CDS Scientific Council meeting 2022

Summary of CDS activities 2021-2022

02 December 2022

1. Introduction

The past year 2021-22 has been another challenging one for CDS because the COVID-19 pandemic has required continual re-planning of work and events. Remote working was required for the first part of 2022 (03 January - 07 February 2022). The eventual easing of restrictions in March 2022 allowed the internal weekly CDS service meetings to return to in-person, yet many of the external community events throughout the year have been subject to uncertainty and changes. There were numerous COVID-related staff absences in the first half of 2022 (as seen at the national level).

CDS activities have continued throughout the year where we have managed as best we can with the various challenges. The overall production of the CDS has remained very high over this period, with some large peaks in user activity connected with major data releases. The core operations for the ingestion of data from publications have involved the treatment of a large volume of data. High priority data sets have been ingested (e.g. Gaia DR3). New contractors have been recruited to work on CDS data ingestion tasks and also on important upgrades of the CDS services and new innovations. CDS has continued to be strongly engaged with the astronomy community in meetings, conferences and training events, many of which were held on-line.

The 50th anniversary of the CDS has been celebrated and advertised in a number of events. A major CNRS prize was awarded to the CDS documentalists.

The scientific, technical and policy environments in which the CDS operates, are evolving rapidly. In the past year the significant changes include: the launch of the French ‘Recherche Data Gouv’ initiative, new developments of the European Open Science Cloud (EOSC), and the establishment of the SKA Science Region Centre prototyping activities. CDS has been active in all these areas in order to follow the developments and to identify the most important to pursue further.

There have been many activities associated with the European projects in which CDS participates, in particular training events and the preparation of deliverables. These projects have been successful with many benefits for CDS, yet they have also been very demanding in terms of reviews, reports and management. We are now entering a phase where the current large projects will end and new projects will be proposed.

Difficulties have been encountered in some of areas of CDS operations. Staffing issues have been temporarily addressed by contractual staff (3 new contracts in 2022), but the 2 departures of permanent staff in 2021 (reported last year) have not been replaced. The loss of the ObAS logistic staff lead, who had the expertise for managing the server room cooling and electrical infrastructure, has had an impact. The CDS is currently operating without an administrative assistant for managing contracts.

In this report we present the status of CDS in Section 2 and the highlights of the 2021-2022 period in Section 3. The activities of the CDS services are in Section 4, and Projects in Section 5. Extra details about the SIMBAD and Aladin activities are included in Appendices. The recommendations of the CDS Scientific council will be addressed in the meeting presentations.

2. Status of the CDS

In 2021 we reported on the successful application of CDS to renew its status as a “Research Infrastructure” on the French National Research Infrastructure Roadmap, the “Feuille de Route”, established by the Ministry of Higher Education and Research (MESR). This roadmap was launched on 8 March 2022 with a national webinar event, and publication of the document[1]. The CDS entry provides a description of the CDS and its role for Open Science, with recognition of the international dimension of the Virtual Observatory via CDS activities in IVOA and Euro-VO. This new edition includes a section on the ‘place of research data in infrastructures’ and a full chapter on “Research Infrastructures and Open Science” which describe the National Pan for Open Science and EOSC, and which references the long term CDS activities in this area.

In 2022 the CDS has been labelled by the Strasbourg University as a ‘Scientific Platform’ in the CORTECS network. This enables CDS to be made visible alongside the other platforms across all scientific domains in CORTECS network pages[2], as well as bringing new resources.

A new French national initiative has been launched called “Recherche Data Gouv[3]” (RDG). It is described as an “ecosystem for sharing and opening research data”. CDS has been named as one of the first set of six “Thematic Reference Centres” in this new initiative. This is a new status for CDS and the details of the expectations and responsibilities are still to be discussed. Some of the activities of thematic centres that are related to the established role of the CDS are:

- define the appropriate international data description standards in their thematic field;
- define and disseminate best practices for data collection, documentation, processing and dissemination in their thematic field;
- compile and manage a list of national and/or international reference data repositories in their thematic field. Researchers in that field will be asked to deposit their data in these repositories which are to be harvested by *Recherche Data Gouv;*
- support inclusion in and compliance with the international ecosystem.

CDS also has a minor role in the Alsace Helpdesk Data management cluster called ADELE which was proposed by the University of Strasbourg.

The CoreTrustSeal certification of the CDS Vizier service came up for renewal in 2022. We decided to extend the application to certify both the VizieR and Aladin services. The application concerned 16 criteria and was submitted (a 34 page document) in October 2022, with the result expected in early 2023.

In 2022 the CDS is being evaluated as part of the evaluation of the Observatoire Astronomique de Strasbourg by the HCERES *(Haut conseil de l'evaluation de la recherche et de l'enseignement supérieur).* These evaluations are usually done on a 5-year timescale, but a delay was introduced due to the COVID pandemic so this is happening in 2022 rather than 2021. The documents for the evaluation were submitted in Summer 2022, and a virtual visit was conducted in October 2022. The result is expected in the coming months. In addition to the HCERES evaluation, a ‘Prospective’ plan for ObAS for the next 5 years has been submitted to the University and CNRS, and interaction with our authorities about this plan is expected in early 2023. The major CDS elements of the plan will be presented at the Scientific Council meeting.

---

[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

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

3. Highlights 2021-2022

50th anniversary of CDS

A number of events have celebrated the 50th anniversary of CDS, 1972-2022. *(See the list of community events in section 3.1).*

CNRS Collective cristal medal 2022

In June 2022 we got the excellent news that the CDS Documentalists were to be awarded the CNRS Cristal Collective medal. This medal *"rewards teams in support functions who have carried out projects with remarkable technical mastery, collective dimension, applications, innovation and reach"*. The ceremony took place on 14 November 2022, in the presence of families, friends, colleagues and alumni.

Fig 1. The CNRS Cristal collective prize awarded to the CDS documentalists on Nov 14, 2022

Gaia Data Release 3 (DR3)

On 13 June, 2022, the **CDS published the ESA Gaia mission 'Data Release 3'** in the CDS services VizieR, Aladin and the CDS catalogue cross-match service. Gaia DR3 contains 1.8 billion astrophysical sources with their positions, parallaxes, proper motions, radial velocities for ~33 million stars and many other source properties. As part of the publicity around DR3 we made a 30m 'galactic panorama' display (made with *hips2fits*) which is currently mounted on the fence of the Observatory.

Fig 2. Galactic Panorama - 30m display of the Gaia density map mounted on the Observatory fence.

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:

- **Astronomical Data Analysis Software and Systems (ADASS) Conference,** Virtual (Hosted by IDIA, RATT, SARAO and SAAO, Cape Town, South Africa), 24-28 October 2021. *(see the CDS-ADASS document for all CDS contributions).*

- **IVOA Interoperability Meeting, Virtual, 2-4 November 2021.**

- **EOSC Policy Event[4], Strasbourg, 3 May 2022.**
  - Invited presentation — *Enabling Open Science, International dimensions — the example of Astronomy (M. Allen and CDS team)*

- **IVOA Interoperability Meeting, on-line, 25-28 May 2022.**

- **European Astronomical Society annual meeting, Valencia, Spain, 27 June - 01 July 2022:**
  - CDS-50 Booth in exhibit hall.
  - CDS 'Lunch Session'.

Highlighted events in the French community:

- **Science Ouverte et le 50e anniversaire du CDS[5] / Open Science and the 50th anniversary of CDS, 01 April 2022.**
  - Launch event for the 50th anniversary with CNRS, CNRS-INSU, UNISTRA, MESR.

- **Gaia DR3 media events - 13 June 2022.**
  - 50th anniversary news integrated into into French media events for the publication of Gaia DR3.

- **French National Astronomy meeting - SF2A 2022, 8 June 2022.**
  - Invited presentation - 50 years of CDS, today and future challenges *(M. Allen and CDS team).*

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

---

[4] https://eosc.eu/events/eosc-policy-event

[5] https://indico.in2p3.fr/event/26208/

4. Activity Report for CDS Services 2021-2022

4.1 CDS Information System

The CDS services continue to have a very high level of use with an average of 2.5 million queries per day from 360K unique IPs/month in 2021-22 (Table 1). Figure 3 shows the monthly queries on a 6 year timescale which indicate a general increase but also fluctuations and peaks of activity. A major peak occurred in July 2022 which is due to HiPS data being accessed when the first JWST data were released (see the Aladin report).

Main Services ➔	SIMBAD	VizieR	Aladin	Total (main services)
users / month	**154k** (+6%)	**45k** (+17%)	**360k** (+53%)	**> 360k***
queries / day	**344k** (-12%)	**432k** (-28%)	**1.7M** (+73%)	**2.5M**
load / day	**7.1 GB** (+32%)		**162 GB (+52%)**	**>169 GB**
data volume	**35 GB** (+28%)	**80 TB** (+51%)	**427 TB (+5%)**	**507 TB**
data content	**13.5 M obj.** (+7%)	**22.8k cats** (+7%)	**1068 HiPS (+6%)**	
reliability	**99.75 %**	**99.11 %**	**99.93 %**	**> 99.11%**

Table 1. CDS statistics Oct 2021 - Sept 2022. (Note that there is overlap of the users/month between the CDS services)*

The reliability remained very high in the past year at >99.11%. 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. Continued implementation of server virtualisation via a cluster of VMWare hypervisors, and Docker containers 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").

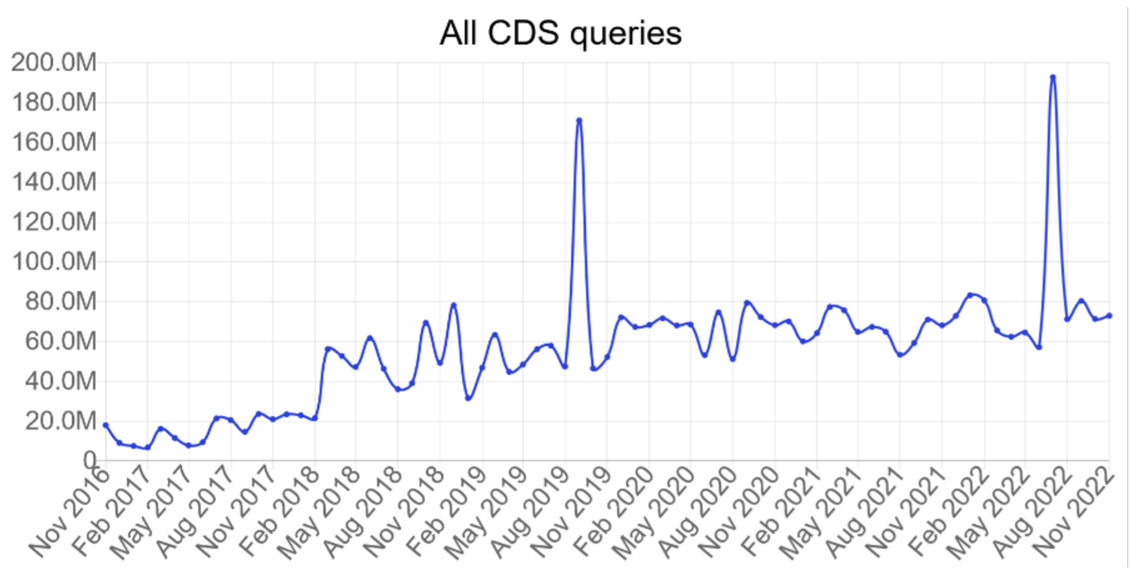

Figure 3. CDS queries on all services 2016-2022

Up-coming changes to the physical hosting of CDS servers

In 2021-22 it has been necessary to rely strongly on these elements in particular because we have needed to respond to changes that have implications for the operational implementation of this architecture.

The main change is that we are now planning for the retirement of CDS installations in the local server room on the ObAS site, much sooner than previously anticipated. This has been driven by a number of factors: ***i)*** the loss of the leader of the ObAS logistics team who had specific expertise for the electrical and cooling systems of the local server room, ***ii)*** technical limitations of the server room and strong encouragement from the University to retire this room (corresponding to University and national level policies of rationalisation/reduction of small server rooms) as well as ***iii)*** the unexpected situation of the energy crisis and the real risk of electricity cuts to the ObAS server room.

As reported in 2021, the physical movement of CDS servers into the University Data Centre is already underway (with for example one duplicate of the All-Sky-Data server is fully operational in the data centre), and this will become the main local site for our installations.

The solution that is emerging for the secondary local site, is a partnership with the IPHC (Institut Pluridisciplinaire Hubert Curien), a mixed research unity of CNRS et Université de Strasbourg UMR7178) on the CNRS Cronenbourg campus. The technical and functional conditions appear a priori to be adapted to our desired level of operations. We aim to begin relocation of servers in the first half of 2023. There will be costs incurred for the relocation, and also continuing costs for the hosting of CDS servers at IPHC which are currently under discussion.

Status of CDS mirror sites

In addition to the local sites in Strasbourg, the CDS has for many years had several mirror sites hosted by partner institutes. SIMBAD is replicated at CfA/Harvard. VizieR currently has operational mirrors at 6 sites (CfA/Harvard, NAO Japan, IUCAA India, Beijing Observatory China, INASAN Russia, IDIA South Africa). CDS also relies on some twenty partner sites (HiPS nodes) for the data used by Aladin. They replicate the most requested HiPS image records, but this represents only a small part of the total volume (<10%).

Updates to the Information System in 2021-22

Among the many updates to the information system done throughout the we can mention the acquisition of a system for the storage of “original and intermediate data” which facilitates the eventual reprocessing of the original data (which is sometimes done for large image surveys). The 800TB of disk storage has been installed at the University Data Centre.

This is distinct from the CDS All-Sky-Data system which is the operational (1.6 PB x 2) data server supporting large numbers of external queries. This system is operating smoothly, but we note that its renewal is coming up soon in ~2025, which will be a major cost item (~500 k€) for which we seek a method to apply for funds. (The original system was funded via a CNRS *mi-lourds* proposal in 2018, plus support from CNES, University of Strasbourg, and a further exceptional CNRS allocation).

In 2021-22 we have made a number of developments and tests of new technologies for the CDS computing infrastructure in close collaboration with the ObAS IT team. A contract engineer (M. Misslin) has helped with a migration of the IRODS system, and has helped set up the global management of virtual machines using Kubernetes.

Renewal of the CDS web pages is currently in progress with the new systems expected to be implemented in 2023.

4.2 SIMBAD

Overview of the SIMBAD Content

SIMBAD is a bibliographic database of astronomical objects of interest. It is the oldest service of the CDS, which continues to evolve and to be heavily used (344,200 requests per day on average in the past year). The team comprises 3 software engineers (plus two participating to special operations), 9 documentalists organised into 3 specialised teams, 6 staff astronomers and 1 post-doc bringing their scientific expertise (5 to 20 % of their time), and 2 staff astronomers leaders of the nomenclature and scientific content. SIMBAD is built from the published literature and very large surveys. It includes all astronomical object types : stars and planets, sets of stars, galaxies, sets of galaxies, interstellar medium, gravitational sources. For each object SIMBAD provides the list of references where it has been cited, the main object type as well as a list of secondary objects types, and cross-identifications with the corresponding list of identifiers.

SIMBAD also provides data collected in papers and surveys, organised into “Basic data” and “Collection of measurements”. The “Basic data” includes: coordinates, proper motions, parallax, HRV/redshift, spectral type, morphological type, size for extended objects, magnitudes. The “Collections of measurements” includes: HRV/redshift, proper motions, parallax, distance, spectral type, stellar fundamental parameters, variability. SIMBAD now contains 13.7 million astronomical objects, 1 million more than a year ago. Table 2 summarises the current content. (*A detailed explanation of the SIMBAD workflow is included in Appendix 1).* The evolution of the number of new references added in SIMBAD per year is shown in Figure 4. It is steadily increasing since 1991.

Overview	N (million)
Objects	13.715
Identifiers	54.449
References	0.411
Citations of objects in articles	30.952
Acronyms	0.015
Basic data	**N (million)**
HRV/redshift	6.395
Proper motions	7.016
Parallax	6.384
Spectral type	0.858
Morphological type	0.145
Magnitudes (UBVRIGJHKugriz)	11.379
Collection of measurements	**N (million)**
HRV/redshift	9.991
Proper motions	13.686
Parallax	14.421
Distance	9.230
Spectral type (spectroscopic only)	1.188
Teff logg Fe/H (spectroscopic only)	2.825
Variability	2.560

Table 2. SIMBAD Content 2022

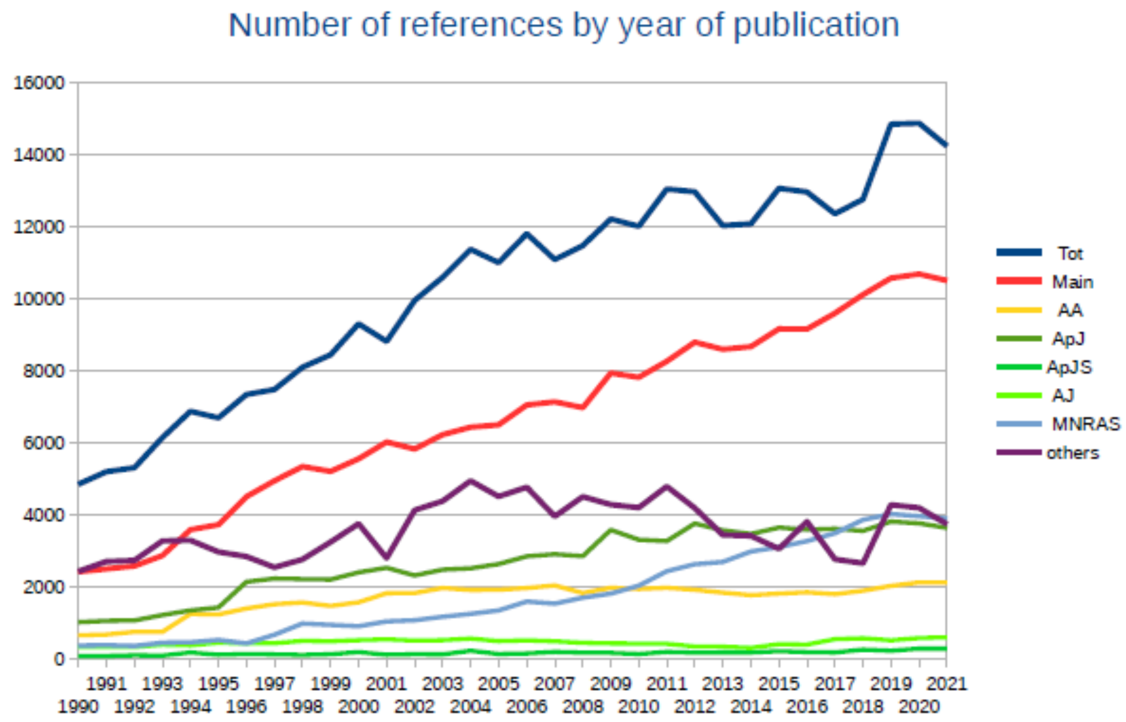

Figure 4. References in SIMBAD 1990-2021

Reorganisation and Maintenance of Objects Types

Last year we announced that the reorganisation of the astronomical objects types in SIMBAD and its integration into the SIMBAD processing software, proving a scientific improvement of the hierarchies and compatibilities of the object types .This year it was successfully integrated into the whole SIMBAD database, which for example now enables a search by criteria or using TAP (Table Access Protocol) queries. The reorganised list of objects types is available with an interactive web page[6] for exploring the hierarchy.

Gaia DR3 in SIMBAD

i) Massive cross-identification of SIMBAD with Gaia (E)DR3

A massive cross-identification of SIMBAD with Gaia EDR3 was performed in February 2022. In total 6.5 million objects have been updated. Another 650,000 objects added between June 2021 and May 2022 were updated in June 2022. Note that we are still ingesting some references based on Gaia DR2, or from other surveys, so that a third iteration will be required. As the number of sources did not change from EDR3 to DR3, all Gaia EDR3 identifiers now appear as Gaia DR3. The Gaia DR3 data stored in SIMBAD are : coordinates, proper motions, parallax, distance if the measurement uncertainty on the parallax is less than 10%, heliocentric velocity, and G magnitude. There are now 7.5 million Gaia DR3 sources in Simbad. It means that 86% of the stars in SIMBAD are identified in Gaia DR3. *(See Appendix 2 for details)*

ii) Gaia DR3 tables of objects of interest

The Gaia DR3 tables were published in VizieR (see VizieR report). These tables were appraised for possible ingestion in SIMBAD at the end of September 2022 : variable stars, non-single stars, and extra-galactic objects. Out of 25 tables, 20 have a priority 1, 3 have a priority 2+, and 2 did not appear relevant for the content of SIMBAD. All pulsating stars, eclipsing binaries, and non-single stars, with a period, will be ingested in SIMBAD. Among the extra-galactic objects, galaxy candidates with a size are being cross-identified with the content of SIMBAD, but no new objects will be created. The processing started mid-October 2022. Table 3 summarises the present status.

---

[6] http://simbad.cds.unistra.fr/guide/otypes.htx

Variability	Priority	N Period>0	Status/Comments
Cepheids and RV Tauri	1	14,987	Done
Main sequence oscillators	1	54,476	Done
Long Period Variables	1	392,240	Being processed (150,000 done)
RR Lyrae	1	271,779	Being processed
Eclipsing Binaries	1	2,190,738	To be processed (compact companion done)
Micro-lensing events	1	363	Done
Rotational Modulation	2+	2,935,262	
Non-single stars	**Priority**	**N Period>0**	**Status/Comments**
Spectrocopic Binaries	1	220,527	Being processed (139,200 done)
Eclipsing Binaries	1	86,918	To be processed
Astrometric Binaries	1	135,760	To be processed
Extra-galactic	**Priority**	**N**	**Status/Comments**
Galaxy candidates	1	424,786	Being processed / Limited to galaxies with a size (RadS>0)

Table 3. Status of Gaia DR3 improvements to SIMBAD

Bibliographical Centre Supervisor

The new process for ingesting references directly from publishers was implemented last year. It has been built to manage the references from the major publishers (EDP, AAS, OUP) and involves 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. Continuation of developments are still in progress to help the curation before the ingestion of bibliographical data in SIMBAD.

Dictionary of Nomenclature

The object identification system (Dictionary of Nomenclature) re-construction will be the major software project in the next years. After the first step of prototyping, that should be heavily tested and validated by both astronomers and documentalists, it will be followed by developments in the core of SIMBAD, and other internal tools. Maintenance and evolution of SIMBAD softwares and libraries should continue in parallel.

Enhancing object distances in SIMBAD

Very large surveys are changing the historical view of SIMBAD. Objects of interest are not anymore only those that have been studied in the literature. Here it should be recalled that the great majority of published articles deal with small numbers of objects, typically a few to a few thousands. Among objects of interest are those with a reasonably well known distance. All stars with a distance less than 100pc (as deduced from Gaia DR3) will be ingested in SIMBAD. For such stars the criteria Parallax / error(Parallax) < 10 is not sufficient as the sample would be contaminated by too many incorrect measurements. We thus apply error(Parallax) < 0.9 as it minimises the number of spurious parallaxes. At present all of the Gaia DR3 stars following this criteria and with Parallax ≥ 30.0 have been ingested.

We also have the project to ingest all SDSS galaxies with a reliable spectroscopic redshift. That would double the number of galaxies with a distance measurement in SIMBAD. The main problem of this project is that the SDSS does not provide sizes and often finds several sources in big galaxies. We are currently studying the criteria that can be applied to solve this problem.

4.3 VizieR

VizieR content

Despite occasional resurgence of COVID infections, the work rhythm in the VizieR team has returned to a the balanced mix of in-person and remote work, and VizieR meetings are held in-person. The team has been enhanced with the recruitment of contractual staff:

- A. Flint, VizieR developer, 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).

- A. Fiallos, VizieR documentalist, in May 2022.

A. Flint is helping to support the renewal of VizieR code, some of which can be several decades old, making it difficult to maintain. The code renewal aims to address separate aspects of the VizieR ecosystem — analysis of the data sent by the authors, metadata generation, insertion in to the database, retrieval, interface and display. Some aspects have already been addressed, such as the textual indexation with ElasticSearch in production since about 2 years, and the global MOC/HEALPix indexation (cf. workflow and technical evolutions below). This very welcome arrival will help the service continue this modernisation effort.

A. Fiallos strengthens the team of documentalists for the processing of tables from the main journals, and is close to completing her 6 months training by VizieR documentalist P. Vannier. Vizier is now (temporarily) supported by 4 dedicated documentalists for the first time in its history. As a result, the number of references processed in the past year is higher than in previous years (+1440), which brings the total number of catalogs in VizieR to 22852. The VizieR volumetry increased from 52.8 TB in 2021 to 85.5TB in 2022, a ~60% increase, driven mainly by the new large catalogs and in particular Gaia DR3.

VizieR Content

	2017	2018	2019	2020	2021	**2022**
Number of Catalogues	16 528	17 673	19189	20289	21412	**22852 (+1440)**

Table 4. VizieR Content 2016 - 2021 Oct 2020 - Sept 2021

Gaia DR3 in VizieR: ingesting and distributing an enormous dataset

A highlight of 2022 was the release of Gaia DR3 on 13 June 2022. The team was on a very tight schedule to complete the ingestion of this enormous dataset. CDS received the data just one month before the public release date (the DPAC plan was to provide it 2 months ahead, but this was delayed). The breadth of the data and its volume put the team under pressure. Ultimately, the data was published in VizieR on-time for the DR3 release date. In total almost 70 individual tables grouped by topic into 6 VizieR catalogs have been made available. Among them are 11 "large tables" (i.e. requiring the large catalogue treatment), so that Gaia DR3 can be considered to be indeed equivalent, in terms of processing by the VizieR team, to 11 large catalogues (in just one month!). Gaia DR3 also pushed the envelope of the large catalogue service capacities. Indeed, the largest Gaia DR3 table we ingested counts almost 12 billion records (the BP/RP spectra), outmatching The Guide Star Catalogue, Version 2.4.2 (GSC2.4.2), at almost 3.5 billion records, which was previously our largest catalogue.

Very large *versus* **long** *versus* **thick catalogs**

Besides this very high intensity episode, 7 other very large catalogues were ingested: SDSS DR16, DES DR2, DESI Legacy Imaging Surveys DR8, GSC2.4.2 (end of October 2021), StarHORSE2021, TIC v8.2, GDR2AP (astrophysical parameters from Gaia DR2 and several other combined surveys).

We expect to ingest in the 2022-2023 period, DES: DESI DR8, DECALS DR9 (or DR10 if it is released soon enough), Pan-STARRS DR2 (detailed definition of tables to transfer in progress), EROS (contact taken, pending sample data), NOAO Source Catalogue (NSC, flat catalogue, the contact is taken, data transfer pending), KIDS DR4, ESO PHASE3: ATLAS-DR4, VPHAS+ DR3.2 , VIKING DR4.

The trend of "long" catalogues continues in 2022: small collaborations, even a single author, working with a large input dataset such as Gaia, are able to produce large catalogues of a billion sources and more, which we must process through the large catalogues pipeline, even though they may have just one added quantity 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 StarHORSE2021, which, as predicted in the 2021 report, released the EDR3 version of their distance estimates.

Another continuing trend is that of "thick catalogues", such as SDSS-RM (Reverberation Mapping, Shen et al. 2019), with 849 objects but 472 columns. Such complexity naturally requires documentalists to spend much more time than average on these catalogues.

VizieR usage

In terms of use of the service there was an average of 432000 queries/day in 2022. The huge majority of queries are directed at VizieR catalogues, and are performed mostly via python, other APIs, and the web forms. The associated data service saw a decrease in usage, while TAPVizieR usage almost doubled over the year. This last point is encouraging and could explain the small drop in total queries, considering that TAPVizier queries are generally more powerful than other simple VizieR queries, although such an interpretation would require a more thorough analysis of the logs.

VizieR Usage-1

	2017	2018	2019	**2020**	**2021**	**2022**
All queries, /day	326 000	368 000	696 000	520 000	**514 000**	**432 000**
Associated Data Service Queries/day	270	80	543	845	**1212**	**811**
VizieR TAP service Queries/day	5 250	3 700	2094	14 000	**23 130**	**43 000**

Table 5. VizieR usage 2017-2022 (queries/day)

Workflow and technical evolutions

- A new syntax corrector has been developed for the parameter files generated by the VizieR documentalists during catalogue ingestion. This tool is based on *antlr* analysing a json version of the parameter file, using a grammar file. It is intended to help improve the ergonomy for documentalists and speed up ingestion by detecting simple errors such as missing or unmatched parentheses, erroneous function calls etc.

- Global indexation: all catalogue footprints are now described using MOCs (Multi-Order Coverage maps), and the computation of these MOCs is now part of the standard workflow for all catalogues.

- The compliance of VizieR with the VO standard VOunits[7] has been improved.

- CoreTrustSeal certification: The application for renewal of the certification was submitted in September 2022. The application includes a more accurate description of the licensing aspect of the data distributed by VizieR.

- An important update to VizieR database system was performed through the migration from Postgres 9.6 (released in 2016) to Postgres 14 (released in 2021). This involves VizieR at CDS and its international mirrors, TAPVizier, and the associated data service. In total 12 databases have been successfully migrated.

Future considerations for the associated data workflow

The VizieR associated data service provides access to non-tabular (FITS format) data associated with a journal publication. A crucial step of this workflow is the mapping of fits keywords to the IVOA Observational Core Data Model (ObsCore). This mapping is currently performed by a module of a SAADA (a tool developed at ObAS by L. Michel and adopted by CDS). The imminent retirement of L. Michel requires us to re-think this strategy and most likely to re-build a mapping engine that can be supported and maintained within CDS. We are currently exploring the development of a new tool would allow us to tailor more closely to VizieR needs.

[7] https://www.ivoa.net/documents/VOUnits/20140523/VOUnits-REC-1.0-20140523.pdf

4.4 CDS X-Match

The X-Match service continues to operate smoothly with the current established code.

Usage Statistics

The service is mostly used via its API with only a small number of queries made though the web interface. Following a large jump in usage last year, the number of API queries is currently ~16k jobs per day. The total number of associations has not changed very much. This is an interesting change in the user behaviour which will be closely followed to understand how it is evolving.

CDS X-Match Service Usage

	2013	2014	2015	2016	2017	2018	2019	2020	2021	**2022**
Web interface (jobs/day)	15	16	20	30	33	40	43	49	60	**42**
HTTP API (jobs/day)	47	50	580	889	1256	2687	2556	7244	21986	**16317**
Associations/day (Web Interface)	13M	~70M	~55M	~104M	~164M	~179M	~44M	~72M	~85M	**~63M**
Associations/day (HTTP API)	298k	~1.6M	~6.6M	~6.7M	17.8M	50.2M	43M	90M	78M	**78M**

Table 6. CDS X-Match service usage statistics 2013-22

Future

Here we list the points we would like to achieve during the coming years, noting that some of these are very ambitious:

- Continue the development of the next generation cross-match service prototype ‘ExXmatch’.
- Redesign the current interactive Web Interface.
- Redesign the structure, API and code of the X-Match service.
  - provide all VizieR columns while allowing the user choose the desired column set.
  - allow post filtering to reduce the size of output file.
  - remove the distinction between large and small jobs by automatically splitting large jobs into smaller, individual jobs.
  - let the user upload large tables (and automatically transform the file on the server to prepare it for efficient cross-matches).
- Provide a command line client for a better programmatic access.
- Allow for cross-matches taking into account proper motions.
- Explore distributed cross-matches for better performances.
- Explore the possibility to let the user make plot from the server data to avoid large file transfers.

4.5 Aladin

In 2022 the Aladin project reached its 30 year milestone. Recalling the earliest publication at the start of the project (Paillou et al. 1994[8]), we can note that the spirit has remained the same (being able to compare data sets, view the context of astronomical objects, integrate CDS databases and connect to international ones,...). Yet, the framework of the project has completely evolved, with two clients (Aladin Desktop and Aladin Lite) and the revolution of hierarchical progressive surveys (HiPS). This evolution has of course been continuous and is still ongoing with major new versions:

Aladin Desktop Version 12

Aladin Desktop V12 was released in Summer 2022. It includes several new features and improvements:

- Photometric tools: HiPS support, pixel extraction by region.
- Planetary data support: HiPS, MOCs, catalogues.
- Spectrum manipulation: spectrum extraction from cubes, CASSIS plugin integration (IRAP development).
- MOC: support of MOC 2.0 standard, STMOC integration (space and time filtering).
- HiPS improvements: faster downloading thanks to gzip/nogzip automatic strategy for FITS tiles, better display in any projection (notably ARC, CAR).
- New supported projections: Mercator, HEALPix.
- FITS improvement: PanSTARRs BOFFSET/BSOFTEN special encoding supported.

The audience for Aladin Desktop remains at a high level of 6560 users/month, 1010 launches/day.

Aladin Lite Version 3

Aladin Lite v3 has been released for beta testing late June 2022. It integrated the following new features:

- Improved display (thanks to GPU rendering using WebGL2).
- Access to FITS HiPS tiles: HiPS FITS tiles can now be loaded and visualised in Aladin Lite, giving access to the whole dynamic range of corresponding HiPS dataset.
- New projections available: SIN (orthographic), AIT (Hammer-Aitoff), MOL (Mollweide), MER (Mercator), ARC (zenithal/azimuthal equidis- tant), TAN (gnomonic), HPX (HEALPix). *See — Figure 5.*
- Coordinate grid displays.
- Access to all 1000+ available image HiPS.
- Access to all VizieR tables.
- Overlay multiple image HiPS.
- Contrast adjustment.
- New colour maps and stretch functions. The user can also change the stretch functions with options of pow2, linear, sqrt, asinh and log.

[8] https://ui.adsabs.harvard.edu/abs/1994IAUS..161..347P/abstract

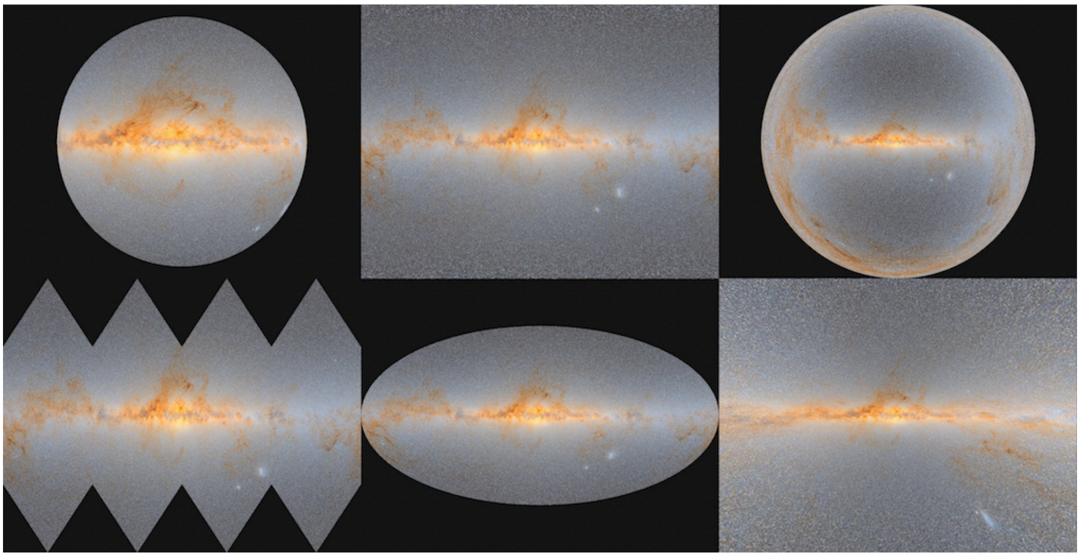

Figure 5. Different projections enabled by Aladin Lite v3 suing the example of the Gaia flux map.

Special attention has been paid to ensure the compatibility of the existing API, as to minimise the cost of migrating to v3. The feedback we received has been useful in tracking down bugs due to inconsistencies across browsers and operating systems. Beta testers have been generally quite enthusiastic about the new version. An official release will occur in January 2023.

While the number of daily startups of the Aladin Lite previewer is 608 startups per day, the number of startups from sites embedding Aladin Lite is increasing significantly as they add up to 30700 startups per day. Among the new sites using Aladin Lite are the SNAD ZTF object viewer, the ALeRCE ZTF Explorer and esawebb.org.

Hierarchical Progressive Surveys

HiPS has now been the default image format for Aladin for a few years. While the main interests at the beginning were to use the hierarchical properties and individual tiles to be able to seamlessly browse from the full sky down to very small regions, the fact that all data sets are stored on exactly the same pixel grid also offers interesting benefits. We advocate that HiPS can be used for scientific analysis directly and not solely for data browsing.

Currently, the HiPS network gives access to 1168 image HiPS (including data cubes and planetary surfaces). We welcome and prefer when observatories hosting the original data sets create their own HiPS and publish it to the HiPS network, and we help them to do so as needed (e.g. LOFAR data set this year). Yet we also continue to create many HiPS at CDS and improve former ones we created according to needs.

New HiPS surveys done at CDS this year include:

- GALEX GR6/7 (co-added surveys, photometric improvement using counts response maps)
- MeerKAT Galactic Centre
- JWST First images
- ZTF color
- Gaia DR3 flux map
- HST 3D-DASH
- HSC Legacy Archive g and r bands
- QUEST Herschel-Planck feathered data • ESO Outreach
- HST Outreach

We have also started to include a photometric correction to the PanSTARRS HiPS. These HiPS are used by the XMM2Athena project which pointed us to a specificity of the data set that was affecting the photometry. One of the PanSTARRS band has been fully reprocessed and is now photometrically precise. Reprocessing of the other bands is planned and this special encoding of FITS keywords is now supported.

The CDS HiPS node manages a very high level of use by the community at 1.7 million tile requests per day. In 2022 we experienced a significant peak in the HiPS usage via the World Wide Telescope (WWT) on July 13 2022 when the first JWST images were released, resulting in ~1000 requests/second. *(See Appendix 3 for the details of how we responded to this peak)*.

Data cube advances

The main focus of the Aladin project has always been toward images. Image data cubes have long been incorporated in Aladin Desktop data sets as well and visible by browsing the velocity/wavelength depth of the cube one plane at a time, in a movie-like mode. While this is useful and builds on the strengths of Aladin image capabilities, it has been clear that exploring data cubes in Aladin is not as seamless and requires new features and developments.

We have strengthened our collaboration with the CASSIS (spectral tool) team in Grenoble, and this resulted in the development of a plugin to extract spectra out of individual FITS data cubes loaded in Aladin. The extracted spectrum (for a position or a region) is sent to CASSIS and user interactions on the spectral side and the spatial side work along thanks to the interoperability of the tools.

We are also using the context of the SKA to prepare Aladin for very large data cubes. Several members of the Aladin team are part of the SKA SRC prototyping activity, working on ‘distributed data visualisation and discovery’. Our implication in SKA is in alignment with our broader needs for advances on large data cubes and while we participate in the SKA prototyping phase, the developments and test we make are based on our current tools and expertise and meant to be generic and useful for any future large data cubes. Among the prototype developments, we are both testing information excerpts services on HiPS cubes (moment maps, on-the-fly spectral extraction, on-the-fly combination of planes in the 3rd dimension) and pushing limits of the HiPS cube format for very large data sets (changes of velocity axis, file format for tiles, importance of tile size, improvements of *hipsgen*, discussion of hierarchical sampling in the third dimension). The challenge of these large cubes should not be understated, and the adaptations to Aladin are not completely clearly set yet, however it is research in-progress and SKA is providing an good incentive to move in that direction, probably opening new paths for large data cubes in many other wavelength ranges.

30 years after its conception, the Aladin project is very active, at the forefront of future challenges and very much used. With the breadth of the project as it is today, we would like to stress that the Aladin team is dependant on current and future human resources. We are happy that M. Baumann is back in the team and we see the impact on Aladin Lite in particular. Yet we are aware this is a temporary contract. Similarly, the implication of the Aladin team in the SKA project and how this is driving developments on the visualisation of cubes is relying on F. Bonnarel who will retire in 2024. As for the HiPS generation, one documentalist M. Buga works part time on this task. There are also contributions from P. Fernique, T. Boch, D. Durand (retired CADC). As usual, careful selection of the data sets we create and functionalities we develop, will be chosen and scaled according to our available resources and the evolving landscape, with the standing goal to implement the most meaningful services on imaging and data discovery, for a wide usage, by the astronomical community at large.

4.6 R&D

In 2021-22 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**. One of them was later hired as an apprentice for one year. 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, providing at the end a team work. After the smaller number of interns in 2021 (7), in 2022 we returned to a normal level with a similar number of interns expected for 2023.

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

- We continued the study called "Inclusive CDS" started in 2021, with the aim to provide access to the services and data to all, including people with disabilities.

- We continued also the work around the query of the CDS services in Natural Language with the aim to integrate it in a Lab section of the CDS website to open it to a larger panel of users.

- An intern worked on the instruments Field of View data model / format in the Aladin Desktop.

- Two interns worked on the X-Match service, involving Rust skills .

- Two interns worked also around the XMM-Newton data through the implementation of a TAP interface for the XMM-Newton mission and an online data reduction service.

- An intern developed a tool to analyse Aladin HiPS logs, resulting in a poster for the ADASS conference about internet speed as inferred from the logs.

- We worked also on a LaTex parser (based on ANTLR) in the frame of VizieR to validate configuration files.

We started also supported some small "research discovering" internships on the topics of ‘Data Lakes’ and the ‘CDS Chatbot’, giving students an experience of being in the research/development environment of the CDS/ObAS.

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 2022 Gilles Landais became the Chair of the Data Curation and Preservation Interest Group following André Schaaff's term in this role. 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
- **Deputy Chair of the Data Access Layer Working Group** (since May 2021) - G. Mantelet
- **Chair of the Data Curation and Preservation Interest Group** - G. Landais
- **Deputy Chair of the Radio Astronomy Interest Group** - F. Bonnarel
- **Chair of the Education Interest Group** - H. Heinl *(CDS-ESCAPE Support Engineer)*
- **Editorial team for the IVOA Newsletter** - S. Amodeo *(CDS-ESCAPE Science Support Postdoc)*

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 will conclude in January 2023. CDS has lead a major work package (WP4, 'CEVO') on *Connecting ESFRI projects to EOSC through VO framework.*

There have been many ESCAPE activities in this final year of the project. A major review of the project was passed in March 2022. The WP4 activities under the responsibility of CDS as the Work Package lead partner have been:

- **D4.5 - Prototype demonstrator for value-added archive services** - *Report[9] delivered January 2022*
- **D4.6 - *2nd School with Interoperable Data***, *On-line hosted by CDS in February 2022 (Report[10] delivered June 2022)*
- **WP4 3rd Technology Forum[11],** *On-line hosted by CDS, 15-16 March 2022*
- **D4.7 - Final Analysis Report[12] on VO data and service integration into EOSC** - *Delivered September 2022.*
- **D4.8 - Final analysis report on IVOA standards and stewardship best practices** - *Delivered November 2022.*

A concluding event, **ESCAPE to the Future**[13]**,** was held in *Brussels 25-26 October 2022.*

---

[9] https://projectescape.eu/sites/default/files/ESCAPE_D4.5-v2.pdf

[10] https://cloud.escape2020.de/index.php/s/rCVs7hvTzUjxXlg

[11] Event page: https://indico.in2p3.fr/event/26364/

[12] https://cloud.escape2020.de/index.php/s/xf3ytXYuI6fwQ5i

[13] Event page including recordings: https://projectescape.eu/agenda-escape-future

Many other ESCAPE meetings of have been organised by F. Bonnarel focusing on VO tools and standards for radio astronomy and high energy astronomy. All the activities are listed on the WP4 wiki pages[14]. The work on radio astronomy was presented at the EAS 2021 and an article is being prepared for Springer special volume on "Data-Intensive Radio Astronomy". Aladin Lite v3, the enhanced tutorials, MOC 2.0 standard are among the results of the project.

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) started in September 2022 to work on this project (and other CDS projects).

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 continues to be hosted by CNRS and to get support from the National Fund for Open Science. The RDA Regional Advisory Board gathers all the "regions" and countries which provide financial or in kind support to RDA. The co-chairs of the Regional Advisory Board are invited members of the RDA Council .

The RDA is supported by the 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 IVOA is for instance taken as one of the examples of "Global Open Research Commons" (GORC) tackled by the RDA GORC Interest Group and a Working Group on GORC Models, which started with in mind generic frameworks such as the EOSC but should also include thematic ones. Work is on-going in the Working Group to define GORC attributes. IVOA characteristics will be taken into account.

F. Genova with colleagues from different regions is also proposing Evaluation of Research as a new topic for the RDA - it is well recognised that a key obstacle to the development of Open Science is the current bibliometric-driven evaluation methodology. The evolution of research evaluation is an important facet of the national and European Open Science Strategy.
A local group "RDA UNISTRA" was set up in the framework of Strasbourg University Open Science Strategy to disseminate information about the RDA in the University community.

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

5.3 XMM2ATHENA

As part of her independent science A. Nebot is part of a successful proposal in the Horizon 2020 call on SPACE-30-SCI-2020: Scientific data exploitation. The project XMM2ATHENA[15] 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. A. 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. The project started on 01 April 2021 and in August 2021 J. Kuuttila joined the team with a two year postdoctoral position, to study the young stellar and binary contributions to the X-ray emission of our Galaxy. Thanks to the extensive use of the CDS services and tools, e.g. extensive use of the cross-match service, VizieR, HiPS and hips2fits we could spot and report on catalogues available to the community but not in VizieR, some missing metadata and/or associated description, errors on the PanSTARRS HiPS which had wrong values,... This interaction lead to improving the CDS services. Unfortunately, J. Kuuttila decided not to extend his contract due to personal reasons, and we are currently looking for a replacement for a 12-24 months postdoc.

[15] http://xmm-ssc.irap.omp.eu/xmm2athena/

Appendix 1. SIMBAD Workflow details

Much care has been taken in the management of the SIMBAD workflow throughout the 2.5 years pandemic period because it involves a number of steps, and coordination between the SIMBAD and VizieR team. We recall the major steps of the SIMBAD workflow:

1. First a new reference is created in the database (with title, authors, abstract, keywords, but no linked objects yet).

2. The reference is processed by a specialised team of 4 documentalists using the software DJIN. Objects in the text, figures, and small tables, are updated or created in the database and linked to the reference. In addition, flags are added if there is a need for a new acronym (step 3): if a large table has to be ingested in the VizieR database (step 4), or has to be appraised (step 5) and processed by another team by script (step 6). The two first steps follow closely the publications are completely up-to-date.

   9 journals are regularly processed in SIMBAD with a priority 1 (A&A, AJ, ApJ +L, ApJS, MNRAS +L, NatAs, Natur, PASP, PASJ, Sci), 12 with a priority 2 (AcA, AN, ARep, AstBu, AstL, ATel, BaltA, NewA, NewAR, RMxAA, RAA, RNAAS), and 9 with a priority 3 (A&ARV, Ap, ARA&AA, ChA&A, JAD, JApA, Msngr, PASA). The evolution of the number of new references added in SIMBAD per year is shown in figure **x**. It is steadily increasing since 1991.

3. Third step : when a new acronym is used by the authors, it is first checked that it is really new rather than incorrectly written or coming from a previous reference or un-useful. If it is really needed, the acronym and its format(s) are documented in the CDS Dictionary of Nomenclature, a default object type is assigned, and it is finally encoded in the database. This team comprises 1 dedicated documentalist, as well as 1 software engineer and 1 astronomer (part-time). There are now 15,221 acronyms encoded in SIMBAD.

If a reference contains large tables of objects, from about 50 to 500,000 sources, there are 3 additional steps:

4. Ingestion in the VizieR database (see the report about VizieR).

5. Appraisal in bi-weekly meetings involving astronomers and documentalists to assess the relevance and the priority for ingestion in SIMBAD. In 2021 we had 58 meetings where 794 articles with tables of objects were appraised. 70% were assigned a priority 1.

6. Processing by script by a team of 4 specialised documentalists using the COSIM software which allows “semi-automatic” cross-identifications, not only based on positions, but comparing various data (like magnitudes) and checking the compatibility between object types. Typically 90% of the objects are updated or created automatically, while 10% need additional expertise from the documentalists or/and the astronomers.

These 4 last steps require much more time than the two first ones. The time between publication and ingestion in SIMBAD is typically 1.5 ±1 year. The delay is however about 1 to 2 year more for the MNRAS ; In 2021, 706 references with table(s) of objects have been ingested. At end of October 2022 they are 589 ; an extrapolation gives an estimate of 707 at the end of the year.

Working meetings mostly involving documentalists and software engineers are organized on a regular basis in order to identify, prepare, and develop improvements of the processing softwares on both a short and long time-scale.

Appendix 2. SIMBAD cross-match with Gaia - the details

SIMBAD being an historical meta-compilation of objects with all astronomical types, it is inhomogeneous by construction. It contains extended objects like clusters of stars/galaxies, or molecular clouds, that certainly have no counterpart in Gaia. It also contains objects in very crowded environments, and there are still stars and galaxies with an inaccurate position. We thus do not apply a blind positional cross-match on the entire database as that would lead to a few 100,000 false cross-identifications.

To perform a massive cross-identification between SIMBAD and Gaia we of course have to rely on a positional cross-match (COSIM cannot be used on such a large number of objects), BUT with conditions.

The main condition is that the coordinates accuracy in SIMBAD should be better than 1" ; all other sources are ignored. For sources whose coordinates were not from Gaia DR2, we applied a radius avoidance of 3", corresponding to the resolution of 2MASS : there should not be any neighbour within 3" in SIMBAD nor in Gaia DR3. This criteria was however removed if the closest Gaia DR3 source was the brightest and if the difference between both G magnitudes was larger than 2.5 (10% of the flux). Then the cross-identification was accepted if the distance between the SIMBAD and Gaia DR3 positions was smaller than 1" ; note that Gaia DR3 coordinates were calculated at epoch 2000.0 to be compared with the content of SIMBAD. For sources that already had a Gaia DR2 position in SIMBAD, a cross-identification was accepted if the distance between Gaia DR2 and DR3 positions was smaller than 0.15", and if there was no neighbour in Gaia DR3 within 1" as some Gaia DR2 sources turned out to be resolved into two objects in Gaia DR3. As the epoch of Gaia DR2 is 2015.5 while the epoch of Gaia DR3 is 2016.0, the previous criteria mean that all stars with a total proper motion larger than 300 mas/yr were missed. They have been ingested separately, so that all Gaia DR3 stars with PM ≥ 300 mas/yr are in SIMBAD.

Appendix 3. CDS HiPS node - response to peak usage

We experienced a significant change in the HiPS usage as the World Wide Telescope application (WWT) now supporting HiPS as of February 2022 and accesses the CDS HiPS. This resulted in a significant usage of our service in July at the time when JWST first images were released.

The WWT page for the JWST first images was using the CDS unWISE HiPS as the default background, resulting in a significant increase of the number of requests (see Figure A3.1 for the number of tiles per seconds at this period). Given the number of requests and the machine capacity, we had to change the web server configuration, which is seen in the ~ 1 hour outage.

This new configuration enabled us to deal with as many as 1000 queries per second. This peak of usage linked to access to JWST first images is one of the highlights of the Aladin project for this year. This is also a tracer of the underlying evolution we have seen in the last decades that Aladin is now serving users way beyond the astronomical community.

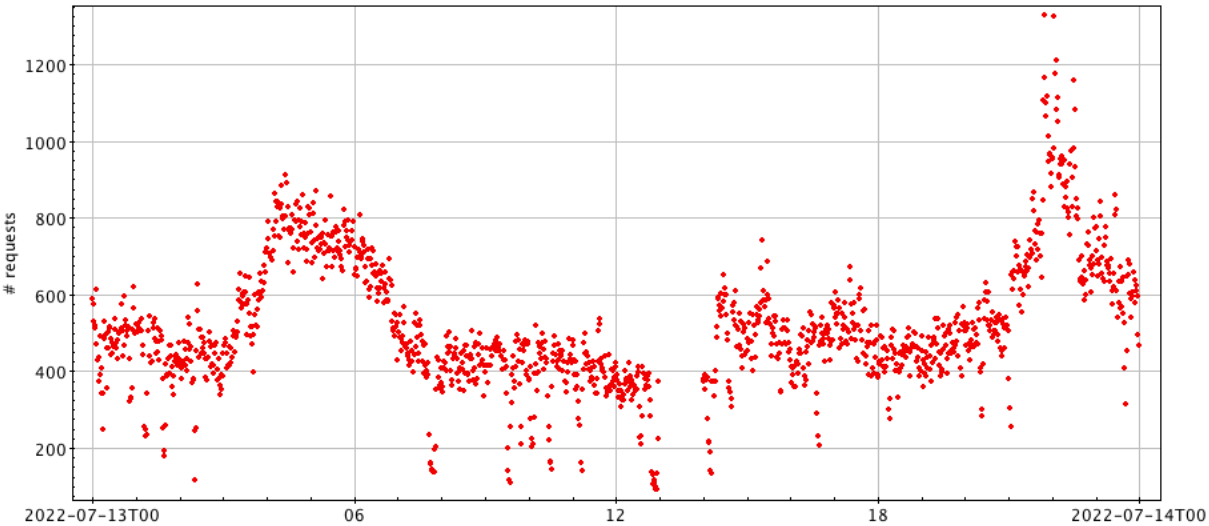

Figure A3.1: Number of HiPS tiles requested on July 13th 2022