Astronomical Data Analysis Software and Systems IX
ASP Conference Series, Vol. 216, 2000
N. Manset, C. Veillet, and D. Crabtree, eds.

SyNAPS: System for Networking Astronomical Publication Services

H. E. Payne, J. E. Huizinga, R. J. Hanisch

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

Abstract. We are developing a Java/XML based distributed electronic preprint service. The goal is to create the tools and infrastructure necessary to allow (astronomical) institutes to create and maintain individual preprint collections while presenting end-users with an integrated preprint system and notification service.

1. Introduction

Most peer-reviewed astronomical journals are now available on the World Wide Web. These electronic journals offer functionality not available 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.

In contrast to the published literature, institutional preprint series have not flourished on the WWW. Some institutes moved their preprint series to the WWW to reduce costs, only to see it perish for lack of visibility. These preprint services were never integrated into an on-line "*new preprints rack*" for easy browsing, or seeing at a glance what was new. The Los Alamos National Laboratory astrophysics preprint archive (astro-ph)[1] provides a highly successful alternative consisting of a single, centralized repository where individual authors can post their work. Together with a notification service, this has become a very useful tool in the fast distribution and dissemination of new results.

However, since preprints are submitted to astro-ph by individual authors, they have lost their institutional association. This has had two effects: institutes have lost a highly visible part of their identity, and readers have lost the imprimatur of institute imposed standards, e.g., requiring preprints to have been accepted for publication in a peer-reviewed journal. Since astro-ph imposes no such restrictions, it has become a hodgepodge of traditional preprints, conference papers, and research notes.

We are developing SyNAPS as an alternative electronic preprint service, built by unifying distributed institutional services (Hanisch et al. 1998). This service presents users with a uniform interface for browsing and searching, and a central notification service, while naturally distributing the maintenance load, and preserving institutional identities and standards. We are also investigating

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

tracking preprints to their final published form, to integrate preprints with the rest of the on-line literature.

While the timely, ephemeral nature of preprints imposes some (interesting) requirements, SyNAPS is well suited to other "gray literature" areas, such as technical reports, observatory reports, and documentation.

2. System Design

SyNAPS is a hierarchical system. A central server acts as a naming authority, granting control over portions of the document identifier namespace to participating institutional servers. The central astronomical preprint server controls the `astro.pp` namespace. The NRAO institutional server controls the `astro.pp.nrao` namespace, and might assign its 1001^{st} preprint the fully qualified identifier (preprint code or "prepcode") `astro.pp.nrao.prep1001`. Similarly, institutional servers can grant control over parts of its namespace to departmental servers, say, and so on. This hierarchy could also be extended upwards to encompass other astronomical services and services in other disciplines. The central server can also maintain an "at large" collection to support authors who might use astro-ph as the repository for their preprints.

Every server knows its parent, its local collections, and its children. To prevent a single point of failure at the top, institutional servers know about each other. At each site, a name resolver routes requests for a given prepcode to the appropriate server, much like domain name resolution.

Each document collection consists of resources available from a Web server. The current system makes explicit provisions for LaTeX, PostScript, Portable Document Format (PDF), and HTML versions of a document. Each document is described by metadata that include title, authors, and abstract. The metadata are used to generate, on-the-fly, the WWW interface to the collection. And the metadata are indexed for fielded searching.

All metadata are passed up the hierarchy. Each site holds all the metadata from servers below it in the hierarchy, and the top server has everything. In this way, we can easily support a central search and notification service. Only requests for preprint source material need to be routed to the appropriate server. Parents can ask their children to send them anything new, and children can signal their parent that they have something new, and these signals can propagate up and down through the entire hierarchy. Metadata are exchanged in XML format over the HTTP protocol.

3. System Components

Every node in the SyNAPS system consists of a Web server, a relational database for metadata, a suite of Java servlets, a site maintenance tool, and a search engine. Servlets handle: (1) communication with the site maintenance tool; (2) server-to-server communication; and (3) generating the web interface to the metadata.

3.1. Metadata

Metadata form the heart of the system. Preprint metadata include title, (contact) authors, e-mail addresses, affiliations, abstract, keywords, associated files, submitted/accepted dates, citation information, and creation and modification timestamps. There are metadata for each collection and each web server, as well. We chose the popular, free MySQL database for the metadata, and use JDBC to connect servlets to the database.

3.2. Preprint Entry and Site Maintenance Tool

The preprint entry and site maintenance tool is a Java client application running on the user's desktop, which communicates via HTTP with Java servlets running on the Web server. New preprint source files can be added to the Web server by hand, or by using the entry tool. In either case, a LaTeX parser extracts the metadata and translates the LaTeX markup to the equivalent XML and HTML. The results are presented in a graphical user interface for editing and verification. When satisfied, the maintainer can upload the metadata to the Web server, where they are stored in the database. It is also possible to modify the metadata for existing preprints, e.g., to enter the URL of the final published version, or to update an earlier draft.

3.3. Server-to-Server Communication

The SyNAPS system is dynamic. Servers communicate, passing metadata for new or modified preprints up, and sending queries and requests for updates down the hierarchy. Any node can ask the central server for the URLs of the institutional servers, and the central server can notify participants that it has moved. All communication is handled by the Java servlets.

3.4. Collection Browsing and Presentation

The SyNAPS network can be entered at each participating site. Users can choose to traverse the hierarchy up to the parent server or down to a child collection, to browse any local preprint collections, or to execute a search. A search can be broadened by passing the query up to the parent server.

All preprint listings and abstract information are generated on the fly from the metadata. All XML entities and markup are replaced with their HTML equivalents. The overall layout of the preprint listings and abstracts is determined by customizable templates, which allow individual institutes to retain their identity within the SyNAPS network.

3.5. Index and Search Tools

Instead of executing database queries in SQL, we are using the Isearch[2] search engine. For indexing, metadata are first parsed, to replace all XML entities with their ASCII equivalents, for more reliable searching. The search engine is run in a servlet Process object, and returns an XML fragment containing a "hit list" of prepcodes. This list is fleshed out with metadata from the databases

[2] `http://www.etymon.com/Isearch/`

and presented to the user as HTML. When a user wants to retrieve the data for one of the hits, a name resolver determines whether the preprint identifier corresponds to a local collection, or should be passed down the hierarchy.

In our previous ADASS contribution (Huizinga et al. 1998) we described a prototype with similar architecture and components. To summarize our progress, Perl CGI scripts were replaced with Java servlets, metadata were rewritten in XML format, and a real relational database replaced directories of metadata files. These changes made the system more robust, more flexible, and much more responsive.

4. Future Work

Our next goals are to repackage the Java code and to deploy some systems in the field on UNIX and Windows platforms. Work is under way to add further site management functionality, allowing users to notify parent servers of new preprints, to trigger a re-indexing for the search engine, to delete old preprint files, etc. We also need to address authentication/authorization control over the site maintenance functions.

We have had preliminary discussions with the American Astronomical Society in anticipation of a system under which the society would keep track of the prepcodes of submissions to its journals. For the first time, this would allow an automatic, reliable method for associating a preprint with the published article it became. This system would have a number of benefits. For SyNAPS users, a search for a preprint would find the published paper instead, even if the preprint had been retired from its SyNAPS server. This makes it trivial to find the proper, permanent citation information for a published preprint. Authors could be assured that citing a preprint by its prepcode would allow readers (and copyeditors) to follow the citation into the published literature. The overall result would be the integration of electronic preprints into the publication process and the on-line literature.

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

References

Huizinga, J. E., Hanisch, R. J., Payne, H. E., & Williamson, R. L. 1998, in ASP Conf. Ser., Vol. 172, Astronomical Data Analysis Software and Systems VIII, ed. David M. Mehringer, Raymond L. Plante, & Douglas A. Roberts (San Francisco: ASP), 255

Hanisch, R. J., Payne, H. E., Huizinga, J. E., Stevens-Rayburn, S., Bouton, E. N., Eichhorn, G., & Boyce, P. B. 1998, in ASP Conf. Ser., Vol. 153, Library and Information Services in Astronomy III, ed. U. Grothkopf, H. Andernach, S. Stevens-Rayburn, & M. Gomez (San Francisco: ASP), 127