Data Interoperability - The CDS Experience

Pierre Fernique, M. G. Allen

Data Interoperability - the CDS experience

Pierre Fernique[1] and Mark G. Allen[1]
[1] *Observatoire astronomique de Strasbourg, Université de Strasbourg, CNRS, UMR 7550, 11 rue de l'Université, F-67000 Strasbourg, France*

Abstract. Provide astronomers with simple, efficient and rapid means of accessing the reference data (bibliography, catalogues, object identifiers, images, spectra, time series, ...) necessary for their research". Since its creation in the early 1970s, the Strasbourg astronomical Data Centre (CDS) has addressed this challenge. The context and technologies have evolved from post mail to touchscreen devices, but the challenge remains: identify, collect, homogenize, describe data, and then redistribute it so that the knowledge can be re-used. What is called nowadays: the FAIR principles. At the heart of this challenge: the data interoperability. Standards, data serialization, metadata, dictionaries, tools, languages, semantics ... Based on our experience at CDS, we will try to identify the good, but also the bad, of the various interoperability solutions that have marked the evolution of CDS: those that work, those that failed. We will try to draw some lessons from the past experiences, and try to consider what the interoperability might enable for future astronomical projects.

1. Introduction

Data interoperability is a subject that has been particularly studied and debated for decades in ADASS conferences. A search of ADS shows that there are more than 150 publications on the subject since 1992. We start by addressing the question, "What is Interoperability?"

From the user's point of view, *"Interoperability is what allows different applications to work as one"*. But behind this apparent observation hides another definition:*"Data interoperability addresses the ability of systems and services that create, exchange and consume data to have clear, shared expectations for the contents, context and meaning of that data."*. [1]

But, once these definitions are established, what are the technical solutions, the appropriate architecture, the best way to reach this goal? The approach we propose in this article will be based on the experience of the Astronomical Data Centre in Strasbourg[2] (Allen 2018). Throughout its almost 50 years of operation, CDS has continually tested and deployed interoperability solutions. Based on this experience, we will try to identify the good, but also the bad, of the various interoperability solutions that have marked the evolution of CDS: those that work, those that failed.

---

[1] Data Interoperability Standards Consortium definition: `https://datainteroperability.org/`.

[2] http://cds.unistra.fr

Our article will be structured in five points: - Why exchange our data? - With whom ? - Which data to exchange? - How to structure them? - How to distribute them?

But as we will base our study on the CDS experience, we have to present briefly this data centre. The Strasbourg Astronomical Data Centre can be summarized using these 4 numbers: three, five hundred, sixteen and two millions: 3 well-known services: SIMBAD – for astronomical objects in literature, VizieR – for cataloges and Aladin – for images. It represents five hundred terabytes of data, sixteen operational servers or web sites spread over the 5 continents and more than 2 million requests per day; that means more than twenty requests each second. One could say that "Interoperability" is a constant reality for CDS.

2. Why exchange our data?

In the context of CDS, it is interesting to come back to the origins: its charter. *"Collect "useful" data on astronomical objects, in electronic form; Improve them by critically evaluating them and combining them; Distribute the results to the international community..."*. We note that in 1972, these concepts were already very close to what is understood today by the FAIR principles: Findable, Accessible, Interoperable, Reusable.

In fact, astronomy has been a discipline particularly favorable to this. First of all we have a long tradition of collaboration. Of course, the need for scientific exchange provides a favorable context but this alone is probably not sufficient. In terms of economic realities, it may in part be due to the significant costs of the instruments that makes sharing favorable or indeed necessary. Compared with other disciplines such as biology or earth sciences, it may be the fact that astronomical data has no real market value that makes it much easier to distribute freely. Concretely, in astronomy, we are in a very favorable win-win situation for interoperability solutions.

3. With whom do we exchange data?

As mentioned above, we are in a favorable situation, but with whom to exchange data. If we take the example of CDS, but it concerns just as well other data centres and archives around the world, we exchange data not only between the various servers for which we are responsible, but also with dozens of other institutes: publishers, other data centres in order to enrich their data and our data. It is a form of reciprocity which means that, for example, ADS collects all of the SIMBAD content it needs every week, or we collect articles from major publishers daily in order to update our databases

In addition, we must also take into account the software clients who access our databases: Aladin, Stellarium, the various Web interfaces from other institutes which require an astronomical position, etc ... With all this put together we find that the CDS servers attract some 2 million requests per day.

4. Which data to exchange?

Now we come to the question of what kind of data to exchange. Our answer has always been the same: everything that is technically possible. But we will see that it has evolved in time. Until the 1990s, we only provided database records: the SIMBAD

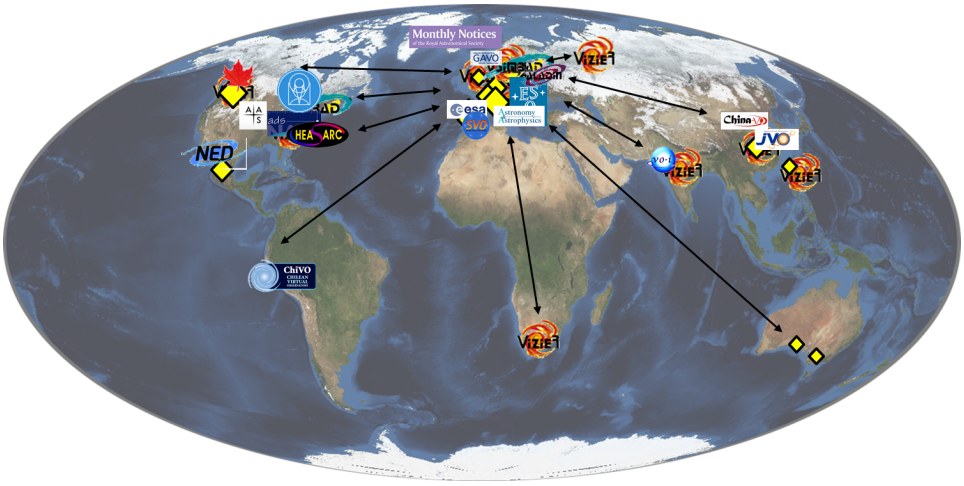

Figure 1. CDS interoperability network: 16 servers/sites managed by CDS + dozens of collaborative institutes/data centres/publishers (especially 18 HiPS nodes)

content. Then at that time, the evolution of technologies, and the creation of the Internet, progressed to allow the distribution of astronomical tables and catalogs such as in the VizieR service. And we had to wait another ten years to be able to exchange the origin of all this data, ie the images, which at CDS was the start of Aladin.

At first glance, one might think that it was the volume that was the discriminating factor that enabled these steps. In fact, the real answer is not so simple. Yes, the network is constantly improving, but the data is also growing larger. And if we compare the growth factors, we realize that the network effectively multiplies its performance by 10 every decade, but faced to a 30 factor for the size of the images (Kaiser 2009). Given these changes, we were certainly in a more favorable situation in terms of transferring individual images in 2000 when we had to exchange 2MASS images, than now when we try to download MegaCAM images. Also a point to underline is that large fraction of world traffic is now done on mobile devices. Not really a very fast technology. And today, half of humanity has network connection of less than 9Mbits/s [3]. Therefore, the answer is not the volume in an absolute way but the time necessary to move from point A to point B such or such data.

So when the data is too big to move, what are our solutions? We mainly use two approaches. i) Either we offer our users to operate directly on the server and only the final result of the request will be transmitted (for example, up until 2006, we had an email query submission service for SIMBAD). Nowadays, we are implementing technologies like jupyter Hub to dynamically perform remote tasks without having to move data. And if we take a step back, we can see that is similar to our old Unix accounts that we offered on our server to our users in the 90s. Python to replace the Shell, the browser panel instead of the VT100 console. But basically it's the same solution: get the jobs done without moving data. ii) The second solution - which can be seen as complementary to the first - consists of moving only the data strictly necessary for the user's needs. It is this principle that we will find in the IVOA TAP protocol

[3]Computed from `https://www.cable.co.uk/broadband/speed/worldwide-speed-league/`

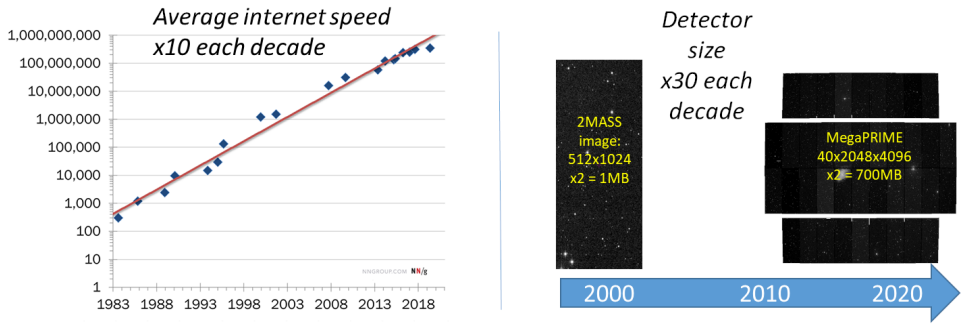

Figure 2. Internet connectivity evolution vs detector size evolution

(Dowler et al. 2010) or in the HiPS standard (Fernique et al. 2017). In the Hierarchical Progressive Surveys mechanism (Fernique et al. 2015), the data (pixels, catalogues) is split into tiles, and only the tiles which are necessary for the user will be downloaded. In 2020, this represents 170,000 billions of pixels from more than a thousand of surveys[4] and 800,000 tiles queries per day. This is clearly the most successful approach. Pre-calculate the data the user wants as far as possible, and make it easy to distribute/stream it at reasonable network speeds.

To finish on this point, we present here the evolution of the statistics of use of SIMBAD over 25 years. We went from 50 requests per day in 1995, 2,000 in 1998, 20,000 in 2004, 200,000 in 2010 and today close to 700,000 requests per day. It is interesting to note that the ability to respond quickly to user requests will create new uses. New clients brings new usages, and the rest of this article will illustrate this point.

5. How to structure them?

So far we have been considering the data, but data is useless if we do not know what it represents. We come back to the definition of interoperability. First and foremost, data interoperability requires good metadata. If we look at the composition of CDS team, we are forty: astronomers, documentalists, engineers; Almost 70% of the work is to associate the right metadata with the data we exchange. This is the heart of the CDS work : to add value.

What do we mean by meta data? These are the four keys of interoperability: *name*, *select*, *characterize*, and *structure*. Below, we expand on each of these points, first we illustrate why interoperability is so important..

Figure 3 shows the Vizier photometry widget (Allen et al. 2014) that anybody may embed in thier web pages. It displays the photometry points extracted from all of the 20,000 tables that VizieR contains. Based on a position, the system is able to plot on the same graph the values in flux versus wavelength from the columns of these tables. To generate this graph, it was necessary to select before the right tables, the right columns, the right units and to associate the right systems - typically filters definitions, and to

[4] `https://aladin.u-strasbg.fr/hips/list`.

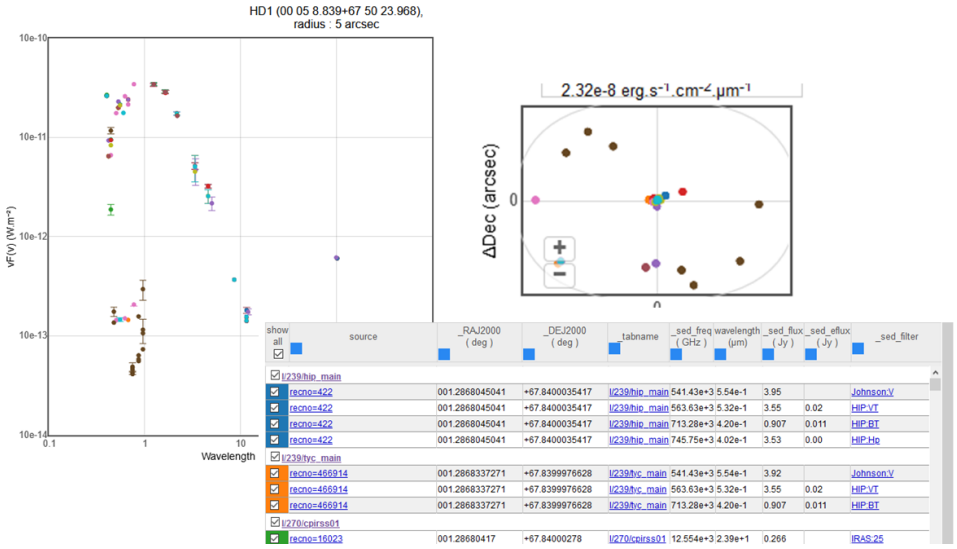

Figure 3. VizieR photometry widget tool allows for easy visualization of photometry points extracted around a given position from photometry-enabled catalogues

combine them. Maybe one day an AI will be able to do this automatically, but right now it is still requires a significant amount of human expertise.

5.1. Name

Naming is the identification of the data. The four naming schemes used at CDS are: **bibcode** for bibliographic references (i.e. 2016A&A...595A...1G), the **catalog identifiers** (i.e. I/350/gaiaedr3), the **DOIs** (i.e.10.1051/0004-6361/201629272) and finally the **IVORNs** (i.e. ivo:/CDS/Simbad). The goal is to be able to identify and list the data resources, for example in a database or in a registry. Without having a robust identification system, interoperability becomes much more difficult. A good system should guarantee that an identifier is: unique, permanent, easy to manipulate (short expression, potentially human readable, ...), shared in the discipline, or even beyond. Also, identifiers should be: adapted the data (data type, granularity, derivable, reusable...), and be easy to create even before publication. This first step allows to register the data resources.

5.2. Select and describe

The data we handle is diverse, and it is often impossible to describe each distinct data item. This is why it is necessary to choose the data elements which will be precisely described. So for example, in Vizier, it is necessary to first characterize the columns of positions, time and flux. In the same way, for SIMBAD, we must first determine in the set of research articles the citations of astronomical objects. This is the heart of the CDS work. Select the good data to be able to provide the good meta data. Of course, this step is very time consuming and even if we have tools to help us to do that, expertise remains the main criterion to get interesting interoperable data. This second step allows to index and compare the data.

5.3. Characterization

Finally, to characterize these selected data, we use several methods. The most complete consists in being able to associate a reference system and a unit. It allows numerical comparisons, but this step is very expensive in terms of expertise. This is why we explored a complementary way allowing to associate a somewhat "fuzzy" description of what it represents, e.g a magnitude, an error, a parallax, etc. These are called Unified Content Descriptors (UCD) (Martinez et al. 2018). This characterization mechanism has been deployed for about fifteen years. It works well, and for instance, the VizieR catalogue selection system can be searched by UCDs. But in fact it is used relatively less than we would have hoped. This is probably due to the inherent imprecision of UCDs. Or maybe it is because users do not fully catch what it represents. This third step allows generic manipulations as presented in the photometric VizieR viewer.

5.4. Structure

There are two approaches for structuring the described data. Either we can manage the data as it was produced, and that is typically the approach of a platform like Zenodo. Or we can map the data in our own structure. And this is the approach adopted by CDS. Mainly because it allows to manipulate a small and well-defined set of metadata. And the clients (tools, API) that will have to read this data will only have to support a single structure. Thus, to provide consistent and homogeneous data and metadata we have opted for constrained structures: predefined fields for SIMBAD, a 2-level structure for the VizieR catalogues and a tiling system for Aladin.

6. How to distribute the data?

The data - named, selected, described and structured - will have to be distributed to ensure the interoperability of the components. There are 5 choices to decide: 1- The medium, 2- The transport protocols, 3- The packaging, 4- The query protocols, 5- The API and clients. The first two have obvious answers: Internet and HTTP. The other three choices are less obvious. In the following we study these elements.

6.1. Packaging

When it is necessary to transport data, either it is the packaging that adapts to the data, or it is the data that adapts to the packaging. In the first case, the data will be move as it is. In the second case it will require conversions and only the data compatible with the packaging could be processed. The choice of CDS is more oriented towards the second option because it allows much greater flexibility, it is scalable, derivable, extensible, easy for the storage and the transport. And finally, it is much simpler for the client.

For instance, if we compare HDF5 and HiPS. We find very few clients capable of understanding the correct HDF5 dialect. Each HDF5 data provider may be tempted to adapt this packaging for their own needs. And in fact, there are more and more HDF5 dialects. In contrast, we will find more than fifteen different clients for the HiPS packaging mechanism - by tiles. And more than twenty data providers of HiPS tiles. This protocol has been standardized in 2017, just 3 years ago. So, the second approach is probably more interoperable for this type of data.

As said before, data without metadata is not very useful, but how do you transfer these metadata? The most practical technique is logically to use the same mechanism

Figure 4. Package adaptability versus Data adaptability

for both. It allows them to be closely associated. Surprisingly, this is not always the first choice of users. For example, there are a large fraction of VizieR users who continue to download catalogues in ASCII, with almost no associated metadata.

In fact, there are at least three possibilities to transfer metadata: either a free vocabulary, for example the FITS keywords. This method quickly turns out to be limited by the impossibility for the clients to effectively process this free vocabulary. It is very heterogeneous, and unclear how to do something with them. A more pragmatic solution, which is the main one used at the CDS, is to use a controlled vocabulary and dedicated fields. We will find this in the IVOA VOTable format (Ochsenbein et al. 2004). Finally, the most comprehensive approach would be to provide the Data Model associated with the data, with two variants: predefined DM, or self describing DM. These last approaches, although tempting, turns out to be very difficult in practice because it requires prior agreement on the Data Model, and on the evolution of the Data Model. And in an international multi-partner context, notably in IVOA, this is clearly a challenge.

6.2. Query protocol

Data distribution requires a request protocol. We can only applaud the efforts of the IVOA to standardize these kinds of protocols (i.e. Cone Search, Table Access Protocol, Simple Image Access, Simple Spectra Access, HiPS, etc). This is clearly the best approach for interoperability even it is not always well adapted to specific need. One can deploy local solution - very efficient but complex to deploy and to evolve, or bipart solution - very well-adapted, but must be reinvented for each collaboration. But in term of interoperability, it is not the same impact. The standardization is definitively the best approach. Or even one can choose proprietary solutions, even opened, but by assuming that the control is out of their hands. There is no guarantee of long-term sustainability.

6.3. APIs and clients

And at the end of the pipe, we have the APIs and clients. This is what the users use to interact with our servers. As we rely on open standardized protocols, we have seen the deployment of a wide choice of clients and APIs, developed by ourselves (CDS toolkit, CDS widgets, Aladin Lite, Aladin Desktop,...) or others (TAP & other IVOA libraries, Astropy & pyVO python libraries, TOPCAT, ESASky, Stellarium, WWT, Digistar,...),

dedicated to several types of audiences: scientists, amateurs, general public. This diversity is really a good indicator of the interoperability of a system. And we are closely monitoring the evolution of these different access modes.

At present, more than 80% of requests on CDS services are not made at the "end of the chain" (on our interactive web pages), but rather come from from APIs, third-party tools or partner institutes in order to re-ingest our data into their databases, and combine them with their data. Regarding the changes in the way SIMBAD is used, we see that the world evolves, the technology evolves and so usages also evolve!

7. Conclusion

By way of conclusion, we go back to the initial question: why exchange data? We wondered about the evolution of the astronomical context. Indeed, astronomical data, which until now had no monetizable value is now paying indirectly: whether through market places, dedicated apps, associated advertisements, or the personal data collected by various platforms. And we are seeing the emergence of new practices. Therefore, will the win-win bet of the data sharing always true ? An open question !

And another big evolution concerns the rise of interdisciplinarity, for instance the European projects: RDA or EOSC who are working on sharing data with more disciplines that just astronomy: with social sciences, medical sciences, etc. Will we collectively succeed in expanding, or invented new structures such as IVOA to cover this new field of interoperability. It's an exciting challenge.

In this article, we presented 50 years of CDS experience on interoperability - at least our feedback on it. Our perspective is probably influenced by the current, rapidly changing environment of data sharing, and of course things are probably more gray than black and white as we have presented.

Acknowledgments. We would like to thank all the people, our colleagues, and our predecessors, who have built these interoperable systems: at CDS of course, but also partners, data centres, archives, publishers, developers, the VO and finally all the actors of the Internet which allow this interoperability to be a reality.

References

Allen, M. 2018, in European Physical Journal Web of Conferences, vol. 186 of European Physical Journal Web of Conferences, 12001

Allen, M. G., Ochsenbein, F., Derriere, S., Boch, T., Fernique, P., & Landais, G. 2014, in Astronomical Data Analysis Software and Systems XXIII, edited by N. Manset, & P. Forshay, vol. 485 of Astronomical Society of the Pacific Conference Series, 219

Dowler, P., Rixon, G., & Tody, D. 2010, Table Access Protocol Version 1.0, IVOA Recommendation 27 March 2010. 1110.0497

Fernique, P., Allen, M., Boch, T., Donaldson, T., Durand, D., Ebisawa, K., Michel, L., Salgado, J., & Stoehr, F. 2017, HiPS - Hierarchical Progressive Survey Version 1.0, Tech. rep.

Fernique, P., Allen, M. G., Boch, T., Oberto, A., Pineau, F. X., Durand, D., Bot, C., Cambrésy, L., Derriere, S., Genova, F., & Bonnarel, F. 2015, A&A, 578, A114

Kaiser, N. 2009, in 2009 IEEE Aerospace conference, 1

Martinez, A. P., Louys, M., Cecconi, B., Derriere, S., Ochsenbein, F., & IVOA Semantic Working Group 2018, The UCD1+ controlled vocabulary Version 1.3 Version 1.3, IVOA Recommendation 27 May 2018

Ochsenbein, F., Williams, R., Davenhall, C., Durand, D., Fernique, P., Giaretta, D., Hanisch, R., McGlynn, T., Szalay, A., Taylor, M. B., & Wicenec, A. 2004, VOTable Format Definition Version 1.1, IVOA Recommendation 11 August 2004