadata generation prototype.

Tool Name	Source / URL	Brief Description	Included in Component
Metadata Harvesting Tool (NLNZ)	http://www.natlib.govt.nz/services/get-advice/digital-libraries/metadata-extraction-tool	Developed by the National Library of New Zealand, the Metadata Extraction Tool is provided free of charge and with its source code. It allows for harvesting document properties of current document formats (e.g. .doc, .pdf, etc) and output is structured in XML format. For each file format a so called 'adapter' is prepared that can be adopted if necessary. In addition new adopters can be incorporated which is especially useful for	Metadata Harvesting
General Architecture for Text Engineering (GATE)	http://gate.ac.uk/overview.html	GATE is an open source and free software for text analysis. GATE includes components for diverse language processing tasks, e.g. parsers, morphology, tagging, information extraction (ANNIE), etc. for various languages.	Seed Enhancement
TopBraid	http://www.topquadrant.com/products/TB_Composer.html	TopBraid Composer is a modelling environment for developing ontologies and building semantic applications. TopBraid Composer is compliant with W3C standards, recommendations and submissions like RDF(S), OWL, SPARQL and SPIN. TopBraid	Metadata Creation; Metadata Update
MeGaWorkbench	https://www.fhnw.ch/personen/jonas-lutz	The MeGaWorkbench is programmed in Java. Its components build the glue between the out of the box tools and it is the backbone of the prototype. The MeGaWorkbench provides a Graphical User Interface to generate, query and update metadata.	Metadata Generation Backbone

Table 38: Software Used in the Metadata Generation Prototype

7.2.1 National Library New Zealand (NLNZ)

In 2003 the National Library of New Zealand (NLNZ) released the first version of the Metadata Extraction Tool. In their paper Kebbell & Campbell (2002) analyzed NLNZ's strategies with respect to the reference model of the Open Archival Information System

(OAIS)[202] and the harvester can be regarded as an continuation and implementation of a part of it. The tool has been redeveloped in 2007 and Version 3 is available as open-source software and can be downloaded from the SourceForge website[203]. The preservation Metadata Extract Tool (for short NLNZ) automatically extracts file attributes from digital files and outputs that data in a standard format (XML). The Metadata Extract Tool includes a number of 'adapters' that readouts file attributes from popular file types. The following adaptors, relevant for my work, are provided:

- Images (stills): BMP, GIF, JPEG, TIFF
- Office documents: MS Word, MS Excel, MS PowerPoint, PDF
- Audio: MP3
- Markup languages: XML.

If a file type is unknown the tool applies a generic adapter, which extracts data common for any given file (such as size, filename, and date created). For the MeGaWorkbench prototype no new adapter has been developed but existing adapters have been modified to harvest all required file attributes. Figure 86 depicts the schema for an adaptor for text files of .xls data format.

[202] The OAIS has been developed by the Council of the Consultative Committee for Space Data Systems (CCSDS), a multi-national forum for the development of communications and data systems standards for spaceflight, founded in 1982. In 2002 OAIS Reference Model (CCSDS, 2012) had the status of Recommended Standard but CCSDS has changed the classification in 2012 from Blue (Recommended Standard) to Magenta (Recommended Practice). URL: http://public.ccsds.org/default.aspx (retrieved: 12.8.2012)
[203] NLNZ Metadata Extraction Tool. URL: http://meta-extractor.sourceforge.net/ (retrieved: 12.8.2012)

```
<File>
  <FileIdentifier>
    <xsl:value-of select="nz_govt_natlib_xsl_XSLTFunctions:determineFileIdentifier(string(MSEXCEL/METADATA/PID),
    string(MSEXCEL/METADATA/OID),string(MSEXCEL/METADATA/FILENAME),string(MSEXCEL/METADATA/FID))"/>
  </FileIdentifier>
  <xsl:for-each select="MSEXCEL/METADATA/PATH">
    <Path>
      <xsl:value-of select="."/>
    </Path>
  </xsl:for-each>
  <Filename>
    <xsl:for-each select="MSEXCEL/METADATA/FILENAME">
      <Name>
        <xsl:value-of select="."/>
      </Name>
    </xsl:for-each>
    <xsl:for-each select="MSEXCEL/METADATA/EXTENSION">
      <Extension>
        <xsl:value-of select="."/>
      </Extension>
    </xsl:for-each>
  </Filename>
  <xsl:for-each select="MSEXCEL/METADATA/FILELENGTH">
    <Size>
      <xsl:value-of select="."/>
    </Size>
  </xsl:for-each>
  <FileDateTime>
    <xsl:for-each select="MSEXCEL/METADATA/DATE">
      <Date format="yyyyMMdd">
        <xsl:value-of select="."/>
      </Date>
    </xsl:for-each>
    <xsl:for-each select="MSEXCEL/METADATA/TIME">
      <Time format="HHmmssSSS">
        <xsl:value-of select="."/>
      </Time>
    </xsl:for-each>
  </FileDateTime>
  <xsl:for-each select="MSEXCEL/METADATA/TYPE">
    <Mimetype>
      <xsl:value-of select="."/>
    </Mimetype>
  </xsl:for-each>
  <FileFormat>
    <Format>
      <xsl:value-of select="string('Microsoft Excel')"/>
    </Format>
    <Version>
      <xsl:value-of select="string('4.1+')"/>
    </Version>
  </FileFormat>
  <Text>
    <CharacterSet>
      <xsl:value-of select="string('UTF-8')"/>
    </CharacterSet>
    <MarkupLanguage>
      <xsl:value-of select="string('unknown')"/>
    </MarkupLanguage>
  </Text>
</File>
```

Figure 86: XML Schema of an NLNZ Adapter for Excel Files

Figure 87 depicts an XML-file with the harvested document properties of a document in pdf format. On the left hand side of the figure four of the related DC metadata are shown. Note, that the harvested author does not become a metadata seed for creator but for contributor and that the file name is considered alternative title (refer to Chapter 6.1.5.2 for details).

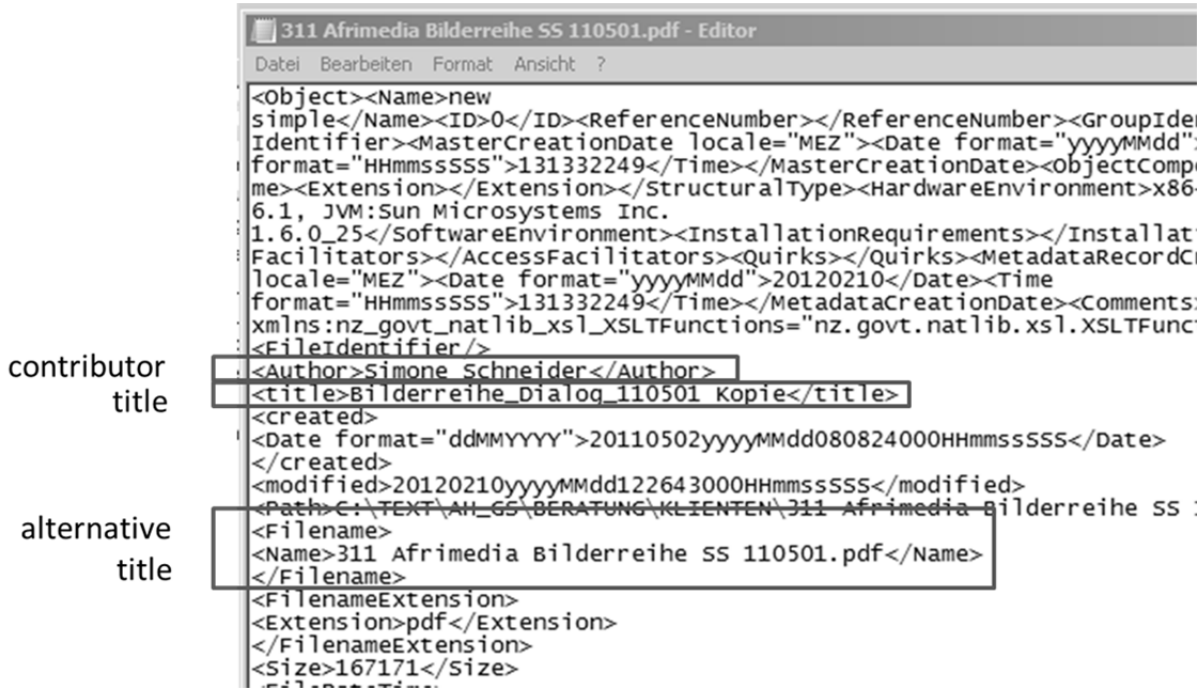


Figure 87: Harvest of a Document

The NLNZ harvester tool has both a Microsoft Windows interface and a UNIX command line interface. This allows for embedding it into the MeGaWorkbench_G component, which calls the harvester, hands over the documents and takes back the results in the form of an XML file. After that the MeGaWorkbench_G component hands it over to GATE to create the metadata seeds.

7.2.2 GATE

The General Architecture for Text Engineering (GATE) is a long-established, well-accepted tool in the scientific community for text analysis of all shapes and sizes[204]. Since 15 years available, "From large corporations to small startups, from €multi-million research consortia to undergraduate projects, our [the GATE] user community is the largest and most diverse of any system of this type, and is spread across all but one of the continent" (GATE, n.d.).
Do Prado & Ferneda (2008) and (Saggion (2008) describes GATE as a tool for natural language processing and information extraction where own applications can be defined with modules like document collections (corpora), tokenizers (converting a sequence of characters into a sequence of tokens), gazetteers (directories), sentence splitter, part of speech tagger (grammatical tagging), named entities transducer (entity identification) and coreference tagger (multiple expressions refer to the same thing).
For automatic metadata generation only a litte part of the comprehensive functionality of GATE is used, namely some ANNIE components. ANNIE, the 'Nearly-New Information Extraction System' is a plug-in for GATE. ANNIE is used for tokenization, sentence splitting and morphological analysis. As the part-of-speech tagging (POS)[205] is currently not supported

[204] GATE Papers. URL: http://gate.ac.uk/gate/doc/papers.html (retrieved: 12.8.2012)
[205] Wikipedia. "Part-of-speech tagging, also called grammatical tagging, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context — i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc. URL: http://en.wikipedia.org/wiki/Part-of-speech_tagging (retrieved: 12.8.2012)

for the german language by GATE itself, a external POS tagger called TreeTagger from the University of Stuttgart[206] has been integrated () i.e. for word-category disambiguation, to identify nouns in a document's file name. Figure 88 depicts the alternative title as shown in Figure 88 plus the metadata seeds gained by using GATE, respectively the enhanced ANNIE.

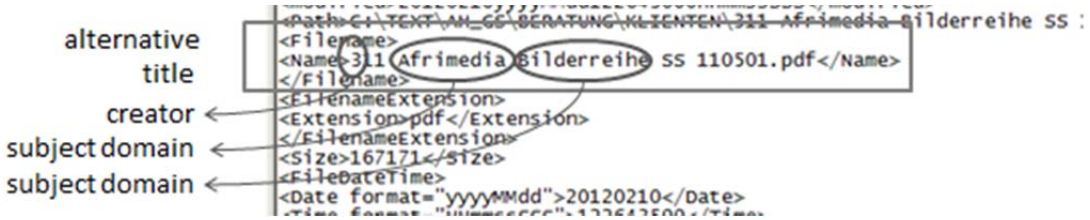


Figure 88: Input for ANNIE

After the metadata seeds are created, the JENA API is used by the MeGaWorkbench to create the respective instances in seEAD.

7.2.3 TopBraid Composer

According to Vavliakis et al. (2011, p 3844) computational limitations in semantic querying and inferencing have drastically reduced the thrust in semantic technologies and tools are needed that "actually 'work' efficient […] allowing the semantic transformation of legacy data, which can then be queried, processed and reasoned upon". As already mentioned above many tools for developing Semantic Web applications are available although Vavliakis et al. (2011) claimed that no system meets business requirements in a user-friendly and easy-to-adapt manner. However, in their paper Suárez-Figueroa et al. (2011) classified the broad landscape of ontology managing software according to their functionalities and briefly explained the various tools. One of them is TopBraid Composer, developed by TopQuadrant[207], which is a modelling environment for developing Semantic Web ontologies and building semantic applications. TopBraid Composer has been reviewed an many scientific papers (amonst others by Waterfeld et al. (2008) and Kandefer & Shapiro (2008)), and the tool's use for building semantic applications has proved appropriate for many scientific projects (amongst others lately by Khan et al. (2011) and Saba & Mohamed (2012)) and it is fully compliant with W3C standards. TopBraid Composer is implemented as an Eclipse plug-in and uses the Jena API[208]. TopBraid Composer supports developing, managing and accessing knowledge models and their instance knowledge bases through graphical user and programming interfaces and provides in-built support of ontology to database mapping, and offers - according to (Yu 2011, p 477) "relatively complete and impressive support". TopBraid also provides easy to use import mechanism that allows adding and mapping further ontologies if needed. Furthermore, TopQuadrant (n.d., p 2) stresses the necessity of Enterprise Architecture descriptions "for realizing the vision of an agile enterprise that can adapt its IT and enterprise models to changing situations and opportunities" and claims that "EA models that are able to be distributed, federated and executable will be essential for realizing the vision of an agile enterprise that can adapt its IT and enterprise models to changing situations and opportunities".

[206] The TreeTagger is a tool for annotating text with part-of-speech and lemma information. It was developed at the Institute for Computational Linguistics of the University of Stuttgart. URL: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ (retrieved: 3.11.2012)
[207] TopQuadrant. URL: http://www.topquadrant.com/products/TB_Composer.html (retrieved: 3.8.2012)
[208] W3C Semantic Web. TopBraid. URL: http://www.w3.org/2001/sw/wiki/TopBraid (retrieved: 6.8.2012)

With respect to the MeGaWorkbench prototype all implementation requirements have been met by TopBraid Composer. Figure 89 depicts a print screen of TopBraid Composer GUI. At the left hand side of the screen a detail of seEAD is shown with the selected concept of `AHSGA_Document`. Representation of the concepts and properties is similar to Protégé. At the right hand side of the screen properties of the `AHSGA_Document` concept are depicted. In the lower part of this side an inferencing rule for this concept is listed (rule AHSGA_IR_1 explained in Chapter 6.1.4). SPIN rules are considered (other) properties of concept, here of `AHSGA_Document`,

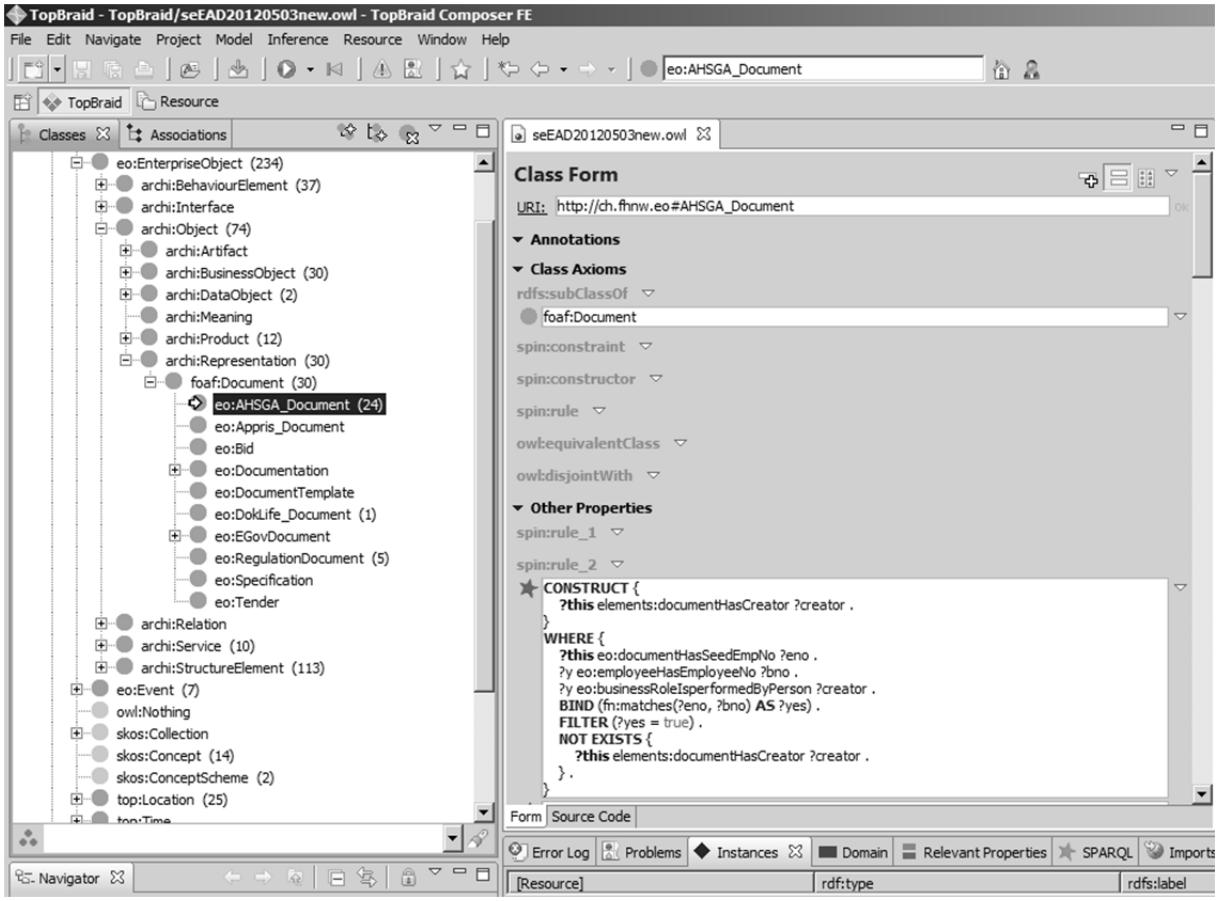


Figure 89: Print Screen of TopBraid Composer

For defining rules sub-properties of the `spin:rule` can be created, for example `spin:rule_1, spin:rule_2`, etc. SPIN rules can be grouped, i.e. more than one rule can instantiate a property like spinrule_2. To enforce the sequence of execution of (groups of) rules a SPIN property is provided, called `nextRuleProperty`. Figure 90 depicts a detail of the rules property form of `spin:rule_2` with `spin:rule_3` to be executed next.

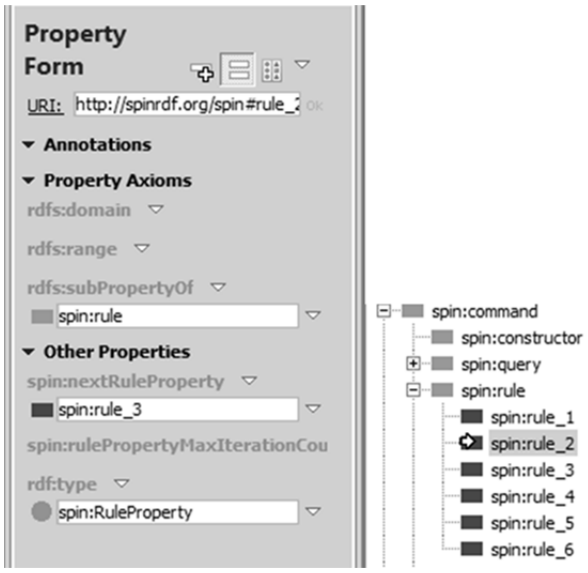


Figure 90: SPIN Rules Properties

When inferencing is triggered all rules – independently from the concept they are defined as properties for – are executed. That is not only the rules for the `AHSGA_Document` concept are executed but also the for the SKOS concept (cf. Chapter 6.1.4, AHSGA_IR 21 – 23).
Furthermore, TopBraid provides a mechanism to encapsulate SPARQL queries so that they can be reused in different contexts. The so called SPIN Template is basically a canned SPARQL query that is parameterized with arguments.

Figure 91 gives an example of a SPIN Template. It selects the documents that have a `skos:narrower` term related to the argument as a document's subject. Such kind of query is used to broaden a query result as the argument is a selected the document's subject (cf. 6.1.5.4 on AHSGA's metadata).

```
SELECT DISTINCT ?arg1 ?doc ?narrower  ?title
WHERE       {
?selDoc eo:documentHasSubjectDomain ?arg1.
?arg1 a skos:Concept .
?arg1 skos:narrower ?narrower .
?doc eo:documentHasSubjectDomain ?narrower .
?doc elements:documentHasTitle ?title .
            }
```

Figure 91: SPIN Template

TopBraid Composer is used for ontology managing, mainly inferencing. For rule definition, SPARQL query creation and testing the GUI of TopBraid is used. Within the MeGaWorkbench prototype the functionality is triggered by the provided programming interface. After the MeGaWorkbench has created the instances of newly constructed metadata seeds the inferencing rules stored in TopBraid are triggered to generate the metadata (candidates).

7.3 *Prototype*

As stated by Houde & Hill (1997, p 368) "Prototypes provide the means for examining design problems and evaluating solutions". Prototyping has been chosen as method to provide a proof of concept of my approach for automatic format-independent metadata generation based on semantically enriched context information. As designed (cf. Chapter 2.2.4) in each loop of

my Action Research study an artefact has been developed, gradually evolving from a model (loop 1), over a demonstrator (loop 2) to en executable prototype (loop 3). As aforementioned, no one tool has supported the iterative design work in all of the important areas of investigation but different tools has been used for different prototyping tasks, an experience also described by Houde & Hill (1997).
The MeGaWorkbench prototype consists of the two components MetadataGeneration and MetdataMangement, adapted to the requirements of the two Action Research partners AHSGA and Symfact. Figure 92 depicts the two components of the prototype. The colour gradient indicates the focus of the respective Action Research partners.

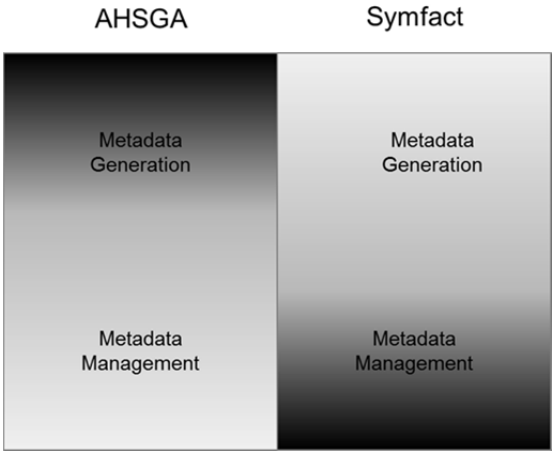


Figure 92: MeGaWorkbench Focus

In the following the behaviour of the prototype for Metadata Generation is briefly described. For the sake of brevity refer to Chapters 5 and 6 for details on the underlying models. In Chapter 7.3.3 specifics of AHSGA's prototype are provided and in Chapter 7.3.4 particularities of Symfact's prototype are given.

7.3.1 MeGaWorkbench Application Profile

The developed prototpye is described according to the "DCMI Tools Application Profile" suggested by Greenberg & Severiens (2007).

Element	Qualifiers/Terms	Value
Creator		Barbara Thönssen
Contributor	Developer Documenter Tester	Jonas Lutz Barbara Thönssen Action Research partners, end-users, software vendors
Date	Created lastModified Issued	30.8-2012 15.10.2012 tbd
Description		Harvests document properties from enterprise documents of common formats (e.g. of MS Office Documents). Creates seeds for inferencing metadata from a documents context, i.e. from related information objects

Element	Qualifiers/Terms	Value
Identifier	Repository	www.thoenssen.ch/megaworkbench
Language		en
Rights	accessRights license	open source GNU General Public License www.gnu.org/copyleft/gpl.html
RightsHolder		Barbara Thönssen
Title		MeGaWorkbench
Type	Software	Prototype
Audience		Developer User
ProgrammingLanguage		Java 7
OperatingSystem		Windows 7

Table 39: MeGaWorkbench Application Profile

Table 39 provides the MeGaWorkbench Application Profile based on Greenberg & Severiens (2007).

7.3.2 Prototype Behaviour

Figure 93 depicts an UML state diagram that is used to model the behaviour of the MeGaWorkbench prototype. The figure shows the various states the metadata generation process runs through, starting from the files loaded into the NLNZ harvester until the metadata are inferred from seEAD.

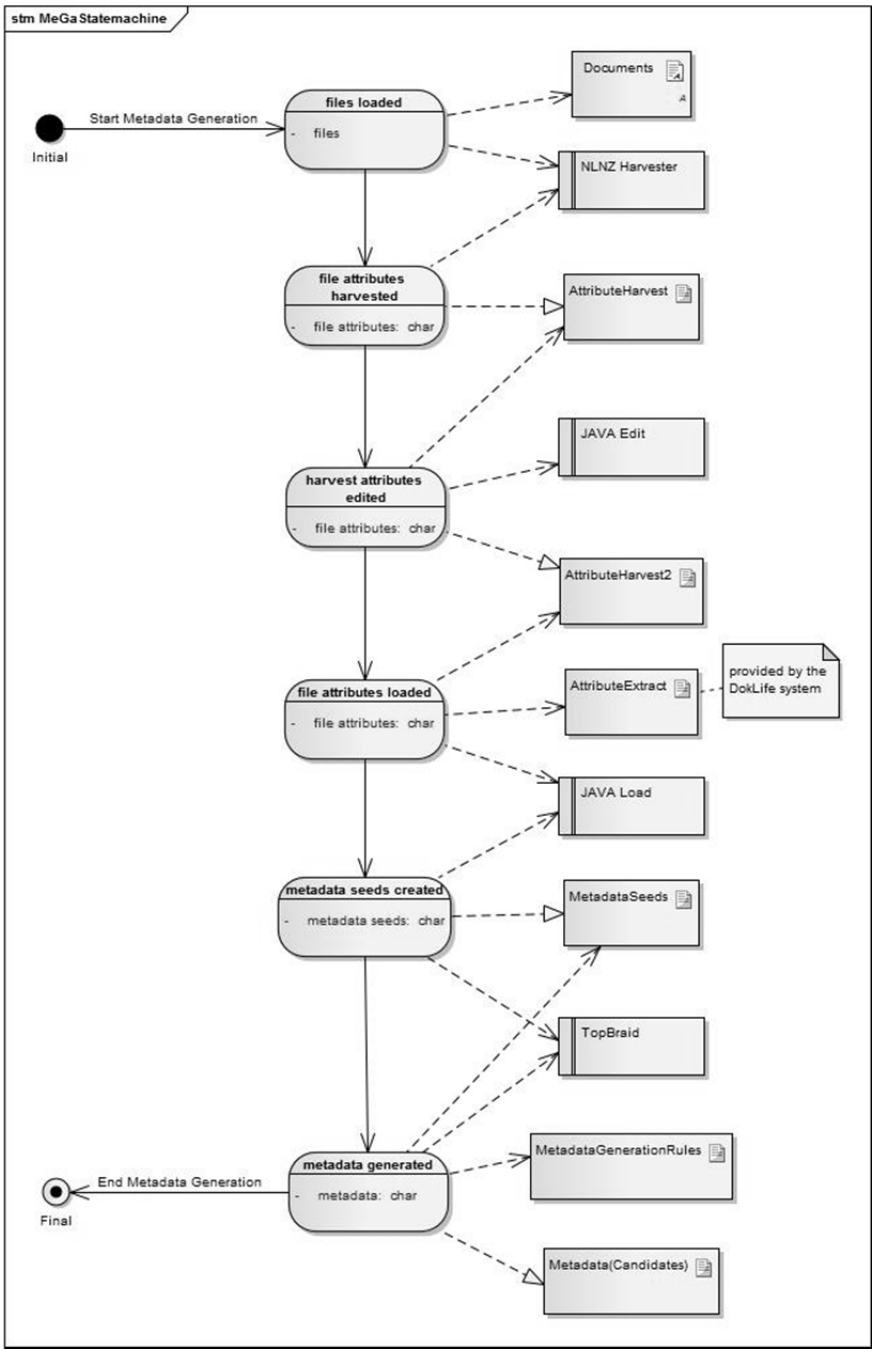


Figure 93: State Diagram for Metadata Generation

The MeGaWorkbench prototype allows for monitoring the generation process as for the main states (document properties harvested, metadata seeds created and metadata generated) log entries are displayed. Table 40 shows snippets of log entries for a test run.

Log Detail	State
Harvest of File 'C:\megaWorkbench\TEXT\AH_V\adressen_mitarbeiter.XLS' Completed Successfully Harvest of File 'C:\megaWorkbench\TEXT\AH_V\beges_Messlatte.jpg' Completed Successfully Harvest of File 'C:\megaWorkbench\TEXT\AH_V\Nationales_HIV_STI-Programm_2011-2015.pdf' Completed Successfully Harvest of File 'C:\megaWorkbench\TEXT\ESCHIRMER\311_Dankesbrief_Winkelriedstiftung.doc' Completed Successfully Harvest of File 'C:\megaWorkbench\TEXT\ESCHIRMER\Arbeitszeiten.xls' Completed Successfully Harvest of File 'C:\megaWorkbench\TEXT\ESCHIRMER\DIALOG_Bild_Seite_222.jpg' Completed Successfully	document properties harvested[209]
C:\megaWorkbench\fileAttributeFiles\syphilis_doppelseite.jpg.xml generated C:\megaWorkbench\fileAttributeFiles\1_Die_sechs_wesentlichen_Arbeitsgebiete_der_AHSGA_mit_Zuteilung.doc.xml generated C:\megaWorkbench\fileAttributeFiles\Abrechnung_2011.xls.xml generated C:\megaWorkbench\fileAttributeFiles\love_Signs.jpg.xml generated C:\megaWorkbench\fileAttributeFiles\310_zeitpunkt_mayakalender.pdf.xml generated C:\megaWorkbench\fileAttributeFiles\398_Sex_und_geistige_Behinderung.DOC.xml generated	metadata seeds are created
C:\megaWorkbench\fileAttributeFiles\Abrechnung_2011.xls.xml generated C:\megaWorkbench\fileAttributeFiles\love_Signs.jpg.xml generated C:\megaWorkbench\fileAttributeFiles\310_zeitpunkt_mayakalender.pdf.xml generated C:\megaWorkbench\fileAttributeFiles\398_Sex_und_geistige_Behinderung.DOC.xml generated Inferred triples: 1658	metadata (candidates) are generated

Table 40: Example of MeGaWorkbench Log Entries for AHSGA Documents

After generation is successfully completed metadata are stored in seEAD. Whereas with the MeGaWorkbench prototype each step of metadata generation is activated manually in a productive system it would be triggered automatically, e.g. by a timer, and be processed transparent for the user.

7.3.3 AHSGA Prototype

For the prototype AHSGA defined a sample of representative documents of all file formats they use (cf. Chapter 4.3.1 for details). A total of 187 documents in 25 folders have been selected and copied to a test environment in order not to jeopardize AHSGA's productive system. Figure 94 depicts the part of AHSGA's file structure simulated at the test system.

Figure 94: Explorer Structure Simulated for Prototyping

[209] As described in Chapter 6.2.5.1 in Symfact's case metadata generation starts with a subset of the extracted and already annotated information created within the DokLife project.

From these directories documents are taken, their document properties are harvested, metadata seeds are created and metadata (candidates) are generated as detailed in the previous chapter.

Whereas the GUI for metadata generation is similar for both Action Research partners, the screens for retrieval differ completely. In AHSGA's case the MeGaWorkbench GUI substitutes the screens for task reporting AHSGA's ITRS by one entry screen as depicted in Figure 95. On the left hand side product independent entry data is depicted whereas on the right hand side data, specifc for the product 'AHSGA Prevention' is shown (refer to 6.1.5.5 for details on AHSGA's ITRS data).

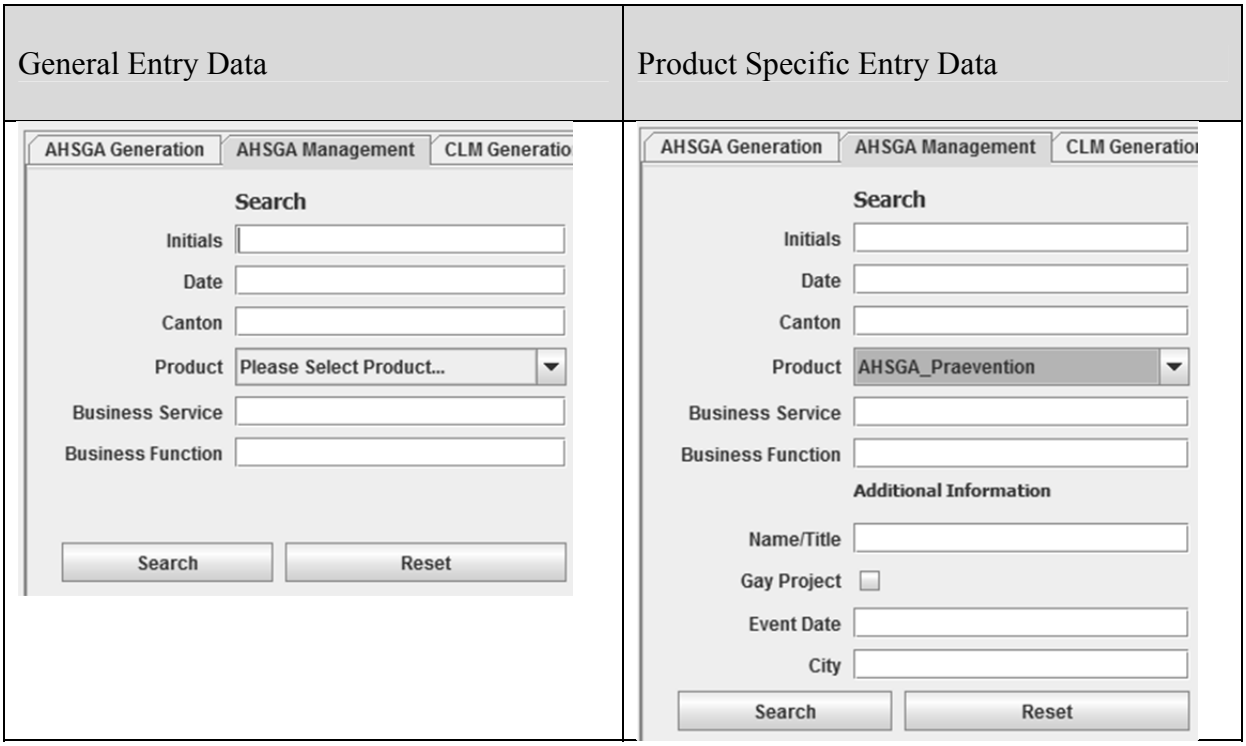


Figure 95: MeGaWorkbench Substitute of AHSGA's ITRS GUI

As already mentioned, the AHSGA part of the MeGaWorkbench prototype simulates the behaviour of ITRS with respect to task reporting but instead of the several screens for task reporting one entry screen is provided for data input.
Figure 96 provides a screen shot of data recorded by a user with the name intials 'ss', for the date '2012-10-23', the product 'AHSGA_Praevention', and additional information 'Isla', depicted at the right side of the figure. The entered data is transformed into query terms for the SPARQL query (in addition the required restriction of only retrieving documents that have been created or updated within the last 90 days is added).
For the recorded task two documents of the test collection are retrieved (depicted at the lower left side of the print screen) and ranked according to AHSGA's requirements. At the right side of the screen the metadata for the selected document is displayed, here for 'imagesCABPO547'. Note that little data could be harvested from the document properties: neither 'author' nor 'title' is available; the file name is meaningless. Thus, neither title nor contributor is available for the displayed document. The creator (❶) was inferred via context information: the user, in the given example 'SimoneSchneider' has the business role 'PadagogueInHumanSexualBehaviour'; this role is responsible for the customer group 'Behinderte' (which is another business role). Since the business role 'Behinderte' could be

derived from the directory in which the document was stored, the employee(s) with the aforementioned responsibility could be inferred. Since the role 'PadagogueInHumanSexualBehaviour' is assigned to three AHSGA employees, task records of each of them would retrieve this document (given the other criteria match, too). Also inferred from the business role 'Behinderte' are the AHSGA's clients who belong to this customer group. Thus, the additional information recorded to the task ('Isla') allows for retrieving the document via the inferred context (❸). ❷ shows another example of how AHSGA's low level governance instruments for information storage are used to determine the context of a document, here the intangible product 'Praevention'.

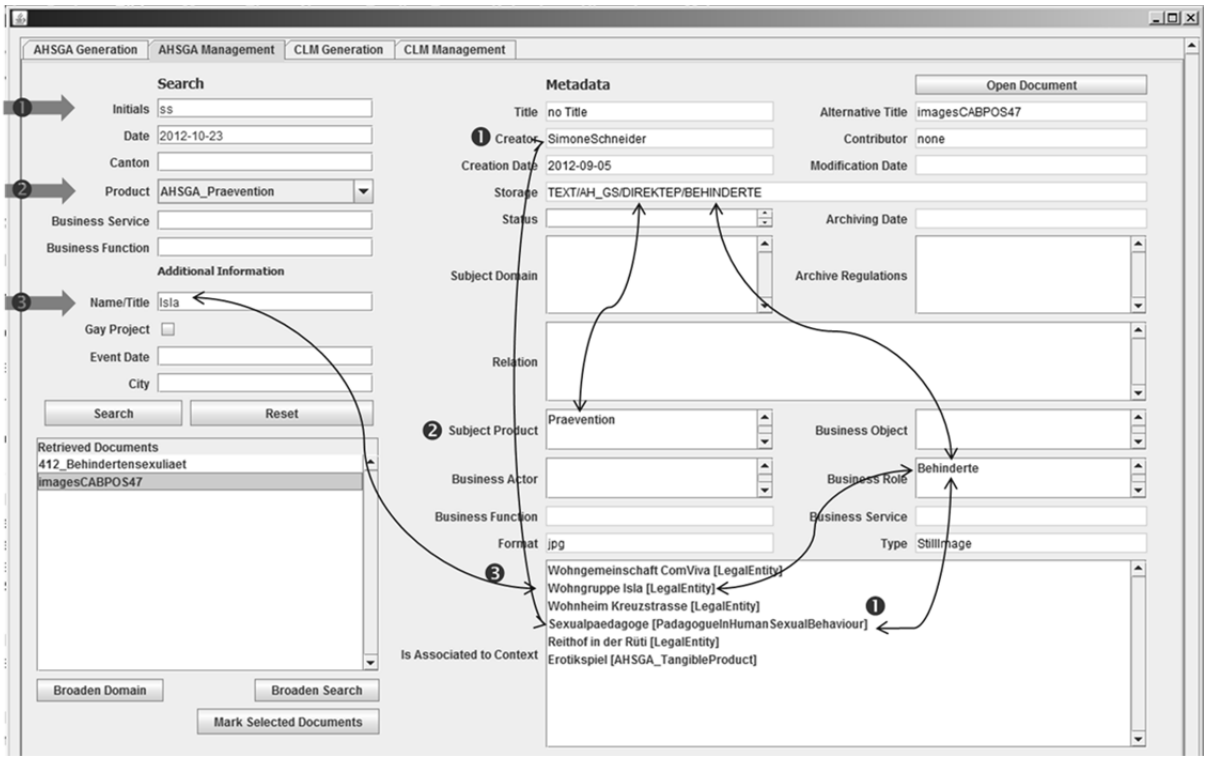


Figure 96: MeGaWorkbench Print Screen 1 for AHSGA

Another example is given in Figure 97 focussing on inferred metadata for the recorded canton 'Appenzell'. The document was stored in the directory related to the business actor 'AusbildungAuboden'; the business actor is located in 'Brunnadern', which is a city in canton 'Appenzell' (❶). The recorded task is titled 'SexualpaedagogischeEinzelarbeit'; related documents were inferred via the context of the business service 'Prävention Schule und Jugend', which is the business function 'SexualpaedagogischeEinzelarbeit' (❷). Search for task related documents was broadened based on the selected document's subject domain 'Sexualitaet'. By inferring the `skos:narrower` property documents having the narrower term in their subject domain metadata were retrieved (❸).

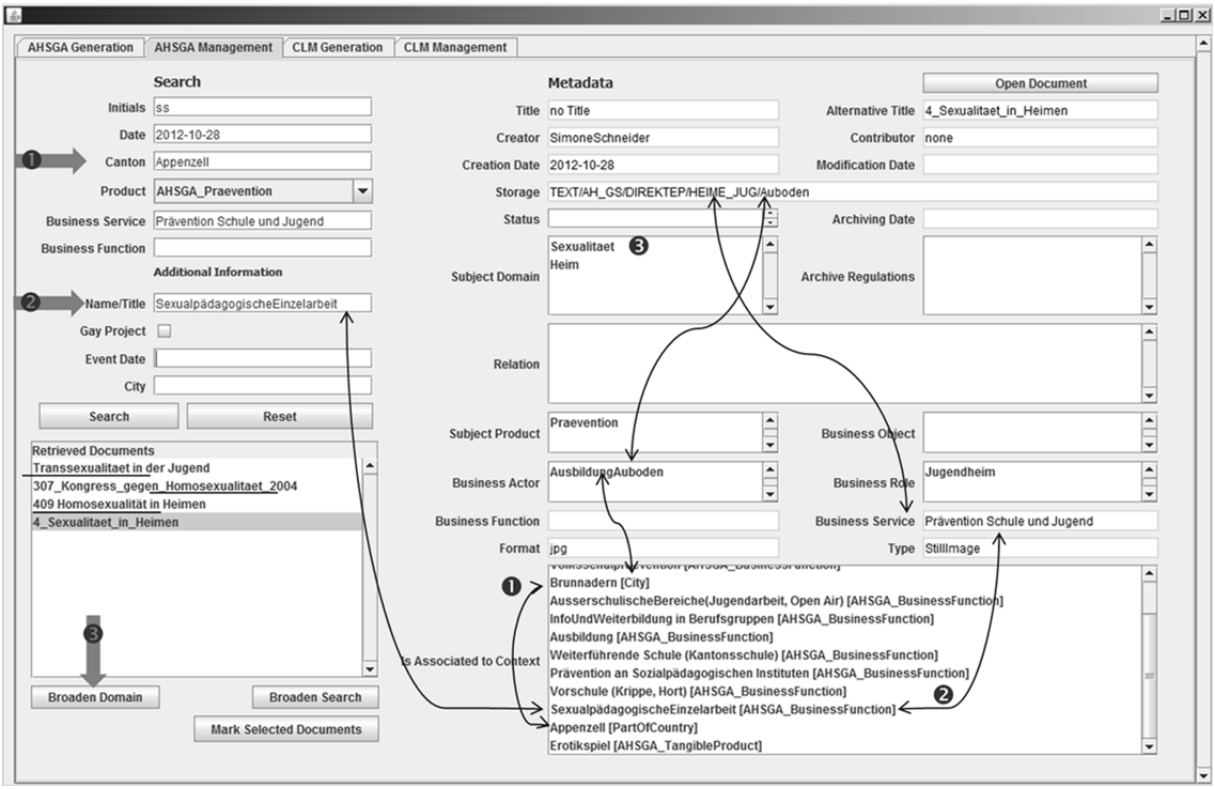


Figure 97: MeGaWorkbench Print Screen 2 for AHSGA

Documents considered relevant to the recorded task can be selected and permanently related to this task. How this functionality can be implemented in AHSGA's Information and Task Recording system was investigated in the third loop of Action Research (cf. Chapter 8.2.1, p 240 ff).

7.3.4 Symfact Prototype

Due to data protection in Symfact's case no real data was taken for proof of concept. Hence, only a small set of data was created for functional testing.

In contrary to AHSGA search for documents is not triggered by manually entering information in a task recording system (then used as search criteria) but information provided by external sources, for example by information service providers like Dun&Bradstreet[210] and Humanitarian Early Warning System[211]. For simplification in the prototype such information is entered via the Graphical User Interface (GUI). Figure 98 gives an example for a company (called 'GiveMeFive') who filed bankruptcy and is a contract partner of the company running the CLM system (called 'DontWorryInsurance'). In the lower left side of the print screen the contract is listed which is affected by the incident. At the right hand side of the figure the information is depicted relevant for identifying affected contracts. For the contracts, concluded between 'GiveMeFive' and 'DontWorryInsurance' the obligations are listed which are due in case of such a business event (❶). The contract status (❷) is set to 'under surveillance' and archiving information is provided (❸), based on general law relevant

[210] Dun&Bradstreet is a business information provider. The company is consortium partner in the APPRIS projekct. URL: http://www.dnb.com/company.html (retrieved: 28.10.2012)
[211] "The IASC Humanitarian Early Warning Service (HEWSweb) is an inter-agency partnership project aimed at establishing a common platform for humanitarian early warnings and forecasts for natural hazards". URL: http://www.hewsweb.org/hp/ (retrieved: 28.12.2012)

inferred from the information about the applicable law, stated in the contract. In addition law, relevant for archiving was inferred that is specific for the industry domain the contractee belongs to (here: 'SwissProductLiabilityLaw' as 'GiveMeFive' is assumed to be a software manufacturer with the associated GICS Code 45103010 (cf. Chapter 5.1.3, p117).

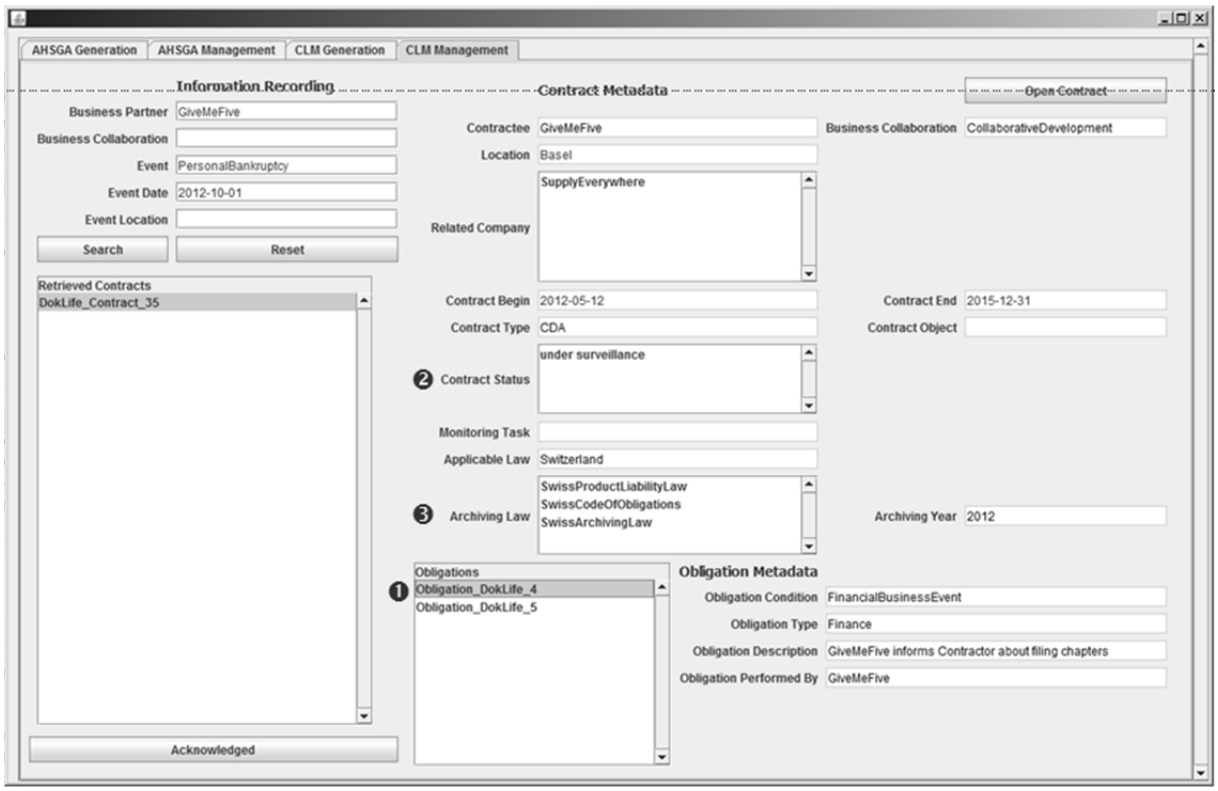


Figure 98: MeGaWorkbench Print Screen 1 for CLM

The second example, depicted in Figure 99, shows information about a force majeure event (here: 'Earthquake'). Since in seEAD events, e.g. earthquake, tsunami, flood are modelled and also information about locations it can be easily determined if a business partner is affected. In the given example, the contractee 'UNICAM', located in Camerino (❷) might be affected from an earthquake that happened in the region Marche (❷). While in the given example information on locations and the event is not particularly detailed, based on GPS and on specific event-related information like the Richter scale, this can be done. Emmenegger et al. (2012) show how the functionality is implemented in the APPRIS project for risk detection in the supply chain. More interesting for the proof of concept here is that additional information regarding the affected contract partner can be used to refine the search. For example, only these business partners shall be considered who have their production plants in the affected area, or we have a certain type of business relations with. In Figure 99 only business partners 'DontWorryInsurance' conducts a 'CollaborativeDevelopment' (❶) shall be considered,

If a business partner is identified who may be affected by the force majeure event, the contracts and the respective obligations are displayed – same as in case of a business event described above. As a result the status of the contract is changed and a monitoring task can be created automatically to track the issue.

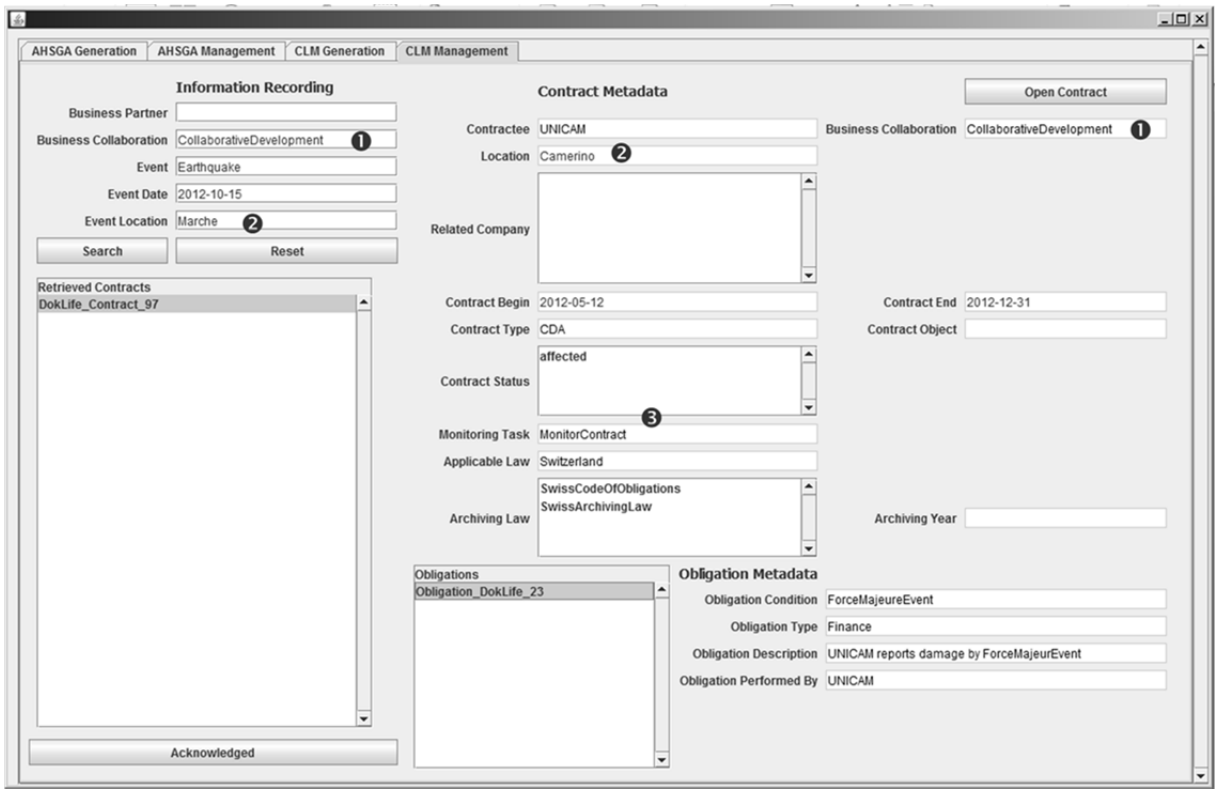


Figure 99: MeGaWorkbench Print Screen 2 for CLM

How this functionality can be used to improve Symfact's Contract Lifecycle Management system was investigated within the third loop of Action Research (cf. Chapter 8.2.2, p 243 ff).

7.4 *mintApproach Findings III*

In the previous chapter the MeGaWorkbench prototype was decribed, intended to provide proof of concept for the mintApproach. It could be shown that automatic metadata generation based on context is possible and appropriate for documents used in an enterprise. Based on the Enterprise Architecture Meta Model (ArchiMEO) two implementations of seEAD have been made, physically combined in one enterprise ontology. This proved, that ArchiMEO can be (re)used without contradiction and complemented by enterprise specifics to provide the required context, machine processable and cognitive adequate for humans.

The fact that for the MeGaWorkbench prototype AHSGA's *and* Symfact's seEAD were represented physically in *one* ontology also proves that the notion of context as "everything that is not text", i.e. as dependend on the point of view but not as a purpose-specific model, provides the flexibility Linnhoff-Popien & Strang (2004) request. Hence, rules can be defined to meet a specific purpose but the seEAD remains the same. Figure 100 sketches an example: Context is dynamically defined by the rules that allows for a purpose specific view on, respecitive use of, a semantically enriched Enterprise Architecture (seEAD) depending on the Architecture Viewpont.

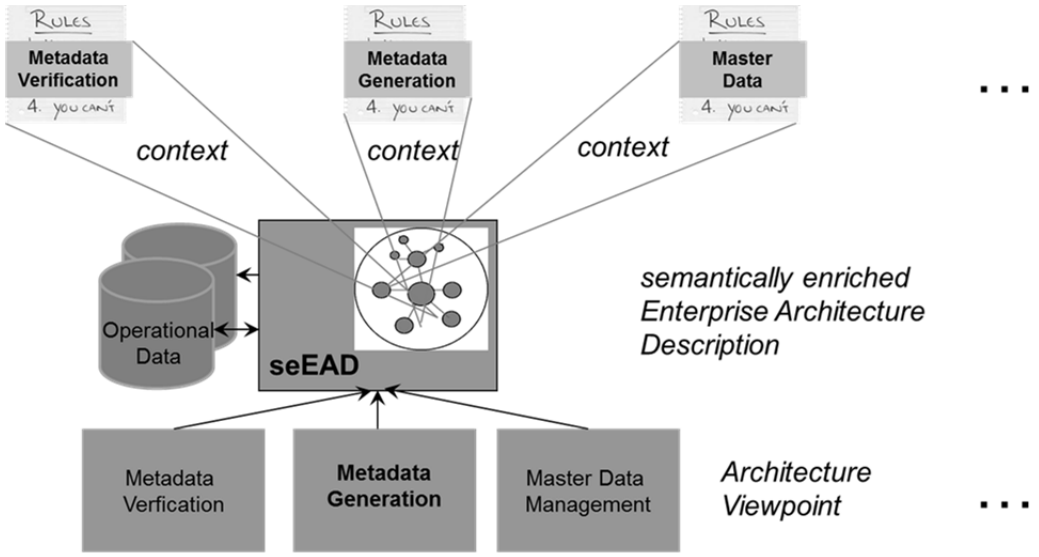


Figure 100: The Notion of Context in the mintApproach

With the MeGaWorkbench prototype it could also be proved that even if not many enterprise objects can be defined as context for documents still reasonable metadata can be generated.

From a technical point of view the mintApproach has proved feasible and the chosen meta model language (RDFS Plus) appropriate for representing the Enterprise Architecture Meta Model (ArchiMEO) and hence, the semantically enhanced Enterprise Architecture Description (seEAD) for AHSGA and Symfact; also the SPIN rules were suitable for automatic metadata generation based on context.

The mintApproach and the MeGaWorkbench prototype were evaluated by practitioners as detailed in the following Chapter 8.

8 Evaluation

Chapter 8 of my thesis provides the results of the evaluation of the mintApproach as illustrated in Figure 101.

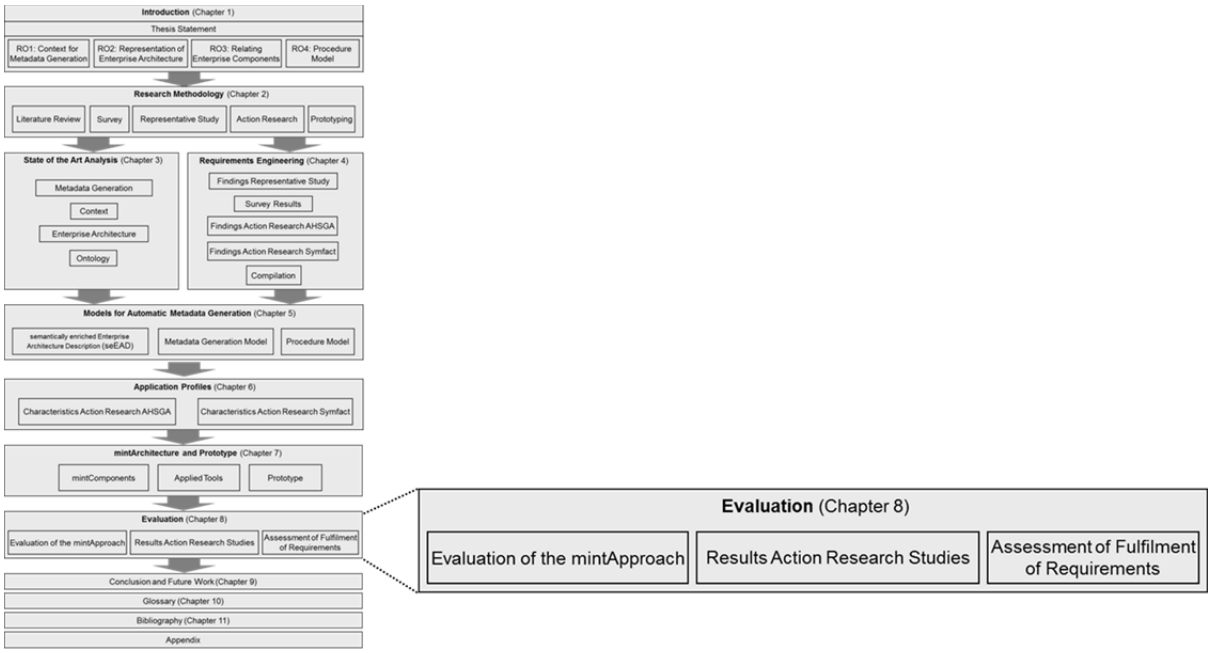


Figure 101: Position of Chapter 8 in the Overall Structure of the Thesis

Review of the mintApproach comprises three parts:

1. Evaluation of the approach using the MeGaWorkbench prototype, described in Chapter 7.3
2. In-depth analysis of the approach within loop 3 of my Action Research studies, as defined in Chapter 2.2.3
3. Theoretical comparison of requests and results, based on the requirements compiled in Chapter 4.5.

Chapter 8 is structured accordingly. It starts with the evaluation of the mintApproach giving the evaluation's subjects and aims. Then I describe the evaluation set-up, the methods I used and the evaluation criteria I applied. After that I introduce the evaluation results. In Chapter 8.2 I describe the findings of the third loop of my Action Research studies. The chapter closes with an assessment of the fulfilment (Chapter 8.3) of the requirements defined in Chapter 4.5.

8.1 *Evaluation of the mintApproach*

For proof of concept of my thesis the MeGaWorkbench (cf. Chapter 7.3) has been developed. As defined in Chapter 2.2.4 prototyping was performed through three rounds associated to the three loops of the Action Research study. Development started with a model (cf. Chapters 4.3.1 and 4.3.2), followed by a demonstrator (cf. Chapters 6.1.6 and 6.2.6) and then the executable prototype (cf. Chapters 7.3.3 and 7.3.4).

The Joint Committee on Standards for Educational Evaluation (2000, p 25) defines evaluation as "the systematic study of the applicability or quality of an item"[212]. Here, the applicability

[212] Original definition in German: "Evaluation: Die systematische Untersuchung der Verwendbarkeit oder Güte eines Gegenstands", translated by me.

of my approach realized in the MeGaWorkbench prototype. That is, the prototype is assessd with respect to the provided functionality of automatic, format-independent metadata generation; non-functional aspects like performance or user-friendliness are not considered as they are immaterial for this evaluation.

In their Handbook of Evaluation Standards the Joint Committee on Standards for Educational Evaluation (2000) provides guidelines for evaluation actitivities I considered for structuring the assessment.

8.1.1 Evaluation Subject and Aims

Subject of the evaluation is the mintApproach using the MeGaWorkbench prototype, following the argument of Church et al. (1986, p 65): "A software prototype is a functionally incomplete model of a proposed system, built to demonstrate feasibility or explore potential requirements".

Goal of the evaluation is to determine the appropriateness, capability and applicability of the mintApproach. The MeGaWorkbench is used to illustrate the mintApproach and thus, to make it easier for the evaluators to assess the underlying scientific concepts.

A second goal of evaluation is to determine the adequateness of the prototype for further development in order for the prototype to evolve into a completed system.

8.1.2 Evaluation Set-up

To meet the aforementioned goals two types of evaluators were chosen: end-users who participated in the survey on metadata in enterprises and software vendors who were interested in the approach. The selected end-users are people I interviewed within the survey I conducted for requirements analysis (cf. Chapter 4.2), and who agreed to participate in subsequent questioning. The participating software vendors are professionals in the domain of information or document management. The Action Research partners took a special role since they actively contributed to the prototyping. Hence, review of the MeGaWorkbench was performed within the third loop of the Action Research study as detailed in Chapter 8.2).

For evaluation I have chosen the qualitative method of in person interviews, triparted into demonstration of the MeGaWorkbench, question and answer sessions and guided interviews. Presentation of the prototype is based on applications scenarios, characteristic of the Action Research partner's business. The interviews were conducted based on a structured questionaire, the question and answer sessions in parallel or after the MeGaWorkbench presentation. Each evaluation lasted one hour. All evaluations were carried out in October 2012.

My assessment follows the procedure for Qualitative Evaluation introduced by Kuchartz et al. (2007). Kuchartz et al. (2007) suggests seven steps – from the definition of the evaluation subject, through code of practice for interviews to considering the results – which I used as guiding principles.

8.1.2.1 Evaluation Criteria

To assess the capability of the mintApproach, prototypically realized in the MeGaWorkbench, I reviewed the problem statement (cf. Chapter 1.1) and extracted a relatively small number of

high-level requirements to serve as key criteria for assessment, as suggested by (Church et al., 1986).

Table 41 lists the requirements that any implementation of the mintApproach should meet to solve the identified problems.

Problem Statement	Key Criteria
Manual metadata creation is too costly with respect to human effort and time and error prone (cf. Chapter 1.1.1)	Metadata can be created automatically
Most people do not like to manually create metadata (cf. Chapter 1.1.1)	Unless otherwise wanted metadata can be created without any human interaction
Full-text indexing is limited to textual documents or transcribed audio files cf. Chapter 1.1.2)	Metadata is created format-independently for all kinds of documents
Automatic extraction of low-level feature from multi-media documents like images and video documents is not useful for content related search (cf. Chapter 1.1.2)	Multi-media (text-) documents can be searched by automatically created metadata
Commercial products cannot get through (cf. Chapter 1.1.3)	Metadata creation can become a integrated part of an already existing enterprise's business information system
Relations between documents and other enterprise objects are not made explicit (cf. Chapter 1.1.4)	Relations between enterprise objects (including documents) are made explicit in an semantically enriched Enterprise Architecture Description
Enterprise governance instruments, like an Enterprise Architecture description, are not 'understandable' by machines (cf. Chapter 1.1.5) and thus not usable for document management	The Enterprise Architecture Description can be processed by machine and is adequately cognitive for humans
Operational data in an enterprise is locked in business applications and cannot be used in an integrated way (cf. Chapter 1.1.5)	Ontological representations can be mapped to operational data for broader use

Table 41: High-Level Requirements

To explore the potential of the mintApproach I considered again work of the TENCompetence project (Grigorov, 2007) and adjusted it to my evaluation needs.

Table 42 lists the adapted attributes used to measure the scope of applicability of the mintApproach.

Attributes for the scope of applicability	Evaluation Criterion **The mintApproach …**	Likart Scale 1 2 3 4 5 ☐ ☐ ☐ ☐ ☐
Practicability	… is well suited to be used in day-to-day business	
Feasibility	… is well suited for further development and implementation in a productive environment	
Relevance	… is relevant for the further development of the domain of • Document Management	
	• Enterprise Architecture Descriptions	
	• Enterprise Repository (operational data integration)	

Significance	… contributes considerably to meet the problem of metadata creation in an enterprise	
Originality	… provides a new solution to solve the problem of metadata generation in an enterprise	
	… is a novel combination of existing techniques	
Impact	… provides an added value as the solution includes an application independent part (the enterprise ontology) which can be used for other purposes, too	

Table 42: Attributes for the Scope of Applicability (based on Grigorov, 2007)

In addition to the applicability of the mintApproach the capability of the approach, visualized in the MeGaWorkbench, was evaluated. Therefore I based it on the ISO 9126 standard[213] which gives guidelines and describes the quality attributes that could be used for the evaluation of a software product. Since the standard has been criticized in several publications, for example by Al-Kilidar et al. (2005) and Botella et al. (2004) inter alia for its incompleteness and ambiguity I also considered work carried out within the European funded IST project TENCompetence[214]. In the TENCompetence project delivery report Grigorov (2007) refined and supplemented the ISO 9126 to determined evaluation criteria for a software product. Because of the specifics of a prototype I selected these criteria applicable to "a functionally incomplete model" as Church et al. (1986) put it.
Table 43 provides a sub-set of the qualitative attributes as suggested by Grigorov (2007) and the respective evaluation criteria to assess the capability of the mintApproach visualized by the MeGaWorkbench prototype.

Quality Attributes	Evaluation Criteria **Capability of the mintApproach to**	Likart Scale[215] 1 2 3 4 5 ☐ ☐ ☐ ☐ ☐
Functionality	… provide functions which meet the needs for document management	
Suitability	… provide an appropriate set of functions for automatic, format-independent metadata generation	
Accuracy	… provide metadata of the type 'descriptive', 'structural' and 'administrative' (as defined by NISO 2004)	
Interoperability	… interact with one or more specified systems	
Functionality compliance	… adhere to standards and conventions	
Usability	… be understood, and appealing to the user	

[213] The ISO/IEC 9126-1:2001 is standard for software engineering and product quality. URL: http://www.iso.org/iso/catalogue_detail.htm?csnumber=22749 (retrieved: 25.8.2012).
[214] The TENCompetence developed models and tools for the creation, storage and exchange of knowledge resources for lifelong competence development. URL: http://cordis.europa.eu/ist/telearn/fp6_tencompetence.htm (retrieved: 25.8.2012)
[215] A Likert scale is a psychometric scale commonly involved in research that employs questionnaires. URL: http://en.wikipedia.org/wiki/Likert_scale (retrieved: 26.8.2012)
The five-level Likert scale I used for evaluation is: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree and 5 = fully agree.

Quality Attributes	Evaluation Criteria **Capability of the mintApproach to**	Likart Scale[215] 1 2 3 4 5 ☐ ☐ ☐ ☐ ☐
Understandability	… enable the user to understand whether the approach is suitable, and how automatic, format-independent metadata generation is achieved	
Attractiveness	… be attractive to the user in order to	
	… improve management of non-textual documents	
	… reduce labour costs for non-productive work	
	… reduce errors on wrong document storage	
	… improve document archiving according to law and regulations	
	… enforce compliance with business rules	
	… better support document lifecycle management (e.g. reducing the risk of not knowing that documents are concerned by an event)	
	… use operational data actively (e.g. data on clients, products, business services)	
Changeability	… be adapted to enterprise specific needs	

Table 43: Quality Attributes (derived from Grigorov, 2007)

The chosen evaluation criteria allow assessing the proposed solution from the software venders' and end-users' viewpoints in terms of its usefulness, feasibility and appropriateness.

8.1.2.2 Evaluation Data

Evaluation of the prototype was performed on the data provided by the Action Research partner (for the AHSGA part) and derived from test data available from the DokLife project (for the CLM part).
As detailed in Chapter 7.3.3, AHSGA defined the most important storage locations and selected a sub-set of their documents from there. In order not to jeopardize productive IT infrastructure the structure of the file system was mirrored on a test system and the documents were imported to the respective directories. This approach led to the problem that due to the copying of the documents to another system creation date for all files was changed to date and time of copying. Hence, creation date was no longer useful for search and thus in the prototype only mutation (which remained unchanged) is considered. For evaluation a total number of 187 documents were copied to the test system. Figure 94 provides the structure of the file system mirrored on the test system.

Due to data security issues no live data was provided by Symfact. Instead test data, which I used for the CLM part of the prototype, was created by the DokLife project team, (cf. Chapter 7.3.4).

8.1.2.3 Application Scenarios for Evaluation

Scenario-based evaluation considers that the mintApproach is not thought of as yet another document management system but integrated into business application systems. Thus the

prototype should be evaluated under the aspect of how well the approach contributes to meeting the business needs rather than focusing on single data result sets retrieved for certain query terms. The application scenarios described below are used as introduction for the evaluators and provide the framework for the evaluation.
First the application scenario for AHSGA is introduced and afterwards two scenarios are described for Symfact.

AS 1: Task Reporting
Application Scenario 1 is derived from AHSGA's motivating scenario (cf. Chapter 6.1).
Initial business situation: An AHSGA employee records information about a task she had performed. The information need is to find documents relevant for this task. For several employees, with different roles (e.g. manager, pedagogue in human sexual behaviour) and different responsibilities, tasks are recorded (cf. Chapter 7.3.3 for examples) and documents were retrieved based on context-related, automatically generated metadata.

AS 2: Environmental Disaster
Application Scenario 2 and 3 are derived from Symfact's motivating scenario (cf. Chapter 6.2).
Initial business situation: In a newspaper an environmental disaster is reported and the information need is to know if business partners are affected and if so, if obligations are due. Therefore news, e.g. published on a newspaper web-site, is analysed using the same text analysis methods used for analysing the contract documents. For simplification in the prototype such information is entered via a Graphical User Interface (GUI) as described in Chapter 7.3.4.
The goal is to find out whether a business partner is affected by the reported disaster, if valid contracts exist with an affected partner and which obligations are due. First information about a disaster in a region is given in which a contract partner exists and the affected contracts and obligations are displayed. Then it is shown how relations between business partners can be used to identify if for example a supplier of a supplier is affected.

AS 3: Bankruptcy
Initial situation: An information service provider gives notice that a company filed bankruptcy in one country but at the same time opens a new production plant in another country. Information need is to know the consequences of these activities.
Analysis procedure: The provided information is analysed and seEAD is queried for contracts containing obligations due in case of bankruptcy for that very business partner. Contracts and obligations are displayed and if the user acknowledges the notification status of the contracts have changed (it is no longer valid); the respective law and retention period is inferred from seEAD and a surveillance task is triggered (cf. Chapter 7.3.4 for a detailed description).

The three application scenarios were introduced at the beginning of the respective part of the MeGaWorkbench presentation.

8.1.3 Applied Evaluation Standards

In order to ensure the high quality of my evaluation I adhered to the evaluation standards provided by the German and Swiss Evaluation Societies as briefly described below. The German Society for Evaluation[216] and the Swiss Evaluation Society[217] provide characteristics

[216] The Deutsche Gesellschaft für Evaluation (DeGEval) is an association of individuals and institutions that are active in the field of evaluation. DeGEval aims for the professionalization of evaluation, the combination of

of a good evaluation, which they call *Standards*. According to the societies an evaluation should be useful, practicable, fair/correct and precise (Gesellschaft für Evaluation 2008, Widmer et al. 2000).
The *Utility Standards* aim to ensure that the evaluation purpose and the information needs of the intended users are aligned. The *Feasibility Standards* want to ensure that an evaluation is planned realistically, well thought out, and cost- and time-conscious. *Propriety Standards* aim for unbiased conduct and disclosure of the results. The *Accuracy Standards* want to ensure that an evaluation produces valid and reliable results and justified conclusions (Joint Committee on Standards for Educational Evaluation, 2000).
Both Evaluation Societies define requirements to be met for each standard. Due to the limited space in this thesis the requirements are not explicitely quoted but
Table **44** briefly explains how the standards were addressed in my evaluation.

Evaluation Standard	How the standard is addressed in the MeGaWorkbench evaluation
Utility Standards	The scope and functionality of the prototype permits the discussion of relevant questions of evaluation while taking into account the interests and needs of the evaluation participants (Widmer et al., 2000).
Feasibility Standards	Evaluation was performed within one hour and triparted into demonstration, question and answer session and interview. Demonstration is based on a story board and interviews are based on a structured questionaire (as listed above).
Propriety Standards	Strengths and weaknesses of the prototype were fully and fairly presented, scutinized, and openly discussed so that the strengths can be further expanded and the problem areas can be treated (Widmer et al., 2000).
Accuracy Standards	Data used for evaluation has been provided by the Action Research partner AHSGA and thus is realistic and reliable. Data used for the CLM part was derived from data created for testing within the DokLife project. Within my thesis the type of data used in the prototype has been sufficiently accurately described, so that the adequacy of the information can be assessed. The prototype is designed and documented in a way that its functionality is traceable and the created results are reproducible. The evaluation results are described and documented clearly and accurately, so that it can be uniquely identified (Widmer et al., 2000).

Table 44: Evaluation Standards and how They are Addressed

8.1.4 Evaluation Results

The evaluation was carried out over a period of four weeks; from begin until end of October 2012. A total of 5 evaluation sets took place with three end-users and two software vendors. In addition, review of the prototype was performed within the Action Research Studies (cf. Chapter 8.2).
The end-users were selected because of their former involvement in my survey on document handling in enterprises and their acceptance of consecutive elective interviews. The software vendors have been chosen because of their expertise in the domain of Information Management or Document Management software.

different perspectives of the evaluation as well as information and exchange on evaluation. URL: http://www.degeval.de/index.php (retrieved: 24.8.2012)

[217] The Swiss Evaluation Society (SEVAL) is a multidisciplinary organisation with the goal to foster the exchange of information and experience in the field of evaluation between politics, administration, academia, NGOs and the private sector. URL: http://www.seval.ch/en/index.cfm (retrieved: 24.8.2012)

Table 45 provides the results of the assessment of the mintApproach. The number of entries for each level is given in the three right columns of the table. Since it turned out that a five level Likart scale is too fine-grained I reduced it to a three level scale with the following spectrum: 1 = disagree, 3 = agree with restriction[218], and 5 = fully agree. The number of entries for each level is given in the three right columns of the table.

QNo	Scope Attributes	Evaluation Criterion **The mintApproach …**	1	3	4
Q1	Practicability	… is well suited to be used in day-to-day business with appropriate adaptation		0	5
Q2	Feasibility	… is well suited for further development and implementation in a productive environment		3	2
	Relevance	… is relevant for the further development of the domain of			
Q3		· Document Management?		1	4
Q4		· Enterprise Architecture Descriptions?		2	3
Q5		· Enterprise Repository (operational data integration)?		1	4
Q6	Significance	… contributes considerably to meet the problem of metadata creation in an enterprise		1	4
Q7 Q8	Originality	… provides a new solution to solve the problem of metadata generation in an enterprise		0	5
		… is a novel combination of existing techniques		0	5
Q9	Opportunity	… provides an added value as the solution includes an application independent part (the enterprise ontology) which can be used for other purposes, too		1	4
Q10	Impact	… contributes considerably to business		1	4

Table 45: Evaluation Results for the Applicability of the mintApproach

Evaluation results are paraphrased below and completed by statements that emerged during question and answering.

All evaluators fully agree that the mintApproach is suited to be used in day-to-day business (Q1). Three evaluators emphasized the value of automatic and unsupervised metadata generation for documents for which in general, little and often meaningless metadata is available as for example images.

It was difficult for the evaluators to assess how well the mintApproach is suited for further development and implementation in a productive environment (Q2). Reservations were made with respect to the use of ontologies in a productive environment. This is explicable since none of the evaluators addressing this aspect, have any experience with ontologies yet.

All but one evaluator approved the relevance of the mintApproach for Document Management (Q3) and Enterprise Repository (Q5). There was uncertainty with respect to

[218] Because the evaluators perceived level 3 of the Likart Scale not as 'neutral', i.e. 'neither agree nor disagree' but as 'I agree but …' I adjusted the meaning to the effective statements, which were made

Enterprise Architecture Descriptions (Q4) due to the fact that Enterprise Architecture was no concern of the evaluator's who agreed with restriction. A question regarding the necessity of having a full-blown Enterprise Architecture Description for the mintApproach was posed. It could be answered by showing that in AHSGA's case only few low-level governance instruments have been represented in seEAD and that this was sufficient for the mintGeneration.

All evaluators affirmed the significance of the mintApproach to handle today's problems of metadata creation for documents in the enterprise (Q6). All but one evaluator agreed that automation of metadata generation – particular for non-textual documents – is wanted and exploiting a documents context for automatic metadata generation appears sensible.

Also all evaluators are in complete agreement that the mintApproach provides a new solution for automatic metadata generation in enterprise by exploiting the documents' context (Q7) and that it is a novel – and reasonable – combination of existing techniques, e.g. harvesting document properties and using ontologies for storing and exploiting background knowledge (Q8).

Question 9 (Q9) was to evaluate the potential of the mintApproach beyond automatic metadata generation. For example the possibilities that result from representing Enterprise Architecture Descriptions semantically enriched in a machine processable way and the flexibility of defining context depending on the viewpoint of stakeholders. The question is related to question 10 and all evaluators discussed both questions at one go.

All evaluators agreed that the mintApproach considerably contributes to business (Q10). Several aspects where discussed in the question and answer session with evaluators. First and foremost the advantage of automatic metadata generation for all types of documents (text and multi-media) was emphasized as a way to handle the huge variety of documents. Next, the possibility to improve document lifecycle management was considered, since external events can trigger retrieval of affected documents and support legally compliant, automatic archiving, based on the generated metadata. Next, the benefit of making the context of documents explicit was regarded for example with respect to changes of a document's metadata. If, for example a contract isn't valid any more because the contract partner filed for bankruptcy and the document's status has changed, this change can automatically trigger a task for monitoring outstanding bills.

Table 46 shows the results of the assessment of quality of the mintApproach, visualized in the MeGaWorkbench prototype. Again the Likart scale was reduced to same spectrum as for quality assessment, and the meaning of the third level was adapted to the evaluators' perception: 1 = disagree, 1 = disagree, 3 = agree with restriction, and 5 = fully agree.

Question No	Quality Attributes	Evaluation Criteria **Capability of the mintApproach to …**	1	3	5
Q11	Functionality	… provide functions which meet the needs for document management		0	5
Q12	Suitability	… provide an appropriate set of functions for automatic metadata generation, search and management		2	3

Question No	Quality Attributes	Evaluation Criteria **Capability of the mintApproach to …**	1	3	5
Q13	Accuracy	… provide metadata of the required type (descriptive, structural and administrative as defined by NISO 2004)		2	3
Q14	Interoperability	… interact with one or more specified systems		1	3
Q15	Functionality compliance	… adhere to regulations in laws		1	4
Q16	Usability	… be understood, and attractive to the user		2	3
Q17	Understandability	… enable the user to understand whether the software is suitable, and how it can be used for automatic metadata generation, search and management		1	4
Q18	Attractiveness	… be attractive to the user		0	5
Q19	Changeability	… be adapted to enterprise specific needs		1	4

Table 46: Evaluation Results for the Capability of the mintApproach

In the following evaluation results are paraphrased and completed by statements that emerged during question and answering.

All evaluators fully agree that the demonstrated functionality of the mintApproach addresses business objectives for document management (Q11).

Suitability of the demonstrated set of functions for automatic metadata generation, search and management was challenged with respect to implementation in a productive environment. None of the evaluators questioned the general approach but didn't feel confident to judge on a prototypical basis (Q12).

Evaluators also had problems to assess accuracy of the generated metadata since they did not know the standards and its use couldn't be visualized in the MeGaWorkbench (Q13).

All evaluators felt confident, that the simulated interoperability can be implemented in a productive system and thus, no specific system for document management is needed (Q14). However, all evaluators also indicate that they believe that true interoperability requires customization effort; one expressed his doubts in lower rank.

Answers to question five (Q15) reassembled the answers to question three since the applied standards did not become visible to the evaluators. The demonstrated possibilities of automating records management based on background knowledge, e.g. belonging of an enterprise to an industrial sector, legal domicile etc. was positively acknowledged. One evaluator expressed his concerns in a lower rank.

In general, with the MeGaWorkbench usability of the mintApproach could be communicated and was acknowledged (Q16). However, due to invisible procedure of metadata inferencing in the MeGaWorkbench and little understanding of semantic technologies, two evaluators agreed to question 6 with restrictions.

All but one evaluator understand the functionality of the MeGaWorkbench and how metadata generation, search and management can be supported (Q17). The evaluator who didn't fully agree expressed concerns about using ontologies in a productive information infrastructure.

The demonstrated possibilities of the use of automatically generated metadata, e.g. to supplement recorded tasks, to support automation of records management and to improve obligation management were considered attractive, i.e. useful and beneficial by all evaluators (Q18).

All but one evaluator considered the MeGaWorkbench adaptable to enterprise specific needs (Q19) and thus, appropriate to be used for systems engineering as suggested in the mintProcedure (cf. Chapter 5.3.3, p 142).

In summary it can be said that the mintApproach, visualized in the MeGaWorkbench prototype, was assessed appropriately for automatic format-independent metadata generation for business documents. Using the context for metadata generation is considered promising, particularly regarding multi-media documents, respectively documents with little, meaningless or even wrong document attributes. The mintApproach is assessed as beneficial for enterprises and contributes significantly to meet business needs for handling the ever-increasing amount of unstructured information with as little human effort as possible.

8.2 *Summary of Action Research Loop 3*

Second part of the review of my approach is done within the third loop of my Action Research studies. In addition to the evaluation of the MeGaWorkbench prototype by the Action Research partners, specifics of the studies have been addressed. The third loop of the studies was executed between August 2012 and October 2012.

8.2.1 Third Loop of Action Research With AHSGA

The third and final loop of Action Research with AHSGA focussed on the assessment of the mintApproach, visualized in the MeGaWorkbench prototype.

8.2.1.1 Results of the Third Loop of Action Research With AHSGA

In the following, results of the third iterative cycle are provided as specified within the Action Research method (cf. Chapter 2.2.3).

1. Presentation and evaluation of the executable prototype
   Since the MeGaWorkbench prototype was an evolutionary development within the Action Research study and thus, per se meets the partner's needs, evaluation with AHSGA focussed on practical implementation and operational use, see sections three and four below.
   The MeGaWorkbench was presented to AHSGA's manager, to a pedagogue in human sexual behaviour and to the manager's assistant. Based on examples of recorded tasks the audience gave, the appropriateness of the retrieved documents was assessed. Except in one case the listed documents were considered correct and documents relevant to the recorded task could be selected. In the aforementioned case the selected document was stored in a wrong directory leading to the generation of false metadata.
   However, despite the throughout positive evaluation of the MeGaWorkbench, it became clear that for a final assessment the mintGeneration ought to be performed in a productive environment.

2. Identified change requests and supplementary requirements
   Since AHSGA's documents are in the German language the standard POS tagger GATE provides, has not been applicable. Instead the TreeTagger from the University of Stuttgart was used for identifying nouns in a document's file name. Since the method remains the same – both for parts of speech segmentation and for composite word building – MeGaWorkbench functionality has not been affected by the change.
   Since the tagger expects proper sentences, problems occurred if file names were written without blanks or other terms segmenting characters like underscores and hyphens. In these cases the subject domain could not be parsed. For testing some of the names have been changed accordingly.

3. Captured questions to be answered
   As required by the Action Research method (cf. Chapter 2.2.3) questions that arose during the development of the MeGaWorkbench prototype are to be answered in collaboration with the Action Research partner. In this loop of the Action Research study, questions regarding implementation of my approach in the ITRS have been analysed and discussed with the developer of the software. Since currently ITRS uses a MS Access Database, data would have to be migrated to a SQL server (for example MySQL to avoid extra licence fees). According to the database developer, this would not require much effort. ITRS provides an API for export and import. Thus, data recorded in ITRS could be passed to the MeGaSystem (ie.e. to a productive implementation of the mintApproach) as designed (cf. Chapter 12.4.12), the query could be executed in seEAD (cf. Chapter 12.4.14) and the result list plus the documents' metadata could be displayed in a separate window (cf. Chapter 12.4.15). After the selection of one or more documents the import into ITRS could be triggered. That is, the MeGaSystem would pass path and filename of the document to ITRS. Within ITRS a link would then be created to these documents. This functionality would replace the existing way of manually browsing the explorer for documents and linking them with a reported task. The ITRS developer considered my approach a valuable enhancement of the ITRS fitting nicely into existing functionality.

   Another question is related to the enterprise specific maintenance of seEAD. Although user-friendly ontology management software is available, as for example the tool TopBraid I used in my work, a knowledge engineer would be necessary to apply new concepts, relations, mappings or rules and to adjust, if required the API to the ITRS. However, this limitation is well understood by AHSGA as it is the same as for ITRS maintenance: changes on the database are done by a database expert.

4. Actions to meet the requirements
   With the developer of AHSGA's ITRS a strategy has been sketched to show how the functionality demonstrated with MeGaWorkbench prototype could be implemented in the productive system. Figure 102 depicts a print screen of AHSGA's ITRS for recording information on a task related to the intangible product 'Prävention' (prevention). Instead of searching for documents manually by clicking the 'Document' button, documents are retrieved automatically – as described in the previous sections – based on the entered data for reporting a task. As shown in Figure 102 as a result the retrieved documents can be listed and selected.

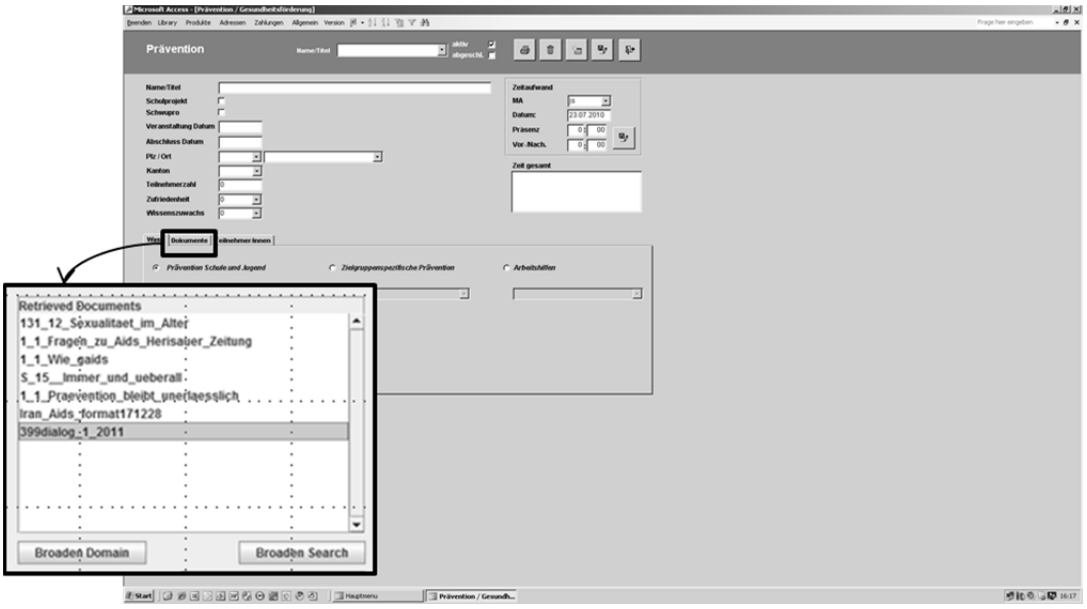


Figure 102: ITRS Print Screen to-be

Suggestions for participants related to a task could be provided, too. By clicking the other tab called 'TeilnehmerInnen' participants could be inferred from seEAD, too, if need be and enhanced information on clients is captured

5. Share with others (departmental meeting, publication, conference, etc.)
   To share results of the third loop of Action Research with others the executable prototype was presented and discussed with the Action Research team on October 11th, 2012.

8.2.1.2 Research Question Addressed Within the Third Loop of Action Research With AHSGA

As proposed in my research design (cf. Chapter 2.1, Table 1) also within the third loop of Action Research with AHSGA, several research questions were addressed.

1. To measure quality of generated metadata AHSGA assessed the documents retrieved for reported tasks in the MeGaWorkbench (answer to RQ 7). It showed that all documents retrieved are related to the task (generated metadata is correct) but that the hit list might get too long in a productive environment, i.e. precision might be compromised. Hence, to determine quality of generated metadata the mintGeneration ought to be tested in productive use.

2. It is not possible to provide a general rule for determining the boundary between enterprise objects represented in a semantically enriched Enterprise Architecture (i.e. in an ontology) and enterprise components. Where to draw the line depends amongst other things on how well both types of representation can be mapped. In the case of enterprise components represented in a relational database, mapping can be done more easily (cf. Chapter 5.1.5) than if represented in a non-relational data structure. If paper copies are the prevalent representation format migrating to a solely ontological representation could be beneficial. As suggested in the procedure model I provided (cf. Chapter 5.3.3) defining the appropriate format for representing enterprise objects is a major task when setting up automatic metadata generation in an enterprise. For productive use of my approach in AHSGA's case low level governance instruments, up to now recorded in paper folders,

would be represented in seEAD – as they already are in the prototype – whereas records of the reported tasks would be kept further in the ITRS database (answer to RQ12).

3. More difficult, also from an enterprise specific point of view, is to determine redundancies of representations of the same enterprise object. That is in AHSGA's case enterprise objects represented in seEAD *and* ITRS, as for example information on clients, products and documents. The problem can partly be solved by technical solutions, e.g. keeping all instances in the productive relational database (cf. Chapter 5.1.5), but redundancies are inevitable with respects to concepts and their relations. In AHSGA's case the issue is somewhat mitigated as core data remain rather constant, i.e. the type of clients (e.g. youth centers, schools, public bodies), the type of offered services, employees' roles, business functions, etc. does not change over years. Hence, the redundancies developed in the course of prototyping were accepted and considered also for a productive implementation (answer to RQ13).

4. As already shown in Chapter 6.1.6 the first two phases of the procedure model (cf. Chapter 5.3) have proved to be suitable for automatic, format-independent metadata generation. In loop three of Action Research with AHSGA the next phase, i.e. 'Realization', has been successfully applied with respect to prototyping and evaluation. As far as applicable in the Action Research study the procedure model proved appropriate for setting-up, conducting and utilizing metadata (answer to RQ14).

5. The MeGaWorkbench prototype is the final stage of evolutionary prototyping as defined in Chapter 2.2.4. Although the value of an executable prototype is widely accepted (cf. Chapter 5.3.3), it clearly is not sufficient to determine how quality of the metadata generation process could be improved. Despite the obvious aspects, like true integration of the functionality into the ITRS, automatic, format-independent metadata generation must be used in productive operation to determine improvements. If so mechanisms to log user interactions, for example for capturing changes of metadata, could be added; the logs could be analysed and based on the results corrections could be made, e.g. adaptions of rules (answer to RQ17).

8.2.2 Third Loop of Action Research With Symfact

The third and final loop of Action Research with Symfact focussed on the assessment of the mintApproach, visualized in the MeGaWorkbench prototype.

8.2.2.1 Results of the Third Loop Action Research With Symfact

In the following results of the third iterative cycle are provided as specified within the Action Research method (cf. Chapter 2.2.3).

1. Presentation of the executable prototype
   Also in Symfact's case the MeGaWorkbench prototype was developed in an iterative process within the Action Research study. Since no real data could be used in the CLM part of the MeGaWorkbench and the functional principles were already assessed on the basis of the demonstrator (cf. Chapter 6.2.6), evaluation of the prototype was forgone.
   However, the MeGaWorkbench was presented to Symfact's manager, chief developer and two software engineers and the benefits of the mintApproach were acknowledged.

2. No change requests or supplementary requirements emerged in the third loop.

3. Captured questions to be answered
   Since in the prototype information about events (force majeure and business events) must be entered manually, ways of automation in a productive system were discussed. The problem has already been addressed within the APPRIS project in which information sources, for example provided by LexisNexis, Dun&Bradstreet and Humanitarian Early Warning System) were integrated through web-services (Emmenegger et al., 2012). Similarly a source-processing engine of the CLM system could monitor and analyse the information sources and extract the data to query seEAD ("Earthquake", "Bankruptcy", "Merger&Acquisition", etc.)

4. No further actions were proposed. Figure 103 depicts how functionality, prototypically implemented in the MeGaWorkbench can be integrated into the CLM system.

Figure 103: CLM Printscreen With Simulated Notification

5. Share with others (departmental meeting, publication, conference, etc.)
   To share results of the third loop of Action Research with others the executable prototype was presented to the Action Research team on October, 11th, 2012.
   A summary of the results has also been published in the DokLife Project Description available at the projects web-site (http://www.doklife.ch/publications/).

   Part of the results have been published in (Thönssen & Lutz 2012).

8.2.2.2 Research Questions Addressed Within the Third Loop of Action Research With Symfact

As provided in my research design (cf. Chapter 2.1, Table 1) also within the third loop of Action Research with Symfact several research questions were addressed.

1. In Symfact's case quality of generated metadata was assessed with respect to its contribution to solve business problems like risk management and records management (answer to RQ 7).
   It could be shown, that quality is appropriate to improve the handling of obligations represented in contract documents by providing the missing link to external business events (e.g. bankruptcy) or force majeure events (e.g. earthquake), and thus contribute significantly to risk managment.
   With the achieved quality of generated metadata records management for contract documents can be improved as well. If a contract partner goes bankrupt all contracts

become invalid. The documents' retention period starts by then which can be triggered automatically based on seEAD.

2. In Symfact's case the boundary between enterprise objects represented in seEAD and enterprise components in the relational database of the CLM system have been stipulated by the DokLife project. As detailed in Chapter 6.2.5 only these metadata are represented in seEAD needed for inferencing, all other metadata is to be stored in the CLM system (answer to RQ12).

3. The remaining redundancy is unavoidable (answer to RQ13). However, dealing with redundancy, respectively with data consistence isn't new and one can draw upon existing approaches. Lee et al. (2006), for example, indicated in their work on an operational product ontology system, two approaches for dealing with ontology-dabase-redundancy: a database-level approach (updates automatically requested on database level) and an application-level approach (update requests triggered by specially designed logging modules).

4. As already shown in Chapter 6.2.6 the first two phases of the procedure model (cf. Chapter 5.3) have proved to be suitable for automatic, format-independent metadata generation. In loop three of Action Research with Symfact the next phase, i.e. 'Realization', has been successfully applied with respect to prototyping and evaluation (answer to RQ14).

5. As for AHSGA, the MeGaWorkbench prototype is the final stage of evoluationary prototyping as defined in Chapter 2.2.4, and suitable for evaluating the approach. To determine how quality of the metadata generation process could be improved automatic, format-independent metadata generation must be used in productive operation. In Symfact's case the existing CLM system should be enhanced by mechanisms to log user interactions, for example for capturing changes of metadata. The logs could be analysed and based on the results the respective rules could be adapted (answer to RQ17).

8.3 *Assessment of the Fulfillment of Requirements*

The third part of the review of my approach is theoretical assessment of the fulfilment of the requirements listed in Chapter 0, that have been derived from literature review (cf. Chapter 3), the Representative Study (cf. Chapter 4.1), the survey on document handling (cf. Chapter 4.2) and particular requirements specified by the two Action Research partners (cf. Chapter 4.3). Table 47 provides the list with a self-assessment of my work summarized in the far right column of the table.

Requ. No	Requirement	Criteria for Measurement	Assessment of Work
Q1	implement and evaluate automatic metadata generation within a real use case	implementation and evaluation is done within two Action Research studies based on real data and documents	evaluation has been done with AHSGA and Symfact as designed (cf. Chapters 8.2.1 and 8.2.2)

Requ. No	Requirement	Criteria for Measurement	Assessment of Work
Q2	build seEAD in the right balance of expressiveness and decidability	seEAD is expressive enough to model the required knowledge but remains decidable	the use of seEAD within the two Action Research studies has led to pragmatic decisions on the representation and rule language of seEAD (cf. Chapter 5.1)
Q3	use standards to build seEAD	seEAD is based on evaluated and approved standards	seEAD is based on the ArchiMate standard and enhanced by the Dublin Core standard (cf. Chapter 5.1.3)
Q4	ensure quality of seEAD by sticking to essence	only those enterprise objects are represented in the ontology and related to enterprise component that are required	when enhancing ArchiMEO to enterprise specific needs I strictly stuck to the fitness-for-purpose principle, introduced by Chen et al. (2008); supported by the chosen modelling procedure based on the methodology of Uschold & Gruninger (1996) only these concepts, relations and rules were modelled that represent business requiremens (cf. Chapters 6.1 and 6.2)
C5	provide stakeholder specific views on seEAD	depending on the stakeholder (e.g. AHSGA or Symfact) other context is used for metadata generation or document lifecycle management	stakeholder specific views on seEAD are provided the individual context models (cf. Chapters 6.1.2 and 6.2.2)
C6	use context of documents for active support of document life-cycle-management	based on context, dependencies or implications of change on documents are identified, e.g. if a product changes what specifications are affected	
C7	adapt and enhance seEAD based on enterprise specific governance instruments	seEAD reflects content of enterprise specific governance instruments like a management handbook	seEAD represents AHSGA's low level governance instruments like the Quality Management Manual, the Organisations Structure etc. (cf. Chapter 4.3.1)
C8	represent governance instruments formally	governance instruments are modelled in an enterprise ontology	representation of AHSGA's low level governance instuments is formalized in RDF-Plus and SPIN rules (cf. Chapter 6.1.4) Furthermore, regulations determine compliant records management, are formally modelled in seEAD and forced by SPIN rules (cf. Chapters 6.1.4 and 6.2.4)

Requ. No	Requirement	Criteria for Measurement	Assessment of Work
C9	relate directory structure to seEAD	storage location of a document is parsed and analysed for information about the document (context) (Soules & Ganger, 2005)	the directory structure is reflected in AHSGA's context model (cf. Chapter 6.1.2)
MD10	provide context for metadata generation	business objects - documents represent – and their relations are formally modelled in the semantically enriched Enterprise Architecture Description (seEAD)	
MD11	automatically generate metadata regardless of a document's type	content related metadata is automatically generated for all kinds of documents (text, image, sound)	metadata can be generated for documents of all common file formats; since the starting point of the procedure is the harvest of document properties and the chosen NLNZ harvester provides an 'universal' adapter for unknown file formats, presumingly at least a minimal set of attributes will be harvested in any case (cf. Chapter 7.2.1)
MD12	harvest document properties of the following file formats doc, pdf, ppt, xls, jpg, gif, png, mp3, mp4	all document properties of the specified formats are harvested	all formats specified by AHSGA are supported (cf. Chapter 7.3.3);
MD13	determine retention period of a document based on qualitative instead of formal criteria	metadata (dceo:archiveDate) is generated automatically based on a document's context (the business object a document represents, the branch of trade the enterprise is in, the law that have been obeyed, etc.)	retention period and determine law is inferred from a document's context and created automatically (cf. rules for AHSAG Chapter 6.1.4: AHSAG_IR15 – IR 17; rules for Symfact Chapter 6.2.4: Symfact_IR2 – IR3)
MD14	specify rules for inferring context to automatically generate metadata	generic rules are specified (like "for all primary context elements of a metadata seed all n-ary context elements are inferred as metadata candidates")	rules are specified for AHSGA (cf Chapter 6.1.4, rules AHSGA_IR8 – AHSGA_IR_12)
MD15	derive rules for analysing file names based on low-level governance instruments (e.g. naming conventions)	rules are defined reflecting naming conventions, e.g. 'if an employee's number in a file name = 1 then 'creator is <employee name>'	rules are specified for AHSGA (cf Chapter 6.1.4, rules AHSGA_IR2 and AHSGA_IR_23)

Requ. No	Requirement	Criteria for Measurement	Assessment of Work
R16	provide interface between seEAD and existing platforms	seEAD can be used by target systems, like AHSGA's Time Recording System, or Symfact's Contract Lifecycle Management System	interface is defined (cf. Chapter 5.1.5) but not implemented in the prototype as the development does not contribute to my thesis
I17	keep metadata generation solution independent from upstream and downstream function	the solution is independent from harvesting or extraction tools (upstream function) and from Information- or Document-Management-Systems (downstream function)	the Metadata Generation Architecture consists of several components which can be used largely independent of each other (cf. Chapter 7.1)
I18	enable machine-processing of seEAD	the semantically enriched enterprise architecture can be used by humans and machines alike	seEAD is machine processable, either stand-alone using an ontology managing system (e.g TopBraid) or embedded in prototype or production system (cf. Chapters 7.2.3 and 7.3)
I19	create metadata with as little user interaction as possible	automatic metadata generation is performed in the background and no extra effort from the user is required	metadata generation is automated transparent for the user and can be performed without any manual interaction; incidental business activities – like reporting on a task – are exploited for implicit management of metadata (cf. Chapter 7.3)

Table 47: Fulfilment of Requirements

As shown above, all requirements specified for automatic, format-independent metadata generation have been met with respect to the defined evaluation criteria.

9 Conclusion

In this chapter the findings of my study and the conclusions I deduced are stated. Also a summary of my contributions to related research and suggestions for further research are provided.

9.1 Summary of Findings

In this dissertation, the aim is to assess how metadata for documents used in an enterprise can be generated automatically regardless of their format and based on their context. Returning to the objectives posed at the beginning of this study, it is now possible to state that a semantically enriched and formally represented enterprise architecture description can provide the context from which metadata can be inferred automatically and un-supervised for enterprises' documents. I constructively proved this using design methodology by developing a new approach – which I call mintApproach, depicted in Figure 104.

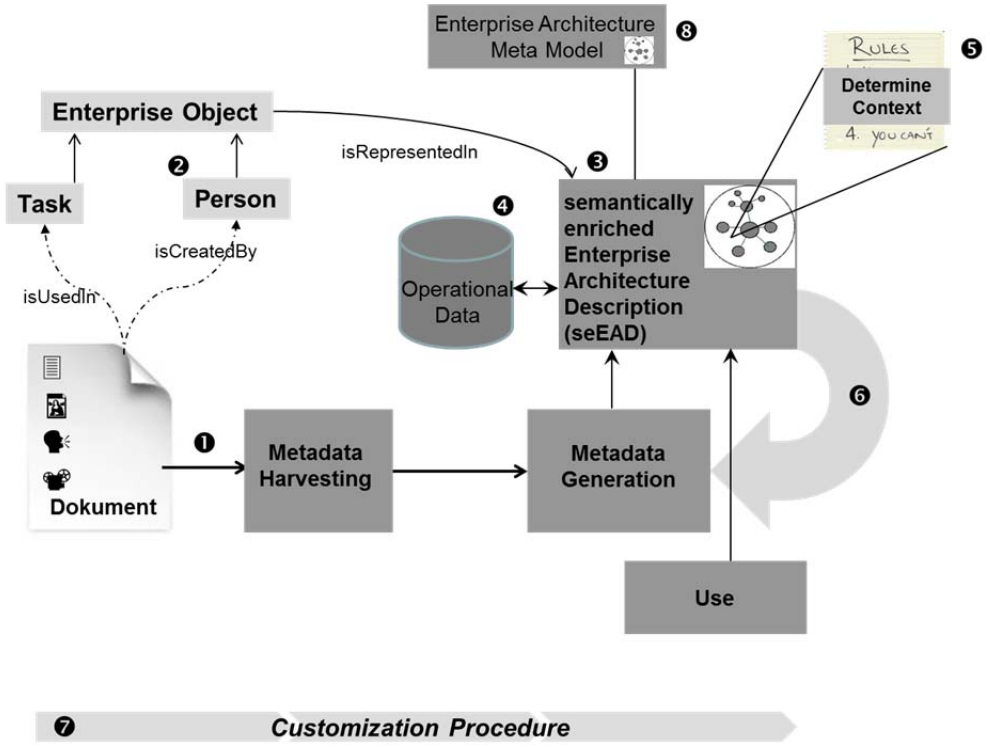


Figure 104: The mintApproach Résumé

My research showed that it is possible to automatically generate metadata for any kind of document (text, image, audio and video), used in an enterprise for business purposes (❶). A document is considered an enterprise object related to other enterprise objects such as the person who created the document and the task for which the document is used (❷). The enterprise objects and their relations are determined in an enterprise architecture, which can be represented in a semantically enriched enterprise architecture description (❸). Since enterprise objects can be also represented in operational data stores, ontological and relational representations of the same enterprise object are related (❹) and build an enterprise repository. The enterprise objects related to a document build the document's context. Based on rules (❺) the document's context can be inferred for automatic metadata generation (❻).

Thus, for example the task a document is used for can become a metadata, as well as the product the creator is responsible for. Because definition of enterprise objects and context used for automatic metadata generation is enterprise specific a customization procedure was investigated (❼). In the following my findings are detailed.

Automatic creation of metadata is an approach designed to meet business needs for managing the ever increasing amount of unstructured information in an enterprise. As the use of multi-media documents increases and automatic content analysis does not provide meaningful information for business use, metadata is particularly important for this type of document form. However, document properties, automatically created by document creation software and operating systems, are few, of limited use for business purpose (e.g. file size, image size) and unreliable. They might be wrong (e.g. the author of a document is not the creator), meaningless (e.g. a randomly generated file name of an image) or even completely missing. Since creation and use of a document takes place within the context of business activities, e.g. a task that is to be performed by a person, this person is then responsible for a product, the product is worked out in a service (❷) I was able to prove that this information can be used for automatic metadata generation.
Context of business documents is composed of enterprise objects related to it. Enterprise objects – like a person, a role, a function, a task etc. – are defined and related to each other according to an enterprise's conception of its business. In an enterprise architecture description this concept is made explicit (❸). What context is, in a specific enterprise and for the business documents of this enterprise, can be defined by rules (❺). This provides huge flexibility since context is determined by its use and not pre-defined in a model.

Since the main goal of the mintApproach is to automatically generate metadata based on the document's context, a machine processable enterprise architecture description is shown to be a suitable basis (❻). Graphical or textual representations are insufficient but semantic technologies satisfy the requirements. Although there is broad consensus that representing Enterprise Architecture knowledge in an ontology is advantageous, enterprise architecture modelling (and describing) and ontology modelling have only recently merged. Regarding an enterprise ontology as a formal representation of enterprise architecture has only become a research topic in the last two years and previously existing approaches were not sufficient for the mintApproach.

Furthermore, only a few approaches consider Enterprise Architecture Frameworks in their work. From the many frameworks available I regard Zachman's framework (Zachman, 2003) particularly well suited to ensure quality of an enterprise ontology with respect to its completeness. The ArchiMate framework (The Open Group, 2009b) is essential because of its architecture description language although its semantics are basically undefined (Ettema & Dietz, 2009).

Analysis of research on representation languages for ontologies has shown that there is no 'silver bullet for formalization'. Even if a computational level of formalization is given, there are several modelling languages – or dialects – that could be selected. For automatic metadata generation based on context I relied on W3C standards for ontology representation (RDFS & OWL), on the W3C recommendation for ontology managing (SPARQL), and on the W3C submission for rule formalization (SPIN).

An enterprise architecture description is regarded as part of an enterprise repository, comprising all enterprise objects constituting an organisation despite their representation. Analysis of methods for linking enterprise objects stored in an ontological representation to

enterprise objects stored in existing non-ontological data-stores led to my decision to use a 'direct and single mapping strategy' and ODBA (ontology-based database access) for querying (❹).

Within the mintApproach I developed several models for automatic, format-independent metadata generation. First and foremost I developed a model for describing enterprise architecture in a way that is machine executable but also cognitively adequate for humans. That is, I developed a semantically enriched Enterprise Architecture Description (seEAD) (❸), which is used to infer a document's context for automatic metadata generation (❻). To decrease ontology development costs and better exploit the potential of a semantically enriched enterprise architecture description I created a meta model (❽) that can be used to model the enterprise-specific semantically-enriched Enterprise Architecture according to the Meta Object Facility specification (OMG, 2011b). Thus, it can serve as 'General Enterprise Model' or 'Core Enterprise Ontology' as suggested by Fox & Gruninger (1998) and Bertolazzi et al. (2001), respectively.

To ensure quality and appropriateness of the Enterprise Architecture Meta Model (ArchiMEO) (❼) I based it on the ArchiMate standard, which has been enhanced by other standards, e.g. Dublin Core and complemented by concepts derived from existing enterprise ontologies and requirements from real life. Contradictions discovered in the ArchiMate standard have been resolved and its semantics have been refined. To formally represent the Enterprise Architecture Meta Model a pragmatic approach was chosen: strong enough to express its content in a sound and formal way but 'light' enough to remain executable. This led to RDFS-Plus as the meta modelling language.

Since the mintApproach provides general models to be customized to enterprise specific needs and requirements in order to determine the context for automatic metadata generation, this task is supported by a procedure model (❼). The procedure model is a modified waterfall method comprising four phases (analysis, modelling, realization and operation). Focus is on mintApproach specific aspects, like assessment of governance instruments to determine documents' context, phrasing competency questions to determine enhancements of seEAD and defining the borderline between enterprise objects stored in the ontology and already existing data stores (❹).

The combination of the results as stated above makes the mintApproach unique.

To validate my findings, the mintApproach was verified for correctness, suitability and practical relevance within two Action Research studies. Embedded in the Action Research studies a procedure model for setting up automatic metadata generation in an enterprise and a prototype were incrementally developed and evaluated by the Action Research partners. The executable prototype was also evaluated by a broader audience of end-users and vendors of information management software. The evaluation gave evidence of the appropriateness of my approach and of its great potential benefit to the users.

Even though this thesis presents a significant contribution to the field of information management in an enterprise, there are some weaknesses and limitations that need to be considered.
In the case of few meaningful document properties, too much metadata might be generated. Thus, additional information on a document's creation – for example that which is derived from the task the document is used for – could provide input for refinement. Further research might also investigate if the type of relations between enterprise objects could be used to

weight created metadata (e.g. the task the creator of a document is considered more important than the organisational unit the person works in), if similarity measures could be added, for example to rank related documents, or if metadata which has been generated traversing two or more paths, could be considered of greater importance.

Although proof of concept was given with the MeGaWorkbench prototype it was evaluated in a test environment as defined with the Action Research partners in order not to jeopardize their productive environment and avoid unnecessary integration costs. Hence, a long-term and broad study of the mintApproach in an enterprise's productive working environment would provide valuable input to improve automatic metadata generation as well as the Enterprise Architecture Meta Model.

9.2 Contribution and Suggestion for Further Research

Recent research on the management of unstructured information in enterprises also considers the context of a document as source for semantic metadata. The approaches differ significantly to mine as the context models that are used are somehow arbitrary and miss a sound foundation like an Enterprise Architecture. The mintApproach helps to improve recent approaches with respect to rectification of documents' metadata. Findings of this study contribute to the understanding and quality of context models in the business domain. For example, instead of controlled vocabularies (e.g. FOAF: Friend Of A Friend) and proprietary ontologies the Enterprise Architecture Meta Model supports a context model that is standard-compliant.

The ArchiMEO ontology is available under the Creative Commones License[219] and can be downloaded from the ArchiMEO web-site: www.ec-ikm.net/archimeo.

The study has also gone some way towards enhancing our understanding of the use of enterprise architecture descriptions on operational level. It contributes to the third wave of Business Process Management (Smith & Fingar, 2003) supporting the synthesis and extension of existing approaches into a unified whole. With the mintApproach documents are described by the generated metadata in a way that allows for identifying and linking them automatically as resources to knowledge intensive business processes. Thus, documents can become part of collaborative process knowledge management and maturing as suggested for example by Brander et al. (2011).

Research based on results of the mintApproach has already started. In the APPRIS project the Enterprise Architecture Meta Model of the mintApproach was taken to model the enterprise architecture description used for risk detection (Emmenegger et al., 2012).

Most recently use of Enterprise Architecture Meta Model of the mintApproach is investigated in the SEEK!sem project. In this project the focus is on determining similarity between enterprise objects, for example between two or more documents, between documents and electronic folders, between folders etc. To do so already existing representations of the considered enterprise objects in the Enterprise Architecture Meta Model will be enhanced by the application specific ones as foreseen in the mintApproach. A number of possible future studies using the same set up are apparent.

[219] Creative Commons. The chosen license allows commercial use and modifications. Modifications are to be published under same license. URL: http://creativecommons.org/ (retrieved: 28.11.2012)

This research will further serve as a base for future studies on the use of an enterprise repository, connecting enterprise objects represented in a semantically enriched enterprise architecture description to operational data stored in business applications. Currently a Master's thesis is on its way, supervised by me, to investigate how such an enterprise repository can support business agility and interoperability.

10 Glossary

The symbol ☞ is used to refer to the preferred term where the explaination can be found.

Term	Definition
administrative metadata	"Metadata used in managing and administering information resources, e.g., location or donor information. Includes rights and access information, data on the creation and preservation of the digital object" (DCMI, 2010)
annotation	1) a comment or instruction 2) the act of adding notes[220] In the digital world annotation is often perceived as a synonym for metadata (Ruvane, 2005)
application profile	"A declaration of the metadata terms an organization, information resource, application, or user community uses in its metadata. In a broader sense, it includes the set of metadata elements, policies, and guidelines defined for a particular application or implementation. The elements may be from one or more element sets, thus allowing a given application to meet its functional requirements by using metadata elements from several element sets including locally defined sets" (DCMI, 2010)
ArchiMate	ArchiMate is a Technical Standard, which has been developed and approved by The Open Group. ArchiMate V1 has been published in 2009 (The Open Group, 2009b); ArchiMate V2 has been published in 2012 (The Open Group, 2012). The role of the ArchiMate Enterprise Architecture Framework is to provide a graphical language for the representation of enterprise architectures (The Open Group, 2012).
ArchiMEO	ArchiMEO is the core ontology of seEAD, which includes the Top Level Ontology and an ontological representation and enhancement of the ArchiMate standard
architecture	According to The Open Group "architecture" has two meanings: 1. "A formal description of a system, or a detailed plan of the system at component level to guide its implementation" 2. "The structure of components, their inter-relationships, and the principles and guidelines governing their design and evolution over time" The Open Group URL: http://www.opengroup.org/architecture/togaf8-doc/arch/ (retrieved: 5.12.2012)

[220] Source: WordNet 3.0 online. URL: http://wordnetweb.princeton.edu/perl/webwn?s=annotation&sub=Search+WordNet&o2=&o0=1&o7=&o5=&o1=1&o6=&o4=&o3=&h= (retrieved: 14.7.2010)

Term	Definition
architecture description	"An architecture description is a formal description of an information system, organized in a way that supports reasoning about the structural properties of the system. It defines the components or building blocks that make up the overall information system, and provides a plan from which products can be procured, and systems developed, that will work together to implement the overall system. It thus enables you to manage your overall IT investment in a way that meets the needs of your business" The Open Group URL: http://www.opengrouporg/architecture/togaf8-doc/arch/ (retrieved: 15.8.2011)
business object	"A business object is defined as a unit of information that has relevance from a business perspective" The Open Group URL: http://www.opengrouporg/archimate/doc/ts_archimate/ (retrieved: 15.8.2011)
CLM	short for ☞ Contract Lifecycle Management
connection feature[221]	Relations between constituting elements (e.g. document properties and metadata elements or storage location / directory path and metadata candidates)
context	context is "everything that is not text" as it depends on the point of view what is considered context and what is not; hence, context of a document are these entities of an enterprise architecture the documented is related with, e.g. business processes, activities, organisational units etc.
contract lifecycle management	Contract Lifecycle Management (CLM) is the same as ☞ Dokument Lifecycle Management but specific for *contract* documents
crosswalk	A table that maps the relationships and equivalencies between two or more metadata schemes. Crosswalks or metadata mapping support the ability of search engines to search effectively across heterogeneous databases (DCMI, 2010)
daily work documents	is a denotation to summarize documents based on templates like minutes
DC	short for ☞ Dublin Core
DLM	short for ☞ Document Lifecycle Management
DCMES	short for ☞ Dublin Core Metadata Element Set
DCMI	short for ☞ Dublin Core Metadata Initiative
descriptive metadata	metadata describing the content of a document, like the DC elements subject, description, relation (Greenberg et al. 2005)
DLM	short for ☞ Document Lifecycle Management
document	'document' is used for all kinds of unstructured information, that is text, audio, video or images.
document lifecycle management	Document Lifecycle Management (DLM) defines the phases a document passes from creation, modification/update to deletion or archiving

[221] Term was introduced by (M. Margaritopoulos et al. 2008) for the "interrelated properties of the resources connected with a relation that specify this connection on the basis of similarities or differences."

Term	Definition
domain ontology	a domain ontology defines the concepts from a given domain (Mascardi et al., 2007)
Dublin Core	The Dublin Core is a metadata element set. It includes all DCMI terms (that is, refinements, encoding schemes, and controlled vocabulary terms) intended to facilitate discovery of resources (DCMI, 2010)
Dublin Core Metadata Element Set	the Dublin Core Metadata Element Set defines 15 core elements that can be used to describe information resources introduced by the Dublin Core Initiative
Dublin Core Metadata Initiative	The Dublin Core Metadata Initiative is the body responsible for the ongoing maintenance of Dublin Core (DCMI, 2010)
EA	short for ☞ enterprise architecture
EAF	short for ☞ enterprise architecture framework
EMO	short for ☞ Enterprise Model Ontology
enterprise	Any collection of organisations that has a common set of goals and/or a single bottom line (Lankhorst, 2009) Here: "enterprise" is used for all types of organisations including non-profit organisations and public administrations
enterprise architecture	Enterprise Architecture (EA) is a coherent whole of principles, methods, and models that are used in the design and realisation of enterprise's organisational structure, business processes, information systems, and infrastructure (Lankhorst, 2009)
enterprise architecture description	"An architecture description is what is written down as a concrete work product. An architecture description (AD) expresses the architecture of a system of interest. An AD could be a document, a repository or a collection of artifacts used to define and document an architecture" (DSCI 2012)
enterprise architecture framework	An Enterprise Architecture Framework (EAF) is a tool which can be used for developing a broad range of different architectures. It should describe a method for designing an information system in terms of a set of building blocks, and for showing how the building blocks fit together. It should contain a set of tools and provide a common vocabulary. It should also include a list of recommended standards and compliant products that can be used to implement the building blocks. (The Open Group URL: http://www.opengroup.org/architecture/togaf8-doc/arch/ (retrieved: 4.3.2011)
enterprise component	an enterprise component is a non-ontological representation of an enterprise object, e.g. a record in a relational database, a paper document or a video file
enterprise object	an enterprise object is considered any entity that is part of an enterprise, like a business process activity, a compliance requirement, an organisational unit, personnel, IT infrastructure, motivation objects or manufactured products, regardless its representation
enterprise model ontology	Enterprise Model Ontology (EMO) adds meta information (e.g. based on Zachmann's framework) to a semantically enriched Enterprise Architecture to improve quality
enterprise ontology	An ontology is an explicit specification of a shared conceptualization (Gruber et al. 1993, Studer et al. 1998) of entities and their structures of the domain "enterprise"

Term	Definition
enterprise repository	An enterprise repository is considered the entirety of explicitly represented information available in an enterprise
format	here: the Dublin Core element used to designate the physical or digital manifestation of the resource (DCMI, 2010)
fully-automatic metadata generation	Complete (or total) reliance on automatic processes to create metadata (Greenberg et al., 2005)
MeGaSystem	MeGaSystem is the conception model of automatic, format-independent metadata generation based contex information in an enterprise
MeGaWorkbench	MeGaWorkbench is the name of the executable prototype developed within my research
metadata	Metadata is "data about data". Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource (NISO, 2004)
metadata candidates	Metadata candidates are instances of classes and properties in seEAD inferred from the context of a document
metadata creation	The act of producing metadata, that can be done by humans or machines (Greenberg & Severiens, 2007); used synonymously to *metadata generation*
metadata extraction	The process of automatically pulling (extracting) metadata from a resource's content. Resource content is mined to produce structured ('labelled') metadata for object representation (Greenberg et al., 2005)
metadata generation	Metadata generation is the act of creating or producing metadata (Greenberg, 2004); used synonymously to *metadata creation*
metadata harvesting	The process of automatically collecting resource metadata already embedded in or associated with a resource. The harvested metadata is originally produced by humans or by fully or semi- automatic processes supported by software (Greenberg et al., 2005) Here the *harvesting* is restricted to elements associated with a resource like document properties.
metadata record	a particular instance that a set of metadata elements is applied to for describing an object (Zeng & Qin, 2008)
metadata repository	A metadata repository is a collection of many metadata records (Hatala & Forth, 2003)
metadata schema	a metadata schema that defines the structure and semantics of metadata elements
metadata seeds	Metadata seeds are instances of classes and properties in seEAD created on the basis of harvested document properties (AttributeHarvest artifact) or content annotations (ContentAnnotations artifact) or on a mix of both.
metadata standards	standardized structure of metadata for the interoperable description of resources, e.g. Dublin Core
OCR	short for ☞ Optical Character Recognition

Term	Definition
ontology	An ontology is an explicit specification of a shared conceptualization (Gruber et al. 1993, Studer et al. 1998); in essence means, providing "a view on how the world or a specific domain is structured as agreed upon by the members of a community" (Buitelaar & Cimiano 2008, p 45). It is a formal, machine-understandable respresentation of a conceptual model, "in which concepts, properties, relationships, funcitons, constriants, and axioms are all explicitly defined" (Baca et al. 2008, p. 82)
optical character recognition	"Optical character recognition, usually abbreviated to OCR, is the mechanical or electronic conversion of scanned images of handwritten, typewritten or printed text into machine-encoded text" Wikipedia: Optical Character Regonition. URL: http://en.wikipedia.org/wiki/Optical_character_recognition (retrieved: 5.12.2012)
OWL	short for ☞ Web Ontology Language
RDF	short for ☞ Resource Description Framework
RDFS	short for ☞ Resource Description Framework Schema
RDFS-Plus	also called RDFS 3.0 is suggested as third version of RDFS, which extends RDFS by a small number of OWL language constructs
RDFS 3.0	also called -> RDFS-Plus
resource description framework	Resource Description Framework (RDF) is a standard model for data interchange on the Web. The RDF data model can be compared to other conceptual modeling approaches like entity-relationship or class diagrams. With RDF statements about resources can be expressed in the form of subject-predicate-object expressions. These expressions are known as *triples* in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object (Decker et al., 2000)
resource description framework schema	"RDFS is the schema language for RDF. […] It provides information about the ways in which we describe our data" (Allemang & Hendler 2008, p 91)
seEAD	short for ☞ semantically enriched Enterprise Architecture Description
semantic annotation	A semantic annotation is additional information that identifies or defines a concept in a semantic model in order to describe part of that document (WSDL Working Group, 2007)
semi-automatic metadata generation	Partial reliance on software to create metadata; a combination of fully-automatic and human processes to create metadata (Greenberg et al., 2005)
semantically enriched enterprise architecture description	semantically enriched Enterprise Architecture Description (seEAD) is an architecture description, which is formally represented (e.g. in an ☞ ontology) and enhanced by information about its meaning

Term	Definition
Simple Knowledge Organization System	The Simple Knowledge Organization System is "a set of specifications for organizing, documenting, and publishing taxonomies, classification schemes, and controlled vocabularies, such as thesauri, subject lists, and glossaries or terminology lists, within an RDF framework (Woodley 2008, p 49)
SKOS	short for ☞ Simple Knowledge Organization System
system properties	metadata that a system automatically generates (e.g., date_created, date_modified, or size), as well as metadata stored in a user profile (e.g., institutional_name or rights) and automatically assigned to documents (Greenberg et al., 2005)
technical metadata	technical metadata is generated automatically by document creation software, for example when nontextual digital resources are created, like file size, or creation date (Greenberg et al., 2005). Technical metadata build a subset of system properties
TOL	short for ☞ Top Level Ontology
top-level ontology	a Top-level Ontology (TOL) comprises generic concepts of the world like time, location and event
type	here: the Dublin Core element used to designate the nature or genre of the content of the resource. Type includes terms describing general categories, functions, genres, or aggregation levels for content (DCMI, 2010)
upper-level ontology	an upper-level ontology defines the very general concepts that are are highly reusable across several domains and applications (Mascardi et al., 2007)
web ontology language	"The OWL Web Ontology Language is designed for use by applications that need to process the content of information instead of just presenting information to humans" (W3C OWL Working Group, 2004)

11 Bibliography

Abramowicz, W., Filipowska, A., Kaczmarek, M., & Kaczmarek, T. (2007). Semantically enhanced Business Process Modelling Notation. *Proceedings of the Workshop on Semantic Business Process and Product Lifecycle Management (SBPM 2007) in conjunction with the 3rd European Semantic Web Conference (ESWC 2007)* (pp. 88-91). Innsbruck, Austria.

Agnoloni, T., Francesconi, E., & Spinosa, P. (2007). xmLegesEditor : an OpenSource Visual XML Editor for supporting Legal National Standards. *Proceedings of the V Legislative XML Workshop*.

Akabuilo, E. (2012). *Towards a Semantic Enterprise Repository with Access to Business Data*. *Applied Sciences*. University of Applied Sciences Northwestern Switzerland.

Akkermans, J. M., & Gordijn, J. (2003). Value-based requirements engineering: exploring innovative e-commerce ideas. *Requirements Engineering*, *8*(2), 114-134. doi:10.1007/s00766-003-0169-x

Al-Kilidar, H., Cox, K., & Kitchenham, B. (2005). The Use and Usefulness of the ISO/IEC 9126 Quality Standard. *Framework*, 126-132.

Albassuny, B. M. (2008). Automatic metadata generation applications: a survey study. *International Journal of Metadata, Semantics and Ontologies*, *3*(4), 260-282.

Allemang, D., & Hendler, J. (2008). *Semantic Web for the Working Ontologist - Effictive Modeling in RDFS and OWL*. United States: Morgan Kaufmann.

Allemang, D., Hodgson, R., & Polikoff, I. (2005). FEA Reference Model Ontologies (FEA RMO). *Development*.

Alter, S. (2002). *Information Systems: Foundation of E-Business* (4th ed.). Upper Saddle River, NJ.

Anderson, J. D., & Pérez-Carballo, J. (2001). The nature of indexing : how humans and machines analyze messages and texts for retrieval. Part I : Research , and the nature of human indexing. *Information Processing and Management*, *37*, 231-254.

Angles, R., & Gutierrez, C. (2008). The Expressive Power of SPARQL. In A. E. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin, & K. Thirunarayan (Eds.), *The Semantic Web - ISWC 2008* (Lecture No., pp. 114 - 150). Springer Berlin / Heidelberg.

Appelt, D. E. (1999). Introduction to information extraction. *Ai Communications*. Menlo Park, CA: Artificial Intelligence Center, SRI International.

Appelt, D. E., & Israel, D. J. (1999). Introduction to Information Extraction Technology. Menlo Park, CA: DFKI. Retrieved from http://www.dfki.de/~neumann/esslli04/reader/overview/IJCAI99.pdf

Arenas, M., Bertails, A., Prud'hommeaux, E., & Sequeda, J. F. (eds. . (2011). A Direct Mapping of Relational Data to RDF. *W3C Working Draft 20 September 2011*. Retrieved from

Arenas, M., Gutierrez, C., & Pérez, J. (2010). On the Semantics of SPARQL. *Semantic Web Information Management* (pp. 281-310). Springer Berlin / Heidelberg.

Ares Casal, J. M., Dieste Tubío, O., García Vázquez, R., López Fernández, M., & Rodríguez Yáñez, S. (1998). Formalising the Software Evaluation Process. *Proceedings of the International Conference of the Chilean Society of Computer Science, SCCC '98*. Antofogasta , Chile.

Athanasiadis, T., Tzouvaras, V., Petridis, K., Precioso, F., Avrithis, Y., & Kompatsiaris, Y. (2005). Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content. *Processings of the 5th International Workshop on Knowledge Markup and Semantic Annotation* (pp. 59-68). Galway, Ireland.

Auer, S., Feigenbaum, L., Miranker, D. P., Fogarolli, A., & Sequeda, J. (2010). Use Cases and Requirements for Mapping Relational Databases to RDF. *W3C Working Draft 2 June 2010*. Retrieved January 26, 2012, from http://www.w3.org/2001/sw/rdb2rdf/use-cases/

Baca, M., Gilliland, A. J., Gill, T., Woodley, M. S., & Wahlen, M. (2008). *Introduction to metadata*. (M. Baca, P. E. Pardo, S. U. Berg, & E. Zozom, Eds.) (2nd ed.). Los Angeles: Getty Research Institute.

Baldauf, M., Dustdar, S., & Rosenberg, F. (2007). A survey on context-aware systems. *International Journal of ad hoc and ubiquitous computing*, *2*(4), 263-277.

Balke, W.-T. (2012). Introduction to Information Extraction: Basic Notions and Current Trends. *Datenbank-Spektrum*, *12*(2), 81-88. doi:10.1007/s13222-012-0090-x

Bao, J. (2008). OWL Full Semantics - RDFCompatible Model-Theoretic Semantics. Retrieved from http://tw.rpi.edu/wiki/Image:2008-09-06_OWL_FULL_Semantics.ppt

Barnes, S.-A., Bradley, C., Brown, A., Cook, J., Kaschig, A., Kunzmann, C., Magenheim, J., et al. (2010). Results of the representative study and refined conceptual knowledge maturing model. Retrieved from http://mature-ip.eu

Barrasa, J., Corcho, O., & Gómez-pérez, A. (2003). Fund Finder : A case study of database-to-ontology mapping Case study. *Semantic Integration Workshop, ISWC 2003*. Florida, USA.

Barrasa, J., Corcho, Ó., & Gómez-Pérez, A. (2004). R2O , an Extensible and Semantically Based Database- to-ontology Mapping Language. *Proceedings of the Second Workshop on Semantic Web and Databases (SWDB 2004)*. Toronto, Canada.

Baskerville, R. L. (1999). Investigating information systems with action research. *Communications of the AIS*, *2*(October). Retrieved from http://portal.acm.org/citation.cfm?id=374476

Baskerville, R. L., & Myers, M. D. (2004). SPECIAL ISSUE ON ACTION RESEARCH IN INFORMATION SYSTEMS: MAKING IS RESEARCH RELEVANT TO PRACTICE — FOREWORD. *MIS Quarterly*, *28*(3), 329-335.

Baskerville, R., Pries-Heje, J., & Venable, J. (2009). Soft design science methodology. *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09* (p. 1). Philadelphia, Pennsylvania: ACM Press. doi:10.1145/1555619.1555631

Bazire, M., & Brézillon, P. (2005). Understanding Context Before Using It. *LNCS Modeling and Using Context*, *3554*, 29 - 40.

Bechhofer, S., Goble, C., & Horrocks, I. (2002). Requirements of Ontology Languages. *Knowledge Management*. Retrieved from http://www.sti-innsbruck.at/fileadmin/documents/deliverables/Ontoweb/D4.1.pdf

Beckett, D., & Grant, J. (2003). SWAD-Europe Deliverable 10.2: Mapping Semantic Web Data with RDBMSes. *Technical Report*. Retrieved January 26, 2012, from http://www.w3.org/2001/sw/Europe/reports/scalable_rdbms_mapping_report/

Beer, W., Christian, V., Ferscha, A., & Mehrmann, L. (2003). Modeling Context-Aware Behavior by Interpreted ECA Rules. *Lecture Notes in Computer Science*, (2790), 1064-1073.

Beisler, A., & Willis, G. (2009). Beyond Theory: Preparing Dublin Core Metadata for OAI-PMH Harvesting. *Journal of Library Metadata*, *9*(1), 65-97. doi:10.1080/19386380903095099

Berners-Lee, T. (1998). Relational Databases on the Semantic Web. *Relational Databases and the Semantic Web (in Design Issues)*. Retrieved January 26, 2012, from http://www.w3.org/DesignIssues/RDB-RDF.html

Bertolazzi, P., Krusich, C., Missikoff, M., & Manzoni, V. (2001). An Approach to the Definition of a Core Enterprise Ontology : CEO. *International Workshop on Open Enterprise Solutions: Systems, Experiences, and Organizations - OES-SEO 2001* (pp. 104-115). Rome.

Bizer, C., & Cyganiak, R. (2007). D2R - Lessons Learned. Retrieved from http://www.w3.org/2007/03/RdfRDB/papers/d2rq-positionpaper/

Bizer, C., & Seaborne, A. (2004). D2RQ – Treating Non-RDF Databases as Virtual RDF Graphs. *World Wide Web Internet And Web Information Systems*.

Bizer, C., Cyganiak, R., Garbers, J., Maresch, O., & Becker, C. (2009). The D2RQ Platform v0.7 - Treating Non-RDF Relational Databases as Virtual RDF Graphs. Berlin. Retrieved from http://www4.wiwiss.fu-berlin.de/bizer/d2rq/

Black, M., Bordwell, S., Brown, A., Caplan, P., Clifton, G., Duerr, R., Guenther, R., et al. (2007). *Final Report of PREIMIS Working Group. Implementing the PREMIS data dictionary : a survey of approaches*. *Context*.

Bock, G. E. (2005). The Case for Semantic Storage. Adding Insight and Expertise to Information Lifecycle Management. Retrieved from http://www.emc.com/services/consulting/application/expertise/information-lifecycle-management.htm

Born, M., Filipowska, A., Kaczmarek, M., Markovic, I., & Starzecka, M. (2008). Business Functions Ontology and its Application in Semantic Business Process Modelling. *ACIS 2008 Proceedings. Paper 110*.

Botella, P., Burgués, X., Carvallo, J. . P., Franch, X., Grau, G., Marco, J., & Quer, C. (2004). ISO/IEC 9126 in practice : what do we need to know? Universitat Politecnica de Catalunya. Retrieved from http://www.essi.upc.edu/~webgessi/publicacions/SMEF%2704-ISO-QualityModels.pdf

Bouquet, P., Ghidini, C., Giunchiglia, F., & Blanzieri, E. (2003). Theories and uses of context in knowledge representation and reasoning. *Journal of Pragmatics*, *35*, 455-484.

Brachmann, R., & Levesque, H. (2004). *Knowledge Representation and Reasoning*. *New York*. New York: Morgan Kaufmann.

Brander, S., Hinkelmann, K., Hu, B., Martin, A., Riss, U. V., Thönssen, B., & Witschel, H. F. (2011). Refining process models through the analysis of informal work practice. *9th International Conference on Business Process Management*. Clermont-Ferrand, France.

Breaux, T. D., & Antón, A. I. (2007). A Systematic Method for Acquiring Regulatory Requirements: A Frame-Based Approach. *6th Workshop on Requirements for High Assurance Systems (RHAS)* (Vol. 1936).

Breaux, T. D., & Antón, A. I. (2008). Automating the Extraction of Rights and Obligations for Regulatory Compliance. *Conceptual Modeling - ER 2008* (Lecture No., Vol. 523, pp. 154-168). Pittsburgh, PA, USA: Springer. Retrieved from http://www.springerlink.com/content/dn7240937t72lx46/

Breaux, T. D., Vail, M. W., & Antón, A. I. (2006). Towards Regulatory Compliance: Extracting Rights and Obligations to Align Requirements with Regulations. *14th IEEE International Requirements Engineering Conference (RE'06)* (pp. 1-10). Minneapolis/St. Paul, Minnesota, USA.

Brown, P. J. (1998). Triggering information by context. London: Springer.

Brun, R. (2010). *Linked Enterprise Models and Objects representing Content and Context of an Enterprise*. University of Applied Sciences Northwestern Switzerland (FHNW). Retrieved from http://www.martenschoenherr.de/pdf/Enterprise_Architecture Frameworks.pdf

Brunner, J.-S., Ma, L., Wang, C., Zhang, L., Wolfson, D. C., Pan, Y., & Srinivas, K. (2007). Explorations in the use of semantic web technologies for product information management. *Proceedings of the 16th international conference on World Wide Web - WWW '07* (p. 747). New York, New York, USA: ACM Press. doi:10.1145/1242572.1242673

Brügmann, H. (2011). *Management of Unstructured Information Using Semantic Metadata*. Friedrich-Alexander-Universität Erlangen-Nürnberg.

Buitelaar, P., & Cimiano, P. (2008). *Ontology learning and population: bridging the gap between text and knowledge*. (P. Buitelaar & P. Cimiano, Eds.)*Knowledge and Information Systems* (Knowledge ., pp. 45-69). IOS Press.

Bézivin, J. (2004). In Search of a Basic Principle for Model Driven Engineering. *The European Journal for the Informatics Professional*, *V*(2), 1-5.

Būmans, G., & Cerans, K. (2010). RDB2OWL : a Practical Approach for Transforming RDB Data into RDF / OWL. *Proceedings of the 6th International Conference on Semantic Systems* (pp. 1-3).

CCSDS. (2012). Reference Model for an Open Archival Information System (OAIS). *Practice*. Washington, DC, USA: CCSDS Secretariat.

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., Rosati, R., et al. (2011). The MASTRO system for ontology-based data access. *Semantic Web*, *2*, 43-53. doi:10.3233/SW-2011-0029

Campbell, D. G. (2002). The Use of the Dublin Core in Web Annotation Programs. Retrieved from http://www.bncf.net/dc2002/program/ft/paper12.pdf

Cardinaels, K., Meire, M., & Duval, E. (2005). Automating metadata generation: the simple indexing interface. Chiba, Japan: ACM New York, NY, USA. Retrieved from http://portal.acm.org/citation.cfm?id=1060825

Cardoso, Y. C. (2010). *Creation and Extension of Ontologies for Describing Communications in the Context of Organizations*. Universidade Nova de Lisboa.

Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., & Wilkinson, K. (2003). *Jena : Implementing the Semantic Web Recommendations Jena : Implementing the Semantic Web Recommendations*. *Digital Media*.

Carvalho, R. F. (2008). Metadata goes where Metadata is : contextual networks in the photographic domain. *Ph.D. Symposium of the 5th European Semantic Web Conference (ESWC 2008)*. Tenerife, Islas Canarias Spain.

Celino, I., Valle, E. D., Cerizza, D., & Turati, A. (2006). Squiggle : a Semantic Search Engine for indexing and retrieval of multimedia content. *Proceedings of First International Workshop on Semantic-enhanced Multimedia Presentation Systems (SEMPS 2006)*. Athens, Greece.

Charvat, J. (2003). *Project Management Methodologies: Selecting, Implementing, and Supporting Methodologies and Processes for Projects*. John Wiley & Sons.

Chen, D., Doumeingts, G., & Vernadat, F. (2008). Architectures for enterprise integration and interoperability : Past , present and future. *Computers in Industry*, *59*, 647-659. doi:10.1016/j.compind.2007.12.016

Chen, H., Finini, T., & Jashi, A. (2004). An ontology for context-aware pervasive computing environments. *The Knowledge Engineering Review*, *18*(3), 197-207.

Chen, Z., & Pooley, R. (2009a). Domain Modeling for Enterprise Information Systems - Formalizing and Extending Zachman Framework Using BWW Ontology. *2009 WRI World Congress on Computer Science and Information Engineering* (pp. 634-643). Los Angeles, CA: Ieee. doi:10.1109/CSIE.2009.934

Chen, Z., & Pooley, R. (2009b). Rediscovering Zachman Framework Using Ontology from a Requirement Engineering Perspective. *33rd Annual IEEE International Computer Software and Applications Conference* (pp. 3-8). Seattle,Washington: Ieee. doi:10.1109/COMPSAC.2009.107

Cheng, Xian Yi, Yang, A. Q., & Cheng, X. Y. (2011). The Study of Ontology Reasoning to Semantic Web. *Advanced Materials Research*, *204-210*, 375-380. doi:10.4028/www.scientific.net/AMR.204-210.375

Church, V. E., Card, D. N., Agresti, W. W., & Jordan, Q. L. (1986). An Approach for Assessing Software Prototypes. *ACM SIGSOFT Software Engineering Notes*, *11*(3), 65-76. Retrieved from http://dl.acm.org/citation.cfm?id=12927

Ciravegna, F., Chapman, S., Dingli, A., & Wilks, Y. (2004). Learning to Harvest Information for the Semantic Web. *The Semantic Web: Research and Applications* (Lecture No.). Berlin / Heidelberg: Springer.

Corcho, O., & Gomez-Perez, A. (2000). A Roadmap to Ontology Specification Language. *Knowledge Engineering and Knowledge Management Methods, Models, and Tools* (Lecture No., pp. 80-06). Springer Berlin / Heidelberg.

Coyle, K., & Baker, T. (2009). *Guidelines for Dublin Core Application Profiles*. *Framework* (pp. 1-14). Retrieved from http://dublincore.org/documents/profile-guidelines/index.shtml

Cronholm, S., & Goldkuhl, G. (2004). Conceptualising Participatory Action Research – Three Different Practices. *Electronic Journal of Business Research Methods*, *2*(2), 47-58.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., et al. (2011). Text Processing with GATE (Version 6). *University of Sheffield Department of Computer Science*. Retrieved August 27, 2011, from http://gate.ac.uk/sale/tao/split.html#splitpa3.html

Currier, S. (2008). Metadata for Learning Resources: An Update on Standards Activity for 2008. *Ariadne*, (55), 1-7. Retrieved from http://www.ariadne.ac.uk/issue55/currier/

DCMI. (2010). Dublin Core Metadata Initiative. Retrieved July 19, 2010, from http://dublincore.org/about-us/

DCMI Usage Board. (2012). DCMI Metadata Terms. *Recommendation*. Retrieved December 1, 2012, from http://dublincore.org/documents/dcmi-type-vocabulary/

DSCI. (2012). ISO/IEC/IEEE 42010 Website. Retrieved October 12, 2012, from http://www.iso-architecture.org/42010/

Daconta, M. C. (2007). *Information As Product. How to Deliver the Right Information, To the Right Person, At the Right Time* (2nd ed.). USA: Outskirts Press Inc.

Das, S., & Srinivasan, J. (2009). Database Technologies for RDF. *Reasoning Web*. Berlin / Heidelberg: Springer.

Das, S., Sundara, S., & Cyganiak, R. (eds. . (2011). R2RM: RDB to RDF Mapping Language. *W3C Working Draft 20 September 2011*. Retrieved January 26, 2012, from http://www.w3.org/TR/r2rml/

Davis, M., King, S., Good, N., & Sarvas, R. (2004). From Context to Content : Leveraging Context to Infer Media Metadata. *Proceedings of the 12th annual ACM international conference on Multimedia* (pp. 188-195).

Dawes, S. S. (2008). Governance in the information age : a research framework for an uncertain future. *Proceedings of the 9th Annual International Digital Government Research Conference* (pp. 290-297). Montreal, Canada. Retrieved from http://portal.acm.org/citation.cfm?id=136783.1367881

De Bruijn, J. (2003). *Using Ontologies - Enabling Knowledge Sharing and Reuse on the Semantic Web*. *October*. Galway, Ireland. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.106.7278&rep=rep1&type=pdf

De Leenheer, P., & Mens, T. (2008). Ontology Evolution. In M. Hepp, P. De Leenheer, A. De Moor, & Y. Sure (Eds.), *Ontology Management - Semantic Web, Semantic Web Services, and Business Applications* (pp. 131-176). SpringerScience + Bsuiness Media Inc.

Decker, S., Dumontier, M., Finin, T., & Horrocks, I. (2008). *An OWL 2 Far*. Retrieved from http://videolectures.net/iswc08_panel_schneider_owl/

Decker, S., Erdmann, M., Fensel, D., & Studer, R. (n.d.). How to use ontobroker? Retrieved February 11, 2012, from http://ksi.cpsc.ucalgary.ca/KAW/KAW98/decker/

Decker, S., Harmelen, F. V., Broekstra, J., Erdmann, M., Fensel, D., Horrocks, I., Klein, M., et al. (2000). The Semantic Web - on the respective Roles of XML and RDF. *Internet Computing, IEEE*, *4*(5), 63 - 73.

Dellschaft, K., & Staab, S. (2008). Strategies for the Evaluation of Ontology Learning. In P. Buitelaar & P. Cimiano (Eds.), *Ontology Learning And Population: Bridging the Gap between Text and Knowledge*. Amsterdam: IOS Press.

Departement of Defense. (2007). DoD Architecture Framework Volume I : Definitions and Guidelines. *Architecture*. U.S. Department of Defense.

Departement of Treasury. (n.d.). Treasury Enterprise Architecture Framework (TEAF). Retrieved December 1, 2012, from

http://www.cioindex.com/enterprise_architecture/articleid/2007/treasury-enterprise-architecture-framework-teaf.aspx

Dey, A. K., & Abowd, G. D. (1999). Towards a Better Understanding of Context and Context-Awareness. *Proceedings of the 1st international sympo-sium on Handheld and Ubiquitous Computing (HUC '99)* (pp. 304-307). The Hague, The Netherlands. Retrieved from http://nl.ijs.si/~damjan/NeOn/99-22.pdf

Dietz, J. L. G. (2006). *Enterprise Ontology. Theory and Methodology*. Berlin Heidelberg: Springer-Verlag.

Dietz, J. L. G., & Hoogervorst, J. A. P. (2008). Enterprise ontology in enterprise engineering. *SAC '08 Proceedings of the 2008 ACM symposium on Applied computing*.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., et al. (2003). A case for automated large-scale semantic annotation. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, *1*, 115-132. doi:10.1016/j.websem.2003.07.006

Do Prado, H. A., & Ferneda, E. (2008). *Emerging technologies of text mining: techniques and applications* (p. 358). IGI Global snippet.

Doctorow, C. (2002). Metacrap: Putting the torch to seven straw men of the meta-utopia. *E-Learning Guru Newsletter*. Retrieved from http://www.e-learningguru.com/articles/metacrap.htm.

Donalies, E. (2005). *Die Wortbildung des Deutschen: ein Überblick*. (U. Hass, W. Kallmeyer, & U. Wassner, Eds.) (Studien zu., p. 192). Tübingen: Gunter Narr Verlag.

Drucker, P. F. (1994). *Post-capitalist society*. New York: HarperBusiness.

Dublin Core Metadata Initiative. (2002). Expressing Qualified Dublin Core in RDF / XML. *DCMI Proposed Recommendation*. Retrieved December 2, 2012, from http://dublincore.org/documents/dcq-rdf-xml/

Dublin Core Metadata Initiative. (2007). DCMI Tools Glossary. *draft version*. Retrieved November 28, 2012, from http://dublincore.org/groups/tools/glossary.shtml

Dublin Core Metadata Initiative. (2012). Dublin Core Metadata Element Set, Version 1.1. Retrieved November 29, 2012, from http://dublincore.org/documents/dces

Eiter, T., Ianni, G., Krennwallner, T., & Polleres, A. (2008). Rules and Ontologies for the Semantic Web. *Reasoning Web* (pp. 1-54). Springer Berlin / Heidelberg.

Eiter, T., Ianni, G., Polleres, A., Schindlauer, R., & Tompits, H. (2006). Reasoning with Rules and Ontologies. *Reasoning Web* (pp. 93-127). Springer.

Emmenegger, S., Laurenzi, E., & Thönssen, B. (2012). IMPROVING SUPPLY-CHAIN-MANAGEMENT BASED ON SEMANTICALLY ENRICHED RISK

DESCRIPTIONS. *Proceedings of 4th Conference on Knowledge Management and Information Sharing (KMIS2012)*. Barcelona, Spain.

Enser, P., & Sandom, C. (2003). Towards a Comprehensive Survey of the Semantic Gap in Visual Image Retrieval. *International Conference on Image and Video Retrieval. Lecture Notes in Computer Science* (Vol. 2728, pp. 291-299).

Enterprise Engineering Institute. (n.d.). Methodology - The benefits of DEMO. Retrieved May 25, 2011, from http://www.demo.nl/methodology

Eriksson, H. (2007). The semantic-document approach to combining documents and ontologies. *International Journal of Human-Computer Studies*, *65*(7), 624-639. Retrieved from http://www.sciencedirect.com/science/article/B6WGR-4NFXDN3-1/2/b2dfe578e6aa84b801e80a84b7888a35

Erling, O., & Mikhailov, I. (2009). RDF Support in the Virtuoso DBMS. *Networked Knowledge - Networked Media Studies in Computational Intelligence* (pp. 7-24).

Ettema, R., & Dietz, J. L. G. (2009). ArchiMate and DEMO – Mates to Date? *Advances in Enterprise Engineering III : Proceedings of the 5th International Workshop, CIAO! 2009, and 5th International Workshop, EOMAS 2009* (pp. 172-186). Amsterdam, The Netherlands: Springer LNCS.

FEAF. (1999). Federal Enterprise Architecture Framework. *Architecture*. Federal CIO Council.

Farrell, J. (ed. ., & Lausen, H. (ed. . (2007). Semantic Annotations for WSDL and XML Schema. *editors' copy*. Retrieved November 28, 2012, from http://www.w3.org/2002/ws/sawsdl/spec/#Terminology

Feldkamp, Daniela, Hinkelmann, K., & Thönssen, B. (2010). Ontologies for E-Government. In M. Healy, A. Kameas, & R. Poli (Eds.), . Heidelberg: Springer.

Fensel, D. (2004). *Ontologies: a silver bullet for knowledge management and electronic commerce*. Springer Berlin / Heidelberg / New York.

Fensel, D., Decker, S., Erdmann, M., & Studer, R. (1998). Ontobroker : The Very High Idea. *Proceedings of the 11. International Flairs Conference (FLAIRS-98)*. Sanibal Island, USA.

Fernandez, M., Gomez-Perez, A., & Juristo, N. (1997). METHONTOLOGY: From Ontological Art towards Ontological Engineering. *Proceedings of the AAAI-97 Spring Symposium Series on Ontological Engineering* (p. pages 33—40.). Stanford, CA, USA.

Filipowska, A., Kaczmarek, M., Kowalkiewicz, M., Markovic, I., & Zhou, Y. (2009). Organizational Ontologies to Support Semantic Business Process Management. *IProceedings of the 4th International Workshop on Semantic Business Process Management - SBPM 09*.

Fink, K., & Grimm, D. (2007). *The Use of Business Process Management during the Implementation of Electronic Records Management Systems*. *Information Strategy*. Department of Information Systems, Univeristy of Innsbruck, Innsbruck, Austria. Retrieved from http://www.google.ch/url?sa=t&source=web&cd=5&ved=0CCkQFjAE&url=http%3A%2F%2Fibis.in.tum.de%2Fmkwi08%2F06_eGovernment%2F01_Fink.pdf&rct=j&q=Gever+switzerland+success&ei=-sgPTPPQPN-P4gayzK2qDA&usg=AFQjCNGmaubAdEahjhaXVBfxzgqCDcg3_Q&sig2=vFnRc-P_8Y2vazCLW3fY8Q

Fox, M. S., & Grüninger, M. (1997). On Ontologies And Enterprise Modelling. *International Conference on Enterprise Integration Modelling Technology 97*. Springer. Retrieved from http://www.eil.utoronto.ca/enterprise-modelling/papers/fox-eimt97.pdf

Fox, M. S., & Grüninger, M. (1998). Enterprise Modeling. *AI Magazine*, *19*(3), 109-121. doi:10.1147/sj.372.0170

Fox, M. S., Barbuceanu, M., & Grüninger, M. (1996). An organisation ontology for enterprise modeling: Preliminary concepts for linking structure and behaviour. *Computers in Industry*, *29*(1-2), 123-134. doi:10.1016/0166-3615(95)00079-8

Fox, M. S., Barbuceanu, M., Grüninger, M., & Lin, J. (1996). An Organization Ontology for Enterprise Modelling. *Simulating Organizations: Computational Models of Institutions and Groups*, (AAAI/MIT Press), 131-152.

Frigg, R., & Hartmann, S. (2012). Models in Science. *Science*. Zalta, Edward N. (ed.). Retrieved from http://plato.stanford.edu/archives/fall2012/entries/models-science/

Fuggetta, A. (2003). Open source software—an evaluation. *Journal of Systems and Software*, *66*, 77-90. doi:10.1016/S0164-1212(02)00065-1

Fürber, C., & Hepp, M. (2010). Using SPARQL and SPIN for Data Quality Management on the Semantic Web 1 Introduction. In W. Abramowicz, R. Tolksdorf, W. Aalst, J. Mylopoulos, & M. Rosemann (Eds.), *Business Information Systems* (Lecture No., pp. 1-12). Springer Berlin / Heidelberg.

GATE. (n.d.). GATE: a full-lifecycle open source solution for text processing. *Introduction to GATE*. Retrieved August 12, 2012, from http://gate.ac.uk/overview.html

Gailly, F., & Poels, G. (2007). Ontology-Driven Business Modelling : Improving the Conceptual Representation of the REA Ontology. *Lecture Notes in Computer Science*, *4801*, 407-422.

Geerts, G. L., & McCarthy, W. E. (2000). *The Ontological Foundation of REA Enterprise Information Systems*. *Business* (pp. 1-34).

Geerts, G. L., & McCarthy, W. E. (2002). An ontological analysis of the economic primitives of the extended-REA enterprise information architecture. *International Journal of Accounting Information Systems*, *3*, 1-16.

Gesellschaft für Evaluation. (2008). Standards für Evaluation. *Evaluation*. Mainz. Retrieved from http://www.alt.degeval.de/calimero/tools/proxy.php?id=19076

Ghawi, R., & Cullot, N. (2007). Database-to-Ontology Mapping Generation for Semantic Interoperability. *Proceedings of the 3rd international workshop on database interoperability (INTERDB 2007)*.

Gilliland, A. J. (2008). Setting the Stage. In M. Baca, P. E. Pardo, S. U. Berg, & E. Zozom (Eds.), *Introduction to metadata* (2nd ed., pp. 1-19). Los Angeles: The Getty Research Institute.

Giunchiglia, F. (1993). Contextual Reasoning. *Epistemologia - Special Issue on "I Linguaggi e le Macchine"*, *16*, 354-364. LAFORIA.

Godby, C. J., Young, J. A., & Childress, E. (2012). D-Lib Magazine December 2004. *D-Lib Magazine*, *10*(12), 1-11.

Gomez-Perez, A., Fernandez-Lopez, M., & Corcho, O. (2004). *Ontological Engineering* (4th ed.). Springer-Verlag London, UK.

Gonsalves, A. (2005). Podcast Users Expected To Reach 60 Million In Five Years. *Information Week - Global CIO*. Retrieved from http://www.informationweek.com/news/global-cio/showArticle.jhtml?articleID=165600711

Goudos, S. K., Peristeras, V., & Tarabanis, K. (2006). Mapping Citizen Profiles to Public Administration Using Ontology Implementations of the Governance Enterprise Architecture (GEA) models. In A. Abecker, G. Mentzas, & L. Stojanovic (Eds.), *Proceedings of the Workshop on Semantic Web for eGovernment 2006*. Budva, Serbia & Montenegro.

Grau, B. C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., & Sattler, U. (2008). OWL 2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web*, *6*(4), 309-322. doi:10.1016/j.websem.2008.05.001

Greenberg, J. (2004). Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. *Journal of Internet Cataloging*, *6*(4), 59 — 82. Retrieved from http://www.informaworld.com/smpp/title~content=t904385370

Greenberg, J., & Severiens, T. (2007). DCMI-Tools: Ontologies for digital application description. *Proceedings of ELPUB Conference* (Vol. 2). Vienna, Austria.

Greenberg, J., Spurgin, K., & Crystal, A. (2005). *Final Report for the AMeGA (Automatic Metadata Generation Applications)* (p. 97). North Carolina. Retrieved from http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf

Greenberg, J., Spurgin, K., & Crystal, A. (2006). Functionalities for automatic metadata generation applications: a survey of metadata experts ' opinions. *International Journal of Metadata, Semantics and Ontologies*, *1*(1).

Grigorov, A. (2007). Milestone M5.8 KRSM first cycle prototype evaluation plan. *System*. TEN Competence.

Gruber, T. R., Guarino, N., & Poli, R. (1993). Toward Principles for the design of Ontologies Used for Knowledge Sharing. Kluwer Academic Publishers. Retrieved from http://www2.umassd.edu/SWAgents/agentdocs/stanford/onto-design.pdf

Grüninger, M., & Fox, M. S. (1995). Methodology for the Design and Evaluation of Ontologies. *Industrial Engineering (1995)*, *95*, 1-10. Retrieved from http://ibict.phlnet.com.br/anexos/grninger95methodology.pdf

Guarino, N. (1998). Formal Ontology in Information Systems. In N. Guarino (Ed.), *Proceedings of The First International Conference (FOIS'98)* (pp. 3-15). Trento, Italy.

Guerra Currie, A.-M., Travis Fricke, C., Diedrick, S. W., Kepper Lagarde, J., & Mycotte, H. O. (2007). System and Method of Generating Automated Document Analysis Tools.

Guha, R. V. (1991). *Contexts: A Formalization and Some Applications*. Stanford University., Palo Alto, Ca. Retrieved from http://www.filosoficas.unam.mx/~morado/TextosAjenos/guha-thesis.pdf

Gómez-Pérez, A., Fernández, M., & de Vicente, A. J. (1996). Towards a Method to Conceptualize Domain Ontologies. *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI'96)*. Budapest, Rumanía. Retrieved from http://oa.upm.es/7228/1/Towards_a_Method_.pdf

Gómez-Pérez, A., Fernández, M., & de Vicente, A. J. (2004). Towards a Method to Conceptualize Domain Ontologies. *In Workshop on Ontological Engineering, ECAI'96*. Berlin / Heidelberg / New York: Springer.

Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2005). *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. London: Springer.

Handschuh, S., Staab, S., & Maedche, A. (2001). CREAM — Creating relational metadata with a component-based , ontology-driven annotation framework. *Proceedings of K-CAP'01 Conference*. Victoria, British Columbia, Canada.

Handschuh, S., Staab, S., & Volz, R. (2003). On Deep Annotation. *Proceedings of the 12th International World Wide Web Conferenc*. Budapest, Hungary.

Hanschke, I. (2009). *Strategic IT Management: A Toolkit for Enterprise Architecture Management*. Berlin Heidelberg: Springer-Verlag.

Hatala, M., & Forth, S. (2003). A COMPREHENSIVE SYSTEM FOR COMPUTER-AIDED METADATA GENERATION. *Proceedings of the 12. International Conference of of The World Wide Web Consortium (WWW2003)*. Budapest, Hungary.

Heflin, J., Hendler, J., & Luke, S. (1999). *SHOE : A Knowledge Representation Language for Internet Applications 1 Introduction*. *Computer* (pp. 1-30). Retrieved from http://drum.lib.umd.edu/handle/1903/1044

Hepp, M., & Roman, D. (2007). An Ontology Framework for Semantic Business Process Management. *Proceedings of Wirtschaftsinformatik 2007* (pp. 1-18). Karlsruhe.

Hert, M., Reif, G., & Gall, H. C. (2001). A Comparison of RDB-to-RDF Mapping Languages. *Proceedings of the 7th International Conference on Semantic Systems* (pp. 25-32). Graz, Austria. Retrieved from http://dl.acm.org/ft_gateway.cfm?id=2063522&ftid=1054499&dwn=1&CFID=81605256&CFTOKEN=14608054

Hillmann, D., Dushay, N., & Phipps, J. (2004). Improving Metadata Quality : Augmentation and Recombination. *Proceedings of the 2004 international conference on Dublin Core and metadata applications: metadata across languages and cultures (DCMI '04)* (pp. 1-8). Retrieved from http://www.ecommons.cornell.edu/bitstream/1813/7897/1/Paper_21.pdf

Hinkelmann, Knut, Merelli, E., & Thönssen, B. (2010). The Role of Content and Context in Enterprise Repositories. *Proceedings of the 2nd International Workshop on Advanced Enterprise Architecture and Repositories - AER 2010*.

Hofstee, E. (2009). *Constructing a Good Dissertation*. EPE.

Holz, H., Maus, H., Bernardi, A., & Rostanin, O. (2005). From Lightweight , Proactive Information Delivery to Business Process-Oriented Knowledge Management. *Knowledge Creation Diffusion Utilization*, *0*(2), 101-127.

Hong, J. I., & Landay, J. A. (2001). An Infrastructure Approach to Context-Aware Computing. *Human-Computer Interaction*, *16*(2), 287-303.

Hong, J.-yi, Suh, E.-ho, & Kim, S.-J. (2009). Context-aware systems: A literature review and classification. *Expert Systems with Applications*, *36*(4), 8509-8522. Elsevier Ltd. doi:10.1016/j.eswa.2008.10.071

Horrocks, I., Patel-Schneider, P. F., & van Harmelen, F. (2003). From SHIQ and RDF to OWL: the making of a Web Ontology Language. *Web Semantics: Science, Services and Agents on the World Wide Web*, *1*(1), 3-31. doi:10.1016/j.websem.2003.07.001

Houde, S., & Hill, C. (1997). What do Prototypes Prototype ? In H. M., L. T., & P. P. (Eds.), *Handbook of Human-Computer Interaction* (2nd ed., pp. 1-1582). Amsterdam: evier Science B. V.

IFIP IFAC Task Force. (1998). GERAM - Generalised Enterprise Reference Architecture and Methodology. Retrieved December 1, 2012, from http://www.ict.griffith.edu.au/~bernus/ims/gm21/tg4wp1/geram/1-6-2/v1.6.2.html

Johner, S. (2011). *Best Tool for Metadata Harvesting*. University of Applied Sciences Northwestern Switzerland.

Johnson, P., Lagerström, R., Närman, P., & Simonsson, M. (2007). Enterprise architecture analysis with extended influence diagrams. *Information Systems Frontiers*, *9*(2-3), 163-180. doi:10.1007/s10796-007-9030-y

Joint Committee on Standards for Educational Evaluation. (2000). *Handbuch der Evaluationsstandards*. (J. R. Sanders, Ed.). Opladen: Leske + Budrich.

Jonkers, H., Lankhorst, M., Buuren, R. V., Hoppenbrouwers, S., & Bonsangue, M. (2004). Concepts for Modelling Enterprise Architectures. *INTERNATIONAL JOURNAL OF COOPERATIVE INFORMATION SYSTEMS*, *13*, 257-288.

Jonsson, M. (2007). Sensing and Making Sense. Designing Middleware for Context Aware Computing. Stockholm, Sweden: The Royal Institute of Technology.

Kandefer, M., & Shapiro, S. C. (2008). Comparing SNePS with Topbraid /Pellet. *Group*. Retrieved from www.cse.buffalo.edu/~shapiro/Papers/tn42.pdf

Kang, D., Lee, J., Choi, S., & Kim, K. (2010). An ontology-based Enterprise Architecture. *Expert Systems With Applications*, *37*(2), 1456-1464. Elsevier Ltd. doi:10.1016/j.eswa.2009.06.073

Karagiannis, Dimitris. (1995). BPMS: business process management systems. *ACM SIGOIS Bulletin*, *16*(1). Retrieved from http://portal.acm.org/citation.cfm?id=209894#

Kasneci, G., Ramanath, M., Suchanek, F., & Weikum, G. (2009). The YAGO-NAGA Approach to Knowledge Discovery. *ACM SIGMOD Record*, *37*(4), 41-47.

Kebbell, A., & Campbell, D. (2002). Managing digital objects and their metadata: challenges and responses. *Discover*.

Khan, A. N., Asghar, S., & Fong, S. (2011). Evaluation of Semantic Web Services: Before and After Applying in Telecommunication. *Journal of Emerging Technologies in Web Intelligence*, *3*(2), 120-135. doi:10.4304/jetwi.3.2.120-135

Kiyavitskaya, N., Krausová, A., & Zannone, N. (2008). Why Eliciting and Managing Legal Requirements Is Hard. *Proceedings of the 2008 Requirements Engineering and Law* (pp. 26-30). Ieee. doi:10.1109/RELAW.2008.10

Kiyavitskaya, N., Zeni, N., Cordy, J. R., Mich, L., & Mylopoulos, J. (2009). Data & Knowledge Engineering Cerno : Light-weight tool support for semantic annotation of textual documents. *Data & Knowledge Engineering*, *68*(12), 1470-1492. Elsevier B.V. doi:10.1016/j.datak.2009.07.012

Klyne, G. (ed. ., & Carroll, J. (ed. . (2002). Resource Description Framework (RDF): Concepts and Abstract Data Model. *W3C Working Draft 29 August 2002*. Retrieved December 2, 2012, from http://www.w3.org/TR/2002/WD-rdf-concepts-20020829/

Knublauch, H. (2009). Composing the Semantic Web. *A tool developer's blog on ontology development for the Semantic Web and beyond*. Retrieved from http://composing-the-semantic-web.blogspot.ch/2009/01/object-oriented-semantic-web-with-spin.html

Kondylakis, Haridimos, Plexousakis, D., Wache, H., Wolff, D., Hinkelmann, K., & Bergmayr, A. (2010). Semantic Technology to enable Business and IT alignment. In Robert Woitsch & A. Micsik (Eds.), *OKM Open Knowledge Models, Workshop W3 at EKAW 2010* (pp. 25-32).

Konstantinou, N., Spanos, D.-E., Chalas, M., Solidakis, E., & Mitrou, N. (2006). VisAVis: An Approach to an Intermediate Layer between Ontologies and Relational Database Contents. *Proceedings of the CAiSE'06 3rd International Workshop on Web Information Systems Modeling (WISM'06)* (pp. 1050-1061). Luxembourg, Luxembourg.

Kontchakov, R., Lutz, C., Toman, D., Wolter, F., & Zakharyaschev, M. (2010). The Combined Approach to Query Answering in DL-Lite. *Proceedings of the Twelfth International Conference on the Principles of Knowledge Representation and Reasoning (KR 2010)* (pp. 247-257). Toronto, Canada.

Kordon, F., & Henkel, J. (2003). An Overview of Rapid System Prototyping Today. *Design Automation for Embedded Systems*, *8*(4), 275-282. doi:10.1023/B:DAEM.0000013062.16911.c5

Korpiää, P. (2005). Blackboard-based software framework and tool for mobile device context awareness. *VTT Publications*, (579).

Kröse, B., Kasteren, T. V., Gibson, C., & Dool, T. van. (2008). CARE: Context Awareness in Residences for Elderly. *Proceedings of ISG'08: The 6th International Conference of the International Society for Gerontechnology* (pp. 101-105). Pisa, Italy.

Krötzsch, M., ul Mehdi, A., & Rudolph, S. (2010). Orel: Database-Driven Reasoning for OWL 2 Profiles. *23rd Int. Workshop on Description Logics (DL2010), CEUR-WS 573* (pp. 114-124). Waterloo, Canada.

Kuchartz, U., Dresing, T., Rädiker, S., & Stefer, C. (2007). *Qualitative Evaluation* (pp. 1-119). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kuster, J., Huber, E., Lippmann, R., Schmid, A., Schneider, E., Witschi, U., & Wüst, R. (2005). *Handbuch Projektmanagement*. Springer.

Lagerström, R., Saat, J., Franke, U., Aier, S., & Ekstedt, M. (2009). Enterprise Meta Modeling Methods – Combining a Stakeholder-Oriented and a Causality-Based Approach. *Enterprise, Business-Process and Information Systems Modeling* (Lecture No., pp. 381-392).

Lagoze, C., & Hunter, J. (2001). The ABC Ontology and Model. *Journal of Digital Information. Special Issue on Metadata: Selected papers from the Dublin Core 2001 Conference* (pp. 18 - 36). Tokyo, Japan.

Lankhorst, M. (2004). *ArchiMate Language Primer*. *Work*. Retrieved from https://doc.telin.nl/dsweb/Get/Document-43839/

Lankhorst, M. (2009). *Enterprise Architecture at Work. Modeling, Communication and Analysis*. (M. Lankhorst, Ed.) (2nd ed.). Berlin / Heidelberg: Springer. Retrieved from

http://www.springerlink.com/content/pwh613/?p=164b8299a56a4610b79e7222b448791d&pi=0

Le Clair, C., & Poore, K. (2008). Enterprise Content Management's Next Step Forward. *Forrester Research*. Retrieved July 24, 2010, from http://www.forrester.com/rb/Research/enterprise_content_managements_next_step_forward/q/id/44243/t/2

Lee, C., Park, S., Lee, D., Lee, J.-won, Jeong, O.-ran, & Lee, S.-goo. (2008). A Comparison of Ontology Reasoning Systems Using Query Sequences. *Proceedings of the 2nd international conference on Ubiquitous information management and communication* (pp. 1-4). Kuala Lumpur, Malaysia.

Lee, T., Lee, I.-hoon, Lee, S., Lee, S.-goo, Kim, D., Chun, J., Lee, H., et al. (2006). Building an operational product ontology system. *Electronic Commerce Research and Applications*, *5*(1), 16-28. doi:10.1016/j.elerap.2005.08.005

Lenat, D. (1998). *The Dimensions of Context-Space*. Austin. Retrieved from http://www.cyc.com/doc/context-space.pdf

Lenat, D. B. (1995). CYC: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, *38*(11), 33-38. doi:10.1145/219717.219745

Leppänen, M. (2005). A Context-Based Enterprise Ontology. In G. Guizzardi & G. Wagner (Eds.), *Proceedings of the EDOC International Workshop on Vocabularies, Ontologies and Rules for the Enterprise (VORTE'05)* (pp. 17-24). Enschede, Netherlands: Springer Berlin.

Leppänen, M. (2006). Towards an Ontology for Information Systems Development. *The 18th Conference on Advanced Information Systems Engineering* (Vol. 35, pp. 273-286). Luxembourg, Luxembourg. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.5785&rep=rep1&type=pdf

Lerner, J., & Tirole, J. (2002). Some Simple Economics of Open Source. *The Journal Of Industrial Economics*, *50*(2), 197-234.

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-Based Multimedia Information Retrieval : State of the Art and Challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, *2*(1), 1-19.

Lichtenstein, S., & Parker, C. (2003). Processes and Impacts of Knowledge Creation in Email. *Australian Conference for Knowledge Management & Intelligent Decision Support. ACKMIDS*. Melbourne, Australia.

Liddy, E. D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ozgencil, N. E., Diekema, A., et al. (2002). Automatic Metadata Generation & Evaluation. *Proceedings of Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 401-402). Tampere, Finland. Retrieved from http://portal.acm.org/citation.cfm?id=564464

Liimatainen, K., Hoffman, M., & Jukka, H. (2007). *Overview of Enterprise Architecture work in 15 countries. Finnish Enterprise Architecture Research Project*. *Finance*. Helsinki.

Linnhoff-Popien, C., & Strang, T. (2004). A Context Modeling Survey. *Graphical Models*, (4), 1-8.

Lux, M. (2006). *Semantic Metadata*. TU Graz.

Maamar, Z., Benslimane, D., & Narendra, N. C. (2006). What Can Context Do for Web Services. *Communications of the ACM*, *49*(12).

Maedche, Alexander, & Staab, S. (2001). Ontology Learning for. *IEEE Intelligent Systems*, *16*(2), 72-79.

Maedche, A., Motik, B., & Stojanovic, L. (2003). Managing multiple and distributed ontologies on the Semantic Web. *The VLDB Journal — The International Journal on Very Large Data Bases*, *12*(4), 286-302.

Maedche, Alexander, Motik, B., Stojanovic, L., Studer, R., & Volz, R. (2003). Ontologies for Enterprise Knowledge Management. *IEEE Intelligent Systems*, *18*(2), 26-33.

Maier, R., & Schmidt, A. (2007). Characterizing Knowledge Maturing: A Conceptual Process Model for Integrating E-Learning and Knowledge Management. *4th Conference Professional Knowledge Management - Experiences and Visions WM '07* (pp. 325-334). Potsdam, Germany: GITO, Belrin.

Malik, S. K., Prakash, N., & Rizvi, S. (2010). Semantic Annotation Framework For Intelligent Information Retrieval Using KIM Architecture. *International Journal of Web & Semantic Technology (IJWest)*, *1*(4), 12-26.

Margaritopoulos, M., Margaritopoulos, T., Kotini, I., & Manitsaris, A. (2008). Automatic metadata generation by utilising pre-existing metadata of related resource. *International Journal of Metadata, Semantics and Ontologies*, *3*(4), 292-304. Retrieved from www.inderscience.com

Martin, A. (2010). *Linked Enterprise Models and Objects providing Context and Content for creating Metadata*. University of Applied Sciences Northwestern Switzerland.

Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., Narayanan, S., et al. (2004). OWL-S: Semantic Markup for Web Services. *W3C Member Submission 22 November 2004*. Retrieved December 1, 2012, from http://www.w3.org/Submission/OWL-S/

Mascardi, V., Cordì, V., & Rosso, P. (2007). *A Comparison of Upper Ontologies (Technical Report DISI-TR-06-21)*. *Span*. Genova, Italy. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.1689&rep=rep1&type=pdf

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., & Schneider, L. (2003). *WonderWeb Deliverable D17 Preliminary Report*. *Communities* (pp. 1-38).

Matthes, D. (2011). *Enterprise Architecture Frameworks Kompendium*. Heidelberg: Springer.

McCarthy, J. (1993). Notes on formalizing contexts. *Science* (pp. 1-13). Chambery, France: Morgan Kaufmann.

McCarthy, W. E. (1982). The REA Accounting Model: A Generalized Framework for Accounting Systems in a Shared Data Environment. *The Accounting Review*, *57*(3).

McCray, A. T., & Gallagher, M. E. (2001). Principles For Digital Library Development. *Communications of the ACM*, *44*(5), 49-54.

Meditskos, G., & Bassiliades, N. (2010). DLEJena: A practical forward-chaining OWL 2 RL reasoner combining Jena and Pellet. *Web Semantics: Science, Services and Agents on the World Wide Web*, *8*(1), 89-94. doi:10.1016/j.websem.2009.11.001

Mei, Q., & Zhai, C. X. (2006). A Mixture Model for Contextual Text Mining. *Contexts* (pp. 649-655). Philadelphia, Pennsylvania, USA: KDD'06.

Mena, T. B., Saoud, N. B.-B., Ahmed, M. B., Pavard, B., & Kokinov D.C.; Roth-Berghofer, Th.R.; Vieu, L., B. . R. (2007). Towards a Methodology for Context Sensitive Systems Development. Roskilde, Denmark: Springer.

Mendling, J., & Nüttgens, M. (2002). Event-driven-process-chain-markup-language (EPML): Anforderungen zur Definition eines XML-Schemas für Ereignisgesteuerte Prozessketten (EPK). *Proceedings of the 1st GI-workshop on business process management with event-driven process chains (EPK 2002)* (pp. 87-93). Trier, Germany.

Minonne, C., Colicchio, C., Litzke, M., & Keller, T. (2011). *Business Process Management 2011 – Status quo und Zukunft Eine empirische Studie im deutschsprachigen Europa*. *Management* (pp. 1-68). Zürich.

Mitschick, A. (2009). *Ontologiebasierte Indexierung und Kontextualisierung multimedialer Dokumente für das persönliche Wissensmanagement*. Universität Dresden.

Mitschick, A., & Meissner, K. (2008). Metadata generation and consolidation within an ontology-based document management system. *International Journal Metadata, Semantics and Ontologies*, *3*(4), 249-259.

Molina, A., Chen, D., Panetto, H., Vernadat, F., & Whiteman, L. (2005). Enterprise Integration and Networking: Issues, Trends and Vision. In P. Bernus & M. S. Fox (Eds.), *Knowledge Sharing in the Integrated Enterprise. Interoperability Strategies for the Enterprise Architect* (pp. 303-313). New York.

Moody, D. L. (2005). Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. *Data & Knowledge Engineering*, *55*(3), 243-276. doi:10.1016/j.datak.2004.12.005

Motik, Boris, & Sattler, U. (2006). A Comparison of Reasoning Techniques for Querying Large Description Logic ABoxes. *13th International Conference on Logic for*

Programming Artificial Intelligence and Reasoning (LPAR 2006) (pp. 227-241). Phnom Penh, Cambodia.

NISO. (2004). Understanding Metadata. Retrieved from http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3-26. doi:10.1075/li.30.1.03nad

OAI. (2008). The Open Archives Initiative Protocol for Metadata Harvesting. Retrieved November 28, 2012, from http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

OMG. (2008). Semantics of Business Vocabulary and Business Rules (SBVR). *Business*. Retrieved from http://www.omg.org/spec/SBVR/1.0/index.htm

OMG. (2011a). Business Process Model and Notation (BPMN). *Business*. Retrieved from http://www.omg.org/spec/BPMN/2.0

OMG. (2011b). OMG Meta Object Facility (MOF) Core Specification. Object Management Group.

OMG. (2012). Service oriented architecture Modeling Language (SoaML) Specification. *Language*.

Ochoa, X., Cardinaels, K., Meire, M., & Duval, E. (2005). Frameworks for the Automatic Indexation of Learning Management Systems Content into Learning Object Repositories. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 1407-1414). Montréal, Kanada. Retrieved from http://www.editlib.org/p/20276

Ohren, O. P. (2005). An Ontological Approach to Characterising Enterprise Architecture Frameworks. In P. Bernus & M. Fox (Eds.), *Knowledge Sharing in the Integrated Enterprise. Interoperability Strategies for the Enterprise Architect* (pp. 131-142). New York: Springer.

Osterwalder, A., & Pigneur, Y. (2002). An e-Business Model Ontology for Modeling e-Business. *Proceedings of the 15th Bled Electronic Commerce Conference*. Bled, Slovenia.

O'Leary, D. E. (2010). Enterprise ontologies: Review and an activity theory approach. *International Journal of Accounting Information Systems*, *11*(4), 336-352. Elsevier Inc. doi:10.1016/j.accinf.2010.09.006

Palmirani, M., & Brighi, R. (2003). Metadata for the legal domain. *Proceedings of the 14th International Workshop on Database and Expert Systems Applications*, 553-558. IEEE Comput. Soc. doi:10.1109/DEXA.2003.1232080

Patton, M., Reynolds, D., Choudhury, G. S., & DiLauro, T. (2004). Toward a Metadata Generation Framework. *D-Lib Magazine*, *10*(11). Retrieved from http://www.dlib.org/dlib/november04/choudhury/11choudhury.html

Pease, A., Niles, I., & Li, J. (2002). The Suggested Upper Merged Ontology : A Large Ontology for the Semantic Web and its Applications. *AAAI Technical Report*, *28*, 7-10.

Peirsman, Y., Deyne, S. D., Heylen, K., & Geeraerts, D. (2008). The Construction and Evaluation of Word Space Models. *Proceedings of the Language Resources and Evaluation Conference (LREC)* (pp. 3082-3088). Marrakech, Morocco.

Phipps, J., Hillmann, D. I., & Paynter, G. (2005). Orchestrating Metadata Enhancement Services: Introducing Lenny. *Proceedings of the Int. Conf. on Dublin Core and Metadata Applications* (pp. 49-58). Madrid, Spain.

Pickert, S. (2008). tun, wenn das Backup-Zeitfenster zu klein für die Datenmenge ist? Retrieved July 24, 2010, from http://www.searchstorage.de/themenbereiche/backup-recovery/fundamente/articles/160519

Pinto, H. S., & Martins, J. P. (2004). Ontologies: How can They be Built? *Knowledge and Information Systems*, *6*(4), 441-464. doi:10.1007/s10115-003-0138-1

Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M., & Rosati, R. (2008). Linking Data to Ontologies. *Journal on data semantics X*. Berlin / Heidelberg: Springer.

Pohl, K., & Rupp, C. (2009). *Basiswissen Requirements Engineering* (1st ed.). Heideberg: dpunkt.verlag.

Powell, A., Nilsson, M., Maeve, A., Johnston, P., & Baker, T. (2007). DCMI Abstract Model. *DCMI Recommendation*. Retrieved December 2, 2012, from http://dublincore.org/documents/abstract-model/index.shtml

Prekop, P., & Burnett, M. (2003). Activities , context and ubiquitous computing. *Computer Communications*, *26*(11), 1168-1176. Retrieved from http://www.sciencedirect.com/science/article/B6TYP-479VG0H-2/2/a2971390fc6e677b6697c72de9b114b0

RDF and SPARQL: Using Semantic Web Technology to Integrate the World's Data. (2007).*Early draft of a submission to the VLDB conference*. Retrieved February 4, 2012, from http://www.w3.org/2007/03/VLDB/

Rahman, A., Hossain, A., Kiringa, I., & El Saddik, A. (2006). Towards an Ontology for MPEG-7 Semantic Descriptions. *Event (London)*. Ottawa: University of Ottawa.

Reeve, L., & Han, H. (2005). Survey of semantic annotation platforms. *Proceedings of the 2005 ACM symposium on Applied computing - SAC '05* (p. 1634). New York, USA: ACM Press. doi:10.1145/1066677.1067049

Riss, U. V., Jurisch, M., & Kaufman, V. (2009). E-mail in Semantic Task Management. *2009 IEEE Conference on Commerce and Enterprise Computing* (pp. 468-475). Ieee. doi:10.1109/CEC.2009.76

Rizzo, G., & Troncy, R. (2011). NERD : Evaluating Named Entity Recognition Tools in the Web of Data. *Workshop on Web Scale Knowledge Extraction (WEKEX'11)* (pp. 1-16). Bonn, Germany.

Ruvane, M. B. (2005). Defining Annotations : a visual (re) interpretation 1. *Proceedings of the American Society for Information Science and Technology* (pp. 1-3).

Saba, F., & Mohamed, Y. (2012). A Semantic Approach to Representation, Sharing and Discovery of Construction Simulation Models. *Construction Research Congress 2012: Construction Challenges in a Flat World* (pp. 591-601). West Lafayette, Indiana, United States: American Society of Civil Engineers.

Sabol, V., Granitzer, M., Tochtermann, K., & Sarka, W. (2005). MISTRAL – MEASURABLE , INTELLIGENT AND RELIABLE SEMANTIC EXTRACTION AND RETRIEVAL OF MULTIMEDIA DATA. *Proceedings of the 2nd European Workshop Integration of Knowledge, Semantics and Digital Media Technology, 2005 (EWIMT 2005)* (pp. 349 - 355). London, UK.

Saggion, H. (2008). Mining Profiles and Definitions with Natural Laguage Processing. In H. A. Do Prado & E. Ferneda (Eds.), *Emergin Technologies of Text Mining. Techniques and Applications*. IGI Global.

Sahoo, S. S., Halb, W., Hellmann, S., Idehen, K., Thibodeau, T. (Jr), Auer, S., Sequeda, J., et al. (2009). A Survey of Current Approaches for Mapping of Relational Databases to RDF.

Saidani, O., & Nurcan, S. (2007). Towards Context Aware Business Process Modelling. *Context*. Retrieved from http://lams.epfl.ch/conference/bpmds07/program/Saidani_33.pdf

Sarawagi, S. (2008). Information Extraction. *Foundations and Trends in Databases archive*, *1*(3), 261-377. doi:10.1561/1500000003

Saunders, M., Lewis, P., & Thornhill, A. (2007). *Research Methods for Business Students* (4th ed.). Pearson Education Limited.

Scheer, A. (2000). *ARIS - Business Process Modeling*. Berlin / Heidelberg: Springer.

Schekkerman, J. (2004). *How to survive in the jungle of Enterprise Architecture Frameworks*. Victoria, B.C.: Trafford.

Schelp, J., & Winter, R. (2009). Language communities in enterprise architecture research. *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09*, 1. New York, New York, USA: ACM Press. doi:10.1145/1555619.1555650

Schilit, B. N., & Theimer, M. M. (1994). Disseminating Active Map Information to Mobile Hosts 1 Introduction. New York.

Schmidt, Albrecht. (2002). *Ubiquitous Computing – Computing in Context Computing in Context*. *Environments*. Lancaster University. Retrieved from http://www.google.ch/url?sa=t&source=web&ct=res&cd=1&ved=0CAwQFjAA&url=http://www.comp.lancs.ac.uk/~albrecht/phd/Albrecht_Schmidt_PhD-Thesis_Ubiquitous-Computing_print1.pdf&rct=j&q=schmidt+ubiquitous+computing+context&ei=BLQSS-DvOtLm-Qb2urS1AQ&usg=AFQjCNFKrKoKhCCCiAlzAsRklkskP4O60A

Schmiemann, M. (2006). SMEs and entrepreneurship in the EU. *Statistics in focus - Industrie, Trade, Services*, *24*, 1-8. Retrieved from http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=DSP_GLOSSARY_NOM_DTL_VIEW&StrNom=CODED2&StrLanguageCode=EN&IntKey=17399050&RdoSearch=BEGIN&TxtSearch=enterprise&CboTheme=&IntCurrentPage=1

Schulz, S., Dittrich, K., König, W., Oberweis, A., Rannenberg, K., & Wahlster, W. (2003). Kontext als Beziehung: Ein Kontextmodell für Mobiles Wissensmanagement (pp. 392-397). Frankfurt am Main: Köllen Druck & Verlag GmbH. Retrieved from http://www.google.ch/url?sa=t&source=web&ct=res&cd=5&ved=0CB8QFjAE&url=http://subs.emis.de/LNI/Proceedings/Proceedings35/GI-Proceedings.35-68.pdf&rct=j&q=lenat+dimensions+context-space&ei=W1ARS9_IM9D7_AbS6-DjBA&usg=AFQjCNGW9t7MhDSu3pceAk4s808tOVeXbA

Schönherr, M. (2004). Enterprise Architecture Frameworks. *Architecture*. Retrieved from

Shi, H., Maly, K., Zeil, S., & Zubair, M. (2011). Comparison of Ontology Reasoning Systems Using Custom Rules. *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*.

Sieber, P. (2009). *Enterprise-Search Katalysator für den internen Informations- und Wissensfluss*. Bern. Retrieved from www.sieberpartners.ch

Simperl, E., Thurlow, I., Davies, J., Warren, P., Dengler, F., Grebelnik, M., Mladenic, D., et al. (2010). Overcoming Information Overload in the Enterprise. *Ieee Internet Computing*, *14*(6), 39-46.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2006). Pellet : A Practical OWL-DL Reasoner. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, (D).

Smith, H., & Fingar, P. (2003). *Business Process Management : The Third Wave About this Book*. *Business* (p. 292). Meghan-Kiffer Press.

Soules, C. A. N., & Ganger, G. R. (2005). Connections: Using Context to Enhance File Search. *ACM SIGOPS Operating Systems Review*, *39*(5), 119-132.

Sowa, J. F., & Zachman, J. A. (1992). Extending and formalizing the framework for information systems architecture. *IBM Systems Journal*, *31*(3), 590 - 616.

Spanos, D.-E., Stavrou, P., & Mitrou, N. (2011). Bringing relational databases into the Semantic Web: A survey. *Semantic Web – Interoperability, Usability, Applicability*, *0*, 1-41. doi:10.3233/SW-2011-0055

Spira, J. B. (2008). Information Overload: Now $900 Billion – What is Your Organization's Exposure? Retrieved July 24, 2010, from http://www.basexblog.com/2008/12/19/information-overload-now-900-billion-what-is-your-organizations-exposure/

Spohr, D., Cimiano, P., McCrae, J., & O'Riain, S. (2012). Using SPIN to Formalise Accounting Regulations on the Semantic Web. *Proceedings of the First International Workshop on Finance and Economics on the Semantic Web (FEOSW 2012)*. Heraklion, Greece.

Spree, U. (2009). Wissensorganisation und Records Management: Was ist der State of the Art ? *Wissenschaft & Praxis*, *60*, 339-354.

Steinbrecher, W., & Müll-Schnurr, M. (2010). *Prozessorientierte Ablage. Dokumentenmanagement-Projekte zum Erfolg führen. Praktischer Leitfaden für die Gestaltung einer modernen Ablagestruktur.* (2nd ed., pp. 1-283). Wiesbaden: Gabler GWV Fachverlage GmbH.

Stracke, C. M. (2010). The Benefits and Future of Standards: Metadata and Beyond. *Proceedings of the 4th International Conference on Metadata and Semantics, MTSR1200* (pp. 354-261).

Stuckenschmidt, Heiner, & Van Harmelen, F. (2001). Ontology-Based Metadata Generation from Semi-Structured Information. *Proceedings of the 1st international conference on Knowledge capture* (pp. 163-170). Victoria, British Columbia, Canada.

Stuckenschmidt, Heiner. (2011). *Ontologien. Konzepte, Technologien und Anwendungen*. (O. P. Günther, W. Karl, R. Lienhart, & K. Zeppenfeld, Eds.)*Informatik im Fokus* (2nd ed., pp. 1-273). Springer Berlin Heidelberg.

Studer, R., Abecker, A., & Decker, S. (1999). Informatik-Methoden für das Wissensmanagement. *System*. Teubner Verlag.

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge Engineering: Principles and methods. *Data & Knowledge Engineering*, *25*, 161-197.

Su, X., & Ilebrekke, L. (2006). A Comparative Study of Ontology Languages and Tools. In L. N. in C. Science (Ed.), *Advanced Information Systems Engineering* (pp. 765-777). Springer Berlin / Heidelberg.

Sure, Y., Staab, S., & Studer, R. (2009). Ontology Engineering Methodology. In S. Staab & R. Studer (Eds.), *Handbook on Ontologies* (pp. 135-152). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-92673-3

Suárez-Figueroa, M. C., García-castro, R., Villazón-Terrazas, B., & Gómez-Pérez, A. (2011). Essentials In Ontology Engineering: Methodologies, Languages, And Tools.

Proceedings of the 2nd Workshop organized by the eeb data models community- CIB conference W078-W012 (pp. 9-21). Sophia Antipolis, Francia. Retrieved from http://oa.upm.es/9739/

Tang, J., Hong, M., Zhang, D., Liang, B., Li, J., Do Prado, H. A., & Ferneda, E. (2008). Information Extraction: Methodologies and Applications (Emergin Te., pp. 1-33). Hershey: Information Science Reference.

Tanner, C. (2009). Bedeutung der elektronischen Archivierung von Geschäftsdokumenten in Schweizer KMU - Eine Expertenbefragung. *Arbeitsberichte der Hochschule für Wirtschaft FHNW*, (13), 35. Retrieved from http://www.ecademy.ch/ecademy/ecadpubli.nsf/id/637

Tennant, R. (2004). Digital Libraries : Metadata's Bitter Harvest. *Library Journal*, 1-2.

The Open Group. (2009a). TOGAF 9. Retrieved from http://www.opengroup.org/togaf/

The Open Group. (2009b). *ArchiMate 1.0 Specification*. *Zhonghua er bi yan hou tou jing wai ke za zhi = Chinese journal of otorhinolaryngology head and neck surgery* (Vol. 42, pp. 1-110). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20087110

The Open Group. (2012). ArchiMate 2.0 Specification. The Open Group. Retrieved from http://pubs.opengroup.org/architecture/archimate2-doc/

Thönssen, Barbara. (2010). An Enterprise Ontology Building the Bases for Automatic Metadata Generation. *Proceedings of the 4th International Conference on Metadata and Semantics, MTSR1200* (pp. 195-210). Madrid.

Thönssen, Barbara. (2011). Formalizing low - level governance instruments for a more holistic approach to automatic metadata generation. *Proceedings of the 5th International Conference on Methodologies, Technologies and Tools enabling e-Government* (pp. 1-12). Camerino, Italy.

Thönssen, Barbara, & Lutz, J. (2012). SEMANTICALLY ENRICHED OBLIGATION MANAGEMENT. An Approach for Improving the Handling of Obligations Represented in Contracts. *Proceedings of 4th Conference on Knowledge Management and Information Sharing (KMIS2012)*. Barcelona, Spain.

Thönssen, Barbara, & Wolff, D. (2010). A broader view on Context Models to support Business Process Agility. In S. Smolnik, F. Teuteberg, & O. Thomas (Eds.), *Semantic Technologies for Business and Information Systems Engineering: Concepts and Applications*.

Thönssen, Barbara, Martin, A., De Rosa, F., Nikles, S., & Wagner, S. (2011). D1 . 2 Software Requirements Specification. *Applied Sciences*.

Thönssen, B., Stojanovic, L., & Pariente, T. (2005). OntoGov - Description of Ontologies (Addendum to D2).

TopQuadrant. (n.d.). Towards Executable Enterprise Models : Building Semantic Enterprise Architecture Solutions with TopBraid Suite ™. *Representations*. Retrieved from http://www.topquadrant.com/docs/whitepapers/WP-BuildingSemanticEASolutions-withTopBraid.pdf

Umar, A., & Zordan, A. (2009). Enterprise Ontologies for Planning and Integration of Business: A Pragmatic Approach. *IEEE Transactions on Engineering Management*, *56*(2), 352-371. doi:10.1109/TEM.2009.2013823

Uren, V., Cimiano, P., Handschuh, S., Vargas-vera, M., Motta, E., & Ciravegna, F. (2006). Semantic annotation for knowledge management : Requirements and a survey of the state of the art. *World Wide Web Internet And Web Information Systems*, *4*, 14-28. doi:10.1016/j.websem.2005.10.002

Uschold, M., & Grüninger, M. (1996). Ontologies : Principles, Methods and Applications. *TECHNICAL REPORT- UNIVERSITY OF EDINBURGH ARTIFICIAL INTELLIGENCE APPLICATIONS INSTITUTE AIAI TR*, (191).

Uschold, M., & King, M. (1995). Towards a Methodology for Building Ontologies. *Methodology*.

Uschold, M., King, M., Moralee, S., & Zorgios, Y. (1997). *The Enterprise Ontology*. *The Knowledge Engineering Review* (Vol. 13).

Valtonen, K., Mäntynen, S., Leppänen, M., & Pulkkinen, M. (2011). Enterprise Architecture Descriptions for Enhancing Local Government Transformation and Coherency Management. *2011 IEEE 15th International Enterprise Distributed Object Computing Conference Workshops*, 360-369. Ieee. doi:10.1109/EDOCW.2011.39

Vassiliadis, P. (2009). A survey of Extract – transform – Load technology. *International Journal*, *5*(September), 1-27.

Vavliakis, K. N., Grollios, T. K., & Mitkas, P. A. (2010). RDOTE - Transforming Relational Databases. *Proceedings of the 9th International Semantic Web Concerence (ISWC)* (pp. 1-4). Shanghai, China.

Vavliakis, K. N., Symeonidis, A. L., Karagiannis, G. T., & Mitkas, P. a. (2011). An integrated framework for enhancing the semantic transformation, editing and querying of relational databases. *Expert Systems with Applications*, *38*(4), 3844-3856. Elsevier Ltd. doi:10.1016/j.eswa.2010.09.045

Vernadat, F. (2010). Technical, semantic and organizational issues of enterprise interoperability and networking. *Annual Reviews in Control*, *34*(1), 139-144. doi:10.1016/j.arcontrol.2010.02.009

Volz, R., Handschuh, S., Staab, S., Stojanovic, L., & Stojanovic, N. (2004). Unveiling the hidden bride: deep annotation for mapping and migrating legacy data to the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, *1*(2), 187-206. doi:10.1016/j.websem.2003.11.005

W3C. (2009). VirtuosoUniversalServer. Retrieved February 5, 2012, from http://www.w3.org/wiki/VirtuosoUniversalServer

W3C. (2010). SKOS Simple Knowledge Organization System - Home Page. Retrieved January 19, 2011, from http://www.w3.org/2004/02/skos/

W3C OWL Working Group. (2004). OWL Web Ontology Language Overview. *W3C Recommendation 10 February 2004*. Retrieved December 5, 2012, from http://www.w3.org/TR/owl-features/

W3C OWL Working Group. (2009). OWL 2 Web Ontology Language Document Overview. *W3C Recommendation 27 October 2009*. Retrieved December 5, 2012, from http://www.w3.org/TR/owl2-overview/

W3C Working Group. (2011). RDB2RDF Implementations. Retrieved January 25, 2012, from http://www.w3.org/2001/sw/rdb2rdf/wiki/Implementations

WSDL Working Group. (2007). Semantic Annotations for WSDL and XML Schema. *Editors' Copy*. Retrieved December 5, 2012, from www.w3.org/2002/ws/sawsdl/spec/

Wache, H., Bergmayr, A., Feldkamp, D., Kondylakis, H., Nikles, S., & Plexousakis, D. (2010). *PlugIT. D3.2 Specification of Next Generation Modelling Framework. WP 3.2 Specification of the Next Generation Modelling Framework. Framework*.

Wache, H., Vögele, U., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., & Hübner, S. (2001). Ontology-Based Integration of Information — A Survey of Existing Approaches. *Proceedings of IJCAI-01 Workshop: Ontolgoies and Information Sharing* (pp. 108-117). Seattle, WA.

Wang, H. H., Dzulkifli, M., & Ismail, N. A. (2010). Semantic Gap in CBIR : Automatic Objects Spatial Relationships Semantic Extraction and Representation. *International Journal Of Image Processing (IJIP)*, *4*(3), 192-204.

Wang, X. H., Gu, T., Zhang, D. Q., & Pung, H. K. (2004). Ontology Based Context Modeling and Reasoning using OWL. *Workshop Proceedings of the 2nd IEEE Conference on Pervasive Computing and Communications* (pp. 18–22). Orlando, FL, USA.

Warner, S. (2001). Exposing and harvesting metadata using the OAI metadata harvesting protocol: A tutorial. *High Energy Physics Libraries Webzine*, (4), 1-13.

Waterfeld, W., Weiten, M., & Haase, P. (2008). Ontology Management Infrastructures. In M. Hepp, P. De Leenheer, A. De Moor, & Y. Sure (Eds.), *Ontology Management - Semantic Web, Semantic Web Services, and Business Applications* (pp. 71-101). SpringerScience + Bsuiness Media Inc.

Weikum, G., & Theobald, M. (2010). From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. *Proc of ACM symposium on principles of database systems (PODS)*. Indianapolis, USA.

Whittaker, S. (2011). Personal Information Management: From Information Consumption to Curation. *Annual review of information science and technology ARIST (2011)*, *45*, 3-62. Retrieved from http://people.ucsc.edu/~swhittak/papers/Information_curation_whittaker.pdf

Whittaker, S., Bellotti, V., & Gzisda, J. (2006). Email in Personal Information Management. *Communications of the ACM*, *49*(1).

Widmer, T., Landert, C., & Bachmann, N. (2000). EVALUATIONS-STANDARDS der Schweizerischen Evaluationsgesellschaft (SEVAL-Standards). *Evaluation*. Schweizerische Evaluationsgesellschaft.

Winograd, T. (2001). Architectures for Context. *Human Computer Interaction*, *16*, 401 - 419.

Woitsch, R., Karagiannis, D., Plexousakis, D., & Hinkelmann, K. (2009). Business and IT alignment: the IT-Socket. *Elektrotechnik und Informationstechnik*, *126*(7), 308-321. doi:10.1007/s00502-009-0660-2

Woodley, M. S. (2008). Crosswalks, Metadata Harvesting, Federated Searching, Metasearching: Using Metadata to Connect Users and Information. In M. Baca, P. E. Pardo, S. U. Berg, & E. Zozom (Eds.), *Elements* (2nd ed., pp. 38-63). Los Angeles: Getty Research Institute. Retrieved from http://hdl.handle.net/10211.2/2001

Woods, W. A. (1975). *What's in a Link: Foundations for Semantic Networks*. *Contract* (pp. 1-76). Cambridge. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA022584

Wu, F., & Weld, D. S. (2007). Autonomously Semantifying Wikipedia. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (ACM 07)* (pp. 41-50).

Yu, L. (2011). *Developer's Guide to the Semantic Web* (online.). A. Springer. Retrieved from <http://lib.myilibrary.com?ID=308236>

Zachman, J. A. (1987). A framework for information systems architecture. *IBM Systems Journal*, *26*(3).

Zachman, J. A. (2003). *The Zachman Framework for Enterprise Architecture™: A Primer for Enterprise Engineering and Manufacturing*. Electronic book published March 2003. Retrieved from www.zachmaninternational.com.

Zachman, J. A. (2008). The Zachmann Framework. Retrieved April 8, 2011, from http://test.zachmaninternational.com/index.php/the-zachman-framework

Zachman, J. A. (2012). The Zachman Framework For Enterprise Architecture. A Primer For Enterprise Engineering And Manufacturing. Retrieved from http://www.zachmanframeworkassociates.com/index.php/ebook-on-enterprise-architecture

Zeng, M. L., & Qin, J. (2008). *Metadata*. London.

Zernik, U. (2008). Device, System And Method For Determining Document Similarities And Differences.

Zhang, G., Jia, S., Wang, Q., & Liu, Q. (2010). REA-based Enterprise Business Domain Ontology Construction. *Journal of Software*, *5*(5), 522-529. doi:10.4304/jsw.5.5.522-529

Zhang, H. (2011). *Personal Information Organization and Re-Access In Computer Folders: A Qualitiative Study of Information Workers*. University of Illinois at Urbana-Champaign. Retrieved from http://scholar.googleusercontent.com/scholar?q=cache:i0-RpKm5-pkJ:scholar.google.com/&hl=de&as_sdt=0

Zhang, Z. (2005). *Ontology Query Languages For The Semantic Web: A Performance Evaluation*. University of Georgia Theses and Dissertations. Retrieved from http://athenaeum.libs.uga.edu/handle/10724/8545

Zöller, B. (2006). *Dokumenten Management Systeme: Hersteller und Produktüberblick* (p. 675). Bonn: VOI Verband Organisations- und Infomationssysteme. Retrieved from www.voi.de

Zöller, B. (2007). ANBIETERSTRATEGIEN UND IHRE BEDEUTUNG FÜR DIE ANWENDER : WOZU BRAUCHT MAN EIGENTLICH NOCH EIN DMS ? *5. Branchenguide zur DMS Expo 2007*, 18-22. Retrieved from http://zoeller.de/archiv/files/veroeffentlichungen/2007/Branchenguide_DMS_Expo_2007.pdf

Zöller, B. (2009). Exzellente Aussichten: Triebfedern des ECM-Marktes. *7. Branchenguide zur DMS Expo 2009*. Retrieved from http://zoeller.de/archiv/files/veroeffentlichungen/2009/Vorwort_DMS_Expo_Brancheguide_2009_BZ.pdf

van Lamsweerde, A. (2009). *Requirements Engineering: From System Goals to UML Models to Software Specifications* (1st ed.). Wiley.

12 Appendix

12.1 *Evaluation of Metadata Harvesters*

As there is no harvester tool prevailing publicly available software and tools, either open source, free ware or share ware, are evaluated for further use in my approach. Acting on the survey study of Albassuny (2008) tools are evaluated available via the web. The assessment was performed by Johner (2011) within a research project, which was part of the Master programme Business Information Systems at FHNW, supervised by me. In the following I summarize the results.

Tools are classified based on the "Taxonomy of Metadata Tool Functionalities" provided by Greenberg & Severiens (2007) but more restrictive confined. Whereas Greenberg & Severiens (2007) regard the class "Metadata Harvesting" similar to the "Metadata Extraction"" class here, ""Metadata Harvesting" is used only for tools reading out metadata *associated* to a document, like file attributes. Thus, a tool like Apache Lucene[222] that extracts metadata from (textual) content of a document *and* from its file attributes is considered of class "Metadata Extraction". Similarly the class "Metadata Creation" is restricted to tools, offering manual or semi-automatic creation of metadata. Furthermore the class "Semantic Annotating" is introduced for tools, enhancing harvesting or extracted metadata by referencing to classes in the ontology and to instances. Unfortunately most if not "all existing semantic annotation systems rely on human intervention at some point in the annotation process" (Reeve & Han, 2005) hence delimitation to ""Metadata Creation" might be blurred.
As my work aims to provide a solution for all document formats (text, still and moving image and audio), without manual metadata creation, only metadata harvesting tools are further considered.

12.1.1 Overview on Harvesting Tools

For finding harvesting tools literature review has been done, for example of the survey of Automatic Metadata Generation Applications conducted by Albassuny (2008). In addition the internet has been queried.
In the following the harvesting tools to be evaluated are briefly introduced.

- Metadata Extraction Tool
  In 2007, the National Library of New Zealand released its current version of the Metadata Extraction Tool. The focus of this application lies in the extraction and preservation of file metadata. The software application exports the file attributes into XML structured files. As an export format, the two options "NLNZ Data Dictionary" and "Extract in Native Form" are provided. The tool is open source and could thus be extended according to an institutions need.

- File Identifier
  File identifier is provided as a free but proprietary tool from the Optima SC Inc.[223] The main purpose of this tool is to identify the file type and for certain formats the embedded metadata. The File Identifier runs as a command line tool According to the developer, the full version supports the Dublin Core metadata descriptors.

[222] Apache Lucene Features. URL: http://lucene.apache.org/java/docs/features.html (retrieved: 13.2.1011)
[223] Obtima SC Inc. URL :http://www.optimasc.com/ (retrieved: 6.9.2011)

- Extract Metadata from Multiple Files Software
  This commercial software was developed by Sobolsoft[224]. Once the repository is scanned, the file metadata are shown as a table. Those results could be exported as a list or as an Excel file. In the free version, however, the metadata can be displayed but not exported properly.

- Metadata Miner Catalogue
  The Metadata Miner was developed by SoftExperience[225] in 2009. It extracts the file metadata and provides various possibilities for generating CSV and XML reports. Together with XSL files, those metadata can be converted into Dublin Core and RDF files. This harvesting tool from SoftExperience supports besides the standard file attributes also custom, user-defined file attributes.

- Embedded Metadata Extraction Tool
  The Embedded Metadata Extraction Tool was developed by ARTstor[226], a non-profit organization which supports the domain of artworks (and collections) through digital initiatives. The EMET tool is specialized in the image formats JPG and TIFF. The harvested information is dumped into a file and can be stored in a location of the user's choice. The tool was chosen for this project due to the multitude of JPG metadata attributes tested.

- InsideCAT
  The harvesting tool InsideCAT[227] is provided by Víctor García Tascón. It is available as a free version or as a commercial, so called professional application. The main purpose of InsideCAT consists in cataloguing discs, i.e. with focus on CDs and DVDs, and thus making the files easier retrievable. The metadata are displayed as a list and can be exported in a single file. Furthermore, queries over the catalogued repository can be run.

- JHOVE
  The "JSTOR /Harvard Object Validation Environment"[228] was developed to verify format consistency of a file prior to archiving. The program determines, whether a file is well-formed, when it was created, it generates a checksum and so on. Besides those features, JHOVE offers modules for different file types and the possibility to dump JPG and PDF file information, i.e. to read out metadata. The information is displayed on a command line interface. JHOVE is provided free of charge as collaborative project between the JSTOR and the Harvard University Library.

Table 48 gives an overview on harvesting tools which have been evaluated.

[224] Sobolsoft. URL: http://www.sobolsoft.com/ (retrieved: 6.9.2011)
[225] SoftExperience. URL: http://peccatte.karefil.com/software/ (retrieved: 6.9.2011)
[226] ARTstor. URL: http://www.artstor.org/index.shtml (retrieved: 6.9.2011)
[227] InsideCAT. URL: http://www.insidecat.biz/disk_cataloguer/disk_cataloguer.htm (6.9.2011)
[228] JHOVE. URL: http://hul.harvard.edu/jhove/ (retrieved: 6.9.2011)

Tool Name	Source / URL	Brief Description
Metadata Extraction Tool	http://www.natlib.govt.nz/services/get-advice/digital-libraries/metadata-extraction-tool	Developed by the National Library of New Zealand, the Metadata Extraction Tool is provided free of charge and with its source code.
File Identifier	http://www.optimasc.com/products/f ileid/index.html	The file identifier from Optima SC is given free of charge as a command line tool. Simple reports can be generated.
Extract Metadata from Multiple Files Software (hereinafter referred to as 'Sobolsoft')	http://www.sobolsoft.com/extracte tadata/	This tool from Sobolsoft is easy to use and can deal with most of the formats. The data can be exported as an Excel file or list. The tool is proprietary.
Metadata Miner Catalogue	http://peccatte.karefil.com/software/ Catalogue/MetadataMiner.htm	The Metadata Miner from SoftExperience is a proprietary tool which supports most of the file attributes and offers various export possibilities.
Embedded Metadata Extraction Tool	http://www.artstor.org/global /g- html/download-emet-public.html	This tool was developed by the Non-Profit organization ARTstor. Due to the organizational goals, i.e. the preservation of art, this tool is specialized on images.
InsideCAT	http://www.insidecat.biz/disk_catalo guer/disk_cataloguer.htm	InsideCAT catalogues the content of discs. It focuses on DVD and CD collections.
JHOVE	http://hul.harvard.edu/jhove/distribu tion.html	JHOVE is a metadata harvesting tool developed for the long time preservation of electronic files. Besides extracting metadata, it does validate the documents and generates a checksum.

Table 48: Evaluated Harvesting Tools

12.1.2 Evaluation Criteria

Evaluation has been clustered according to three aspects: Attributes Aspect, the Functional Aspect and the Non-Functional Aspect.

- The Attribute Aspect

  Albassuny (2008) puts a strong emphasis on the supported attributes. This criterion will test, whether a given attribute was identified by the harvesting tools. All the tested attributes are listed in Chapter 12.1.2.

  In Table 49: Schema for Attribute Evaluation are the details listed.

  ApplicationAttribute

Description	For each format, the attributes to be tested are identified. This criterion indicates for each attribute, whether the harvesting tool was able to harvest it.	
Values	yes	The given harvesting tool was able to extract this attribute
	no	The harvesting tool was not able to identify it.

Table 49: Schema for Attribute Evaluation

- The Functional View

  Ares Casal et al. (1998) propose to conduct a functional analysis of the software application and then derive the criteria based on this analysis. For the purpose of my work, the main functions of a metadata harvesting tool are the extraction of appropriate metadata and the provision of those data in well-structured format. Thus, the following criteria are proposed as detailed in Table 50: Schema for Export Syntax Evaluation, Table 51: Schema for Crosswalk Capabilities Evaluation, Table 52: Schema for Ease of Use Evaluation and Table 53: Schema for Customizing and Filtering Options Evaluation:

  Export Syntax

Description	Ability of the application to export and structure the collected data	
Values	none	The tool cannot export the data
	Export	Harvested data is exported as plain text file
	CVS	Harvested data is exported as CSV or Excel File
	XML	Harvested data is exported in XML format
	RDF	Harvested data is exported in XML/RDF format

Table 50: Schema for Export Syntax Evaluation

CrosswalkCapabilities

Description	The tool is able to consolidate the different metadata and describe them in a standardized manner using Dublin Core (DC)	
Values	none	The tool cannot consolidate the different metadata and describe them in a standardized manner using Dublin Core
	own	The tool uses an own description schema
	DC	Harvested data is described according to Dublin Core

Table 51: Schema for Crosswalk Capabilities Evaluation

Ease of Use

Description	How easy the tool is to use	
Values	easy	The tool has a GUI and is self-explicable
	medium	The tool may have a GUI but is quite complex in usage
	hard	The tool does not have a GUI; additional technologies have to be installed or at least knowledge of how to use the command line is necessary

Table 52: Schema for Ease of Use Evaluation

Customizingand Filtering Options

Description	Evaluates, whether the harvesting tool can be customized and adapted according to the repository	
Values	none	The repository's files are harvested as-is
	filter	Certain filtering criteria can be defined, such as which formats and attributes to be harvested
	strategy	The harvest can be customized; at least a schedule can be set up Preferably, the crawl strategy can be defined, e.g. stating only to harvest update

Table 53: Schema for Customizing and Filtering Options Evaluation

- The Non-functional Aspect

  While functional criteria evaluate the functionalities and use-cases of a given tool, non-functional criteria focus on formal aspects. The evaluation of non-functional aspects focuses on accessibility and costs. As most likely a harvester needs to be adapted for the metadata generation prototype, e.g. with respect to file formats, the tool must at least be customizable. As it is a prototype that is developed – not a full blown application – it should be free of costs.

  In Table 54: Schema for Adaptability Evaluation and Table 55: Schema for Cost Evaluation are the details listed.

Propriety vs. Open Source

Description	This criterion indicates, whether the application can be extended (Open Source) or if it is only available as a binary file (Proprietary). This distinction is of importance if new or application specific file formats should be harvested or the harvested metadata should be in a certain format, e.g. for importing into a legacy system	
Values	Open Source (OS)	The tool is Open Source; a source file is available for further development
	no	The software is proprietary and cannot be modified or at least customized

Table 54: Schema for Adaptability Evaluation

Free vs. Commercial

Description	This criterion clarifies, whether the tool can be used at no charge or not (i.e. only commercial version is available). If a commercial and a cost-free version exists, the commercial version is evaluated because of assumed larger functionality	
Values	free	Full functionality is provided without charge
	commercial	Full functionality is provided with costs

Table 55: Schema for Cost Evaluation

Table 56 summarizes the evaluation criteria presented above.

Aspect	Criterion No	Criterion	Brief Explanation	Source
Attribute	1	Application Attribute	The testing of the application attributes is the core of this evaluation. This criterion is important in order to determine, which formats and what kind of attributes a harvesting tool supports. A list of all tested attributes can be found in Chapter 12.1.4	(Albassuny, 2008)
Functional	2	Export Syntax	The ability to export is an essential precondition for future transfers to other PCs and further processing. This criteria shows, which kind of syntax / structure is applied	(Ares Casal et al., 1998)

Aspect	Criterion No	Criterion	Brief Explanation	Source
Functional	3	Crosswalk Capabilities	If the export is done using an established description schema, like for instance Dublin Core, the result can be easier interpreted	(Ares Casal et al., 1998)
Functional	4	Ease of use	For the end-user, the ease of use is an important criteria in order to exploit all available functionalities	(Ares Casal et al., 1998)
Functional	5	Customizing and Filtering Options	The capability of customizing or filtering the harvest according to the repositories structure helps to save scarce resources such as processing time	(Ares Casal et al., 1998)
Non-Functional	6	Proprietary vs. Open source	This criterion is important in order to know, whether the source code is provided and if the tool can be further extended	(Fuggetta, 2003)
Non-Functional	7	Free vs. Commercial	Depending on the budget of the organization, a free application might be preferable. It is important to note that free does not automatically mean open source	(Lerner & Tirole, 2002)

Table 56: Evaluation Criteria for Harvesting Tools

12.1.3 Evaluation Environment and Procedure

The test environment consists of two components, the operating system the tools are implemented on and the repository of test documents. For testing a virtual machine with Windows XP Professional (Microsoft, 2011) and service pack 2 has been set up

As shown in the representative study (cf. Chapter 4.1) as well as in the survey (cf. Chapter 4.2.1.2) Microsoft Office file formats, i.e. DOC, XLS, PPT and VSD are the predominantly used format for written texts, calculations and graphs in enterprise. It is assumed that with increasing use of Windows 7 the new file format equivalents, i.e. DOCX, XLSX, PPTX and VDX are used, too. Additionally, PDF and PDF-A were included for file formats which can be used across platforms and for archiving. Considered applications and file formats are listed in Table 57.

Document Creation	File Type-Standard
Adobe Acrobat Reader	PDF
Adobe Acrobat Reader	PDF-A
Microsoft Excel	XLS
Microsoft Excel 2010	XLSX

Document Creation	File Type-Standard
Microsoft PowerPoint	PPT
Microsoft PowerPoint	PPTX
Microsoft Visio	VSD
Microsoft Visio 2010	VDX
Microsoft Word	DOC
Microsoft Word 2010	DOCX
n.a.	MP3
n.a.	MP4
amongst other MS	GIF
n.a.	JPG-EXIF
n.a.	JPG-IPTC
amongst other MS	PNG

Table 57: Applications and File Formats

Besides the file attributes provided by document creation software, harvesting of the Windows XP Operating System (OS) properties were evaluated as well.

For evaluation the following procedure has been applied for all tested harvesting tools:

1. Install the identified tools in the test-environment
2. Create a test-repository with all relevant file types
3. Establish a file list with all attributes and generate a unique value for each attribute
4. Harvest the metadata with the installed tools. Store the harvesting results into one or several files.
5. Document results in a matrix: Attributes x - Harvesting Tools
6. Evaluate the functional and non-functional criteria and document them in a separate list.

12.1.4 Evaluation Results

Evaluation results are structured according to the evaluation aspects introduced above. The results are clustered according to three aspects introduced above.

The attributes aspect contains 263 file attributes. Only 116 out of these 263 attributes were harvested at least by one tool, 147 were not addressed by any tool. The functional aspect is about the qualitative criteria, for instance the cross-walk capability; non-functional aspects highlight accessibility or costs.

- Results for Attribute Aspect

In the table below (Table 58: Evaluation Results for File Attributes), all fields attributes are summarized that have been harvested by at least one of the evaluated harvesting tools. In the rightmost column, a score is provided, i.e. it is counted by how many tools an attribute has been harvested. As NLNZ' harvester is the tool chosen for the metadata generation prototype that column is highlighted.

File Type-Standard	Attributes	Metadata Extractor	Embedded Metadata Extractor	File Identifier	InsideCAT	JHOVE	Metadata Miner Catalogue	Sobolsoft	Score
DOC	Title	Yes	No	Yes	Yes	No	Yes	Yes	5
DOC	Author	Yes	No	Yes	No	No	Yes	Yes	4
DOC	Subject	Yes	No	No	No	No	Yes	Yes	3
DOC	Keywords	Yes	No	Yes	No	No	Yes	No	3
DOC	Category	No	No	No	No	No	Yes	Yes	2
DOC	Comments	Yes	No	No	No	No	Yes	Yes	3
DOC	Manager	No	No	No	No	No	Yes	No	1
DOC	Company	No	No	No	No	No	Yes	No	1
DOC	Adapt (custom	No	No	No	No	No	Yes	No	1
DOCX	Title	No	No	No	No	No	No	Yes	1
DOCX	Author	No	No	No	No	No	No	Yes	1
DOCX	Subject	No	No	No	No	No	No	Yes	1
DOCX	Category	No	No	No	No	No	No	Yes	1
DOCX	Comments	No	No	No	No	No	No	Yes	1
IPG-IPTC	Headline	No	Yes	No	No	No	Yes	No	2
IPG-IPTC	Date Created	No	Yes	No	No	No	Yes	No	2
IPG-IPTC	Caption/ Abstract	No	Yes	No	No	No	Yes	No	2
IPG-IPTC	Keywords	No	Yes	Yes	No	No	Yes	No	3
IPG-IPTC	By-line	No	Yes	Yes	No	No	Yes	No	3
IPG-IPTC	By-line Title	No	Yes	No	No	No	Yes	No	2
IPG-IPTC	Copyright	No	Yes	No	No	No	Yes	No	2
IPG-IPTC	Credit	No	Yes	No	No	No	Yes	No	2
IPG-IPTC	Contact	No	Yes	No	No	No	Yes	No	2

File Type-Standard	Attributes	Metadata Extractor	Embedded Metadata Extractor	File Identifier	InsideCAT	JHOVE	Metadata Miner Catalogue	Sobolsoft	Score
JPG-IPTC	Object Name	No	Yes	Yes	No	No	Yes	Yes	4
JPG-IPTC	Time Created	No	Yes	No	No	No	Yes	No	2
JPG-IPTC	City	No	Yes	No	No	No	Yes	No	2
JPG-IPTC	Sublocation	No	Yes	No	No	No	Yes	No	2
JPG-IPTC	Province/ State	No	Yes	No	No	No	Yes	No	2
JPG-IPTC	Country/ Primary Location Code	No	Yes	No	No	Yes	Yes	No	3
JPG-IPTC	Country/ Primary Location Name	No	Yes	No	No	No	Yes	No	2
JPG-IPTC	Source	No	Yes	No	No	No	Yes	No	2
JPG-IPTC	Originating	No	Yes	No	No	No	Yes	No	2
JPG-IPTC	Edit Status	No	Yes	No	No	No	Yes	No	2
JPG-IPTC	Original Transmission	No	Yes	No	No	No	Yes	No	2
JPG-IPTC	Writer/ Editor	No	Yes	No	No	No	Yes	No	2
JPG-IPTC	Category	No	Yes	No	No	No	Yes	No	2
JPG-IPTC	Supplemental	No	Yes	No	No	No	Yes	No	2
MP3	Title	No	No	Yes	No	No	No	Yes	2
MP3	Year	No	No	Yes	Yes	Yes	No	Yes	4
MP3	Contributing	No	No	Yes	No	No	No	Yes	2
MP3	Album	No	No	No	No	No	No	Yes	1
MP3	Genre	No	No	No	No	No	No	Yes	1
MP3	Track Number	No	No	No	No	No	No	Yes	1

File Type-Standard	Attributes	Metadata Extractor	Embedded Metadata Extractor	File Identifier	InsideCAT	JHOVE	Metadata Miner Catalogue	Sobolsoft	Score
MP3	Comments	No	No	No	No	No	No	Yes	1
MP4	Year	No	No	No	No	Yes	No	No	1
MP4	Frame height	No	No	Yes	No	No	Yes	No	2
MP4	Channels	No	No	No	No	No	No	Yes	1
PDF	Title	No	No	No	No	Yes	Yes	Yes	3
PDF	Author	No	No	No	No	Yes	Yes	Yes	3
PDF	Subject	No	No	No	No	Yes	Yes	Yes	3
PDF	Keywords	No	No	No	No	Yes	Yes	No	2
PDF	PDF Version	Yes	No	No	No	Yes	Yes	No	3
PPT	Title	Yes	No	Yes	No	No	Yes	Yes	4
PPT	Author	Yes	No	Yes	No	No	Yes	Yes	4
PPT	Subject	Yes	No	No	No	No	Yes	Yes	3
PPT	Keywords	Yes	No	Yes	No	No	Yes	No	3
PPT	Category	No	No	No	No	No	Yes	Yes	2
PPT	Comments	Yes	No	No	No	No	Yes	Yes	3
PPT	Manager	No	No	No	No	No	Yes	No	1
PPT	Company	No	No	No	No	No	Yes	No	1
PPT	Adapt (custom keywords)	No	No	No	No	No	Yes	No	1
PPTX	Title	No	No	No	No	No	No	Yes	1
PPTX	Author	No	No	No	No	No	No	Yes	1
PPTX	Subject	No	No	No	No	No	No	Yes	1
PPTX	Category	No	No	No	No	No	No	Yes	1
PPTX	Comments	No	No	No	No	No	No	Yes	1

File Type-Standard	Attributes	Metadata Extractor	Embedded Metadata Extractor	File Identifier	InsideCAT	JHOVE	Metadata Miner Catalogue	Sobolsoft	Score
XLS	Title	Yes	No	Yes	No	No	Yes	Yes	4
XLS	Author	Yes	No	Yes	No	No	Yes	Yes	4
XLS	Subject	Yes	No	No	No	No	Yes	Yes	3
XLS	Keywords	Yes	No	Yes	No	No	Yes	No	3
XLS	Category	No	No	No	No	No	Yes	Yes	2
XLS	Comments	Yes	No	No	No	No	Yes	Yes	3
XLS	Manager	No	No	No	No	No	Yes	No	1
XLS	Company	No	No	No	No	No	Yes	No	1
XLS	Adapt (custom keywords)	No	No	No	No	No	Yes	No	1
XLSX	Title	No	No	No	No	No	No	Yes	1
XLSX	Author	No	No	No	No	No	No	Yes	1
XLSX	Subject	No	No	No	No	No	No	Yes	1
XLSX	Category	No	No	No	No	No	No	Yes	1
XLSX	Comments	No	No	No	No	No	No	Yes	1
OS	Type of File	No	No	No	No	No	No	Yes	1
OS	Location	Yes	Yes	Yes	No	Yes	No	Yes	5
OS	Created	No	No	No	No	No	No	Yes	1
OS	Modified	Yes	No	No	No	Yes	Yes	Yes	4
OS	File Name	Yes	No	Yes	No	Yes	Yes	Yes	5
JPG-Properties	Dimensions	No	No	No	No	No	No	Yes	1
VSD	Title	No	No	Yes	No	No	Yes	Yes	3
VSD	Author	No	No	Yes	No	No	Yes	No	2

File Type-Standard	Attributes	Metadata Extractor	Embedded Metadata Extractor	File Identifier	InsideCAT	JHOVE	Metadata Miner Catalogue	Sobolsoft	Score
VSD	Subject	No	No	No	No	No	Yes	Yes	2
VSD	Keywords	No	No	Yes	No	No	Yes	No	2
VSD	Category	No	No	No	No	No	Yes	Yes	2
VSD	Comments	No	No	No	No	No	Yes	Yes	2
VSD	Manager	No	No	No	No	No	Yes	No	1
VSD	Company	No	No	No	No	No	Yes	No	1
JPG-EXIF	Image Description	Yes	Yes	No	No	No	No	No	2
JPG-EXIF	Make	Yes	Yes	No	Yes	No	No	No	3
JPG-EXIF	Camera Model	Yes	No	No	Yes	No	No	Yes	3
JPG-EXIF	X Resolution	Yes	Yes	No	Yes	Yes	Yes	No	5
JPG-EXIF	Y Resolution	Yes	Yes	No	Yes	Yes	Yes	No	5
JPG-EXIF	Software	Yes	No	No	No	No	No	No	1
JPG-EXIF	Modify Date	Yes	Yes	No	Yes	No	No	No	3
JPG-EXIF	Exposure Time	No	No	No	Yes	No	No	No	1
JPG-EXIF	Exif Version	No	Yes	No	No	No	No	No	1
JPG-EXIF	Date/Time	Yes	Yes	No	Yes	No	No	No	3
JPG-EXIF	Create Date	Yes	Yes	No	Yes	No	No	No	3
JPG-EXIF	Brightness Value	No	Yes	No	No	No	No	No	1
JPG-EXIF	Light Source	Yes	Yes	Yes	No	No	No	No	3
JPG-EXIF	Minolta Quality	Yes	No	No	No	No	No	No	1
JPG-EXIF	Flash Mode	No	No	No	No	No	Yes	No	1
JPG-EXIF	Exif Image Width	Yes	No	Yes	No	Yes	No	Yes	4
JPG-EXIF	Exif Image Height	Yes	No	Yes	No	Yes	No	Yes	4
JPG-EXIF	Custom Rendered	No	No	No	No	No	Yes	No	1

File Type-Standard	Attributes	Metadata Extractor	Embedded Metadata Extractor	File Identifier	InsideCAT	JHOVE	Metadata Miner Catalogue	Sobolsoft	Score
JPG-EXIF	Sharpness	No	No	No	No	No	Yes	No	1
JPG-EXIF	PrintIM Version	Yes	No	Yes	No	Yes	No	Yes	4
JPG-EXIF	Image Size	No	No	Yes	No	Yes	No	No	2
JPG-EXIF	Shutter Speed	No	No	No	Yes	No	No	No	1

Table 58: Evaluation Results for File Attributes

- Results for Functional Aspect

  In Table 59: Evaluation Results for Functional Aspects are the details listed.

Tool	Export Syntax	Crosswalk Capabilities	Ease of Use	Customiz-ation and Filtering	Remarks
Embedded Metadata Extraction Tool	export	none	easy	none	
File Identifier	XML	Dublin Core	hard	none	This command line crawler states to have cross-walk capabilities to Dublin Core. However, the free version used for the test does not support this feature.
InsideCAT	export	none	easy	filter	
Metadata Extractor	"Native" format	none	medium	strategy	This application provides two export formats, both XML based. Crosswalk to Dublin Core is not implemented
	"NLNZ DTD"				
JHOVE	export	none	hard	none	
Sobolsoft	export	none	easy	none	Good commercial tool for metadata extraction; easy to use but many limitations in the free version, e.g. No export capabilities.

Tool	Export Syntax	Crosswalk Capabilities	Ease of Use	Customiz-ation and Filtering	Remarks
Metadata Miner Catalogue	XML	Dublin Core	medium	strategy	This harvester offers various cross-walking possibilities with XSL files, including Dublin Core

Table 59: Evaluation Results for Functional Aspects

- Results for Non-Functional Aspects

  In Table 60: Evaluation Results for Non-functional Aspects are the details listed.

Tool	Proprietary vs. Open Source	Free vs. Commercial	Remarks
Embedded Metadta Extraction Tool	proprietary	Free	Program was developed for an art collection; thus, it focuses on pictures.
File Identifier	proprietary	commercial	
InsideCAT	proprietary	commercial	Mainly focuses on disc cataloguing such as DVDs or CD s.
Metadata Extractor	OS	Free	possibility to create new adaptors for extraction
JHOVE	OS	Free	Difficult to install file validator. More suitable for digital archiving than for metadata harvesting.
Sobolsoft	proprietary	commercial	
Metadata Miner Catalogue	proprietary	commercial	

Table 60: Evaluation Results for Non-functional Aspects

- Summary of Evaluation Results

Supported Formats and Attributes

Concerning the number of recognized attributes, the harvester from SoftExperience obtained the highest success rate, followed by Sobolsoft. The statistics are presented in the table below (Table 61: Attribute Scores per Harvesting Tool); the column "Attribute Score" indicates, how many attributes were detected, regardless of the format.

Tool	Attribute Score
Metadata Miner Catalogue (SoftExperience)	71
Sobolsoft	56
Embedded Metadata Extraction Tool (ARTstor)	34
Metadata Extraction Tool (National Library of New Zealand, for short NLNZ)	33
File Identifier (Optima SC)	26
JHOVE	17
InsideCat	11

Table 61: Attribute Scores per Harvesting Tool

Concerning the support of attributes and formats it can be concluded that while the harvesting tool of SoftExperience examines the files more in depth and is able to harvest the highest number of attributes, Sobolsoft recognizes a bigger variety of different formats. The harvesters of NLNZ and ARTstor could be considered as second-best alternatives. Although ARTstor's Embedded Metadata Extraction Tool harvests one file attribute more than NLNZ' Metadata Extractor the result should be put into perspective as the ARTstor tool is limited to multi-media file attributes. While Optima, NLNZ and SoftExperience provide support for "old" Office documents, Sobolsoft is the only tool evaluated that supports the latest Office formats such as DOCX, XLSX and PPTX.

VDX, i.e. the XML Visio file format is not harvest by any of the tools.

Export and Crosswalks

In general, all tools provide export functions for harvested metadata. More than half of the tools support a structured way of export, using the CSV or XML format. NLNZ for instance, supports two XML-based crosswalk possibilities: the "NLNZ Data Dictionary" and the "Extract in Native Form". While the former is a defined mapping to a NLNZ schema, the latter is closer to the harvested file format. Sobolsoft is able to harmonize and export the results in a table format, using its own descriptors.
Whereas all tools allow for export, only two provide crosswalks to Dublin Core, namely the Metadata Miner Catalogue and File Identifier. Metadata Miner Catalgoue crosswalks can be realized by using an XSL file which, in conjunction with the original XML results,

displays the harvested metadata mapped to a chosen format. The Metadata Miner Catalogue has the biggest range of crosswalk possibilities.
In the case of the File Identifier, the export and crosswalk feature could not be tested in the cost-free version. In the manual, a report option mapping the harvesting results to Dublin Core is mentioned.

Customizing and Filtering Options

Two tools, InsideCat and Metadata Miner Catalogue, allow to filter the harvesting results. While the former only allows for filtering the results, the latter can be configured prior to the harvest, i.e. it can be defined which formats and attributes should be harvested. NLNZ' Metadata Extractor can be configured in such a way, that it starts the harvest at a certain time.
None of the other tools allow for any configuration.

Open Source and Commercialism

From all the evaluated tools, only NLNZ' Metadata Extractor and JHOVE are pure open source projects. For these tools source code is available and can be modified as needed[229]. The Embedded Metadata Extraction Tool is free of charge but proprietary. Full functionality of all other harvesting tools is available with costs.

Ease of Use

Most of the tools can be installed and used without any major problem. The main challenges occurred with command line tools such as JHOVE and File Identifier. Both harvesters offer special command line syntax; JHOVE in addition needs configuration before running. In general, command line tools are slightly more difficult to get used to.

Best Practices and Possible Enhancements

The harvesting tool should be easy installable, usable and maintainable. It should customizable, offering features to configure the harvest, i.e. which folders, formats and file attributes to harvest. Additionally, a tool should offer the possibility to schedule the harvests and to record, which documents were already harvested and in consequence only consider new or recently modified files.
In addition extensibility and adaptability of tools would be desirable. For file format recognition and attribute harvesting, the support of additional plug-ins or modules dedicated to a certain format would be a good enhancement. With respect to this NLNZ' Metadata Extractor goes in that direction. Unfortunately is the functionality not supported by an appropriate user interface.
Concerning the export structure and crosswalks, the SoftExperience's Metadata Miner Catalogue can be seen as a best practice example. It stores the metadata in an XML file and provides differently mapped XSL files with it. This system would allow mapping the harvesting results to a specific description schema, using the syntax they need. Such functionality could be applied to all harvesting tools which produce an XML-based output.

[229] The Metadata Extraction Tool is written in Java and XML and available under the Apache Public License (version 2), http://meta-extractor.sourceforge.net/ (retrieved: 6.9.2011).
JHOVE is available under the GNU Library or Lesser General Public License (LGPL), http://sourceforge.net/projects/jhove/ (retrieved: 6.9.2011)

However, which harvesting tool is 'the best' depends first and foremost on the needs of the enterprise: the document creation tools the company uses and thus, the file types and formats that must be supported. Second, the file attributes that should be used for search and third, if only automatically created file attributes are to be harvested or user defined attributes, too. Finally, the needed representation (format) of the results is an important decision criterion.

- Metadata Miner Catalogue
  This application is the best tool if the repositories mainly consist of documents in "old" Office or multi-media file formats (e.g. doc, vsd, jpg) and the formats will not change – at least in the immediate future. The tool recognizes many file attributes (including the user-defined ones) and is able to export the data using XML syntax into Dublin Core.
  Unfortunately SoftExperience does not offer the possibility to define new harvesting schema and thus, the tool is inflexible with respect to supporting new file formats. However, its strength lies in the provided crosswalk templates and the possibility to create own mappings. Metadata Miner Catalogue is proprietary and commercial.

- Sobolsoft
  This harvesting tool is the ideal choice if many different formats, such as "old" and "new" (XML based) Office files must be harvested. The harvester recognizes most formats but is limited in export functionality. It only supports export into an list or an Excel table and offers no crosswalk possibilities. The Sobolsoft tool is proprietary and commercial.

- Metadata Extraction Tool
  If enterprise specific adaptations are needed, e.g. because file formats are changing often or specific file formats need to be supported, NLNZ' Metadata Extraction Tool is recommended. Besides its flexibility and extensibility the tool offers the possibility to create user defined file schemas (so called adapters) for harvesting and export schemas for interchange. The native output format is based on XML which could also be mapped using the XSL files (as done by SoftExperience). Due to the fact that the Metadata Extractor is open source, it can be easily integrated into other applications and is available free of charge.

- Embedded Metadata Extraction Tool, JHOVE and InsideCAT are not considered suitable in the business context because of their specialization in images, digital preservation and DVDs/CDs. The harvesting tool File Identifier of Optima SC can be neglected as all of its features are implemented better in one of the other tools.

12.1.5 Conclusion

For building the metadata generation prototype the Metadata Extraction Tool provided by the National Library of New Zealand is considered the best.

The Metadata Extraction Tool

- supports all file formats and attributes created by applications and operating system used by AHSGA
- can be enhanced if necessary (e.g. creating new schema for harvesting) as its source code is available
- provide result files in an XML structure and support the flexible creation of new description schemas

- can be easily integrated in another application and
- is free of charge.

12.2 Excerpt of MATURE Representative Study

12.2.1 Codes of Software

In Table 62: Coding of Software **(Barnes et al. 2010, p202)** are the details listed.

Code	Example	Description
collaboration_tool.conferencing.audio	"audio conferencing"; "Voice over IP"	audio conference call with two or more participants
collaboration_tool.conferencing.desktop	"web meetings", "we use virtual rooms for desktop sharing"; MS Live Meeting	desktop sharing and videoconference tools accessed by employee from own workplace/desktop
collaboration_tool.conferencing.video	"we use videoconferences to transfer knowledge"	dedicated videoconferencing system, sometimes located in special conference rooms
collaboration_tool.generic	clearspace	tool supporting collaboration of team members
collaboration_tool.instant messenger	Skype; Lotus Sametime	tool for chat video and audio calls used at own workplace
collaboration_tool.peer_to_peer	MS Groove	tool supporting collaboration of team members that is peer to peer based
custom.generic	"adapted systems are used"	adapted or (self)developed software is used and was not specified by interviewee
custom.nonproductive_traningsystem	"we teach people by using a copy of the productive system"	custom nonproductive system that is a mirror of the productive system and is used to train people
custom.search_engine	"we use a search function that was developed by our IT department"	custom built search engine to find digital artefacts
desktoppublishing.generic	quark express; Adobe INDESIGN	tools to supporting the creation of publication documents
desktoppublishing.pdf	Adobe Acrobat, Adobe Reader, PDF Creator	tools supporting the creation of PDF documents
dms.adapted	We have aligned the functionality of the DMS to our needs	Document management system that was adapted according to organisation's requirements
DMS.generic	"we introduced a DMS"	a not specified document management system
elearning_tool	WBT, e-learning	specific type of tools used for training of employees at their desktops
elearning_tool.custom	"Learning Content Management System (have their own tailored system)"	specific custom type of tools used for training of employees at their desktops

Code	Example	Description
elearning_tool.flash	WBT, elearning based on flash technology	flash-based tools used for training of employees at their desktops
ERPcontrolling	SAP R/3 CO	ERP software with focus on supporting controlling
ERPCRM	SAP CRM; salesforce	ERP software with focus on supporting customer relationship management
ERPfinance	SAP FI; Sage finance software (for smaller organisations)	ERP Software with focus on supporting finance
ERPfinance.custom	financiero	custom made ERP software, focus on finance
ERPgeneric	SAP ERP	used to manage internal and external resources of the organisation (not specified by interviewee)
ERPhealth_care	SAP IS-H, or hospital management system	ERP software with focus on supporting management of resources in hospitals
ERPhuman_resources	SAP HCM	ERP software with focus on supporting human resources
ERPlegal	MILES33	ERP software with focus on supporting legal
ERPplant_maintenance	SAP PM	ERP software with focus on supporting plant maintenance
ERPprocurement	SAP MM	ERP software with focus on supporting procurement
ERPprod_planning	SAP APO, SAP PP/DS	ERP software with focus on supporting production planning
extranet.generic	extranet	organisational network based on internet architecture that is extended to users outside the company
filebrowser	Windows Explorer, "Filesystem"	tool used to navigate through file systems on own desktop or on network share
graphic_editing_program.generic	photoshop	tool for creating and manipulating images
ide.software_development	eclipse	integrated development environment for creating (platform dependent) software
ide.web_publishing	macromedia dreamweaver	integrated development environment for creating comprehensive web pages
informally.all_allowed	"no restrictions at all via central IT"	use this code to show that everything is allowed and therefore no software is really 'informal'

Code	Example	Description
informally.not_allowed	"informally software is not allowed" or "users have not the rights to install"	use this code to show that informally software is not allowed
informally.not_allowed.but_used	"informally software is not allowed but used on private laptop brought to organisation"	code to show that informally software is not allowed but used nevertheless
informally.not_existent	_	use this if no codes or comments made by interviewer
internet.generic	internet, Internet Explorer, portal	generic service or website in the internet, accessed via specific client (browser)
internet.RSS_feeds	RSS feed	software based on a standard for accessing news on webpages
internet.social_software	people search in internet forums; linkedIn; Xing	software (platform) aimed at managing contacts and networking with people
internet.WCMS	RedDot, Typo3	web content management system to maintain internet web pages
internet.WCMS.wiki	MediaWiki, Confluence	Type of CMS for collaborative editing of contents
intranet.form	intranet forms	Specific form which is accessible via intranet of the organisation
intranet.generic	intranet, Internet Explorer, sharepoint, portal	organisational network, based on internet architecture; accessed via client (browser)
intranet.social_software	"we have introduced a knowledge forum"; blog, tagging environment	software (platform) aimed at managing contacts and networking with people within the intranet of the company
intranet.wcms	web content management system to maintain intranet web pages; wiki published on intranet; wordpress	web content management system to maintain digital contents web pages in the intranet
intranet.wcms.wiki	wiki on the intranet	Type of CMS for collaborative editing of contents in the intranet
ITSM_tool	tool to support IT service management	tool for supporting management tasks aligned to IT service management
kms.generic	Centra Knowledge Center	systems who's primary focus in on improving the handling of knowledge
kms.skill_management	skill management system	system for managing and skills of employees

Code	Example	Description
media.video	video files, flash films	multimedia contents (for training purposes)
MIS.generic	Management Information System	System supporting managers and keeping them up to date
modeling_tool.CAM.CNC	CNC	Computer aided manufacturing using CNC code to drive numerically controlled machine tools
modeling_tool.CAM.generic	CAM	Computer aided manufacturing
modeling_tool.design_and_engineering	CAD, 3D Drawing SW	tools for designing and modeling mechanical/electronical parts
modeling_tool.enterprise	ARIS toolset	tools for modeling processes, organisational structure, etc
modeling_tool.generic	Visio	modelling tool for multiple purposes
modeling_tool.mind_maps	MindManager, FreeMind	tools for modeling mindmaps
modeling_tool.simulation	we use a simulation software to show process performance	used to simulate
office.database	Microsoft Access	database application with a primary focus on desktop use
office.generic	Microsoft Office	office application usually containing software for word processing, spreadsheets and presentations
office.generic.web_based	Google Docs	office software which is based on web technology rather than being platform dependant
office.notes	MS One Note	software for storing notes
office.presentation	Microsoft Powerpoint	application for creating electronic presentations
office.spreadsheet	Microsoft Excel	application for managing table-based data
office.spreadsheet.adapted	Macros developed with Microsoft Excel	parts of code using a programming language within a spreadsheet application
office.word_processing	Microsoft Word	primarily used for creating and editing text-based documents
office.word_processing.forms	Forms used in Microsoft Word	forms which are created using a word processing software
open.source.generic	"different open source software"	Software which is available in source code
PIM.adapted	Lotus Noted specifically adapted in order to use it for collaborative management of ideas and proposals	use this code if a personal information managment tool was adapted for specific use

Code	Example	Description
PIM.add_on	Xobni	add on to a PIM
PIM.generic	Microsoft Outlook, Lotus Notes	personal information management tool
PIM.mail	mailsystem, Microsoft Outlook for mail	personal information management tool (used mail)
PIM.newsreader	usenet	software for reading nntp based messages
PIM.sms	sms	personal information management tool (used short message service)
project_management_tool .adapted	"we use Redmine which is originally a project managment tool for idea management"	use this code if a project management tool was adapted to a specific use
project_management_tool .generic	MS Project; backlog (scrum --> mostly excel based)	use this code for a project management tool
simulation.generic	simulation tool	tool for modeling and running simulations of real world processes
suggestion_system.custom	“we have developed our own suggestion system”	software used for collecting and managing ideas and suggestions of employees which was customized
suggestion_system.generic	suggestion system; idea management system	software used for collecting and managing ideas and suggestions of employees
trouble_ticket_system	trouble ticket solutions; "if there are ideas for optimization, that we open a change request - you can do it via a ticket system, which we also have, you can give requirements to our helpdesk and ask questions"	software used to track trouble tickets

Table 62: Coding of Software (Barnes et al. 2010, p202)

12.2.2 Digital Resources Used in Knowledge Maturing Activities

In Table 63: Use of Digital Resources in Knowledge Maturing Activities **(Barnes et al. 2010, p34)** are the details listed.

Knowledge Maturing Activitiy	example provided by interviewer
find relevant digital resources	Search for information, e.g. documents, web pages or images.
embed information at individual or organisational level	Include the information into one’s own knowledge base, which could be a (personal or shared) file system, a (personal/team/corporate) wiki, or similar.

Knowledge Maturing Activitiy	example provided by interviewer
keep up-to-date with organisation-related knowledge	making sure that oneself or another person stays up-to-date regarding a certain topic
familiarise oneself with new information	Making oneself familiar with e.g. a topic or a community or processes
reorganise information at individual or organisational level	Restructure collections (file systems, wikis, …), consolidate different approaches to collective structuring, removing outdated items, improving findability through assigning metadata, "gardening" of wikis, vocabularies etc., rearrange contents or files, clean-up work spaces and assure quality of a collection of digital resources
reflect on and refine work practices or processes	This reflects process maturing from discovery of task or process patterns, the analysis thereof to improving practices and/or processes. The knowledge maturing activity thus comprises practices (i.e. not formally specified), procedures (informal or endorsed) as well as processes (specified, defined)
create and co-develop digital resources	Generate new or update existing contents by oneself or together with others. Note: co-development is a form of collaboration.
share and release digital resources	Share denotes the informal, release the formal or official part of granting access to contents for a specified or unspecified group of people.
restrict access and protect digital resources	Restricting access to contents.
find people with particular knowledge or expertise	identify a contact person, e.g. by skills
communicate with people	interact with others, e.g. face-to-face, by phone, by mail
assess, verify and rate information	Evaluate contents with respect to certain quality criteria like accurateness, up-to-dateness, usefulness or people with respect to their capacities or behaviour

Table 63: Use of Digital Resources in Knowledge Maturing Activities (Barnes et al. 2010, p34)

12.2.3 Use of Digital Resources as Knowledge Maturing Indicator

Use of digital resources in an enterprise has been considered Knowledge Maturing (KM) Indicator. Table 64: Knowledge Maturing Indicators – as Used in Representative Study **(Barnes et al. 2010, p 36)** gives examples of digital resources investigated in the study.

KM Indicator	example
has been accepted into a restricted domain	article published on company's intranet
has become part of a guideline or has become standard	pdf file became part of user manual
has not been changed for a long period after intensive editing	wiki article remains unchanged since its last major editing

KM Indicator	example
was selected from a range of resources	specific document was chosen out of list of search results
became part of a collection of similar information	folder containing documents on the same topic
was created/refined in a meeting	word document reworked during project meeting
was prepared for a meeting	PowerPoint presentation prepared for project meeting
was created by integrating parts of other digital resources	presentation created using information from two sources
was made accessible to a different user group	access to a document restricted to administrative users
was presented to an influential audience	report presented to the board of directors
is referred to by another resource	wiki article referred within a protocol
has been the subject of many discussions	several emails sent between parties about structure of document

Table 64: Knowledge Maturing Indicators – as Used in Representative Study (Barnes et al. 2010, p 36)

12.3 *Questionnaire on Document Handling in Enterprises*

Questions and pre-defined answers conducted in the survey on document handling in enterprise as detailed in Table 65: Questionnaire on Document Handling in Enterprise.

Question	No	Possible Answers
Do you use a tool to manage/search for electronic documents you use/create at work?	1.1	Document Management System (DMS)
	1.2	Web Content Management System (CMS), e.g. for Web sites
	1.3	Enterprise Content Management System (ECMS)
	1.4	Records Management System (RMS), e.g. for electronic archiving
	1.5	Enterprise Resource Planning System (ERP), e.g. for accounting documents or HR documents
	1.6	other (e.g. PICASA for images or an ACCESS-DB for DVDs)
	1.7	no
What document forms do you work with?	2.1	Text
	2.2	Image
	2.3	Video / podCasts / DVD
	2.4	Audio
	2.5	other (what?)
What document creation software do you use?	3.1	MS Office
	3.2	Enterprise specific software (legacy systems)
	3.3	ERP system
	3.4	CAD
	3.5	other (which?)
Do you use templates for electronic document creation? (e.g. an application form or design templates for presentations)	4.1	yes
	4.2	no
How many templates do you use?	5.1	up to 10
	5.2	11-30
	5.3	more than 30
Document creation software automatically adds attributes to a document, like creation date or file seize. Do you know that you can add more attributes to describe the documents (e.g. for search)	6.1	yes
	6.2	no
	6.3	heard about but never tried
	6.4	tried but didn't work as expected
Where do you store the documents?	7.1	on my personal computer
	7.2	on a server ('myDirectory')

Question	No	Possible Answers
	7.3	on a server accessible for 'all'
	7.4	in a system, e.g. a DMS
	7.5	others (where?)
Does your organization define the storage structure, or parts of it?	8.1	yes
	8.2	no
	8.3	partially (e.g. upper structure)
How is the directory structure organized?	9.1	organisational structure (e.g. departments)
	9.2	business aspects (e.g. projects, customers, suppliers)
	9.3	spatial aspects (e.g. countries, regions)
	9.4	temporal aspects (e.g. year, month)
	9.5	other criteria (which?)
Does the structure correlates to a filing structure (filing plan)?	10.1	yes
	10.2	no
How do you search for a document?	11.1	Browsing the directory structure
	11.2	using file search functions (e.g. windows search function)
	11.3	using a tool for desktop search (e.g. Google)
What attributes/terms do you use to search for documents?	12.1	date
	12.2	author / creator
	12.3	filename
	12.4	document format (e.g. pdf, doc)
	12.5	other (what?)
With which attributes would you like to search (but currently can't, e.g. subject, document type like report)?	13.1	title
	13.2	description
	13.3	type
	13.4	subject
	13.5	source
	13.6	relation
	13.7	coverage
	13.8	creator
	13.9	publisher
	13.10	rights
	13.11	contributor
	13.12	date
	13.13	format (e.g. pdf, doc)
	13.14	identifier
	13.15	language
	13.16	other
Does your organization define	14.1	yes

Question	No	Possible Answers
naming conventions for file names?	14.2	no
Do you know which of the electronic documents you handle are legally binding? (e.g. a project offer you sent out via mail?)	15.1	yes
	15.2	no
Where are legally binding documents stored?	16.1	paper prints in folders
	16.2	digital signed copies locally (e.g. in 'my folder')
	16.3	digital signed copies centrally (e.g. in an electronic archiving system)
	16.4	within the creating system (e.g. a legacy system, ERP system)
	16.5	where they are (no specific treatment)
	16.6	do not know
Does your organization use governance instruments?	17.1	Balanced Score Card
	17.2	EFQM
	17.3	ISO 9001
	17.4	Enterprise Architecture
	17.5	other (what?)
	17.6	no
Does your organisation uses tools for skills/experience management? (e.g.: how can you find out if someone else in your enterprise is working on a topic that could be interesting for you?)	18.1	"Yellow Pages"
	18.2	(Enterprise) Face Book
	18.3	(Enterprise) Blog
	18.4	(Enterpise) Wiki
	18.5	we all know us personally so everybody knows who's doing what
	18.6	other (what?)
	18.7	no
What do you like managing the documents as you currently do? (for example if you use the explorer structure that no extra effort for storing is needed or, if you use a DMS search is better supported)		
What would improve your document management?		

Table 65: Questionnaire on Document Handling in Enterprise

12.4 *Use Cases for Automatic, Format-independent Metadata Generation Based on Context*

12.4.1 UC1 Modify Directory

In Table 66: UC1 Modify Directory are the details listed.

Section	Content
Identifier	UC1
Name	Modify Directory
Description	A user creates, updates or a deletes a document stored in the directory that is set up for harvesting
Triggering event	File manipulation
Actors	Business User (or system, in case a modification is performed by an application)
Pre-condition	-
Post-condition	UC1.1 is executed
Result	New documents are stored in, modified documents are updated in and deleted documents are removed from the directory
Main scenario	A user creates, updates or a deletes a document stored in the monitored directory
Alternative scenarios (optional)	-
Exceptional scenarios (optional)	-

Table 66: UC1 Modify Directory

12.4.2 UC1.1 Create Delete List

In Table 67: UC1.1 Create Delete List are the details listed.

Section	Content
Identifier	UC1.1
Name	Create Delete List
Description	In case a document is deleted by a user an entry in a delete list is made
Triggering event	Execution of the use case is triggered by deletes of documents in the monitored directory
Actors	System

Section	Content
Pre-condition	UC1 is executed
Post-condition	The name of the deleted document is stored in the delete list
Result	For every deleted document an entry with the document's name is created in the delete list
Main scenario	A document stored in a monitored directory of an enterprise's file system is deleted
Alternative scenarios (optional)	-
Exceptional scenarios (optional)	-

Table 67: UC1.1 Create Delete List

12.4.3 UC2 Generate Metadata

In Table 68: UC2 Generate Metadata are the details listed.

Section	Content
Identifier	UC2
Name	Generate Metadata
Description	The MeGaSystem generates metadata automatically based on harvested document properties or extracted information
Triggering event	Execution of the use case is triggered by a timer, e.g. every two hours or at 2 a.m.
Actors	MeGaSystem
Pre-condition	Application specific rules for metadata generation are defined and executable
Post-condition	For all documents of the monitored directory instances in seEAD exist
Result	For each document metadata are created and stored in the corresponding instances in seEAD;
Main scenario	For newly created or newly updated documents metadata are to be created
Alternative scenarios (optional)	For operational use of the MeGaSystem instead of a timer, creation or update of a document could trigger the generation of metadata candidates immediately
Exceptional scenarios (optional)	-

Table 68: UC2 Generate Metadata

12.4.4 UC2.1 Prepare Generation

In Table 69: UC2.1 Prepare Generation are the details listed.

Section	Content
Identifier	UC2.1
Name	Prepare Generation
Description	Metadata generation is prepared, i.e. all documents for which metadata is to be generated are checked for existing metadata candidates in seEAD and if so, these candidates are deleted
Triggering event	Execution of the use case is triggered by a timer or manually
Actors	MeGaSystem
Pre-condition	UC2 is executed
Post-condition	for all documents for which metadata is to be generated existing metadata candidates are removed from seEAD
Result	No metadata candidates for documents for which metadata is to be generated exist in seEAD
Main scenario	Already existing metadata candidates are removed before metadata generation is executed
Alternative scenarios (optional)	Metadata generation for a certain (set of) document(s) is disabled in case a user does not wish an update for any reason
Exceptional scenarios (optional)	-

Table 69: UC2.1 Prepare Generation

12.4.5 UC2.2 Harvest Document Properties

In Table 70: UC2.2 Harvest Document Properties are the details listed.

Section	Content
Identifier	UC2.2
Name	Harvest Document properties
Description	The MeGaSystem harvests document properties for documents in the monitored directory and stores them in related XML-files
Triggering event	Execution of the use case is triggered by a timer or manually
Actors	MeGaSystem
Pre-condition	UC2 is executed

Section	Content
Post-condition	Systems have harvested document properties for newly created or newly updated documents
Result	For each document a related xml-file is created containing the harvested document properties
Main scenario	For a newly created or newly updated document metadata creation is prepared
Alternative scenarios (optional)	-
Exceptional scenarios (optional)	-

Table 70: UC2.2 Harvest Document Properties

12.4.6 UC2.3 Create Metadata Seeds

In Table 71: UC2.3 Create Metadata Seeds are the details listed.

Section	Content
Identifier	UC2.3
Name	Create Metadata Seeds
Description	For harvested document properties or extracted information the MeGaSystem creates metadata seeds; Metadata seeds are instances of classes and properties in seEAD created on the basis of harvested file or content annotations or on a mix of both
Triggering event	Execution of the use case is either triggered after harvesting or information extraction is completed
Actors	MeGaSystem
Pre-condition	UC2 and (UC2.2 or UC3) is executed
Post-condition	MeGaSystem has created metadata seeds and stored in seEAD
Result	For each document, respectively each harvested attribute or annotated extracted information instances are created in seEAD
Main scenario	For harvested document properties metadata seeds are created
Alternative scenarios (optional)	For extracted and annotated information metadata seeds are created
Exceptional scenarios (optional)	-

Table 71: UC2.3 Create Metadata Seeds

12.4.7 UC2.4 Create Metadata

In Table 72: UC2.5 Create Metadata are the details listed.

Section	Content
Identifier	UC2.4
Name	Create Metadata
Description	Metadata are created from metadata seeds inferring primary context information of a document
Triggering event	UC2 or UC7
Actors	MeGaSystem
Pre-condition	UC2 and UC2.3 has been performed and metadata seeds have been generated;
Post-condition	MeGaSystem has created metadata and stored the corresponding instances in seEAD
Result	Systems has created metadata based on the primary context of a document Metadata is stored in the corresponding instances in seEAD
Main scenario	For a newly created or newly updated document metadata are to be created
Alternative scenarios (optional)	Metadata has been modified (UC7) and based on the modification new metadata must be inferred, e.g. the metadata ContractEnd has been updated the retention period must be inferred again
Exceptional scenarios (optional)	-

Table 72: UC2.5 Create Metadata

12.4.8 UC2.5 Create Metadata Candidates

In Table 73: UC2.4 Create Metadata Candidates are the details listed.

Section	Content
Identifier	UC2.5
Name	Create Metadata Candidates
Description	The MeGaSystem creates metadata candidates on the basis of a document's secondary to n-ary context information

Section	Content
Triggering event	UC2.4
Actors	MeGaSystem
Pre-condition	Application specific rules for metadata candidate generation are defined and executable; UC2, UC2.1, UC2.3 has been executed
Post-condition	Systems have created metadata candidates on the basis of a document's secondary to n-ary context information
Result	For each document all metadata candidates are created and stored in the corresponding instances in seEAD
Main scenario	For a newly created or newly updated document metadata candidates are to be created
Alternative scenarios (optional)	-
Exceptional scenarios (optional)	-

Table 73: UC2.4 Create Metadata Candidates

12.4.9 UC3 Extract Information

In Table 74: UC3 Extract Information are the details listed.

Section	Content
Identifier	UC3
Name	Extract Information
Description	Instead or in addition to metadata harvesting information can be extracted from the content of a document, e.g. from a text document; Extracted information is annotated and stored in XML-files
Triggering event	Execution of the use case is triggered by a user request
Actors	DokLifeSystem
Pre-condition	-
Post-condition	Extracted and annotated information is stored in an XML-file;
Result	An XML-file is created with annotated information that has been extracted from (text) documents
Main scenario	Information is extracted from the content of (text) documents

Section	Content
Alternative scenarios (optional)	UC4.1 may be performed in case extracted information is to be mapped to ontologically represented metadata
Exceptional scenarios (optional)	-

Table 74: UC3 Extract Information

12.4.10 UC4 Manage Enterprise Objects

In Table 75: UC4 Manage Enterprise Objects are the details listed.

Section	Content
Identifier	UC4
Name	Manage enterprise objects
Description	Representations of enterprise objects stored in a relational database of a business informations system are mapped to their ontological representation in seEAD. According to Barrasa et al. (2004) mapping can be defined as a set of correspondence that relates the vocabulary of a relational database schema with that of an ontology.
Triggering event	Execution of the use case is triggered by a third party system or a business user
Actors	BusinessInformationSystem and BusinessUser
Pre-condition	-
Post-condition	UC4.1 or UC 4.2 is executed
Result	Mapping between parts of the database schema and concepts of seEAD is set-up
Main scenario	Enterprise objects represented in seEAD can be mapped to records of a relational database of a business information system, e.g. of ITRS or CLM.
Alternative scenarios (optional)	-
Exceptional scenarios (optional)	-

Table 75: UC4 Manage Enterprise Objects

12.4.11 UC4.1 Map Metadata

In Table 76: UC4.1 Map Metadata are the details listed.

Section	Content	
Identifier	UC4.1	
Name	Map Metadata	
Description	Metadata represented as instances in the ontology are mapped to attribute values represented in a relational database	
Triggering event	Execution of the use case is triggered by a third party system	
Actors	BusinessInformationSystem	
Pre-condition	Database to ontology mapping has been set-up. Mapping between ontologically represented entities in seEAD and relationally represented entities in a business information system has been set-up	
Post-condition	-	
Result	An executable query has been generated that can be performed in seEAD	
Main scenario	Related to a query initiated by a target system an answer set is created	
Alternative scenarios (optional)	-	
Exceptional scenario	*Description*	Attribute values to be mapped to instances in seEAD are created
	Triggering event	Mapping cannot be performed
	Actors	BusinessInformationSystem
	Pre-condition	Attribute values to be mapped to instances do not exist already and must be created first
	Result	Attribute values to be mapped to instances in seEAD exist

Table 76: UC4.1 Map Metadata

12.4.12 UC4.2 Search Enterprise Object

In Table 77: UC4.2 Search Enterprise Object are the details listed.

Section	Content
Identifier	UC4.2
Name	Search enterprise object

Section	Content
Description	An enterprise object, either a business object or a representation of it, is searched for; for that a query request from a third party system is transformed into a query based on pre-defined templates
Triggering event	Execution of the use case is triggered by a query request from a third party system
Actors	BusinessInformationSystem, Business User
Pre-condition	Pre-defined templates for query transformation
Post-condition	The request is transformed into a query that can be run in seEAD
Result	An executable query has been generated that can be performed in seEAD
Main scenario	Related to a query initiated by a target system an answer set is created
Alternative scenarios (optional)	-
Exceptional scenarios (optional)	-

Table 77: UC4.2 Search Enterprise Object

12.4.13 UC4.3 Request Update

In Table 78: UC4.3 Request Update are the details listed.

Section	Content
Identifier	UC4.3
Name	Request Update
Description	Metadata is modified based on a user request
Triggering event	Execution of the use case is triggered by an request for updating metadata from a third party system
Actors	BusinessInformationSystem
Pre-condition	Pre-defined templates for request transformation
Post-condition	The request is transformed into a procedure/service that can be executed to update seEAD
Result	Modification service for update request has been created
Main scenario	Metadata (candidates) are to be updated

Section	Content
Alternative scenarios (optional)	Specific user interface is used for update request
Exceptional scenarios (optional)	-

Table 78: UC4.3 Request Update

12.4.14 UC5 Query seEAD

In Table 79: UC5 Query seEAD are the details listed.

Section	Content
Identifier	UC5
Name	Query seEAD
Description	A query is performed for the requested enterprise object, either a business object or a representation of it
Triggering event	UC4.2 or manual query creation
Actors	MeGaSystem
Pre-condition	UC2
Post-condition	A query for the requested information has been executed
Result	An answer set for the executed query has been created
Main scenario	Execution of a query in seEAD
Alternative scenarios (optional)	A query is created directly in seEAD, e.g. using the Protégé sparql plug-in
Exceptional scenarios (optional)	-

Table 79: UC5 Query seEAD

12.4.15 UC5.1 Provide Result List

In Table 80: UC5.1 Provide Result List are the details listed.

Section	Content
Identifier	UC5.1
Name	Provide Result List
Description	Results of the query are transformed into user readable format

Section	Content
Triggering event	UC5
Actors	MeGaSystem
Pre-condition	UC5
Post-condition	Results of the query have been transformed into a format that is (more easily) readable by business users
Result	A user readable result list of retrieved business objects or their representations has been created
Main scenario	Results of the query are transformed into more easily readable format for humans
Alternative scenarios (optional)	Results of the query are transformed into an XML-schema for export, respectively import into a third party system UC4.1 may be performed in case extracted information is to be mapped to ontologically represented metadata
Exceptional scenarios (optional)	-

Table 80: UC5.1 Provide Result List

12.4.16 UC6 Modify Metadata (Candidates)

In Table 81: UC6 Modify Metadata (Candidates) are the details listed.

Section	Content
Identifier	UC6
Name	Modify Metadata (Candidates)
Description	Metadata or metadata candidates (stored in seEAD) are modified, including update and delete
Triggering event	Request for modification sent by third party system
Actors	MeGaSystem
Pre-condition	UC4.3
Post-condition	Metadata is updated or deleted
Result	Metadata in seEAD is up to date
Main scenario	Metadata (candidates) are modified on request of a third party system
Alternative scenarios (optional)	Metadata is modified directly in seEAD
Exceptional scenarios (optional)	-

Table 81: UC6 Modify Metadata (Candidates)

12.5 *AHSGA Ancillary Information*

List of recorded products, services and functions in AHSGA's Information and Task Management System as detailed in Table 82: Overview on AHSGA's ITRS Records.

IntangibleProducts	Angebot Dienstleistungsprodukt	BusinessService	Leistungen	BehaviourElement BusinessFunctions
AdvisoryAndInformation	Fachberatung und Information sowie individuelle Beratung (FB&IB)	Helpdesk	Allgemeine Auskünfte und Fachinfos	GeneralInformationAndAdvise
AdvisoryAndInformation	FB&IB	Helpdesk	Support Fachpers./SchülerInnen	SupportProfessionals&Students
AdvisoryAndInformation	FB&IB	Helpdesk	Fachberatung und Information	InformationAndHotline
AdvisoryAndInformation	FB&IB	Helpdesk	Individuelle Beratung	Counseling
Prevention	Prävention	School&JuvenilePrevention	Prävention Schule und Jugend	PreschoolPrevention
Prevention	Prävention	School&JuvenilePrevention	Prävention Schule und Jugend	ElementarySchoolPrevention
Prevention	Prävention	School&JuvenilePrevention	Prävention Schule und Jugend	SecondarySchoolPrevention
Prevention	Prävention	School&JuvenilePrevention	Prävention Schule und Jugend	SocialEducationalInstitutes
Prevention	Prävention	School&JuvenilePrevention	Prävention Schule und Jugend	Education
Prevention	Prävention	School&JuvenilePrevention	Prävention Schule und Jugend	Info&TrainingInOccupationalGroups
Prevention	Prävention	School&JuvenilePrevention	Prävention Schule und Jugend	ExtracurricularDomain
Prevention	Prävention	School&JuvenilePrevention	Prävention Schule und Jugend	SexualPadagogicsIndividualWork
Prevention	Prävention	TargetGroupSpecificPrevention	Zielgruppenspezifische Prävention	PrisionPrevention
Prevention	Prävention	TargetGroupSpecificPrevention	Zielgruppenspezifische Prävention	Migration

IntangibleProducts	Angebot Dienstleistungsprodukt	BusinessService	Leistungen	BehaviourElement BusinessFunctions
Prevention	Prävention	TargetGroupSpecificPrevention	Zielgruppenspezifische Prävention	MSM-Project
Prevention	Prävention	TargetGroupSpecificPrevention	Zielgruppenspezifische Prävention	DonJuan-Project
Prevention	Prävention	ToolsForPrevention	Arbeitshilfen	WorkingAndTeachingAids
Prevention	Prävention	ToolsForPrevention	Arbeitshilfen	Material4Prevention
InfoAndPublicRelations	Informations- und Öffentlichkeitsarbeit	PublicRelations	Öffentlichkeitsarbeit	DIALOG&AHSGAPublications
InfoAndPublicRelations	Informations- und Öffentlichkeitsarbeit	PublicRelations	Öffentlichkeitsarbeit	MediaWork
InfoAndPublicRelations	Informations- und Öffentlichkeitsarbeit	PublicRelations	Öffentlichkeitsarbeit	PublicInformation&SmallExibitions
InfoAndPublicRelations	Informations- und Öffentlichkeitsarbeit	PublicRelations	Öffentlichkeitsarbeit	PublicActions
InfoAndPublicRelations	Informations- und Öffentlichkeitsarbeit	Networking	Vernetzung	Networking
		NonBillableServices	Nicht verrechenbare Leistungen	HumanResourceManagement
		NonBillableServices	Nicht verrechenbare Leistungen	Strategic&ConceptualWork
		NonBillableServices	Nicht verrechenbare Leistungen	Accounting
		NonBillableServices	Nicht verrechenbare Leistungen	Accounting
		NonBillableServices	Nicht verrechenbare Leistungen	Accounting
		NonBillableServices	Nicht verrechenbare Leistungen	ICT_Support
		NonBillableServices	Nicht verrechenbare Leistungen	OrganisationalWork

IntangibleProducts	Angebot Dienstleistungsprodukt	BusinessService	Leistungen	BehaviourElement BusinessFunctions
		NonBillableServices	Nicht verrechenbare Leistungen	InternalInformation&Meetings
		NonBillableServices	Nicht verrechenbare Leistungen	ContractingBody
		NonBillableServices	Nicht verrechenbare Leistungen	Patrons
		NonBillableServices	Nicht verrechenbare Leistungen	QualityManagement
		NonBillableServices	Nicht verrechenbare Leistungen	BoardMeeting
		NonBillableServices	Nicht verrechenbare Leistungen	Retraite(BoardOnly)
		NonBillableServices	Nicht verrechenbare Leistungen	Membership&DonationsSupport
		NonBillableServices	Nicht verrechenbare Leistungen	Reading
		NonBillableServices	Nicht verrechenbare Leistungen	DataAnalysis
		NonBillableServices	Nicht verrechenbare Leistungen	Emergency&Solifond
		NonBillableServices	Nicht verrechenbare Leistungen	LivingHIV
		NonBillableServices	Nicht verrechenbare Leistungen	LHIVE
		NonBillableServices	Nicht verrechenbare Leistungen	Logistik
		NonBillableServices	Nicht verrechenbare Leistungen	Administration&Statistics
		NonBillableServices	Nicht verrechenbare Leistungen	InventoryManagement

Table 82: Overview on AHSGA's ITRS Records

12.6 *Symfact Ancillary Information*

Within the DokLife project values have been defined for data used in the MeGaWorkbench prototype as detailed in Table 83: Ancillary Information for Symfact.

DokLife Entity	Values	Use in MeGaWorkbench
Trigger	Event, Time, Repeating	not used in
Obligation Type	C&C, Compliance, Report, Finance, HR, Legal, Operation	
Contract Type	License, NDA, CDA, Maintenance, Outorsourcing, Support	
Obligation Condition	ForeMajeure FinancialBusinessEvent	eo:obligationHasCondition ?condition

Table 83: Ancillary Information for Symfact