VizieR catalogue system certified by the Data Seal of Approval

Gilles Landais[1,⋆], *Françoise* Genova[1,⋆⋆], *Jean-Yves* Hangouët[1,⋆⋆⋆], and VizieR Team[1]
[1]*Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, F-67000 Strasbourg, France*

Abstract. Astronomy is a fertile environment with collaborations leading to the elaboration of disciplinary standards. Deploying standards in data centers can be beneficial for long term preservation and it puts the discipline in the open data era.
VizieR [1] is the CDS service dedicated to catalogues. It gathers data and tables, but also associated data like spectra or images coming from published papers or collaborations with various agencies. VizieR gives access to data through services to the astronomical community and preserves data for the long term. Several years ago, it became clear that it was worth applying for an external evaluation to certify that VizieR is a trustable data repository.
The Data Seal of Approval (DSA) [2] is granted to repositories that are committed to archiving and providing access to scholarly research in a sustainable way. CDS applied to the DSA in particular for the VizieR catalogue service. The VizieR information system is based on accurate data documentation provided by specialists. The data life cycle from ingestion to access and preservation is presented, with reference to the Open Archive Information System (OAIS [3]).

1 Open Data in VizieR

1.1 Role of the data center

The CDS VizieR catalogue service, created in 1995, provides the astronomical community with free access to data. Ingestion workflows have been in action since the beginning. VizieR uses a standardized description of the resources which is provided by specialized documentalists, in some cases with an initial description provided by the author or the journal.

The CDS is clearly a major actor in Open Data in astronomy and beyond. Open data is currently a "hot" concept [4], demanded by authorities for publicly funded science data.

The first mission of an Open Data Center consists of preserving the data and giving free access. In return users have to cite the data origin (author, article) with a persistent identifier. VizieR provides a bibcode (the bibliography identifier used by the astronomical community [5]) and internal identifiers which are both persistent and enable citations and access. Furthermore, a Virtual Observatory

⋆e-mail: gilles.landais@astro.unistra.fr ORCID: 0000-0003-4868-5873
⋆⋆orcid: 0000-0002-6318-5028
⋆⋆⋆orcid: 0000-0001-7338-8075

identifier (ivoid) is assigned for tables, enabling data reuse in a larger context.

The second mission consists of providing data that is useful and usable for research. Usability is well described by the FAIR (Findable, Accessible, Interoperable, Reusable) [6] principles, and leads data centers to implement vocabularies, formats, protocols, identifiers, etc.

The data quality depends on the assigned meta-data. In VizieR we distinguish "basic" meta-data and "rich" meta-data.

- Basic meta-data is the minimal description required for understanding the data.
  e.g. for a table : column description, units, etc.
- Rich meta-data enables data to be reused by software.
  In astronomy, the IVOA (International Virtual Observatory Alliance) elaborates advanced standards in particular for registries, data models, protocols and vocabularies which enable FAIR interoperability. Providing data in the Virtual Observatory framework should be compulsory today for an astronomical data center.

1.2 The Open data challenges

VizieR, as is the case for every data center, is between the data producers and the consumers which each have their own demands.

The producers for VizieR are the agencies which run space and ground-based telescopes, teams which produce large surveys, and the academic journals, with huge data diversity and increasing volume, especially for large surveys catalogues with tables over a billion of records (e.g. Gaia mission).

The consumers are the astronomers, research pipelines or software which require data and meta-data quality that we have to ensure. Quality has a cost. For example, it takes resources to find the information necessary to produce the rich meta-data of the Virtual Observatory standards.

Responding to these demands is a challenge, which becomes possible through interaction and cooperation with others. Firstly, cooperation with data producers is needed to improve the ingestion pipeline and build the initial knowledge about the data. Secondly, CDS long term success demonstrates that pooling different skills is beneficial: at CDS, astronomers, documentalists and engineers work together.

These collaborations locally and with data producers are needed to jump on the big data bandwagon, which is moving faster and faster.

VizieR is a reference service for catalogue dissemination recognized by its funding bodies, its partners and the global astronomical community. The next step was to get certification that VizieR is a trustable repository through an external evaluation. The CDS had already been certified as a member of the ICSU World Data System (WDS - http://www.icsu-wds.org/ [7]) community of trusted data services for Global Science. We decided in 2014 to submit an application to the Data Seal of Approval for VizieR, which fulfils preservation tasks in addition to data dissemination.

2 The Data Seal of Approval

2.1 The criteria

The Data Seal of Approval (DSA) is dedicated to scholarly research data repositories. It is a trust-based certification, simpler to get than the ISO one. It does not include audit visits but is based on a

self-assessment of 16 criteria, which is then examined by the DSA Board which grants the Seal. The label is provided for 3 years before needing renewal.

The DSA worked with the WDS under the auspices of the Research Data Alliance to align their certification criteria and processes, resulting in an updated catalogue of common criteria in 2016, which better underlines the importance of data reuse. The two organisations established a common Certification Board in 2017 a common Certification Board.

The DSA criteria cover different aspects of the information system:

- Organization : these criteria include the mission of the data center, the continuity of access strategy, etc.
- Data management : there are several criteria to describe the workflows from input to output including identifiers and meta-data assignment, integrity checking and availability of scientific expertise.
- Technology and security is composed of 3 criteria on data preservation aspects, architecture, monitoring and preservation plans.

2.2 DSA evaluation

The DSA proposes an auto-evaluation with a compliance scale composed of 5 levels depending on the implementation status for each criterion. Among the levels, one is "Not applicable." For instance, in VizieR the criteria concerning confidentiality and ethics is not applicable because VizieR data are not subject to disclosure risks.

The DSA board (now the common DSA-WDS one) examines the responses and the documentation provided by the data center. Peer evaluation is performed by one member of DSA and one from WDS. The reviewers write a report to the Board which decides whether to award the Seal.

During the VizieR evaluation, which was under the sole DSA Board responsibility at that time, we got comments from the DSA reviewers on our initial application. It was helpful assistance to improve the documentation. Finally, we submitted a more complete response with linked documentation provided on our website. The description of the CDS processes in the OAIS framework described in the next section was among the documents provided for review.

Writing the documents submitted for review was a team effort at CDS. The initial work was mainly conducted by the technical staff: software engineers and people in charge of the network and system infrastructure. The documents were completed and reviewed by documentalists and the CDS governance.

For the renewal this year, the documentalists contributed to the writing of the documents, which resulted in a more precise description of the workflows.

The DSA review was also a guideline. It helped us to improve our Information System by providing a good list of relevant questions. It resulted in some improvements. As an example, checksums are now computed for each table ingested in the pipeline and verified during the data life cycle.

2.3 Describing the Information System with OAIS

DSA proposes to use OAIS (Open Archive Information System) to describe the data center processes. OAIS is registered as an ISO norm (14721:2012); it defines a reference model and an editing guideline to describe an Information System (including the system, architecture, staff).

OAIS is dedicated to open archives. Data and meta-data are inseparable in the OAIS concepts, they are gathered into information packages and evolve together during the whole data life cycle.

In the next sections we will describe some of the OAIS concepts:

- the Information System described in six OAIS entities (section 3.1)
- the three types of information packages of the OAIS norm (section 4)

3 VizieR described in the OAIS norm

3.1 The OAIS entities

The OAIS norm divides the Information System into six entities:

- "Entry", "Archival storage" and "Access" are the entities related to the workflows which manage data and meta-data through the different steps of the data life cycle from ingestion to output. "Archival Storage" also includes preservation and mirroring aspects.

- The processes are conducted by the "Data management" entity which includes rich meta-data assignment and submission validation.

- The workflows are supervised by the "Preservation Planning" entity which assures the migration and monitoring strategy, redundancies and technology.

- Finally, the "Administration" entity administrates the whole system. It includes the governance of the data center, the relations with producers and users, etc. The Administration also validates the technology proposed by the Preservation planning entity, etc.

The six entities are completed by "Management" which is outside of the Information System. At CDS, we have the Scientific Council composed by an international committee that evaluates the CDS scientific impact. They audit the CDS each year and write a report which is a guideline for the next year.

Figure 1 shows the VizieR information System described in the OAIS-like norm.

VizieR benefits from for than 20 years of experience, complemented by the even longer term expertise of the CDS in astronomical data management and dissemination which has been built up since its creation in 1972. In particular VizieR benefits from well established workflows for data description and of CDS expertise in the implementation of the Virtual Observatory standards. Both were key elements of VizieR in the description used to obtain certification.

3.2 VizieR Workflows according to the CDS data curation

The data provenance as well as the content (including meta-data) are defined by the Administration entity of the CDS, which is composed of the astronomers and the governing body.

Depending of the origin, the data are submitted by authors or documentalists into the "Entry" entity. Documentalists are the building block of the data ingestion. They offer assistance to authors for their data submission and then check and complete the description or fully write the description. Finally, the astronomers validate the documentalist's work and allow access to the public database.

Data curation includes homogenization of the formats and assignment of basic and rich meta-data. There is also a data quality validation for each document ingested.

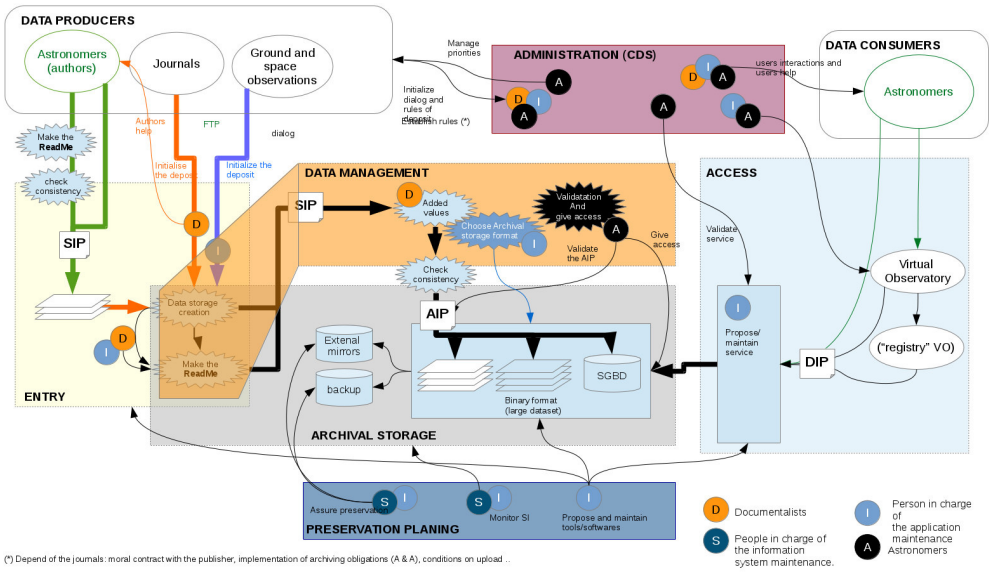

Figure 1. The VizieR Information System described with OAIS

4 The VizieR Information Package life cycle in OAIS norm

4.1 Input Data

The input data, or the "Submit Information Package" (SIP) in the OAIS norm, is the data submitted by authors or generated by documentalists.

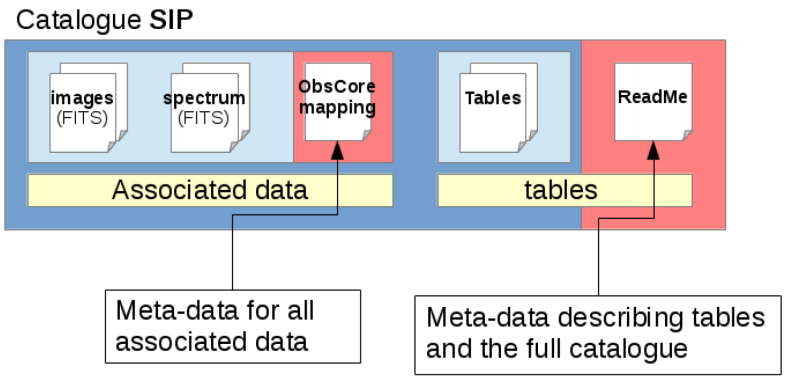

Figure 2. The VizieR Submission Information Package (SIP)

In VizieR, the meta-data of the SIP are gathered in the ReadMe file. The ReadMe file describes the catalogue and the tables with meta-data including the authors, the abstract, references links, columns descriptions, etc.

Associated data [8] such as images or spectra, often in FITS format, are described with the ObsCore Data Model [9] of the Virtual Observatory. The ObsCore meta-data of the FITS documents is separated from the ReadMe and provided as key-value mapping.

The tables are submitted in the CDS ascii format with aligned columns. The CDS provides tools for authors to migrate CSV, FITS or VOTable into the CDS ascii format. This format is adapted for long term preservation; it is independent of the technology and is described with a byte-by-byte section in the ReadMe file. All the columns of the tables are described with their unit, data type and a description text.

4.2 Archival Storage

Data are preserved by the "Archival Storage" entity. Data and meta-data are the Archival Information Package (AIP) described in the OAIS norm.

The AIP contains all data and meta-data available locally. In VizieR, it includes the basic and rich meta-data with the identifiers, but also some private data useful to rebuild the system, like the history of the ingestion processes, the email correspondence with authors, etc.

The AIP are the result of curation by the CDS documentalists. This process is the core of VizieR, and is required to provide useful data and added-value.

Among the meta-data, a configuration file completes the ReadMe file. It includes:

- the United Content Description (UCD [10]) for each column which are used by VO enabled software;
- links to other VizieR catalogues or external links;
- target name resolution to add positional computed columns;
- plot visualization for time series;
- photometric system and filters for photometry columns.

Finding the information is not trivial, for example looking for the photometric meta-data requires particular attention. Providing this kind of information cannot be automatized and the role of specialized documentalists is crucial.

4.3 Data Provision

The data provided by VizieR are the Diffusion Information Package (DIP) described in the OAIS norm.

In OAIS, the DIP includes the information package and the dissemination protocols. The Virtual Observatory standards including data model, formats and protocols constitute a framework which is a good example of what can be a DIP.
As an example, the VO registries [11] index the astronomical resources in a common list of resources independent of the data center. All resources have an identifier (ivoid) and are linked with a description. The registry technology is based on the OAI-PMH protocol, a standard used for instance in the digital library context or in EUDAT B2FIND registry (https://www.eudat.eu/services/b2find), which allows federation with other resources.

5 Conclusion

VizieR DSA certification was facilitated by more than 20 years of experience with well established workflows which are already in place for data management and description. The implementation of the Virtual Observatory standards was the second pillar that enabled us to document the CDS response to the criteria. These standards match well with the dissemination requirements of the OAIS norm.

The documentation in VizieR is provided by specialized documentalists. Their work is crucial to the high quality of the service and to gathering the information needed to build and rebuild the system.

The full certification process required about 3 person month. It was team work to which all the team members contributed. While there was no immediate update of the system there was better documentation, which was necessary to obtain certification. This demonstrated the high quality of the processes built over time, and the detailed assessment of each criterion led to technical improvements. Some of them are still in development today. One of the most interesting output was the end-to-end description of the processes, which led to a better understanding by all of the whole system, and of their role in it.

We would also like to express our gratitude to the VizieR team for their work: the documentalists P.Vannier, M.Brouty, E.Perret, S.Guehenneux, T.Pouvreau, the astronomers P.Ocvirk and C.Bot and the engineers T.Boch and F.X.Pineau

Also, many thanks to F.Ochsenbein who built VizieR.

References

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

[2] I. Dillo and L. de Leeuw, https://www.datasealofapproval.org

[3] CCSDS, *Reference model for an Open Archival Information System* (OAIS), Magenta Book. (Issue 1) (2002)

[4] F. Genova, *Data as an infrastructure: CDS, the Virtual Observatory, astronomy, and beyond*, These Proceedings (2018)

[5] M. Schmitz, G. Helou, C. Lague et al., *NED and SIMBAD conventions for bibliographic reference coding* in *Information & On-line Data in Astronomy*, edited by D. Egret and M.A. Albrecht. Astrophysics and Space Science Library, Vol. 203. (1995) p. 259-270, ISBN: 978-94-010-4178-2

[6] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg et al., Scientific Data 3, art. num. 160018 (2016)i DOI: https://doi.org/10.1038/sdata.2016.18

[7] I. Gärtner-Roer, S. Harrison, S. Sorvari, Geophysical Research Abstracts **19**, EGU2017-4917, 2017, EGU General Assembly (2017)

[8] P. Ocvirk, G. Landais, L. Michel et al., *Associated data: indexation, discovery, challenges and roles*, These Proceedings (2018)

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

[10] S. Derrière, A. Preite Martinez, R. Williams et al., *An IVOA Standard for Unified Content Descriptors Version 1.10*, ADS bibcode: 2005ivoa.spec.0819D (2005)

[11] M. Demleitner, P. Harrison, M. Molinaro et al., *IVOA Registry Relational Schema Version 1.0*, ADS bibcode: 2014ivoa.spec.1208D (2014)