Organization of VizieR's Catalogs Archival

Table of Contents

Foreword

In this document, we are describing the organization of VizieR archives and how the different pipelines interact with the tools and staff of the CDS. We were inspired by the OAIS organization which offers a well adapted organizational model, without however pretending that the Vizier archival is an OAIS archival.

This document is based on OAIS material from the French Archives website
http://www.archivesdefrance.culture.gouv.fr/
http://pin.association-aristote.fr/lib/exe/fetch.php/public/documents/norme_oais_version_francaise.pdf

OAIS definition: Open Archival Information System
The term 'Open' in OAIS is used to imply that this Recommendation and future related Recommendations and standards are developed in open forums, and it does not imply that access to the Archive is unrestricted

Archive definition:
An organization that intends to preserve information for access and use by a Designated Community.

This term covers the whole of activities, from the initial archival management to the preservation and access of archived files.

Therefore, the archive ensures sustainability of data up to preservation and access to archived files but also preservation, with data, of all necessary information to its comprehension and use (metadata).

"Long-Term" definition:
A period of time long enough to be concerned about the impacts of changing technologies, including support for new media and data formats, and of a changing Designated Community, on the information being held in an OAIS. This period extends into the indefinite future.

OAIS packages definition:
The OAIS norm has 3 package format:

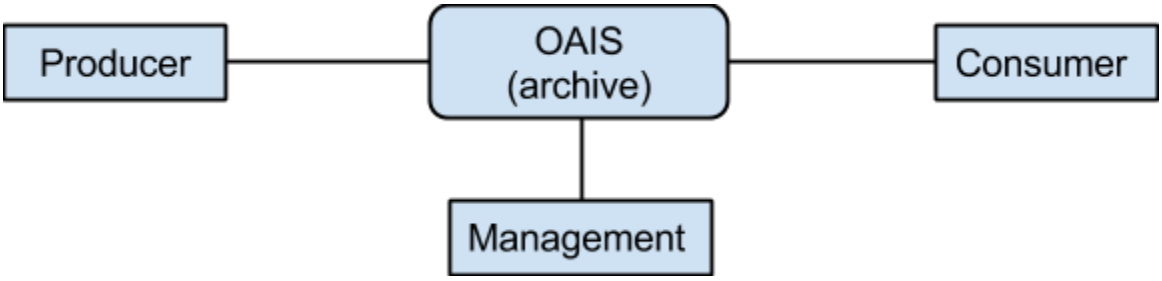
- **SIP**: Submission Information Package
Data delivered by the producer. This data contain the content + a part of metadata useful for archival.

- **AIP**: Archival Information Package
Data archived in the Information System only accessible in the OAIS.

- **DIP**: Dissemination Information Package
Gather distributed information.

Environment applied to VizieR archives

The diagram below reproduce the workflow of an OAIS organization:



The archive

The VizieR archive consists of elaborated scientific data (*science-ready*) from journals or data of ground and space observations.

This data can be:
▪ tabular data from observation catalogs, compilations, models, etc.
▪ associated data: spectrums, images, etc.

The producer

The data producer can be:
▪ scientist(s) behind the publication of the astronomical catalog to ingest into VizieR.
▪ the ground and space observatories, agencies, or teams which elaborate the results of their observations or of their modelisations.

The user

The VizieR data is intended for the scientific community. The **knowledge database** (OAIS concept) regroup the astronomical concepts required by the Astronomer’s profession as well as the knowledge of standard formats used by the community (FITS format for example) which can be exploited by well known software or libraries of the field.

The management

The CDS is an infrastructure of the French Institute National des Sciences de l'Univers of the Centre National de la Recherche Scientifique, managed in agreement with the University of Strasbourg.

A scientific council (6 French and 6 foreign members) examine its activities every year.

The notion of “open” archive

We’re examining here, how the notion of “open archive” can be included in VizieR.

▪ The CDS Council (see The management) is a source of propositions. Its executive role of guiding the choices of the CDS is fundamental.
▪ The Virtual Observatory (VO), which aims at providing seamless access to astronomy on-line resources by defining interoperability standards and protocol, is a strong component regarding the orientation of the choices of formats and technologies (http://www.ivoa.net/).
▪ Users benefit from customer service (“questions”) which allows a direct link with the CDS staff and the demonstration of services.
▪ The CDS listens to the users, proposing solutions to the main astronomical journals: recommendations, help and tools for the electronic publication of scientific data. The CDS has been collaborating for more than 20 years, particularly, with the international journal “Astronomy & Astrophysics”, but also with the other major academic journals.
▪ The CDS takes part in international meetings (l’AAS, Interoperability,...) ans demonstrate its services during sessions or on demo booths.

Data exchanges

Note: the notion of OAIS packages designate the data format used in the exchanges between the OAIS (VizieR) and the external participants (producers, users).

In VizieR’s case, the packages (SIP,AIP) consist of a group of files: tabular data, other type of data files (images, spectra,...) and the ReadMe file (describing the whole).

Information provided by the producer (SIP):

The OAIS standard differentiates the content and the meta-data specific to the sustainability of the object to archive. Both are necessary for the OAIS storage.

In VizieR’s case, the information content consists of tabular data and associated files which complies the format described on the page: http://cds.u-strasbg.fr/doc/catstd-2.htx

The sustainability information (meta-data) is provided by the producer (see section Standards for data deposit intended for producers). Those information are then checked and completed by authorized CDS staff (AIP data).

The VizieR meta-data is regrouped in a “ReadMe” file which includes:

▪ the catalog’s name, its author(s), the year of the article's publication, as well as its bibliographical references (including year of publication and bibliographical code)
▪ keywords used for researching catalogs
▪ Catalog’s tables and associated files description (title, file names and characteristics like file number of lines in tables)

▪ for each table, column descriptions: column names, type of data and units used; the written descriptions associated to each column contain necessary details for a scientific exploitation of the tabulated values.

Archived information (AIP)

These information are only available through VizieR.
Ingested information (tabular data + associated data + ReadMe file) are analyzed by competent and authorized CDS staff. This analysis builds new data and meta-data files used by the ingesting software in the Information System.

Additional meta-data

Additional meta-data consist of:

▪ Add-ons in the ReadMe file: additional standardized keywords, data origin…
▪ ASCII file (.Summary or .status) which regroups complementary information
- standardized descriptions (Unified Contents Descriptors) recognized by the Virtual Observatory
- additions of positional columns (to place the astronomical objects in the sky)
- internal links (to other catalogs with added data stored in VizieR), links to objects of the SIMBAD database, or links to external resources (defined by URLs)
- descriptions of operations of tables fusion (merge) or juncture.

▪ SHELL unix scripts to treat additional data like spectrums or time series in graphic forms.
▪ (exceptionally) SQL scripts accompany the final formatting of results
▪ a dated log file relating the operations performed on the catalogs. For each action; the name of the executant and the execution date are indicated.

Operation example: insertion in VizieR, data made public, modification of data, meta-data…

New data

In parallel, results files emanate from original data in a format suitable to data research. It is tabulated data with **unmodified content**, but which could have undergone transformations in their representation or changed their format (for example, several FITS files can provide one table only). The exit format can be ASCII files or proprietary binary files (in the case of large volume, this format allows very fast queries).

Then, the ingest process uses a relational database to store the catalog's meta-data. Data itself is stored as database tables or binaries files as explained above.

Note: Concerning publications data, VizieR commit to conserve ingested tabular data in an ASCII format. This file format isn't necessarily the one of the original file.

Constitution of the final catalog (AIP)

▪ original data (ASCII files, images,...)
▪ database tables OR binary files
▪ new ASCII format files (if necessary)

▪ registration of meta-data in VizieR database
▪ a meta-data file in ASCII format for the reconstruction of the catalog
▪ the ReadMe file
▪ scripts for the representation of certain additional data like spectra, light curves
▪ private data like emails with the authors
▪ log files dating the principal added actions (accompanied by the name of the authorized person); the objective being to log the actions performed during the pipeline.

Distributed information (DIP)

Distributed information is the meta-data searchable through services maintained and developed by the CDS. It principally consists of:

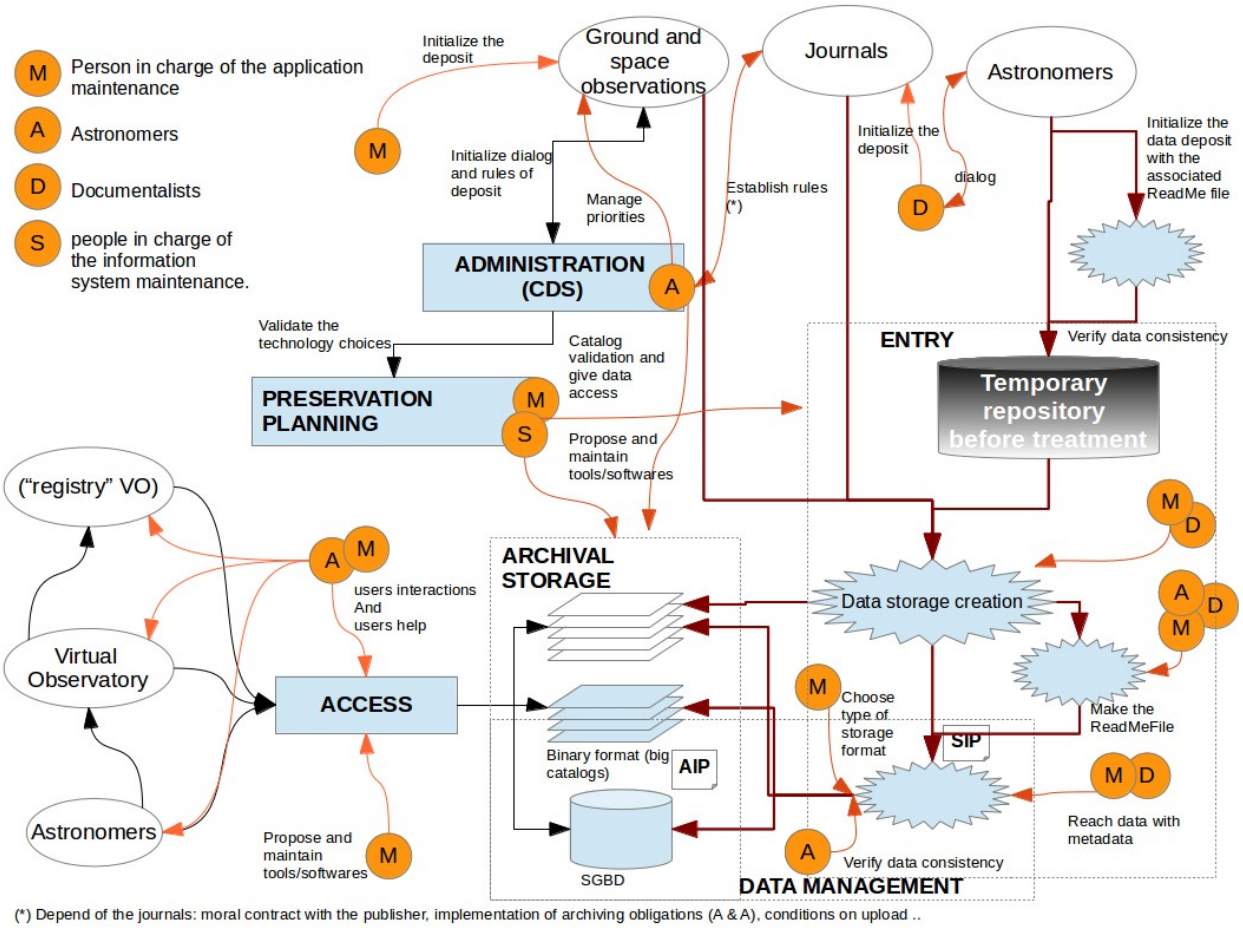
▪ multi-criteria web forms (http://vizier.u-strasbg.fr)
▪ services meeting the Virtual Observatory standards according to protocols defined by the IVOA (http://www.ivoa.net/ http://tapvizier.u-strasbg.fr/adql/)
▪ dedicated services as for the access to the large catalogs (http://cds.u-strasbg.fr/resources/doku.php?id=cdsclient)
▪ FTP service

Note: Original data of some ancient large catalogs is not guaranteed by the CDS, the CDS having not been responsible of the archival of original data at the time of the catalog ingestion. In that case, the access to the original is specified in the ReadMe file available to the user.

Description of VizieR pipeline

VizieR pipeline

In this section, we are describing how the organization of VizieR can be represented in an OAIS context.



Data reception

Note: This pipeline phase corresponds to the "INGEST" entity of an OAIS organization.
The treatment is in charge of the data reception (see Information provided by the producer (SIP)) and its development in the Information System (see Archived information (AIP)).

An authorized person takes care of data which can be deposited in a local FTP host, from a remote website or, in case of large catalogs, on a disk provided by the data producer (see section Standards for data deposit intended for producers). Data received is original data usually including an ASCII file, files in FITS format or binary format (for some large catalogs). Depending on the process chosen by the data producer, original data can be accompanied by a standardized file called "ReadMe". In that case, an authorized person will have to build this file using tools similar to those proposed to the producer by

searching the information in literature associated to the catalog. In all cases the file is signed by the creator.

The second step consist of adding meta-data as explained in the section “Archived Information (AIP)”.

Lastly, the authorized person generates the archival storage and workspace which remains private until the end of the pipeline. This step is dated and signed by the person.

Note: the authorized person can be:
▪ a documentalist
▪ an astronomer
▪ a person in charge of the application maintenance (for large catalogs)

Archives storage

Note: This pipeline phase corresponds to the “ARCHIVAL STORAGE” entity of an OAIS organization.

The archives storage process consists of:

▪ The storage space created in INGEST containing original data.
▪ Useful and enhanced data (SGBD and owner binary format) and meta-data updated by the data ingestion process developed by the CDS.
▪ Ingestion and/or modification processes of meta-data of a catalog (historized, signed and dated).
▪ Redundancy and backup.

The storage content consists of catalogs or AIP information (see Constitution of the final catalog (AIP)).

Digital storage

The entire storage uses a level 5 or 6 file system. A daily backup is done and monitored by a Nagios system (controllers state, power supply, discs, etc.).

Data duplication

▪ Public original data is copied onto a mirror website (the CfA: Center for Astrophysics, USA)
▪ the private archives are kept only at CDS
▪ the entire archives are duplicated onto a disk array, located in another building than where the data servers are.
▪ Data and meta-data (SGBD or owner format) are replicated onto 8 mirror sites.
This replication is specific to each mirror which may only have a partial replication, managed by a CDS personnel.

Recovery service

Redundancy of data and meta-data is available locally or on remote mirrors.

▪ Meta-data, stored in databases, have a local base at CDS used as replication. Furthermore, mirrors can be used. Those databases contain all meta-data et can therefore be used to rebuild a damaged base.

▪ Catalogs from journals stored in databases are replicated in another base in CDS as well as on at least one mirror. Redundancy can then be used to rebuild a damaged base.
▪ Large catalogs have at least one copy on one mirror.

In all three cases, a manual mechanism allows the configuration of data source used by VizieR software (in some cases the process of the source's choice is computerized).

All services are supervised by a dedicated Nagios server and a "GLU" mechanism. The GLU mechanism specific to CDS allows a better use of services by using resources redundancy and by activating the fastest responding service.

Finally, TAP/ADQL VizieR service possesses its own database and therefore constitute an other data access.

Data enhancement

Note: This pipeline phase corresponds to the "DATA MANAGEMENT" of an OAIS organization.

An authorized person takes care of the archive to build exploitable data. She/He uses the storage space generated by the "INGEST" phase.

Processed data is listed in the "Constitution of the final catalog (AIP)" section.

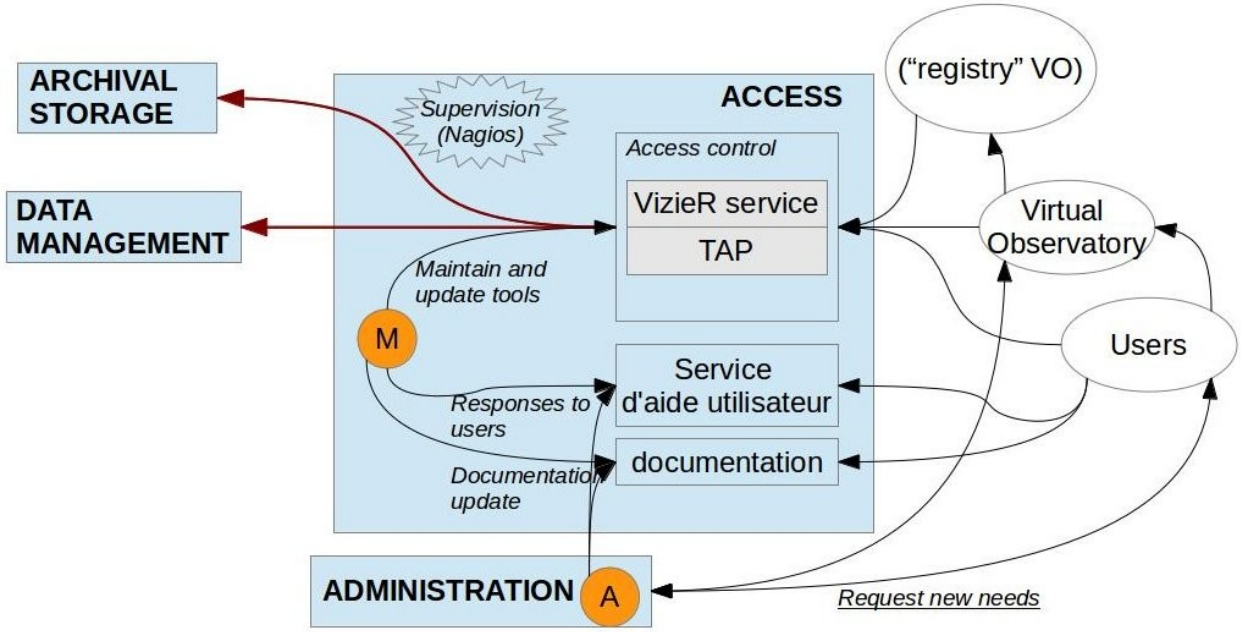
The VizieR pipeline updates VizieR meta-data. Those meta-data consist of added positional index, authors listing, keywords, photometry, etc.

The "ReadMe" file, as well as ingested data in the information system are validated by an astronomer.

Note: the authorized person can be:
▪ a documentalist
▪ an astronomer
▪ a person in charge of the application maintenance (for large catalogs)

Data access

Note: This pipeline phase corresponds to the "ACCESS" entity of an OAIS organization.



VizieR offers 3 archives access (see Distributed information (DIP):) through software developed and maintained at the CDS.

A customer service connects users with developers and astronomers.

Note: the authorized persons can be:
▪ an astronomer (for customer service and taking care of functional requests emanating from the Virtual Observatory or customer service)
▪ a person in charge of the application maintenance (for tools maintenance and taking care of new technologies or protocols emanating from the Virtual Observatory)

Astronomers part in VizieR archival

Note: This pipeline phase corresponds to the “ADMINISTRATION” entity of an OAIS organization.

Astronomers constitute the members of what we would call the “administration” entity of the OAIS organization.

Their part consist of:

▪ validating data/meta-data added by authorized personnel
▪ sharing their knowledge to enhance the archives (example: adding links to an external website of large surveys, offering relevant plots to illustrate catalogs)
▪ maintaining contact with data producers: main astronomical journals, space centers and large observatories producing large catalogs.
▪ maintaining contact with data users (example: taking care of functional requests sent to the hotline)

- participating in workshops, discussing data access protocols standardization (IVOA members)
- validating applications suggested by the people in charge of the information system maintenance.

Data sustainability

Note: This phase corresponds to the “PRESERVATION PLANNING” of an OAIS organization.

Data sustainability aims to a long-term preservation of the VizieR archives. It requires actions on both storage formats used on the base and the tools available that need updated based on new technologies or their obsolescence.

Distributed data sustainability from journals: at the entry point in the data processing pipeline, the accepted formats are standard formats (FITS, ASCII, …) which are kept in their own storage space. The use of open standards constitute a security for the long term reading of those documents.

Data sustainability from large catalogs: large catalogs’ original data is preserved as much as possible. However, those data formats can be binary and therefore, only exploitable through the producer center knowledge: the preservation of original data (when binary) is to be discussed with the producers. Original data is then transformed into a binary format specific to the CDS: the CDS performs an evolutionary maintenance of the access and creation tools of its formats.

Meta-data sustainability: meta-data of a catalog are stored in a dedicated space as ASCII files, making them independent of any technology. They are accompanied by a “ReadMe” file and a log file of the principal pipeline steps and updates of the catalog.

Used tools sustainability: the CDS uses open source softwares (PostgreSQL, Linux, GNU, …)

Data sustainability (see Archives storage)

Archives Interoperability: The Virtual Observatory, in which VizieR is involved, assures a data access and a data interoperability with other archives centers. For example, output format VOTable is a standard ingestible by the virtual observatory tools. Those standard tools offer gateways possibilities between data centers.

VizieR responsibility in the archival

We are listing below the mandatory responsibilities for an OAIS type archival:

1. Negociate with information producers and accept appropriate information on their part.

The CDS favors the dialog with journals which the catalogs are from. It can lead to regulated data deposit or to the definition of the CDS responsibility in archiving. In that case, the CDS abides by the international rules in effect.

The CDS takes into account the articles release dates before publishing the catalogs.

Large catalogs data are directly discussed with the producers to specify the most relevant data to publish in VizieR.

2. Acquire a sufficient knowledge of the given information, to guarantee its sustainability.

The CDS is accountable for the data content it publishes: which gives it the rights to modify the storage format based on the technological novelty or obsolescences.

3. Define, itself, or in collaboration with others, which communities must constitute the target users community able to understand the given information.

The target community is defined by CDS authorities, the “Centre National de la Recherche Scientifique” (CNRS) and the University of Strasbourg.
VizieR archives are primarily for professional astronomers.

4. Insure that the preserved information is immediately comprehensible by the target users community.

Meta-data enhances the content of catalogs and allows an indexation in line with the astronomer's research.

5. Apply a strategy and documented proceedings guaranteeing the preservation of information against all contingency within reasonable limits.

See section “Archive storage, Data sustainability”.

6. Allow the distribution of information.

See sections “Archives interoperability”, “Interfacing standards with external data centers”.

7. Authenticated copy of the original or allowing to get hold of the original.

The “ReadMe” file describes files (associated or tabular) of the catalog.

- Data from journals: original published data is preserved.

Note: Regarding data from publications, VizieR commits to keep input tabular data in an ASCII format. This file's format is not necessarily the ones of the original file but the content is fully preserved.
▪ For large catalogs: the "ReadMe" file indicates the data center which produced the original. Original data can, according to the agreement by the producer, be, or not, archived in VizieR.

8. Make preserved information available to the target users community

see section "Data Sustainability".

Among other things, the standards of the virtual observatory guarantee an exploitable output by tools developed externally (see Archives Interoperability).

Procedures in use

Here is a non-exhaustive list of proceeding or procedures in use for VizieR.

Procedures for data deposit intended for producers

VizieR data is from publications in astronomical journals, results of space or ground observations provided by national or international centers (ESA, ESO, …).

Procedures for published data (from journals)

If the CDS signs a contract with an editor, data are only published at CDS and it's the editor who takes care of the data processing (ex: A&A). He uses for that the data input tool available at http://cds.u-strasbg.fr/vizier/submit.htx .

Agreements with the editor can be accompanied by recommendations and help for the publication.

When no agreement exists, the CDS abides by the international rules in effect. The CDS take into account the publications dates before publishing the catalogs.

Procedures for data formats provided by large catalogs producers

The large catalogs volume from space missions and ground based surveys require a special treatment, currently customized for each catalog.

- Meta-data is built by an authorized CDS astronomer.
- Data content is defined with the data producer.

Particular fate for public data formats which aren't subjected to discussions with large catalogs producers

In the case of public data, the CDS can retrieve said data on producing organisms websites and then launch the input pipeline in the OAIS under the following conditions:

- a documentation is available for the columns explanation
- an effective device allows the data recovery (good network liaison for example)

The CDS builds meta-data file "ReadMe" associated with the catalog, reserving the choice of data to publish:

- choice of columns
- description of the catalog, tables, columns
- data enhancement (internal/external links, graphs)
- addition of keywords
- transcription of the rules of use and acknowledgement stated by the producing center

Original data are preserved as much as possible and if applicable, the “ReadMe” file will indicate the organization responsible for the original data archival.

Procedures of data deposit from journals

Data deposit methods available for the data producer are:

1. deposit service (http://cdsarc.u-strasbg.fr/cgi-bin/Submit) which includes meta-data deposit according to the format specified by “Standards for published data (from journals)”
2. secured FTP liaison (sftp)
3. email to a authorized CDS personnel

Furthermore, authorized CDS staff (documentalists or astronomers) can search for public data in licensed journals.
The deposit service (1.) is highly recommended. It allows a faster ingestion of data , creation of the bibliographic code and VizieR pipeline. This method is required for publication coming from A&A.

Procedures of data deposit for large catalogs

A discussion between the CDS and producing center representatives is encouraged. Where appropriate, for large surveys with public content, authorized CDS staff search the information on data producer websites.

- The request of a large catalog ingestion can be initiated by the data producing center or by CDS staff.
- It is usually followed by a discussion in order to:
    - remind the CDS role in the archival and data provision (http://cdsweb.u-strasbg.fr/about)
    - establish the data distribution policy: private data, date of distribution to users
    - establish terms of data recovery (SIP): reception date, pipeline and update frequency, recovery method (ftp, rsync, copy on disk, etc…)
    - An explanation allowing the scientific exploitation of data is required
    - agreeing on data to distribute (DIP). For tabular data, the respective managers select the columns to distribute.
    - Define acknowledgements as well as possible enhancements by the CDS.

Interfacing with external data centers

The CDS offers an access to its services via different Virtual Observatory standards. For VizieR, those standards consist of:

- a possibility of data retrieval in VOTable format (http://www.ivoa.net/documents/VOTable/)
- tool usage like websamp (http://astrojs.github.io/sampjs/)
- *conesearch* protocols (http://ivoa.net/documents/REC/DAL/ConeSearch-20080222.html) or TAP (http://www.ivoa.net/documents/TAP/)

Note: on-going studies are made to propose compatible output with the Virtual Observatory for images, spectra and temporal series.

Procedures of data's numerical identification

Each catalog is named in a unique way according to the journal the article has been published in or according to the data type. This naming is used by journals (see A&A website).

A description of the naming is available at: http://cds.u-strasbg.fr/doc/catstd-2.htx

Note: This nomenclature allows to find data via FTP.

The bibliographic link is added to the ReadMe file when it is known. In VizieR, it is the main "bibcode" number from which the data is obtained (see http://cds.u-strasbg.fr/simbad/refcode/section3_2.html)

Procedures of protocols used to search for archived data

Tools developed at CDS allow the extraction of information from meta-data.

VizieR uses the Virtual Observatory standards to index those data:

▪ UCD (Unified Content Descriptor) standard is used to described the columns in a standardized way.
▪ A pipeline updates the Virtual Observatory "registry of resources" for a data visibility within the Virtual Observatory.

Procedures for data archival not yet published

If the article is yet to be published, VizieR allows a pre-ingestion of the catalog: preserving it in a private state or with limited access. The catalog will be made public:
▪ after publication of the article in the journal.
▪ in agreement with the producer for large surveys.