Astronomical Data Analysis Software and Systems VIII
ASP Conference Series, Vol. 172, 1999
D. M. Mehringer, R. L. Plante, and D. A. Roberts, eds.

The Distributed Astronomical Preprint Service

J. E. Huizinga, Robert J. Hanisch, H. E. Payne, R. L. Williamson

Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218

Abstract. We have developed a prototype distributed preprint service. The system is hierarchical, with identical software at each node. Site maintenance tools written in Java allow information providers to enter new preprint metadata and to update them when preprints are published. A web site at each node is built on-the-fly from the metadata, allowing both browsing and searching. Parent servers can gather metadata from child servers, allowing the central server to index everything, and to notice new entries. Queries return preprint identifiers ("prepcodes") which resolve to on-line journals in the case of published preprints.

1. Introduction

Most new peer-reviewed astronomy research literature is now available in electronic form on the World Wide Web. Electronic journals offer features not possible in paper versions: searches, cross-reference links, forward references, and machine readable tables. Electronic submission has streamlined journal production, and dramatically reduced the time between acceptance and publication. Agreements between the publishers of the major astronomy journals allow links from one journal to another, in a collaboration called *Urania*[1].

Against this background, the astronomical preprint has made a much less successful transition to electronic distribution. Preprints are the historical channel for rapid communication of results in astronomy. Preprints have traditionally been produced in an institutional framework to demonstrate an organization's quality and vitality. But isolated institutional preprint web sites make it difficult for a user to discover new and interesting preprints. The centralized Los Alamos National Laboratory astrophysics preprint archive[2] is one solution of the discovery problem. But this service remains to be integrated into the on-line astronomical literature identified with *Urania*.

We are developing a distributed system for integrating preprint collections maintained at participating institutions. Institutional participation provides the best assurance of correctness and currency, distributes the workload and resource requirements, and satisfies an institution's desire to display its own achievements. The system will allow users to locate documents anywhere in

[1] `http://www.aas.org/Urania/`

[2] `http://xxx.lanl.gov/archive/astro-ph/`

the collection by means of a single query, or to search through the holdings at a single institution. The system integrates preprints into the on-line literature by tracking a preprint through to final publication: a hyperlink to a preprint will resolve to the on-line final paper, once it is published. The system could eventually be integrated with journal production by providing manuscripts for referees and production staff. In return, journals could automatically supply the final URL of each preprint. At the moment, however, we are working on a system to facilitate and preserve the preprint tracking being done by hand by librarians.

2. A Prototype System

Our prototype preprint service is a hierarchical system. A central authority tracks participating institutions. Each institution can maintain a number of document collections, and may also grant authority over parts of its "preprint namespace" to subordinate or "child" servers, at the departmental level, say, and so on. Each node in the prototype system has identical software, although "lite" sites are also envisioned.

The practical key to the *Urania* system was a standard naming scheme, known as "bibcodes," for journal articles. Our preprint identifiers, or "prepcodes," reflect the hierarchical nature of the system, and are modeled on the HANDLE[3] system, used by the NCSTRL[4] project, for example. Preprint identifiers consist of four components:

`astro.pp.stsci.stsci1281`

discipline doctype institution prepid

The first field allows for collections in other disciplines. Doctypes might include other "gray literature" categories, like technical reports and observatory manuals. The institution code is assigned by the central authority, to avoid duplication, while the prepid field is largely arbitrary.

Metadata for each document are keyed by preprint identifier. Metadata include a title, authors and affiliations, an abstract, pointers to author contributed TeX and PostScript files, if present, and the location of the final published paper. The metadata are indexed for searching, and used to dynamically generate most web pages. Metadata are in a format that is nearly XML. Other metadata files describe each collection, and encapsulate parent-child relationships.

At each node of the system, a name resolver provides access to the metadata associated with each prepcode in its assigned name space. Prepcodes at child nodes are passed down the hierarchy to child servers, and unrecognized prepcodes are passed up the hierarchy to the parent server. The central server forwards valid prepcodes to the appropriate institutional server, and catches invalid prepcodes.

---

[3] `http://www.handle.net/`

[4] `http://www.ncstrl.org/`

3. Prototype System Components

We envision a system with four components: (1) tools that information providers can use to update and maintain their document collections, (2) tools for creating, maintaining, and searching the indexes, (3) a user query interface, and (4) a notification service. Our prototype system contains versions of the first three:

Site Maintenance and Preprint Entry Tools. Preprint entry tools have been implemented in Java for use as applications or as Web applets. They allow easy entry of new preprints into an institute's collection, and allow entries to be updated, e.g., to reflect publication. Web site pages are generated on-the-fly from the metadata. Page layout is template-driven, allowing easy customization of the appearance at each institution. In the prototype, page generation is handled by Perl CGI scripts, with metadata maintained in DBM databases. Conversion to Java servlets and a database like MySQL are being considered.

Index and Search Tools Preprint metadata are indexed for fielded and full text searches with CNIDR's Isite[5] package, familiar to us from our ASDS (Hanisch, Payne, & Hayes 1994) project. A search returns preprint identifiers of matching documents, which are passed to a name resolver to obtain the document. The name resolver is implemented as a Perl CGI script. Prepcode passing is accomplished by having the name resolver return HTTP redirect instructions to the user's web browser.

A parent server can collect metadata from its children, allowing it to create an index spanning all collections in its name space, and allowing the central authority to index all collections. A user viewing an institutional site and wishing to search a larger collection of documents is directed to the search page at the parent site. This mechanism could be the basis of a notification service, which would take note of new items.

User Query Interface The prototype query interface is similar to our ASDS query interface, and makes use of Isite's "virtual database" concept, which allows a single query to be presented to a set of physically distinct index databases as if they were a single database. At each level of the system hierarchy, an index is constructed for each local collection and for each child site. A dynamically created search page gives users the option of searching all of these databases or any one of them.

The overall architecture is sketched in Fig. 1. The "Top Server" will have many child "Institute Servers," each of which can also be a parent of servers at a lower (departmental, for example) level. That is, the diagram can be expanded both horizontally and vertically, as indicated by dots. Arrows indicate the exchange of metadata and name resolver requests between parent and child, and the user's ability to query the central server or local servers. Site management tools are shown in the upper right of the figure.

---

[5] `http://www.cnidr.org/ir/isite.html`

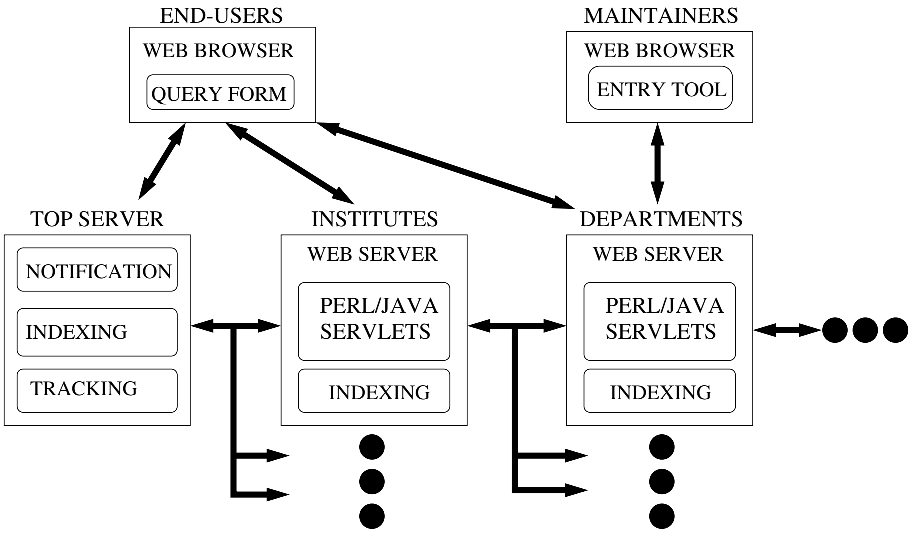

Figure 1. The architecture of a distributed preprint service.

4. Future Work

The prepcode passing between name resolvers is not robust against an intermediate resolver being down. We are considering exchanging metadata about the collections, so that each name resolver can pass prepcodes directly to the responsible server, perhaps using GLU (Fernique, Ochsenbein, & Wenger 1998). Other concerns include ensuring that the system will work in the Windows NT environment, used by many librarians (GLU and our customized Isite are currently unavailable on this platform), database issues, and an all-Java implementation.

Acknowledgments. This project is funded by the NASA Applied Information Systems Research Program (NRA 96-OSS-10) under grant NAG5-3942.

References

Fernique, P., Ochsenbein, F., & Wenger, M. 1998, in ASP Conf. Ser., Vol. 145, Astronomical Data Analysis Software and Systems VII, ed. R. Albrecht, R. N. Hook, & H. A. Bushouse (San Francisco: ASP), 466

Hanisch, R. J., Payne, H. E., & Hayes, J. J. E. 1994, in ASP Conf. Ser., Vol. 61, Astronomical Data Analysis Software and Systems III, ed. D. R. Crabtree, R. J. Hanisch, & J. Barnes (San Francisco: ASP), 41