Library and Information Services in Astronomy III
ASP Conference Series, Vol. 153, 1998
U. Grothkopf, H. Andernach, S. Stevens-Rayburn, and M. Gomez (eds.)

Information Extraction: New Developments in Astronomical Information Retrieval for Electronic Publications

Soizick Lesteven

CDS - Observatoire Astronomique, 11, rue de l'Université, 67000 Strasbourg, France

F. Bonnarel, P. Dubois, D. Egret, P. Fernique, F. Genova, F. Murtagh, F. Ochsenbein, M. Wenger

CDS - Observatoire Astronomique, 11, rue de l'Université, 67000 Strasbourg, France

Abstract. The explosion of on-line services and the rapid evolution in information technology, with the advent of the WWW, gives its full dimension to the electronic publication. Electronic publication has to be conceived with links to external resources (databases, bibliographic services) and with intelligent information retrieval tools. To provide links one needs to recognize the relevant information from a document, and to connect this information to the proper distributed resource. Recognition is the first step; the whole procedure may include the validation for correctness and completeness, and the addition of dynamic links to distributed services. In addition, publication in electronic form permits new methods to access published information.
Several activities take place at CDS in this context:
• CDS develops and maintains links to and from other distributed services (CDS services with electronic publications and the ADS);
• CDS develops and maintains services which give access to published information (the VizieR catalogue browser for published tables or SIMBAD which tracks object citations in papers);
• CDS develops information retrieval tools (the bibliographic maps or tools to automatically recognize object names in a text).
All these developments require close connections with the distributed services (editors, database managers, service managers, ...). A few examples will be presented.

1. Introduction

The rapid evolution in information technology and the explosion of on-line services are bringing important modifications in the way scientists collect information for their research. The availability of data and of scientific literature on the WWW makes it possible to interlink resources on the network, thus giving a highly value-added service for research purposes. In the astronomical community, the scientific literature and the data that support the research are

well-defined and electronically available. Some interoperabilities already exist between data resources (coming from data centers as CDS, NED and observatory archives), the ADS abstract service and electronic publications (ApJ, AJ, PASP, A&A, NewA, ...). Electronic publishing begins to be conceived with extensive links both within the document and to external resources.

To provide links one needs to extract relevant information from the document and to connect this information to the proper distributed resource. Information extraction is a really complex process that should not be underestimated. Recognition is the first step of the whole procedure; it can be relatively straightforward when the information is tagged in the text or corresponds to a standard format (tables, bibcodes) but it can be more complex when the data is heterogeneous (e.g. astronomical object names). The second step is the validation of the extracted information. The validation process should ensure the correctness of the information but also its completeness. This procedure should be completed manually by an expert. The third step of the extraction procedure is the addition of dynamic links to the distributed services. One may have to build procedures that take into consideration the fact that these services can evolve later on, whereas the published text has to remain unchanged.

Several CDS activities take place in this context. CDS has developed and maintains services which give access to published information, links between distributed services, information retrieval tools, and automatic recognition and extraction tools. A few applications will be presented in the following. The first one concerns the published tabular data and is already operational. The second one is an extraction tool for astronomical object names and is under construction. These two applications will illustrate the complexity of the information extraction process.

2. Published tabular data

2.1. Description

The CDS collects and distributes astronomical data catalogues, related to observations of stars and galaxies, and other galactic and extragalactic objects. Catalogues about the solar system bodies and atomic data are also included. Since January 1993, tables from articles published in Astronomy & Astrophysics are prepared and made available on-line at CDS, by agreement with the editor. Tables from the AAS CD-ROMs were also made available on-line by CDS by agreement with the AAS. Tables from some other major journals are also available. The number of tables is continuously increasing, in May 1998, one counts 2178 published tables from the major astronomical journals (Table 1). The CDS offers two ways to retrieve the catalogued data:

- The "Astronomer's Bazaar" (http://cdsweb.u-strasbg.fr/Cats.html) describes all catalogues stored at CDS, which can be copied via anonymous FTP.
- "VizieR", allows to access the most complete library of published astronomical catalogues and tables organized in a self-documented database (http://vizier.u-strasbg.fr/). VizieR is an excellent example of a new and powerful method to access published electronic information.

Astronomy and Astrophysics	375 catalogues
Astronomy and Astrophysics Supplement Series	843 catalogues
Astronomical Journal	346 catalogues
Astrophysical Journal	122 catalogues
Astrophysical Journal Supplement Series	267 catalogues
Publications of the Astronomical Society of the Pacific	52 catalogues
Other major journals	173 catalogues

Table 1. Number of published tables from the major astronomical journals available from the CDS catalogue service.

These services give access both to astronomical catalogues and published tables, thanks to the definition of a common standard which is used both by data centers and publishers.

2.2. CDS standards for catalogues

In order to facilitate the usage of the data in a large variety of contexts and the data processing, F. Ochsenbein (1994) proposed a Standard Description for Astronomical Catalogues. This standard documentation (accessible on the Web at http://vizier.u-strasbg.fr/doc/catstd.htx) is now shared with other astronomical catalogue producers. The description gives the signification and the format of the tables thus allowing easy extraction of the data from the tables. The standardization plays now a key role for exchange of tabular data between different partners by allowing:

- Data validation (from information given in the description)
  Edition of excerpts of tables to check their validity
- Transformation into other formats (FITS, Fortran, ..)
- Automated integration into the VizieR database providing access to all facilities

2.3. Data surfing

From an on-line article (for instance a paper in A&AS) the reader can get direct access to data tables available at CDS and to the facilities offered by the catalogue database. Reversely, from the electronic tables, one can access the corresponding on-line article when it exists. The interconnectivity between the electronic tables and the other astronomical services on the network is already running and new links will be added in the near future.

This interconnectivity is based on another standard: the 19-digit bibcode (http://cdsweb.u-strasbg.fr/simbad/refcode.html). First developed as a result of the cooperation between NED and CDS to provide a unique and readable representation of a bibliographic reference, it has become a standard code also – with minor variations – for ADS and other bibliographic services, in particular on-line journals. This code facilitates the exchange and can be automatically created. It makes the interconnectibility feasible between all bibliographic services and bibliographic data producers.

3. Astronomical object names

3.1. Description

When an astronomer reads an on-line article where an astronomical object name is cited, he/she would frequently like to get more information about it. Presently, this can be done by opening a new window on the screen and connecting to SIMBAD or NED and sending the appropriate request. Hypertext features of the Web allow in principle a much easier approach where just by clicking on the displayed name one would receive that information directly. The link can be completely transparent for the users, they don't need to know where the information is located and how the object has to be written for the query.

SIMBAD and NED also have to manage this information. The work is done manually by the bibliographers who read publications. Every time they recognize an object name in the article (title, abstract, table, ...), they update the databases. This means that they find how the object name has to be written in the database, and whether or not the object is already in the database. If this is not the case, they create the new object name. Then they link the reference to the existing or new object and add some basic data (coordinates, magnitudes) when known. The maintenance of that information is done by the SIMBAD and NED teams.

To help the bibliographers, and to allow direct access to an astronomical database from a electronic article text, we have begun to develop tools to automatically recognize the astronomical object names in texts. The problem is not trivial because an object name may be very complex, and it can be written in many different ways. Moreover, new acronyms are created on a regular basis for newly published lists.

3.2. Dictionary of the Nomenclature of Celestial Objects

An astronomical object name may be short, long, structured or not: examples are *Orion Nebula, the Superantennae, DR21(OH), CCDM J00335+4509BC, NGC 1866, QSO 0347-3819, Cl* NGC 2419 SAW V18, T Tau N,* etc. The extraction of all these kinds of names in a text is not straightforward; the way these objects are written is heterogeneous and varies from one paper to another, or even within a given paper.

To provide an automatic extraction tool, we have developed a software based on the "Dictionary of the Nomenclature of Celestial Objects" (Lortet et al. 1994). This dictionary is a reference work which tracks all designations quoted in the literature; it is available on the Web at http://vizier.u-strasbg.fr/cgi-bin/Dic.

A designation is a structured name basically made of an *acronym* and a *numbering* which are both strings of alphanumeric characters. The structure of the numbering is called the format. Examples of formats are *NNN* for a running number as in NGC, $\pm$*DD NNNN* for a running number in a declination zone as in BD, *JHHMMm+DDMMAAA* for J2000 coordinates as in CCDM, *FFF-NNN* for a running number in a field as in ESO, etc. A *specifier* can be added. There should be one object per designation but unfortunately many exceptions to this rule are found in the published literature. The Dictionary provides full references and usages of the different acronyms. An example, corresponding to

the ESO acronym, is given in Table 2. Furthermore, the Dictionary also gives the corresponding names in SIMBAD. It presently contains more than 5000 acronyms and it is updated on a regular basis.

Acronym	Use	Format	Year	1st Author	Obj. Type
ESO	ESO	FFF-TTT NN	1981	HOLMBERG E.B.+	(Opt)
	ESO	FFF-NNN			
ESO	ESO	HHMMSS+DDMM.m	1982	LAUBERTS A.	(Opt)
(ESO)	Ruiz	FFF-NNNA	1988	RUIZ M.T.+	*
	Ruiz	FFF-NNNW			
ESO-Halpha		NNN	1992	REIPURTH B.+	Em. *
ESO-HA	ESO-Halpha	NNN	1994	PETTERSSON B.+	Em. *
ESO-LV	ESO-LV	FFF-NNNN	1989	LAUBERTS A.+	G

Table 2. Dictionary of the Nomenclature of Celestial Objects: Result of a query for the acronym "ESO"

3.3. Tagging

Electronic publications are more and more conceived with links to external resources. Different publishers try to integrate direct links from object names in on-line articles to external information about the object. Two different approaches appear. The first one is implemented by the journal "New Astronomy". Some object names are selected in the article by the publisher, and a link to SIMBAD is included after validation by the SIMBAD team. The second approach is implemented by the journal "Astronomy and Astrophysics". A LaTeX macro has been created by the publisher, allowing authors to tag object names in their article. Some control tools have to be developed to help the authors and maintain the correctness of the link. Furthermore, validation by an expert will have to be performed to ensure the validity of the object name.

3.4. A name recognition tool

Another way to extract object names is to develop an automatic recognition tool. The tool is based on the "Dictionary of the Nomenclature of Celestial Objects". It is written in C language and uses rules (written with regular expressions). Each designation, coming from the dictionary, is automatically translated into a rule. The text is searched for the set of rules thus collected. Identifiers are retrieved. As already discussed, an astronomical object name can be a complex expression, a validation by an expert remains necessary to ensure the correctness and completeness of the recognition. Some of the inaccurate recognitions can be detected by filtering the results (for example, space mission names, spectral types, atomic or molecular species can easily be confused with object names). At the end of the process the object names are tagged in the text. Furthermore, a link between the name found in the literature and the SIMBAD name can be created.

3.5. Astronomical database update

When the above tool will be operational, the automatic identification of an astronomical object name will help the bibliographers who update astronomical databases (SIMBAD, NED). Their work will evolve to be more focussed on value-added activities such as validation of the names proposed by automatic tools or by the authors, and checking of the SIMBAD/NED syntax for the object name link. In addition, the following tasks will remain:

- Detect cross-identifications
- Add or improve data (coordinates, magnitudes, spectral types, redshifts, etc.)
- Detect and create new acronyms
- Control and detect inconsistencies.

The maintenance of the accuracy of links should not be underestimated. The links will have to survive changes in the database, while the article itself will by principle remain unchanged.

3.6. Conclusion

Astronomical Object name extraction is a complex process. The automatic recognition has to evolve with nomenclature. The "Dictionary of the Nomenclature of Celestial Objects" is updated on a regular basis. Automatic recognition becomes still more complicated when object names do not respect the different rules of the nomenclature. Specifications concerning designations have been defined by the Task Group on Astronomical Designations of IAU Commission 5, who wrote a document giving recommendations, definitions and examples. This document is available on-line (http://cdsweb.u-strasbg.fr/iau-spec.html).

Astronomical object name extraction can be done by the authors and publishers in parallel with automatic techniques, but an expert will still have to control that information. The links between the literature and the databases have to be maintained to ensure the correctness of the information. New tools have to be developed. As this process is shared between the authors, publishers and database managers, cooperation is essential.

4. Bibliographical surfing

The relationship between various services dealing with bibliography is shown in Figure 1, as seen from CDS. The interesting characteristic of links through the WWW is that each participating service itself may well be in the middle of such a plot.

4.1. GLU

To obtain a good interoperability between the different astronomical services, one needs to maintain all links one has created. This is a real challenge for database managers. In this context, the CDS has developed the GLU system

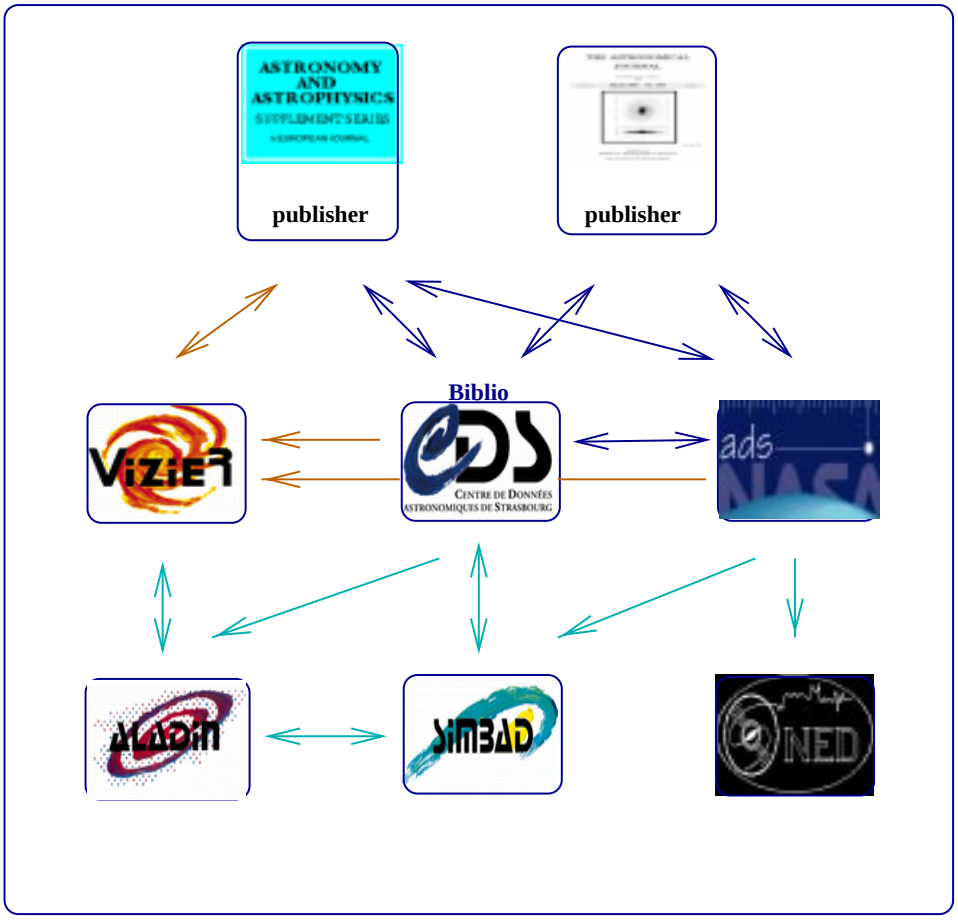

Figure 1. Interconnectivity between bibliographic services.

(Générateur de Liens Uniformes). This tool generates automatically hypertext links, avoiding the well-known drawbacks of hard-coded URLs, which are often not modified when the target address changes or even when small modifications affect a script generating the answers to an http request. To realize this purpose, the GLU implements two concepts:

- The GLU dictionary which is a compilation of symbolic names with their corresponding URLs and the way the parameters have to be written;
- The GLU resolver which replaces symbolic names and their associated parameters with relevant URLs on the fly.

Using this tool, data managers can forget the URLs and just use their symbolic names in all their Web documents. So, the GLU system is particularly adapted to design cooperative Web services, allowing to generate links to other services which always remain up-to-date. The GLU is already used, for example, for the bibliographic surfing in astronomy between the different CDS services and ADS, in the AstroBrowse NASA initiative, etc.

4.2. Bibliographical map

Another new application recently developed by the CDS is a visual tool, a bibliographic map, that allows to retrieve papers relevant to a domain. This map is based on a neural network analysis of the keywords associated to the articles. Documents having similar contents are clustered in the same area of the map. By clicking on a dot of the map, one can retrieve similar documents which are relevant to the request. A more complete description of that tool is presented by P. Poinçot in these proceedings (page 85) (http://simbad.u-strasbg.fr/A+A/map.pl).

5. Conclusions

Information extraction allows bibliographic and data surfing between all the services that deal with astronomy. It provides a high added value for reseach purposes.
Information extraction tools rely on standards shared at the astronomical community level. The links need to be permanently resolved requiring close cooperation between all the services (publishers, databases, authors, ...) shown in Figure 1.
In the future, information extraction should be improved and diversified to other type of information (magnitudes, coordinates, space missions, ...).

References

Fernique, P., Ochsenbein, F. & Wenger, M. 1998, CDS GLU, a tool for managing heterogeneous distributed Web services, in Astronomical Data Analysis Software and Systems VII, ASP Conf. Ser., Vol. 145, R. Albrecht, R. N. Hook & H. A. Bushouse, eds., (San Francisco: ASP), 466

Genova, F., Bartlett, J., Bonnarel, F., Dubois, P., Egret, D., Fernique, P., Jasniewicz, G., Lesteven, S., Ochsenbein, F. & Wenger, M. 1998, The CDS information hub, in Astronomical Data Analysis Software and Systems VII: ASP Conf. Ser., Vol. 145, R. Albrecht, R. N. Hook & H. A. Bushouse, eds., (San Francisco: ASP), 470

Lortet, M.-C., Borde, S. & Ochsenbein, F. 1994, Second Reference Dictionary of the Nomenclature of Celestial Objects, A&AS, 107, 193

Ochsenbein, F. 1994, Adopted Standards for Catalogues at CDS, Bull. Inform. CDS, 44, 19

Ochsenbein, F. 1997, Published Tabular Data, Baltic Astronomy 6, 221

Poinçot, P., Lesteven, S. & Murtagh, F. 1998, A spatial user interface to the astronomical literature, A&AS, 130, 183

Schmitz M. et al., 1995, NED and SIMBAD Conventions for Bibliographic Reference Coding, in Information & On-line Data in Astronomy, D. Egret & M. A. Albrecht, eds., (Dordrecht: Kluwer Acad. Publ.), 259