ASTRONOMY & ASTROPHYSICS
SUPPLEMENT SERIES

Astron. Astrophys. Suppl. Ser. **143**, 137–143 (2000)

Search and discovery tools for astronomical on-line resources and services

D. Egret[1], **R.J. Hanisch**[2], **and F. Murtagh**[1,3]

[1] CDS, Observatoire astronomique de Strasbourg, UMR 7550, 11 rue de l'Université, F-67000 Strasbourg, France
[2] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, U.S.A
[3] School of Computer Science, The Queen's University of Belfast, Belfast BT7 1NN, Northern Ireland

Received January 6; accepted January 17, 2000

Abstract. A growing number of astronomical resources and data or information services are made available through the Internet. However valuable information is frequently hidden in a deluge of non-pertinent or non up-to-date documents. At a first level, compilations of astronomical resources provide help for selecting relevant sites. Combining yellow-page services and meta-databases of active pointers may be an efficient solution to the data retrieval problem. Responses generated by submission of queries to a set of heterogeneous resources are difficult to merge or cross-match, because different data providers generally use different data formats: new endeavors are under way to tackle this problem. We review the technical challenges involved in trying to provide general search and discovery tools, and to integrate them through upper level interfaces.

Key words: astronomical databases: miscellaneous

1. Introduction

How to help the users find their way through the jungle of information services is a question which has been raised since the early development of the WWW (see e.g., Egret 1994), when it became clear that a big centralized system was not the efficient way to go.

Obviously the World Wide Web is a very powerful medium for the development of distributed resources: on the one hand the WWW provides a common medium for all information providers – the language is flexible enough so that it does not bring unbearable constraints on existing databases – on the other hand the distributed hypertextual approach opens the way to navigation and links between services (provided a minimum of coordination can be achieved). Let us note that it has been already widely demonstrated that coordinating spirit is not out of reach in a small community such as astronomy, largely sheltered from commercial influence.

Searching for a resource (either already visited, or unknown but expected), or browsing lists of existing services in order to discover new tools of interest implies a need for query strategies that cannot generally be managed at the level of a single data provider.

There is a need for road-guides pointing to the most useful resources, or to compilations or databases where information can be found about these resources. Such guides have been made in the past, and are of very practical help for the novice as well as the trained user, for example: Andernach et al. (1994), Egret & Heck (1995), Egret & Albrecht (1995), Heck (1997), Grothkopf (1995), Andernach (1999).

In the present paper our aim is to address the questions related to the collection, integration and interfacing of the wealth of astronomical Internet resources, and also to describe some strategies that have to be developed for building cooperative tools which will be essential in the research environment of the decade to come.

2. Compilations of astronomical Internet resources

At a first level, the user looking for new sources of information can consult compilations of existing resources. Examples of such databases, or yellow-page services are given in this section.

2.1. The StarPages

*Star*s Family* is the generic name for a collection of directories, dictionaries and databases which has been described in details by Heck (1995a) who has been building up their contents for more than twenty-five years.

Correspondence to: Daniel.Egret@astro.u-strasbg.fr

These very exhaustive data sets are carefully updated and validated, thus constituting a gold mine for professional, amateur astronomers, and more generally all those who are curious of space-related activities, and want to locate existing resources.

The Star*s Family of products can be queried on-line from the CDS Web site (Strasbourg, France) under the generic name of StarPages[1]. It includes the following databases:

StarWorlds: a directory of astronomy, space sciences, and related organizations (Heck et al. 1994); it includes URLs of Web sites when available, as well as e-mail addresses; unlike most of the services mentioned in the present paper, it is not restricted to describing on-line resources, but also lists directory entries for organizations which do not provide any on-line information.

StarHeads: individual Web pages essentially of astronomers and related space scientists (Heck 1995b).

StarBits: a very comprehensive dictionary of abbreviations, acronyms, contractions, and symbols used in astronomy and space sciences (Heck 1995b).

All three databases are associated with a query engine based on character string searches. Filters prevent extraction of too large subsets of the database.

2.2. *AstroWeb*

AstroWeb (Jackson et al. 1994) is a collection of pointers to astronomically relevant information resources available on the Internet. The browse mode of AstroWeb opens a window on the efforts currently developed – in some cases, unfortunately, in a rather disorganized way – for making astronomically related, and hopefully pertinent, information available on-line through the World Wide Web.

AstroWeb is maintained by a small consortium of individuals located at CDS, STScI, MSSSO, NRAO, and Vilspa. The master database is currently hosted at CDS[2] (after having been for a long time at STScI), and all the above-mentioned places, as well as the Institute of Astronomy, Cambridge, host a mirror copy with customized presentation of the same data.

Each URL is checked by a robot on a daily basis to ensure aliveness of all referenced resources. The resource descriptions are usually submitted by the person or organization responsible for the resource, but are checked and eventually modified by one of the consortium members. The search engine is a WAIS search index. The index is constructed from the resource descriptions, and also includes all the words contained in the referenced home page. This latter feature is quite powerful for bringing new names of projects, topics, research groups, very quickly to the index.

Table 1 lists the resources present in the AstroWeb database in December 1999.

3. Current status of on-line astronomy resources

Following the classification scheme adopted by AstroWeb, we will outline in this section the current status of the main categories of on-line astronomy resources, pointing to meta-resources (i.e. organized lists of resources) when they are available.

3.1. *Organizations*

Most of the active astronomical organizations (institutes, astronomy departments, etc.) now have home pages on the Internet. StarWorlds[3] is currently the most comprehensive searchable directory of such resources; it can be queried by names, keywords, or character strings. For browsing lists sorted by alphabetical order or by country, see AstroWeb (Sect. 2.2). National or international organizations also maintain useful lists.

3.2. *Observational projects and missions*

It is now difficult to envisage an observational project without a web site. As they are more dynamic and often involve multiple organizations or institutions, the best way to find them may be to use one of the powerful commercial search engines that routinely index millions of web pages on the Internet.

The indexing system of AstroWeb may also be helpful, especially when it is important to limit the investigation domain to astronomy, or to keep track of new emerging projects.

3.3. *Data and information systems*

Astronomy data and information centers are becoming increasingly interconnected, with both explicit links to other relevant resources and automatic cross-links that may be invoked transparently to the end-user. Section 5 describes current efforts to provide interoperability within astrophysics (*Astrobrowse*) and across the space sciences (*ISAIA*).

1 http://cdsweb.u-strasbg.fr/starpages.html

2 http://cdsweb.u-strasbg.fr/astroweb.html

3 http://cdsweb.u-strasbg.fr/starworlds.html

Table 1. Resources listed in the AstroWeb database (December 1999). The number of resources (Web sites) in each category is given between parentheses. A number of resources appear in more than one category

Organizations	Astronomy Departments (508)
	Professional and Amateur Organizations (159)
	Space Agencies and Organizations (46)
Observing	Observatories and Telescopes (328)
resources	Telescope Observing Schedules (25)
	Meteorological Information (10)
	Astronomical Survey Projects (65)
Data resources	Data and Archive Centers (145)
	Astronomy Information Systems (39)
Abstracts,	Bibliographical Services (29)
Publications,	Astronomical Journals and Publications (90)
Libraries	Astronomy & astrophysics preprints (58)
	Abstracts of Astronomical Publications (29)
	Conference Proceedings (45)
	Astronomy-related Libraries (48)
	Other Library resources (11)
People-related	Personal Web pages (800)
Resources	People (lists) (14)
	Jobs (37)
	Conferences and Meetings (45)
	Newsgroups (31)
	Mailing Lists (16)
Software	Astronomy software servers (129)
Computer	Document Preparation Tools (9)
Science	Overviews & technical notes for protocols (11)
	Computer Science-related Resources (33)
Research areas	Radio Astronomy (109)
Astronomy	Optical Astronomy (178)
Space Physics	High-Energy Astronomy (77)
	Space Astronomy (175)
	Solar Astronomy (77)
	Planetary Astronomy (64)
	History of Astronomy (21)
	Earth, Ocean, Atmosphere, Space Sciences (41)
	Physics-related Resources (91)
Educational	Professional and Amateur Organizations (159)
resources	Educational resources (240)
	Astronomy Pictures (105)
Miscellaneous	Primary Lists of Astronomy Resources (10)
	Other lists of astronomy resources (78)
	Miscellaneous Resources (137)

3.4. Bibliographic resources

Here also a virtual network is being organized, as exemplified by the *Urania*[4] initiative, or by the coordinated efforts to create links between ADS and other services (Kurtz et al. 2000). Note that many of the bibliographical resources are electronic journals for which a subscription may be required.

3.5. People-related resources

Some databases (RGO E-mail directory[5], StarHeads[6]) follow the development of electronic mail addresses and personal Web pages. Directories from national or international societies (e.g., AAS, EAS, IAU) are also generally very carefully kept up to date.

The database of meetings and conferences maintained by CFHT[7] is very complete and well organized. Astronomical societies also maintain their own lists.

3.6. Astronomical software

The Astronomical Software and Documentation Service (ASDS[8]) is a network service that allows users to locate existing astronomical software, associated technical documentation, and information about telescopes and astronomical instrumentation (Payne et al. 1996). ASDS originated as a service devoted entirely to astronomical software packages and their associated on-line documentation and was originally called the Astronomical Software Directory Service. Much code is rewritten these days, not because anyone has found a fundamentally better way to solve the problem, but because developers simply don't know who has already done it, whether the code runs on the system they have available, or where to get it if it does. That is the problem that ASDS was intended to solve.

In 1998 the scope of ASDS was expanded to include astronomical observing sites and their associated telescope and instrument manuals, taken from a listing maintained at CFHT. The service was renamed at this point.

3.7. Educational resources

Education and public outreach have always been a strong concern in astronomy, but the importance of this activity is growing at a higher rate, with the advent of the World Wide Web.

It is difficult to give general rules for such a wide field, going far beyond the limits of astronomical institutions.

[4] http://www.aas.org/Urania/
[5] http://star-www.rl.ac.uk/astrolist/astrosearch.html
[6] http://cdsweb.u-strasbg.fr/starheads.hml
[7] http://cadcwww.dao.nrc.ca/meetings/meetings.html
[8] http://asds.stsci.edu/

Let us just say that we expect to see in the future an increasing rôle of educational institutions (planetariums, or outreach departments of big societies or institutions), for conveying general astronomy knowledge, or news about recent discoveries, to the general public.

The yellow-page services mentioned above do keep lists of the most important education services.

4. Towards a global index of astronomical resources

In the following we will focus on Internet resources that actually provide data, of any kind, as opposed to those describing or documenting an institution or a research project, without giving access to any data set or archive.

One main trend is certainly the increase of interconnections between distributed on-line services, the "Weaving of the Astronomy Web" (which was the title of a Conference organized in Strasbourg by Egret & Heck 1995).

More generally, with the development of the Internet, and of a large number of on-line services giving access to data or information, it is clear that tools giving coordinated access to distributed services are needed. This is, for instance, the concern expressed by NASA through the Astrobrowse project (Heikkila et al. 1999).

In this section we will first describe a tool for managing a "metadata" dictionary of astronomy information services (GLU); then we will show how the existence of such a metadatabase can be used for building efficient search and discovery tools.

4.1. The CDS GLU

The CDS (Centre de Données astronomiques de Strasbourg) has recently developed a tool for managing remote links in a context of distributed heterogeneous services (GLU[9], Générateur de Liens Uniformes, i.e. Uniform Link Generator; Fernique et al. 1998).

First developed for ensuring efficient interoperability of the several services existing at CDS (VizieR, Simbad, Aladin, bibliography, etc.; see Genova et al. 2000), this tool has also been designed for maintaining addresses (URLs) of distributed services (ADS, NED, etc.).

A key element of the system is the "GLU dictionary" maintained by the data providers contributing to the system, and distributed to all sites of a given domain. This dictionary contains knowledge about the participating services (URLs, syntax and semantics of input fields, descriptions, etc.), so that it is possible to generate automatically a correct query for submission to a remote database.

The service provider (data center, archive manager, or webmaster of an astronomical institute) can use GLU for coding a query, taking benefit of the easy update of the system: knowing which service to call, and which answer to expect from this service, the programmer does not have to worry about the precise address of the remote service at a given time, nor of the detailed syntax of the query (expected format of the equatorial coordinates, etc.).

4.2. New search and discovery tools

The example of GLU demonstrates the usefulness of storing into a database the knowledge about information services (their address, purpose, domain of coverage, query syntax, etc.). In a second step, such a database can be queried when the challenge is to provide information about whom is providing what, for a given object, region of the sky, or domain of interest.

Several projects are working toward providing general solutions.

4.2.1. Astrobrowse

Astrobrowse is a project that began within the United States astrophysics community, primarily within NASA data centers, for developing a user agent which significantly streamlines the process of locating astronomical data on the web. Several prototype implementations are already available[10]. With any of these prototypes, a user can already query thousands of resources without having to deal with out-of-date URLs, or spend time figuring out how to use each resource's unique input formats. Given a user's selection of web-based astronomical databases and an object name or coordinates, Astrobrowse will send queries to all databases identified as containing potentially relevant data. It provides links to these resources and allows the user to browse results from each query. Astrobrowse does not recognize, however, when a query yields a null result, nor does it integrate query results into a common format to enable intercomparison.

4.2.2. AstroGLU

Consider the following scenario: we have a data item I (for example an author's name, the position or name of an astronomical object, a bibliographical reference, etc.), and we would like to know more about it, but we do not know a priori which service S to contact, and what are the different data types D which can be requested. This scenario is typical of a scientist exploring new domains as part of a research procedure.

The GLU dictionary can actually be used for helping to solve this question: the dictionary can be considered as a reference directory, storing the knowledge about all services accepting data item I as input, for retrieving data D_1 or D_2. For example, we can easily obtain from

[9] http://simbad.u-strasbg.fr/glu/glu.htx

[10] http://heasarc.gsfc.nasa.gov/ab/

such a dictionary the list of all services accepting an author's name as input; information which can be accessed, in return, may be an abstract (service ADS), a preprint (LANL/astro-ph), the author's address (RGO e-mail directory) or personal Web page (StarHeads), etc.

Based on such a system, it becomes possible to create automatically a simple interface guiding the user towards any of the services described in the dictionary.

This idea has been developed as a prototype tool, under the name of AstroGLU[11] (Egret et al. 1998). The aim of this tool is to help the users find their way among several dozens (for the moment) of possible actions or services. A number of compromises have to be taken between providing the user with the full information (which would be too abundant and thus unusable), and preparing digest lists (which implies hiding some amount of auxiliary information and making somewhat subjective selections).

A resulting issue is the fact that the system puts on the same line services which have very different quantitative or qualitative characteristics. AstroGLU has no efficient ways yet to provide the user with a hierarchy of services, as a gastronomic guide would do for restaurants. This might come to be a necessity in the future, as more and more services become (and remain) available.

5. Towards an integration of distributed data and information services

To go further, one needs to be able to integrate the result of queries provided by heterogeneous services. This is the aim of the ISAIA (Integrated System for Archival Information Access) project[12] (Hanisch 2000a, 2000b).

The key objective of the project is to develop an interdisciplinary data location and integration service for space sciences. Building upon existing data services and communications protocols, this service will allow users to transparently query a large variety of distributed heterogeneous Web-based resources (catalogs, data, computational resources, bibliographic references, etc.) from a single interface. The service will collect responses from various resources and integrate them in a seamless fashion for display and manipulation by the user.

Because the scope of ISAIA is intended to span the space sciences – astrophysics, planetary science, solar physics, and space physics – it is necessary to find a way to standardize the descriptions of data attributes that are needed in order to formulate queries. The ISAIA approach is based on the concept of *profiles*. Profiles map generic concepts and terms onto mission or dataset specific attributes. Users may make general queries across multiple disciplines by using the generic terms of the highest level profile, or make more specific queries within subdisciplines using terms from more detailed subprofiles.

The profiles play three critical and interconnected roles:

1. They identify appropriate resources (catalogs, mission datasets, bibliographic databases): the *resource profile*;
2. They enable generic queries to be mapped unambiguously onto resource-specific queries: the *query profile*;
3. They enable query responses to be tagged by content type and integrated into a common presentation format: the *response profile*.

The resource, query, and response profiles are all aspects of a common database of resource attributes. Current plans call for these profiles to be expressed using XML (eXtensible Markup Language, an emerging standard which allows embedding of logical markup tags within a document) and to be maintained as a distributed database using the CDS GLU facility.

The profile concept is critical to a distributed data service where one cannot expect data providers to modify their internal systems or services to accommodate some externally imposed standard. The profiles act as a thin, lightweight interface between the distributed service and the existing specific services. Ideally the service-specific profile implementations are maintained in a fully distributed fashion, with each data or service provider running a GLU daemon in which that site's services are fully described and updated as necessary. Static services or services with insufficient staff resources to maintain a local GLU implementation can still be included, however, as long as their profiles are included elsewhere in the distributed resource database. The profile concept is not unique to space science, but would apply equally well to any distributed data service in which a common user interface is desired to locate information in related yet traditionally separate disciplines.

6. Information clustering and advanced user interfaces

A major challenge in current information systems research is to find efficient ways for users to be able to visualize the contents and understand the correlations within large databases. The technologies being developed are likely to be applicable to astronomical information systems. For example, information retrieval by means of "semantic road maps" was first detailed in Doyle (1961), using a powerful spatial metaphor which lends itself quite well to modern distributed computing environments such as the Web. The Kohonen self-organizing feature map (SOM; Kohonen 1982) method is an effective means towards this end of a visual information retrieval user interface.

[11] http://simbad.u-strasbg.fr/glu/cgi-bin/astroglu.pl
[12] http://heasarc.gsfc.nasa.gov/isaia/

Fig. 1. Visual interactive user interface to a set of articles from the journal *Astronomy and Astrophysics*. Original in color

6.1. Interfacing datasets with a Self-organizing Map

The Kohonen map is, at heart, k-means clustering with the additional constraint that cluster centers be located on a regular grid (or some other topographic structure) and furthermore their location on the grid be monotonically related to pairwise proximity (Murtagh & Hernández-Pajares 1995).

A regular grid is quite convenient for an output representation space, as it maps conveniently onto a visual user interface. In a web context, it can easily be made interactive and responsive.

Figure 1 shows an example of such a visual and interactive user interface map, in the context of a set of journal articles described by their keywords. Color is related to density of document clusters located at regularly spaced nodes of the map, and some of these nodes/clusters are annotated. The map is installed on the Web as a clickable image map, with CGI programs accessing lists of documents and – through further links – in many cases, the full documents. In the example shown, the user has queried a node and results are seen in the right-hand panel. Such maps are maintained for (currently) 12000 articles from the *Astrophysical Journal*, 7000 from *Astronomy and Astrophysics*, over 2000 astronomical catalogs, and other data holdings. More information on the design of this visual interface and user assessment can be found in Poinçot et al. (1998, 2000).

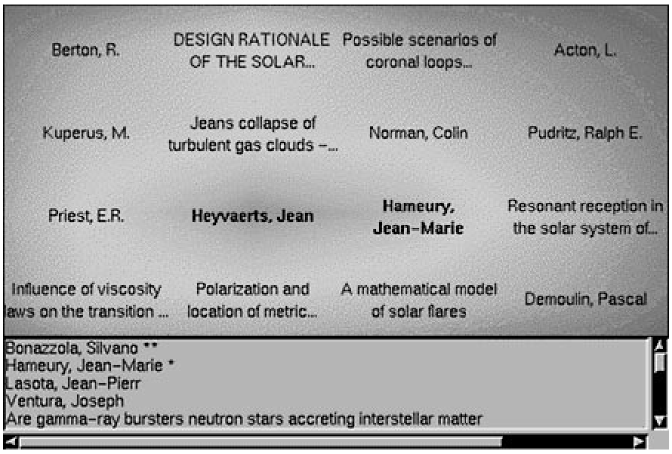

Fig. 2. Visual interactive user interfaces, based on graph edges. Vertices are author names, article titles and (not shown here) astronomical object names. Map for astronomer Jean Heyvaerts. Original in color

6.2. Hyperlink clustering

Guillaume & Murtagh (2000) have recently developed a Java-based visualization tool for hyperlink-based data, in XML, consisting of astronomers, astronomical object names, article titles, and possibly other objects (images, tables, etc.). Through weighting, the various types of links could be prioritized. An iterative refinement algorithm was developed to map the nodes (objects) to a regular grid of cells, which, as for the Kohonen SOM map, are clickable and provide access to the data represented by the cluster. Figure 2 shows an example for an astronomer (Prof. Jean Heyvaerts, Strasbourg Astronomical Observatory).

These new cluster-based visual user interfaces are not computationally demanding. In general they cannot be created in real time, but they are scalable in the sense that many tens of thousands of documents or other objects can be easily handled. Document management (see e.g. Cartia[13]) is less the motivation as is instead the interactive user interface.

Further information on these visual user interfaces can be found in Guillaume (2000) and Poinçot (1999).

6.3. Future developments for advanced interfaces

Two directions of development are planned in the near future. Firstly, visual user interfaces need to be coupled together. A comprehensive "master" map is one possibility, but this has the disadvantage of centralized control and/or configuration control. Another possibility is to develop a protocol such that a map can refer a user to other maps in appropriate circumstances. Such a protocol was developed

[13] http://www.cartia.com/

a number of years ago in a system called Ingrid[14] developed by P. Francis at NTT Software Labs in Tokyo (see Guillaume 2000). However this work has been reoriented since then.

Modern middleware tools may offer the following solution. This is to define an information sharing bus, which will connect distributed information maps. It will be interesting to look at the advantages of CORBA (Common Object Request Broker Architecture) or, more likely, EJB (Enterprise Java Beans), for ensuring this interoperability infrastructure (Lunney & McCaughey 2000).

A second development path is to note the clustering which is at the core of these visual user interfaces and to ask whether this can be further enhanced to facilitate construction of query and response agents. It is clear to anyone who uses Internet search engines such as AltaVista, Lycos, etc. that clustering of results is very desirable. A good example of such clustering of search results in practice is the Ask Jeeves search engine[15]. The query interface, additionally, is a natural language one, another plus.

7. Conclusion

The on-line "Virtual Observatory" is currently under construction with on-line archives and services potentially giving access to a huge quantity of scientific information: its services will allow astronomers to select the information of interest for their research, and to access original data, observatory archives and results published in journals. Search and discovery tools currently in development will be of vital importance to make all the observational data and information available to the widest scientific community.

Acknowledgements. CDS acknowledges the support of INSU-CNRS, the Centre National d'Etudes Spatiales (CNES), and Université Louis Pasteur.

References

Andernach H., Hanisch R.J., Murtagh F., 1994, PASP 106, 1190

Andernach H., 1999, in Astrophysics with Large Databases in the Internet Age, Proc. IXth Canary Islands Winter School on Astrophysics, Kidger M., Pérez-Fournon I., Sánchez F. (eds.). Cambridge University Press, p. 1

Doyle L.B., 1961, J. ACM 8, 553

Egret D., 1994, in Astronomical Data Analysis Software and Systems III, ASP Conf. Ser. 61, 14

Egret D., Albrecht M., 1995, Information & On-line Data in Astronomy. Kluwer Academic Publ., Dordrecht

Egret D., Heck A., 1995, Weaving the Astronomy Web, Vistas Astron. 39, 1–127

Egret D., Fernique P., Genova F., 1998, in Astronomical Data Analysis Software and Systems VII, ASP Conf. Ser. 145, p. 416 (AstroGlu)

Fernique P., Ochsenbein F., Wenger M., 1998, in Astronomical Data Analysis Software and Systems VII, ASP Conf. Ser. 145, 466 (GLU)

Genova F., Egret D., Bienaymé O., et al., 2000, A&AS 143, 1 (CDS)

Grothkopf U., 1995, Vistas Astron. 38, 401

Guillaume D., Murtagh F., 2000, "Clustering of XML documents", Computer Physics Communications (in press)

Guillaume D., 2000, Ph.D. Thesis, Université Louis Pasteur, Strasbourg

Hanisch R.J., 2000a, Computer Physics Communications (in press)

Hanisch R.J., 2000b, in ADASS IX, ASP Conf. Ser. (in press)

Heck A., 1995a, in Information & On-line Data in Astronomy, Egret D. and Albrecht M.A. (eds.), p. 195 (Star*s Family)

Heck A., 1995b, A&AS 109, 265 (StarHeads)

Heck A., 1997, Electronic Publishing for Physics and Astronomy, Astrophys. Space Science 247. Kluwer, Dordrecht

Heck A., Egret D., Ochsenbein F., 1994, A&AS 108, 447 (StarWorlds – StarBits)

Heikkila C.W., McGlynn T.A., White N.E., 1999, in Astronomical Data Analysis Software and Systems VIII, ASP Conf. Ser. 172, 221 (Astrobrowse)

Jackson R., Wells D., Adorf H.M., et al., 1994, A&AS 108, 235 (AstroWeb)

Kohonen T., 1982, Biol. Cyber. 43, 59

Kurtz M., Eichhorn G., Accomazzi A., et al., 2000, A&AS 143, 41 (ADS)

Lunney T.F., McCaughey A.J., 1999, Computer Physics Communications (in press)

Murtagh F., Hernández-Pajares M., 1995, J. Class. 12, 165

Payne H.E., Hanisch R.J., Warnock A., 1996, in Astronomical Data Analysis Software and Systems V, ASP Conf. Ser. 101, 577

Poinçot Ph., Lesteven S., Murtagh F., 1998, A&AS 130, 183

Poinçot Ph., Lesteven S., Murtagh F., 2000, J. Am. Soc. Info. Sci. (submitted)

Poinçot Ph., 1999, Ph.D. Thesis, Université Louis Pasteur, Strasbourg

[14] http://www.ingrid.org/

[15] http://www.askjeeves.com/