UNIVERSITÀ DEGLI STUDI DI ROMA
LA SAPIENZA

DOCTORAL THESIS

A deep X-ray view of Stripe-82: improving the data legacy in the search for new Blazars

---

Author:
Carlos Henrique
BRANDT

Supervisor:
Paolo GIOMMI

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

ICRANet, Physics department

July 9, 2018

Declaration of Authorship

I, Carlos Henrique BRANDT, declare that this thesis titled, "A deep X-ray view of Stripe-82:
improving the data legacy
in the search for new Blazars" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

_____

Date:

_____

"La calma è la virtù dei forti."

– Barista romano

UNIVERSITÀ DEGLI STUDI DI ROMA LA SAPIENZA

Abstract

Faculty Name
Physics department

Doctor of Philosophy

**A deep X-ray view of Stripe-82:
improving the data legacy
in the search for new Blazars**

by Carlos Henrique BRANDT

In the era of big data, multi-messenger astrophysics and abundant computational resources, strategic uses of the available resources are key to address current data analysis demands. In this work, we developed a novel technological approach to a fully automated data processing pipeline for Swift-XRT observations, where all images ever observed by the satellite are downloaded and combined to provide the deepest view of the Swift x-ray sky; Sources are automatically identified and their fluxes are measured in four different bands. The pipeline runs autonomously, implementing a truly portable model, finally uploading the results to a central VO-compliant server to build a *science-ready*, *continuously-updated* photometric catalog. We applied the Swift-DeepSky pipeline to the whole Stripe-82 region of the sky to build the deepest X-ray sources catalog to the region; down to $\approx 2 \times 10^{-16}\ erg\ s^{-1} cm^{-2}$ (0.2-10 keV). Such catalog was used to the identification of Blazar candidates detected only after the DeepSky pipeline.

Acknowledgements

First of all, I would like to thank my country, Brazil, the Brazilian society, and our funding agency CAPES for the opportunity and support. I also thank the University of Rome La Sapienza and ICRANet for such effort.

I thank Paolo Giommi for the uncountable lessons given and trust placed in me. I'm talking about a distinguished scientist that will sit next to you if necessary to explain fundamental physics as well as high performance computing; or have a discussion about next-level science and how we should help our society to become a better place; or food, as every good italian. I have always been lucky to have great supervisors, but Paolo became also a *mentor*, an example to follow. In such atmosphere I'd like to thank Yu-Ling Chang and Bruno Arsioli, Paolo' pupils that are also, not only great researchers but, amazing people.

I thank a particular group of researchers at CBPF, Brazil, Ulisses Barres, the brothers Marcio and Marcelo Portes and Martin Makler, for their current or earlier support in my career and example of passionate and responsible public servants. This is the group I hope to help build the future of science and technology of my country.

I would like also to thank Remo Ruffini for the opportunity of this PhD and in providing the interface to top-level discussions. At ICRANet I had the opportunity to meet scientists from all continents, which has been an incredible experience. A big thanks to the staff of the institute, in particular the girls from secretariat and the IT guys, always available to help and look after the students.

Friends. I have done great friends in Italy, people who taught me about people; a lot. In particular, I want to thank Ingrid Meika, Aurora Ianni, Ruben Abuhazi, Riccardo Dornio, Martina Pentimali and Salameh Shatnawi.

A special, deep, beloved thanks goes to Susan Higashi, who supported and cared about me with enormous patience and love; much more than I can perceive nowadays, I know, has been taught.

Finally, Family. There are no words to express my gratitude to my parents and siblings. They are everything that matters.

Much have been learned.

Contents

List of Figures

List of Tables

Aos meus avós – Oma, Opa, Vó e Vô.

Chapter 1

Introduction

Astrophysics lives its golden era of data, with multiple ground- and space-based instruments surveying the sky in all different wavelenghts – from gamma-rays, to X-rays, ultraviolet, optical, infrared, to radio band – as well as observatories for astro-particles and gravitational waves brought astrophysics to a position of extreme wealth which we now need to work it out to extract the the best of knowledge from it to feed it back to the cycle of development.

To handle such plurality in data sets, computational infrastructure is a fundamental component of this discussion. Means to store, process and share data efficiently are critical to the mission of extracting and allowing others to analyze such data. Once data is being handled efficiently, the analysis of data through statistical and physical modeling is to be done with much more efficiency also.

Some efforts have been implemented in the last years to address the issue, the most prominent being the Virtual Observatories (VO), by the International Virtual Observatory Alliance (IVOA[1]). Most recently, the Open Universe initiative (OUN[2]) brought the discussion to the United Nations arguing that it is not only of the interest of the astronomical community but of the whole society to address this issue: on top of the VO achievements and individual solutions we, as a united society, organize a common agenda to effectively put in practice the well succeeded results and learn from each other – academia and industry – the failures.

The argument in place is about high-level data (also called *science-ready* data): easily accessible data, ready to be used for modelling and empirical inference. Data's ultimate goal is to be used in its full extent, which in practice it means to be frequently handled from various different perspectives. There are two major components under the *high-level data* discussion: the data itself – how comprehensive it is – and the software to handle it.

[1] http://www.ivoa.net/
[2] http://www.openuniverse.asi.it/

To tackle the problem with a specific science case, we developed a pipeline to detect and measure x-ray sources from the *Neil Gehrels Swift Observatory* (former *Swift Observatory*, hereafter *Swift*) XRT instrument. The pipeline, called *Swift DeepSky*, handles all the processing from data download, detection and measurements, to data publication in the Virtual Observatory network if required by the user. After a basic set of input parameters, the pipeline delivers to the user a table of measurements ready for scientific use.

As a science case, we applied the *Swift DeepSky* to the Stripe82 region, a prominent multi-wavelength field of the sky, in the search for blazars. Besides the pipeline, a set of software tools have been developed to access, correlate and publish data where we applied technical concepts we believe improve the everyday work we handle.

The work here presented has application in the Brazilian Science Data Center (BSDC), an infrastructure project started during this doctorate to develop the elements of high-level data. Together with the Open Universe initiative, we are bringing the discussion here in place to the Brazilian society so that, in the near future, Brazil can not only become an important node in the astronomical data science network, but make this development a product for the society as a whole, beyond the academical walls.

The next sections of this introduction will present is some greater details the the elements this work builds upon and applies to: the scientific data access discussion and the Stripe82 data collection. Then, in the chapter 'Swift DeepSky' we describe the pipeline implemented and the production of the *SDS82* catalog. The chapter 'Brazilian Science Data Center' presents the infrastructure design and its implementation so far. Finally, in the conclusions, we summarise the work done.

1.1 Stripe 82

The Stripe82 is a 275 deg^2 stripe along the celestial equator, $-50 < RA < 60$, $-1.25 < Dec < 1.25$ ($\approx$ 1% of the sky), imaged more than 100 times as part of the Sloan Digital Sky Survey (SDSS) Supernovae Legacy survey (Annis et al., 2011; Abazajian and Survey, 2008). The name comes from SDSS' sky survey plan, which cover the sky along *stripes* and that one is the *82* in the schema.

Repeated visits to the region created a collection in the archives of SDSS appealing to deep extragalactic studies, but by co-adding the Stripe82 images Annis et al. (2011) and Jiang et al. (2014) assembled a data set ~ 2 *mag* deeper

than the usual single pass data ($mag_r \sim 22.4$) and with a median seeing of $\approx 1.1''$, providing a unique field for the study of sensible objects like faint and distant quasars (Mao and Zhang, 2016).

The coverage strategy applied to the Stripe82 – considering also its particular position in the Sky, being visible by telescopes in North and South hemispheres – made it particularly interesting for other observatories to follow-up. Stripe82 is emerging as the first of a new generation of $\Omega \sim 100\, deg^2$ deep extragalactic survey fields, with an impressive array of multi-wavelength observations already available or in progress. The wealth of multi-wavelength data in Stripe82 is unparalleled among extragalactic fields of comparable size. Besides dedicated surveys covering the Stripe, large area surveys cover the area in considerable extent: *GALEX* All-Sky Survey and Medium Imaging Survey (Martin et al., 2005) covered much of the field in the ultra-violet; UKIDSS (Lawrence et al., 2007) has targeted the Stripe as part of its Large Area Survey; parts of the Stripe82 are also in the footprint of FIRST, in the radio band.

At radio wavelengths, Hodge et al., 2011 provided a sources catalog containing $\sim 18K$ at $1.4GHz$ radio sources from observations of the *Very Large Array* with 1.8" spatial resolution. The catalog covers $92deg^2$ of the Stripe82 down to $52\mu Jy/beam$, three times deeper than the previous, well known FIRST catalog (Becker, White, and Helfand, 1995) also covering the region.

Timlin et al., 2016 conducted deep infrared observations with the *Spitzer Space Telescope* in the *Spitzer IRAC Equatorial Survey* (SpIES), covering around one-third of the Stripe ($\sim 115deg^2$) in 3.6 and $4.5\mu m$ bands, between $-0.85 < Dec < 0.85$, $-30 < RA < 35$ and $13.9 < RA < 27.2$, down to AB magnitudes $m_{3.6} = 21.9$ and $m_{4.5} = 22$. SpIES provided 3 catalogs, for each combination of Spitzer-IRAC Channel-1 ($3.6\mu m$) and Channel-2 ($4.5\mu m$), each one providing more than ~ 6 million sources with spatial resolution better than (FWHM) $\sim 2''$.

In the far-infrared, Viero et al., 2013 observed the region with the *Herschel Space Observatory* SPIRE instrument to cover $\sim 80deg^2$ of the Stripe82 in the *Herschel Stripe82 Survey* (HerS). Observations go down to an average depth of 13, 12.9, 14.8 $mJy/beam$ at 250, 350, $500\mu m$, respectively. The band-merged catalog provided by HerS contains ~ 33000 sources.

High-resolution observations and deep sources catalog in optical was provided by the CS82 collaboration (Soo et al., 2018; Charbonnier et al., 2017) using the *Canada-France-Hawaii Telescope* covering $\sim 170deg^2$ of the Stripe down to magnitude $m_i = 24.1$ in the i-band with a 0.6" median seeing.

LaMassa et al., 2015 (see also LaMassa et al., 2012; LaMassa et al., 2013) combined *Chandra* archival data and new *XMM-Newton* observations to compiled three source catalogs in X-ray ($0.5 < E(keV) < 10$), covering $\sim 31 deg^2$, providing energy flux to a total of 6181 sources down to $\sim 10^{-15}\, erg\, s^{-1}\, cm^{-2}$.

Besides dedicated works to cover the Stripe82, a number of all-sky surveys of mapped either partially or entirely the region. The Dark Energy Survey (DES) covered $200\ deg^2$ down to magnitude $m_r \approx 24.1$ (()Abbott2018); GALEX (Martin et al., 2005) observed $200\ deg^2$ down to magnitude $m_{NUV} \approx 23$; UKIDSS (Lawrence et al., 2007) covered the region in the infrared down to magnitude $m_J \approx 20.5$; While (**Lang2014**) reprocessed WISE data to provide higher resolution coadded images. Figure 1.1 present the depths (in $erg\ s^{-1} cm^{-2}$) of the main surveys present to date covering the Stripe82 region. The figure takes Timlin et al., 2016 (Table 2) as good summary of such list and considers also serendipitous reference the author crossed through.

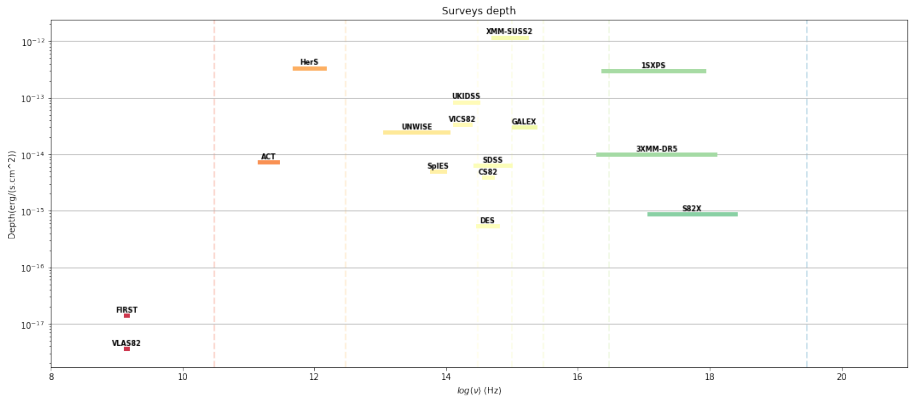


FIGURE 1.1: Compilation of current surveys covering the Stripe82, depth (in $erg\ s^{-1} cm^{-2}$) and observed waelengths

Multi-wavelength surveys are key to the study of Active Galactic Nuclei (AGN), as AGNs emit energy throughout the whole electromagnetic spectrum. In particular to our interest, deep, multi-wavelength, wide-area surveys are key to the study of *blazars*.

To improve the Stripe82 data collections at the high energy range we created a deep x-ray catalog from the *Neil Gehrels Swift Observatory* observations, with data from the XRT instrument – $0.2 < E(keV) < 12$. We designed a pipeline to combine all observations ever done by the Swift-XRT instrument (hereafter, Swift) to have the deepest catalog possible from Swift. We visited every $12'$ Swift field and combined all observations overlapping in that field, detected and measured the flux in three independent wavebands: *soft*, *medium*, *hard* x-ray bands, besides the *full* band covering the entire energy range.

1.2 Blazars

The current paradigm for Active Galactic Nuclei (AGN) considers a central engine – a supermassive black hole ($\gtrsim 10^6 M_\odot$) – being feed by a surrounding accretion disk. Matter in the accretion disk comes from a thick torus, which is the structure delimiting the AGN food supply. Surrounding this central engine there are clouds of dust more-or-less distant from the black hole which are perceived by spectral emission lines more-or-less broad. The clouds are would be trapped by the strong gravitational field and heated by the radiation coming out from the accretion disk while the matter (in the disk) is accelerated towards the black hole. In some objects we may observe a strong relativistic jet coming out from the central engine, roughly perpendicular to the disk; Such objects are called Radio-loud AGNs (**Urry1995**). Figure 1.2 offers a schematic view of such model.

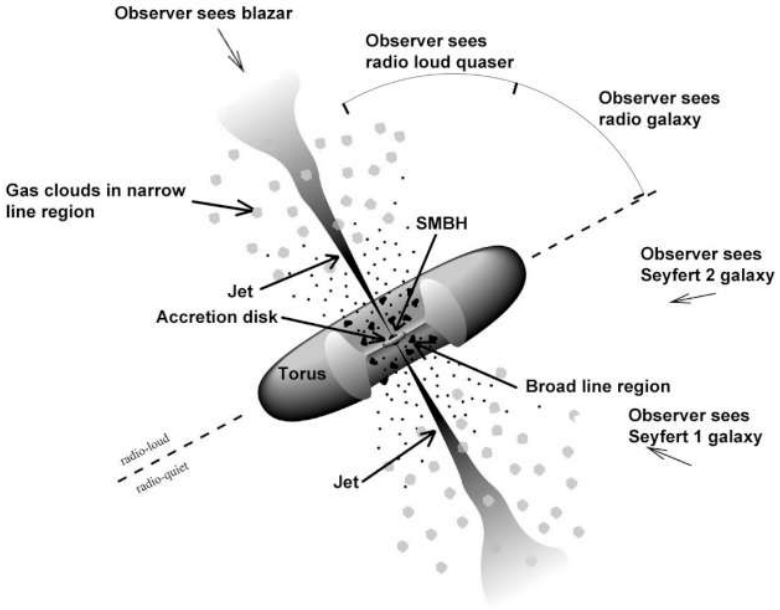


FIGURE 1.2: Depending on the orientation between the observer and the AGN/galaxy different properties are observed. The Unified Model for AGNs states that different "types" of AGNs are effectively the result of which part of the system we are being able to observe. The figure presents a simple and clear view of observer's point of view and type of object detected. Credit: GSFC/NASA Fermi collaboration website (`https://fermi.gsfc.nasa.gov/science/eteu/agn/`)

Depending on the orientation relative to us of such system we will be able to observe different properties. For instance, if we are looking such system from the side (as in figure 1.2), the very central part and the accretion disk are going to be hidden by the torus, we should be able to see narrow emission

lines coming from the most distant clouds as well as both lobes (North/-South) from the relativistic jet in the Radio band.

In this model, blazars represent the fraction of AGNs with their jets aligned towards us – more precisely: at relatively small ($\lesssim 20 - 30°$) angles to the line of sight. This produces strong amplification of the continuum emission ("relativistic beaming") when viewed face-on. Radio-loud galaxies constitute a small fraction ($\sim$ 10%) of the current known AGN population, around half of them are blazars (Padovani, 2017), making blazars a rare class of objects in our sky.

The blazar class includes flat-spectrum radio quasars (FSRQ) and BL Lacertae objects (BL Lac). The main difference between the two classes lies in their optical spectrum features: FSRQ presenting strong, broad emission lines supported by a non-thermal continuum, while BL Lacs present weak to no features at all, only the non-thermal continuum.

Blazars are characterized by emission of non-thermal radiation along a large spectral range, from radio to γ-rays, and possibly to ultra high energies (Padovani et al., 2016). The overall Spectral Energy Distribution (SED, $log(\nu f_{(\nu)})$ vs. $log(\nu)$ plane) of blazars is described by two humps, a low-energy and a high-energy one. The peak of the low-energy hump can occur at widely different frequencies (ν^S_{peak}), ranging from about $\sim 10^{12.5} Hz$ (Infrared) to $\sim 10^{18.5} Hz$ (X-ray). The high-energy hump has a peak energy frequency somewhere at $\sim 10^{20} Hz$ to $\sim 10^{26} Hz$. Depending on the frequency of the low-energy hump BL Lacs can be divided in Low Synchrotron Peak (LSP or LBL) sources ($\nu^S_{peak} < 10^{14} Hz$), Intermediate energy peak (ISP or IBL) sources ($10^{14} \nu^S_{peak} < 10^{15} Hz$), or High energy peak (HSP or HBL) sources ($\nu^S_{peak} > 10^{15} Hz$). Figure 1.3 (Padovani, 2017) present an schematic view of AGN – jetted and non-jetted – profiles. In the present picture, the first bump is associated with non-thermal emission from the Synchroton radiation, originated from the jet's relativistic charged particles moving in the magnetic field, while the second bump is understood as the result of low energy photons that are Inverse Compton scattered to higher energies by the beam of relativistic particles.

Typically with continuous bolometric luminosity $10^{45} \sim 10^{49} erg/s$, blazars are also known for displaying strong variability, with $\nu f_{(\nu)}$ variability up to $10^3 - 10^4$ times in timescales of weeks-to-days ($0.01 - 0.001 pc$). In figures 1.4 and 1.5 present two examples of a LBL – the 3C279 blazar (Webb et al., 1990) – and a HBL – the Mrk421 blazar (Barres de Almeida et al., 2017) –, where we can visualy extrapolate the datapoints to see the low- and high-energy

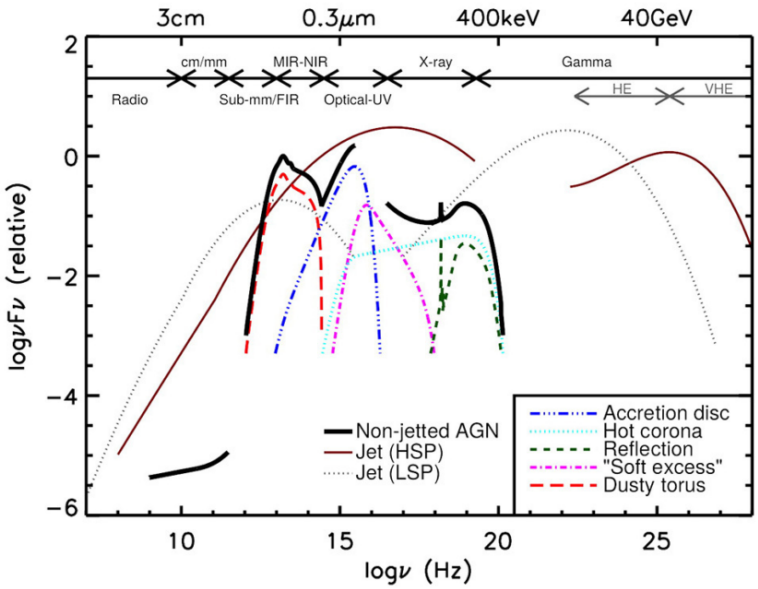


FIGURE 1.3: Spectral Energy Distribution continuum components for jetted and non-jetted AGNs. Credit: Padovani, 2017.

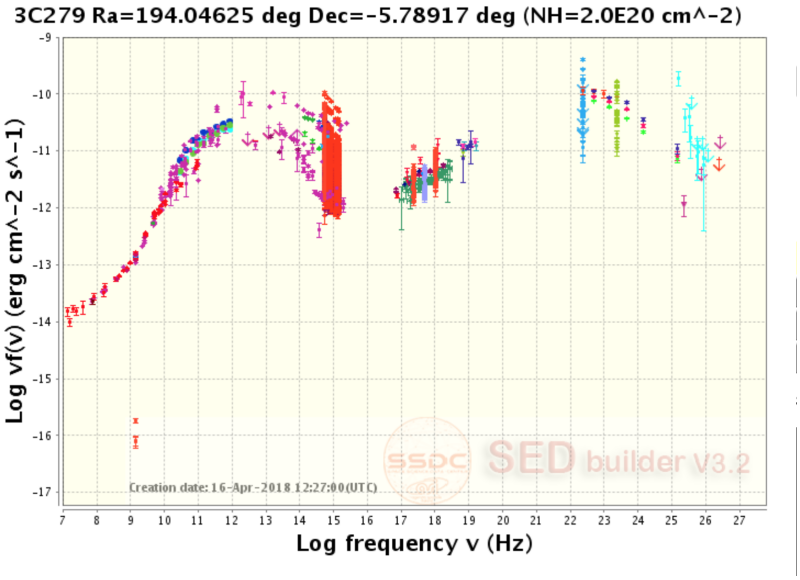


FIGURE 1.4: Spectral Energy Distribution of 3C279, a LSP blazar

humps as well as their variability (figures created with ASDC SED tool[3]).

1.3 Data accessibility

Astrophysicists have been extremely successful during the last decades in designing observational facilities, collecting a wide range of signals with ever

[3]https://tools.asdc.asi.it/SED/

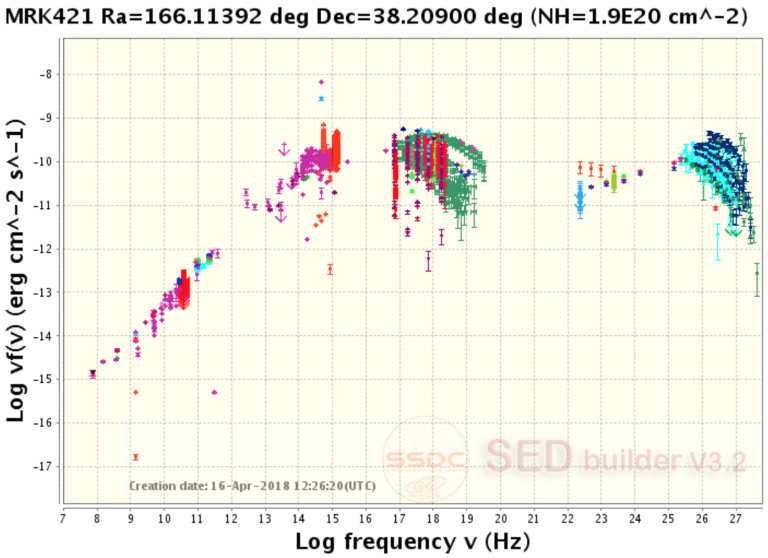


FIGURE 1.5: Spectral Energy Distribution of Markarian 421, a HSP blazar

more sensitivity, which boosted the number of publications – and ultimately, the knowledge – of the community as a whole. The amount and rate of data accumulation is, naturally, at its peak as we keep pushing technology speed and quality on data acquisition, soon projects like the LSST Robertson et al., 2017 will cross the rate mark of 10 TB (terabytes) per night. Plus, besides fresh new data arriving every day to active missions' databases, a wealthy collection of past missions and data analysis results stored in institutes and universities across the globe, such data provides a unique temporal record of astrophysical events along the recent history of physics research.

Astronomical datasets are not only getting bigger, but also more complex. *Multi-messenger* astronomy is an emerging field of data analysis to relate astronomical events handled to us by different physical *messengers*: photons, gravitational waves, neutrinos, cosmic rays.

In such scenario, data accessibility becomes a crucial discussion (Wilkinson et al., 2016). The main questions in place are (*i*) how to organize such diverse data collection in a meaningful way, (*ii*) how to publish the datasets clearly regarding their content, (*iii*) how to consume (*i.e*, query, analyze) such data. Ultimately, as scientists, we want to address the task: *"query a distributed petabyte-scale heterogeneous data base, build a sample of sensible data, extract information"*

As highlighted by Wilkinson et al. (2016), the data access discussion is not particular to astronomy, but most of science fields. And a good part of the issue is obviously because datasets are becoming bigger and bigger, but

mainly because we are moving to a new kind of publication media: from paper-format articles to digital documents. And this "simple" shift demands a whole new schema to make scientific resources available.

The astronomical community has organized itself to address the task a decade ago with the establishment of the International Virtual Observatory Alliance (IVOA[4]), and most recently the Open Universe initiative[5] is proposing to the United Nations an extension of this discussion to the benefit of not only the astronomical community but to the broader society.

1.3.1 The Brazilian Science Data Center

Aligned to the concepts about data accessibility, the *Brazilian Science Data Center* (BSDC) is a project started during this work with the ultimate goal to become a *de facto* interface – practical, useful, accessible interface – for the Brazilian community (not only, but primarily) to *science-ready* data. It is a project under construction at the Brazilian Center for Physics Research (CBPF) which builds upon the experience and is being developed in close collaboration with ASDC, the science data center of the Italian Space Agency (ASI), where the concept of a science-ready data center was originally advanced. The BSDC shares on the ethos of and supports the Open Universe initiative and in this context is also supported by the Brazilian Space Agency (AEB).

The *Open Universe*[6] is a recent initiative aimed at greatly expanding the availability and accessibility to space science data, extending the potential of scientific discovery to new participants in all parts of the world. The initiative was proposed by Italy and presented to the United Nations Committee on the Peaceful Uses of Outer Space (COPUOS) in June 2016. Open Universe guideline is to promote and create means to guarantee high-level space science data to be public and available by simple, transparent means so that a large part of the worldwide society can directly benefit from it. Not only scientist to be benefit, but also the general public through the educational of outer space data and data analysis tools. The ultimate goal is to boost the overall knowledge of society through all different uses of the data by promoting the inclusion of different groups, other than researchers only, in the appealing discussion of the observation of the Universe.

[4] http://www.ivoa.net/
[5] www.unoosa.org/oosa/en/oosadoc/data/documents/2016/aac.1052016crp/aac.1052016crp.6_0.html
[6] http://www.openuniverse.asi.it/

As the Open Universe initiative (OUN) is proposing the discussion at a global level, trying to address the different cultural and economical instances of society, the BSDC is going to address analogous contrast in social parameters and continent-scale distances. Which brings BSDC and OUN to a very particular partnership where BSDC should provide practical, field results for theoretical discussions promoted by OUN.

The availability, accessibility and quality of data have an impact on all these groups, harnessing their potential and enabling them to enhance their contribution to the global stock of common knowledge. The moment is particularly important as data storage, computing power and connectivity are broadly available to billions of people across all different scales, from supercomputers to smartphones. At the same time the scope of human cultural and intellectual exchanges has broadened, as the current age of big data sharing and open source gathers pace.

It is well acknowledge today in the scientific community the benefits in productivity and innovation driven by open data access. However there is a considerable unevenness in the interfaces provided by outer space data providers. In the next years, considering the exponencial grow in overall data archives, both from the increase in capabilities to analyse existent data and by the new generation of observatories to come, it will be fundamental to the health of science to consolidate, standardize and expand services, promoting a significant inspirational data-driven surge in training, education and discovery.

From ever since the beginning of the Open Universe proposal to the United Nations, Brazil has been represented in the discussions through the active participation of BSDC members (Carlos H Brandt and Ulisses Barres, Almeida et al., 2017) in the various meetings held to discuss the initiative To summarise the theory, or foundations of the initiative, follows the digest of those meetings.

From Committee on the Peaceful Uses of Outer Space (2016):

> The initiative intends to foster and spread the culture of space science and astronomy across different countries. It will pursue several interrelated tasks to the benefit of all actual and potential users of space science data, namely:
>
> 1. Promoting the robust provision and permanent preservation of science-ready data;
> 2. Advancing calibration quality and statistical integrity;

3. Fostering the development of new centralized services, both large and small, to exploit the interconnectedness of the modern Internet through new web-ready data;
4. Increasing web transparency to space science data;
5. Advocating the need for current and future projects to recognise the essential equality of hardware and software and incorporate centralized high-specification end-to-end analytics into cost envelopes;
6. Promoting active engagement of the Committee on the Peaceful Uses of Outer Space and other relevant national and international organizations towards tangible actions in this domain.

Efforts to standardize and promote high-level data and data access services have been carried out in the last decades, among them the International Virtual Observatory Alliance (IVOA), the International Planetary Data Alliance, the Planetary Data System of the National Aeronautics and Space Administration (NASA), the Virtual Solar Observatory and, with a focus on interdisciplinary standards, the Research Data Alliance. As a quite specific example – data file format – but also a very good one we can cite the FITS (Flexible Image Transport System) format which homogenized at great extend the exchange and archival of data between individuals.

Open Universe wants to push the philosophy and all the experience acquired by the community in past and current efforts towards higher standards in data services and transparency. Such effort is necessary in order to satisfy the needs of not only those target groups, but anyone interested in astronomy and space science. Space science data could be seen then not only as our magnificent participation in the Universe, but also a mean to promote education and creativity across the Internet to whoever feels like doing it. It is then understood the importance to involve not only the astronomical community but other stakeholders, so that efforts are not duplicated and experiences are shared to optimize both sides towards better data access.

In (Committee on the Peaceful Uses of Outer Space, 2018) it was expressed that the final calibrated data, together with complete ancillary data that characterize the observations, should be stored in online archives, following established standards, and that the data should be made available to the public, without the need of further data processing, after the required proprietary periods.

The Open Universe initiative, then, plains to engage with a wide user base, including the various target groups identified, ranging from the research community, higher and secondary education, citizen and amateur scientists, industry and other potential end users (Committee on the Peaceful Uses of Outer Space, 2017). The Brazilian Science Data Center, further presented in section 4, is preparing itself to provide access to data, high-level software and a platform for the Brazilian and International community, and must focus to outreach the Brazilian society.

Virtual Observatory

The Virtual Observatory (VO) is a network of astronomical data collections organised in so called *services*, distributed around the world. VO participants share the vision that astronomical datasets and other resources should work as a seamless whole. To that goal, the development of the various scientific and technological aspects to make VO possible is coordinated by the International Virtual Observatory Alliance (IVOA). IVOA is an open initiative for discussing and sharing VO ideas and technology, and body for promoting and publicizing the VO.

The Virtual Observatory program started in 2002 with the formation of IVOA with the National Virtual Observatory (from USA), the Astrophysical Virtual Observatory (from ESO) and the AstroGrid (from UK) as founding partners. The IVOA then grew to include other national projects and international organizations like the W3C (World-Wide Web Consortium) and the International Astronomical Union (IAU). From which (W3C) the working structure was adopted and (IAU) standard recommendations could get support (Hanisch et al., 2015).

VO can be seen as a collection of *resources*, defined as (Hanisch, IVOA, and NVO, 2007) "VO elements that can be described in therms of who curates or maintains it and which can be given a name and a unique identifier". In practice, the VO *resources* (or *data services*) are provided by data centers in a distributed configuration, worldwide. Each resource is published in a *registry*, which has the role to broadcast the resources exist and their metadata. Much like the World Wide Web hyperlinks structure, this schema allows resources to be discoverable and, then, reachable by astronomers (Demleitner et al., 2014a; Demleitner et al., 2014b).

The figure 1.6 (Arviset, Gaudet, and IVOA, 2010) depicts the objective user interface on one side, the plural data archives on the other side and the VO standards and services in between.

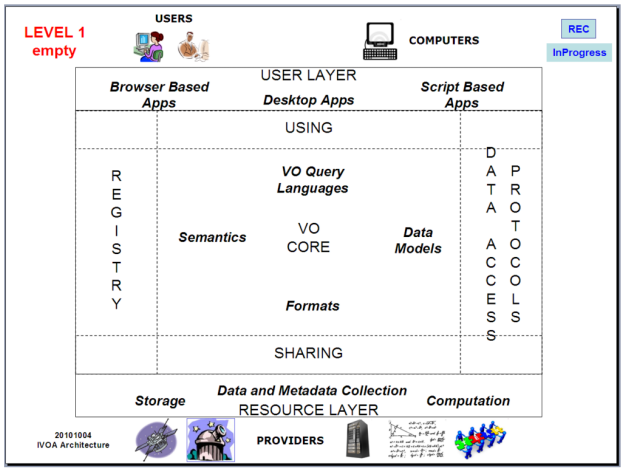


FIGURE 1.6: Virtual Observatory software resources harvesting infrastructure. Credit: (Arviset, Gaudet, and IVOA, 2010)

Key components of the VO infrastructure include the resources registry, the data access layer protocols and applications and application programming interfaces (API). On the data providers side there is a set of publishing registries that will communicate among them on sharing resources and identifiers. A registry is a database of resource records – *i.e,* data collections descriptions and services metadata – in the VO. Figure 1.7, (Demleitner et al., 2014a), represent the data discovery activity involving registries and client applications.

On the user side, the client applications implement a generic interface to search the publishing registries, there is no central or preferential server they clients know about, but all the standard protocols in between the parts that abstract location and internal implementations. Which is to say that any application following the IVOA standards will integrate seamlessly to the network, as well as any data service, will be able to promptly publish and communicate in the network.

The underlying infrastructure, registries, data centers and databases, is transparent to the user. Astronomers will typically interrogate multiple services when searching for a particular kind of content. This is made possible by standardization of data models and exchange methods.

The formal products that IVOA provides to the community are *standards* establishing the interface between providers (*e.g,* services providing spectral data) and consumers (users applications) and among services themselves (*e.g,* harvesting resources). One of the policies during the process of development and eventual adoption of a new standard is to have two reference

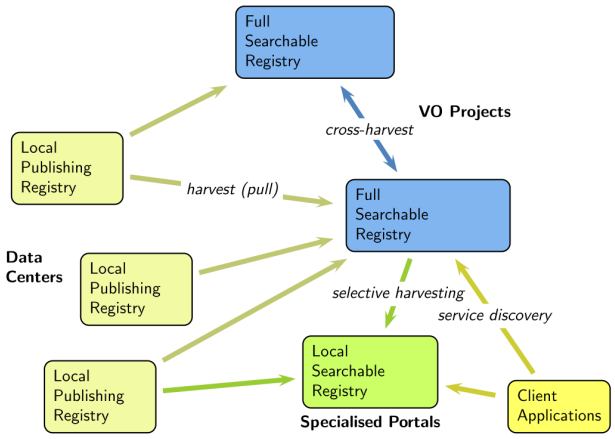


FIGURE 1.7: Virtual Observatory resources harvesting infrastructure. Credit: Demleitner et al., 2014a

implementations to it. The objective of such policy is to ensure that the standard can effectively work and that its application is useful to the research community.

The standards released or under discussion by IVOA are available at `http://www.ivoa.net/documents/index.html`. Each recommendation document goes through a cycle of discussions where description goes to the details to allow a better implementation and fulfill the compatibility space. It is not IVOA's mission to provide the software, only the standards. Nevertheless, the Alliance do publish the list of software developed by the community in the VO framework:
`http://www.ivoa.net/astronomers/applications.html`.

During this thesis we used the server-side data publication package developed by the German Astronomical Virtual Observatory (GAVO) DaCHS (Demleitner et al., 2014b). And, to the client-side, we implemented tools to handle VO data collections and services discovery.

Chapter 2

Swift DeepSky project

The amount of time a telescope observe a particular region of the sky dictates the amount of information that can be retrieved from that particular region. An astonishing demonstration of the power of extended integration time was given by the Hubble Space Telescope in 1995; the telescope observed a small region of the sky for 10 consecutive days generating more than 300 images, which were then combined into what was called the Hubble Deep Field, the deepest observation done to that date.

Objects with apparent luminosity too low to be significantly detected in a single exposure may show out when many observations are co-added. Clearly, co-addition is possible only when the telescope has visited a given region of the sky multiple times, which happens in three situations: (i) the observed region is signed to a series of dedicated time project, like the Hubble Deep Field, (ii) the region is part of a wide field survey footprint periodically visited, as in the Sloan Digital Sky Survey, or (iii) the telescope has long been collecting data that overlaps become a feature, which is the scenario we are exploring with the Swift Telescope.

Swift primary goal is to investigate Gamma-Ray Bursts (GRB), the telescope carries three detectors: the Burst Alert Telescope (BAT[1]), which triggers the whole telescope's attention whenever a GRB is detected; the X-Ray Telescope (XRT,[2]), that follows the subsequent emission of the GRB; finally the UltraViolet-Optical Telescope (UVOT[3]) responsible for registering the uv/optical GRB afterglow. Although its priority is GRB events, Swift will follow a schedule of observations of x-ray sources whenever GRBs are not on the sight.

Willing to create the deepest catalog of Swift X-ray data that could be dynamically updated whenever new observations arrived, we developed the

[1]https://swift.gsfc.nasa.gov/about_swift/bat_desc.html
[2]https://swift.gsfc.nasa.gov/about_swift/xrt_desc.html
[3]https://swift.gsfc.nasa.gov/about_swift/uvot_desc.html

Swift-DeepSky pipeline. The pipeline will combine *all* observations taken by Swift with its XRT instrument in Photon-Count mode since it started operating, in 2004. The *co-added* image will be used to detect X-ray sources in the field and then proceed with a series of flux measurements (`countrates` and $\nu F\nu$ fluxes).

For the first 7 years of Swift operation, from 2005 to 2011, D'Elia et al. (2013) provided the `1SWXRT` with positions and flux measurements for all point-like sources detected in XRT Photon-Count mode observations with exposure time longer than $500 seconds$. While Evans et al. (2013) published the `1SXPS` catalog with $\sim 150k$ sources from all observations made by Swift-XRT during its first 8 years of operation.

The *Swift-DeepSky* goes one step further by providing the *software* – open source, clearly – and a mechanism to have a central, VO-compliant catalog of the *DeepSky* measurements keeps an up-to-date version of itself, *forever* – until Swift finishes its lifetime. This work combines both methodologies – D'Elia et al.; Evans et al. – for XRT data reduction in a steady, scalable software solution where software design is a major component of this work results.

Rational

The pipeline is the implementation of a conceptual solution for high-level user interface for scientific pipelines. We bring in collection cutting edge technologies to provide secure, fully automated data analysis software to address also non-technical users, besides providing results producibility, which we believe to be an obstacle for science development. The goal of the *DeepSky* is to deliver a reliable and stable Swift-XRT data reduction tool at the same time that it keeps a living catalog of such data, that grows and updates itself at each use.

The pipeline combines all Swift XRT events observed in Photon Count mode in observations longer than 100 *seconds* in a region of $12'$ around a given position of the Sky. HEASoft[4] tools are used to combine multiple observations and extract physical information from the objects observed. The detection of source are done using the `Full` bandpass, 0.3-10 keV. We then measure the photon flux of each detected source using HEASoft `Sosta` in three intervals: `Soft` (0.3-1 keV), `Medium` (1-2 keV), `Hard` (2-10 keV). Whenever a source is not identified in one of the bands, an upper limit is estimated using the local background level and effective exposure time. Energy fluxes

[4] https://heasarc.nasa.gov/lheasoft/

(νF_ν) are corrected after our galaxy absorption, considering Milky-Way's hydrogen column (NH) in the line-of-sight to the source and a spectral slope considering instrumentation effects. At the end of the processing, the pipeline *may* upload the outputs to a central server where the results will be merged to the *Swift-DeepSky* primary sources catalog.

We applied the pipeline to the entire region covered by the Stripe82 and generated a unique deep x-ray sources catalog for the region. For the sake of clarity and to eventually motivate the reader in using *DeepSky* , the process of *efficiently* surveying a region of the sky, the steps of such processing are explained in details as well as supportive tools are equally made public.

In the next sections we describe the *DeepSky* pipeline work-flow in its technical aspects as well as the methodology on surveying a large area of the sky to generate a catalog of unique sources. To the technological aspect of the work, we will also present the software design and infra-structure adopted to the publication of the *DeepSky* pipeline.

2.1 The pipeline

DeepSky combines all Swift-XRT observations for a given region of the sky. The region is defined by a central coordinate and a radius. The pipeline combines all events and exposure-maps centered in the corresponding field and detect the objects using all events in the XRT energy range – 0.3-10 keV. For each object detected, the pipeline then do a series of flux measurements in three energy sub-ranges and Swift's *full* energy band:

Band	Energy range (keV)
Full	0.3 - 10 keV
Soft	0.3 - 1 keV
Medium	1 - 2 keV
Hard	2 - 10 keV

TABLE 2.1: *Swift-DeepSky* x-ray energy bands

2.1.1 Processing stages

Figure 2.1 presents the *DeepSky* pipeline workflow, it is composed by six conceptual blocks. The user provides a position on the Sky – Right Ascension and Declination –, optionally a radius defining the surrounding region of interest (default is $R = 12'$), and the pipeline will proceed through the following steps:

1. Search for Swift observations in that region
2. Combine (textiti.e, co-add) all observations
3. Detect objects using all events ($0.3 < E_{(keV)} < 10$)
4. Measure each object fluxes (in three x-ray bands)
5. Estimate spectral energy slope
6. Convert count rates to $\nu F\nu$ (νF_ν) flux

At the end of the processing the pipeline outputs the flux measurements in two sibling tables (*i.e,* same number of objects, respecting the same order) containing (*i*) photon fluxes and (*ii*) νF_ν energy fluxes. Provides also the x-ray events image with each detected object labeled and the corresponding exposure-map for visual inspection. As well as important for the sake of transparency, *all* temporary files used during the processing are kept in a separated folder.

Swift observations

The Swift Master Table[5] is the record table of all Swift observations, it contains informations like central coordinates (RA,Dec), start and end time-stamp, instrument used (BAT, XRT, UVOT), observation mode (PC, WT), unique observation identifier, for each observation done by the satellite. This table, maintained by the Swift data center, is where the pipeline starts.

The pipeline starts by querying the Master Table for observations done by the Swift-XRT telescope in Photon Counting[6] mode in a given region of the sky. The region is defined by the user through an (RA,Dec) position of the sky and a radius to consider: all observations with (central) coordinates within the the region will be evaluated. The pipeline offers a default radius value of $12'$ since this is Swift-XRT field-of-view, meaning that the position of interest for the user may have been covered by any observation within $12'$.

In practical terms, the pipeline starts with a pair of coordinates or object name of interest – in that case, *DeepSky* will ask CDS/SIMBAD[7] for the corresponding (RA ,DEC) position:

```
# swift_deepsky --object GRB151001A
```

[5]https://heasarc.gsfc.nasa.gov/W3Browse/swift/swiftmastr.html
[6]http://www.swift.ac.uk/analysis/xrt/modes.php
[7]http://simbad.u-strasbg.fr/simbad/

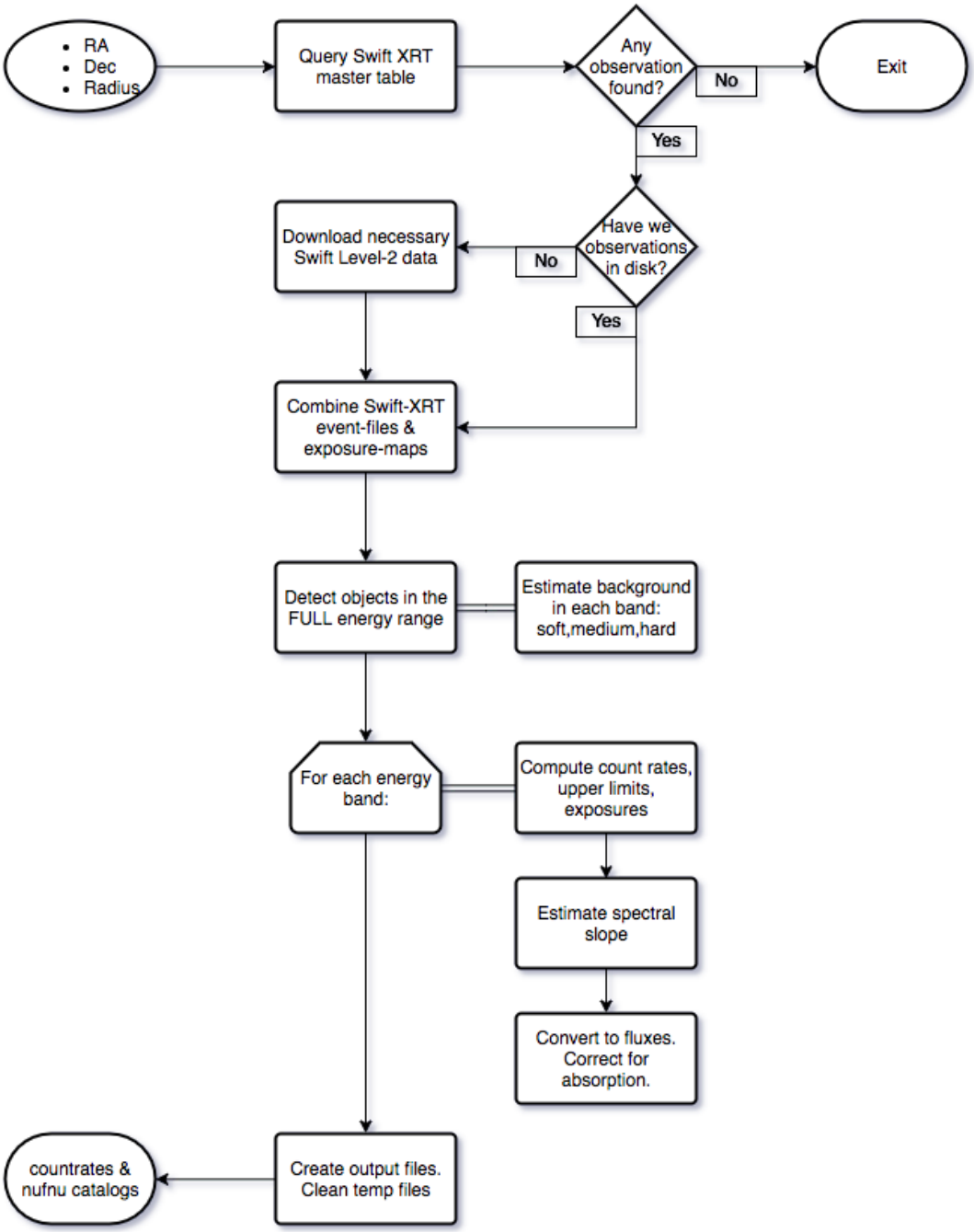


FIGURE 2.1: Swift DeepSky workflow

Once the list of observations potentially covering the given position is retrieved, the respective data is downloaded from the Italian Swift Archive[8]. The pipeline uses the archived level-2 data, in particular level-2 `PC` event-files (`OBSID/xrt/event/*pc*`) and exposure-maps (`OBSID/xrt/products/*pc_ex*`).

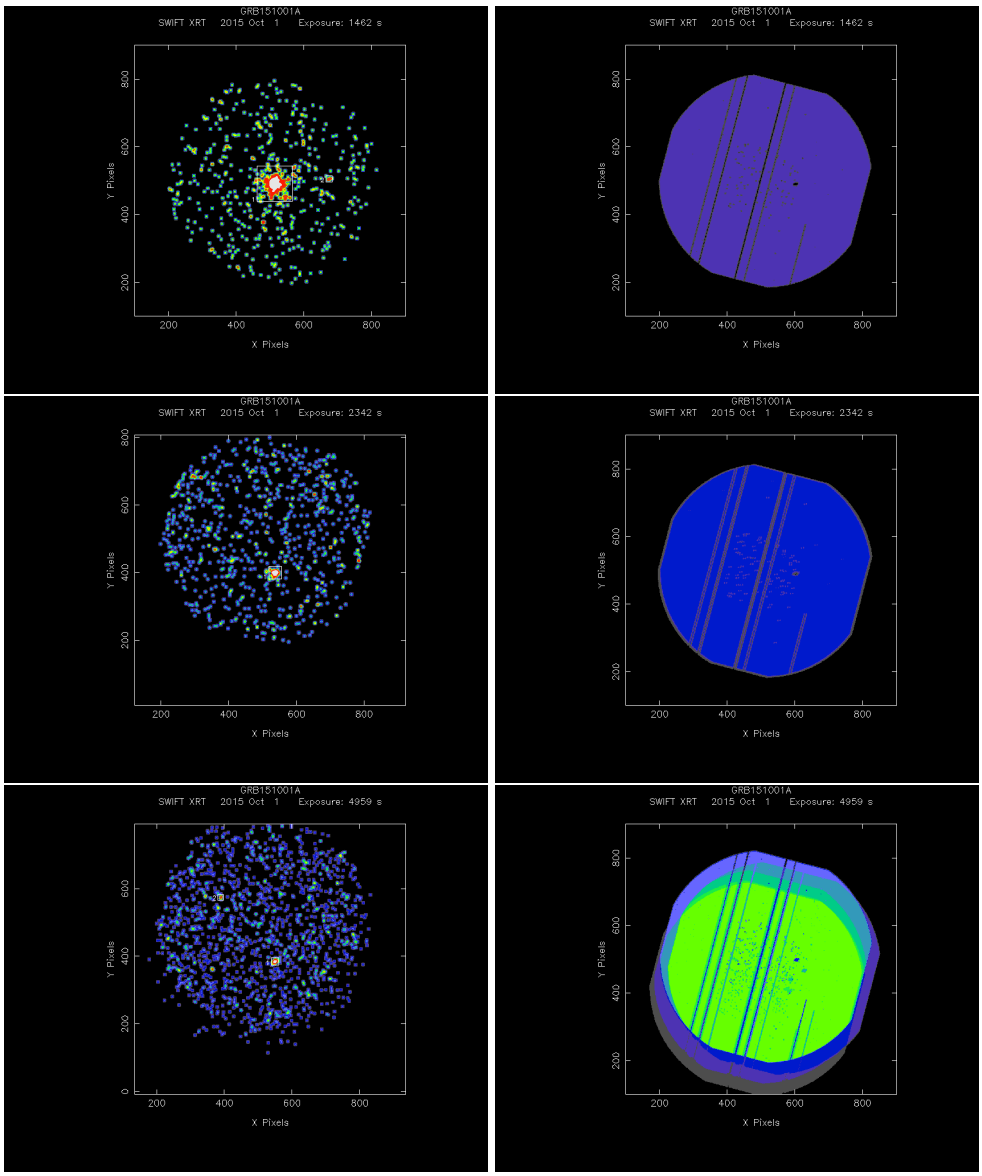


FIGURE 2.2: Swift XRT observation images example

Combining observations

With all event-files and exposure-maps in hands we combine them to build one unique events-file and exposures-map. We first use HEASoft Xselect[9] to extract all events in *good time intervals* from each observation and

[8] http://www.ssdc.asi.it/mmia/index.php?mission=swiftmastr
[9] https://heasarc.gsfc.nasa.gov/ftools/xselect/

write them all in one list of events. Analogously, HEASoft Ximage[10] is used to sum each observation's exposure-map.

At this point we have the two files – events list and exposures-map – that will be used for the rest of the pipeline.

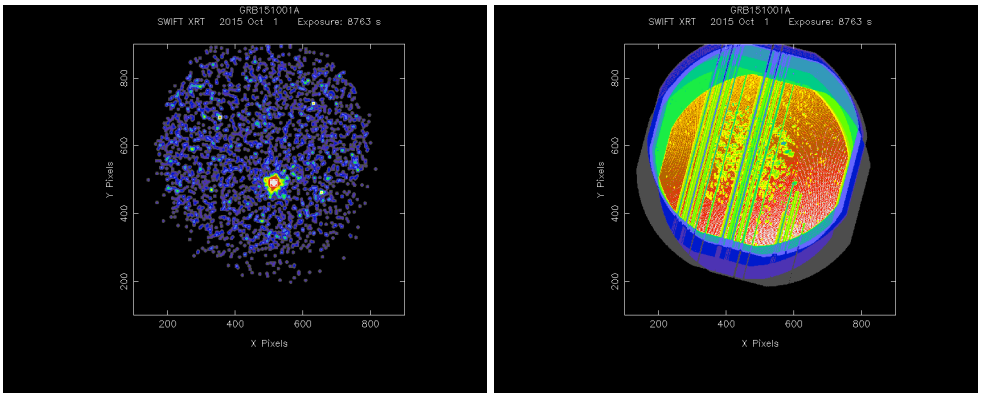


FIGURE 2.3: Swift XRT combined observations image example

Objects detection

Objects detection are done considering all events registered by Swift-XRT, in the `Full` (XRT) energy range – $0.3 - 10keV$. HEASoft XImage's `detect` routine is used for sources identification, background and countrates estimates.

The `detect` algorithm estimates the image background from a set of small boxes across the image. After a sliding-cell traversed the field looking for excesses, the objects are detected in a boxes of size such that its signal-to-noise ratio is optimized. When we initialize the events and exposure images in `detect` we define a an area of 800-x-800 pixels in the image, which represent $\sim 16'$ on the sky. The size was chosen arbitrarily to respect Swift XRT field of view ($12'$) and include extra field of the sky, but not still keep it at a minimum because of performance (the process of combining the images is computationally expensive and scaled with the image size).

In figure 2.4 we see the objects detected in our example case. We can compare this figure to those in figure panel 2.2 to see how the detections may change. These – from the combined image – are the detected sources we will carry for the next steps of the pipeline.

To illustrate how detection results are given by `detect`, it is presented below the photon flux estimates given by the routine:

[10] https://heasarc.gsfc.nasa.gov/xanadu/ximage/ximage.html

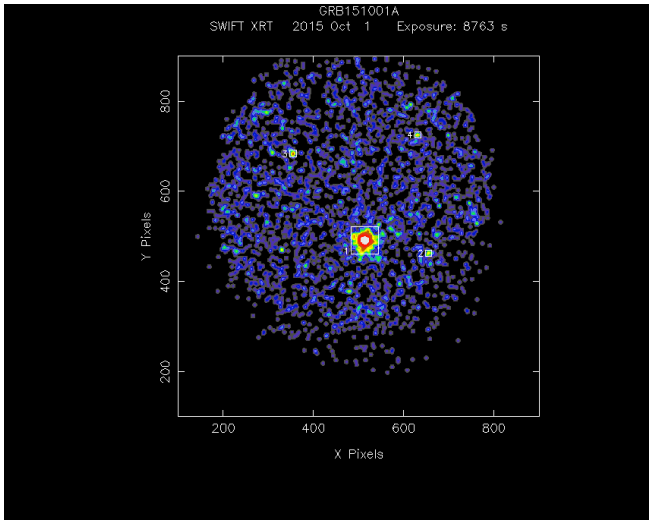


FIGURE 2.4: Swift XRT combined observations detection example

```
! Field Name     : GRB151001A
! Instrument     : SWIFT XRT
! No of sources  : 4
! Exposure (sec) : 8763.0135
! Input file     : GRB151001A_sum.evt
! Image zoom     : 1.0000
! Back/orig-pix/s: 8.6523602E-07
! Equinox        : 2000
! RA  Image Center: 233.73870
! Dec Image Center: 10.972480
! Start Time : 2015-10-01T15:05:48.00
! End Time   : 2015-10-02T02:07:39.00
!  #   count/s    err         pixel      Exp  RA(2000)   Dec(2000)   Err  H-Box
!                           x     y     corr                        rad  (sec)    prob    snr
    1 2.03E-01+/-5.2E-03 514.05896 491.67648 8374.13 15 34 55.118 +10 58 00.128  -1     72 0.000E+00 3.940E+01
    2 1.43E-03+/-5.3E-04 655.42859 462.85715 8755.90 15 34 32.489 +10 56 52.130  -1     15 5.746E-06 2.669E+00
    3 2.39E-03+/-7.3E-04 354.76923 683.92310 8089.25 15 35 20.626 +11 05 33.257  -1     18 1.445E-08 3.302E+00
    4 1.37E-03+/-5.7E-04 631.57141 725.14288 8758.83 15 34 36.296 +11 07 10.435  -1     15 7.914E-05 2.426E+00
```

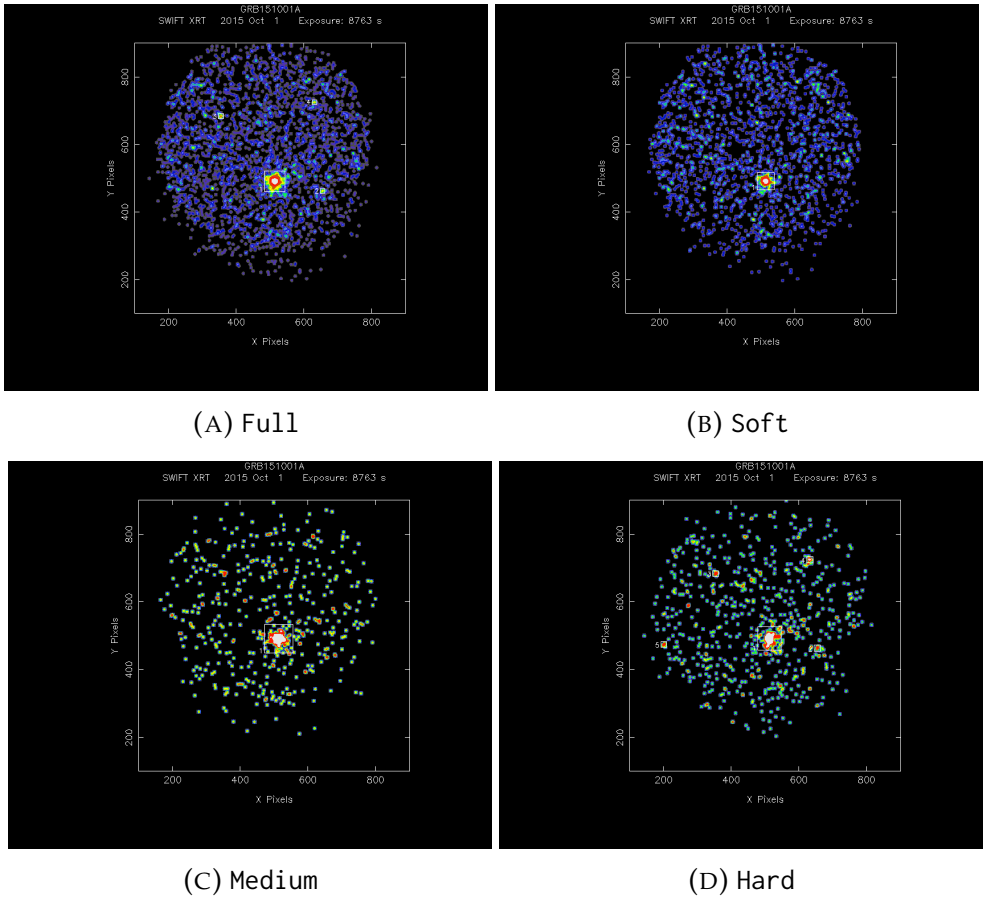
Flux measurement

After the detection of objects using all events collected, we then re-use `detect` considering only the events in the `Soft ,Medium ,Hard` energy range. We do so to estimate the background at each band, which will be used in the next step to estimate each object's signal (photon counts) using another algorithm.

The combined list of events appear at each x-ray band as shown in the figure panel 2.5

Photon flux measurement

Using the background measurements in each band and the detected objects positions we now will measure the corresponding photon fluxes using



(A) `Full` (B) `Soft`

(C) `Medium` (D) `Hard`

FIGURE 2.5: Swift XRT combined observations per-band example

XImage's `sosta` tool.

Sosta measures objects count rates weighted by the exposure map previously created, and the background previously computed by detect. It is worth to remember that objects –and their corresponding position – were detected using the full energy band, and background measurements in each band.

For each (object) position Sosta consider the events within a small region around it. The region is dynamically defined by the amount of *encircled energy fraction* (eef), which is a given parameter defined by the pipeline based on the `detect`-estimated intensity.

When the events accounted do not result in significant statistics to result in a flux measurement, an *upper limit* estimate is provided when sufficient photons can not be associated to the sources, the background (noise) level and the exposure time. The upper limit estimates, flux measurements and, together with the exposure times, all those values populate our output files. Table 2.2 presents what are the measured results for our example.

Column name	Source 1	Source 2	Source 3	Source 4
RA	15:34:55.118	15:34:32.489	15:35:20.626	15:34:36.296
DEC	+10:58:00.128	+10:56:52.130	+11:05:33.257	+11:07:10.435
countrates [0.3-10keV]	2.030E-01	1.430E-03	2.390E-03	1.370E-03
countrates error [0.3-10keV]	5.200E-03	5.300E-04	7.300E-04	5.700E-04
exposure-time(s)	8374.1	8755.9	8089.2	8758.8
countrates [0.3-1keV]	7.120E-02	4.086E-04	7.964E-04	4.565E-04
countrates error [0.3-1keV]	3.102E-03	2.863E-04	4.247E-04	-5.694E-01
upper limit [0.3-1keV]	-999	-999	-999	1.291E-03
countrates [1-2keV]	7.195E-02	4.086E-04	5.311E-04	3.044E-04
countrates error [1-2keV]	3.102E-03	2.863E-04	3.451E-04	-5.694E-01
upper limit [1-2keV]	-999	-999	-999	1.291E-03
countrates [2-10keV]	5.986E-02	6.129E-04	1.062E-03	1.370E-03
countrates error [2-10keV]	2.828E-03	3.472E-04	4.911E-04	5.700E-04
upper limit [2-10keV]	-999	-999	-999	-999

TABLE 2.2: *Swift-DeepSky* detected objects countrates

Energy flux measurements

At the last measurement stage the pipeline transforms photon flux to energy νF_ν flux, the integrated flux at a pre-defined effective frequency in each band. To compute the νF_ν flux we have to estimate spectral slope for each source across the energy band. The final energy fluxes are then corrected by our galaxy's absorption.

For each energy band, the effective frequency is defined as:

Band	Effective energy (keV)
`Full`	3 keV
`Soft`	0.5 keV
`Medium`	1.5 keV
`Hard`	4.5 keV

TABLE 2.3: *Swift-DeepSky* x-ray bands effective energy

The amount of dust absorbing x-ray light is computed using HEASoft `NH`tool, which estimates the density of Hydrogen atoms column along the line of sight for each source.

Based on the `NH`value, the count rates between `Soft + Medium` (combined) and `Hard` bands, and Swift-XRT instrument sensitivity in each band, the spectral slope is calculated. With the energy slope in hands, each source has its `countrates` measurement transformed to energy νF_ν flux in $erg/s/cm^2$. In table 2.4 we see the results of energy fulx computed for our running example.

Column name	Source 1	Source 2	Source 3	Source 4
RA	15:34:55.118	15:34:32.489	15:35:20.626	15:34:36.296
DEC	+10:58:00.128	+10:56:52.130	+11:05:33.257	+11:07:10.435
NH	3.11E+20	3.08E+20	3.19E+20	3.13E+20
energy-slope	0.734	0.8	0.8	0.8
energy-slope error	+0.04/-0.04	-999/-999	-999/-999	-999/-999
exposure-time	8374.1	8755.9	8089.2	8758.8
nufnu [3keV]	2.97984e-12	1.98813e-14	3.33525e-14	1.90786e-14
nufnu error [3keV]	7.63308e-14	7.36859e-15	1.01871e-14	7.93782e-15
nufnu [0.5keV]	1.75878e-12	1.02097e-14	2.00685e-14	1.14504e-14
nufnu error [0.5keV]	7.66256e-14	7.15378e-15	1.0702e-14	-1.42823e-11
upper limit [0.5keV]	-999	-999	-999	3.23822e-14
nufnu [1.5keV]	2.57653e-12	1.45474e-14	1.89332e-14	1.08439e-14
nufnu error [1.5keV]	1.11083e-13	1.01931e-14	1.23025e-14	-2.02843e-11
upper limit [1.5keV]	-999	-999	-999	4.59906e-14
nufnu [4.5keV]	3.28326e-12	3.30378e-14	5.72567e-14	7.38553e-14
nufnu error [4.5keV]	1.55113e-13	1.87155e-14	2.64772e-14	3.07281e-14
upper limit [4.5keV]	-999	-999	-999	-999

TABLE 2.4: *Swift-DeepSky* detected objects fluxes

2.1.2 Results

The pipeline primary output are the *countrates* and νF_ν flux tables, like tables 2.2 and 2.4 previously exemplified (except that here, in this document, they are transposed to properly fit the page width). The fluxes, respective error and upper limits are provided for the all bands – `Soft` , `Medium` , `Hard` – as well as the `Full` band. Total exposure time, spectral energy slope and hydrogen column density (`NH`) are also included.

Besides the flux tables, the combined exposure-map and events-file in FITS format – suitable for further analysis – and their `.gif` versions, as well as the *log* file with each processing steps information. All provided for the sake of complete transparency easy access and visual inspection whenever required. Are also provided in the output directory, properly encapsulated in a (tarball) file "`tmp.tgz`", *all* the intermediate files used during the processing. Which allows the user to control every step of the pipeline and eventually double check the results.

2.2 Creating a living catalog

In accordance to our *science-ready* and *high-level services* guidelines we have implemented an `upload` option to the *DeepSky* pipeline to allow results to be shared as soon as they are produced, transparently, to a worldwide

audience. The idea to have the public, VO-compliant catalog always up-do-date, the updates made by the users themselves on demand. Eventually, the whole sky Swift DeepSky database will be available, again, up-to-date by the users worldwide.

Figure 2.6 depicts the flow of results, from the user that generated the last results, to our servers, back to the users through a VO service and online table.

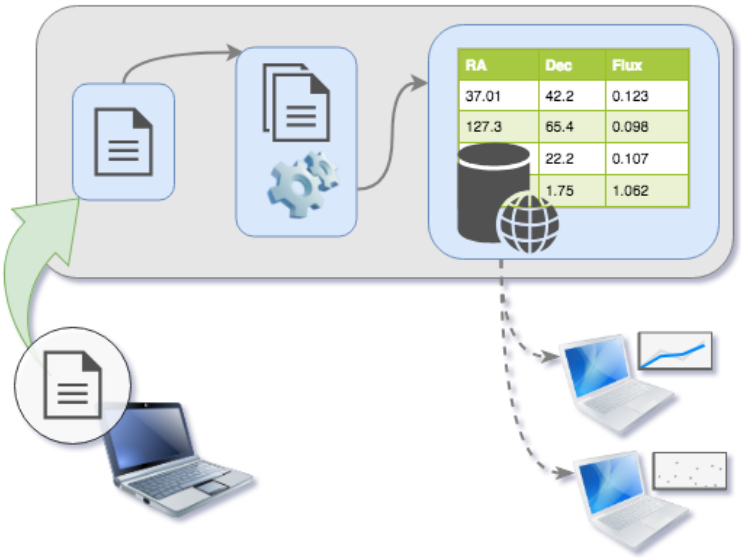


FIGURE 2.6: Workflow of the *DeepSky* living catalog: an user runs the pipeline to a specific region of their interest, results are transparently uploaded to a dedicated server where they are combined with existing table. The VO-compliant catalog then exposes the updated catalog, publicly available to other users

The upload of the pipeline output is done anonymously and explicitly under the user request (through the explicit use of `--upload` argument).

The output files are compressed and securely uploaded to a dedicated server. When the output files arrive to the server, they verified against the already existing database of previous results to verify for duplicates. If a particular source is already in the published results to parameters are compared between them: total exposure time and full-band flux measured; If the corresponding values are greater in the recently arrived data, then the database is updated to include (and substitute) the new measurements, otherwise the left unchanged. If the new data is not a duplication, it is simply included in the database.

2.3 Surveying the Stripe82

In this section we describe the process of covering a large contiguous region of the sky with discrete steps in the RA,Dec plane. We will do so for the Stripe82 field, used for the creation of our SwiftDS-82 catalog. We will go through the following steps:

- define the coordinates to visit and coverage area (*pointings*);
- define an efficient way to process all pointings;
- aggregate the results to compile a unique catalog.

Mapping the sky with HEALpix

The region coverage task is basically the classical optimization problem of covering a rectangular area with circles in such a way that there is no gaps in between the circles and a minimal overlapping area between the circles.

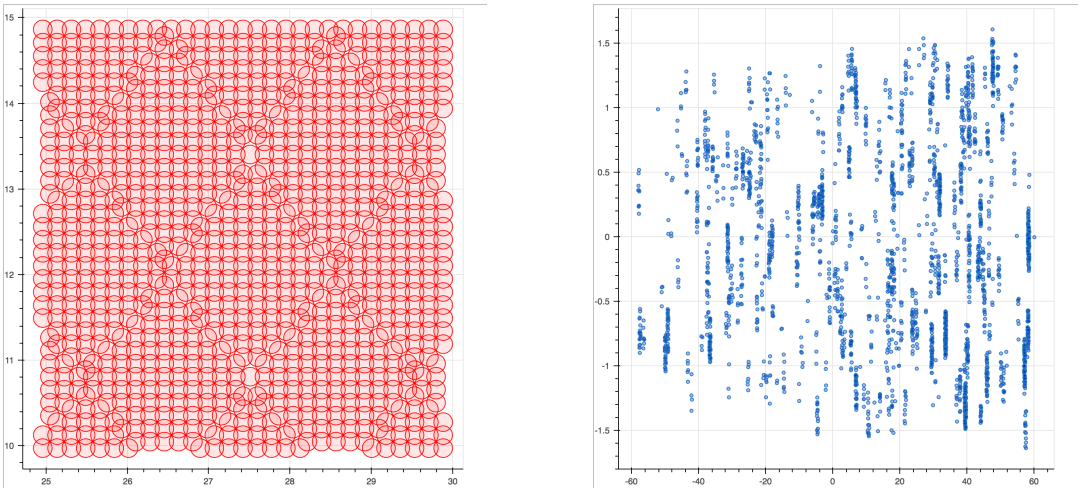
Clearly we want to completely cover the region of interest, but the optimization regarding the overlapping area is not of major concern. It is important to define a methodology to avoid redundant processing so that computational resources are used wisely, but some overlapping excess of overlapping is actually required to compensate for Swift's square images and non-uniform coverage.

We have chosen to use Healpix tesselation schema as it provides the mechanism to define a regularly spaced coordinates grid. The schema defined is multi-dimensional tree-like structure where at each level there will be $12 + 2^{level+1}$ non-overlapping cells covering the sky. At each level, HEALpix grid cells –called diamonds– have an equal area.

Notice that at each level the grid cells have a pre-defined size, which reduces approximately by half at each level up. The factor two –like in a Quadtree– comes from each diamonds being split in four each time we go one level up. Another important characteristic about HEALpix schema is in its cells positioning: the coordinates system is fixed; meaning that a position in the sky will always be represented by the same HEALpix element in a given resolution (level), independent of the surrounding data or platform in use.

The software implemented for this task is published as a small package called moca, it is based in healpy and inspired by mocpy. The package provides an interactive document in its `docs` folder for reference.

To build our list of pointings we queried the Swift Master table for all observations inside the Stripe82 region -60 < RA(deg) < 60 and -1.25 < Dec(deg)



(A) (B)

FIGURE 2.7: HEALPix pointings map example when covering a hypothetical contiguous region (A) and the representation of the Swift observations over the Stripe82 (B).

< 1.5. Each position observed by Swift is associated to a Healpix element, at a given level. Duplicated elements are not considered as our goal here is to define (unique) positions to visit. When all observations have been associated to its respective Healpix element, the inverse transform – *i.e*, from Healpix elements to coordinates – is taken to build up the list of pointings to collect data from.

The Healpix *level* is defined based on Swift XRT field-of-view. Since we want to completely cover the region, the steps between our pointings cannot be bigger than our observations field-of-view. Swift XRT has a FoV of $12'$, Healpix levels 9 and 8 provide pixels with sizes $6.87'$ and $13.74'$, respectively. For our purpose, since we better have extra overlap than gaps, *level 9* is the appropriate level to used when defining the pointings.

When calling the pipeline with the positions from this process and the original radius from Swift-XRT FoV – $12'$ –, adjacent pointings will overlap. Although this is not optimal from the processing point-of-view, it is necessary to guarantee our results the best signal-to-noise by combining all possible events in that area.

Figure 2.7a illustrates the coverage of a small region following the algorithm described above. And figure 2.7b presents all pointings defined to run *DeepSky* over.

The resulting list of pointings contains 699 entries, from ~7000 observations. Now that we have the list of positions we want to visit, we have to

define how we will do it, considering that the processing of ~7000 observations is a quite time consuming task.

Parallel running the pipeline

Covering the Stripe82 $\sim$ 700 pointings were defined, and since they are all independent from each other we may use a parallel strategy to reduce the amount of real time consumed. I will also take the chance to present the software packaging adopted, which makes the setup and parallel run a straightforward process.

Each pointing is processed independently of other runs. Such independence between the runs makes place to a simple parallelism strategy known as *bag of tasks*: a set of independent jobs, homogeneous or not, that may run in parallel in a FIFO (first in, first out) queue structure.

In the of *DeepSky* , the most demanding computational resource is CPU, memory, disk and network I/O present modest use. CPU bottleneck is good because it is an easy to acquire and easy to control resource.

To control the execution of these $\sim$ 700 jobs a easily portable queue system was implemented. It implements a FIFO queue system to which the user input a list of tasks to run and the number `N` of maximum running jobs allowed. The queue system then will keep `N` running jobs, feeding the next in the waiting line, until all (700, for instance) jobs have been completed.

The queue system used was implemented using `Bash` scripting and may be downloaded from Github, Simplest-Ever-Queue system. Many queue systems are publicly available, but none of them is simple enough to *just use*; they all need more-or-less complex setups and they are usually focused on larger, distributed high performance systems. My goal was to provide a simple queue system that anyone with access to a multiprocessor machine could use it effortlessly.

Aggregating the results

Each successful run of the *DeepSky* pipeline will output a set of catalogs, images, log files, etc. Of major interest are the catalogs – photon and energy flux catalogs –, containing measurements for each detected object.

When multiple different runs share an overlapping area and objects are detected in such region we will end with multiple sets of measurements from the same object. From the multiple runs, to have a unique list of sources we

have then to clean out the duplicates. In this Stripe82 processing, concatenating the output of all ~700 runs will generate one big table with inevitably many duplicates.

The removal of duplicated objects is done through a cross-matching, where we basically search for objects that are too close to be two different sources. The definition of this *confusion distance* is usually associated with the instrument's point spread function, at least the PSF is a good first approach since a point source will not be better defined than that. Whenever two (or more) objects follow within this tolerance distance one of them is kept and other(s) are discarded.

Objects within the confusion distance are filtered after their `Full` band signal-to-noise ratio (SNR): the entry with higher SNR is set as the primary source and goes to the final catalog of unique sources.

Finally, the *Swift-DeepSky* over Stripe82 produced a flux catalog with 2755 (unique) sources. In the next section we will look some properties of this final catalog.

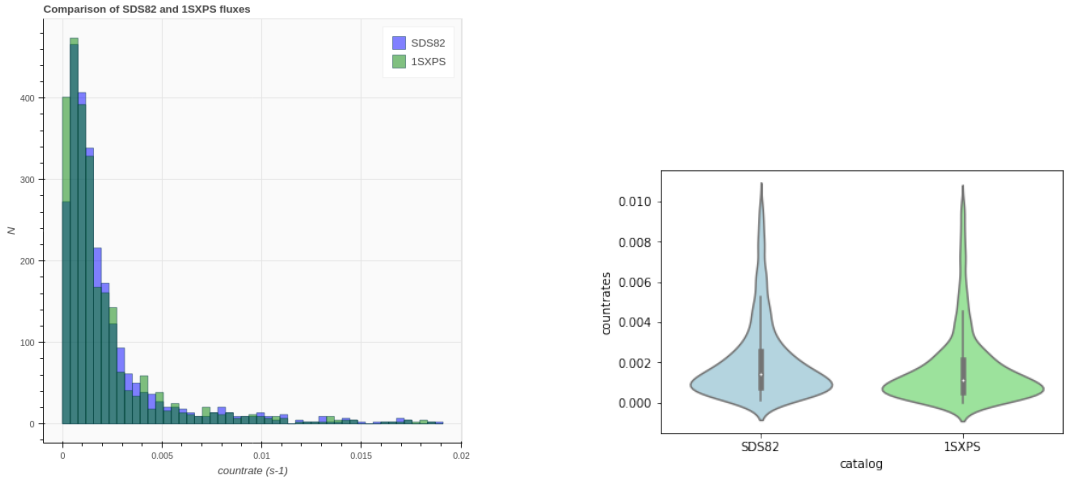
2.3.1 Checking the results

The 1SXPS catalog (Evans et al., 2013) provide an all-sky deep view of the Swift-XRT sky using the first 8 years of observations. That work took a different approach, using data from XRT's Window Timing (WT) mode to investigate variability also, but one of their results is an integrated flux like we have done in this work.

To check how our results compare to (Evans et al., 2013) we cross-matched the catalogs to a distance of 5″, which is the average position error in the 1SXPS. Figures 2.8a and 2.8b present the countrates (flux) distribution side-by-side where we see the catalogs mostly agree, 1SXPS apparently showing an excess in faint sources. A more qualitative visualization of that comparison though may be seen in figures 2.9a and 2.9b, where we see SDS82 recovering more photons from objects, naturally as exposure time is bigger, and in overall agreement with their results.

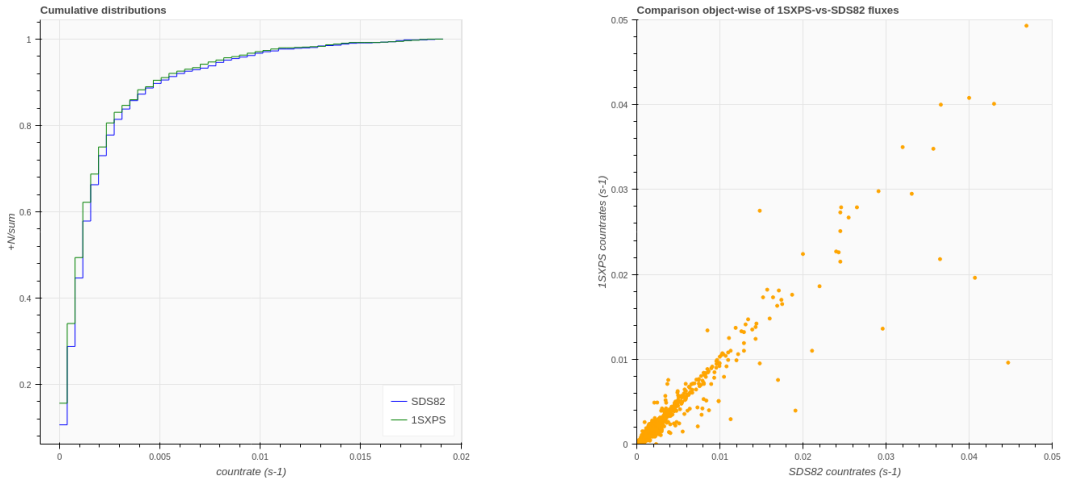
2.4 Pipeline distribution

The pipeline is publicly available and maintained as an open source project. It is distributed using a novel software technology, where the concept of *software portability* is implemented at its highest level.



(A) Histogram (B) Violin-plot

FIGURE 2.8: Countrates distribution (ct/s-1) of catalogs SDS82 and 1SXPS, figure (A) presents the well-known histogram view where we can see the distributions overlap, while figure (B) offers a violin-plot where we see individual distributions density.



(A) Cumulative Distribution (B) Scatter plot

FIGURE 2.9: Countrates comparative plots SDS82 *vs* 1SXPS

Many concepts may characterise a software, *portability* is the one that qualifies whether a software can run in different platforms. A non-portable software is one that runs in one, specific operating system or architecture (*i.e*, platform); on the other hand, a portable software may run in different platforms.

Regarding *portability*, in very recent years the landscape of computing has been significantly changed with the development of linux containers[11]. Containers are the top level of virtualisation technologies, which allows us to

[11]https://en.wikipedia.org/wiki/LXC

mimic an entire environment around a software so that the bare system underneath it can be highly abstracted. This paradigm removes the weight of portability from the (core) software, which not only simplifies the development but also promotes the focus on developing core functionalities for the software. In Section 4.1.1 we go in some deeper details about Containers, in particular their implementation interface Docker (containers).

The *DeepSky* pipeline is distributed in a (Docker[12]) container, which provides the user a *ready-to-use* software package. Everything necessary for the pipeline to run is packaged together – including the HEASoft tools. And the package will run in any platform, although it has been developed for linux systems, because of the virtualisation framework Docker provides, implemented a very efficient abstraction layer between Windows and MacOS systems to make use of Linux containers.

By using such technology and a public code base, we address the portability or transparency issues about reproducibility of scientific results. Another aspect we care about is that of productivity: the use of containers allow the users of this package to spend virtually *zero* time in setting it up, ready for science.

[12]https://en.wikipedia.org/wiki/Docker_(software)

Chapter 3

The SDS82 catalog

Swift-DeepSky outputs a table with the measured fluxes (νF_ν fluxes and countrates) in the `Full` band, `Soft` , `Medium` and `Hard` . In the Stripe82 region, these catalog contain 2755 unique sources. The *flux* catalog provides also total exposure time, energy slope, and NH used during the conversion from countrates to νF_ν flux.

The catalog reaches 5σ flux limits of 4.04×10^{-15}, 4.96×10^{-16}, 1.20×10^{-15} and 7.67×10^{-16} $erg.s^{-1}.cm^{-2}$ in the `Full` , `Soft` , `Medium` , `Hard` , respectively (figure 3.1). Table 3.1 summarizes the numbers of the catalog.

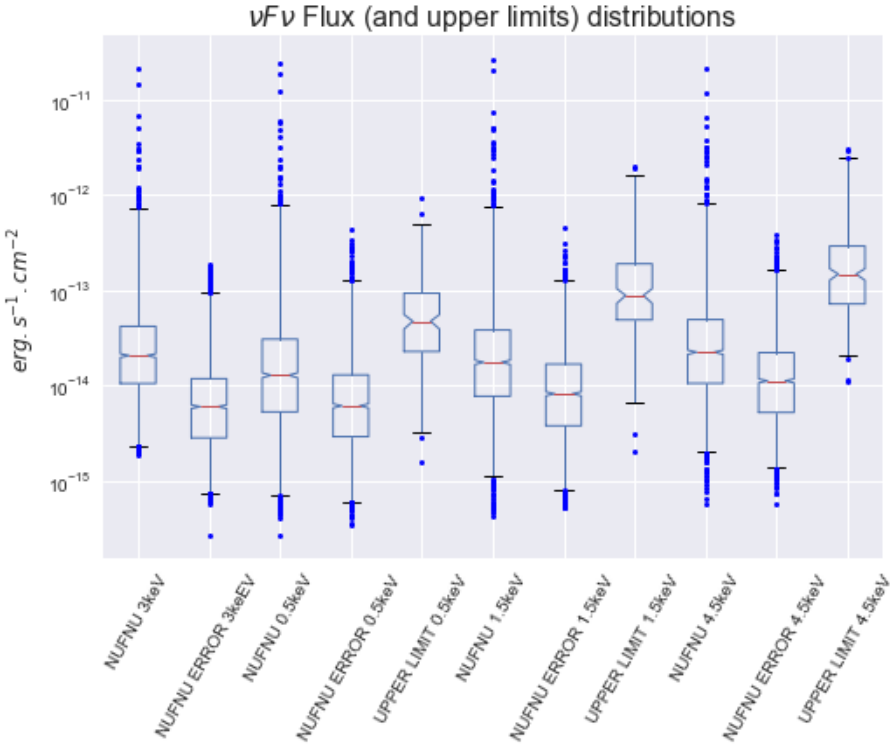


FIGURE 3.1: SDS82 νF_ν fluxes distribution

While figure 3.2 presents the sensitivity of the catalog from a different perspective by showing the `Full` band countrates behavior along the total

exposure time.

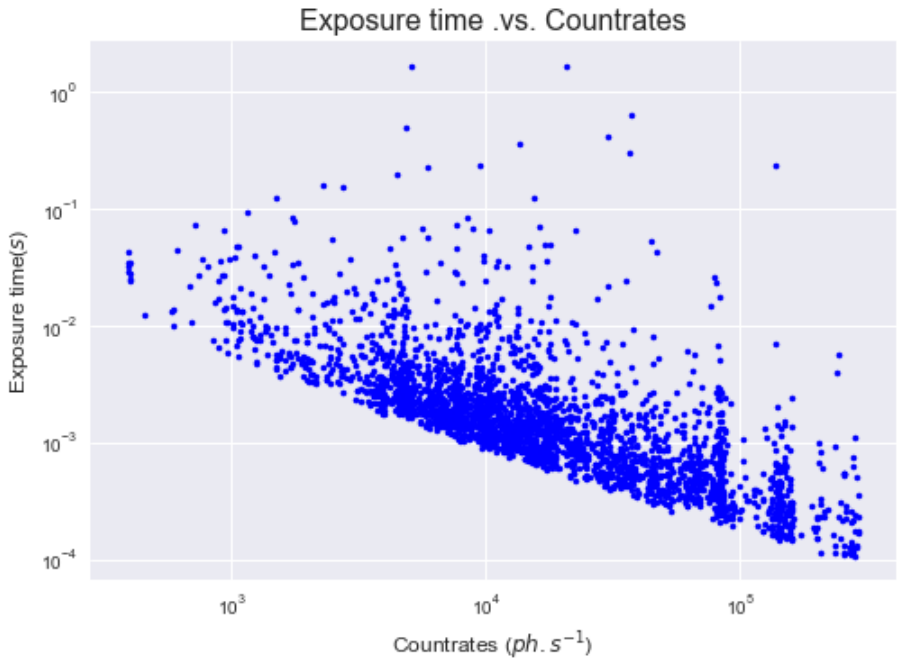


FIGURE 3.2: SDS Countrates *vs* Exposure time

The SDS82 catalog is available through the Virtual Observatories network as well as BSDC's VO website[1].

3.1 Blazars in SDS82

The search for blazars is a particular application for *Swift-DeepSky* data at it may help to reveal distant objects non detected previously by other, shallower studies. Particularly interesting for high energy studies is the class of High Synchrotron Peak (HSP) blazars as they are among the most energetic objects in the Universe emitting photons beyond TeV.

In the SDS82 footprint there are 33 known 5BZCAT (Massaro et al., 2015) blazars, of which 17 are known to be HSP sources (Chang and Giommi, in preparation). Table 3.2 lists these known blazars together with there designation in BZCAT and HSP catalogs.

We then applied the VOU-Blazars tool to search for new blazars in SDS82 sources.

[1] http://vo.bsdc.icranet.org/sds82/q/cone/form

Column	mean	std	min	25%	50%	75%	max
NH	4.6e+20	2.477e+20	1.81e+20	2.82e+20	3.6e+20	5.845e+20	1.19e+21
Energy slope	0.8006	0.1464	-1.723	0.8	0.8	0.8	3.476
Exposure time(s)	35410	49480	387.5	7300	14720	42360	296000
νF_ν 3keV	7.46e-14	5.413e-13	1.901e-15	1.057e-14	2.073e-14	4.204e-14	2.095e-11
νF_ν error 3keV	1.102e-14	1.745e-14	2.729e-16	2.843e-15	6.035e-15	1.185e-14	1.905e-13
νF_ν 0.5keV	7.309e-14	6.749e-13	2.717e-16	5.32e-15	1.29e-14	3.082e-14	2.42e-11
νF_ν error 0.5keV	1.362e-14	2.725e-14	3.451e-16	2.925e-15	6.162e-15	1.301e-14	4.352e-13
Upper limit 0.5keV	8.518e-14	1.146e-13	1.599e-15	2.287e-14	4.752e-14	9.361e-14	9.195e-13
νF_ν 1.5keV	7.854e-14	6.977e-13	4.296e-16	7.774e-15	1.763e-14	3.844e-14	2.623e-11
νF_ν error 1.5keV	1.566e-14	2.579e-14	5.271e-16	3.805e-15	8.241e-15	1.688e-14	4.566e-13
Upper limit 1.5keV	2.029e-13	3.366e-13	2.038e-15	4.909e-14	8.864e-14	1.902e-13	2.031e-12
νF_ν 4.5keV	8e-14	5.314e-13	5.692e-16	1.061e-14	2.251e-14	4.946e-14	2.114e-11
νF_ν error 4.5keV	2.011e-14	3.043e-14	5.721e-16	5.254e-15	1.126e-14	2.224e-14	3.84e-13
Upper limit 4.5keV	3.012e-13	4.685e-13	1.114e-14	7.212e-14	1.483e-13	2.92e-13	3.095e-12
countrates 0.3-10keV	0.006005	0.05047	0.00011	0.0007165	0.00145	0.0029	1.68
countrates error 0.3-10keV	0.0007836	0.001306	3.6e-05	0.00019	0.00042	0.00082	0.019
countrates 0.3-1keV	0.002545	0.02152	1.022e-05	0.0001855	0.0004715	0.001142	0.68
countrates error 0.3-1keV	0.0005026	0.00103	1.161e-05	9.902e-05	0.0002294	0.0004949	0.014
Upper limit 0.3-1keV	0.003134	0.004319	6.406e-05	0.0008502	0.001703	0.003324	0.03663
countrates 1-2keV	0.002212	0.01988	1.207e-05	0.0002145	0.0004884	0.00107	0.7504
countrates error 1-2keV	0.0004357	0.0007269	1.474e-05	0.0001043	0.0002304	0.0004686	0.01306
Upper limit 1-2keV	0.005649	0.009462	5.723e-05	0.001337	0.002445	0.00512	0.05748
countrates 2-10keV	0.001522	0.01057	1.056e-05	0.000196	0.0004161	0.0009164	0.4076
countrates error 2-10keV	0.0003744	0.0005739	1.36e-05	9.702e-05	0.0002081	0.000413	0.0076
Upper limit 2-10keV	0.005582	0.008693	0.0002055	0.001335	0.00275	0.005406	0.05748

TABLE 3.1: Distribution of values in the SDS82 catalog.

3.1.1 VOU-Blazars

Blazars may be identified by the slope between their x-ray and radio emission. Using this empirical evidence the tool VOU-Blazars (Y. Chang, in preparation) has been developed to search for blazars in VO catalogs. The tool select objects that present a Radio-to-X-ray energy flux characteristic of blazars and evaluates the chances of the each object being a LSP, ISP or HSP blazar.

Given a position (RA,DEC) in the sky and a search radius, VOU-Blazars selects all objects in the field presenting Radio emission. For each Radio source, a X-ray counterpart is searched and the ratio between the corresponding νF_ν fluxes is evaluated to have a first an indication of the kind of object we are dealing with. Table 3.3 presents the criteria for decision on what kind of blazar candidate VOU-Blazars evaluates.

After the first screening based on the Radio/X-ray fluxes ratio (RXR), the tool proceeds to collect and cross-match the potential candidates with

TABLE 3.2: Known blazars in SDS82 catalog.

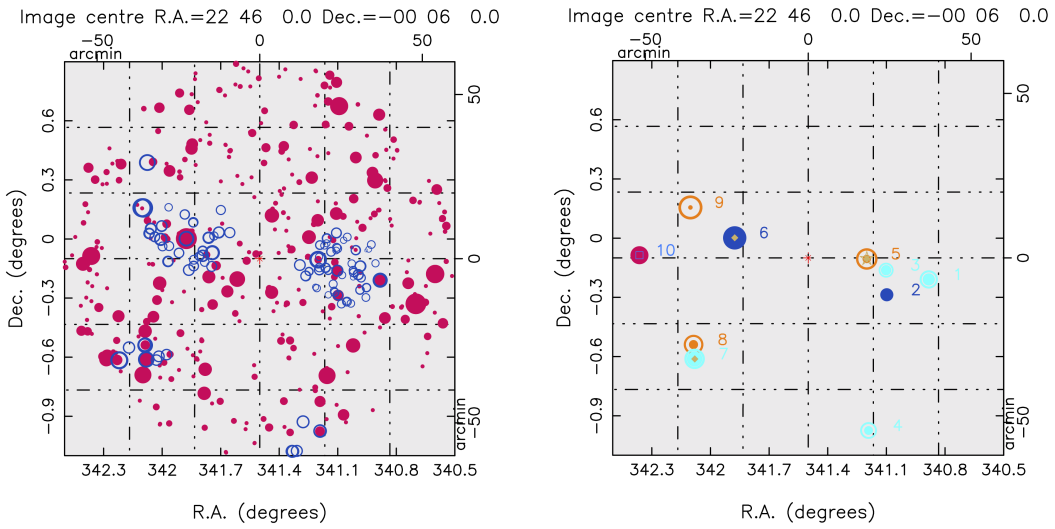
NAME	RA	DEC	$NAME_{5bzcat}$	RA_{5bzcat}	DEC_{5bzcat}	z_{5bzcat}	Source classification	$NAME_{3HSP}$	RA_{3HSP}	DEC_{3HSP}	z_{3HSP}
SDSX J0002.0-0024.0	0.737912	-0.413523	5BZBJ0002-0024	0.73817	-0.41308	0.523	BL Lac				
SDSX J0011.0+0057.0	2.87614	0.964622	5BZQJ0011+0057	2.87667	0.96444	1.492	QSO RLoud flat radio sp.				
SDSX J0016.0-0015.0	4.04579	-0.253309	5BZQJ0016-0015	4.04621	-0.25344	1.577	QSO RLoud flat radio sp.				
SDSX J0022.0+0006.0	5.50436	0.116323	5BZGJ0022+0006	5.50396	0.11611	0.306	BL Lac-galaxy dominated	3HSP J002200.9+000657	5.5039788	0.1160734	0.306
SDSX J0108.0-0037.0	17.1117	-0.623032	5BZQJ0108-0037	17.11183	-0.62339	1.374	QSO RLoud flat radio sp.				
SDSX J0125.0-0005.0	21.3707	-0.0984472	5BZQJ0125-0005	21.37017	-0.09886	1.077	QSO RLoud flat radio sp.				
SDSX J0148.0+0129.0	27.1409	1.48359	5BZBJ0148+0129	27.14079	1.48369	0	BL Lac	3HSP J014833.7+012901	27.1408395	1.4837295	0.94
SDSX J0158.0+0101.0	29.7194	1.02572	5BZUJ0158+0101	29.71988	1.02578	0.952	Blazar Uncertain type				
SDSX J0201.0+0033.0	30.2756	0.566597	5BZBJ0201+0034	30.27575	0.56672	0.298	BL Lac	3HSP J020106.1+003400	30.275698	0.5667576	0.298
SDSX J0243.0+0046.0	40.7618	0.774161	5BZBJ0243+0046	40.76221	0.77425	0.409	BL Lac	3HSP J024302.9+004627	40.7622065	0.7742376	0.4
SDSX J0259.0-0019.0	44.8689	-0.333232	5BZQJ0259-0020	44.86883	-0.33333	2	QSO RLoud flat radio sp.				
SDSX J0301.0+0118.0	45.348	1.30878	5BZQJ0301+0118	45.34838	1.31	1.221	QSO RLoud flat radio sp.				
SDSX J0304.0-0054.0	46.1405	-0.901947	5BZBJ0304-0054	46.1415	-0.90128	0.511	BL Lac	3HSP J030433.9-005404	46.1414732	-0.9012285	0.511
SDSX J0323.0-0108.0	50.9862	-1.14152	5BZBJ0323-0108	50.986	-1.14156	0.392	BL Lac	3HSP J032356.5-010833	50.9855352	-1.1427243	0.392
SDSX J0323.0-0111.0	50.9318	-1.19662	5BZBJ0323-0111	50.93175	-1.19614	0	BL Lac	3HSP J032343.6-011146	50.9317492	-1.1961639	0.44
SDSX J2014.3-0047.0	303.619	-0.78975	5BZBJ2014-0047	303.61933	-0.78967	0	BL Lac	3HSP J201428.6-004722	303.6193604	-0.7896846	0.231
SDSX J2055.3-0021.0	313.868	-0.354781	5BZBJ2055-0021	313.86758	-0.35478	0	BL Lac	3HSP J205528.2-002116	313.8675988	-0.3547826	0.44
SDSX J2118.3+0013.0	319.572	0.221357	5BZQJ2118+0013	319.57246	0.22131	0.463	QSO RLoud flat radio sp.				
SDSX J2129.3+0035.0	322.418	0.590902	5BZUJ2129+0035	322.41954	0.59108	0.426	Blazar Uncertain type	3HSP J212940.6+003527	322.4194413	0.5910023	0.426
SDSX J2136.3+0041.0	324.16	0.698517	5BZQJ2136+0041	324.16079	0.69839	1.941	QSO RLoud flat radio sp.				
SDSX J2153.3-0042.0	328.272	-0.708726	5BZBJ2153-0042	328.27225	-0.7085	0.341	BL Lac	3HSP J215305.3-004230	328.2721903	-0.7084445	0.341
SDSX J2156.3-0037.0	329.062	-0.618034	5BZUJ2156-0037	329.0615	-0.61794	0.495	Blazar Uncertain type				
SDSX J2206.3-0031.0	331.68	-0.517617	5BZBJ2206-0031	331.68038	-0.51736	0	BL Lac				
SDSX J2211.3-0003.0	332.785	-0.0506528	5BZBJ2211-0003	332.78475	-0.05069	0.362	BL Lac	3HSP J221108.3-000302	332.7847611	-0.0506922	0.362
SDSX J2211.3-0023.0	332.791	-0.391599	5BZGJ2211-0023	332.79117	-0.39094	0.448	BL Lac-galaxy dominated	3HSP J221109.8-002327	332.7911597	-0.3909912	0.447
SDSX J2223.3+0102.0	335.872	1.04023	5BZBJ2223+0102	335.87321	1.04075	0	BL Lac	3HSP J222329.5+010226	335.8732192	1.0407506	0.51
SDSX J2226.3+0052.0	336.694	0.86978	5BZQJ2226+0052	336.69392	0.86981	2.262	QSO RLoud flat radio sp.				
SDSX J2244.3-0006.0	341.201	-0.105253	5BZBJ2244-0006	341.20038	-0.10542	0	BL Lac	3HSP J224448.0-000619	341.2004064	-0.1054444	0.7
SDSX J2247.3+0000.0	341.875	0.00165417	5BZBJ2247+0000	341.87583	0.00181	0	BL Lac				
SDSX J2248.3-0036.0	342.081	-0.611605	5BZGJ2248-0036	342.081	-0.61158	0.212	BL Lac-galaxy dominated				
SDSX J2254.3+0054.0	343.518	0.906085	5BZBJ2254+0054	343.51837	0.90583	0	BL Lac				
SDSX J2319.3-0116.0	349.969	-1.27346	5BZGJ2319-0116	349.97017	-1.27411	0.284	BL Lac-galaxy dominated	3HSP J231952.8-011626	349.9701518	-1.2740842	0.28
SDSX J2356.3-0023.0	359.016	-0.398281	5BZBJ2356-0023	359.01671	-0.39828	0.283	BL Lac	3HSP J235604.0-002353	359.0167358	-0.3982722	0.283

X-ray/Radio α slope	Candidate type
$0.78 < \alpha < 0.95$	LSP
$0.42 < \alpha < 0.78$	ISP/HSP
$\alpha < 0.42$	non-jetted AGN

TABLE 3.3: X-ray to Radio flux ratio for blazar candidates classification applied by VOU-Blazars

other multi-wavelength catalogs. In the second phase, the fluxes in different wavebands are retrieved to build each source's Spectral Energy Distribution (SED). VOU-Blazars also consults catalogs dedicated to quasars and blazars, in particular, 5BZCAT (Massaro et al., 2015), CRATES (**crates**) and 3HSP (Chang and Giommi, in preparation) catalogs, to graphically indicate it to the user.

At the end of the first phase, VOU-Blazars presents to the user two (graphical) plots indicating the candidates found (figure 3.3b and the map with Radio and X-ray sources found in the requested field (figure 3.3a.



(A) Radio/X-ray sources (B) Blazar candidates

FIGURE 3.3: VOU-Blazars first-phase results: sky maps showing Radio and X-ray sources (A) and known blazars and blazar candidates in the field (B).

Figures 3.3a and 3.3b provide a complete example of the quality of information VOU-Blazars provides the user at the first phase. On the left (3.3b) we see colorful labeled points indicating blazar candidates in the 4 $arcmin^2$ field around position $RA = 341.5(deg)$, $Dec = -0.1(deg)$. Colors and symbols we can see in this figure are explained in table 3.4, they indicate whether

the corresponding object was previously known (symbols) and if it is a good candidate for LSP, ISP or HSP (colors).

Symbol	Meaning
diamond	5BZCAT blazar
star	3HSP blazar
square	CRATES source
circle *size*	Radio *intensity*

Color	Meaning
orange	HSP candidate
cyan	ISP candidate
blue	LSP candidate
red	Radio-only detection

TABLE 3.4: VOU-Blazars candidates plot colors and symbols

At this point – second phase – the user may request the SED of each candidate. VOU-Blazars then queries more then 30 catalogs, from different wavelengths (table 3.5) to bring to the user the location of the sources cross-matched and the SED assembled. The complete list of catalogs used by VOU-Blazars is listed in table 3.5.

The figures panel 3.4 present the SED and location map of sources 5-10 seen in figure 3.3b.

The SED and source location plots from VOU-Blazars are powerful instruments to have a good idea about the object at hand, from them we may eliminate objects clearly out of our interest or of low quality data or, in case very good photometric data is available to VOU-Blazars , even conclude for a blazar . Usually though, VOU-Blazars is not conclusive and we will make use of other tools and datasets to have a precise view of the object, potential candidates will then be further analysed.

Typically, we will use the *Open Universe* portal[2] to have a follow-up, a complementary view of a candidate. Through the portal interface we have access to all major astronomical data access services available internet-wide, all linked to the object (name or position) you first-type when you enter the portal. Particularly useful for this work are the interfaces with *SSDC SED tool*[3], *NED*[4], *SDSS SkyServer*[5] and *Aladin*[6]. These services provide us general information from previous studies (*NED*), where we can have a broad idea of what the object has gone through so far; A closer look of the object appearance in Optical (*SDSS SkyServer*), specially if a spectrum is available to give us a detailed view of its identity; A image gallery of the Space around the object (*Aladin*), helping us to identify a possible contamination of VOU-Blazars

[2]http://www.openuniverse.asi.it/
[3]https://tools.asdc.asi.it/SED/
[4]https://ned.ipac.caltech.edu/
[5]http://skyserver.sdss.org/dr13/en/tools/chart/navi.aspx
[6]http://aladin.u-strasbg.fr/AladinLite/

Catalog	Waveband/Information	Reference
SUMSS	Radio 843 MHz	Mauch et al. (2003)
NVSS	Radio 1.4 GHz	Condon et al. (1998)
FIRST	Radio 1.4 GHz	White et al. (1997)
NORTH20	Radio 1.4 GHz	White and Becker (1992)
PMN	Radio 4.85 GHz	Wright et al. (1994)
GB87	Radio 4.85 GHz	Gregory and Condon (1991)
GB6	Radio 4.85 GHz	Gregory et al. (1996)
ATPMN	Radio 4.8 and 8.6 GHz	McConnell et al. (2012)
AT20G	Radio 20 GHz	Murphy et al. (2010)
PCCS44	Radio/mm 44 GHz	Ade et al. (2014)
PCCS70	Radio/mm 70 GHz	Ade et al. (2014)
PCCS100	Radio/mm 100 GHz	Ade et al. (2014)
PCCS143	Radio/mm 143 GHz	Ade et al. (2014)
PCCS217	Radio/mm 217 GHz	Ade et al. (2014)
PCCS353	Radio/mm 353 GHz	Ade et al. (2014)
SPIRE500	Sub-mm 500 μm	Schulz et al. (2017)
SPIRE350	Sub-mm 350 μm	Schulz et al. (2017)
SPIRE250	Sub-mm 250 μm	Schulz et al. (2017)
PACS70	Infrared 70 μm	Marton et al. (2017)
WISE	Infrared 3.4, 4.6, 12, and 22 μm	Wright et al. (2010)
2MASS	Infrared 1.25, 1.65 and 2.16 μm	Skrutskie et al. (2006)
SDSS-DR14	Optical ugriz	Abolfathi et al. (2017)
HSTGSC	Optical UBVRI	Lasker et al. (2008)
PanSTARRS	Optical grizy	Flewelling et al. (2016)
GAIA	Optical G (white)	Prusti et al. (2016)
GALEX	Ultra-violet FUV, NUV	Martin et al. (2005)
XMMOM	UV/Optical UVW2,UVM2,UVW1,U,B,V	Page et al. (2012)
UVOT	UV/Optical UVW2,UVM2,UVW1,U,B,V	Yershov (2014)
CMA	X-ray 0.05-2.0 keV	Giommi et al. (1991)
RASS	X-ray 0.1-2.4 keV	Boller et al. (2016)
WGACAT	X-ray 0.1-2.4 keV	White, Giommi, and Angelini (1994)
BMW	X-ray 0.1-2.4 keV	Panzera et al. (2003)
IPC2E	X-ray 0.4-4 keV	Harris et al. (1994)
IPCSL	X-ray 0.4-4 keV	Elvis et al. (1992)
CHANDRA	X-ray 0.1-10 keV	Evans et al. (2010)
SDS82	X-ray 0.3-10 keV	This thesis
SXPS	X-ray 0.3-10 keV	Evans et al. (2013)
XMMSL	X-ray 0.2-12 keV	Saxton et al. (2008)
3XMM	X-ray 0.2-12 keV	Rosen et al. (2016)
BAT105	X-ray/Gamma 14-195 keV	Oh et al. (2018)
3FGL	Gamma-ray 0.1-300 GeV	Acero et al. (2015)
3FHL	Gamma-ray 10-2000 GeV	Collaboration (2017)
XRTSPEC		
1BIGB		
5BZCAT	Blazars	Massaro et al. (2015)
3HSP	HSP Blazar candidates	Chang and Giommi (in preparation)
CRATES	Flat Spectrum Radio Sources	Healey et al. (2007)
ZWCLUSTERS	Galaxy clusters	Zwicky et al. (1961)
PSZ2	Galaxy clusters	Collaboration (2016)
ABELL	Galaxy clusters	Abell, Corwin, Harold G., and Olowin (1989)
MCXC	Galaxy clusters	Piffaretti et al. (2011)
SDSSWHL	Galaxy clusters	Wen, Han, and Liu (2012)
SWXCS	Galaxy clusters	Liu et al. (2015)
PULSAR	Pulsars	Bock (2014)
F2PSR	Pulsars	Collaboration (2013)

TABLE 3.5: Catalogs (VO) used by VOU-Blazars

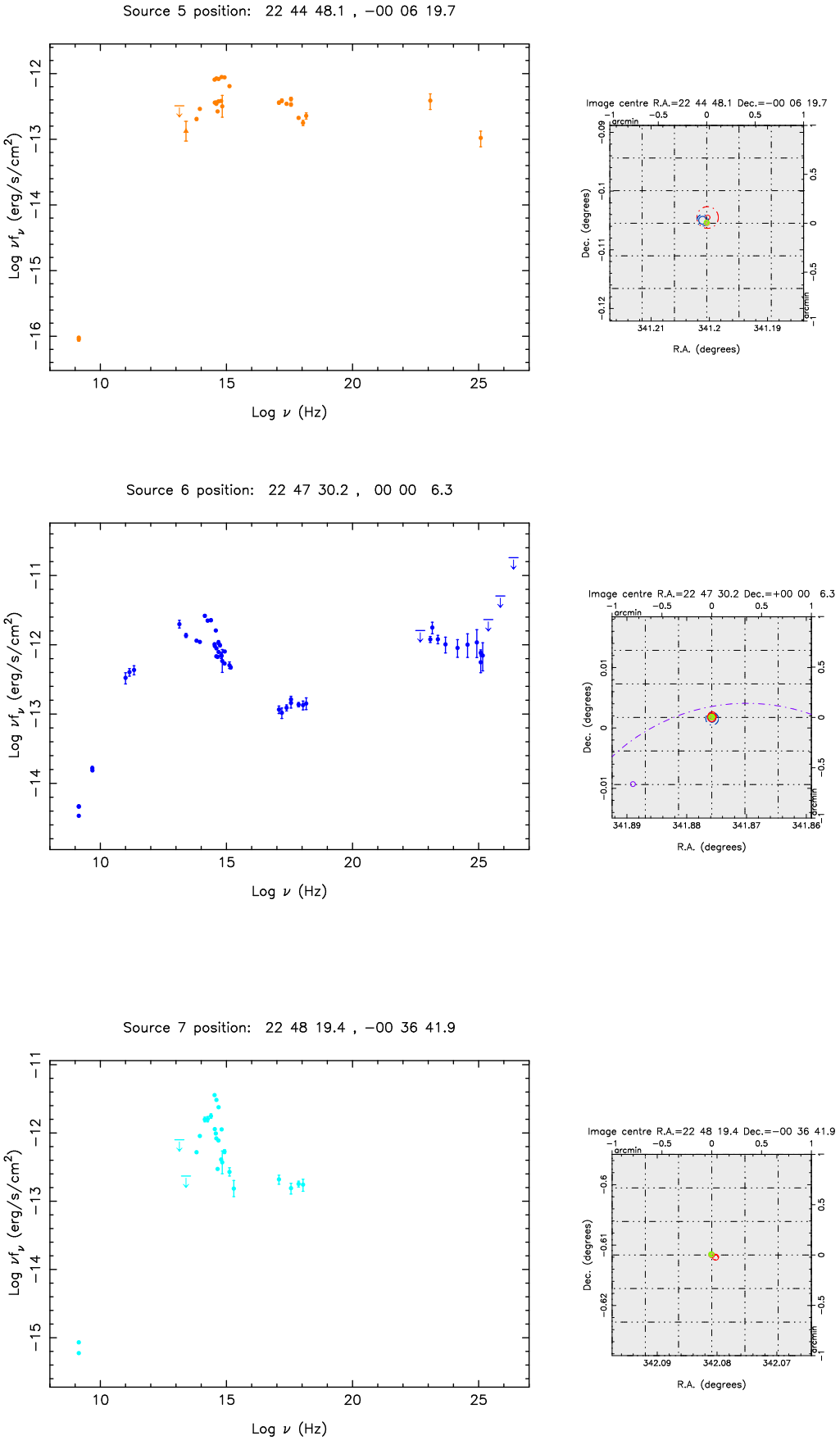


FIGURE 3.4: VOU-Blazars candidates 5, 6, 7 from figure 3.3

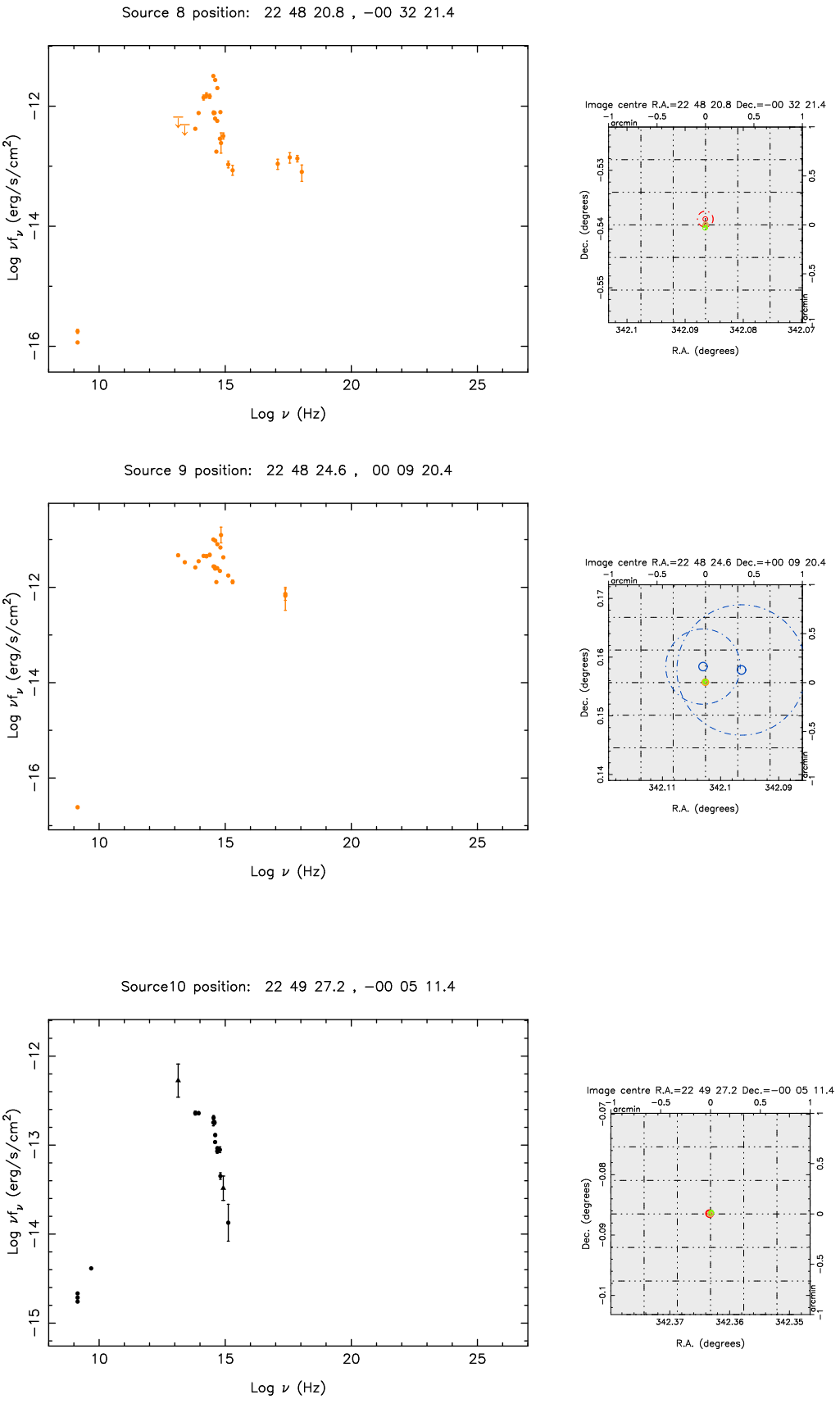


FIGURE 3.5: VOU-Blazars candidates 8, 9, 10 from figure 3.3

data by nearby objects; And finally we may have a complete view of a full featured SED interface through *SSDC SED* tool.

An VOU-Blazars candidate follow-up

For this exercise, let us take the object '9' from figure 3.3b, with VOU-Blazars 's SED output in figure 3.5. This object, at position `22h48m24.6s, +00d09m20.4s` (*RA* 342.1026 deg, *DEC* 0.15567 deg) is quite dubious as it has (*i*) the correlated X-ray emission somewhat displaced (figure 3.3a, although the positional error circles do match), (*ii*) the IR/Optical/UV emission looks noisy (could be optical variabilty, or confusion with nearby objects), (*iii*) the Radio flux is quite low.

We then go to the *Open Universe* portal and query the position. From the services highlighted previously we get the following information (figure **??** persent the respective images):

- *NED* says we are dealing with a Seyfert 1 galaxy at redshift $z = 0.053709$;
- *SDSS* has its spectrum, labeled as "*QSO Starburst Broadline*" object;
- *Aladin* shows images from different surveys, different wavebands;
- *SSDC* presents a broader view of the object and possibly surroundings,

from where we conclude the object '9', `22h48m24.6s, +00d09m20.4s`, is a Seyfert 1 galaxy (*i.e,* non-jetted AGN).

3.1.2 New Blazar candidates after SDS82

To verify the impact *Swift-DeepSky* can make to high energy astrophysics, in particular on the interest of blazars studies, we specifically searched for blazars (using VOU-Blazars) among objects that have no X-ray emission other than *Swift-DeepSky* [7].

To get there, we setup VOU-Blazars to get data from the SDS82 catalog available through BSDC-VO conesearch service[8]. Then we ran VOU-Blazars over the very same fields used for the generation of SDS82 catalog – same position and radius.

Over the whole SDS82 footprint, 300 blazar candidates were found by the tool alone. We then selected those VOU-Blazars candidates where *only* SDS82

[7] Considering the current catalogs in VOU-Blazars (table 3.5)
[8] http://vo.bsdc.icranet.org/sds82/q/cone/form

data was available among the X-ray catalogs counterparts, which left us with an expressive list of 65 sources. Each of the 65 sources were inspected following the procedure described in the previous Section 3.1.1 to understand if such sources are potential HSP blazar candidates. 3.6.

TABLE 3.6: SDS82 blazar candidates inspection summary

RA	Dec	z	Comments	Known
1.44055	-0.5804		Group of galaxies	
2.66172	-0.86275	1.51382	FSRQ candidate	
13.64979	-0.03183	1.158(p)	BLLAC LSP candidate	
16.671	-0.86419	0.86979	Too few data points	
16.87579	-0.80214	1.882	Unknown AGN	
17.11182	-0.62339		FSRQ	yes
18.15025	0.2052		Too few data points	
20.51987	0.69389		FSRQ candidate	
22.88712	0.55589	0.07932	Unknown AGN	
24.81063	-0.78764	0.42986	BLLAC LSP candidate	
28.7511	0.83093	0.82587	FSRQ candidate	
29.63553	-0.7107	2.61297	BL QSO	
29.71418	-0.81924	2.85811	BL QSO	
29.71983	1.02599	1.61	BLLAC LSP candidate	
30.55956	-0.29674		BLLAC LSP candidate	
31.17359	0.83278	0.627	BLLAC HSP candidate	
37.11625	-1.1795	1.13	BL QSO	
39.49495	-1.15307	0.25121	Unknown AGN	
39.84849	1.09397	0.00898	Beautiful Edge-on galaxy	
39.96684	-0.85066	1.64604	QSO	
43.37334	-0.23486	0.02876	Seyfert 2	
44.05857	0.6614	0.91707	BL QSO	
44.10684	-1.31986	2.491	QSO candidate	
45.34835	1.31016	1.221	FSRQ	yes
45.4208	1.26419	1.77358	QSO	
45.9208	-0.95196	1.80511	BL QSO	
47.8281	1.31466		non-jet AGN	
49.40681	1.31718		Unknown type	
50.16968	-1.02256	1.18564	Too few data points	
57.0961	-0.81041		Unknown AGN	
57.26806	-1.03796		Too few data points	
57.50275	-1.46556	0.04110	Edge-on galaxy	
58.10566	-0.65656		Too few data points	
58.57137	-0.16614		Unknown type	
302.61854	-0.79139		Unknown type	
310.4728	0.4861	0.39803	FSRQ	
310.92567	0.02195	0.405(p)	ISP candidate	
314.95312	0.6394		QSO	
316.17875	1.13635	2.065(p)	BLLAC LSP candidate	FSRS
322.81437	0.67082	1.496	BL QSO	
323.45244	-0.69751	1.63	BLLAC LSP candidate	
328.55353	0.07315	0.2168	HSP candidate	
329.97691	-0.36395	1.96541	FSRQ candidate	
330.14146	-0.47269	1.20571	FSRQ candidate	
331.68529	-0.65168		BL QSO	
333.28546	0.61556	1.21473	Too few data points	
335.13587	0.42667	4.21036	BL QSO	
337.28988	0.24732	0.05487	non-jet AGN	
337.37326	-0.14586		BLLAC LSP candidate	
337.85872	0.07664	0.9331	FSRQ candidate	
340.6006	0.92031	3.98553	Unknown type	
340.6112	1.19536	0.04749	Seyfert 2/Galaxy merge	
342.08091	-0.61164	0.21234	HSP candidate	yes
349.95737	0.06468	0.18524	LSP candidate	
354.00654	0.1781		Unknown AGN	
355.44293	-1.35199		Beautiful Face-on galaxy	
357.12675	0.65516	1.99956	FSRQ candidate	
357.15495	0.79121		Too few data points	

3.2 SDS82 value-added catalog

As a side product, due to the various various minor developments a thesis goes through, we correlated the SDS82 catalog to other two catalogs, the Optical 'CS82' (Soo et al., 2018) and 'Stripe82X' (LaMassa et al., 2015) catalogs. The CS82 catalog was cross-matched using the a Maximum Likelihood Ratio (MLE) method (Sutherland and Saunders, 1992; LaMassa et al., 2015), Stripe82X using the Great-Circle (GC), distance-based method. For these cross-matchings tasks we developed the `xmatch`[9] tool implementing the MLE method.

LaMassa et al. (2015) combined archival XMM–Newton and Chandra observations as well as dedicated observations from XMM–Newton to cover $\sim$ 31 deg^2 of the Stripe82 to create the Stripe82X catalog. In total, the Stripe82X catalog contains 3362 unique X-ray sources matched to multi-wavelength catalogs using the MLE algorithm.

The CS82 Survey is a joint Canada-France-Brazil project to observe the Stripe82 using the MegaCam at the Canada France Hawaii Telescope (CFHT). It surveyed approximately 170 deg^2 of the equatorial Stripe82 area. It is a relatively deep survey that maps down to magnitude 24.1 in the i-band for a point-like source detection at 5σ. The survey was able to produce final images with mean seeing of 0.6".

The catalog we use in this work was the one generated and used by Soo et al. (2018) and Charbonnier et al. (2017) in their studies on galaxy morphology. Besides flux measurements, the catalog provides the redshift measurements (spectroscopic or photometric) for each object as presented in Soo et al. (2018). More than 6 million objects, extended and point-like sources, compose the CS82 catalog.

3.2.1 Cross-matching astronomical catalogs

Cross-matching is the process of relating objects between two (or more) astronomical catalogs. Considering that both catalogs contain references of objects from the same region of the sky, the goal is to identify which are the references pointing to the same objects.

For a proper explanation, let us consider two astronomical catalogs – where rows contain properties of astronomical objects, columns organize those properties and each (real) astronomical object can be present no more than once in each catalog. To picture a very simple situation, without loosing

[9] https://github.com/chbrandt/xmatch

in generality, we can think of such catalogs as being the products of optical and x-ray observations of a certain region of the sky. It it is expected – say the *null-hypothesis* – that not all but some of the objects are in both catalogs. Notice, though, that the catalogs – as data structures – are not required to share anyother structural property like number or order of rows and columns. This is a typical scenario astronomers handle a cross-matching.

The process of finding the objects shared by both catalogs is called cross-matching. In extragalactic astronomy, in practice the objects – galaxies, QSOs – do not move; which brings their position in the sky to be used as an identifier. The very basic parameters to be used for the cross-matching is then Right Ascension and Declination: at each of those catalogs, the objects that are in the same position of the sky are said to be the same. The result of this process is a *cross-matched* catalog, containing the matched objects and the merge of columns (*i.e.*, properties) from both catalogs.

Although it looks like a simple subject, cross-matching is a long-standing issue in astronomy. And it is quite easy to see that once we realize how observational effects (*e.g.*, astronomical seeing) affect, for instance, an object's position measurement. The uncertainties added to the measurements cause the same object to show up slightly different positions in each catalog; given that we have to match not exact values, but coordinates that should match within a tolerance value ϵ – also called *error radius* or *search radius*. Intrinsic astrophysical effects can also cause the same astronomical object not to match between catalogs of different wavebands, for example, Radio Lobes generated by an Active Galactic Nuclei (AGN) can cause a mismatch between a radio and an optical catalogs.

In this work we implemented a maximum likelihood estimator (MLE, Sutherland and Saunders, 1992) during the process of cross-matching catalogs from different wavebands where the surface density and astrometric precision are sufficiently different and the trivial position-matching becomes ambiguous. That is the case, for instance, when cross-matching the Swift x-ray catalog with the CS82 optical catalog. Within the Swift (positional) error radius we will usual find more than three optical candidates, to identify the correct counterpart we may use other properties of the objects to improve our results.

3.2.2 Maximum Likelihood Estimator(MLE)

MLE is applied by LaMassa et al. (2015) to find the correct – or most probable – counterpart to their x-ray sources. MLE was first proposed by Sutherland and Saunders (1992) and is being adopted as a better alternative to the simplistic *great-circle* algorithm.

What MLE does is to estimate how probable a given counterpart candidate is to be real counterpart from a source in its vicinity. The method was developed having in mind that multiple candidates can be nearby in the (RA,Dec) sky-projected plan. Accordingly, the method includes the ancillary magnitudes as a third component to help differentiating background objects from candidate(s).

Rational

Consider the situation where there is a source S (observed by instrument A) and in the vicinities, within a distance Δd_i, there are N objects (obj_1, obj_2, ..., obj_N) that were observed by a different instrument (B). Also from catalog B, but further distant from S there are M objects (obj_1, obj_2, ..., obj_M) that can not be considered candidates to S, but will be used as background sample. The M objects lie beyond the distance Δd_i up to a distance Δd_o, $\Delta d_o > \Delta d_i$. The question we want to answer is "*which of the objects observed by 'B' is in fact 'S', observed by 'A'*? Notice that instruments A and B present physical effects that lead to uncorrelated errors and different image resolutions, which means that S and its (true) counterpart may *not* be the nearest one.

The distance Δd_i from S is considered to be the "vicinity", and objects inside this distance are considered are called the *ancillary* objects, candidates to the (true) counterpart. The objects from sample M will be called *background* objects, they compose the sample of objects observed by B not considered as S counterpart.

The MLE method will eventually the samples and give a score called *Reliability* (R) to each of counterpart candidate by considering non only the separation distance, but an extra *feature*, the brightness of each object. The Reliability is the probability of being the true counterpart, and is given by:

$$R_j = \frac{LR_j}{\sum_j LR_j + (1 - Q)}$$

The central figure in MLE is the likelihood ratio, LR:

$$LR_j = \frac{q(m)f(r)}{n(m)}$$

.

$f(r)$ is the statistical *prior* regarding the position of the candidate object relative to the source. The $f(r)$ function is modeled as a bi-dimensional Gaussian with σ being the quadrature sum of source's positional error and objects' average positional error:

$$f(r) = \frac{1}{2\pi\sigma} exp^{-r^2/2\sigma^2};$$

$$\sigma = \frac{1}{2}\left[\sqrt{\sigma_{\alpha_S}^2 + \sigma_{\delta_S}^2} + \sqrt{\sigma_{\alpha_O}^2 + \sigma_{\delta_O}^2}\right]$$

The $q(m)$ factor is the likelihood of the object being a good candidate given its brightness. It is computed by drawing the ancillary objects normalised magnitude distribution and subtracting from it the normalised background magnitude distribution.

Finally, $n(m)$ is the surface density of background objects with magnitude m. It is computed by counting the number of background objects per magnitude bin per square-degree; normalised by the number of objects.

The algorithm

To compute MLE quantities we need to define the background and ancillary samples. To do that we will define the *search radius* (r_s) – from where the *ancillary* sample will come out – and the *inner & outer radii* (r_i, r_o) defining the background sample.

Search radius: There are different ways to estimate the (best) *search radius*. Typically, the instrument's (nominal) error radius, systematic plus statistical, is used, as in LaMassa et al. (2015). Timlin et al. (2016) have used the Rayleigh Criterion to estimate such radius, a physical limitation on resolving close by objects; similarly, the overall PSF (FWHM) is a valid estimator. Another way of estimating r_s, data driven, is by directly estimating the typical distance between the objects in each catalog.

Analogously, we have to define the *inner and outer radii*, from the primary source, of the annulus defining the background region. Truly speaking, the background region does *not* need to be drawn as an annulus centered in

the source, but that is a common, straightforward choice for sampling background sources. It is important to notice that the background region should avoid other nearby sources ancillary sample, which is to say that the (annulus) region should not intersect with another source's search area.

- ancillary search radius: r_s
- background annulus radii: $r_i < r_o$

Samples definition: Once we have the radii defined we cross-match the catalogs to define the *ancillary* and *background* samples. At this point, each target source has two lists of objects related to it:

- Target source:

  ancillary sample (within Rs)

  background sample (between Ri and Ro)

  We then define global $q(m)$ and $n(m)$.

- Estimate magnitude distributions

  $n(m)$: background surface brightness distribution

  $q(m)$: ancillary brightness distribution

  Q: expected counterpart recover rate (*efficiency*)

Radial prior: The radial profile $f(r) \propto \sigma^{-1} \exp^{-r^2/\sigma^2}$ is defined for each source, for σ is a function of the source and ancillary objects positional errors, σ_s and σ_o, resp.:

$$\sigma = \sqrt{\frac{\sigma_s^2 + \sigma_o^2}{2}}$$

If the positional errors are well behaved – i.e, their variance is small –, we may approximate $f(r)$ as a global function. We may consider σ_s and σ_o as the mean of the respective positional errors.

- Compute mean positional errors
- primary sources catalog
- ancillary objects
- define $f(r)$

Likelood Ratio threshold: The LR-threshold, LR_{th}, is the minimum value an ancillary object may score to be considered a counterpart candidate. There are different ways to compute LR_{th}, the simplest one is based on the reliability parameter in a assyntotic case: consider there is only *one* ancillary object within the *search radius* around a source; in this case we would expect such object to be the true source counterpart. Considering the Reliability parameter, (R) a probability score, $R_j = 0.5$ is the minimal (reasonable) value for such parameter so that the object can be considered a candidate. Using the definition of R above we should have:

$$0.5 = \frac{LR_{th}}{LR_{th} + (1 - Q)}$$

$$LR_{th} = \frac{0.5(Q - 1)}{-0.5}$$

$$LR_{th} = 1 - Q$$

Counterpart evaluation: Now that we have all the ingredients in place we may visit each target source and their ancillary candidates to evaluate each one.

For each source,

```
-> Loop over the respective ancillary sample:
   -> Evaluate each object's LR
   -> Remove objects with LR_j < LR_th
   -> Sum all ancillaries' LR_j
   -> Loop over all candidates:
      -> compute R_j
   -> Chose max(R_j) as counterpart
```

3.2.3 Comparison of matching results: GC versus MLE

To verify how the *minimum likelihood estimator* and *great circle* cross-matching results change we compare them using the data from the Swift DeepSky - Stripe82 (SDS82) and the CFHT-Stripe82 (CS82) catalogs.

The purely positional cross-matching considered a search radius of $rs = 6''$ around the target catalog (SDS82), this value corresponds to the size of the mean Swift-XRT point-spread-function.

From the original 2755 sources in SDS82 , 1105 found a match in CS82. The average distance between counterparts is 2.9″, with mode 2.6″. The standard deviation of the distribution is 1.5″.

Figure 3.6 presents the distribution of CS82 magnitude at hand, `MAG_AUTO`, from the GC matched sources (blue) and a background sample (yellow). The background sample include all sources within a radius $ro = 6 * rs$ around each SDS82 target source.

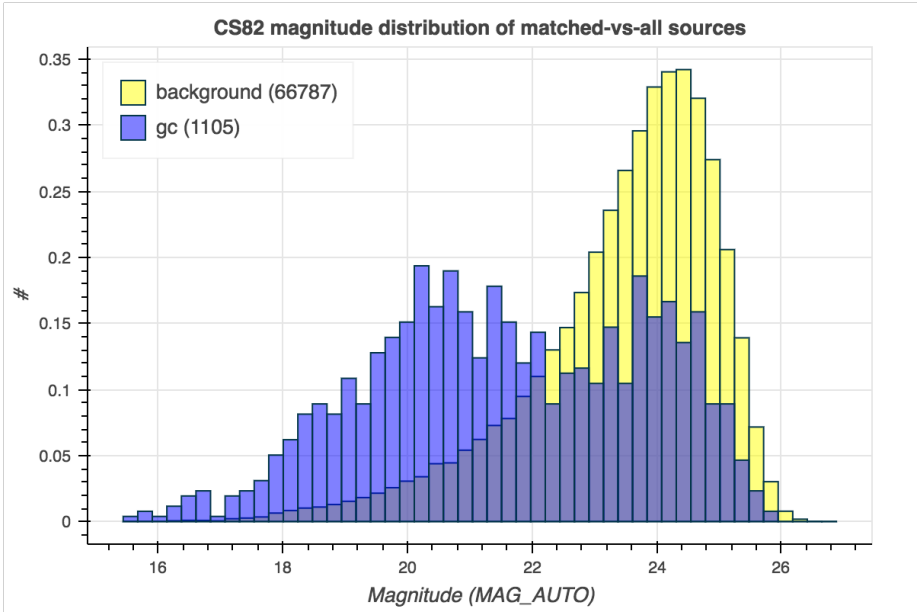


FIGURE 3.6: CS82 magnitude distribution of matched-vs-all sources

Likewise, the CS82 magnitude (`MAG_AUTO`) distribution from the MLE is shown for the *matched* sources in comparison to the *background* sample in figure 3.7.

We can see in figure 3.8 the distribution of x-ray flux from SDS82 from all sources with a matching counterpart (blue) and those without a counterpart (yellow).

Figure 3.9a shows the distribution of the *full* band flux of the 1105 matched sources and the 1659 sources that didn't find a pair. Figures 3.9b, 3.9c and 3.9d, shows the distribution of the soft, medium and hard x-ray bands fluxes, respectively, for *matched* and *non-matched* sources.

While 3.10a and 3.10b present the distributions of the hydrogen column `NH` and exposure time of the (non-)matched samples.

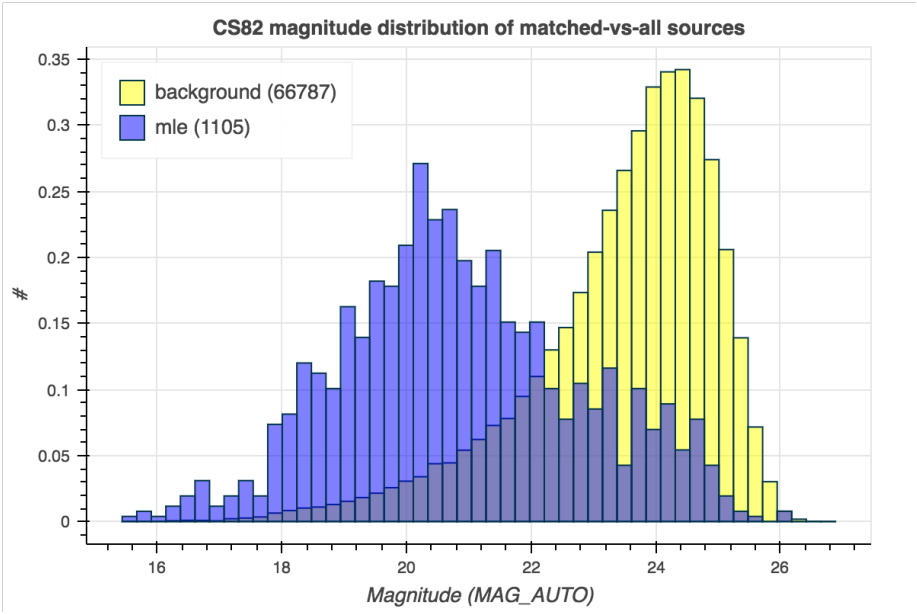


FIGURE 3.7: MAG_AUTO distribution for MLE cross-matched samples

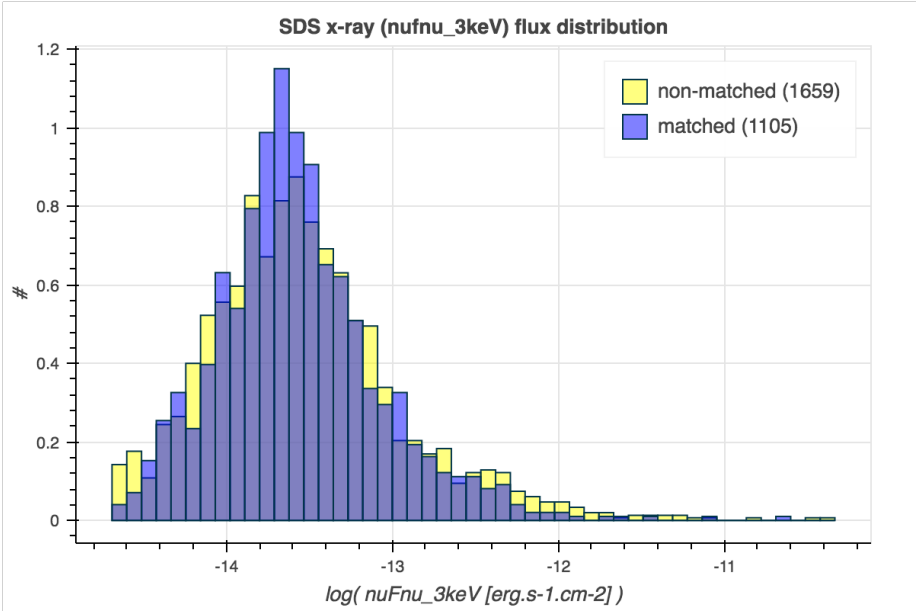
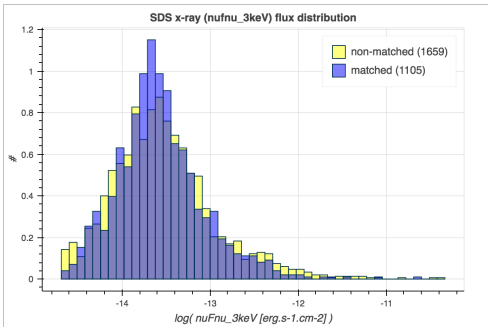
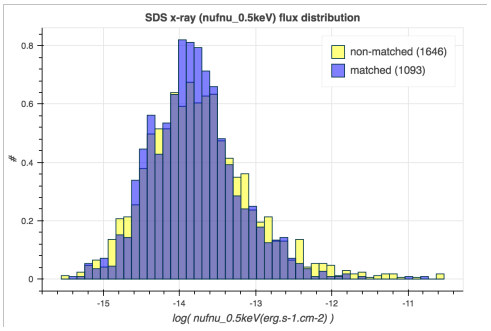


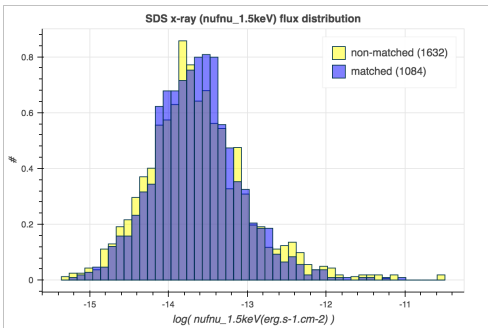
FIGURE 3.8: SDS82 X-ray flux distribution of matched and non-matched sources



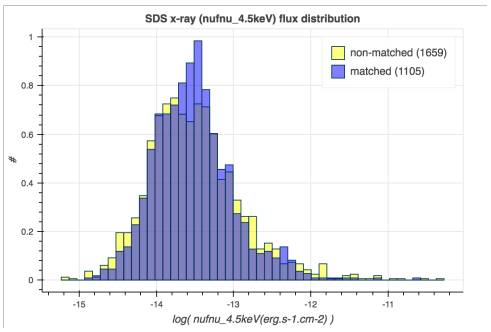
(A) Flux 3keV



(B) Flux 0.5keV

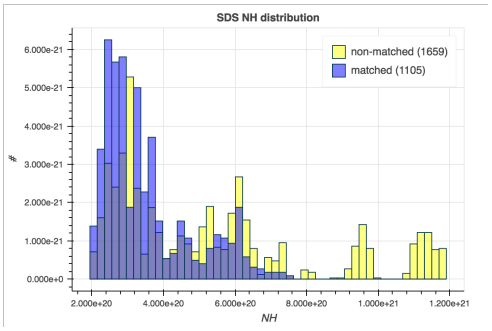


(C) Flux 1.5keV

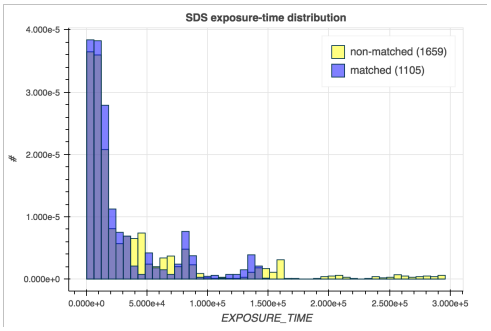


(D) Flux 4.5keV

FIGURE 3.9: SDS82 flux distributions for matched and non-matched sources



(A) NH



(B) Exposure-time

FIGURE 3.10: SDS82 distributions for matched and non-matched sources

Chapter 4

Brazilian Science Data Center

The Brazilian Science Data Center (BSDC) is a project to implement (*i*) a database for high energy astrophysics, (*ii*) multi-wavelength data exploration tools, (*iii*) an interface to reach out nearby communities (*e.g*, computer scientists). BSDC wants to bring a technological development work case to the hands of Brazilians and engage the country in the international discussion about high-level data access and exploration. The project grew out of this doctorate from the experience at the ASI Space Science Data Center and the discussions around Virtual Observatory and next-level data analysis interface.

BSDC is developing innovative tools for data analysis as well as easy-to-use, automated pipelines for data publishing. This PhD thesis carries the first set of services and technologies there in to provide distributed and seamless data analysis to the Brazilian community.

The nature of BSDC is based on the data access discussion (1.3) and intimately engaged with the Open Universe initiative (1.3.1). We, in Brazil, not only foresee the scientific benefits of a more fluent use of data but need to optimize the interfaces to education. Science in Brazil is done by a very small group of highly educated people, for the development of the society two things have to happen: the inclusion of more people in the technological and scientific (considering the argument in place) layer and (again, technological and scientific in this matter) production to be improved.

Concerning astronomical data, there are two kinds of users, data provider and data consumer. The former aims to publish data, make their data available in a standard format so that science can be done with it. Consumer is the user looking for data suitable for their science case.

A third kind of user, guided by their technological curiosity is the software (or systems) developer. This group is looking for challenging technological cases to solve.

BSDC is looking to provide interfaces for all those kinds of users not only because science can benefit from the different expertises working together but also because we understand the particular and appealing use case in our hands for their technical improvement.

In terms of the components, BSDC can be seen as follows:

- data : scientific products originating from observations and processing;
- services : interfaces to access resources, data, software and documents;
- pipelines : software for processing specific data input to output;
- infra-structure : underlying systems.

Guidelines

As BSDC develops its concepts around high-level data products, and the services to put those concepts in practice, it is understood that automation and monitoring are key concepts to our software systems. These concepts are important to deliver a more stable, smoother system as well as provide the bases to a scalable infra-structure.

From the accessibility point of view, the user interface is the most important aspect to focus on, the users have to interact with a simple, straightforward to have their job done.

In our collaboration with the VERITAS project we have implemented our first version of a fully automated system for data publication, where data is uploaded to our systems, validated and published in VO Spectra service and web interface. The goal is to keep an updated database of blazars spectra observed by VERITAS, delivered by the project whenever they have new data available.

As part of BSDC guidelines, to have the data published through VO is a requirement and part of the system was already in place from previous experiences; for instance, the MAGIC database. To automate the entire workflow – from data transfer, to processing, to its publication – and provide a seamless user interface were requirements I imposed to workout the underlying functionalities and to have more time to develop other interesting solutions.

Distributed collaboration

An key characteristic of the BSDC collaboration is it is distributed. Which is a trend in modern projects as internationalization has become (and is ever more) a common workspace.

What that means in practical terms is that we need to think of a system to support everyday work elements for a remote, distributed audience. In the case of BSDC, we may group the material elements in:

- software
- data
- documents

Portable software

One important aspect realized is to have the software that runs in the BSDC systems is to have it public and portable, meaning that users may have the software running on their own machines. That will help the software to be developed in cooperation with the users when they run it on their owns and – for those with software development knowledge – improve pieces of the software and give it back. It also helps to have software components in a modular design, independent from other components, still needed for the integrated solution but stable to work alone.

For example, the Assai tool (under development 4.3.2) is composed (as we see in figure 4.5) by *data query*, *database*, *transform data* and *visualization* blocks. The *data query* component depends only on a set of datasets, and the user may provide the data in whatever structure he wants (VO catalogs services or even CSV files), just having to fit the *database* interface expected; and doing so the user will have a simple tool to search and filter data in a given position of the sky. The same happens to the *transform data* and *visualization* components. For instance, the *visualization* software – which is mainly a plot canvas with extra capabilities – may be user in its own, it does not need to know what is the software stack above it, but only need a table with data values.

That can be accomplished by developing software components based in (*i*) high-level languages (Python, for instance), (*ii*) well documented and structured software and (*iii*) adopting packaging systems that allow greater flexibility and portability. To the latter, in particular, we may adopt new technologies like Docker containers (see 4.1.1) to allow *zero-installation* software packages.

4.1 Software solutions

During this work a set of software packages were developed to handle everyday data analysis tasks on data retrieval, high-energy analysis, as well as data publication. The ultimate goal behind each software was to provide a high-level, easy-to-use interface.

Each development exercised a different aspects of the science and technology behind computational astrophysics as we see key points to leverage the data accessibility and data analysis production we are engaged in.

For instance, the `docker-heasoft` package applied a novel technological approach to for portable packaging of a rather complex toolkit. In `xmatch` we implemented a non-trivial catalogs cross-match algorithm using likelihood ratio. While `eada` is a utility for VO conesearch services, astronomer's everyday task.

All tools are publicly available through the Github platform:

- docker-heasoft [1]: HEASoft package with Docker containers;
- xmatch [2]: cross-match using Maximum Likelihood Estimator;
- eada [3]: VO conesearch services discovery and catalogs query.

4.1.1 Linux containers for science

Linux containers provide an environment close to a full Linux operating system, like in a Virtual Machine but without the overhead of running an independent kernel and simulating the hardware. A single kernel, from the host system, can run any number of containers – just like it would with bare applications. Particularly interesting about containers is that they run isolated from the other containers (if any) and applications, in a "black box". Figure 4.1 presents the difference between a virtual machines setup from containers, and the isolation between the components.

Linux containers were made possible after the Linux kernel cgroups[4] feature. Cgroups provide isolation of group of processes resources, to simplify their accounting and management. Containers then make use of such feature to run application bundles independent of other packages from the host system, only the kernel and allocated resources.

[1] https://github.com/chbrandt/docker-heasoft
[2] https://github.com/chbrandt/xmatch
[3] https://github.com/chbrandt/eada
[4] https://www.kernel.org/doc/Documentation/cgroup-v1/cgroups.txt

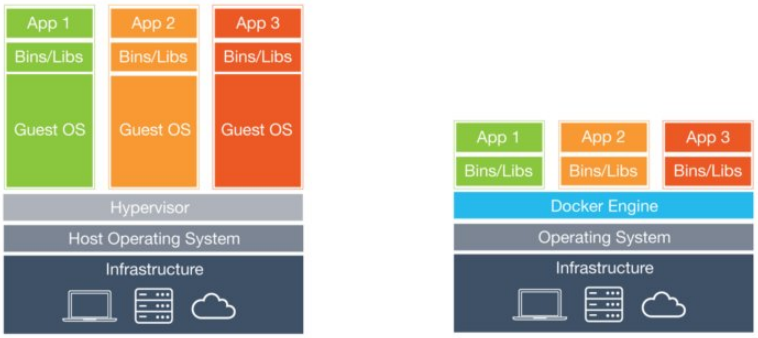


FIGURE 4.1: Difference between VMs and containers.
Credit: *Docker Inc.*

Containers are broadly used in the industry as a highly scalable solution and efficient deploy mechanism. In our case, as scientists, we saw containers as a practical solution for sharing complex setup packages. Another great application for containers is the archival – an executable archival – for old applications, software dependent on outdated dependencies.

Docker. A particular development of the containers technology was done by Docker Inc[5]. Docker containers got a faster acceptance by the information technology community by providing a easy to use interface, compatibility with Windows and MacOS, and high-level services such as cloud repositories.

Of particular interest for us is the *portability* feature of Docker. Docker allows a package to run in all major platforms: Linux, Windows and MacOS. Which gives a new level for the concept of portability and significantly reduces the costs of software development.

I experimented containers with a mixture of applications, archival, web servers, graphical and scientific pipelines. All of them available at the public repository DockerHub[6]. In what follows I present the packaging of HEASoft[7] and DaCHS[8] using docker containers as a simple and easy-to-use solution for scientific packages distribution.

HEASoft

HEASoft is a package for High Energy Astrophysics data analysis. The package provides a wide range of tools to process x-ray and gamma-ray data

[5]https://www.docker.com/what-container
[6]https://hub.docker.com/u/chbrandt/
[7]https://heasarc.gsfc.nasa.gov/lheasoft
[8]http://dachs-doc.readthedocs.io/

from raw data reduction, to model fitting, to visualization. It is composed by many individual components developed by many scientists and provided as a package by NASA. The programming languages used in each tool may be different, the source codes are written in Fortran, C, C++, Python, Perl and others, which make the setup of the package complex and may be troublesome for some users.

Providing a harsh experience for the users at the very beginning may be a *show stopper*, certainly to be something to avoid. Another drawback from heterogeneous and complex packages is the risk of not compatible to a particular operating system as the packages dependencies are usually many and should comply all together in each system.

Pre-compiled packages and virtualization are two features that Linux containers bring together to provide a light solution for software portability.

To overcome the setup of HEASoft and CALDB we applied the Docker containers technology to provide a *ready-to-use* HEASoft version.

The `heasoft` container setup was designed to work seamlessly from the user perspective. The user may run it as any other container and after its instantiation find himself inside it and operate normally. Or, the user may run a `install` script that exposes HEASoft tools – like `nh`, `ximage`, for example – to the host system, *faking* the real location of that binary: it will work as if it was installed in the host system, but in reality it will run from inside the container.

Figure 4.2 depicts the layers between the system's command-line interface and the container: the *entrypoint* layer parse and adapt the environment accordingly to the arguments given by the *docker-interface*, on the other hand, the interface knows about the container internals so that the command given to the container is reasonable.

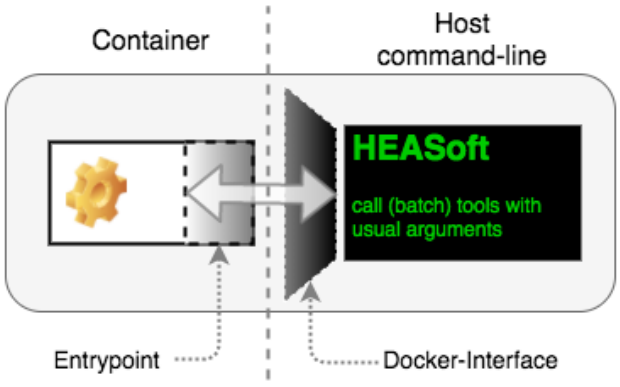


FIGURE 4.2: Docker HEASoft container abstraction layers

The `heasoft` container can be found at:

- https://hub.docker.com/r/chbrandt/heasoft/.

The latest version contain HEASoft version `6.15`, as soon as the bundle finds its way to a broad audience, all versions of HEASoft will be released in the very same format.

The recipe and all the details to build the container (including the automated setup of HEASoft and CalDB themselves) and the `install.sh` script cited above to *abstract* host system binaries can be found in the public Github repository[9].

DaCHS

DaCHS (Demleitner et al., 2014a) is a Virtual Observatory (VO) publication service developed by the German Astrophysical Virtual Observatory (GAVO). DaCHS is a high-level astronomical data base manager compliant with IVOA standards, implementing stable protocols such as VO's Simple Conesearch (SCS) and Simple Spectra Access (SSA).

DaCHS uses Postgres to store and manage one or more datasets, each being published through a intuitive web interface and/or one or more VO services. For instance, the Brazilian Science Data Center (BSDC) publishes a set of astronomical catalogs through SCS and SSAP services, as well as HTTP interface, using DaCHS.

Publishing a dataset through DaCHS involves the definition of a *Resource Descriptor*, which is a XML file containing metadata and settings for the appropriate publication services for the corresponding data set. Such file is used for ingesting data in Postgres as well defining the interfaces to access such data.

DaCHS is a remarkably stable software, packaged for Debian Linux servers the setup on those systems is straightforward. One aspect of DaCHS that may be troublesome, specially for newcomers, is the definition of the RD files as they are powerful in abstracting the database management but complex in details.

In the case of DaCHS, docker containers were used to provide portability to other systems, like MacOS, and an efficient sandbox environment to support new data resource setup process: build, test, adjust, test. Very efficiently, as fast as the usual DaCHS daemon would take to launch, the container version instantiated and used for tests however necessary without any risk of

[9] https://github.com/chbrandt/docker-heasoft

damaging the other datasets that eventually be running on the host server. The DaCHS container will respond to HTTP requests through (default) port 80 as it would from a bare system.

The `docker-dachs` container is publicly available at:

- https://hub.docker.com/r/chbrandt/dachs/

The setup of the DaCHS container is a bit different and is provided in two different flavors: one that runs Postgres and the DaCHS server in the same container instance, the other will run two container instances – Postgres and DaCHS – bound by `compose` setup, which orchestrate them to work together. The first option – both services in the same container instance – was designed to be used as a simple test environment for dataset schemas. The `compose` option may be used when the DaCHS server only needs to be configured or tested while the database stays untouched.

Figure 4.3 depicts the schema and user interaction for the `docker-dachs` container.

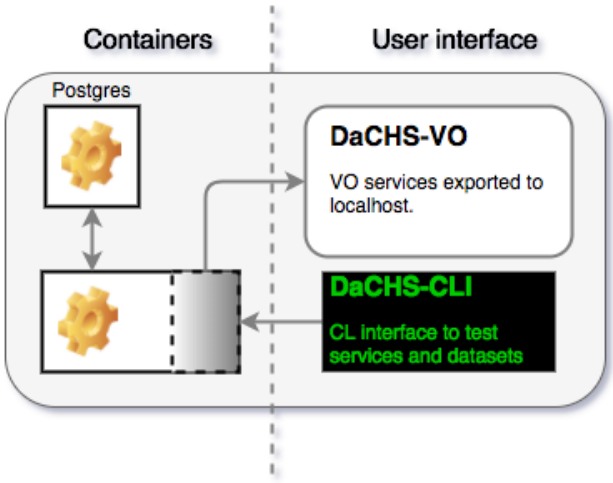


FIGURE 4.3: Docker-DaCHS containers interaction

4.1.2 EADA

The *External Archives Data Access* (EADA) is a tool to search for catalogs in the Virtual Observatory (VO) network and download their data. The tool is written in Python and provides a command line interface to ease its use and provide wider usability.

EADA is built on top of PyVO, a python library implementing lower level communication interfaces between VO services In EADA I implemented the

interface to handle users queries for specific data types: flux, position, waveband, etc.

EADA provides three services to search VO resources:

- `servsearch`: searches for *conesearch* services
- `conesearch`: searches for data (aka, objects) in *SCS* services
- `specsearch`: searches for data in *SSAP* services

Simple Conesearch Services (SCS) is a VO protocol to provide a simple interface to positional queries. Cone searches is the most popular operation we do in astronomical research, given a (R.A, Dec.) position in the sky and a surrounding area around it (radius) we get the information about the object(s) found. A SCS services will retrieve zero, one or more entries from the database according to the number (of objects) found in the specified search region.

The VO Simple Spectra Access Protocol (SSAP) offers a similar query interface the data retrieved is a table of spectral data (*e.g*, frequency and flux) for each object found in the database.

4.2 Very High Energy data publication

One of the BSDC mission is the provision of a science-ready Very High Energy (VHE) dataset. We are building it through automated pipelines taking care from the data acquisition to data publication. The *publication pipeline* goal is to create an easy interface to VHE projects (*e.g,* VERITAS collaboration) for the publication of their data, and make such process as efficient and informative as possible through fully automated workflows.

The basic idea of the publication pipeline is to have the three conceptual blocks orchestrated to have an homogeneous interface for data reception, processing, and data publication:

1. input data files interface;
2. verify and transform to *science-ready* format;
3. publish the data through VO services.

Where input data files may change from project to project, but not necessarily the verification or publication stages. Likewise for the second and

third blocks. To have this design established is important to allow stable extensions of the service as well as consistent monitoring of each step.

The pipeline is in operation for the VERITAS (Weekes et al., 2002) collaboration in a fully automated way. A first version of the pipeline has been implemented and for MAGIC (Cortina, 2005) data, this one working in a semi-automated model. In each case they have input data files in different formats and transformed to homogeneous data formats at the end, for publication through VO Simple Spectral Access Protocol (SSAP).

When we talk about *science-ready* data here we mean *(i)* data in useful physical units – flux data in $erg\ s^{-1}cm^{-2}$ –, *(ii)* data easily accessible – VO SSAP service and web page –, *(iii)* transparency regarding data origin and processing. Those details are carefully carried out in our pipelines by publishing complete information on each spectrum and also linking each resource to external and high-level tools like the ASDC SED tool.

An important feature of the output is to include the reference article where the corresponding data has been published. This is important so that the user accessing that data can have a direct way to access the scientific details behind the data. And that is one of the reasons MAGIC and VERITAS were used. Another reason is the way the (spectral) data and metadata arrive to our systems: the MAGIC data is periodically downloaded from MAGIC webpage, while the VERITAS collaboration upload their data to our archive. This difference allowed us to develop solutions to deliver the same output format for two quite different outputs, teaching us important lessons on the development of autonomous systems for data processing. The next sections detail the implementation of each solution.

VERITAS publication pipeline

The data transfer – input – from VERITAS spectra to BSDC server is done through a dedicated channel provided by the Syncthing files synchronization service. Syncthing keeps a secure channel alive between two servers, that effectively appears as a directory in the servers file system hierarchy. Whenever new data has to be transfered it is sufficient to copy the corresponding file to the configured directory and it will be sent to BSDC's server.

When a new data file is transfered, an existing file is updated, or a file is removed a action is triggered in our data server to apply an according modification to BSDC's database.

The processing step is responsible to transform the data file received to a format suitable to be ingested in our database and ultimately publish the

spectral information and its metadata as a VO SSAP service. During the formatting process, data and metadata are verified following an standard predefined between VERITAS and BSDC. The input data file format can be seen in appendix A.

Every modification in the data set is tracked by a version control system, Git[10], so that information is never lost; We also use it as backup utility, since we keep a remote server in live sync after each modification. If by any reason the processing step fails to complete – because of missing metadata, for example – the processing log is promptly given back to the data provider through the same channel spectra data is transfered.

Finally, the database and exposed spectra table through VO-SSAP, as well as a web interface[11] are updated in real time. The data is published through the GAVO-DaCHS[12] server.

Figure 4.4 gives a graphic representation of the described system. The system was implemented using Bash and Python programming language, source code is publicly available at the github repository:

- https://github.com/CBDC/veritas

MAGIC publication pipeline

The publication pipeline for MAGIC spectra works in an *active* fashion, as we define it. The pipeline periodically access MAGIC-PIC webpage[13], where data files and articles are published, parse the page to collect the metadata for each entry and download the file linked. Then each set of data – file and metadata – goes through the processing step where it will be verified and transformed to, eventually, be published through our VO service at *http://vo.bsdc.icranet.org/magic/q/web/form*.

Unlike VERITAS data, MAGIC spectral measurements do not follow a standard regarding their units, flux values may be in $phTeV^{-1}s^{-1}m^{-2}$ or $ergs^{-1}cm^{-2}$, for example. Some spectra flux are in $phs^{-1}m^{-2}$, which prohibit an automatic units conversion. And in other cases the metadata indicates a suitable physical units, but values are found to be either out of scale or, apparently, different units actually.

[10] https://git-scm.com/
[11] http://vo.bsdc.icranet.org/veritas/q/web/form
[12] http://dachs-doc.readthedocs.io/
[13] http://vobs.magic.pic.es/fits/

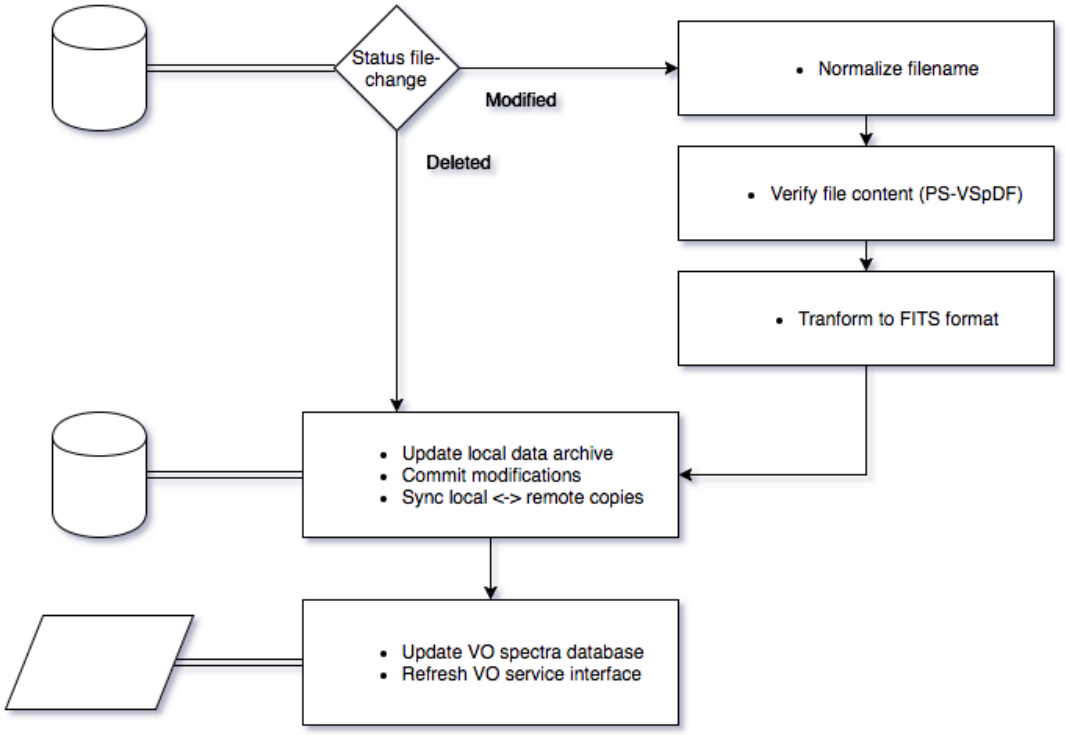


FIGURE 4.4: VERITAS processing workflow

4.3 Tools currently in development

4.3.1 UCDT: handling IVOA UCDs

Unified Content Descriptor (UCD[14]) is a low-level data descriptor for astronomical data. UCDs are used to describe data sets using a specific dictionary and grammatic in such a way that UCDs can be understood by machines.

In a catalog, column's metadata is typically composed by:

Metadata field	Metadata content
name	a unique name for a column in the table
description	a short explanation of column's content
unit	the physical unit the data is represented
type	the data type (*e.g*, numeric, string, boolean)
null	value use for non-existent entries

From the above list, it is reasonable to say the most important metadata field is `description` as it explains what is a column's content about. Considering the description is complete, is is possible for an astronomer to – with

[14]http://www.ivoa.net/documents/latest/UCD.html

some extra work – infer the other metadata fields. Clearly, for technical reasons and robustness of a dataset (*i.e*, catalog), all the metadata fields should be filled.

IVOA proposed this extra metadata field called `ucd` to accomplish machine-level understanding of the data. As it is the purpose `description` for humans, `ucd` is meant to (ideally) uniquely describe a column's content in a low-level so that data inspection can be done automatically by software.

The `ucdt` tool, available at `://github.com/EmptySpace/ucdt'`, creates an interfaces to access elements within a dataset (*e.g*, columns in a catalog) through UCDs, instead of the traditional "column name" based. Which means, astronomers can access a data column by its column meaning, like "position" or "flux", instead of a variable, dataset specific column name.

`ucdt` is implemented in python and can be used as an API and as such be embeded in general data handling routines. The library implements a tree structure to support the hierarchical nature of UCDs. For example, suppose we have a table from which we want to access the "photometric columns". (There are many specifications of "photometric measurements", in astrophysics and UCDs do consider that.) The table contains among other columns:

index	Name	UCD
0	Mag_x	`phot.mag`
1	x_flux	`phot.flux`
2	nufnu_x	`phot.flux.density`

In `ucdt`, such columns are represented as:

```
(Photometry)
- phot
| - flux : 1
| | - density : 2
| - mag : 0
```

And by querying `ucdt` for the "flux density" data, the position `[2]` of the respective column comes to our knowledge, while if we want all photometric columns, we can ask equally simple:

```
>>> UCDT.get('phot.density')
[2]

>>> UCDT.get('phot')
[0, 1, 2]
```

UCD standard recommends the the extension of the pre-defined dictionary to accomplish special cases not defined in the IVOA official vocabulary. In this aspect, the official list of words (see [15]) is said to belong to the namespace "`ivoa`". The UCD dictionary can be extended under other namespaces. The `ucdt` package allow the addition of other namepaces to the environment by using a plugin-style set of dictionaries. This features allows very flexible uses of the UCD metadata to handle particular or sensible data mining scenarios.

4.3.2 Assai: a portable SED builder

To visualize spectral energy distributions (SEDs) we developed the Advanced Spectral Analysis Interface (ASSAI) (pronounced as the brazilian-portuguese word *açaí*). *Assai*is a so called *SED builder* software, it is being developed in Python and Javascript for interactive data visualization through web browsers.

Currently the tool implements the query for flux data in VO conesearch (*scs*) services, its visualization and the interactive selection for subsequent analysis.

The workflow (or interactive actions) `Assai` is currently designed to accomplish are:

- define the conesearch services to query/retrieve data
- search a given region of the sky (input: RA, Dec, radius)
- plot $\nu F\nu$ flux *vs* frequency
- select/unselect data points
- download SED data and catalogs metadata

The primary goal of BSDC's SED builder is to be an exploratory tool. So it must provide the output (*i.e*, datapoints) through a graphical and interactive interface as well as provide the mechanism to upload/download datasets. As part of BSDC's guidelines, the tool must be VO-compliant. Figure 4.5 depicts the schema of *Assai*.

- Input, `object name/position` : If the user has given an object name, it has to be translated into the corresponding coordinates.

[15] http://www.ivoa.net/documents/latest/UCDlist.html

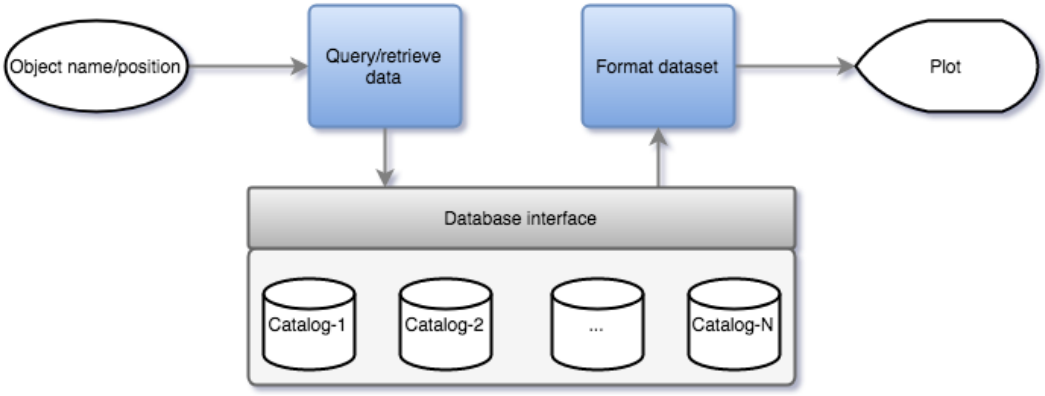


FIGURE 4.5: *Assai*software design schema

- `Query/retrieve data` block: ask the database for all entries around a (RA,Dec) position, select columns with flux measurements. Data returned may have zero or more datapoints.
- `Format dataset` block: transform each catalog's dataset into a regular table of *frequency*, *flux*.
- Output, `data visualization` : present the datapoints in a *flux-vs-frequency* interactive plot. The user can select/unselect datapoints, access per-point information like source catalog or angular separation from input coordinates, and finally download the set of datapoints and respective catalogs metadata.

Database interface

The database interface is the layer a tool (*e.g*, the SED builder) communicates with to query/retrieve data. The interface is meant to abstract the underlying database implementation, allowing each catalog to have its data stored in different format or technologies.

The catalogs `(1, 2, ... N)` in figure 4.5 should be considered services, independent from each other. The interface knows how to query each one through a standard set of settings, where the settings for each catalog are organized in a configuration file. Each catalog's configuration file and auxiliary files are placed in their own (home) directory.

The main type of catalog service we want to work with is IVOA's Simple Conesearch Services (SCS), where inclusion or removal of a SCS catalog is as simple as including (or removing) a configuration file in a structure of directories. But each catalog service can have its data stored in different technologies. For example, out of 10 catalogs services we could have one using

a relational DBMS, another one a NoSQL system, a third saved in a CSV text file, and the remaining seven using external VO services.

Independent of the way data is managed in each catalog internals, the interface is homogeneous. Each service should provide the following interface methods and metadata:

- `search` : given a sky coordinate and search radius, return zero or more matching entries, with their flux measurements;
- `frequency` : given the column name(s) associated with the flux measurement(s), return the corresponding frequency;
- `metadata` : return the metadata associated to each column.

This design simplifies the overall data base maintenance and improves its (horizontal) scalability.

In such schema, a VO catalog service is defined by the layout in figure 4.6.

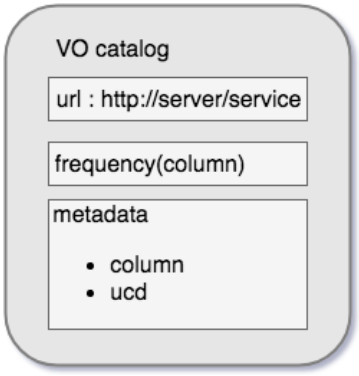


FIGURE 4.6: VO catalog service for SED builder

Chapter 5

Conclusion

In this work we have addressed some issues of data accessibility by using the SDS82 catalog and its application to the search for blazars as a science case. This thesis provided practical elements to the development of the *Brazilian Science Data Center* (BSDC) and the *Open Universe* initiative (OUN) through the implementation of VO[1]-compliant data access software.

This thesis offers to the astronomical community (*i*) a dynamically updated deep X-ray catalog, kept up-to-date by a collaborative model for ever the Swift satellite operates; (*ii*) an automated Swift-XRT data analysis pipeline implementing a technological model focused on portability and the user interface; (*iii*) a software infra-structure model to portable data analysis tools. Particularly interesting to the Brazilian community, we established a close collaboration with the VERITAS collaboration through the implementation of a fully automated data publication pipeline, also applied to the MAGIC public spectra database.

The SDS82 [2] catalog is the product of the *Swift-DeepSky* pipeline we developed to identify X-ray sources and measure their fluxes in co-added images of the Swift-XRT instrument. The pipeline outputs (`.gif`) images and (`.csv`) tables for the detected sources and their measurements in reasonable units (*cgs*) ready to be used in a Spectral Energy Distribution (SED), for instance. The catalog was created by processing *all* images of the Swift XRT archive within the Stripe-82 region ($\approx$ 1% of the sky) – since the satellite started operating (2004) up to 2018. Finally, we applied the catalog to the search of blazars in the Stripe82 using an algorithm based on SEDs for automated selection of possible blazar candidates, the VOU-Blazars (Y. Chang, in preparation). The catalog played a unique role to 69 (out of 300) blazar candidates for being their unique X-ray counterpart, from which we found 17 promising

[1] Virtual Observatory
[2] http://vo.bsdc.icranet.org/sds82/q/cone/info

HSP blazar candidates; eventually added to the 3HSP list of blazars (Chang and Giommi, in preparation).

The *DeepSky* [3] pipeline uses Swift-XRT observations to autonomously detect sources and provide flux measurements on four bands – `Full` (0.3 < E(keV) < 10), `Soft` (0.3 < E(keV) < 1), `Medium` (1 < E(keV) < 2), `Hard` (2 < E(keV) < 10) – for a given region of the sky. The analysed images are the combination of *all* XRT Photon Counting images (and respective exposure-maps) that overlap in the region of interest to achieve the deepest possible view of the field. In the case of SDS82 catalog we were able to recover fluxes for 2755 sources as low as $\sim 10^{-16} erg\ s^{-1}\ cm^{-2}\ \nu F_\nu$ fluxes.

Selection and download of the images are carried out by the pipeline alone so that the user does not have to interact with the Swift data archive neither understand its, rather technical, structure. The results of the processing are simple, commonly used files to allow their reading by everywhere available tools (*e.g,* text editors). *Par défaut*, the results will also be uploaded to a dedicated BSDC server where they will be automatically published in the VO network.

In the data access to the Virtual Observatories network we developed the EADA tool to discover conesearch services based on IVOA resources UCDs, physical units and possibilly keywords of particular interest. The tool has been used in the VOU-Blazars pipeline to retrieve multi-wavelength data for the discovery of HSP blazar candidates in the building of the 3HSP catalog. The tool is also responsible for the access to VO catalogs in the SSDC SED[4] tool. Among the catalogs the SED tool makes use of we will find the catalogs published by BSDC-VO[5] portal, implemented as part of this thesis with the support of the German Astrophysical Virtual Observatory (GAVO) DaCHS system.

In the set of catalogs publishes by BSDC, the VERITAS spectra database are the result of an automated publication pipeline[6] developed to directly connect the VERITAS collaboration servers to BSDC servers and make the publication of their high energy data to the VO network a seamless process. To that end, we designed a standard data format in agreement with the Gamma-ray data formats[7] and an stable upload interface. The system

[3] https://github.com/chbrandt/swift_deepsky
[4] https://tools.asdc.asi.it/SED/
[5] http://vo.bsdc.icranet.org/
[6] https://github.com/CBDC/veritas
[7] https://gamma-cat.readthedocs.io/

implemented is restrictive to badly formatted data files to allow only minimally qualified data to eventually go public through our VO SSAP service.

Finally, this thesis also presented side products that support the main results here discussed. The xmatch tool implements the non-trivial Maximum Likelihood Estimator algorithm. Docker containers to leverage the use of software with non-trivial install process or in test-production systems, to which we make the case for the HEASoft package and the DaCHS system.

Final words

The data accessibility discussion is a long – and probably endless – road with many branches along the way; it is paved with different materials, but no holes in it – as far as I can see. We took the branch of data processing interface, and helped building with automation and lowering some barriers across the data analysis workflow. The works that remained on hold for practical reasons will soon be retaken by me and the Brazilian Science Data Center, under the auspices of the Open Universe initiative.

It is particularly important – besides interesting – we are engaging Brazil in such high-level discussions about the technology, science and education surrounding astronomical data; Especially by actively participating in efforts such as Virtual Observatories and Open Universe. We feel we have a great opportunity, again, to put the country in the first line of the discussion and help boosting our educational and economical standards.

Appendix A

BSDC-VERITAS spectra data format

This document presents what came to be the ‘version 3‘ of VERITAS-BSDC data format. The data format present below is the result of an interactive process between VERITAS and BSDC trying to accomplish easiness of use and standards adopted in the gamma-ray community[1].

The progress of this process can be seen in the previous alike documents:

- data format - v2, presented below in section A.1.1
- data format - v1[2]

A.1 Data format - v3

After the last interaction (section A.1.1), some modifications over metadata keywords were applied to accomplish the processing data files need to follow.

The modifications proposed here were motivated mainly to keep metadata as clear and clean as possible; some changes were motivated by the processing of data itself – for the data is transformed to FITS before being published.

After the example below, the data format structure – without content – is proposed for a better understand of what is essential and what is not in such version of the format.

Example

The following example is the same (data file) used in the previous document, the `Mrk421_2008_highA` observation(Acciari et al., 2011) file:

[1] https://gamma-cat.readthedocs.io/
[2] https://github.com/CBDC/veritas/blob/master/docs/data_formatting-v1.rst

```
# %ECSV 0.9
# ---
# meta: !!omap
# - OBJECT: Mrk 421
#
# - DESCRIBE:
#    Spectral points for multiwavelength campaign;
#    Observations taken between 2008 January 01 and 2008 June 05;
#    Flux sensitivity 0.8e-10 < flux(E>1TeV) < 1.1e-10
#
# - MJD:
#    START: 54502.46971 # unit=day
#    END: 54622.18955   # unit=day
#
# - ARTICLE:
#    label: Ap.J. 738, 25 (2011)
#    url: http://iopscience.iop.org/0004-637X/738/1/25/
#    arxiv: http://arxiv.org/abs/1106.1210
#    ads: http://adsabs.harvard.edu/abs/2011ApJ...738...25A
#
# - COMMENTS:
#    Name: Mrk421_2008_highA
#    Tag: highA
#    Redshift: 0.031
#    LiveTime: 1.4  # unit=hour
#    Significance: 73.0
#
# - SED_TYPE: diff_flux_points
#
# datatype:
# - name: e_ref
#   unit: TeV
#   datatype: float64
# - name: dnde
#   unit: ph / (m2 TeV s)
#   datatype: float64
# - name: dnde_errn
```

```
#   unit: ph / (m2 TeV s)
#   datatype: float64
# - name: dnde_errp
#   unit: ph / (m2 TeV s)
#   datatype: float64
#
e_ref dnde       dnde_errn  dnde_errp
0.275 1.702E-005 3.295E-006 3.295E-006
0.340 1.289E-005 1.106E-006 1.106E-006
0.420 8.821E-006 6.072E-007 6.072E-007
0.519 5.777E-006 3.697E-007 3.697E-007
0.642 3.509E-006 2.351E-007 2.351E-007
0.793 2.151E-006 1.525E-007 1.525E-007
0.980 1.302E-006 1.024E-007 1.024E-007
1.212 6.273E-007 6.117E-008 6.117E-008
1.498 3.310E-007 3.853E-008 3.853E-008
1.851 1.661E-007 2.401E-008 2.401E-008
2.288 1.124E-007 1.732E-008 1.732E-008
2.828 6.158E-008 1.138E-008 1.138E-008
3.496 3.347E-008 7.427E-009 7.427E-009
4.321 1.160E-008 4.031E-009 4.031E-009
5.342 5.230E-009 2.371E-009 2.371E-009
```

The format (v3)

In the following, consider `value` between '< >' as the value to be substituted. Notice the indentation, it is essential for parsing the information correctly.

```
# %ECSV 0.9
# ---
# meta: !!omap
# - OBJECT:   <name of the object>
#
# - DESCRIBE:
#    <multiple line description of the data>
#    <free-form content; just has to follow the block indentation>
```

```
#
# - MJD:
#    START: <start of observation in 'mjd' (unit:days)>
#    END:   <end of observation in 'mjd' (unit:days)>
#
# - ARTICLE:
#    label: <bibcode or alike>
#   url:   <any url important for the user to understand the data>
#    arxiv: <if published, the article's arXiv url>
#    ads:   <if published, the article's ads reference url>
#
# - COMMENTS:
#    Name:          <a label, typically the file rootname>
#    Tag:           <a short, contiguous label>
#    Redshift:      <z>
#    LiveTime:      <observation time in hours>
#    Significance:  <significance value>
#
# - SED_TYPE: diff_flux_points
#
# datatype:
# - name: e_ref
#   unit: TeV
#   datatype: float64
# - name: dnde
#   unit: ph / (m2 TeV s)
#   datatype: float64
# - name: dnde_errn
#   unit: ph / (m2 TeV s)
#   datatype: float64
# - name: dnde_errp
#   unit: ph / (m2 TeV s)
#   datatype: float64
#
e_ref dnde dnde_errn dnde_errp
<...> <...> <...> <...>
```

A.1.1 Data format - v2

After a first proposal for VERITAS data (spectra, for instance) files format, we evolved the format to include some features from SED data format for gamma-ray astronomy[3].

SED types

One important feature noticed is the inclusion and adoption of standards established by the gamma-ray astronomy effort on data format, in particular, in what regards the spectrum/SED files.

VERITAS *spectrum* files provide *differential flux* measurements; columns represent:

- the 'energy' from which flux measurement is referred to
- the 'differential flux' measured
- the 'asymmetric negative flux error'
- the 'asymmetric negative flux error'

According to those standards, we can improve the use of such data by including the keyword `SED_TYPE` in the file headers. The standard proposes the value `dnde`, we found the below value more informative:

```
SED_TYPE = diff_flux_points
```

And columns go labeled after the standard:

- `e_ref`: the reference measurement energy
- `dnde`: differencial flux values
- `dnde_errn`: negative flux error
- `dnde_errp`: positive flux error

ECSV format

The file format ECSV (version 0.9), as proposed by Astropy's APE-6[4] is to be used as it is a good compromise for human readability and metadata- rich format for lightweight data files.

[3] https://gamma-cat.readthedocs.io/
[4] https://github.com/astropy/astropy-APEs/blob/master/APE6.rst

The *Extended CSV* lies over YAML[5] data (serialization) format. In what follows, it is described the new proposal for data format to be used with VERITAS spectra; an example file is taken as example.

The v2 format

Let us consider the file for Markarian 421 in a high state between '2008-02-06' and '2008-06-05' (Acciari et al., 2011).

```
# \%ECSV 0.9
# ---
# meta: !!omap
# - object: Mrk 421
#
# - description:
#    Spectral points for multiwavelength campaign;
#    Observations taken between 2008 January 01 and 2008 June 05;
#    Flux sensitivity 0.8e-10 < flux(E>1TeV) < 1.1e-10
#
# - mjd:
#    start: 54502.46971
#    end: 54622.18955
#
# - article:
#    label: Ap.J. 738, 25 (2011)
#    url: http://iopscience.iop.org/0004-637X/738/1/25/
#    arxiv: http://arxiv.org/abs/1106.1210
#    ads: http://adsabs.harvard.edu/abs/2011ApJ...738...25A
#
# - comments:
#    - Name=Mrk421_2008_highA
#    - z=0.031
#    - LiveTime(h)=1.4
#    - significance=73.0
#
# - SED_TYPE: diff_flux_points
#
```

[5] http://yaml.org/

```
# datatype:
# - name: e_ref
#   unit: TeV
#   datatype: float64
# - name: dnde
#   unit: ph / (m2 TeV s)
#   datatype: float64
# - name: dnde_errn
#   unit: ph / (m2 TeV s)
#   datatype: float64
# - name: dnde_errp
#   unit: ph / (m2 TeV s)
#   datatype: float64
#
e_ref dnde       dnde_errn  dnde_errp
0.275 1.702E-005 3.295E-006 3.295E-006
0.340 1.289E-005 1.106E-006 1.106E-006
0.420 8.821E-006 6.072E-007 6.072E-007
0.519 5.777E-006 3.697E-007 3.697E-007
0.642 3.509E-006 2.351E-007 2.351E-007
0.793 2.151E-006 1.525E-007 1.525E-007
0.980 1.302E-006 1.024E-007 1.024E-007
1.212 6.273E-007 6.117E-008 6.117E-008
1.498 3.310E-007 3.853E-008 3.853E-008
1.851 1.661E-007 2.401E-008 2.401E-008
2.288 1.124E-007 1.732E-008 1.732E-008
2.828 6.158E-008 1.138E-008 1.138E-008
3.496 3.347E-008 7.427E-009 7.427E-009
4.321 1.160E-008 4.031E-009 4.031E-009
5.342 5.230E-009 2.371E-009 2.371E-009
```

The format is organized as follows.

Mandatory directive, as the very first two lines of the file:

```
# %ECSV 0.9
# ---
```

As well as `# datatype:` (below), `meta` are mandatory (first-level) collections.

```
# meta: !!omap
```

In fact, `meta` and `datatype` are the only two first-level blocks that ECSV-0.9 accepts. Notice the argument `!!omap`, it is a mandatory *tag* (in yaml's jargon) for Astropy to succeed in reading it (probably a bug).

'meta' section

Begin of 'meta' section.

```
# meta: !!omap
```

Notice that '!!omap' is mandatory.

```
# - object: Mrk 421
```

'object' is the object's designation. The name of the object is meant to be used to cross-correlate with other databases and as such must be broadly recognized. The 'object' name should be recognized by Simbad[6].

```
# - description:
#    Spectral points for multiwavelength campaign;
#    Observations taken between 2008 January 01 and 2008 June 05;
#    Flux sensitivity 0.8e-10 < flux(E>1TeV) < 1.1e-10
```

'description' is a free-form paragraph used to briefly describe the content of the file.

```
# - mjd:
#    start: 54502.46971
#    end: 54622.18955
```

The (spectra) data points reported do not pursue a MJD (or MJD-range). Instead, data points are grouped (in data files, like this one) according to the object's activity and the period of time. That said, the `mjd` (values, mjd-range) is reported to all the spectrum data points, all together; 'start' and 'end'.

```
# - article:
#    label: Ap.J. 738, 25 (2011)
#    url: http://iopscience.iop.org/0004-637X/738/1/25/
#    arxiv: http://arxiv.org/abs/1106.1210
#    ads: http://adsabs.harvard.edu/abs/2011ApJ...738...25A
```

[6] http://simbad.u-strasbg.fr/simbad/sim-fid

'article', 'label' is a higher-level designation of the article. Whereas 'url' holds the address of the (if published) journal; 'arxiv' and 'ads' (for ADS-Harvard) are relevant for open access.

```
# - comments:
#    - Name = Mrk421_2008_highA
#    - z = 0.031
#    - LiveTime(h) = 1.4
#    - significance = 73.0
```

'comments' are suggested to be placed as list items (preceded by '-') if they are short and detached. Otherwise, like in 'description', 'comments' can be a paragraph, contiguous block of text spanning multiple lines to form a higher-level note about the data.

```
# - SED_TYPE: diff_flux_points
```

Following the SED standard, 'SED_TYPE' defines the type of spectrum we should expect from the data points in the table. The following options are supported:

- 'diff_flux_points' (synonym for 'dnde')

'datatype' section

Begin of 'datatype' section.

```
# datatype:
```

```
# - name: e_ref
#   unit: TeV
#   datatype: float64
# - name: dnde
#   unit: ph / (m2 TeV s)
#   datatype: float64
# - name: dnde_errn
#   unit: ph / (m2 TeV s)
#   datatype: float64
# - name: dnde_errp
#   unit: ph / (m2 TeV s)
#   datatype: float64
```

Columns 'e_ref', 'dnde' are mandatory for 'SED_TYPE = dnde'. 'unit' information for each column are mandatory as well. 'datatype' for each column is a Astropy/ECSV requirement.

A.2 Summary

Data files were defined and tested to support the requirements of the Gamma-ray community and have had features that extend its usability to our particular collaboration. Attention to the SED standard been built by the gamma-ray community is an important aspect to merge efforts on better describing datasets, and helps in building a homogeneous, high-level interface for data access. Regarding the ECSV format, we could arrange our information using the Astropy's format so that metadata stays clear and readability is gain through the use of a stable, broadly used library.

Bibliography

Abazajian, K. and for the Sloan Digital Sky Survey (2008). "The Seventh Data Release of the Sloan Digital Sky Survey". In: *Astrophys. J. Suppl. Ser.* 182.2, pp. 543–558. arXiv: 0812.0649.

Abell, George O, Jr. Corwin, Harold G., and Ronald P Olowin (1989). "A catalog of rich clusters of galaxies". In: *Astrophys. J. Suppl. Ser.* 70, p. 1.

Abolfathi, Bela et al. (2017). "The Fourteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the extended Baryon Oscillation Spectroscopic Survey and from the second phase of the Apache Point Observatory Galactic Evolution Experiment". In: arXiv: 1707.09322.

Acciari, V. A. et al. (2011). "TeV and multi-wavelength observations of Mrk421 in 2006-2008". In: *Astrophys. J.* 738.1.

Acero, F. et al. (2015). "Fermi Large Area Telescope Third Source Catalog". In: *Astrophys. Journal, Suppl. Ser.* 218.2.

Ade, P. A. R. et al. (2014). "Planck 2013 results. XXVIII. The Planck Catalogue of Compact Sources". In: *Astron. Astrophys.* 571, A28. arXiv: 1303.5088.

Almeida, Ulisses Barres de et al. (2017). "The Brazilian Science Data Center (BSDC)". In: *IWARA*. Vol. 0. arXiv: 1702.06828.

Annis, James et al. (2011). "The SDSS Coadd: 275 deg^2 of Deep SDSS Imaging on Stripe 82". In: *Astrophys. J.* 794.2, p. 120. arXiv: 1111.6619.

Arviset, Christophe, Severin Gaudet, and Technical Coordination Group IVOA (2010). *IVOA Architecture*.

Barres de Almeida, Ulisses et al. (2017). "Long-Term Multi-Band and Polarimetric View of Mkn 421: Motivations for an Integrated Open-Data Platform for Blazar Optical Polarimetry". In: *Galaxies* 5.4, p. 90.

Becker, Robert H., Richard L. White, and David J. Helfand (1995). "The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters". In: *Astrophys. J.* 450, p. 559.

Bock, Douglas C.-J. (2014). "The Australia telescope national facility". In: *2014 XXXIth URSI Gen. Assem. Sci. Symp. (URSI GASS)*, pp. 1–1. arXiv: 0412641 [astro-ph].

Boller, Th. et al. (2016). "Second ROSAT all-sky survey (2RXS) source catalogue". In: 103, pp. 1–26. arXiv: 1609.09244.

Chang, Y. and P. Giommi (in preparation). In:

Charbonnier, Aldée et al. (2017). "The abundance of compact quiescent galaxies since z ˜ 0.6". In: *Mon. Not. R. Astron. Soc.* 469.4, pp. 4523–4536. arXiv: 1701.03471.

Collaboration, Planck (2016). "Astrophysics Special feature Planck 2015 results". In: *Astron. Astrophys.* 13, pp. 1–39. arXiv: 1502.01589.

Collaboration, The Fermi-LAT (2013). "The Second Fermi Large Area Telescope Catalog of Gamma-ray Pulsars". In: 17. arXiv: 1305.4385.

— (2017). "3FHL: The Third Catalog of Hard Fermi-LAT Sources". In: arXiv: 1702.00664.

Committee on the Peaceful Uses of Outer Space (2016). *"Open Universe" proposal , an initiative under the auspices of the Committee on the Peaceful Uses of Outer Space for expanding availability of and accessibility to open source space science data.* Tech. rep. June.

— (2017). *Report on the Expert Meeting on preparation of the United Nations / Italy Workshop on the Open Universe Initiative*. Tech. rep. June.

— (2018). *Report on the United Nations/Italy Workshop on the Open Universe initiative.* Tech. rep. December 2017.

Condon, J J et al. (1998). "1.4 GHz NRAO VLA Sky Survey (NVSS)". In: *Astron. J.* 8065.5, pp. 1693–1716.

Cortina, Juan (2005). "Status and First Results of the Magic Telescope". In: *Astrophys. Space Sci.* 297.1-4, pp. 245–255. arXiv: 0407475 [astro-ph].

D'Elia, V. et al. (2013). "The seven year Swift-XRT point source catalog (1SWXRT)". In: *Astron. Astrophys.* 551, A142. arXiv: 1302.7113.

Demleitner, Markus et al. (2014a). "The Virtual Observatory Registry". In: *Astron. Comput.* 7, pp. 101–107. arXiv: 1407.3083.

Demleitner, Markus et al. (2014b). "Virtual observatory publishing with DaCHS". In: *Astron. Comput.* 7-8.2, pp. 27–36. arXiv: 1408.5733.

Elvis, Martin et al. (1992). "The Einstein Slew Survey". In: *Astrophys. J. Suppl. Ser.* 80, p. 257.

Evans, Ian N. et al. (2010). "The Chandra source catalog". In: *Astrophys. Journal, Suppl. Ser.* 189.1, pp. 37–82. arXiv: 1005.4665.

Evans, P. a. et al. (2013). "1SXPS: A deep Swift X-ray Telescope point source catalog with light curves and spectra". In: *Astrophys. J. Suppl. Ser.* 210.1, p. 8. arXiv: 1311.5368.

Flewelling, H. A. et al. (2016). "The Pan-STARRS1 Database and Data Products". In: arXiv: 1612.05243.

Giommi, P. et al. (1991). "The EXOSAT high Galactic latitude survey". In: *Astrophys. J.* 378, p. 77.

Gregory, P. C. and J J Condon (1991). "The 87GB catalog of radio sources covering delta between O and + 75 deg at 4.85 GHz". In: *Astrophys. J. Suppl. Ser.* 75, p. 1011.

Gregory, P. C. et al. (1996). "The GB6 Catalog of Radio Sources". In: *Astrophys. J. Suppl. Ser.* 103, p. 427.

Hanisch, R. J. et al. (2015). "The Virtual Astronomical Observatory: Re-engineering Access to Astronomical Data". In: *Astron. Comput.* 11.PB. arXiv: 1504.02133.

Hanisch, Robert, Resource Registry Working Group IVOA, and Metadata Working Group NVO (2007). *Resource Metadata for the Virtual Observatory*.

Harris, D.E. et al. (1994). *EINSTEIN Observatory catalog of IPC X-ray sources*.

Healey, Stephen E. et al. (2007). "CRATES: An All-Sky Survey of Flat-Spectrum Radio Sources". In: *Astrophys. J. Suppl. Ser.* 171.1, pp. 61–71. arXiv: 0702346 [astro-ph].

Hodge, J. A. et al. (2011). "High-resolution VLA Imaging of SDSS Stripe 82 at 1.4 GHz". In: *Astron. J.* 142.1, p. 3. arXiv: 1103.5749.

Jiang, Linhua et al. (2014). "The Sloan Digital Sky Survey Stripe 82 Imaging Data: Depth-Optimized Co-adds Over 300 Deg^2 in Five Filters". In: *Astrophys. J. Suppl. Ser.* 213.1, p. 12. arXiv: 1405.7382.

LaMassa, Stephanie M. et al. (2012). "Finding Rare AGN: X-ray Number Counts of Chandra Sources in Stripe 82". In: *Mon. Not. R. Astron. Soc.* 432.2, pp. 1351–1360. arXiv: 1210.0550.

LaMassa, Stephanie M. et al. (2013). "Finding Rare AGN: XMM-Newton and Chandra Observations of SDSS Stripe 82". In: *Mon. Not. R. Astron. Soc.* 436.4, pp. 3581–3601. arXiv: 1309.7048.

LaMassa, Stephanie M. et al. (2015). "The 31 Deg^2 Release of the Stripe 82 X-ray Survey: The Point Source Catalog". In: *Astrophys. J.* 817.2, p. 172. arXiv: 1510.00852.

Lasker, Barry M. et al. (2008). "The second-generation guide star catalog: Description and properties". In: *Astron. J.* 136.2, pp. 735–766. arXiv: 0807.2522.

Lawrence, A. et al. (2007). "The UKIRT infrared deep sky survey (UKIDSS)". In: *Mon. Not. R. Astron. Soc.* 379.4, pp. 1599–1617. arXiv: 0604426 [astro-ph].

Liu, Teng et al. (2015). "The swift X-ray telescope cluster survey. III. Cluster catalog from 2005-2012 archival data". In: *Astrophys. Journal, Suppl. Ser.* 216.2. arXiv: 1503.04051.

Mao, Lisheng and Xuemei Zhang (2016). "Long-term optical variability properties of blazars in the SDSS Stripe 82". In: *Astrophys. Space Sci.* 361.10, p. 345.

Martin, D. Christopher et al. (2005). "The Galaxy Evolution Explorer : A Space Ultraviolet Survey Mission". In: *Astrophys. J.* 619.1, pp. L1–L6. arXiv: 0411302 [astro-ph].

Marton, G et al. (2017). "The Herschel/PACS Point Source Catalogue Explanatory Supplement". In: pp. 1–56. arXiv: 1705.05693.

Massaro, E. et al. (2015). "The 5th edition of the Roma-BZCAT. A short presentation". In: *Astrophys. Space Sci.* 357.1, pp. 1–5. arXiv: 1502.07755.

Mauch, T et al. (2003). "SUMSS: a wide-field radio imaging survey of the southern sky - II. The source catalogue". In: *Mon. Not. R. Astron. Soc.* 342.4, pp. 1117–1130.

McConnell, D. et al. (2012). "ATPMN: Accurate positions and flux densities at 5 and 8GHz for 8385 sources from the PMN survey". In: *Mon. Not. R. Astron. Soc.* 422.2, pp. 1527–1545. arXiv: 1202.2625.

Murphy, Tara et al. (2010). "The Australia Telescope 20 GHz Survey: The source catalogue". In: *Mon. Not. R. Astron. Soc.* 402.4, pp. 2403–2423. arXiv: 0911.0002.

Oh, Kyuseok et al. (2018). "The 105-Month Swift -BAT All-sky Hard X-Ray Survey". In: *Astrophys. J. Suppl. Ser.* 235.1, p. 4. arXiv: 1801.01882.

Padovani, P. et al. (2016). "Extreme blazars as counterparts of IceCube astrophysical neutrinos". In: *Mon. Not. R. Astron. Soc.* 457.4, pp. 3582–3592. arXiv: 1601.06550.

Padovani, Paolo (2017). "Active Galactic Nuclei at All Wavelengths and from All Angles". In: *Front. Astron. Sp. Sci.* 4.November, pp. 1–7.

Page, M. J. et al. (2012). "The XMM-Newton serendipitous ultraviolet source survey catalogue". In: *Mon. Not. R. Astron. Soc.* 426.2, pp. 903–926. arXiv: 1207.5182 [astro-ph.CO].

Panzera, M R et al. (2003). "The Brera Multi-scale Wavelet ROSAT HRI source catalogue". In: *Astron. Astrophys.* 399.1, pp. 351–364.

Piffaretti, R. et al. (2011). "The MCXC: a meta-catalogue of x-ray detected clusters of galaxies". In: *Astron. Astrophys.* 534, A109. arXiv: 1007.1916.

Prusti, T. et al. (2016). "The Gaia mission". In: *Astron. Astrophys.* 595, A1. arXiv: 1609.04153.

Robertson, Brant E. et al. (2017). "Large Synoptic Survey Telescope Galaxies Science Roadmap". In: pp. 1–3. arXiv: 1708.01617.

Rosen, S. R. et al. (2016). "The XMM-Newton Serendipitous Survey. VII. The Third XMM-Newton Serendipitous Source Catalogue". In: *a&a* 590.3, A1. arXiv: 1504.07051.

Saxton, R.˜D. et al. (2008). "The first XMM-Newton slew survey catalogue: XMMSL1". In: *Aap* 480, pp. 611–622. arXiv: 0801.3732.

Schulz, Bernhard et al. (2017). "SPIRE Point Source Catalog Explanatory Supplement". In: arXiv: 1706.00448.

Skrutskie, M. F. et al. (2006). "The Two Micron All Sky Survey (2MASS)". In: *Astron. J.* 131.2, pp. 1163–1183. arXiv: 1708.02010.

Soo, John Y. H. et al. (2018). "Morpho-z: improving photometric redshifts with galaxy morphology". In: *Mon. Not. R. Astron. Soc.* 475.3, pp. 3613–3632. arXiv: 1707.03169.

Sutherland, W. and W. Saunders (1992). "On the likelihood ratio for source identification". In: *Mon. Not. R. Astron. Soc.* 259.3, pp. 413–420.

Timlin, John D. et al. (2016). "SpIES: THE SPITZER IRAC EQUATORIAL SURVEY". In: *Astrophys. J. Suppl. Ser.* 225.1, p. 1. arXiv: 1603.08488.

Viero, M. P. et al. (2013). "The Herschel Stripe 82 Survey (HerS): Maps and Early Catalog". In: *Astrophys. J. Suppl. Ser.* 210.2, p. 22. arXiv: 1308.4399.

Webb, J. R. et al. (1990). "The 1987-1990 optical outburst of the OVV quasar 3C 279". In: *Astron. J.* 100.5, p. 1452.

Weekes, T.C et al. (2002). "VERITAS: the Very Energetic Radiation Imaging Telescope Array System". In: *Astropart. Phys.* 17.2, pp. 221–243. arXiv: 0108478 [astro-ph].

Wen, Z. L., J. L. Han, and F. S. Liu (2012). "A catalog of 132,684 clusters of galaxies identified from Sloan Digital Sky Survey III". In: 34. arXiv: 1202.6424.

White, N. E., P. Giommi, and L Angelini (1994). *The WGA Catalog of ROSAT Point Sources*.

White, Richard L and Robert H Becker (1992). "A new catalog of 30,239 1.4 GHz sources". In: *Astrophys. J. Suppl. Ser.* 79, p. 331.

White, Richard L. et al. (1997). "A Catalog of 1.4 GHz Radio Sources from the FIRST Survey". In: *Astrophys. J.* 475.2, pp. 479–493.

Wilkinson, Mark D. et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Sci. Data* 3, p. 160018.

Wright, Alan E. et al. (1994). "The Parkes-MIT-NRAO (PMN) surveys. 2: Source catalog for the southern survey (delta greater than -87.5 deg and

less than -37 deg)". In: *Astrophys. J. Suppl. Ser.* 91, p. 111. arXiv: arXiv:1011.1669v3.

Wright, Edward L. et al. (2010). "The Wide-field Infrared Survey Explorer (wise): Mission description and initial on-orbit performance". In: *Astron. J.* 140.6, pp. 1868–1881. arXiv: 1008.0031.

Y. Chang P. Giommi, C. Brandt (in preparation). In:

Yershov, V. N. (2014). "Serendipitous UV source catalogues for 10 years of XMM and 5 years of Swift". In: *Astrophys. Space Sci.* 354.1, pp. 97–101.

Zwicky, F. et al. (1961). *Catalogue of galaxies and of clusters of galaxies, Vol. I*. California Institute of Technology.