BibCat: The Chandra Data Archive Bibliography Cataloging System

S. Winkelman, A. Rots

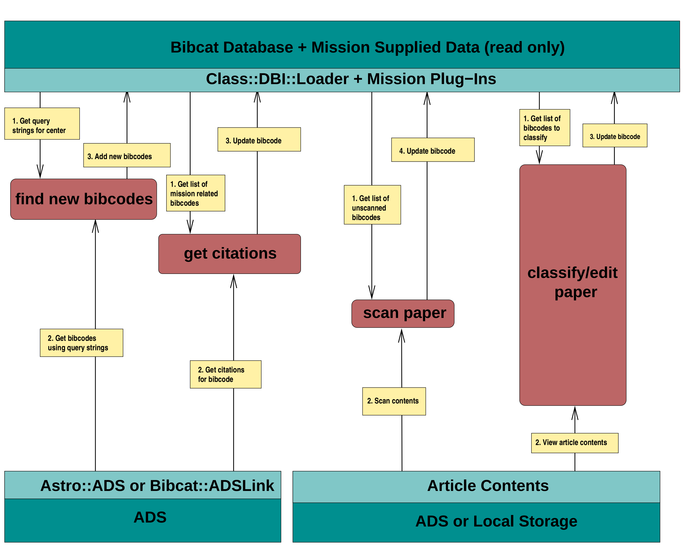
CXC/Smithsonian Astrophysical Observatory

INTRODUCTION

The Chandra Data Archive (CDA) has been tracking publications in refereed journals and on-line conference proceedings that are based on Chandra observations since early in the mission. Over the years this database and its associated tools have expanded dramatically. In this paper we describe our newly renovated bibliography architecture with an emphasis on new features which have been added including: auto-scan capabilities to reduce in an automated fashion the number of papers which need to be manually classified and to flag keywords (such as observatory names or surveys) used within papers; multi-user classification allowing quality assurance checks; multi-observatory capabilities allowing multiple facilities to use the same database independently; and plug-in support allowing access to associated observatory data to more fully describe data links in papers.

The usefulness of some of these features speak for themselves, but others are not so obvious. As an example, we intend to use the multi-observatory functionality to apply separate classification schemes to papers relating to the CDA and the Chandra Source Catalog and potentially to other observatories at the Center for Astrophysics. The data mining aspects of the auto-scanning capabilities can be used for many purposes such as: improving searching for Chandra related papers from both ADS and our bibliography search pages or linking papers to grants for internal uses.

Acknowledgment: This work is supported by NASA contract NAS8-03060.



Graphical representation of the basic steps and communication streams involved in classifying papers for a bibliography catalog.

FROM ADS TO BIBCAT TO USERS

The BibCat was designed to accomodate two goals of the CDA regarding Chandra-related papers: track papers for Chandra science statistics and flag information in papers for research purposes.

Find new papers: We use the ADS as our source for all potential publications, rather than going to each publisher independently. We need to retrieve new bibcodes and account for bibcodes which have changed names or have been retired from ADS.

Download papers: The method for getting the papers is publisher dependent and requires a library of code to access their holdings.

Convert papers: Papers are generally pdf or html. Converting the pdf's to text poses a number of problems with both text-flow issues and scanned image papers.

Scan paper: The converted papers need to be scanned for auto-elimination; meta-data collection; and possibly for a first guess at classification

Classify papers: Classification of papers requires people to peruse papers and then record their judgments in the database.

Record citation counts for papers: Citations counts need to be updated from ADS on a regular basis.

Interfaces to access database: Interfaces are needed at every step of the process. They encompass scripts for maintaining the database, GUI's for classifying bibcodes and maintaining the database, web interfaces for searching our database, and SOAP services for ADS to harvest information from our database. Our current search interface can be found at http://cxc.harvard.edu/cgi-gen/cda/bibliography

BIBCAT REQUIREMENTS

The BibCat is a combination of a database and tools for populating and accessing the data. The database provides flexibility to the classification process. This is key to maintaining a bibliographic catalog.

- Accommodate multiple observatories/data centers for observatory-related meta-data collection
- Accommodate multiple classifications per paper per observatory for quality assurance
- Maintain complete tracking and history of bibcodes through every step of the process
- Link bibcodes to observatory-specific data
- Flexibility in adding new flags and lists of meta-data
- Ability to link papers to archival data
- Ability to plug-in observatory-specific modules which are recognized by the BibCat

THE DATABASE

To allow for the flexibility needed in the requirements, the database tables fall into three categories: 1) tables describing the physical publication including publication date, publication type, refereed, title, authors, keywords, and links to other papers (blue box); 2) tables describing the data center meta-data collection including definitions of the categories and flags attached to papers (green box); and 3) tables recording the classification of papers by individual classifiers (purple box).