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

Distributed Information Search and Retrieval for Astronomical Resource Discovery and Data Mining

Fionn Murtagh

Observatoire de Strasbourg, Université Louis Pasteur, Strasbourg, France; Faculty of Informatics, University of Ulster, Londonderry, Northern Ireland

Damien Guillaume

Observatoire de Strasbourg, Université Louis Pasteur, Strasbourg, France

Abstract. Information search and retrieval has become by nature a distributed task. We look at tools and techniques which are of importance in this area. Current technological evolution can be summarized as the growing stability and cohesiveness of distributed architectures of searchable objects. The objects themselves are more often than not multimedia, including published articles or grey literature reports, yellow page services, image data, catalogs, presentation and online display materials, and "operations" information such as scheduling and publicly accessible proposal information. The evolution towards distributed architectures, protocols and formats, and the direction of our own work, are focussed on in this paper.

1. Needs and Existing Solutions in Astronomy

Astronomy is characterized by massive data collections which are often incompatible, from many different sources, and stored in many different ways [1]. The user – astronomer or librarian – needs access in a common environment.

The traditional solution to this issue is to have centralized databases available. This provides the common environment needed, and helps with quality control. Major databases can be considered, such as Simbad [2], NED [3] and LEDA [4] for astronomical objects, VizieR [5] for tables or ADS [6] for astronomical articles. Collecting data, storing them with one format in one database, and providing user interfaces, all require very considerable human resources. For the user, problems may arise when accessing data from different databases, because the data on a given subject reside on different servers with different meta-information and formatting. These problems are partially overcome by linkages between the data collections. For example, links between Simbad, NED, VizieR and ADS allow retrieval of articles based on an object which was cited.

It is also possible to use a more direct protocol with NED and Simbad to retrieve information about an object, where one system can send a request to another and then format the response in an appropriate way. Such peer-to-peer linkage is used by different servers, for example, to find all names for an astronomical object. When such peer-based protocols are used, the overall system is commonly referred to as a federated database.

To improve the flexibility and the reliability of HTTP links, a package called the Uniform Link Generator, GLU, has been developed to keep a distributed list of available data, based on their URLs [7]. GLU tags are used to reference data, and these tags are resolved into URLs automatically.

A standard for passing parameters to CGI [8] programs (Common Gateway Interface, the common means of providing a Web server with the ability to execute programs to carry out any desired task) for astronomy has been defined, and is called the Astronomical Server URL, ASU [9]. This standard extends a URL which points to a catalog or database to allow access to specified parts of these datasets. Such access requires a program to be run. In the case of a database, this program involves the submission of an SQL (Standard Query Language - the most usual interface language for relational databases) query.

Text indexing and retrieval is often supported by public domain toolsets, which may support the Z39.50 information retrieval protocol or some related but different variant of this protocol. Isite and Isearch are part of one such package which is available from the Center for Networked Information Discovery and Retrieval (CNIDR). The context of such indexing and search toolsets is set out in Murtagh [10]. Gateways to Web servers are available for such text-supporting packages, which replace the erstwhile WAIS user interfaces.

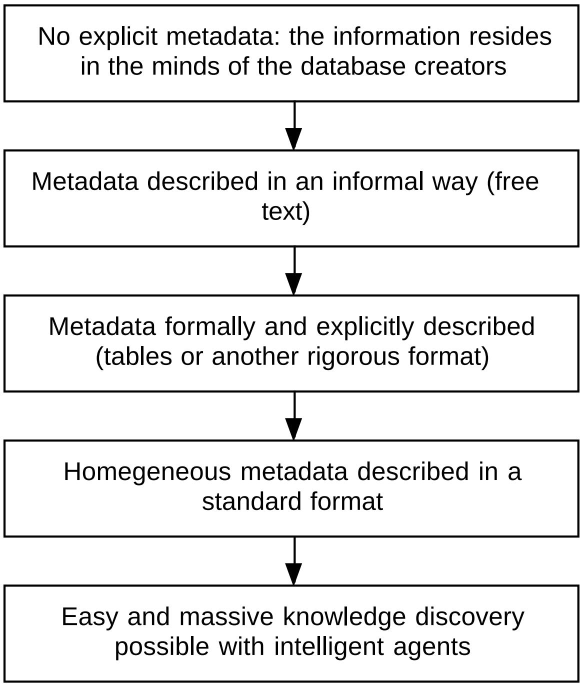
2. Metadata - Descriptive and Content Markup

Today's intelligent agents and knowledge discovery programs in astronomy are restricted to a small amount of data, because of the lack or heterogeneity of metadata. These programs need to have data described formally and precisely, ideally in a standardized format, to allow subsequently for statistical processing and cross-correlating of documents.

Currently, all the major steps in the evolution towards standardized metadata exist. In particular in the library database area, Dublin Core [11] is widely used, and has been extended with the Warwick Framework [12]. Distributed environmental and governmental information has employed GILS, the Global Information Locator Service [13], which incorporates such metadata standards. The ESPRIT ADDSIA (Access to Distributed Databases for Statistical Information and Analysis) project [14] seeks, among other objectives, to define and apply metadata to official and economic statistical micro- and macrodata.

With regard to descriptive metadata information, usually a database is in one state and can be improved by changing to the next state. We will look at this process.

STEPS OF METADATA ORGANISATION IN A DATABASE


No explicit metadata: the information resides in the minds of the database creators
Metadata described in an informal way (free text)
Metadata formally and explicitly described (tables or another rigorous format)
Homegeneous metadata described in a standard format
Easy and massive knowledge discovery possible with intelligent agents

- The first step is a minimum for any database, because forgetting it could result in the loss of metadata, for example if the creators of the database leave. This risk should not be taken at any price.
- The second step, formalizing all metadata information, is the most difficult, but results in better quality and self-consistency for the database.
- Deciding on a standard and adopting it can be painful, and will not necessarily be seen as an improvement for a given institute, but is very rewarding for the more general community, and can cater for new developments in the discipline.
- Finally, all this work will enable very important developments in knowledge discovery and many other types of data processing. The preceding steps may not be seen as necessary, but without a standardized way of describing metadata the subsequent processing programs can only use a limited amount of homogeneous data, restricted to a given institute for example.

Currently, improvements are planned in regard to many databases, but this work is often delayed by other projects having results apparently available more quickly. This leads to the question of the importance of metadata in astronomy, and the answer is that it is very important, probably more so than in many other disciplines: astronomical data are varied, but they have numerous links in them and they are often needed by different geographically-dispersed groups at the same time and for different purposes. So, working on the question of metadata for any computer-available data collection is not a waste of time.

A comprehensive format which can be used as a basis for metadata is SGML [15], Standard Generalized Markup Language. SGML allows the definition of markup languages with a Document Type Definition (DTD) file describing the syntax of the language. It enables the creation of appropriate formats in a very flexible way, and facilitates automatic understanding and reading of the SGML files.

The main limitation of SGML is that it only describes text files. The advantage is clearly that no special reader is necessary for viewing an SGML file, but on the other hand binary files have the advantage of being compact and are read faster. The usual solution, if binary files are necessary, is to put references to these files in the SGML files. Generally speaking, SGML is especially useful for exchanging a limited amount of data with lots of metadata. Its role is not limited to simply supporting a specific application on a specific platform.

In spite of the ASCII limitation, SGML is currently the best format for most data. It is an international standard (ISO 8879), is widely used (a list of companies and organisations using it would be far too extensive to list) and its use continues to spread in all computing activities.

SGML is

- flexible: it is simple to modify a DTD and the documents using it accordingly. It is also easy to create a new version of a DTD and keep it compatible with the old versions, for example by adding a tag.
- rigorous: despite the flexibility, an SGML document must always conform to a DTD, and the validating parsers verify automatically this conformance.
- easily readable: anyone understanding a markup language such as HTML can understand the meaning of an SGML document. As it is in text form, it requires no special editor.
- personalizable for multi-users: it is possible, for example, to add specific tags in a document which are not defined in the DTD. It is also possible to use different DTDs and to keep global consistency between them.
- content-oriented: an SGML document usually concentrates on the content of the text, not on its appearance: for instance, a title would be included in `<TITLE>` tags, not in `<I> <B> <Font="Helvetica"> <Size=24>`.
- easily portable: as it is content-oriented, it is simple, after the creation of an SGML document, to convert it into different interface-oriented formats such as HTML or Postscript.

The success of SGML is not only visible from the existence of many DTDs used in various areas. It is also supported by a lot of software, for parsing and browsing. This means that a parser for a new SGML-based language can use the object classes of a general SGML parser. It becomes also quickly readable with a nice interface, with existing SGML browsers. The use of SGML becomes especially easy in a Web environment, and can be easily processed with Java or Perl, but it is in no sense limited to being processed by these languages.

As the overall definition of SGML is rather complex, a subset of it has been defined to make its use even more easy: XML, Extensible Markup Language [16]. XML simplifies SGML, and is thus easier to implement. It will soon be supported by Web browsers such as Netscape and Internet Explorer, and other uses of it are currently under development.

We are currently defining a new XML language called AML, "Astronomical Markup Language", aimed at being a standard exchange format for metadata in astronomy. AML now supports the following objects (in the object-oriented sense): astronomical object, article, table, set of tables, image, person. This means that all these objects can be described with the same language, allowing easier establishing of links between them, and the creation of programs handling all these objects with the same user interface. AML examples and a user interface for browsing them (developed in Java) can be seen at `http://www.infm.ulst.ac.uk/~damien/these`.

The creation of this language is a prerequisite for further work on information retrieval in astronomy, and it will allow the use of the same information by both humans and intelligent agents: the intelligent agents can use AML with an XML parser, while a browser for AML documents will be used to create an associated user interface.

3. New Directions in Computer Science - Distributed Computing

Some new or not-so-new concepts or tools used in our current work on information retrieval and resource discovery are: object-oriented concepts, Java, CORBA and distributed search. CORBA, Common Object Request Broker Architecture [17], is a standard for communication between distributed objects on the Internet (or an Intranet): it allows the transparent call of an object's procedure from another object, possibly implemented in another language, on another system, and from another server. What we mean by "distributed search" is a search that uses different data located on different servers. Usually search engines on the Web use, on the contrary, centralized databases, sometimes with mirrors at different locations, but they are not truly distributed.

With progress in Internet technologies, it is possible to build an open architecture to exchange data and processes. CORBA enables the creation of distributed applications, with a Java user interface and different processes executed on different servers: the availability is "universal", in the sense that both human users and intelligent agents can work on the same data. With SGML documents, which have their structure described in a DTD file, the agents can indiscriminately use data from different sources.

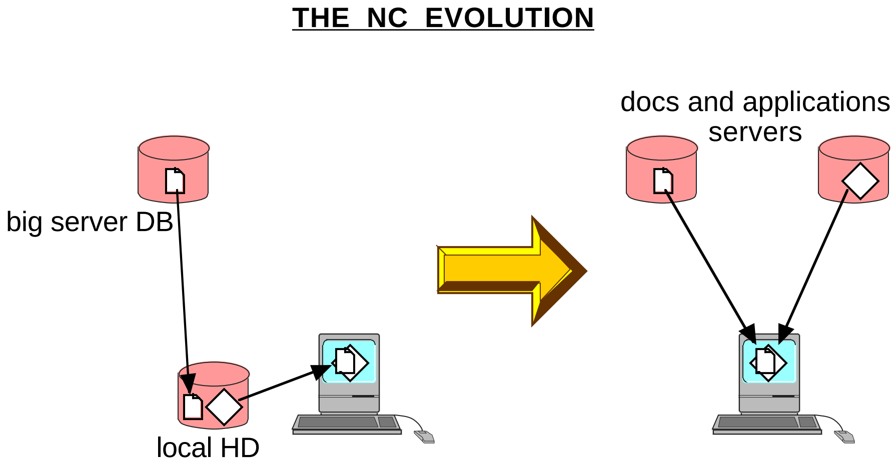
3.1. The Network Computer Concept

Nowadays, to use a file available on the network, we have to:

- download the file
- have a local program to open and process the file
- locally open the file with the program

What could be done (especially with a network computer) is:

- use a Java browser on a navigator to read and process the distant file referenced by a URL



In this way, there is no need for a local program, no compatibility or version problem, and no protection problems (the Java browser can use a specific server to do lengthy processing on a file, thus allowing access control by the server). The use of a system like CORBA also allows reuse of programs or subroutines on the network, with an object-oriented architecture.

In addition, the use of SGML files is an important issue for information retrieval and allows an interface to be defined more easily for these files. Using SGML content-oriented files to produce an interface (for example with HTML) is better than to mix the content and the interface, for instance by adding special comments in an HTML file.

3.2. Distributed Search Engines

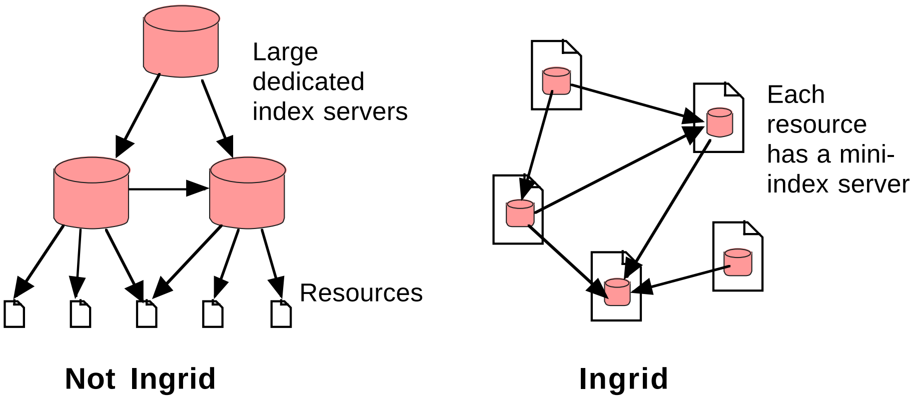
Given many different SGML document collections on many different servers, what is the best way to do a search? There is no clear answer yet, and there are many search engine experiments trying to solve this problem.

Distributed searches have many advantages, such as allowing the control of the indexing and updating of the data in a given server on this server. There are however different stages in the "distributedness" of a search engine.

One solution is a fully distributed search, with the metadata distributed in different sites and linked together by a grid: an experiment like this has already been conducted by NTT Software Labs, with the name of Ingrid [18].

Ingrid is also quite specific in the way that its grid is self-organising: when a document is added, new links are created between the new document and related documents, so that new information, residing in the grid topology, is created. Ingrid is still at an experimental state, and does not recognise SGML documents, so an important task here would be to render Ingrid "SGML-aware".

Another, less distributed, solution used in Harvest [19], is to collect the indexing data in a distributed way, but to handle the search itself from a centralized database with all indexing data.



Ingrid and Harvest are both divided into 3 programs: the publisher (or gatherer), which reads a file, extracts the meta-information, puts them in a specific file and sends this file to the servers; the navigator, which provides a user interface for a query and manages the search by querying the servers; and the server (or broker), which manages a metadata database and processes queries on it.

An alternative solution, not at all distributed, would be for example to construct a cross-correlation matrix on a super-computer: this is the centralized approach, and the problem here is to deal with billions (or more) documents, whereas the distributed approach avoids this problem.

4. Some Current Work

4.1. AML Version 1.0

Version 1.0 of AML aims at providing a working system for astronomers. Other goals for AML are: easy conversion to and from other formats like FITS [20], easy and flexible new datatype definition, precise metadata, customization facilities, and compatibility with XML and SGML parsers and browsers. AML has all the necessary properties to become a widely-used standard.

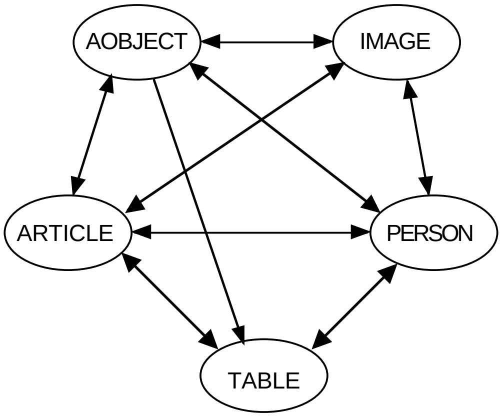
4.2. A user interface for searching and browsing AML documents

Applications to read and modify SGML files already exist. However, interfaces to display and manage queries for SGML files are very limited. It is also advantageous to have a specific interface for browsing and searching SGML files having a specific DTD like AML. This is also something to be created, in a way which could be reused to create other DTD-specific interfaces.

Java is currently the best programming language for creating browsing interfaces, because Java applets can be loaded by Web browsers, Java is not platform-dependent, it is possible to use CORBA with it (for example with the Java Development Kit jdk1.2), and it is a good object-oriented language. The only drawback of Java is its speed, but this is not important for the user interface. Complex processing like astronomical image processing can be done on local servers with faster platform-dependent languages communicating with the applets.

4.3. Knowledge discovery using the AML links

An AML document can include links to other AML documents, with their URLs. For instance, an article can have links to its authors, or to the astronomical objects it relates to. These links can then be used to discover relationships between different AML objects, by the use of clustering techniques. They can also be used for an extended search: a user can be interested in the documents on a certain type of object, related to another given object (for example, the authors having co-written an article with someone, or the articles indirectly linked with another article).



5. Conclusion

The tasks to be carried out in the immediate future are clear. As mentioned, the aim is to have an infrastructure which can be effectively and efficiently processed by human agents or by intelligent software agents. We are seeking to make use of a lot of the semantic structure in our data, through metadata, and not to ignore it as other systems have done in the past. Astronomical data and information is not limited to textual and document materials, even if these are undoubtedly very important. Therefore we must cater for multimedia data right from the start. The great landmasses of knowledge discovery and of data mining can be clearly demarcated in the background. These are the ultimate goals, and motivation, for this work.

References

[1] Astrophysics and Algorithms: Workshop on Massive Astronomical Data Sets May 6-8 1998, Princeton University
`http://dimacs.rutgers.edu/Workshops/Astro`

[2] SIMBAD: Set of Identifications, Measurements, and Bibliography for Astronomical Data

`http://simbad.u-strasbg.fr/Simbad`

[3] NED: NASA/IPAC Extragalactic Database

`http://nedwww.ipac.caltech.edu/`

[4] LEDA: Lyon/Meudon Extragalactic Database

`http://www-obs.univ-lyon1.fr/leda/home_leda.html`

[5] VizieR: CDS service for astronomical catalogues

`http://vizier.u-strasbg.fr/cgi-bin/VizieR`

[6] ADS: The NASA Astrophysics Data System

`http://adswww.harvard.edu/`

[7] GLU: Uniform Link Generator

`http://simbad.u-strasbg.fr/demo/glu.html`

[8] CGI: Common Gateway Interface

`http://hoohoo.ncsa.uiuc.edu/cgi/overview.html`

[9] ASU: Astronomical Server URL

`http://vizier.u-strasbg.fr/doc/asu.html`

[10] F.D. Murtagh, 1997, The technologies which underpin electronic publishing, in A. Heck, ed., *Electronic Publishing for Physics and Astronomy*, Dordrecht: Kluwer, 31-40.

[11] A Syntax for Dublin Core Metadata

`http://www.uic.edu/~cmsmcq/tech/metadata.syntax.html`

[12] Warwick Framework

`http://www.ukoln.ac.uk/metadata/resources/wf.html`

[13] GILS: Global Information Locator Service

`http://www.usgs.gov/public/gils`

[14] ADDSIA: Access to Distributed Databases for Statistical Information and Analysis

`http://www.ed.ac.uk/~addsia`

[15] SGML: Standard Generalized Markup Language

`http://www.sil.org/sgml/sgml.html`

A Gentle Introduction to SGML

`http://etext.virginia.edu/bin/tei-tocs?div=DIV1&id=SG`

[16] XML: Extensible Markup Language

`http://www.sil.org/sgml/xml.html`

[17] CORBA: Common Object Request Broker Architecture

`http://www.acl.lanl.gov/CORBA/`

[18] Ingrid: World-Wide Information Discovery

`http://www.ingrid.org`

[19] Harvest: Information Discovery and Access System

`http://harvest.transarc.com/`

[20] FITS: Flexible Image Transport System

`http://www.cv.nrao.edu/fits/`

6. Other acronyms

- AML: Astronomical Markup Language
  `http://www.infm.ulst.ac.uk/∼damien/these`
- CNIDR: Center for Networked Information Discovery and Retrieval
  `http://www.cnidr.org`
- DTD: Document Type Definition (see SGML [15])
- HTML: HyperText Markup Language
- HTTP: HyperText Transfer Protocol
- NC: Network Computer
- Z39.50: a standard for textual data indexation
  `http://lcweb.loc.gov/z3950/agency/`