Astronomical Data Analysis Software and Systems XIII
ASP Conference Series, Vol. 314, 2004
F. Ochsenbein, M. Allen, and D. Egret, eds.

The Chandra Bibliography Database

Arnold H. Rots, Sherry L. Winkelman, Sarah E. Blecksmith, John D. Bright

Harvard-Smithsonian Center for Astrophysics, 60 Garden Street MS 67, Cambridge, MA 02138, U.S.A.

Stéphane Paltani

Observatoire de Marseille

Abstract. Early in the mission, the Chandra Data Archive started the development of a bibliography database, tracking publications in refereed journals and on-line conference proceedings that are based on Chandra observations, allowing our users to link directly to articles in the ADS from our archive, and to link to the relevant data in the archive from the ADS entries. Subsequently, we have been working closely with the ADS and other data centers, in the context of the ADEC-ITWG, on standardizing the literature-data linking. We have also extended our bibliography database to include all Chandra-related articles and we are also keeping track of the number of citations of each paper. Obviously, in addition to providing valuable services to our users, this database allows us to extract a wide variety of statistical information. The project comprises five components: the bibliography database-proper, a maintenance database, an interactive maintenance tool, a user browsing interface, and a web services component for exchanging information with the ADS. All of these elements are nearly mission-independent and we intend make the package as a whole available for use by other data centers. The capabilities thus provided represent support for an essential component of the Virtual Observatory.

1. Introduction and Existing Capabilities

For the past two years the Chandra Data Archive (CDA[1]) has been building a database linking articles in the literature to Chandra datasets being presented in those articles as a service to our own users and to users of the ADS[2].

Currently, on the side of the CDA users are able to link from observations to articles in the ADS, while there are also some links to scattered publications. On the part of the ADS, users can link from articles presenting Chandra observations

[1] `http://cxc.harvard.edu/cda`

[2] `http://adswww.harvard.edu/`

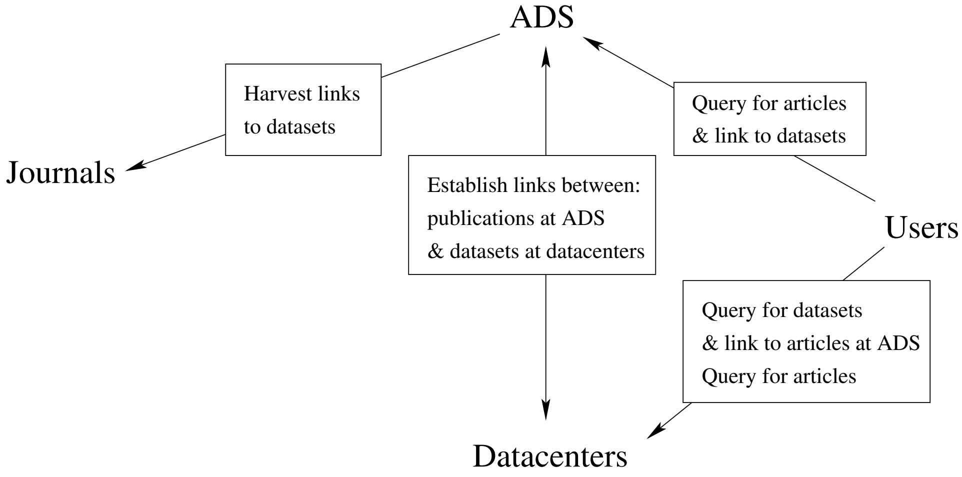
ADS
Harvest links to datasets
Journals
Establish links between: publications at ADS & datasets at datacenters
Query for articles & link to datasets
Users
Query for datasets & link to articles at ADS Query for articles
Datacenters

Figure 1. Planned query mechanisms to allow users flexible access to data and publications.

to those datasets in the archive and select articles with general Chandra mission tags.

These are valuable services but not as general as one would like and very labor-intensive to maintain: to date humans have inspected 11,000 articles to be judged on relevance and manually made 2400 links between the literature and the archive's datasets. The objective is to move to a system where the links can be harvested automatically and users are provided with access to data and published articles through the ADS and through the data centers as illustrated in Fig. 1. We will outline in the following sections the steps that will lead to such improved user-friendliness.

2. Identifiers and Automatic Linking

The ADS, the data centers, and the US journal editors are working on a proposed agreement that will enable authors to insert links to archived datasets directly into their manuscripts. These identifiers will be IVOA (International Virtual Observatory Alliance) compliant and consist of a name space (ivo:), an authority Id (ADS), a data collection (e.g., Sa.CXO), and a dataset name. For more detailed information, see Accomazzi & Eichhorn (2004).

The objective is to allow the data centers and the ADS to harvest the links between the literature and the archived datasets automatically, thus eliminating a large component of the labor that is currently invested in bibliographic databases associated with archives.

3. Database Extension

Whereas our database originally only contained links between datasets and the bibcodes of journal articles and conference proceeding papers that presented these datasets, we have been extending it in three directions. First, we now include five subject categories of papers: referring to specific observations (the original database); referring to published results; predicting Chandra results; referring to instrumentation, software, or operations; other. Second, we now include all other types of publications, with the exception of preprints. We would be prepared to include Astro-ph articles if the data identifier links could be harvested from that site; but it does not seem likely that this will happen anytime soon. Third, we are including a large variety of of attributes for each article in our database, such as: subject category; kind of publication (book, journal, proceedings, thesis, circular, review, newsletter, internal note); type of publication (article, abstract, memo, data, erratum, title only, electronic); number of citations; keywords; date of publication; refereed or not.

To this end we have expanded our database to 10 tables:

- **BibTable** is the main table that holds one record for each article with all simple attributes, some of them encoded.
- **ObsId** contains the links between bibcodes and single-observation datasets. Each ObsId may refer to more than one record in the BibTable and each record in the BibTable may refer to more than one ObsId record. ObsIds link to the observation catalog (a separate database) and to proposals for more information.
- **Subjects** contains the description of the five encoded subject categories.
- **Datasets** are conglomerates of observations that belong together and are represented by a single identifier, rather than (potentially) a large number of identifiers.
- **DatasetObsIds** enumerates the ObsIds contained in each Datasets record.
- **URLs** provides a mapping between BibTable entries and articles that are maintained as specific URLs.
- **StdKeywords** contains the standard keywords for each article.
- **StdKeywordCategories** holds the descriptions of the standard journal keyword categories.
- **StdKeywordList** is the list of canonical standard keywords as adopted by the major journals.
- **CustomKeywords** holds the keywords invented by individual authors.

We manage the entry of new records into the bibliography database through a dedicated database which is filled through automated queries to the ADS database. Attributes are filled in through a GUI and the entries are migrated to BibTable upon completion.

In addition, automated ADS queries update the number of citations for each article and check the continued validity of all bibcodes.

Table 1 provides some statistics on our database as of ADASS XIII.

4. Services

We are developing the following services:

- Exchange of information with the ADS, in particular the harvesting of Bibcode-Dataset Identifier pairs in both directions.
- Provide access to datasets through either a Dataset Identifier or a Bibcode.
- Provide information to the ADS on Bibcodes that are not related to specific observations.

Table 1. Number of articles in the CDA bibliography database as of early October 2003

Category	1999	2000	2001	2002	2003	Total	No. Cit.
Observations	53	284	485	485	352	1659	5639
Refer to obs.	9	94	333	499	322	1257	5300
Instr., s/w, ops.	34	141	124	69	18	386	1362
Predict result	11	67	21	14	21	135	306
Unclassified	15	90	70	29	40	244	663
Total	122	676	1033	1097	753	3681	13270
Total Reviewed	1011	2507	2735	2758	1859	10870	

- Provide access to publications through queries from our archive. We are developing a specialized literature query interface related to the Chandra mission, allowing users to search for articles on the basis of criteria that are specific to Chandra.
- Derive metrics through queries to the bibliography database (standardized metrics as well as custom requests).

5. Conclusion

We have developed a comprehensive database design that is capable of tracking almost all mission-related publications and preserving all relevant information. Added to this are a database and GUI that make maintenance, particularly data entry, as painless as possible. Our services include cross-linking with the ADS, literature search from our archive, and metrics.

The entire package is reasonably mission-independent and available to other data centers.

Acknowledgments. This work is supported by NASA contract NAS 8-39073 (CXC).

References

Accomazzi, A. & Eichhorn, G. 2004, this volume, 181