CDS Scientific Council meeting 2023

Summary of CDS activities 2022-2023

23 November 2023

1. Introduction

The period 2022-23 has been very busy for the CDS. The most important news is that two new permanent positions were attributed to CDS in 2023 and successful recruitments have been made (see 2.1 for details).

The core CDS operations for the ingestion of data from publications have involved the treatment of a large volume of data, and new versions of services have been released. High priority data sets have been ingested (e.g. Gaia Focused Data products). CDS has continued to be strongly engaged with the astronomy community in meetings, workshops and conferences, some of which were held in-person for the first time since the pandemic period. We have also interacted with all of the CDS partners (ESO, ESA, CNES, SAO/ADS, NED and the main astronomy journals). The CDS also became active in the SKA Science Region Centre prototyping activities.

Some of the CDS 50th anniversary events carried over into 2023, in particular the CDS-50 exhibit booth at the January 2023 AAS meeting in Seattle, and the 'Open Science in Astronomy' workshop at the French national SF2A conference in Strasbourg, June 2023.

The changes in the scientific, technical and policy environments in which the CDS operates have had a large influence on the activities in the past year. CDS has participated in many events of the the French 'Recherche Data Gouv' initiative in its first year of development. We have also followed and participated in the implementation of the European Open Science Cloud (EOSC) via the ESCAPE and EOSC Future projects where we have helped the integration of the VO framework into these more generic open science systems. We also followed the global open science developments in the Research Data Alliance (RDA) and the World Data System (WDS).

A number of the European projects in which CDS participates reached their final stages, with a lot of effort applied to the preparation of deliverables and reports. These projects have been successful with many benefits for CDS, yet they have also been very demanding in terms of reports and management. A new large European proposal was prepared and submitted, with an unfortunately negative result. Other national level proposals have been made with results pending.

In this report we present the status of CDS in Section 2 and the highlights of the 2022-2023 period in Section 3. The activities of the CDS services are in Section 4 based on short reports provided by the different service teams. News about recent and current projects is outlined in Section 5. Section 6 provides a response to the recommendations of the CDS Scientific council from 2022.

2. Status of the CDS

CDS maintains its status as a "Research Infrastructure" on the French National Research Infrastructure Roadmap[1], the "Feuille de Route", established by the Ministry of Higher Education and Research (MESR). This was last renewed in 2021.

CDS is a "Service National d'Observation" (SNO) defined by the INSU Scientific advisory board, of type AA-ANO–5 with ObAS as the coordinating OSU. CDS has the status as a 'Scientific Platform' in the CORTECS network[2] of the Strasbourg University. CDS is named as a "Thematic Reference Centre" in the French national initiative called "Recherche Data Gouv[3]" which is an "ecosystem for

[1]French version - https://www.enseignementsup-recherche.gouv.fr/sites/default/files/2022-03/feuille-de-route-nationale-des-infrastructures-de-recherche---2021-v2--17318.pdf

[2] https://cortecs.unistra.fr/plateforme/?tx_ameosplatforms_platformviewerdetail%5Bplatform%5D=172&cHash=cdde73cf884808b85990e907cb29c804

[3] English version: https://recherche.data.gouv.fr/en , French version: https://recherche.data.gouv.fr/fr

sharing and opening research data" (CDS is beginning to set up the activities associated with this status and a proposal has been submitted in a call of the National Fund for Open Science to support some activities. CDS also has a minor role in the Alsace Helpdesk Data management cluster called ADELE managed at the Strasbourg University.

CDS applied in October 2022 for the CoreTrustSeal certification of the CDS Vizier and Aladin services. The application concerned 16 criteria (a 34 page document was submitted). The initial result was received in February 2023, and we responded to the 2 referee's comments in May 2023. We await the final approval. This process is unexpectedly long, but we expect it to be finalised before the end of 2023.

In 2022 the CDS was evaluated as part of the Observatoire Astronomique de Strasbourg by the HCERES *(Haut conseil de l'evaluation de la recherche et de l'enseignement supérieur).* The result of the evaluation is available publicly on the HCERES site[4] (in French only). Responses to some of the HCERES recommendations will be addressed during the council meeting presentation.

2.1 CDS Personnel

This year has involved a number of important changes in the personnel attributed to the CDS:

Two new positions:

Following discussions with CNRS-INSU about the CDS resources in early 2023, two permanent positions were opened by CNRS in 2023; a documentalist position and a research engineer position. These positions have now been filled by Katia van der Woerd as a documentalist, and Matthieu Baumann as a research engineer, both of whom were previously employed on CDS contract positions. The opening of these positions and the successful recruitments represent a major success, and we thank everyone who helped to make it happen.

Changes in permanent staff:

An administrative assistant (C. Steyer) has been transferred by CNRS into ObAS to support the CDS (an 80% share of a position shared with EOST) in October 2023. This is very welcome since this role has been vacant for 1 year. This situation will however finish at the end of 2023 because C. Steyer will take a leave of absence. We hope that CNRS will maintain this position for ObAS/CDS. Another change in the permanent staff is the departure of Evelyne Son (documentalist) which is expected at the end of 2023.

Changes in contract staff:

There have also been a number of changes in the contract staff over the year. The contracts for a system engineer (M. Misslin) and a project engineer (H. Heinl) came to an end. Three other contractors finished (or will finish) their contracts earlier than expected in 2023; S. Amodeo (CDS Postdoc), A. Fiallos (documentalist) and A. Flint (research engineer).

A recruitment of a 'Open Science Researcher' contractor was made in August 2023 (A. Gonneau) following her short-term contract on the EOSC Future project. Also an 'Open Science Engineer' contract position was taken up by M. Marchand following her role as a project engineer in the EOSC Future project.

A contract documentalist position is currently open.

[4] https://www.hceres.fr/fr/rechercher-une-publication/obas-observatoire-astronomique-de-strasbourg

3. Highlights 2022-2023

Open Science in Astronomy workshop, and planetarium display at SF2A 2023.

CDS led the organisation of the ‘Open Science in Astronomy’ workshop at the national French astronomy meeting in Strasbourg in June 2023. Also an interactive display was made of research data on the dome of the new Strasbourg Planetarium during conference.

Aladin Lite version 3.

Aladin Lite version 3 was released in January 2023 (Fig 1). This major new version has an improved display thanks to GPU rendering with WebGL2, access to FITS HiPS tiles, multiple projections and many other new features. Aladin Lite v3 was supported in part by the ESCAPE project.

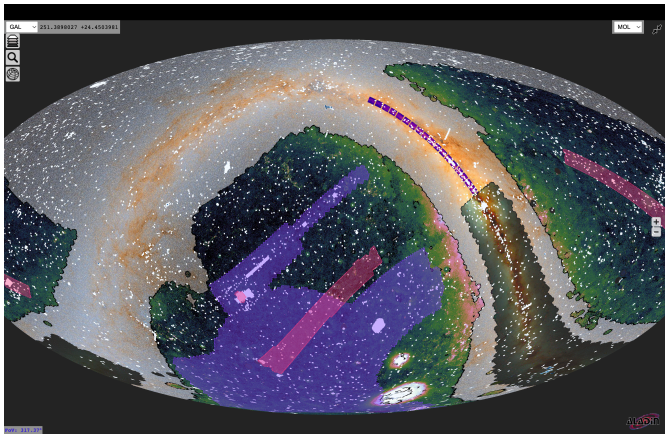

Figure 1.

The new Aladin Lite version 3 showing HiPS of the Gaia DR3 flux map, DECaPS DR2, ATLASGAL APEX, and the DESI Legacy survey. MOC overlays are shown for DES DR2 (violet), VISTA Viking (pink) and XMM (white).

Accelerated growth of the SIMBAD database.

The CDS SIMBAD database grew by an exceptional 3 million astronomical objects in 2022-23 due to the ingestion of a number of large surveys. This reflects an important (and expected) change in astronomy surveys which increasingly include spectroscopically characterised sources that make them suitable for inclusion in SIMBAD.

Relocation of the CDS servers in the UNISTRA data centre and IPHC server room.

The final moves of CDS servers out of the ObAS server room into larger mutualised facilities has been completed in 2023 (Fig 2). The historical server room in the basement of the South Building at ObAS has served the CDS well, but we now have more room to grow and greater infrastructure stability.

Figure 2.

T. Keller (ObAS Information System engineer) during the major move of CDS servers to the University of Strasbourg Data Centre and to the IPHC server room.

3.1 Highlights of CDS in the Community

The CDS has actively participated in a number of astronomy community events where we have interacted with researchers, partners, journals/publishers, other data centres, observatories, missions and projects.

Highlighted international community events:

- **IVOA Interoperability Meeting,** Virtual, 17-20 October 2022.

- **Astronomical Data Analysis Software and Systems (ADASS) Conference,** Virtual (Hosted by Canadian Astronomy Data Centre, University of Toronto and the University of Victoria), 31 October - 4 November 2022.

- **American Astronomical Society (AAS) winter meeting, Seattle, USA, January 2023**

CDS Exhibit booth & Exhibit Theatre presentation (M. Allen, M. Baumann, T. Boch, M. Buga)

- **IVOA Interoperability Meeting,** Bologna, Italy, 7-12 May 2023.

- **European Astronomical Society annual meeting, Krakow, Poland, 10-14 July 2023:**
  - **CDS stand** (M. Allen, S. Derriere, P. Fernique, A. Gonneau, M. Marchand, G. Monari, A. Oberto, B. Vollmer)
  - Special Session: ***"Science with the Virtual Observatory: status, success cases, the future"***
    - *Invited presentations:* M. Allen, P. Fernique, *Posters:* A. Gonneau, M. Marchand

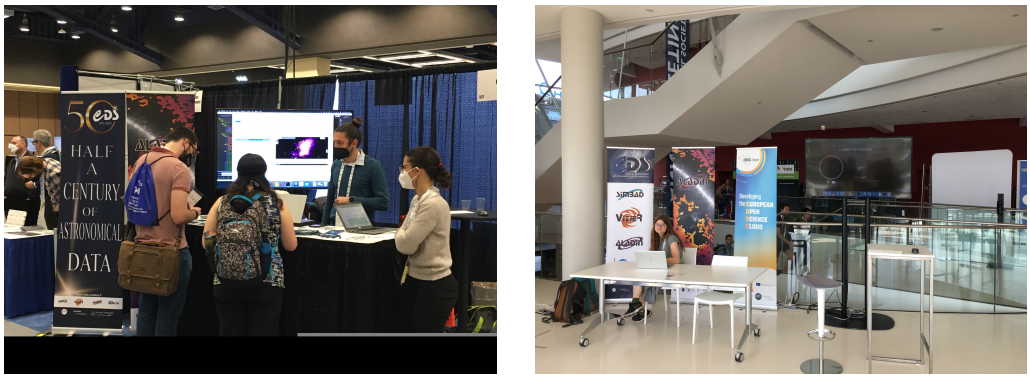

Figure 3. CDS Exhibit booths at the AAS 2023 meeting (left) and EAS 2023 Conference (right).

Highlighted events in the French community:

- **ASNUM2022:** Conference of the 'Action Spécifique Numérique', Lyon, 12-16 December 2022. Invited talk: *50 years of CDS, CDS today and future challenges*, M. Allen.

- **Technical Workshop EOSC-France**, Strasbourg, 24-26 January 2023. Invited talk: *Experience of CDS and ESCAPE-VO on-boarding to EOSC*. (M. Allen, M. Marchand, S. Derriere, S. Amodeo, H. Heinl, A. Schaaff, M. Molinaro (ESCAPE/INAF))

- **Spring-time of data** ('Printemps de la donnée') webinar presentation on Standards and Meta-data (Aline Eisele, Emmanuelle Perret, Mihaela Buga, Soizick Lesteven), 13 June 2023.

- **Recherche Data Gouv seminar**, 7 June 2023, Presentation of Thematic Reference Centres (Aline Eisele, Caroline Bot, Soizick Lesteven)

Social media communication channels: Facebook (@CDSportal) and Twitter (@CdSportal).

4. Activity Report for CDS Services 2022-2023

4.1 CDS Information System

The CDS services continue to have a very high level of use with an average of 3.0 million queries per day from 390K unique IPs/month in 2022-23 (Table 1). We continue to aim for 24/7 operations, but supported by personnel with a classic 5-day work week. Fig 4 shows the monthly queries on a 9 year timescale showing a general increase but also fluctuations and peaks of activity, and Fig 5 shows the global distribution.

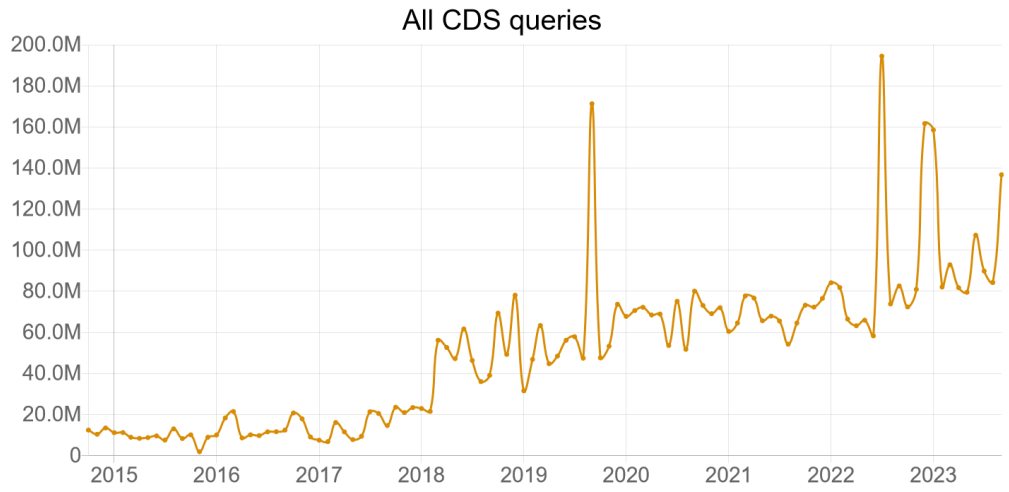

Figure 4. Queries per month on all CDS services 2015-23

The reliability remained high in the past year at >97.94%. The architecture of the information system is designed to support this high level of performance, with the four main elements:

1. CDS installations in 2 geographically distinct local server rooms.
2. Support of external mirror sites (in France and in other countries).
3. Server virtualisation, via a cluster of VMWare hypervisors, and Docker encapsulations, to host the virtual machines providing the services.
4. Data storage on disks, in the form of 2 RAID6 bays synchronised on the two local sites ("CDS All-Sky-Data system").

Main Services →	**SIMBAD**	**VizieR**	**Aladin**	**Total** (main services)
users / month	**190k** (+23%)	**61k** (+35%)	**390k** (+8%)	**> 390k***
queries / day	**449k** (+30%)	**727k** (+100%)	**1.8M** (+5%)	**3.0M**
load / day	**6.1 GB** (-13%)		**263 GB** (+62%)	**>269 GB**
data volume	**44 GB** (+28%)	**77 TB** (-3%**)	**713 TB** (+39%)	**790 TB**
data content	**16.9 M obj.** (+17%)	**24.3k cats** (+6%)	**1203 HiPS** (+16%)	
reliability	**99.82 %** (+0.08%)	**97.94 %** (-1.2%)	**99.84 %** (-0.09%)	**> 97.94%**

Table 1. CDS statistics Oct 2022 - Sept 2023. (Note that there is overlap of the users/month between the CDS services, ** Some redundant files cleaned resulting in reduced overall volume.)*

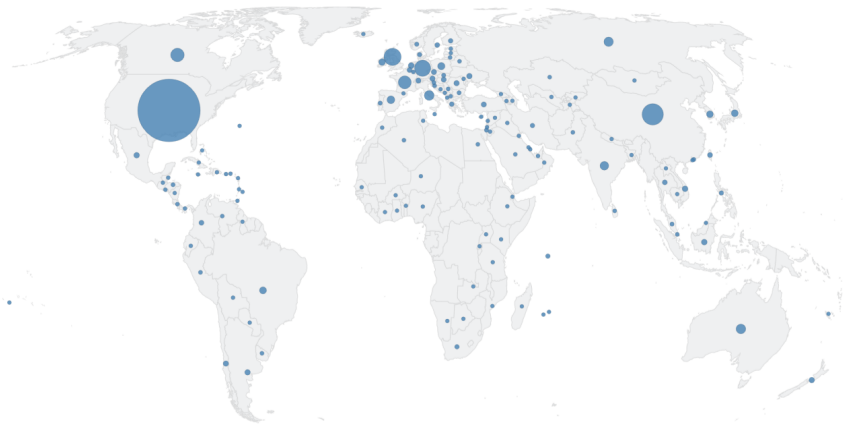

Figure 5. Global distribution of queries on SIMBAD 2022-23

Physical hosting of CDS servers

Local sites: A major change was implemented in 2023 so that the CDS servers are now hosted in 2 server rooms in Strasbourg, one at the Strasbourg University Data Centre (since 2020) and the other at the IPHC (Institut Pluridisciplinaire Hubert Curien), a mixed research unit of CNRS and Université de Strasbourg UMR7178) on the CNRS Cronenbourg campus (since April 2023). The final move took place in June 2023, and now the server room at the Strasbourg Observatory has been definitively closed. Usage of the 2 server rooms is now governed by contractual agreements.

Partner mirror sites: In addition to the local sites in Strasbourg, the CDS has for many years several mirror sites hosted by partner institutes. SIMBAD is replicated in the US (CfA/Harvard), and VizieR is partially replicated at 7 sites (US, Canada, India, China, Russia, South Africa, Japan). These servers are acquired and maintained by the partner institutes; only the software management and data availability are the responsibility of the CDS staff. The CDS also relies on some twenty partner sites (HiPS nodes) for the data used by Aladin. They replicate the most requested HiPS image records, although currently these represent only a small part of the total volume (<10%). The management of these HiPS mirror sites is not the responsibility of the CDS, but it has been CDS strategy to promote the creation of these sites as it improves the Aladin users' experience.

Operational storage space

In terms of data storage space, the CDS services are not homogeneous. The major part (>89%) of the volume is for the Aladin service, and then VizieR (9%) and SIMBAD (<1%). The evolution of storage needs is directly linked to the volume of astronomical data published by the discipline (bibliographic data, tables, catalogues and image surveys), and the capacity for CDS to process and ingest the data. Currently, the total data distributed by the CDS represents 790 TB. The evolution over time of the CDS volume (Fig 6) shows rapid growth up to 2020 (doubling every 18 months) then a slowdown in 2021-22 two years, followed by a ramping up in 2022 and onwards. The current storage system, set up in 2018 (CNRS-INSU *'mi-lourds'* funding) offers 1.6 TB of storage space, duplicated on the 2 local sites.

Provisional plan: The "CDS All-Sky-Data" system storage equipment will come to the end of its nominal service life in 2025 although this may be prolonged until 2027 (via a warranty extension purchase). A new CDS All-Sky-Data 2 system is currently being specified to replace/complement the existing one from 2024. The total target is ~5 Petabytes (replicated). We have had detailed interactions with HP and DELL for specifications and costs of a new system. This will be a major renewal of the CDS storage and discussions about the funding are on-going with CNRS and CNES.

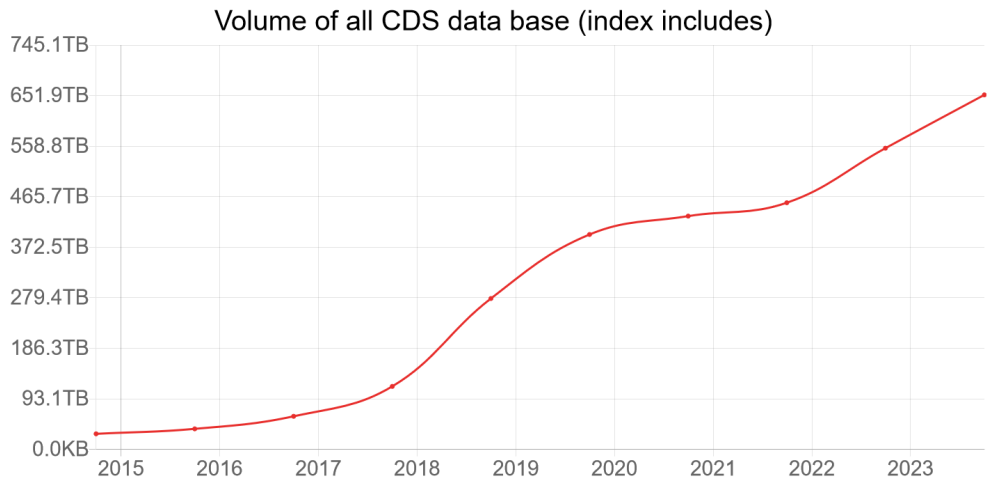

Figure 6. Total volume of data made available via all CDS services 2015-2023

The high cost of disk storage is also motivating us to find new ways of working within more limited disk space. The current operations make full use of the All-Sky-Data storage system, making the most of all the volume available to make the operations as efficient as possible (and not leaving large amounts of disk space un-used). The CDS operations require significant redundancy which we currently measure to be around 5 times the volume of the publicly available data. This includes the local mirror copy, RAID6 redundancy space, snap-shots, file system organisation, temporary computing space, and backup of original data. We are studying several alternatives to try to reduce this redundancy factor without compromising security or performance, in particular by the use of appropriate compression algorithms for some HiPS surveys.

Operational servers

The operational services (web servers) of the CDS require relatively modest computing power, and the CPUs are "on average" used well within their capacity. The X-Match service is the most demanding of fast computing power and disk access. However, this cannot be undersized because the availability constraint implies the capacity to absorb "peak" requests, i.e. a large number of requests simultaneously, up to several hundred per second. Thus, to meet this usage pattern, and to facilitate distribution over distinct geographical sites in the long term, the technical solution adopted in recent years has been to virtualise the servers as far as possible on VMWARE hypervisors. In 2023, two virtual machine servers have been added to the 3 existing ones, 3 of them located at the University Data Centre, the other 2 in the IPHC server room.

Evolution of mirror servers: The mirror sites are strategically important and we maintain the partnerships necessary for the smooth running of these mirror sites. It would beneficial if these mirror sites are upgraded to be HiPS nodes.

Software and development

Development principle: CDS services (SIMBAD, VizieR, Aladin, etc.) rely on software, databases and tools for the management and distribution of data, as well as for the data ingestion workflow. The strategy of the CDS, forged over several years of experience, is to rely, as much as possible, on open source software supported by a dynamic community (e.g. Postgres, astropy, etc.), and to create "in-house" developments as soon as libraries and tools particularly specific to our activity are concerned (e.g. management of spatial indexes, astronomical libraries, bibliography processing, etc.). This approach makes it possible, in the long term, to guarantee knowledge and mastery of the codes, to encourage innovation, while avoiding spreading efforts over related areas.

The use of paid software within the CDS is done only in exceptional circumstances, justified by the absence of an equivalent free solution (e.g. no comparable alternative to VMWare). These paying solutions are considered as transitory, while ensuring that alternative solutions can be implemented without difficulty (e.g. Sybase to Postgres). On the other hand, in order to guarantee knowledge and continuity in the long term, the CDS strategy is to rely on the skills of at least two developers per service.

Information system: CDS services are currently based on 20 components that constitute the CDS information system (database, server codes, client back-end, software required for ingestion workflows, etc.). The constant evolution of these components is essential to ensure that they are in line with current needs (evolution of the characteristics and volume of data), the technical resources available (network, machines, storage) and the human resources in charge of development and operational monitoring (permanent staff, fixed-term contract, retirement, etc.). For each component, we evaluate its sustainability and, depending on this, we plan its renewal, evolution or withdrawal. Recent developments in the information system are shown in the Fig 7 below and include:

- The deployment of the new journal article management system. The "BCS" is replacing the old system (parfile), the transition is ongoing.
- The total re-coding of the large catalogue access tool (QATSS tools), now based on a more robust and faster technology (CGI/RUST).
- The redesign of the CDS website (currently in the deployment phase).

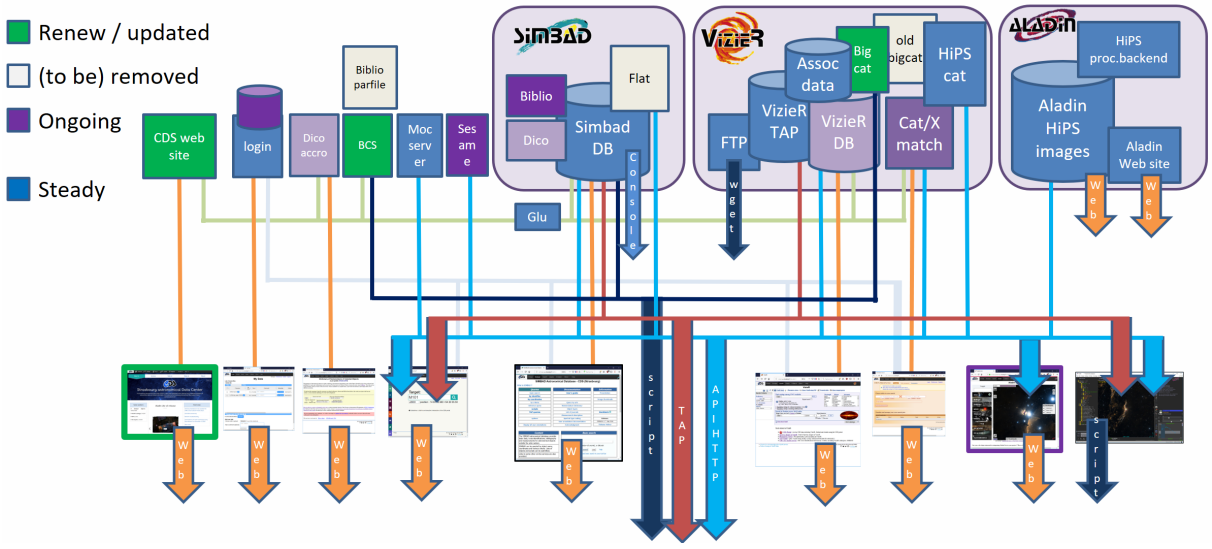

Figure 7. A representation of the CDS architecture showing the status of the 20 main components

4.2 SIMBAD

Overview of the Workflow and SIMBAD Content

SIMBAD is a bibliographic database of astronomical objects of interest. It is the oldest service of the CDS. In 2022-23 it got 449k requests per day on average. The team comprises 3 software engineers (plus two participating in special operations), 9 documentalists divided in 3 specialized teams (plus 1 for 10 % of her time), 6 staff astronomers (5 to 20 % of their time) and 2 staff astronomers lead the nomenclature and scientific content. SIMBAD is a meta-compilation built from the published literature and very large surveys. It includes all object types : stars, exoplanets, sets of stars, galaxies, sets of galaxies, interstellar medium, gravitation.

In 2022 we have added ~15000 new references, mostly from the main journals - A&A, ApJ, ApJS, AJ, MNRAS – as shown in the Figure 8. The main journals involved 10921 references in 2022 (10485 in 2021), and 3487 references from other journals were treated in 2022 (4043 in 2021). Since 2007 we decided to decrease the number of other journals treated (74 journals in 2007 → 23 in 2022) . There are several reasons for this: some journals no longer exist (IBVS, GCNR, etc.) and some other are no longer analysed due to lack of time (CBET, SerAJ, RMxAC, …).

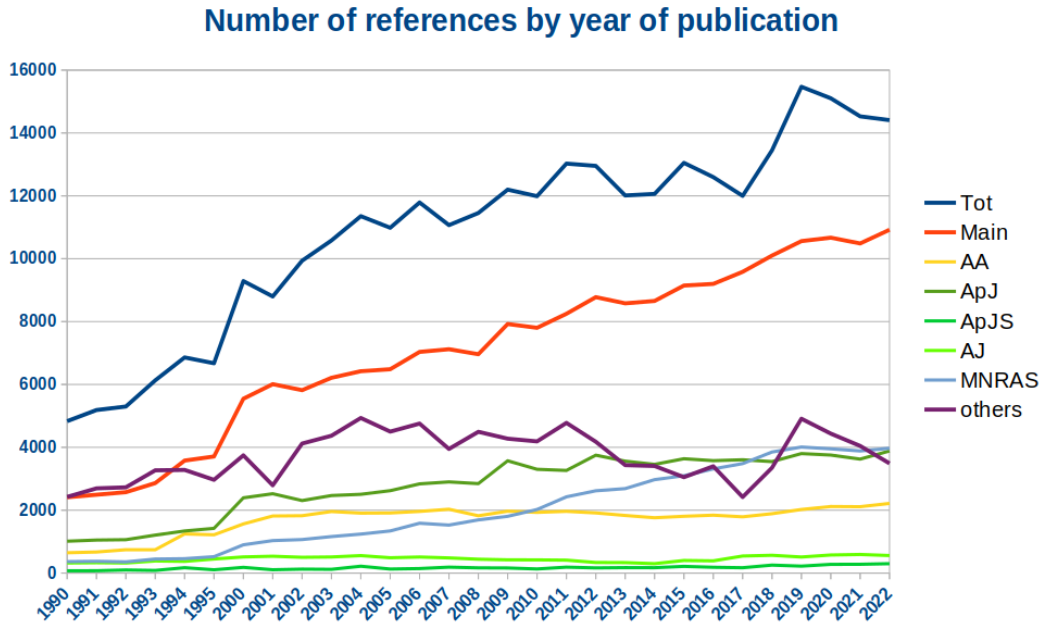

Figure 8. Number of references (journal articles) treated for SIMBAD, 1990 - 2022

Journal references are processed by a specialised team of documentalists with the help of the software DJIN (identifiers recognition). References with tables of objects are flagged to be possibly ingested into the VizieR database (relatively large tables) or to be possibly ingested in SIMBAD without being recorded in VizieR (relatively small tables). Once a large table has been ingested in VizieR it gets a new flag for possible ingestion in SIMBAD. In bi-weekly appraisal meetings (scientists and documentalists) the large and small tables of objects are prioritised for ingestion in SIMBAD, mostly on the basis of scientific and sometimes technical criteria. The workflow is currently very efficient and is catching up the back-log: from January to October 2023, we received 1142 references to be appraised, but more (1354) were done; of these, 66 % were assigned a priority 1 for ingestion, and 9 % a priority 2; the rest either was already done by the DJIN team (no additional data), or are not relevant for SIMBAD. Tables of objects are then ingested in SIMBAD by another team of specialised documentalists with the help of the cross-identification software COSIM, their own expertise, and sometimes the expertise of the astronomers for difficult cases. For the same period of time 901 tables of all sizes (from ~10 to >1 million objects) have been processed (compared to 766 for the whole year 2022).

For each astronomical object, SIMBAD provides the list of references where it has been cited, the main object type as well as a list of secondary object types, and cross-identifications with the corresponding list of identifiers. It also provides data collected in papers and surveys: basic data that appear at the top of the HTML page (coordinates, proper motions, parallax, HRV/redshift, spectral type, morphological type, size for extended objects, magnitudes), and collections of measurements that are at the bottom of the HTML page (HRV/redshift, proper motions, parallax, distance, spectral type, stellar fundamental parameters, variability). SIMBAD now contains 16.9 million objects, a **spectacular increase of nearly 3 million in one year.** It mostly comes from the ingestion of the SDSS DR16 QSO catalogue, APOGEE DR16, GALAH DR3, as well from several articles listing the members of new galactic open clusters (based on Gaia). Table 2 below gives the statistics of the SIMBAD content from 2022-23, as well as the increase in one year.

Overview	N (million)	+ in 1 yr
Objects	16.946	2.977
Identifiers	61.474	6.645
References	0.427	0.015
Citations of objects in articles	38.134	7.104
Basic data		
HRV/redshift	7.636	1.136
Proper motions	9.065	1.788
Parallax	7.713	1.141
Spectral type	0.946	0.088
Morphological type	0.147	0.003
Collection of measurements		
HRV/redshift	12.290	2.187
Proper motions	15.819	1.865
Parallax	15.884	1.249
Distance	9.543	0.281
Spectral type (spectroscopic only)	2.019	0.797
Teff logg Fe/H (spectroscopic only)	4.120	1.293
Variability	3.283	0.153

Table 2. SIMBAD content statistics 2022 and 2023.

Bibliographical Center Supervisor

The new process (developed last years) for ingesting references directly from publishers is now in full operation. It has been built to manage the references from the major publishers (EDP, IOP, OUP) starting from the download of full articles and attached data, conversion into an internal and unique format easily usable by our software. It enables the data to be internally accessible for the needs of the SIMBAD and VizieR teams. Developments have continued, focused on editors without a full access to the XML version of articles (Science, Nature, GCN, ...) and some improvements have been made for a more friendly user interface. The procedures have been adapted, and now take into account many special cases that exist in the different publisher's systems.

Dictionary of Nomenclature

The object identification system (Dictionary of Nomenclature) re-construction will be the major software project in the next years. A number of verifications, and modifications have been processed to prepare the evolution. Some developments have started for a new dedicated database and a new web site.

Maintenance and evolution of SIMBAD API

A new VO-compliant search by cone around a position has been released and developed with recent framework as a micro-service. We are starting new evolution of python libraries, rebuilding fully the code to offer more functionalities and evolution using more stable Simbad TAP interface.

4.3 VizieR

Human resources

There will be a strong reduction in the human resources available for the VizieR team in the coming year 2024:

- Alicia Vanhulle, was recruited as a contract VizieR research engineer, in May 2022, as an answer to an important and identified need (cf “A strategic emergency: the need for an additional VizieR engineer” in 2021 CS report). She will leave CDS in December 2023 for a new position at IPHC.
- Ana Fiallos, recruited as a contract VizieR documentalist, in May 2022, departs in Nov 2023.
- Coralie Fix, recruited as a contract VizieR documentalist, in Sept 2019, departs in Aug 2024.

VizieR content

As a result of being supported by 4 documentalists (though not perennially, as explained above), the number of references processed in the past is higher than in previous years (+1419), which brings the total number of catalogs in VizieR on 1 October 2023 to 24271 (Table 3).

VizieR Content (2018-2023)

	2018	2019	2020	2021	2022	**2023**
Number of Catalogues	17673	19189	20289	21412	22852	**24271 (+1419)**

Table 3.

Gaia Focused Products (I/361): continuing the Gaia legacy

The Gaia Focused Products Release (FPR) dataset was released on 10 October 2023. It improves on a number of aspects upon Gaia DR3 by taking into account a 66 months timeline instead of 34 for Gaia DR3. The release contains mainly:

1. Astrometry and photometry from engineering images taken in the omega Centauri region.
2. The first results of quasars' environment analysis for gravitational lenses search.
3. Extended radial velocity epoch data for Long Period Variables.
4. Diffuse Interstellar Bands from aggregated RVS spectra.
5. Updated astrometry for Solar System objects.

Gaia FPR appears in VizieR as a catalogue of 11 individual tables, 3 of which are large catalogues with up to 171 million sources.

Very large *versus* long *versus* thick catalogs

Besides Gaia FPR, 6 other very large catalogs were ingested. 3 are Gaia-related as well: Carrasco+2023 (II/374), providing a RGB photometric calibration of 213 million Gaia stars, and a stellar variability catalog of 145 million Gaia stars (Maiz Appellaniz+ 2023, J/A+A/677/A137). The third one, Holl+2023 (J/A+A/674/A25), provides information about systematic errors in Gaia DR3, in particular spurious periods and scan-angle-dependent signals.

Two ESO Phase 3 large catalogs were processed: VMC DR6 (II/375) and VVV DR4.2 (II/376). The redshift determinations of DESI DR8 was ingested as well (VII/292). Finally, the simulated

eROSITA catalog was ingested (J/A+A/665/A78). Although we do not prioritise simulated data, our relationship with A&A mandates that we make this catalogue available through VizieR.

Several more large catalogues are expected to be ingested in the coming year: SDSS DR17, DECALS DR9/10, KIDS DR4, and ESO PHASE III ATLAS DR4, VPHAS+ DR3.2, and VIKING R4 are being considered. Additional large catalogs are in the works, such as EROS: we are currently transferring/ingesting the data..

- We are in contact with the ZTF collaboration for ZTF DR19, though further work is currently delayed until the collaboration can give us a recipe for building the urls for the light curves, which will not be hosted at CDS but remain on ZTF servers. Since this aspect is still evolving at ZTF, this is on standby.
- Subaru Hyper Suprime Cam PDR3: the Subaru directorate has been contacted to approve (or refuse) the access and distribution by and through CDS. Also on standby.
- Pan-STARRS DR2: we have just received a link from Gene Magnier to download the internal processing database. We are pursuing this at the time of writing.

The trend of “long” catalogs continues: small collaborations, even a single author, working with a large input dataset such as Gaia, are able to produce large catalogs of up to a billion sources and more, which we must process through the large catalogs pipeline, even though they may have just a few added quantities with respect to the reference catalogue. These are “long” catalogues, i.e. they can have a billion lines but only a handful of columns. An example this year is Maiz Appellaniz+ 2023, J/A+A/677/A137, providing stellar variability statistics for Gaia DR3.

Another continuing trend is that of “thick catalogs”, requiring often >5hours of work from the documentalist in charge for many reasons (selection, recovery, formatting, correction of tables with a large number of columns). An example of that is the catalogue of planetary nebulae detected by GALEX and corollary optical surveys from Gomez-Munoz+2023, with a total of 432 columns. Such complexity peaks naturally require documentalists to spend much more time than average on these catalogs.

VizieR usage

In terms of use of the service there was an average of 750k queries/day in the 2022-23 period (Table 4), in strong increase (doubling) with respect to previous year, and with a huge peak in usage around December 2022. The huge majority of queries are performed mostly via python, then other APIs, and the web forms. Queries of TAPVizier also increased by more than 50%. The use of VizieR for publishing data was also highlighted in a tutorial prepared in the EOSC Future project called “The journey of your data through the Virtual Observatory and the European Open Science Cloud” (see section 5.1).

VizieR Usage

	2018	2019	2020	2021	2022	2023
All queries, /day	368 000	696 000	520 000	514 000	422 000	**750 900**
Associated Data Service Queries/day	80	543	845	1212	811	**345**
VizieR TAP service Queries/day	3 700	2094	14 000	23 130	43 000	**69 400**

Table 4.

Workflow and technical evolutions

- Alicia Flint has worked with Gilles Landais to update the UCD tools of the VizieR workflow. The new software proposes a synthetic, in-editor view of possible UCD1+ matching the columns explanations, with scores, allowing to choose one from a list or simply replace the propositions with the documentalist's choice. We expect this new tool to improve the ergonomy of that aspect of the workflow, which should speed up UCD attributions.

- Global indexation: in 2022, we started to describe all catalogue footprints using MOCs (Multi-Order Coverage), and the computation of these MOCs became part of the standard workflow for all catalogues. This is now baseline workflow and it continues to function nominally. The result is an improved homogeneity of footprint descriptions across services, in particular VizieR and Aladin.

- QAT2S is the new architecture for large catalogue queries, developed by F-X. Pineau. QAT2S is written in Rust, and aims at improving service availability and stability. The previous architecture used 1 Java daemon receiving and processing all queries. If one query was problematic (too big), it could block the whole service. The new architecture circumvents this problem by instanciating 1 QAT2S process per query. If a query is problematic, only that process is affected, while the service remains operational. A corresponding improvement in stability has been observed since the deployment of this new framework.

4.4 CDS X-Match

The X-Match service continues to operate smoothly with the current established code. The underlying IRODS data storage for the service has been migrated from a physical machine into a virtual server (consistent with all other CDS services).

This year, most efforts have again been put into the new VizieR large catalogues query tool (QAT2S). Those developments will benefit the development of the new X-Match service code that will also switch from Java to Rust.

Usage Statistics

The service is mostly used via its API (~11k jobs/day - see Table 5) via python or tools such as TOPCAT. Only a small number of queries made through the web interface (45 jobs/day). The total number of associations/day is around 100M in both the web interface and the API, reflecting the fact that the web interface is used for a smaller number of large jobs whereas the API is dominated by a large number of small cross-match jobs.

CDS X-Match Service Usage

	2014	2015	2016	2017	2018	2019	2020	2021	2022	**2023**
Web interface (jobs/day)	16	20	30	33	40	43	49	60	41	**45**
HTTP API (jobs/day)	50	580	889	1256	2687	2556	7244	21986	14432	**11256**
Associations/day (Web Interface)	~70M	~55M	~104M	~164M	~179M	~44M	~72M	~85M	~66M	**111M**
Associations/day (HTTP API)	~1.6M	~6.6M	~6.7M	17.8M	50.2M	43M	90M	78M	78M	**105M**

Table 5. CDS X-Match service usage statistics 2014-23

4.5 Aladin

The Aladin project is centered around visualisation and browsing of the sky thanks to image surveys and the possibility to overlay various kinds of information. The project encompasses two visualisation clients (Aladin Desktop and Aladin Lite) as well as an ensemble of Hierarchical Progressive Surveys (HiPS). Aladin keeps on growing and to be heavily used as can be witnessed by the continuous usage: we observe an average of 1.8 million queries per day on the Aladin services (+5% since last year).

Aladin Lite is a lightweight version of the Aladin tool, running in the browser and geared towards simple visualisation of the sky. It is powered by WebGL technology and currently works in any modern browser. Recent developments aim at introducing more functionalities to this tool while keeping lightweight and embeddable in any browser.

Aladin Desktop is a java tool enabling a more in depth exploration of image surveys on the sky but requires a classical installation on the user machine. Aladin Desktop enables the location of data of interest on the sky, access and exploration of distributed data sets and visualisation of multi-wavelength data and overlay of additional information from various databases. The goal is to be able to provide visualisation and simple analysis tools for data exploration in Aladin Desktop itself or in coordination with other VO-compatible tools that can be interconnected.

Both Aladin Desktop and Aladin Lite are designed to visualise Hierarchical Progressive Surveys. HiPS are either image or data cubes that are stored and accessed in a hierarchical way and follow the HEALPix tessellation of the sky. The HiPS standard was developed at CDS but has become an international standard and is now widely used. HiPS can be generated thanks to the CDS HiPSgen tool. Image surveys accessible through Aladin have been generated in house or by other entities, and are registered in the HiPS network.

Different additional software related to the ingestion, creation, access of HiPS are also developed and shared with the community (e.g. HiPS2FITS to generate FITS cutouts out of a HiPS, ipyaladin to include Aladin in a python notebook, etc.).

Software updates

Aladin Lite: A new version of Aladin Lite, version 3, was released in January 2023. It adds new projections, support FITS tiles for HiPS and FITS image support. Aladin Lite v3 also benefits to the ipyaladin tool and has helped to fix bugs.

HiPSgen: A tool that can create a HiPS out of a collection of FITS files. This year, a new data release of HiPSgen was done, together with a HiPSgen manual. A new method to adjust photometry cuts while generating a HiPS was developed. This new method computes brightness cuts by regions and is especially efficient for image surveys that are not continuous and for which the background is not homogeneous (e.g. Hubble Space Telescope images).

Aladin Desktop: Following needed improvements in the SKA context, a support for multiple outputs of DataLink queries is now possible. With the evolution of Apple security measures on external software, we are encountering difficulties with the packaging of Aladin Desktop for Mac users. This issue is linked to the registration of trusted software to the apple store and while we had support from the university, it has been difficult to find information for java applications.

HiPS

The number of HiPS published by CDS has increased by 15%. 43 new HiPS datasets have been generated at CDS, which represent a volume increase of 38%, for 200 additional TB. The increase

in volume is mainly driven by the DESI Imaging Legacy Survey, a new optical reference survey, covering half of the sky at a resolution of 250 mas. Four individual HiPS have been generated (filters g, r, i, z), along with a colour version built from the 4 filters. Other significant HiPS produced at CDS in the last year include: JWST, DES DR2, ESO outreach, DECaPS DR2, Fornax Deep Survey. We also host 4 KiDS HiPS datasets that have been initially generated by the PI of the survey. The Hubble Space Telescope HiPS have been updated in May 2023, with the help of D. Durand (CADC affiliated), to take into account the new observations to update/replace the previous HiPS version.

We have carried out a series of actions to make HiPS more visible:
- Individual HiPS are now published in the IVOA registry, making them findable by VO clients querying the VO registry.
- On the curation side, we have started to mark some HiPS as 'deprecated' (for instance older versions of surveys), in order to provide users with a clear up-to-date view of available data.
- A new landing page, embedding Aladin Lite v3, has been deployed for CDS HiPS.

Following the council's recommendations, we have written a « HiPS ingestion strategy » document summarizing our criteria to select and prioritize image datasets to be ingested as HiPS.
We have also started a study to mint DOIs for our published HiPS. This action will allow for easy long-term citation of HiPS datasets, and will improve their FAIRness.

Projects/collaborations around Aladin

SKA Science Regional Centre (SRC) Prototyping, Orange team:
The Aladin team has started to be involved in the development of the SKA SRC. In particular, François Bonnarel and Matthieu Baumann are part of the Orange team which deals with the visualization. Developments related to DataLink, ObsCore and SODA have been made in line with the needs identified by the SKA Orange team leader. These developments also align with needs of the astronomical community at large. This context also led us to think and test our limits on how to deal with very large datacubes. Tests of on-the-fly services to produce moment maps are a promising avenue. This also led to exploratory discussions on how to produce a hierarchical 3rd dimension for the HiPS cubes.

ESA-Sky collaboration:
We have renewed our interactions with the ESA-Sky team in the context of the new Aladin Lite v3 release. We are helping the ESA-Sky team on compatibility issues with the new version of Aladin Lite. On the other hand, the ESA-Sky team is contributing to the Aladin Lite code by providing customized features (e.g. polygon selections).

Europlanet:
In the context of the Europlanet project, new functionalities can be mentioned: a feature name resolver and a ***"what is this ?"*** feature, similar to the SIMBAD pointer but for planetary surface features.

Communication

Alongside the evolution of the services and data, we have continued to promote our services to the widest community and communicate on the new developments. Aladin Lite has been on-boarded on the ESCAPE 'Open Source Software Repository' (OSSR) as part of the ESCAPE project. Aladin Lite and Aladin Desktop were registered to a campaign called *"connaître et rendre visible les logiciels de la recherche" (to know and to make visible research softwares)*. Talks and posters were made at ADASS conference in 2022 (2 talks), at the IVOA Interop, at the ASOV days, at the EAS conference and JupyterCon. Aladin was also presented at the AAS and SF2A CDS booths.

Two different internships were directly related to test developments for Aladin: one for field maps with Aladin Lite and one for displaying HiPS in a virtual reality helmet with Aladin Lite as well.

4.6 R&D

Since the last Scientific Council we have pursued a varied R&D program with both operational actions and exploratory work to prepare the future. This work was carried out by CDS software engineers with the help of **10** interns/apprentices. A new apprentice was hired in September 23 for one year after his internship. Engineers from the GALHECOS team and from the Observatory IT Support were also involved in some actions. It is also important to mention that in several cases, the expertise of astronomers and documentalists is also very helpful, demonstrating R&D across the whole CDS team.

The topics of the R&D program that were supported with interns includes:

- Work around the query of the CDS services in Natural Language with a different approach taking into account the advent of OpenAI ChatGPT at the end of the previous internship.
- We restarted also the experiments around Virtual Reality.
- Work around Data Model annotations in the VOTables.
- R&D on including frequency as a dimension (electromagnetic axis) in MOC coverage maps.
- Development of browser plug-ins to convert HTML webpages in CDS/XML to help with the processing of papers in the CDS DJIN tool.
- The short '*discovery'* interships were more dedicated to experiment quickly with things like the opportunity to work on Alexa skills dedicated to astronomical data querying.

AI should be an important field of investigation in the next internships.

5. Projects

5.1 Virtual Observatory and Open Science projects

IVOA

CDS continues to play a leading role in the development of the Virtual Observatory, in particular via various working group chair positions. In 2023 Pierre Fernique became the Vice Chair of the Time Domain Interest Group. The full list of responsibilities are listed below:

- **Executive Board member for EuroVO** - M. Allen
- **Chair of the Committee for Science Priorities** (since May 2021) - A. Nebot
- **Vice Chair of the Data Access Layer Working Group** (since May 2021) - G. Mantelet
- **Chair of the Data Curation and Preservation Interest Group** - G. Landais
- **Vice Chair of the Radio Astronomy Interest Group** - F. Bonnarel
- **Vice Chair of the Time Domain Interest Group** - P. Fernique
- **Chair of the Education Interest Group** - H. Heinl *(until May 2023)*
- **Editorial team for the IVOA Newsletter** - S. Amodeo *(until May 2023)*

ESCAPE

The **ESCAPE** (**E**uropean **S**cience **C**luster of **A**stronomy & **P**article physics **E**SFRI research infrastructures) project which started in February 2019 concluded in January 2023. CDS has lead a one of the major work packages (WP4, 'CEVO') on *Connecting ESFRI projects to EOSC through VO framework.* There were many ESCAPE activities in this final part of the project. All the deliverables and events are listed on the WP4 wiki pages[5]. Aladin Lite v3, enhanced tutorials, MOC 2.0 standard are among the results of the project. A demonstration of notebook tutorials will be made during the council meeting.

European Project - EOSC-Future

This project involves 40 M€ of funding, with CDS being a small partner (20 PM) for involvement in test science cases and training activities to enable community use of open science resources. The activities are being done in coordination with the ESCAPE project. A contract engineer (M. Marchand) worked on the project September 2022 - August, and a short term postdoc contract (A. Gonneau) was hired March-July 2023 to work on this project (and continued on to other CDS projects afterwards).

One of the results of this project is a tutorial about publishing data in astronomy in the context of the EOSC. This tutorial highlights the role of the CDS VizieR service for publishing data according to the FAIR principles, and then having the records of those data publications propagated through the IVOA registry, the EUDAT B2FIND aggregator, and being made available in the EOSC Portal (Figure 9). The tutorial is hosted on the CDS GitHub (link here[6]) and is itself 'on-boarded' to the EOSC so that it is now part of the EOSC training materials, and is findable on the EOSC Portal.

[5] https://wiki.escape2020.de/index.php/WP4_-_CEVO

[6] https://cds-astro.github.io/a-FAIR-journey-for-astronomical-data/

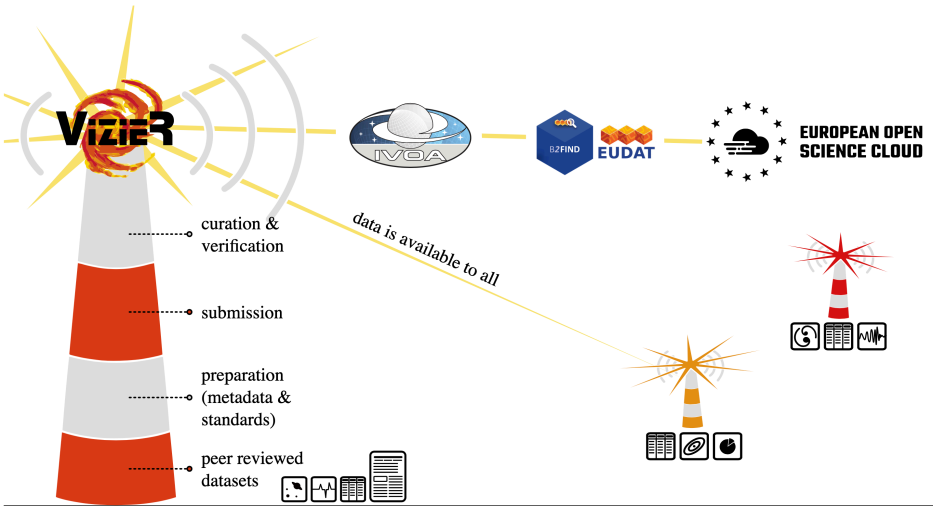

Figure 9. A diagram from the tutorial on "The journey of your data through the Virtual Observatory and the European Open Science Cloud".

EuroPlanet Project / EuroPlanet Research Infrastructure 2020-2024

CDS is participating in the **Europlanet 2024 Research Infrastructure (EPN-2024-RI)** which started in February 2020. The CDS effort in this project has been applied to improving the CDS tools for applications toward planetary data (e.g. catalogues in VizieR and visualisation with Aladin and Aladin Lite).

5.2 The Research Data Alliance

The CDS continued to be active in the Research Data Alliance (RDA). F. Genova continues to co-lead RDA France with L. El Khouri (Direction des Donnéees Ouvertes de la Recherche, CNRS), and to co-chair RDA Regional Advisory Board and Regional Board. RDA France continue to be hosted by CNRS and RDA and RDA France to get support from the National Fund for Open Science. The Regional Advisory Board gathers all the "regions" and countries which provide financial or in kind support to RDA, and the Regional Board all the "regions" interested in the RDA with different levels of interest. The co-chairs of the Regional Advisory Board are invited members of the RDA Council .

The RDA is supported by the French National Plan for Open Science because it is the international forum for discussion and elaboration of good practices for the data aspects of Open Science, and also because it is a powerful tool for the acculturation of the national community to these aspects of Open Science. More than 1000 people with very different profiles are in touch with RDA France. It is also important that the astronomical community stays informed of the RDA work and participates when relevant, to share our knowledge and to make sure that the RDA recommendations are acceptable for us.

The RDA "Global Open Research Commons" (GORC) International Model Working Group recently produced its output, The Global Open Research Commons International Model, Version 1, of which F. Genova is a co-author. The work started with generic frameworks in mind such as the EOSC but the IVOA was fully considered as a GORC example and its characteristics taken into account during the work on the model. F. Genova is also involved in the RDA Data Repository Attributes Working Group, with in mind CDS attributes.

F. Genova with colleagues from different regions proposed the Evaluation of Research Interest Group, to identify the RDA role in the crowded international landscape of initiatives dealing with the necessary evolution of research assessment. The theme is a priority of the national and European Open Science strategies - it is well recognized that a key obstacle to the development of Open

Science is the current bibliometric-driven evaluation methodology. The Interest Group was endorsed by the RDA Council in June 2023 and is starting its activities.

The local group "RDA UNISTRA" set up in 2022 in the framework of Strasbourg University (Unistra) Open Science Strategy to disseminate information about the RDA in the University community and to enable Unistra participants in the RDA to share their knowledge about RDA activities has been regularly meeting since then. In addition, F. Genova was recently appointed *Chargée de mission Ouverture des données et évolution de l'évaluation de la recherche* (advisor on data opening and research assessment evaluation) at Unistra, which reinforces the links between the CDS and the University activities on Open Science.

5.3 XMM2ATHENA

As part of her independent science Ada Nebot is part of a successful proposal in the Horizon 2020 call on SPACE-30-SCI-2020: Scientific data exploitation, XMM2ATHENA[7]. This project brings together members of the XMM-Newton Science ground segment, key members of the Athena Science ground segment, and other members of the X-ray community with complimentary skills to develop and test new methods and software to allow the community to follow the X-ray transient sky in quasi-real time, identify multi-wavelength/messenger counterparts of the sources detected with XMM-Newton and determine their nature using advanced machine learning methods and probe the faintest sources, hitherto undetected, using innovative stacking and detection algorithms. These methods will then be integrated into the Athena software, currently at the beginning of the developmental phase and the newly detected/identified sources will enhance our preparation of the X-ray sky that will be observed with Athena. Ada Nebot leads WP2 (Multi-wavelength/messenger counterparts) bringing together her experience with the CDS services and tools and her expertise in the multi-wavelength characterisation of X-ray sources, experience gained as former member of the XMM-Survey Science Center. This WP will advance algorithms designed in the framework of the former ARCHES project that ObAS led, in particular the ARCHES cross-matching tool developed and maintained by Francois-Xavier Pineau[8], which provides a probabilistic multi-catalogue positional matching.

The project started on 01 April 2021, and Pooja Sharma joined the team as a postdoc in May 2023 (replacing Jere Kuuttila who left end of July 2022 after a 12 month contract) to study the young stellar and binary contributions to the X-ray emission of our Galaxy. Since the beginning of the project we have contributed to the creation of the DR13 XMM-Newton catalogue, which contains more than 900 000 X-ray detections. In particular, within the activities of the WP2 we have produced statistical multi-wavelength associations of more that 200 000 X-ray sources and produced spectral energy distributions from X-ray to radio. Simplified catalogues can be downloaded as FITS files, and full data products can be accessed via TAP service using the XCatDB[9] and through VizieR following FAIR principles. Individual SEDs can be searched for, accessed and downloaded as fits and png files through different web services, e.g. XCatDB and the new XMM-Newton SED finder service[10], that L. Michel from the GALHECOS team created, updates and maintains. Results will be presented at a conference we are organising within the project that will take place in Toulouse end of February 2024[11]. The project was initially funded for a total of 36 months and we have recently been granted with a 6 months no cost extension, so the project will continue until September 2024.

---

[7] XMM2ATHENA http://xmm-ssc.irap.omp.eu/xmm2athena/, 2023AN….34420102W

[8] ARCHES xmatch tool http://serendib.unistra.fr/ARCHESWebService/index.html, 2017A&A...597A..89P

[9] XCatDB https://xcatdb.unistra.fr/4xmmdr13/index.html

[10] XMM SED finder: https://xcatdb.unistra.fr/sedfinder/

[11] https://xmm2athena.sciencesconf.org/

6. Responses to 2022 Recommendations of the CDS Science Council

Recommendation:

"We recommend CDS develops plans for collaborative ventures for deploying machine learning technologies."

Response:

We thank for the council for this recommendation, there are very big developments happening for AI, in particular the large language models (LLM) and the emergence of ChatGPT for example. We expect these new kinds of technologies to have an impact on different aspects of the CDS work, and we are seeking ways to build our knowledge and understand how to benefit. The CDS R&D program has touched on these topics, and we recall that the CDS ChatBot prototype is an ongoing project. Pierre Ocvirk has tested various uses in the context of Vizier treatment of catalogues, and CDS has also been involved in the Deep Learning project within the ESCAPE project (with ESO and HiTS as partners). We are still however at the very early stages and we have not yet had the capacity to strongly engage with the potential partners within CNRS or UNISTRA. Some of the candidates for scientific positions would have brought in expertise, but these potential recruitments have not eventuated. We have however engaged with our close partner SAO/ADS on these topics and discussions at the AAS meeting (Jan 2023), ADASS (2023) and during a recent visit to ADS (November 2023) are leading to CDS participation in AI focused activities that have been advanced by the ADS team. We expect this to develop further in 2024 with the idea of joining ADS initiated workshops on these topics in mid-2024. We note also that CDS is participating in a new initiative at ObAS of an AI working group. This recommendation will be closely followed into the future as AI has the potential to change many aspects of information processing.

Recommendation:

We recommend CDS continues to follow developments in the French and European research landscapes in terms of digital infrastructures (e.g. CNRS plans to develop a combined HPC / HPDA offer for national research infrastructures, EOSC, EuroHPC) and to take advantage, where appropriate, of the opportunities offered."

Response:

CDS is in contact with representatives of these infrastructures in some of the new activities undertaken in the last year, in particular the CDS participation in the MESR "Groupe Thématique Infrastructures de Services aux Données (ISD)" (Thematic Group on Data Infrastructure and Services). This enables us to follow the developments and also to make CDS visible in this national level group. CDS is also strongly involved in EOSC at the French and European level through the ESCAPE project (now collaboration), the EOSC Future project, and participation in one of the EOSC Association Task Forces (M. Allen member of the "Researcher Engagement and adoption Task Force" 2021-2023).

Recommendation:

We recommend that CDS makes a conscious choice for an informal channel of communication with the user community, stopping short of a user committee. This could be as simple as e.g. a whiteboard at AAS, EAS and ADASS, where there is an opportunity to educate the community on the services available, combined with a constantly-open channel such as querying the community on social media. We see this as fending off the future possibility of disengagement with the community, rather than solving any obvious current problem.

Response:

The CDS has had an active social media presence in the past year, although not as active as in the 50th anniversary year which also included many Gaia data release events. We do however expect a shift of emphasis away from Twitter/X because the use of this platform is becoming more difficult in that users without accounts may not be able to see posts etc., as well as a concern for the integrity of this platform.

We have however been very engaged in live astronomy community events as highlighted in section 3.1, with specific events proposed and co-led by CDS at the National SF2A, and European EAS conferences.

Recommendation:

Given the expected growth of data from new instruments in the coming years and of the user community, it becomes more and more important to define procedures to take decisions and criteria for prioritising and selecting (for example) data to ingest and feature requests to include. We recommend that the processes for prioritisation and decision making are made clear to the CDS Scientific Council. We were pleased to see plans for SKA regional centres with CDS HIPS nodes being deployed remotely (rather than CDS being an SKA regional centre itself).

Response:

In response to this recommendation and also to the discussion on this topic at the 2022 Council meeting, we have formalised the strategy for the ingestion of large survey data in HiPS format into the CDS All-Sky-Data system which serves as the major global ‘HiPS node’. For VizieR and SIMBAD the policy is already well established with the various journals, and arrangements with agencies/observatories such as ESA and ESO. The ingestion of HiPS data is a much more recent activity, which has been set up by the CDS development of HiPS and subsequent IVOA standardisation. We are witnessing a strong growth of the HiPS network, into which any astronomy data centre can publish, but we note that the CDS is still the major node. We have produced an CDS internal document on the ‘HiPS ingestion strategy’ which is included in Appendix 1 of this report. We have also added information to the public Aladin FAQ, and we intend to make this information more visible in the coming year on the CDS web pages and also in our interactions in the community. We expect that the publication of HiPS data will become important with EUCLID, LSST, SKA data, so we wish to make the CDS approach to HiPS publication well known in the community to maintain a leading position.

Recommendation:

CDS currently has four staff positions vacant or imminently vacant due to staff turnover: a documentalist, an Aladin visualisation engineer (to replace a 2024 retirement), a Vizier engineer, and an administrative assistant. We feel it is critical that at least the first three are replaced in the near term in order for CDS to maintain its international leadership role in open science. There is no prospect that the workload will be temporary, due to the exponential growth in astronomical data, so we feel these positions should be permanent. The capacity for CDS to plug the gaps with contractors will be decreasing foreseeably over the next three years. For the longer term, the strategy outlined in the Strasbourg Observatory “Prospective” document makes the case for "an urgent need for 1-2 permanent documentalist positions in the next 5 years" on the grounds of the increase in data volumes; the Scientific Council agrees and endorses this aspiration.

Response:

The needs and priorities for these four staff positions were presented to CNRS-INSU in January 2023, and as described in section 2.1, this resulted in two CNRS positions being opened in 2023 for a documentalist, and for a research engineer focusing on visualisation. Successful recruitments have been made via CNRS competition (‘concours’) and these two new permanent staff members take up their positions on 1 Dec 2023. This is a major success and we thank the council for its support, and the CNRS for opening the positions.

A position for a Vizier engineer was also included in the ‘complementary list’ of CNRS competitions in 2023, in a category available to candidates with a handicap. Complementary list competitions only become open in the exceptional circumstance that another competition fails, and as far as we know this is not the case in 2023. CDS did have a candidate who would have applied if this competition was opened, but this candidate has now been recruited by another CNRS lab. We will continue to request a Vizier engineer position.

As described in section 2.1, an administrative assistant (Christophe Steyer) has been transferred by CNRS into ObAS to support the CDS (an 80% share of a position shared with EOST) in October 2023. This situation will however finish at the end of 2023 because of a leave of absence, and we hope that CNRS will maintain this position for ObAS/CDS.

Recommendation:

The Aladin All-sky service relies on a 5 Pb storage that will need to be replaced in the next few years. As replacing such a large amount of disk may be problematic, we recommend that CDS discuss upstream with INSU to plan a solution.

Response:

The needs for a major renewal of the CDS All-Sky-Data system were presented at the meeting with CNRS-INSU in January 2023. This item has also been indicated on our previsional CNRS budget requests for a number of years showing the need for funding in 2024-25 for the renewal. We have been in discussion with CNRS-INSU and CNES about the request and we hope to have confirmations about the 2024 budget before the end of 2023.

Recommendation:

We recommend that the CNRS, and especially its national computing centre IDRIS, help CDS to investigate the resilience of CDS to cybersecurity attacks, check its processes and where applicable / if available obtain official certification.

Response:

Given that some international astronomy infrastructures have been affected by serious attacks, we have made extra efforts to assess the possible threats to the CDS infrastructure. We have followed all the local UNISTRA guidelines, and in particular:

- The CDS works in close collaboration with the Chief Information Security Officers of our supervisory bodies (CNRS and University of Strasbourg) to guarantee the best possible availability of our infrastructures and the integrity of our data.
- We can now rely on 2 hosting sites within each of our supervisory bodies (UNISTRA Datacenter and IPHC server room) on 2 distant campuses, so we can further improve our resilience in the event of a major disaster on one of these sites.
- With regard to threats linked to network intrusions, a segmentation of our networks has been carried out and by the end of 2024 we will be deploying two additional firewalls validated by the ANSSI, which will provide continuity of service and dedicated protection for the public network that enables the community to access CDS resources.
- Regular audits (once or twice a month) are carried out on our servers to detect and proactively remedy security flaws.
- Several years ago, we deployed a monitoring tool (Security Onion) that enables us to identify network threats and abnormal behaviour on our servers and workstations.

The recommendation to seek national level help and certification is welcome, and while it has not been followed up in 2023 we will follow-up in 2024.

Appendix 1. HiPS Ingestion Strategy Document

HiPS ingestion strategy

Caroline Bot, Thomas Boch and the Aladin team

November 21, 2023

1 Introduction

In an ideal world, we aim for having a collection of Hierarchical surveys that would represent the whole landscape of images in astronomy: the largest available sky coverage (whether observations are scattered pointed fields or global sky surveys), images at all wavelengths, images at all resolutions up to the finest ones, images of all depth up to the deepest ones, ... In practice, while we keep that aim, we have to make choices according to the availability of the data as well as resources (human, storage, computing time,...) and this document aims at defining our strategy, laying down our guidelines and criteria when ingesting images into HiPS.

One fundamental criteria we have chosen refers to the quality: we create HiPS out of data sets that have been published, i.e. there is a paper associated to each dataset we transform. In practice, this means for example that images taken by amateur astronomers that do not have an associated publication are not ingested. Example: The Mellinger survey is available as a HiPS, while images from Ciel Austral group were not ingested.

Another fundamental guideline is that we always favor situations where experts of the datasets are making the HiPS. We will ingest images for which the data providers are not making a HiPS themselves, but we always prefer to encourage that the people who know their data best make the image selection, documentation and HiPS creation. We therefore are happy to help teams that would need advice or feedback and will favor this kind of interaction. Our goal and missions are clearly not to create at CDS all the possible HiPS from all existing datasets, even if we had the resources to do so. This is also a way to improve the quality of the HiPS (ideally made by the people who know best) as well as a criteria to select datasets to ingest, and a way to foster the usage of HiPS by a wider community.

In practice, there is however a large need that CDS creates a given amount of HiPS directly. Even if we presently manage to do most of the HiPS we would like to ingest, even if sometimes with a long delay and different priorities, we are already not able to create all the HiPS we would like to. The pressure is likely to increase even more in the (near) future and this implies to put priorities. In all cases, our criteria to decide whether we want to ingest a HiPS or not will be the same as they are right now and this is what we detail in this document. This decision procedure is made of two parts:

- to ingest a dataset, we need to be aware that it exists and could make a potential HiPS. We describe our strategy for dataset watch in section 2.
- we put priorities on a potential dataset to ingest by balancing different aspects that we describe in section 3

This summarizes the strategy we have to watch for, prioritize and eventually select the image data sets that are converted into Hierarchical Progressive Surveys.

2 Keeping up to date with existing and new surveys

In order to be informed about new surveys and datasets of interest, we make use of several channels :

- all members of the Aladin team (as well as CDS) go to *conferences* and gather informations on current missions and surveys, different available image data sets becoming public, ... This includes scientific conferences, community conferences (e.g. AAS, EAS, SF2A) or data and

service-centered conferences (e.g. ADASS, IVOA Interoperability). This is a good way to keep track of the projects and interests of the astronomical community at large

- we also receive a given number of *newsletters*, including from large astronomical data centers (e.g. ESO, IPAC, MAST, ESA). This is a good way to be aware of large datasets or surveys becoming public.
- some of us follow different *social media*, which is a good way to keep track of press releases, highlights, or interesting/trendy images and news
- part of the data watch is also done *as part of other activities at CDS*. For example, scientists at CDS can be aware of interesting data sets simply by checking the literature, either for their own research, or through the validation of tables for VizieR (this includes associated data in VizieR but not only), or through the weekly bibliography check that happens for Simbad (=g meetings). This information of potentially interesting image surveys is then passed on to the Aladin team.
- we can also be informed of image data sets of interest *from external users*, either by PIs who contact us directly or through users looking for a specific image survey and requesting it through cds-question hotline or direct contact.

Despite these various ways to check for image data sets of interests, we are aware that it is difficult to be complete. We hope that the variety of information sources and people at CDS that perform this watch (either naturally on their daily routine or purposefully), we are not missing datasets that would be of high priority to ingest.

3 Selection and prioritization of datasets to ingest

For any new identified image dataset that can potentially become a HiPS, an evaluation is done of whether a HiPS should be made and how high a priority it is. In defining this selection and priorities, several aspects are taken into account and weighted before making a decision:

- *How does it completement existing HiPS data sets?* This includes arguments like the sky coverage, the availability of other HiPS in the same wavelength range, whether it brings a higher resolution or a deeper intensity view, ...
- *What is the added value of doing this HiPS given the already existing image data sets?* The added value can include cases like gaining the effect of a global view that is not available with the collection of small individual images. It can also be the fact to add a new data release of a dataset for which we already have a HiPS, or to redo a HiPS to be able to have access to FITS images and hence pixel values and dynamics with respect to an existing HiPS with only PNG files.
- *Is there a specific opportunity?* This includes cases like someone willing to invest a certain amount of effort for a specific survey (e.g. D. Durand collaboration for HST and JWST HiPS), a special moment to highlight our services (e.g. JWST early release data sets) or an internal interest (e.g. redoing the GALEX surveys as a testbed for filtering tools).
- *How easy/difficult is it to do?* E.g. a survey distributed as a HEALPix file is a straightforward product to add and is usually an easy addition to the collection for a small amount of computation and storage resources. On the other hand, very large optical surveys may require months of computation and some image data sets may require specific knowledge about the data (e.g. which files, pixel units, filters, keywords for blank pixels or instrumental effects, ...). Again, we reinforce the idea that we always favor the case where experts of the datasets or providers of the datasets make the HiPS. Hence the last question we ask ourselves:
- *Can we encourage and help others to do this HiPS?*

With all these elements to weight on, we assign a priority degree to the considered data set and compare it with other surveys that are currently in our waiting list. In practice, this means that different data sets will take different times to get ingested and become a HiPS. As the data amounts continues to rise, we will eventually set a limit at which some datasets will not be considered. At the moment, this has mainly happened when no publication was associated or if the HiPS was already available elsewhere (even in a different format, e.g. PNG only).