EPJ Web of Conferences **186**, 02002 (2018) https://doi.org/10.1051/epjconf/201818602002
*LISA VIII*

# Associated data: Indexation, discovery, challenges and roles

*Pierre* Ocvirk[1,⋆], *Gilles* Landais[1], *Laurent* Michel[1], *Heddy* Arab[1], *Sylvain* Guehenneux[1], *Thomas* Boch[1], *Marianne* Brouty[1], *Emmanuelle* Perret[1], *François-Xavier* Pineau[1], *Tiphaine* Pouvreau[1], and *Patricia* Vannier[1]

[1]*Observatoire astronomique de Strasbourg, Université de Strasbourg, 11 rue de l'Université, 67000 Strasbourg*

**Abstract.** Astronomers are nowadays required by their funding agencies to make the data obtained through public-financed means (ground and space observatories and labs) available to the public and the community at large. This is a fundamental step in enabling the open science paradigm the astronomical community is striving for. In other words, tabular data (catalogs) arriving to CDS for ingestion into its databases, in particular VizieR, is more and more frequently accompanied by the reduced observed dataset (spectra, images, data cubes, time series). While the benefits of making this associated data available are obvious, the task is very challenging: in this context "big data" takes the meaning of "extremely heterogeneous data", with a diversity of formats and practices among astronomers, even within the FITS standard. Providing librarians with efficient tools to index this data and generate the relevant metadata is therefore paramount.

## 1 Vizier

VizieR is the reference database for tabular data from astronomical catalogues and tables published in scientific papers. It contains reference catalogs such as Gaia and SDSS, and also provides access to more advanced data products, such as the CoRoT time series [1]. Its workflow and usage is described in [2]. Tables are stored, along with homogeneous "metadata". This metadata consists of a range of information describing the data, such as the magnitude system used, the coordinate system and its epoch, the units of the columns, the type of data (magnitudes, sizes, positions...), stored as Unified Content Descriptors. This set of metadata can be queried, allowing users to discover catalogs relevant for their research. Another strength of this approach is that VizieR allows for instance to perform cone queries in all of the >14 000 catalogs stored in one simple click. Vizier is fully VO-compatible and communicates with a number of important astronomical resources and software such as the CDS SIMBAD database, Aladin and TOPCAT. These features have made VizieR a very popular resource among astronomers, and as a result, there have been 300 000 queries/day during the last year. Thousands of astronomers have decided that VizieR is the right place to make their data available to the community, and institutional data providers such as ESO and ESA have also trusted the service to distribute their surveys.

Significant developments to improve VizieR pipeline for non-tabular data attached to papers were pursued, based on the ObsCore VO standard [3]. Also an important agreement was made with the

⋆e-mail: pierre.ocvirk@astro.unistra.fr ORCID: 0000-0002-8488-504X

AAS to use ObsCore as a common description framework. The VizieR associated data service has been been developed using the Saada data publishing tool [4] in collaboration with Laurent Michel (High Energy Team), and has been available in test mode in 2016. Among the datasets already implemented are the CoRoT data (150,000 time series), and the LAMOST DR1 spectra. These are queriable with an interactive interface, and a feature of the service is that these are also made automatically available to the VO using the relevant IVOA standards.

## 2 Associated data: the challenge of automatic indexation

### 2.1 Metadata and ObsCore data model

By "associated data" we mean here, non-tabular data, such as spectra, images, data cubes obtained by integral field spectroscopy or mm/submm arrays such as ALMA, associated to the catalog published in Vizier. Our main goal is to make this data discoverable, for instance using a positional query of the type: "select all optical spectra within 1 arcmin of the center of NGC300", through a dedicated service. The positional information is therefore primordial, as is often the case in astronomy. Therefore this positional information must be used for our indexation, along with other important properties of the observation. We have chosen the VO ObsCore data model to store the associated data's metadata, as it offered the best compromise between completeness and simplicity, while covering the most important aspects of the metadata we wanted to capture (name, position, wavelength range, time).

### 2.2 Heterogeneous FITS headers

At the moment we consider as associated data only reduced data, in the form of FITS files, because of the presence of directly accessible metadata in the FITS headers. In these FITS headers, we try to find, in order of priority, the pointing of the observation, the time of observation, the electromagnetic wavelength range, and the instrument in use. Although FITS headers can be quite verbose, they are extremely heterogeneous. Indeed, telescope design and software, instruments data reduction pipelines, guidelines and tasks in MIDAS or IRAF and custom user-written routines have an impact on the content of the FITS headers and keywords. Therefore, the latter have evolved significantly since the introduction of the FITS standard. For instance, among the common keywords for the target name, we find : "TARGET", "OBJECT", "NAME", and all possible abbreviations and combinations of these basic words, linked with possibly hyphens or underscores, "OBJ", "OBJ_NAME", "TARG-NAME", "OBJNAME", etc ... The same holds for the rest of the metadata, i.e. coordinates in space and time, and although lists of commonly used or recommended keywords exist (such as HEASARC's), there is no sign that astronomers nor engineers will restrict their creativity to stay within the existing dictionary of keywords. Therefore, we try to guess the mapping from the FITS keywords to the ObsCore data model as well as possible, and the mapping algorithm will for instance try to find the name within a curated collection of keywords. The main engine for this mapping is built in Saada, a publishing tool for catalogs developed by L. Michel at CDS. Because of the huge diversity of FITS formats, the automatic mapping from FITS keywords to ObsCore data model is often incomplete and must be completed by hand, i.e. requiring the expertise of a documentalist or an astronomer. This is, at the moment, the main factor preventing the full deployment of the associated data pipeline. CDS does not currently (as of August 2017) the manpower to properly index all of the incoming associated data, although work is being done to make the mapping software more robust, so as to limit human intervention in the process as much as possible.

### 2.3 Authors to the rescue

Since the most accurate knowledge of the data resides with the authors themselves, the VizieR catalog submission webpage (http://cdsarc.u-strasbg.fr/vizier.submit/) has been significantly overhauled to allow users to deposit and index their associated data along with the catalogs. Once the associated data is uploaded, the automatic mapping engine provides a first guess, which the author can check, edit and correct if need be, as shown on the left of Fig. 1. Since this feature is pretty recent, we cannot tell yet how successful it will be at properly indexing the data and removing some of the indexation workload from the shoulders of the CDS staff. Only time will tell.

## 3 The Vizier associated data service

Once the mapping of FITS keywords onto the ObsCore data model has been successfully performed, the associated data is ingested into a dedicated database which can then be queried.

The new service has been released to provide easy and interoperable access to data associated with journal publications. These associated data (images, spectra, time series) have always been part of VizieR in addition to the catalogue collection, but they are now much easier to find and use. The Saada database engine is used in collaboration with L. Michel (OAS XMM-SSC). The data can be queried by an interactive interface and the service automatically publishes the data via VO protocols. Important data sets already in the system include 170,000 CoRoT time series and 2.4 million LAMOST DR1 spectra. The service can be accessed at http://cdsarc.u-strasbg.fr/assocdata/, leading to the page shown in Fig. 2.

The benefits of a dedicated service for the associated data are three-fold:

- Better indexation and therefore discoverability and reusability thanks to homogeneous metadata using the ObsCore data model.
- The metadata can be queried for discovery using VO protocols: SIA [5], SSA [6], ObsTAP.
- Consequently the service can be queried by VO tools: plat, CASSIS, TOPCAT, Aladin,as shown in Fig. 3.

## 4 Conclusion

While the service is very new, we plan to advertise it widely among the astronomical community and hope its use will become standard practice, just as the publication of catalogs in VizieR has become. Work will continue to improve the robustness of the automatic mapping of FITS keywords to the ObsCore data model.

## References

[1] S. Chaintreuil, A. Bellucci, F. Baudin, et al., *II.5 Where to find the CoRoT data?* (2016), p. 109

[2] F. Ochsenbein, P. Bauer, J. Marcout, A&AS **143**, 23 (2000)

[3] D. Tody, A. Micol, D. Durand, et al., *Observation Data Model Core Components, its Implementation in the Table Access Protocol Version 1.0*, IVOA Recommendation 28 October 2011 (2011), 1111.1758

[4] L. Michel, P. Bantzhaff, C. Frère, et al., *A New Web Interface for Saada*, in *Astronomical Data Analysis Software and Systems XXI*, edited by P. Ballester, D. Egret, N. Lorente (2012), Vol. 461 of *Astronomical Society of the Pacific Conference Series*, p. 415

[5] P. Dowler, D. Tody, F. Bonnarel, ArXiv e-prints (2016), 1601.00519
[6] D. Tody, M. Dolensky, J. McDowell, et al., *Simple Spectral Access Protocol Version 1.1*, IVOA Recommendation 10 February 2012 (2012), 1203.5725

**Figure 1.** *Bottom:* supervised mapping of the FITS metadata to ObsCore using the interface built in the overhauled VizieR catalog submission web page. *Top:* main panel of the CDS-side mapping software for the documentalists.

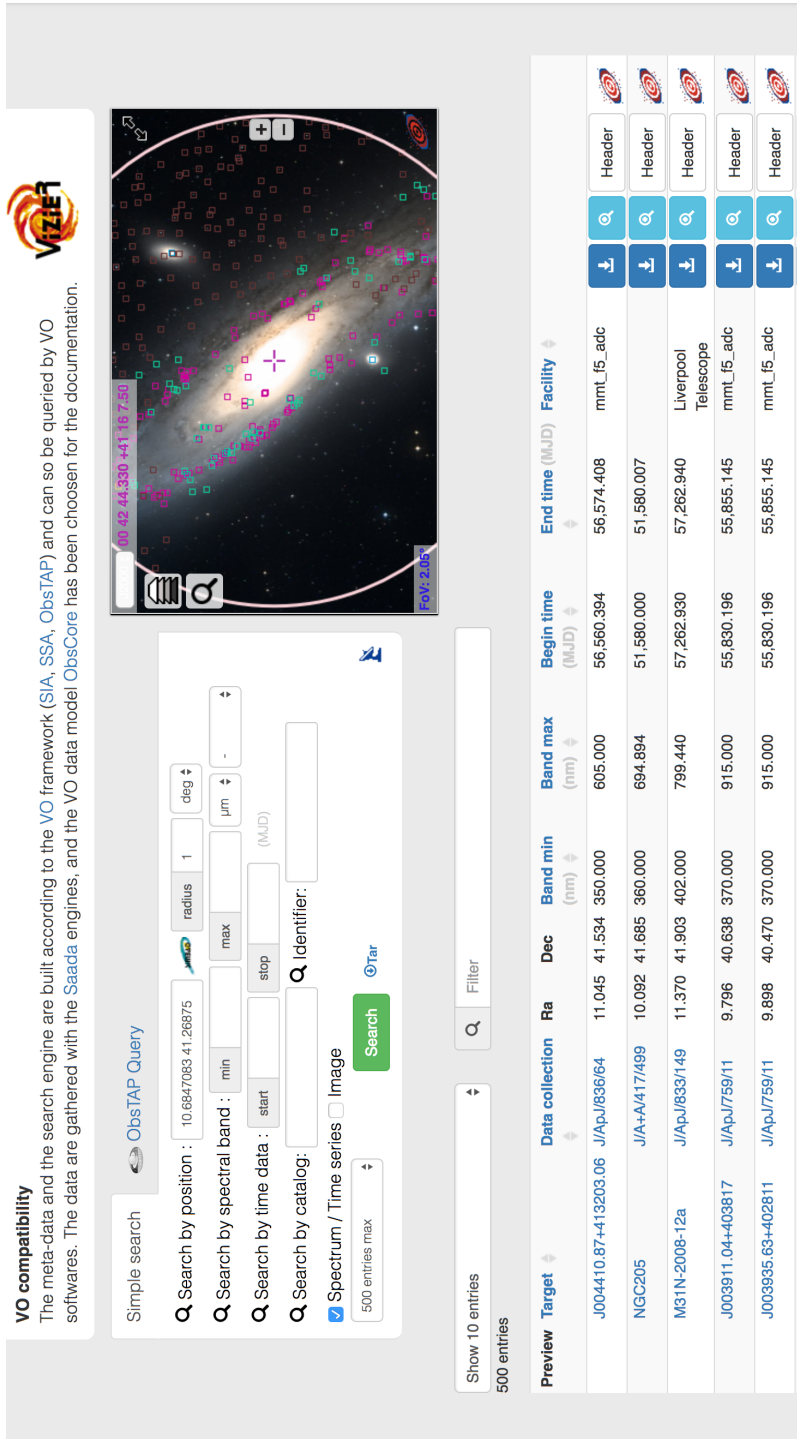


**Figure 2.** Homepage of the dedicated associated data service (http://cdsarc.u-strasbg.fr/assocdata/).

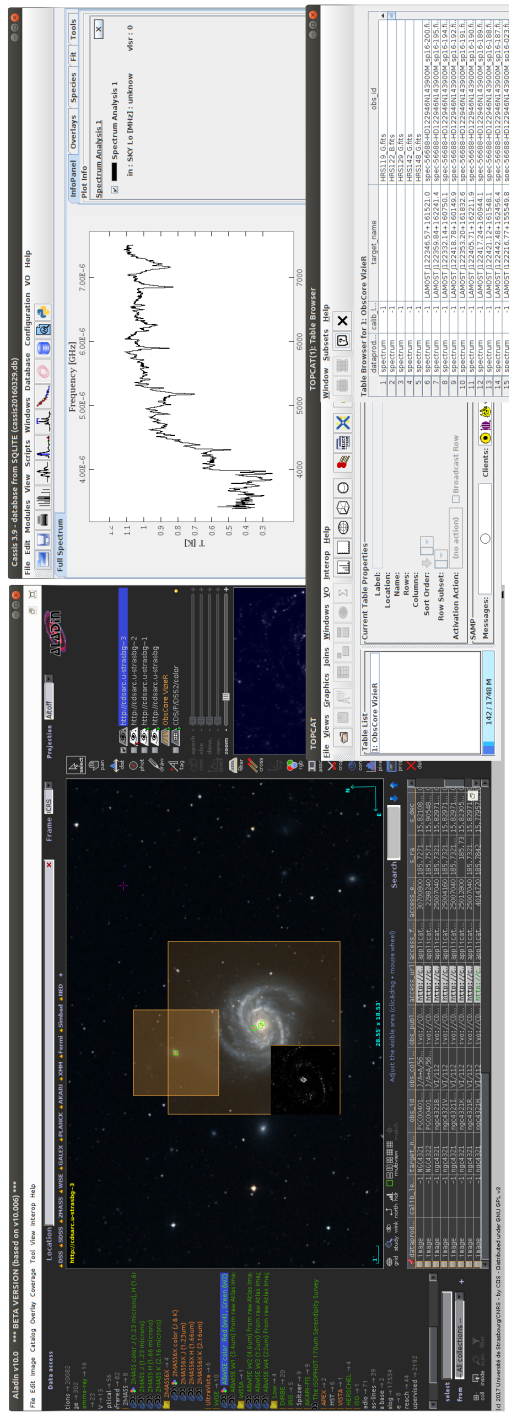

**Figure 3.** Examples of VO tools accessing the associated data service (Aladin, CASSIS, TOPCAT).