CDS Scientific Council meeting 2015
Summary of CDS activities 2014-2015

7 October 2015

TABLE OF CONTENT

HIGHLIGHTS DURING THE PERIOD

Significant progress continues to be made in the different aspects of the CDS activities, some visible from the users, others pertaining to our internal procedures or linked to our collaborations.

One major point since the last regular meeting of the Council was of course the change of director, in which the Council was closely involved. Mark Allen, the fifth director of the CDS, succeeded Françoise Genova. The transition was prepared for several months and was effective on 1 September 2015.

This report has been prepared in collaboration by Mark Allen and Françoise Genova. It discusses the status of the current strategic axes. Mark Allen is preparing a detailed work plan for period from the end of 2015 through 2016, which is being discussed with the team. An updated strategy for the next four years will be presented at the 2016 CDS Council meeting. The report thus focusses on the CDS activities during the last year. Continuity will of course be ensured but the plan is still being discussed.

Other "structuring" highlights:

- Another major evolution of CDS is the progressive departure of several key CDS staff, who had been present since the early 70s and were at the origin of the pioneering role played by the CDS for the astronomical community and beyond. Marc Wenger, the engineer in charge of SIMBAD, retired in September 2014, and François Ochsenbein, the astronomer in charge of VizieR, has been on emeritus status for two years. The new leadership of SIMBAD and VizieR are in place and ensures full responsibility of the services.
- The CDS was proposed to be maintained as a "Research Infrastructure" on the National Research Infrastructure Roadmap established by the Ministry of National Education, Higher Education and Research (MENESR), which is presently being updated. The proposed list is currently being examined by the relevant

Committees and the final decision will be known before the end of the year. More importance has been given to aspects linked to data in the Roadmap update and the CDS is perfectly in line with this priority.

- The last Euro-VO project coordinated by the CDS, CoSADIE, ended in February 2015. A paper describing how Euro-VO coordinated the VO activities in Europe was published in Astronomy and Computing[1]. A new phase of the European Virtual Observatory began in May 2015, with the start of the ASTERICS cluster (Astronomy ESFRI and Research Infrastructure Cluster), a 15 M€ project proposed by a consortium led by Mike Garrett (ASTRON) to a European Commission Call on 2 September 2014. CDS leads WP4, *Data Access, Discovery and Interoperability*, which gathers VO teams and teams from the large ESFRI and ESFRI-like projects to optimize the usage of the data through the Virtual Observatory. WP4 is in charge of about one third of the budget (4.5 M€ over four years). More details about ASTERICS WP4 and CDS role in it will be given below.
- The software engineer position opened by the University of Strasbourg for the replacement of Marc Wenger was recently fulfilled with the recruitment of François-Xavier Pineau. CNRS opened a competition for a "documentaliste" position, which is on-going.
- CDS has been playing a leadership role in the development, implementation and dissemination of the so-called "hierarchical progressive surveys" (HiPS) based on the HEALPix sky tessellation. This hierarchical approach to Big Data in astronomy is fully scalable and easy to implement. A paper[2] published in Astronomy and Astrophysics summarizes the approach. Tools are provided to allow data producers to build their own HiPS data, and a network of "HiPS nodes" is being set up, involving major partners including ESA and other data providers. A strategy is defined to fully include these data in the Virtual Observatory through a "light" formalisation.
- The new ADS interface now allows us to count the number of papers in which the services are cited, not only in the acknowledgement section but in the full text. **In 2014, 679 refereed papers cited the word SIMBAD, 306 the word VizieR, and 58 the word Aladin** (in reference to our service). The CDS is fully recognized by the national authorities as a research infrastructure, but the French astronomical community has difficulties to understand that the CDS services are "scientific productions". We hope that these figures, which are closer to the usual criteria used to measure scientific activity than the usage statistics, will help. This is particularly important for the careers of CDS scientists.

CDS Services

- Several new important features were implemented in SIMBAD, in particular to improve the integration in the VO (the web pages are SAMP compliant), and as expected to improve the links with VizieR with the inclusion of the "Photometry viewer" in the SIMBAD web pages. We continue to improve the user experience, and in particular a new functionality, SimWatch, allows users to be informed about new papers on their favourite objects.
- A new VizieR home page was released at the beginning of 2015. Significant developments to improve VizieR pipeline for non-tabular data attached to papers were pursued, based on the ObsCore VO standard[3]. The database is currently being built. As a test it currently contains the CoRoT data (150 000 time series).
- The CDS Cross-match Service has reached a stable operational phase following several years of development.

[1] Genova, F., Allen, M.G., Arviset, C., Lawrence, A. Pasian, F. Solano, E., and Wambsganss, J: Euro-VO coordination of virtual observatory activities in Europe, Astronomy and Computing, v. 11, p. 155-160.

[2] Fernique, P., Allen, M.G., Boch, T., Oberto, A., Pineau, F.-X., Durand, D., Bot, C., Cambrésy, L., Derriere, S., Genova, F., and Bonnarel, F : Hierarchical progressive surveys. Multi-resolution HEALPix data structures for astronomical images, catalogues, and 3-dimensional data cubes, Astronomy and Astrophysics, v. 578, A114.

[3] http://www.ivoa.net/documents/ObsCore/20111028/

- The interface of VizieR with TOPCat, one of the VO success stories, was improved by implementing a direct link with the CDS cross-match service. This is a significant optimization of the number of queries and of the data flow which circulates on the network, since cross-identification results are provided instead of the two full catalogues to be cross-identified. This brings the computation near the data, which is one of the current Big Data trends. As announced last year, this optimization has the consequence of decreasing the number of queries in VizieR, and increases significantly the usage of the cross-match service.
- New Aladin web pages were released in November 2014. The fast development of HiPS and of Aladin Lite, which is more and more implemented in the web pages of services external to CDS, continues to structure the activity.
- The database contents of all the CDS databases continue to grow fast, as does the workload on the team, which was already under stress before. This is true for SIMBAD and VizieR, including ingestion of very large catalogues, but also for Aladin image database in which more and more image surveys are implemented.

Projects, Collaborations and R&D

- Our participation in the Arches[4] project produced a leading edge cross-match software, allowing computation of mathc probabilities for an arbitrary number of catalogues. The tool is being tested with the different scientific applications included in the project, and it will be released in the CDS cross-match service. The work was successfully presented in an oral paper at the 2014 ADASS meeting, and the corresponding paper in the Proceedings[5] can be used as a reference for the moment. A paper describing the method will be submitted to *Astronomy & Astrophysics* in the coming months.
- The CDS continued to be very active in the IVOA, in particular for the definition of the set of Data Access Layer standards which are being developed to deal better with multi-dimensional data, an IVOA priority well in line with the work programme of ASTERICS.
- Among the highlights from the recent R&D internships: assessment of Big Data technologies, namely Hadoop/Spark, with preparation for a detailed study of these technologies for cross-match and multi-dimensional visualisation applications.

CDS STRATEGY: STATUS

The current strategic plan of the CDS was defined in 2011, on a schedule aligned with the cycle of evaluations of the Observatoire de Strasbourg by the national evaluation agency, which occurs now every five years. We are near the end of this reference period. Also as explained the new director is defining the CDS work plan for the end of 2015 and 2016, and will submit a new strategic plan to the Council in 2016. It is thus useful to present a short report on the status of the 2011 strategic axes, which is given below. Significant progress can be reported for all of them. Information presented during the previous years will be reproduced when relevant to give the full picture.

The high level strategy drivers presented at the 2011 Scientific Council meeting were:

- Maintain the services at the highest possible level in terms of content and functionalities;
- Add functions to the core services in line with our expertise, the users' needs and R&D results;
- Take into account the change in scale of CDS activities due to the increase of publication volume and to the advent of many very large surveys.

[4] http://www.arches-fp7.eu/

[5] Pineau, F.-X., Boch, T., and Derriere, S : Towards next-gen catalogue cross-match service.

The strategic axes identified in 2011 were of different types. (i) Those linked to the *evolution of astronomy* were to accompany the very large survey era; to put our expertise at the service of Gaia usage by the community; the construction of Spectral Energy Distributions; data cubes and polarimetry. (ii) The main driver for *technological evolution* was identified to be the new Web 2.0/3.0 paradigm. (iii) For the *VO aspects*, CDS strategy was VO implementation in the CDS services seen as a priority because they are major building blocks of the VO, continuing to update the VO framework and to disseminate the VO knowledge in the astronomical community, and looking for a framework to pursue outreach towards education. (iv) The possible new role of CDS in the fast evolving landscape of *scientific data curation* was to be assessed.

Status of CDS strategic axes 2011-2015/2016

1) *Accompanying the evolution of astronomy*

- The large survey era

CDS reputation is now very well established for the distribution of catalogue data. A collaboration is set up with ESO for the distribution of their large survey catalogues. Long term collaboration has been on-going with ESA, for instance the recent distribution of Planck catalogues in VizieR, and collaborative work for the provision of Herschel catalogues. CDS is an official member of Gaia project and will distribute the mission catalogues. Preliminary discussions were held with the Pan-STARRS project and with the French LSST teams.

The CDS expertise in HEALPix sky tessellation is a strategic element for tackling large surveys, both for catalogues and for images. Both Aladin and VizieR ingestion pipelines were deeply updated to facilitate survey ingestion. HEALPix and HiPS are better and better recognized, leading to a very significant increase in the number of sky surveys stored in Aladin (currently 236 amounting to 50 TB of data). The new capacity to link images to their progenitors is also a good incentive, since it builds a powerful link to access original images from the original archive or via their metadata. Access is also given to data cubes. Ingestion of very large surveys in VizieR is also now much better streamlined in VizieR, although a specific work is still needed for a proper description. CDS also shares this expertise by providing tools to create HEALPix data sets, by its contribution to the definition of relevant IVOA standards and with the leadership of the IVOA Application Working Group.

In addition to distribution of large survey data, properly described for reuse, CDS is also providing specific added value by making the data interoperable and useable in tools. E.g. one important element of the CDS large survey landscape is the CDS cross-match service, which is currently the best on the market for cross-identification of very large catalogues. As explained, it is now interfaced directly with TOPCat, giving access to all TOPCat functionalities for managing tabular data.

In addition, the strategic axis on large surveys in fact meets the fourth one, which addresses the Data Centre role in the evolving landscape of scientific data curation – in the context of *Open Data – Open Science*. For the CDS, "Big Data" such as the very large surveys, and smaller, "Long Tail" data, should be seen by the users on equal grounds. Long tail data in astronomy are mostly the results of research by teams and individuals. They are very diverse and heterogeneous, which is one of the characteristics of Big Data, and they are reused when available, which is more and more the case for data attached to publication, thanks to the journals and to VizieR. The current evolution of VizieR, to deal better with the different kinds of data attached to publications as well as with large survey catalogues, will also empower a more efficient usage of large surveys.

- <u>CDS participation in the Gaia project</u>

As explained, CDS has been an official member of the so-called "CU9" since the constitution of the consortium. CDS will distribute Gaia result catalogues. The first release GDR1 is foreseen for July 2016. CDS will receive the data in advance to release it at the same time as ESA. An end-to-end test of the ingestion procedure will be held in December 2015.

- <u>Construction of Spectral Energy Distributions</u>

CDS "Photometric viewer", which browses all VizieR tables for photometric data and reconstructs a curve with the photometric points, was released in 2013 as a widget in VizieR. It is also now implemented in Aladin and in SIMBAD.

- <u>Data cubes and polarimetry</u>

Aladin is now able to visualize data cubes and polarization maps. Dealing properly with data cubes is one of the priorities of the IVOA to tackle the data from current and future large projects – and of ASTERICS WP4 for the same reason.

Another important contribution of CDS is its leading role in the definition of IVOA standards, in particular on the suite of Data Access Layer (DAL) standards which are being developed to deal with multi-dimensional data. CDS staff contribute as standard authors, and also as IVOA DAL Working Group lead, successively vice-chair then chair.

2) Technological evolution

In 2011, the main driver for technological evolution was identified as the Web 2.0/3.0 paradigm.

The Web 2.0 approach is centered on the user, who is considered as a co-developer. The implementation of the possibility for users to post comments attached to objects in SIMBAD or catalogues in VizieR is a typical Web 2.0 functionality. The development of widgets allowing one to include specific "windows" on CDS data in any web page is a first step towards the implementation of modular interfaces. Aladin Lite and the VizieR "Photometric Viewer" are indeed included in several services, including for Aladin Lite more and more external services. The possibility for users to open a user account to store for instance cross-identification results as well as to identify him or herself for posting comments is a first step towards a MyCDS personal functionality.

The web 3.0 approach combines the semantic web, mobility and universality (independence with respect to access methods and exploitation systems). The work done on new user interfaces (mobile interfaces, service mashup through the CDS portal) can be seen to belong to that domain. Several projects proposed to the Agence Nationale de la Recherche Calls to support in particular advanced usage of the semantic technologies were not successful, and the reports we got demonstrated that this type of project is not understood at all by the reviewers. Priority was given to the exploration of new interfaces, in view of the growing success of multitouch screens. Virtual Reality techniques have been a topic of R&D for CDS for a number of years.

The update of the CDS Portal is under way with a study of modular components.

The domain relies on technologies in constant evolution, and it is mandatory to evaluate possibly interesting technologies when they emerge. The R&D strategy is thus constantly evolving since technology evolves very quickly in these domains. Many R&D studies were performed since 2011 in other technological domains than web 2.0/3.0. Priorities are evolving in particular with the emergence of Science 2.0/Open Data/Big Data concepts.

CDS put special care into evaluating the relevance of the so-called "Big Data" technologies for our needs. For the moment we continue to use SQL-like methods, and as explained we make full usage of the "Big Data" capacities of HiPS, which is fully scalable. But we make sure to follow the technological evolutions around Big Data to seize opportunities.
During the last years for instance, the following Big Data technologies to browse, explore and visualize large tabular datasets were explored:

- Evaluation of the Apache Solr platform (2013)
- Web applications allowing one to create a SPLOM (Scatter Plot Matrix) from any VOTable (2014). The different views of the matrix are linked allowing for easy exploration of the parameter space.
- For larger datasets, we developed our own version of the Nanocubes[6] data structure, allowing fast interactive visualization of a catalogue with hundred million rows for a few attributes (2014). HiPS heatmaps are dynamically created server-side according to user-chosen criteria and visualized in Aladin Lite.
- Evaluation of the usage of Hadoop/Spark, for a detailed study with the cross-match application (2015).

3) Participation in the VO

For the participation of CDS in the VO, the priority was to implement the CDS services in the VO. This has been done, in particular with the release of a TAP interface for VizieR and SIMBAD, which allows users to build complex queries, and the "SAMPification" of VizieR and SIMBAD web pages, allowing users to send query results to VO services such as TOPCat.

One can also note the close collaboration with ESA for the implementation of their new Multi-Mission User Interface, using several elements of the CDS services and tools, and with several other agencies and laboratories for the implementation of Aladin Lite in their web pages.

The participation of the CDS in the development of the VO framework of standards and tools continues to have a high impact. The list of standards with at least one CDS author, the talks presented by CDS members during the last Interoperability meetings, and Working Group, Interest Group and standing Committee leadership, are detailed in the companion document "CDS participation in IVOA". Since the beginning of the VO, 53 standards have been recommended by the IVOA, including the updates. 24 of these have at least one author from CDS and 14 have one CDS editor. CDS participated in standards produced by all IVOA Working Groups except the Registry and Time Domain ones. CDS staff also led or are leading the Applications, Data Access Layer, Data Model, Grid and Web Services, Semantics and VOTable (now dormant) Working Groups, the Data Curation and Preservation Interest Group, and the Science Priorities and Standard and Processes Standing Committees.

4) CDS role in the fast evolving landscape of scientific data curation

The landscape of scientific data sharing is evolving very fast, and astronomy and the CDS remain at the forefront. *Open Data – Open Science* are currently buzz words and even the G8 Ministries of Research made a very strong statement[7] on a set of principles for open scientific research data on 12 June 2013.

In this fast evolving context, CDS already plays a major role for astronomy, in particular through the collection of data attached to publications in VizieR. In addition to preserving

[6] http://www.nanocubes.net/

[7] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf (Section 3)

"trustable" research results validated by a publication and properly described, it allows discovery and usage, in a framework, VizieR, which also contains very large catalogues, and is also interfaced with the Virtual Observatory which allows combined usage with all the other VO resources.

The first agreement between CDS and a journal was established in 1993 with *Astronomy and Astrophysics*. Data attached to all major astronomical journals are now available in VizieR. It is important to note that even before the real beginning of the "Open Data" rush about 15% of VizieR "catalogues" contained non-tabular data (spectra, time series, images, data cubes).

Significant work was devoted during the last years to improve the management on non-tabular data. It was decided, after a discussion with the AAS in January 2014, to use the IVOA ObsCore standard as a basis for dataset metadata. A database system based on SAADA[8], developed by Laurent Michel from Strasbourg Observatory High Energy Astrophysics team, was built. The first data set, the CoRoT collection of time series, is being ingested. A user interface for loading data sets and relevant metadata (thus keeping the data producer's knowledge about the data) is also being implemented. The point here is not to say that CDS will take care of all data attached to astronomical publications, but to build a technical framework allowing us to optimize the management, discovery and usage of these data. This is taken into account in particular in the framework of our relationship with *Astronomy and Astrophysics*.

CDS is aware of the data curation landscape. The starting points are of course the activities directly devoted to build the data content and to curate and disseminate data at best. But other elements are prominent in this fast evolving landscape.

"Trust" is a key element in the Open data landscape. One key element of Data Management Plans, which are more and more required by funding agencies in applications, is that the data resulting from the project will be deposited in a "trusted repository". The astronomy community trusts CDS, as demonstrated by the huge usage of our services, but in the current landscape an external evaluation is more and more mandatory. This is why CDS decided to submit applications to be reviewed by respected international bodies. This resulted successively in getting the "World Data System" (WDS) label in 2011 and then the "Data Seal of Approval" (DSA) in August 2014. The DSA label was pursued in addition to the WDS one because DSA is recognized as the first level label in the European certification landscape. It is also fully international. CDS was the first natural science repository to obtain the DSA, and only the second French data service to get it, the first one being the CINES, which has a national mandate to preserve data from Universities and Research. The certification process was managed by including the whole VizieR team, and allowed us to check our processes in details. It is interesting to note that we did not have to change the processes, just to describe them in detail.

Preservation and certification are very hot topics for data providers currently. CDS staff members are often invited to share their expertise in front of very different audiences. Françoise Genova is now an active member of the DSA/WDS Certification Working Group of the Research Data Alliance, which is aligning the DSA and WDS sets of criteria to facilitate adoption. She will be a member of the DSA Board starting 1 January 2016.

The next hot topic with respect to the general data curation landscape is the definition and implementation of Digital Object Identifiers for CDS data, and more generally for astronomy. This will be discussed during the Data Curation and Preservation session at the next IVOA meeting. We have persistent identifiers already, the point is to include them in general directories. There may be scalability problems, which will have to be assessed, for the data from space and ground based telescopes. VizieR data sets may be

[8] http://saada.unistra.fr/saada/

more straightforward, one point being to preserve the relationship between data in VizieR and the publication to which it is linked.

STATUS OF SC 2014 RECOMMENDATIONS

This year there was no high level recommendations by the Council directly relevant to this activity report. Two points can be noted however.

It is a real relief that the two positions discussed in the Council reports in 2013 and 2014, the University one for the replacement of Marc Wenger, and the CNRS one following the retirement of Gratiane Chassagnard in Paris, were opened and are or will be fulfilled. As always the staff situation remains fragile, and the backup in software development expertise which had been ensured for SIMBAD and VizieR is not available any more with the retirements of Marc Wenger and (soon) of François Ochsenbein, but at least the immediate threats have been lifted, thanks to the CNRS and the University.

For information, 3 PhDs and 1 post-doctoral contract have been on-going at CDS during the reference period:

PhDs

François Nehlig (Oct. 2012 – Sept. 2015): Galaxy evolution in the Virgo cluster (with Bernd Vollmer)
Maxime Beuret (Oct. 2013 – Sept. 2016): Stellar formation, interactions with the ISM (with Laurent Cambrésy)
Jérémy Chastenet (Oct. 2014 – Sept. 2016): Bayesian analysis of dust grain properties variations in resolved nearby galaxies (with Caroline Bot and Karl Gordon, STScI)

Post-doc

David Eden (July 2014 – July 2015)

Heddy Harab (Jan. 2016 -)

3D extinction in the Galaxy (with Laurent Cambrésy, on the Via Lactea project)

SUMMARY OF ACTIVITIES, SEPTEMBER 2014 – SEPTEMBER 2015

Complementary information can also be found in the previous sections of this report, and in the companion documents.

The services

SIMBAD

October 4th, 2015, SIMBAD contained 7,998,221 objects, 22,322,732 identifiers, 308,588 bibliographic references, and 12,126,329 citations (7,556,225 objects, 18,563,653 identifiers, 294,449 bibliographic references, and 10,749,766 citations in papers on 1st September 2014 and 7 342 000 objects, 18 162 000 identifiers, 285 000 bibliographic references, 10 000 000 citations of objects in papers on October 1st, 2013).

The mean number of queries on SIMBAD each day during the last year have been around 500 000, about 10% for internal use at CDS, 54% for web requests and 42% for Sesame (queries from external services).

SIMBAD content: cross-match tool

The COSIM tool, which replaces the historical software ("raccord") which was used to cross-match lists of objects with SIMBAD taking into account as many parameters as possible to prepare object ingestion, is now fully operational and used by the documentalistes in their daily work, after a period of intensive tests and updates from usage feedback. With COSIM, the documentaliste decides, according to the situation, which offset is acceptable for coordinates, velocity, magnitudes etc., and which object types are compatible. Rapidly and with a very low risk of errors, COSIM sorts objects into good and bad candidates, while bringing to light borderline candidates and possible merges among existing objects in SIMBAD, thus allowing the documentalistes to concentrate their attention on "difficult" individual cases which require all their expertise.

SIMBAD content: specific operations to improve the quality of the content

The daily update of the database of course goes on. Here only specific, significant programmes are listed. Some of these programmes are long term ones and require manpower, but they are pursued because they improve significantly the database quality and scientific relevance.

Objects with a position already better than 1''

A new cross-match of the SIMBAD content with the 2MASS catalogue has been performed. It has used part of the software of the Xmatch service, plus very strict criteria to avoid false cross-identifications. More than 1.5 millions 2MASS identifiers have been added, with an improvement of the astrometry in most cases. Next year a similar operation will be done with the UCAC4 and UCAC2 catalogues.

Objects without coordinates

There were almost 200,000 objects without coordinates in SIMBAD. This is part of the historical burden we still have to deal with. Note that for more than 5 years it is not allowed anymore to create an object in SIMBAD without position. About 125,000 of these lost objects have been investigated.

Those with relevant scientific information have been properly identified (gamma ray-burst, precise objet types, more than 5 references, radial velocity measurements, components of double stars, etc...). Those without any relevant information have been deleted (mostly uncharacterized stars in globular clusters).

Objects with inaccurate coordinates

In general, there is a continuous effort to improve the identification and the astrometry of historical objects in SIMBAD. We often get some input from users. Last year we especially worked on the oldest high proper motion stars (Wolf), emission-line stars, Herbig Ae/Be stars, white dwarfs, and Zwicky galaxies. A special effort is also ongoing to clean double stars in SIMBAD (32,000 components out of 46,000 have been cleaned), and to get back accurate spectral and morphological types with their reference.

Revision of the list of object types in SIMBAD

The list of SIMBAD object types has been for long available from the main SIMBAD web page[9]. It is used as documentation for both the users and the CDS team. It is a hierarchical list, in fact the algorithm itself used in the update and the COSIM software.

[9] http://simbad.u-strasbg.fr/simbad/sim-display?data=otypes

As explained last year, this linear hierarchical list can no longer take into account the complexity of relations between object types, and is not user-friendly as a documentation. A new reorganized list of SIMBAD object types was built with an astronomical point of view. It is much more user-friendly, both for the users and the CDS team. Short definitions are already available for quite a few types. This new list is not available yet to the users, but it is available for the CDS team in the internal wiki. A new system of priority and compatibility relations between the object types has been built based on the astronomical expertise of the whole team. Full discussion of the list and relations and implementation has taken more time than foreseen last year, but it is expected that the list will be operational and implemented in COSIM in the next months.

New functionalities

The following functionalities were successively released during the last period:

- VizieR Photometry Viewer was included in SIMBAD pages in November 2014. This is an important milestone to improve the links between the two services, and also as a demonstration of the usage of widgets to create modular interfaces.
- Many SIMBAD objects have many papers attached, which makes the search for the most relevant papers somehow difficult. The possibility to sort references using the location of the object citations in the paper had been implemented in 2013. A new web page in SIMBAD released in February 2015 expands this functionality: it gives a score to each paper attached to the object, using multiple conditions (locations where the object is cited, as before, but also citation counts, the number of objects in the paper, etc).
- SIMBAD interface with the VO was improved: the web pages were SAMPified, which allows users to send the result of a SIMBAD query to other VO-enabled applications like TOPCat and Aladin (May 2015); an improved TAP integration including new features is in beta-test.
- Direct links to the reference are now available for object types (June 2015).

The SimWatch service described in the highlights, which allows users to track new papers citing their favourite objects, is ready for release.

Other software/hardware evolutions

- Beginning of MOC compliance (i.e. implementation of HEALPix sky tessellation): the point is to query SIMBAD from a MOC and to export results in MOCs
- Improvement of the exchanges between ADS and SIMBAD

Future evolutions in preparation

- The visualisation of hierarchy links
- SIMBAD database will be replicated to face overload. At present a limitation of burst queries was enforced.

Marc Wenger's retirement

Of course, Marc Wenger's retirement after 41 years of faithful service at CDS, which was effective before the end of September 2014, has been a major milestone in SIMBAD's and CDS's lifes. The transition had been very well prepared and it went smoothly.

VizieR

October 4st, 2015, VizieR contained 14 065 catalogues (12 691 in September 2014, 11 579 catalogues in October 2013, 10 360 in September 2012). Four very large catalogues were included since the last council meeting, URAT1, VVV, IPHAS2 and

VIKING, plus a catalogue from The Astrophysical Journal with 116 million records which was ingested through the large catalogue pipeline. Several of these catalogues are ESO Public Surveys, and were made available at CDS in the framework of the ESO/CDS collaboration.

There have been 300 000 queries/day during the last year, which is as announced to the Scientific Council last year significantly lower than the previous year, which was an average of 530 000 queries/day. This is the beneficial consequence of the decision to interface queries directly with the CDS cross-match service when relevant. Should the positions sent to the cross-match in 2015 be submitted through multiple cone searches (which was the old way of doing it, especially from Topcat/STILTS), it would have represented a major load on VizieR with 3 million requests per day!

There has been no specific work during the year on VizieR contents, but the notes taken during the weekly meetings show that in fact each catalogue is somehow specific.

A new mirror was set up in South Africa at the SAAO/NRF in April 2014, the first one in Africa and in the southern hemisphere.

"Additional data" in VizieR

This important evolution aims at making non-tabular data attached to papers fully discoverable in the VO though searches for images or spectra in a given direction. The data is currently only discoverable through the VizieR information (which appears as information from a catalogue) retrieved by a cone search of the VO. It progressed significantly during the last year. The activities and current status are described above. A full database is in beta-test and the first content has been uploaded. A prototype user interface has been developed. Tools are being developed to help CDS documentalistes to populate the documentation. The next phase will be to create the final databases including the local and production ones, to populate them, and to integrate the processes into the pipeline.

The assessment made during the study shows that when populating the database many issues with metadata curation will have to be dealt with. The system will gather as much metadata as possible using the FITS headers, but it is often not sufficient and sometimes wrong information is provided. The new user interface developed for authors to submit their data to CDS aims at improving the situation for new datasets, since it requires the authors to provide the ObsCore elements relevant to their data. Although authors can be careless in those matters, they indeed have the best knowledge of how the data was created, and the objective is to capture this knowledge. Attaching exact metadata as complete as possible is the key quality challenge of this endeavour.

Data Seal of Approval

Certification is a hot topic for data repositories, and the work done to obtain the Data Seal of Approval has been presented in many different contexts. The way the VizieR technical lead proceeded, by involving the whole VizieR team for a deep review of all the procedures, is one of the important messages we pass during the presentations : getting certification is not an easy task, but aside from the importance of being certified by an external authority it can also have very positive consequences internally if the procedure is managed well.

Discussions with the AAS

Discussions with the AAS to improve the pipelines for tables attached to papers published in the AAS journals.

Other software/hardware evolutions

The release of a new Vizier front page already cited in the highlights, gives a better visibility to documentation and to the companion services (TAPVizieR, SED viewer, etc). In addition one can note the following evolutions:

- Migration of the VizieR database server
- Improvement of VizieR availability through software and technical measures
- Improvement of compatibility with the VO following discussions at the IVOA and the findings of IVOA validators
- Replacement of the VizieR machine and extension of disk space. VizieR data is now under the homogenized back-up system.

The UK mirror is on line again. Recovery actions are on-going for the CADC mirror, which has been off for one year after a problem on the Canadian side.

One important technical evolution in the near future will be the complete abandon of the historical commercial Sybase database system, still used locally for table ingestion, for PostgreSQL.

Aladin

Aladin usage (hosts) is clearly moving from Aladin Desktop (33%) to Aladin Lite (66%). Usage of the Aladin database is now 25% for Aladin Lite and 75% for Aladin Desktop. Usage statistics are tricky because a fraction of usage moved from Aladin Java to Aladin Lite, and the HEALPix (HiPS) database is more and more used whereas the comparison with the "traditional" usage is not easy. Also, users can use Aladin without accessing our database and we cannot track this usage.

Nevertheless the usage is continuing to increase: usage (queries) increased by +25% with respect to 2014, and hosts increased by 34% thanks to Aladin Lite. Aladin Lite is fully integrated in SIMBAD and VizieR, and it is more and more used in institutional archives (ESAC, Jaxa, ESO for the Outreach images, etc). The replacement of Java applet technology (used in Aladin Desktop) is a reality.

In September 2015, Aladin contains 236 surveys for a volume of 50 TB (175 surveys for a volume of 45 TB in August 2014, 128 surveys with a volume of 30 TB in August 2013, 81 surveys for 19 TB in August 2012).

As explained, Aladin evolution is really shaped by HiPS. More surveys have been ingested. The tool provided to create HiPS files (HiPSGen) was made more efficient (1 day computation to create 0.5 TB pixels) through a significant effort in software development. There are also more different types of HiPS surveys, including data cubes and catalogues.

Many international collaborations are being built around HiPS. There are currently 6 HiPS providers and more are expected. The creation of a "HiPS network" of HiPS providers at the national and international levels is on-going. The first steps towards a "light" inclusion of HiPS among the IVOA standards were set up at the Sesto Interoperability meeting in June 2015, and discussed further in Strasbourg in September 2015 during the first ASTERICS Technology Forum. The effort will be pursued during the Sydney Interoperability meeting (October 2015).

Utilisation of HiPS for outreach has also been explored through the "Arches walker" programme and a collaboration with the Strasbourg Planetarium using a planetarium fisheye projection.

New version of Aladin Desktop

The new version of Aladin Desktop will be released before the end of 2015 (the previous one, Aladin 8, was launched in March 2014). It will include a dynamic HiPS tree to indicate which HiPS tiles have data in the current field, homogenised HiPS metadata based on the ObsCore IVOA standard, an adaptive HiPS grid, better generation of MOC. Also included are contrast control, a different look of the full screen mode giving direct access to control widgets, and other new features, including a better integration of IVOA compliant files. The new features are described at http://aladin.unistra.fr/java/nph-aladin.pl?frame=downloading.

Other software/hardware evolutions

- Significant effort in software development for HiPS, including MOC server evolution to include dynamic HiPS tree and homogeneous metadata.
- New Aladin web pages were released in November 2014.
- The CDS HiPS server will be brought from 2x100TB to 2x200TB (including backups).

The cross-match service

As explained, the CDS Cross-Match service has reached a stable, operational phase with increasing usage.

674 jobs were submitted through the web interface per month in 2015 (vs. 492 in 2014 and 450 in 2013) by a total of 900 different IP addresses (slight increase in average when compared to 2014). These 674 jobs have produced 13.8 billion links (approximately 3 TB).

The main usage through the API (programmatic access) comes from Topcat and STILTS (more than 90% of all requests). Other tools accessing the service include Astropy and the VAO/MAST portal. As announced last year, access to the cross-match API has been integrated in the Astroquery package of the popular AstroPy[10] Python library. 46,800 queries have been submitted from 778 different IPs (in 2014: 18,000 requests from 406 IPs). One sees here the effect of the direct interfacing of TOPCat/STILTS with the service as a replacement of cone search on the individual tables in VizieR.

Overall, 1.3 billion positions have been submitted to the service in 2015. As explained, should these positions be submitted through multiple cone searches, it would represent a major load on VizieR with 3 million requests per day !

As explained in the CDS report to the Council last year,

> One collateral consequence is that the number of queries on SIMBAD and VizieR will likely decrease, eventually significantly. But this is a GOOD thing in terms of rationalizing the service usage. All the Big Data actors are aiming to move to "smart queries" to avoid overburdening their systems. Let's say that in collaboration with TopCat, and using the capacities offered by the VO, CDS is also implementing "smart queries" to serve its users better.

New VizieR tables keep being integrated at the same pace as they are ingested in VizieR: tables smaller than 10 million rows are automatically integrated, integration of larger catalogues is part of the large catalogues pipeline.

[10] http://www.astropy.org/

Other aspects of the cross-match API

Access to the cross-match API has been integrated in the Astroquery package of the popular AstroPy[11] Python library.

Cross-match in the Arches project

As explained in the previous year Council meetings, CDS is taking the opportunity of its participation in the Arches project to assess new, more advanced cross-identification methods. The project aims at providing a multi-catalogue cross-match framework providing probabilities of cross-identification. The challenge was two-fold: first, existing statistical methods had to be generalized to handle more than two catalogues, then a flexible, highly configurable and still efficient tool had to be developed. E.g., taking into account extended objects and proper-motions required to study and use new indexing data structures. As explained in the Highlights section, the new method is fully operational and is being tested on the scientific cases developed in Arches. A paper describing the method will be submitted to *Astronomy and Astrophysics* in the coming months. The statistical cross-match method will be included in the CDS cross-match service. One challenge is to implement an output which helps to user to use the results (for instance for 6 catalogues, the output file contains 869 probabilities).

R&D

R&D activities continue to be varied and versatile. A list of the internships overseen by CDS staff since the last Council meeting is provided among the SC documents. They cover a large range of aspects of interest for the CDS, such as usage of virtual reality for visualizing astronomical data and simulations (with possible applications to education and outreach) also inputs to the CDS strategic axes, such as the usage of big data technologies for very large surveys or how to deal better with data attached to publications, which is as explained CDS contribution to the dissemination of "long tail" research data.

A list of the R&D programmes performed through internships is provided in a separate document. Studies linked to attached data in VizieR and the ARCHES developments on statistical cross-matching were addressed above. One can note in particular among the R&D activities during the last period:

- The development of a MOC server, which has been in production since June 2015. It is used by the beta release of the next version of Aladin Desktop and by the VO VizieR Registry (7000 requests / day for the moment, a stress test was performed with 1.5 millions requests / day without any problem). Related paper in *Astronomy and Astrophysics*: "Efficient data structures for masks on 2D grids", Reinecke, M.; Hivon, E., http://adsabs.harvard.edu/abs/2015A%26A...580A.132R
- MOCPy, a Python library to handle MOCs
- 3D Visualization of various astronomical data in a Web browser (BoF and poster in ADASS Sydney)
- Experiments of the possible contributions of Information Extraction tools (especially GROBID, https://hal.inria.fr/inria-00493437) to the CDS
- Access in Aladin to IVOA SIAV2 compliant servers
- Development of the Graphic Charter of the CDS to improve the CDS graphic identity and to provide templates for communication materials.

[11] http://www.astropy.org/

Projects

Virtual Observatory, CoSADIE and ASTERICS

Implementation of CDS services in the VO

The implementation of CDS services in the VO, which has been identified as a priority, was already operational, but it has been updated, in particular for TAP queries, as explained for VizieR, to take into account feedback from validators and discussions at the IVOA meeting, and for SIMBAD. SIMBAD and VizieR pages are SAMPified, i.e. query results can be sent seamlessly to VO tools such as Aladin or TOPCat.

Update of the VO framework

The list of new IVOA standards developed with CDS participation is given in a companion document. CDS continued to play a leading role in the IVOA activities and structure. CDS staff currently chair two IVOA Working Groups, *Applications WG* and *Data Access WG*, one chairs the *Data Curation & Preservation* Interest Group, and an associate member of the CDS chairs of the *Semantics* Working Group. The chairs of the *Standing Committee on Science Priorities* and of the *Standing Committee on Standards & Processes* also belong to the CDS. CDS staff were authors of the *DataLink* standard that was accepted as an IVOA Recommendations in 2015.

One key activity in IVOA is currently to upgrade the IVOA framework to optimize the usage of multi-dimensional data, to optimize the usage of large projects, which is perfectly in line with the ASTERICS work programme. A "caravan" of data access layer standards is currently on its way to Recommendation, and the first element, DataLink, was endorsed as IVOA Recommendation in June 2015. The other elements of the caravan are AccessData, which is currently in the Working Draft phase, and the Simple Image Access Protocol V2.

As explained above HiPS will be a topic for discussion at the next Interoperability meeting. The point is here to get an endorsement for HEALPix, as one sky tessellation, to be taken into account by the IVOA, and to develop the elements of the IVOA framework necessary to make it effective, in particular to allow a "HiPS" registry entry for HiPS servers to fully integrate participants in the HiPS network in the IVOA.

Euro-VO activities

The CoSADIE project, the fourth EC-funded project to coordinate European VO activities, ended in February 2015 after a six month extension to the initial two-year duration. CoSADIE included three strands of work: "Increasing awareness and gathering requirements from the user and provider communities" (INTA and GAVO), "Coordinating technical activities and defining the technical needs to maintain the VO Framework" (UEDIN), "Outreach towards education and the general public interested in astronomy" (INAF), and assessed the strategies, governance and financial sustainability of the European Virtual Observatory. A detailed assessment of all the elements of Euro-VO sustainability was performed (Genova et al., 2015), including in particular a detailed assessment of technological sustainability (Allen et al., 2015). The Euro-VO partners, INSU/CDS, INAF, INTA, and the Universities of Edinburgh (UEDIN) and Heidelberg (UHEI), signed a MoU to indicate their willingness to continue to develop the Euro-VO together. As explained, a refereed paper summarizing the Euro-VO history and results, and the methods developed, was published. The project has been working in close collaboration with the Astronet ERA-NET, which gathers European funding agencies. Its sustainability assessment was fully endorsed by Astronet, and Euro-VO sustainability is one of the items of the work programme of the successor of the Astronet ERA-NET, which ended in July 2015.

Euro-VO is entering a new phase with the ASTERICS project, which started on 1 May 2015. The ASTERICS Work Package 4, *Data Access, Discovery and Interoperability*, gathers the European VO teams with the ESFRI and ESFRI-like projects, CTA, SKA, KM3Net, and the Einstein Telescope. It is important to note that the project gathers astronomy and astroparticle physics, in particular to tackle the aspects linked to new messengers, and that pathfinder projects are associated with the ESFRIs to be able to deal with real data (HESS, MAGIC and VERITAS for CTA, LOFAR and JIVE for SKA, ANTARES for KM3Net, and VIRGO/LIGO for the ET). ESO (VLT-E/ELT) is associated with the project, and ASTERICS WP4 continues to work closely with ESAC. The project tackles the three pillars of Euro-VO coordination identified by the suite of Euro-VO projects : support to data providers for their uptake of the VO framework (led by INAF and UHEI), support to the astronomical community (led by CDS and INTA), and technological work to update the Euro-VO framework of standards and tools following the requirements and feedback gathered by the two first activities (led by CNRS and UEDIN). The first ASTERICS WP4 event, the first ASTERICS WP4 Technology Forum, was organised by CDS in Strasbourg in September 2015[12]. One striking point was the high level of reuse of VO standards and tools. The next meetings will be the ESFRI Forum, a key milestone to gather the ESFRI requirements, and the first ASTERICS School, organised respectively by INAF and INTA in December 2015. Project milestones not organised by the project include of course the IVOA interoperability meetings, and also the plenary meetings of the Research Data Alliance.

ARCHES – Astronomical Resource Cross-matching for High Energy Studies

The project[13], initially scheduled for 32 months beginning in January 2013 and extended to the end of 2015, is coordinated by Christian Motch, from the High Energy Astrophysics team of Strasbourg Observatory. It focuses on the X-ray survey catalogue data from the XMM-Newton mission, namely the 3XMM release. As explained a state-of-the-art cross-identification tool was produced by the CDS. Used together with extensive archival resources, it produces well-characterised multi-wavelength data in the form of spectral energy distributions for large sets of objects. The CDS is also involved in the dissemination package, hence the R&D work on the ARCHES Walker. The resulting catalogues will be included in VizieR and the cross-match tool in the CDS Cross-Match service.

ASTRODEEP – Unveiling the power of the deepest images of the Universe

The project[14], funded by the European Commission for 4 years starting in January 2013, is coordinated by Adriano Fontana (INAF Roma). The project aims at getting the best scientific return from the exploitation of the deepest sky surveys available to date, studying the birth and early phase of galaxy evolution. The initial goals of the project were the development and testing of algorithms and software, data reduction and release, and scientific validation and analysis. A number of deep to very-deep sky surveys are used, including GOODS (South and North), CANDELS, HUDFs, covering a wide range in wavelength. A new algorithm named TPHOT has been developed for source extraction on these images, based on PSF-fitting with priors, and is being applied to multi-wavelength images in the GOODS fields to produce new reference source catalogues.

The role of the CDS is to develop a dedicated portal for the internal validation and manipulation of the data, but also to publish the final catalogues through VizieR, and images through Aladin. The portal is based on Aladin Lite for displaying and exploring images in the browser, and linking of heterogeneous data products (catalogues, individual thumbnails, spectra,...) is done using the Saada portal.

[12] https://www.astron.nl/asterics/doku.php?id=open:wp4:wp4techforum1
[13] http://www.arches-fp7.eu/
[14] http://www.oa-roma.inaf.it/astrodeep/

VIALACTEA - The Milky Way as a Star Formation Engine

The project[15], funded by the European Commission for 3 years starting in October 2013, is coordinated by Sergio Molinaro (INAF Roma). CDS participates in the scientific aspects of the project, which aims at exploiting the combination of all the new-generation Infrared to Radio surveys of the Galactic Plane from space missions and ground-based facilities. By using a novel data and science analysis paradigm based on 3D visual analytics and data mining framework, the project will build and deliver a quantitative 3D model of our Galaxy as a star formation engine that will be used as a template for external galaxies and study star formation across the cosmic time. The main task for CDS in the project is to deliver a 3D extinction cube of the Galactic plane. A first version has already been provided to the consortium this month (Oct. 2015).

The Research Data Alliance

In parallel, the evolution of European programmes is towards "e-Infrastructure Commons", a common framework for the European network, computing and data infrastructures. This supports CDS strategy to be involved in the Research Data Alliance[16] (RDA) at the European and international level. The RDA is a recent but high profile organisation which aims at defining building blocks of the data infrastructure facilitating data sharing.

The Research Data Alliance, created in March 2013 with support of the European Commission, NSF and Australia, continued to grow. F. Genova is one of the members of RDA Technical Advisory Board, and liaison is built with the IVOA. Astronomy has historically been at the forefront for the sharing and reuse of data, and thanks to the Virtual Observatory its data holdings are a rare example of a global interoperable data infrastructure. The lessons we learnt and our requirements have to be taken into account by the RDA. F. Genova was invited to participate in the series of European projects set up in support to the RDA. The third one began on 1 September 2015.

Lessons learnt when building the IVOA are shared with the RDA through F. Genova's participation in the IVOA Technical Advisory Board. Work performed in the IVOA and ASTERICS on Provenance, in particular using the complex CTA use case, was successfully presented at the last RDA Plenary in the frame of the RDA Provenance Interest Group. Also, the IVOA Registry of Resources is being entered in the B2FIND "generic" registry of the European EUDAT project, which is important to demonstrate that the astronomy well developed, global and interoperable data infrastructure is not an isolated island but can interoperate with the generic data framework.

[15] http://cordis.europa.eu/project/rcn/188856_en.html

[16] http://rd-alliance.org/