Data valorisation in astronomy

From information networking to Virtual Observatory

Françoise GENOVA[1]

[1]CDS

Observatoire astronomique de Strasbourg, UMR 7550

11 rue de l'Université, 67000 Strasbourg, France

genova@astro.u-strasbg.fr

Abstract - Astronomy is at the forefront for on-line data distribution, from observations to results published in journals, with also widely used value-added services such as those developed by the *Centre de Données astronomiques de Strasbourg* (CDS), NASA bibliographic database ADS or NED. Information networking is also particularly developed, and users can easily navigate from one service to another using Web links. This has been made possible by partnership between service providers from different horizons: a set of *de facto* disciplinary standards has been defined and is shared by all partners. Several standards have been defined in the domain of bibliographic information. The journals, ADS, CDS, and NED share for instance the bib/refcode, a 19-character description of a published reference. Recently Observatory archives have joined the partnership, by providing, for all the papers using their data sets, links to the original data kept in their archives. Beyond these strong foundations, the concept of Virtual Observatory has been emerging in the last years, aiming at building new tools, protocols and collaborations to realize the full scientific potential of the wealth of ground- and space-based observations, including the very large surveys, in the context of the development of GRID. Several R&T/Phase A projects have been accepted in 2001, the European *Astrophysical Virtual Observatory*, USA *National Virtual* Observatory, UK *AstroGrid*, and several other national projects are accepted or being prepared. Common metadata and exchange standards are the key to a really global Virtual Observatory, and an international collaboration is under way in this domain.

Résumé – La discipline astronomie a développé très rapidement des services en ligne, qui donnent accès à toutes sortes d'information, allant des observations dans les archives d'observatoires aux résultats publiés dans les journaux, avec également des services à forte valeur ajoutée, tels ceux développés par le *Centre de Données astronomiques de Strasbourg* (CDS), ADS, la base de données bibliographique de la NASA, ou la base extragalactique NED. L'information est mise en réseau, et les utilisateurs peuvent aisément 'naviguer' d'un service à l'autre en utilisant les liens hypertextes. Cela a été mis en place grâce au partenariat des différents acteurs, qui ont défini et qui partagent un ensemble de standards disciplinaires de fait. Plusieurs standards ont été définis dans le domaine de l'information bibliographique. Les journaux, ADS, le CDS et NED partagent par exemple le bibcode, une description en 19 caractères d'une référence publiée. Les archives d'observatoires ont récemment rejoint cette collaboration, et fournissent, pour tous les articles utilisant un ensemble de données, un lien avec les observations originales. Sur ces bases solides, s'est construit depuis quelques années le concept d'Observatoire Virtuel, pour mettre en œuvre de nouveaux outils, de nouveaux protocoles et de nouvelles collaborations afin d'optimiser le retour scientifique des nouvelles observations sol et spatiales, y compris les très grands relevés du ciel, dans le contexte du développement du GRID. Plusieurs projets de R&T/Phase A ont démarré en 2001, l'*Astrophysical Virtual Observatory* en Europe, le *National Virtual Observatory* aux Etats-Unis, *AstroGrid* en Grande Bretagne, et plusieurs autres projets nationaux sont acceptés ou en préparation. La mise en commun des metadonnées et des standards d'échange est la clé de la mise en œuvre d'un Observatoire Virtuel réellement global, et les projets collaborent sur ce thème au niveau international.

1. Introduction

Astronomy relies on long-term observations of variable phenomena and of a large number of very diverse objects. Conserving and reusing data is the key for major scientific objectives, such as the identification of new object types, the description of their properties, or the study of variability and evolution, all requiring statistical studies on large data sets. Observations at different wavelengths, with different techniques, are used to understand the physical phenomena at work in the objects, which means that heterogeneous data have to be used together. In addition, astronomers often perform their observations with large ground- and space-based observatories, and large teams also develop large surveys of the sky. Reusing the data from these large projects for new scientific objectives is also necessary to optimise their scientific return.

Conservation and distribution of data is an old endeavour for astronomers, which have been sharing catalogues of objects for thousands of years. It was learnt long ago that data must be properly documented to be reusable. At present the information volume and complexity is rapidly increasing, and information is heterogeneous and distributed.

Astronomy has rapidly taken the best advantage of the new technical possibilities offered by the WWW, in terms of information distribution, of integration of data with documentation, and of navigation between on-line services. The increased technical capacities for storing and managing large volumes of complex information are also widely used.

These new tools are useful and appealing, but a fundamental point remains careful work on the service contents and functionalities, and information validation. The success of astronomy information networking relies on dedicated data centres and specialized teams, which produce high quality services and work together to define disciplinary exchange standards. They have succeeded in networking astronomical information, from observations in observatory archives to results published in electronic journals. These teams, and newcomers often close to the IST community, now work together to go further and build the International Astronomical Virtual Observatory.

2. The astronomy bibliographic network

Astronomers are able to share data (images, spectra), because they share a common data description called FITS (*Flexible Image Transport System*). This standard was defined more than 30 years ago by radio-astronomers, who wished to use common tools to deal with their observations. The FITS rules are endorsed by the *International Astronomical Union*, and codified into a formal standard by the NASA/Science Office of Standards and Technology. Documentation can be retrieved from http://fits.gsfc.nasa.gov/ and from http://www.cv.nrao.edu/fits/. Thanks to the existence of FITS, observations from any instrument can be used by any astronomer, and many generic tools for data visualisation and transformation have been developed.

Aside from this widespread standard, other *de facto* standards have been developed, in particular to deal with bibliographic information. One example is the *bibcode*, developed by the *Centre de Données astronomiques de Strasbourg* (CDS) and *NASA Extragalactic Database* (NED) long before the advent of the www, because the two data centres exchanged information about bibliography. The *bibcode* is a 19-character, human readable description of a published reference: for instance, `1991A&A…246..24M` describes a Letter to the Editor published in the European journal *Astronomy and Astrophysics*, volume 246, page 24; the first author’s name is Motch, hence the M. The *bibcode* has been extended and is widely used by the NASA reference astronomy bibliography database, the *Astrophysics Data System* (ADS), and has been adopted by the astronomy journals, when they implemented an electronic version. In all astronomy electronic journals, each reference in the reference list of a paper links to the corresponding entry in the ADS. ADS links to the original on-line paper at the publisher site, when available, to a scanned version of the article, produced by the ADS, for older papers, to information from compilation databases such as SIMBAD and NED (list of astronomical objects cited in the papers, with active links to the database contents for these objects), to on-line data for the paper, in particular tables (as explained below). Recently, observatory archives have begun to use the *bibcode* to link, through the ADS, individual observations to the papers using them. ADS uses the information provided by the archives to implement a link between publications and the original observations. The *bibcode* is thus heavily used to build networking of bibliographic information, which is widely used by the astronomy scientific community. One advantage is that it is easy to build, it is however better adapted to publications in journals and more difficult to use for other publications (e.g. books, electronic publications with no page number,…).

The pre-existence of a disciplinary *de facto* standard, and the partnership between all service providers from different horizons, publishers, data centres, and now observatory archives, has thus permitted the rapid implementation of information networking. The correspondence with emerging general standards such as the DOI will for the moment be made through correspondence tables. Astronomy journals use DOI indexing but continue to provide *bibcodes*.

Tables published in article give another example of collaboration, and of the importance of *de facto* standards. The collaboration with the journal *Astronomy and Astrophysics* and CDS was settled as early as 1993, several years before the journal developed a fully electronic version. CDS implements on-line the abstracts and tables, in close cooperation with the scientific editor [1]. The standard description of tabular information proposed by CDS in 1994 has since then been accepted by other reference journals and data centres. The *ReadMe* is an ascii description of sets of tables. It contains in particular a `Byte-per-byte description`, which details the table structure in terms of format, units, column naming or labels, existence of data (possibility of unspecified or *null* values), and brief explanations. This description links the physical contents of the table to its scientific contents (e.g., in table 1, column 2, from bytes 12 to 13, with format I2, has unit h (hour), label Rah, and is the Right Ascension in equinox 1950, epoch 1983.5). It is now one of the important standards of astronomy, shared by journals and data centres, allowing for data exchange, format transformation and data checking – for tables published in journals for instance, it allows quality checks complementary to the one made by the referee. With this description, all information in tabular format can be accessed with generic tools: catalogues, published tables, and also the catalogues of the large surveys and the list of observations in archives. Tables published in journal become usable data. Examples of homogeneous access to heterogeneous information using this data description are given for instance in the CDS services: the VizieR catalogue Browser, which allows browsing across more than 3500 catalogues, published tables, surveys and lists of observation in archives; and the Aladin image tool, which permits overlay of tabular information on images of the sky. A first XML layout of astronomical tables, *astrores*, is described in [2]. VizieR table output in *astrores* format have readily been used by other services such as HEASARC Browse and IPAC OASIS image tool.

3. Towards the Virtual Observatory

Disciplinary data centres (IPAC, HEASARC, STScI/MAST) and ground-based and space observatories (CADC, CFHT, ESO, Chandra, ISO, XMM, …) are also very active partners in the networking of information and in the definition of standards. Beyond these strong foundations, the concept of Virtual Observatory has been emerging very rapidly in the last years, in particular with two conferences held in 2000: *Virtual Observatory of the Future* (CalTech, [2]) and *Mining the Sky* (Garching [3]). The Virtual Observatory is a science-driven project, aiming at realizing the full scientific potential of astronomical data, in the context of rapidly increasing volume and complexity of available information from observatory and surveys, and of the development of new techniques such as the GRID. Several R&T/Phase A projects have begun in 2001, the *Astrophysical Virtual Observatory*, funded by the European Commission (PI: P. Quinn, ESO), the *National Virtual Observatory*, funded by NSF, *AstroGrid*, funded by PPARC, and several other national projects have been proposed or accepted in 2002 (in Canada, India, Russia, Germany, Japan, …). These projects are collaborating together and an *International Virtual Observatory Alliance* was established in June 2002 during the conference *Toward an International Virtual Observatory*. Common metadata and standards are the key for a truly international Virtual Observatory, and it is highly significant that the first international milestone of the VO has been the definition of a standard, VOTable. An international *Interoperability Working Group*, first settled by the OPTICON (*OPTical Infrared Coordination Network*) European Thematic Network, and extended to include international participants from all the VO projects, holds regular meetings to discuss these topics, and mailing lists discussions are very active.

VOTable is an XML format for tabular data, defined by a wide international collaboration. It has been designed to take advantage of XML properties (data and associated metadata in a single document, encapsulation,…), and also to minimize the “tagging overhead”, using the property of tabular data where columns are assumed to be homogeneous in terms of their associated metadata. It is also designed to be compatible with existing FITS data, and allows FITS or BINARY data to be embedded in the document or remote.

Current discussions between the VO projects include the description of resources for resource discovery purposes, and semantic description of contents. The semantic definition of quantities in astronomic tables has been studied a few years ago in the frame of the *ESO-CDS Data Mining project*. A tree of *Uniform Content Descriptors* (UCDs) has been defined, by organising hierarchically the contents of individual table fields (columns) from the CDS catalogue collection (more than 100,000 columns) [5]. The diversity of information in catalogues and tables makes it an excellent starting point to describe the semantic content of astronomy. Several other efforts are going on, to develop adequate descriptions of different kinds of information, for instance space-time coordinates [6], or Image/Archive data models ([7], [8]). The evaluation of the usage of UCDs with additional more detailed descriptions is evaluated, and questions such as definition, curation and maintenance of UCDs are discussed ([9]).

4. Conclusion

Thanks to a long history of collaboration to define disciplinary standards, astronomy has been able to develop a set of networked services which is an everyday tool for scientists. New standards are discussed at the international level to develop new services and tools in the frame of the GRID and of the Virtual Observatory

Links

ADS	http://adswww.harvard.edu
CDS	http://cdsweb.u-strasbg.fr
NED	http://nedwww.ipac.caltech.edu/
ESA	http://www.esa.int
ESO	http://www.eso.org
IPAC	http://www.ipac.caltech.edu/
HEASARC	http://heasarc.gsfc.nasa.gov/
STScI	http://www.stsci.edu/resources/
CADC	http://cadcwww.dao.nrc.ca/
Chandra	http://cxc.harvard.edu/
CFHT	http://www.cfht.hawaii.edu/
AVO	http://www.eso.org/avo/
NVO	http://www.us-vo.org/
AstroGrid	http://www.astrogrid.org/
VO Conference, Garching 2002	http://www.eso.org/gen-fac/meetings/vo2002/
OPTICON	http://www.astro-opticon.org/
VOTable documentation	http://cdsweb.u-strasbg.fr/doc/VOTable/
VOTable discussion	http://archives.us-vo.org/VOTable/
UCD	http://cdsweb.u-strasbg.fr/UCD/

And many others…

References

[1] F. Ochsenebein and J. Lequeux. *Vistas in Astron.* **39**, 227, 1995.

[2] F. Ochsenbein, M. Albrecht, A. Brighton, P. Fernique, D. Guillaume, R.J. Hanisch, E. Shaya, and A. Wicenec. *ADASS IX*, *A.S.P. Conf. Ser.* **216**, P. 83, 2000.

[3] *Virtual Observatories of the future*. *A.S.P. Conf. Ser.* **225**, 2001.

[4] *Mining the Sky*. *ESO Astrophys. Symp.*, Springer Verlag, 2001.

[5] S. Derriere, et al. "Data Mining Facilities", in *Toward and International Virtual Observatory*, *ESO Astrophysics Symp.*, Springer Verlag, in press.

[6] A. Rots. "Space-Time metadata for the Virtual Observatory", in *Toward and International Virtual Observatory*, *ESO Astrophysics Symp.*, Springer Verlag, in press.

[7] M. Louys, et al.. "IDHA Image archive model", in *Toward and International Virtual Observatory*, *ESO Astrophysics Symp.*, Springer Verlag, in press.

[8] J. McDowell. "Towards an image data model for the Virtual Observatory", in *Toward and International Virtual Observatory*, *ESO Astrophysics Symp.*, Springer Verlag, in press.

[9] *SPIE Conference 4846*. August 2002, in press.

SUMMARY SESSION 5