EPJ Web of Conferences **186**, 04001 (2018) https://doi.org/10.1051/epjconf/201818604001
LISA VIII

Growing a Bibliography

Sherry Winkelman[1,⋆], *Arnold* Rots[1], *Raffaele* D'Abrusco[1], *Glenn* Becker[1], *Sinh* Thong[1], and *Michael* McCollough[1]

[1] *Harvard-Smithsonian Center for Astrophysics*

Abstract. The Chandra Data Archive (CDA) has been tracking publications based on Chandra observations in journals and on-line conference proceedings since early in the mission. Our goals are two-fold: 1) provide a means for Chandra users to search literature on Chandra-related papers to further their scientific research; and 2) provide a means for measuring the science produced from Chandra data. Over the years the database and its associated tools have expanded dramatically. In this paper I will give a history of the development of the bibliography with a focus on the human capital involved, along with the skill sets and management structures developed which allow us to maintain a very rich and extensive bibliography with a limited number of full time employees (FTEs). I will also cover how the diverse metadata collected has made the Chandra bibliography an essential resource in managing the Chandra X-ray Center.
This work has been supported by NASA under contract NAS 8-03060 to the Smithsonian Astrophysical Observatory for operation of the Chandra X-ray Center. It depends critically on the services provided by the ADS.

1 Introduction

The Chandra X-ray Observatory (CXO) was launched on July 23, 1999 and is the third of NASA's Great Observatories. It has two instruments which can be combined with two gratings to create 6 instrument modes. Chandra's data archive is a repository for Chandra data as well as a source of information for all Chandra-related activities such as: proposal submission and management; mission planning; data processing, including the Chandra Source Catalog; and data retrieval. The Chandra archive operations group is responsible for ensuring the integrity of data in the CDA; maintaining the Chandra Observation Catalog; curating the Chandra Bibliography; and tracking downloads of Chandra data. The focus of this paper is the growth, development, and use of the Chandra bibliography: Section 2 describes the development of the Chandra bibliography; Section 3 reviews the management tools and skills sets needed to develop and curate the bibliography; Section 4 explores some of the ways the Chandra bibliography has been used; and Section 5 gives some suggestions of what will be next for the Chandra bibliography.

2 Chandra Bibliography

The purpose of the Chandra Bibliography has always been two-fold: 1) provide a means for astronomers to search literature on Chandra-related papers to further their scientific research; and 2)

⋆e-mail: swinkelman@cfa.harvard.edu ORCID: 0000-0001-7354-6221

provide tools for measuring the science produced from Chandra data in the context of the astronomical literature as a whole. Since its inception, the Chandra Bibliography has linked Chandra data to the papers which analyze the data and shared the links with ADS. Throughout the mission, we have been curating the bibliography and providing database and application development for the bibliography with <= 1 FTE. A plot of FTEs required to develop and curate the Chandra Bibliography over the course of the mission is shown in Figure 1.

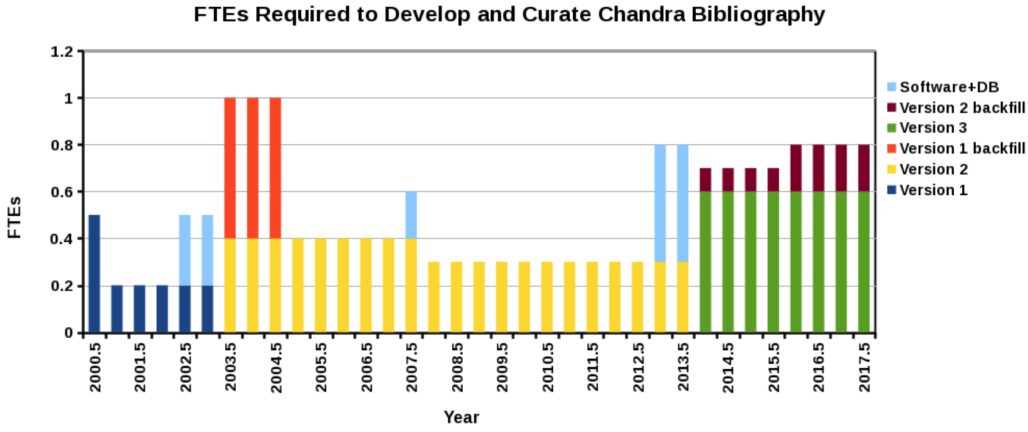

Figure 1. FTEs required to develop and curate the Chandra bibliography.

The bibliography has four basic categories of Chandra related papers. Each category has metadata which describes the Chandra relation within the paper in more detail. The general categories are:

- **Chandra Science Paper (CSP):** Chandra data contributes significantly to the science in the paper
- **Chandra Observatory Paper (CXO):** Chandra instruments, software, or operations
- **Miscellaneous Chandra Papers (Misc CXC):** Papers which refer to the Chandra X-ray Observatory in some way but do not rise to the level of CSP or CXO
- **Not Chandra Related:** Chandra/CXO is in the text of the paper, but the paper does not fit into any other category

Over the years the Chandra bibliography has undergone two major expansions. Each expansion in metadata was accompanied by new classification tools; added complexity to the bibliography database; and a backfill effort to bring the entire bibliography up to the new version. As can be seen in Figure 1, backfilling for Version 2 took 1.5 years operating at the 1 FTE level. Due to the size of the bibliography and the complexity of Version 3, the backfill efforts for this new version will be much greater. To date, we have spent an average of 0.17 FTEs per year in the backfill effort for four years and have completed the metadata for $\sim$ 10% of the backfill. We will need to enlist the help of more FTE's to complete this backfill effort.

The Chandra bibliography has a very rich set of metadata which are described in detail in Table 1 for each version of the bibliography. A GUI was developed to aid classifiers in compiling the metadata attached to papers. Figure 2 provides screenshots of several aspects of the current bibliography classification interface. The left panel shows the publication and status information window which also contains buttons for accessing the rest of the metadata options for a paper. The right panel shows the summary of all flags selected for the paper.

Table 1. Description of the three versions of the Chandra bibliography

Version 1: September 2000 - June 2004
Paper Related Flags:
pub_form: article
Chandra Related Flags:
CSP with direct analysis of data
Data links for all papers

Version 2: April 2003 - present
Paper Related Flags:
pub_form: title, abstract, article
pub_type: journal, proceeding, review, thesis, circular
refereed: Y or N
publication date as noted by ADS
citation count
date found on ADS
Chandra Related Flags:
CSP with direct analysis of data including data links and links to defined datasets
CSP with indirect analysis of data, but no data links
Misc CXC and CXO papers included
Instrument flags: ACIS, HRC, HETG, LETG, HRMA, PCAD, EPHIN
Software and operations flags
Chandra Source Catalog added in 2009
Chandra science context: theory, follow-up analysis, multi-observatory
Status Related Flags:
Person responsible for classifying
Status: in progress, published, request to author
Comment field

Version 3: October 2013 - present
Paper Related Flags: same as V2
Chandra Related Flags: Same as V2 plus
7-9 defined justification flags for each category plus ability to add in a custom justification
Chandra surveys used in paper
Status Related Flags: Same as V2 plus
a detailed history section to annotate the classification process
Data Related Flags: Attached to each ObsId
How data was located: ObsId given, ChaSer, contacted author, reference in paper, etc.
Where in paper lead was located: main body of text, table, figure caption, etc.
Type of analysis: direct, indirect: multiwave, indirect: follow-up, indirect: theory
Reference bibcode if ObsId location came from another paper
Acknowledgment Related Flags:
Grant source: CXC with (in)valid grant, NASA with (in)valid grant, SAO with (in)valid grant, etc.
Data: Chandra data, archive or source catalog
Team: Chandra Director's Office, Mission Planning, science team, etc.
Software: CIAO, ChIPs, Sherpa, Chart
Miscellaneous: support from CXC, acknowledgment with no Chandra connection, no acknowledgment section
Grant Related Flags: Attached to each grant
Grant number as stated in text
Official grant number
Related to proposal: direct, substantial, tangential, questionable, none
General Content Related Flags:
Content flags: data analysis, theory and/or computation, new survey or catalog, etc
Observatories by waveband
Publication Date Flags:
Publisher dates: received, revised, accepted, published, conference date
ArXiv dates: up four versions

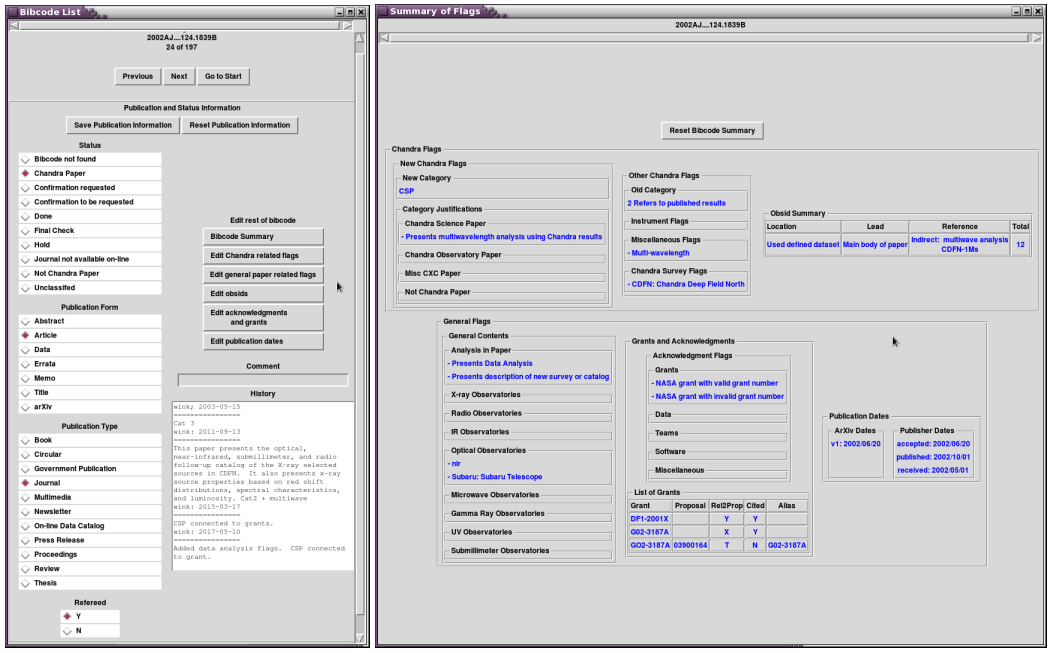

Figure 2. Screenshots of the bibliography editing GUI. The left panel is the publication and status information window. The right panel shows the summary window.

3 Management Tools and Skill Sets

Accuracy and consistency throughout the bibliography is critical. An essential component for maintaining that consistency as classifiers change is providing well documented metadata descriptions and classification procedures. Curators also need good tools to support their efforts and these tools may need to be provided by the curators themselves. We have developed applications which cover: preprocessing of bibcodes so we do not look at papers which are not relevant to us; classifying bibcodes so we can visualize the metadata we are attaching to papers; and post-processing of the bibliography to ensure data integrity and to produce metrics and other products useful to our users.

Data linking also requires good tools to search for data as they are described in the literature to match with the data as they are presented by the data repository. These tools are often available from the repository, but curators need to understand how to use those search tools effectively to match data descriptions in papers with data descriptions in the repository.

The most important skills for classifying and publishing papers are having: a good understanding of the tools available for curating papers; an acute attention to detail; and an understanding of the scientific process as it is described in astronomical papers. Applications development requires knowledge of: a programming language such as Perl, Python, or Java; database design and query structures; and web-application development tools. It is important that management understand and buy into the fact that both elements need to be supported in order for a bibliography to be maintained and effective.

In the 15+ years that the Chandra bibliography has been around, we have had twelve different classifiers/developers with the current Chandra bibliography team consisting of three classifiers and a manager/publisher/applications developer. Two of the classifiers can also provide application development support, while the manager/publisher also provides classification support when needed.

As our bibliography has grown and become more complex, we have found the following management structures to be of great use for curating the Chandra bibliography:

- defining separate roles for publishing and classifying papers
- automating actions with scripts whenever possible and as soon as possible
- providing feedback loops for maintaining and improving documentation and classification procedures
- setting reasonable schedules and priorities for backfill when adding new metadata fields
- consulting with a person with an astronomy research background to clarify how to classify papers which do not fit easily into your classification scheme
- advocating the inclusion of the bibliography in observatory tools and services
- anticipating future needs of users of your bibliography, particularly management, so you can prepare to handle more complex and detailed queries to your bibliography (they will come).

4 Uses of the Chandra Bibliography

As stated earlier, the Chandra Bibliography has two objectives: assist astronomers in their research and provide measures on the science output of the observatory. Implementation of these goals can be separated into three categories: interfaces; metrics; and assessing observatory program allocations.

4.1 Interfaces with the Chandra Bibliography

In addition to providing data links to ADS, the CDA offers a number of services to the astronomy community which tie into the Chandra bibliography: 1) the Bibliography Search Pages (http://cxc.harvard.edu/cgi-gen/cda/bibliography) for querying the bibliography (Figure 3); 2) a CXO Paper Repository (http://cxc.harvard.edu/cda/cxo_papers/cxo_papers.html) providing access to papers on the instruments, software and operations of the observatory (Figure 3); 3) the Chandra Proposal Search pages (http://cxc.harvard.edu/soft/propsearch/prop_search.html) to aid in proposal preparation (Figure 4); and 4) ChaSer (http://cda.harvard.edu/chaser/) for searching the CDA (Figure 4).

4.2 Metrics

We use the Chandra Bibliography to provide a number of metrics to various levels of management to give a sense of the science productivity of the observatory based on the number of Chandra Science Papers published, number of citations to those papers, etc. We have also developed metrics based on the data links in our bibliography. These metrics give us insight into the archival use of Chandra data and are less sensitive to observatory-specific characteristics than are the traditional metrics. Reference [1] gives a complete description of the data-centric metrics we have been using to measure speed of publication, fraction of observing time published, and archival usage. Recently we have been working on identifying high impact astronomy papers which are based on Chandra data and have a paper in this volume discussing our findings [3].

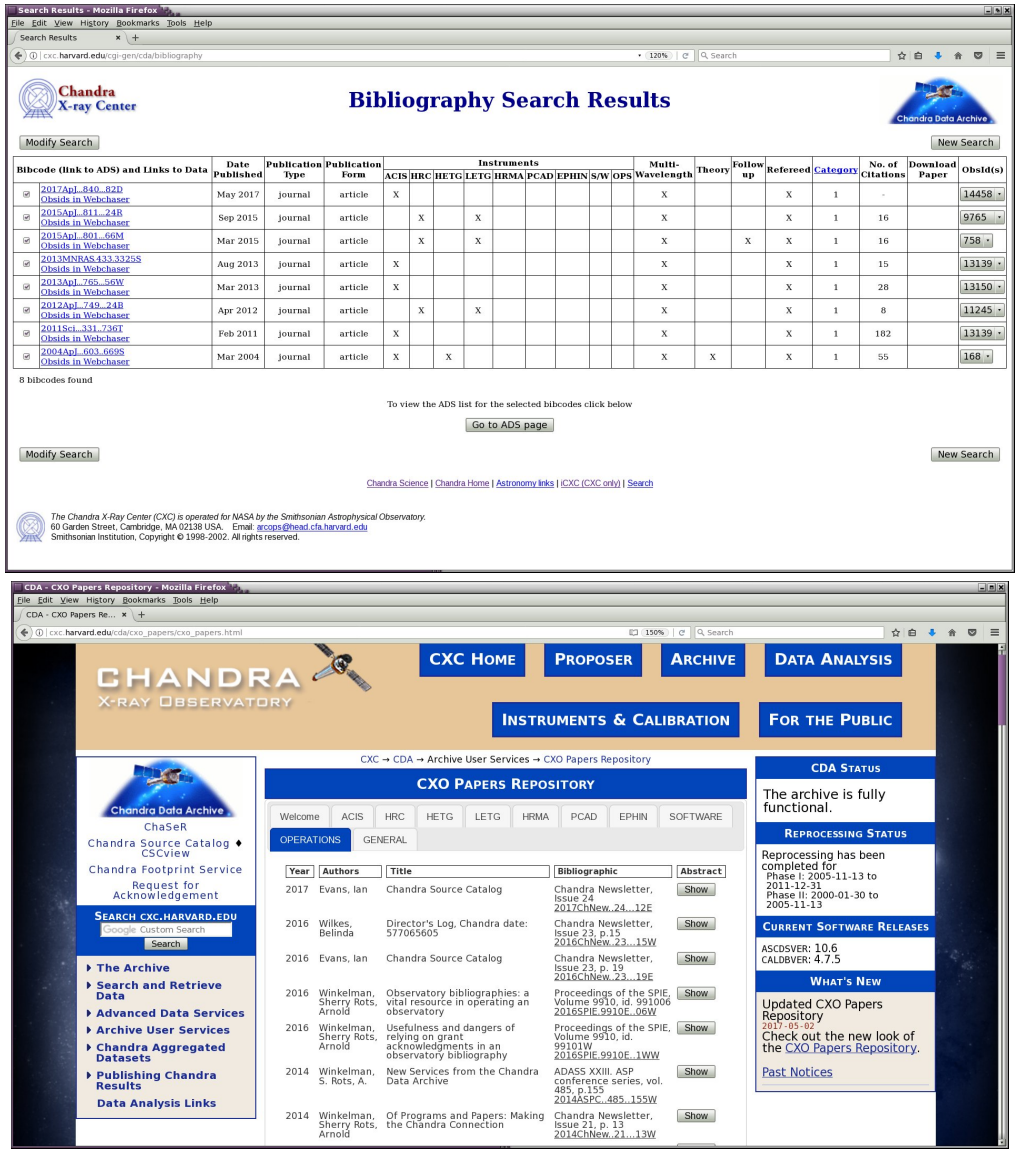

Figure 3. Screenshots of interfaces which access the Chandra Bibliography. The top panel shows search results from the Bibliography Search Pages for refereed journal articles which are CSPs with multi-observatory analysis of the Crab Nebula. The bottom panel displays the CXO papers focused on observatory operations.

4.3 Assessing Observatory Program Allocations

The CDA is routinely asked to provide input to the Chandra Users Committee (CUC), an advisory committee to the Chandra Director's Office representing the principal groups concerned with Chandra development and the general Chandra user community, regarding the productivity of various types of observing and funding programs offered by the CXC. The Chandra bibliography is heavily relied

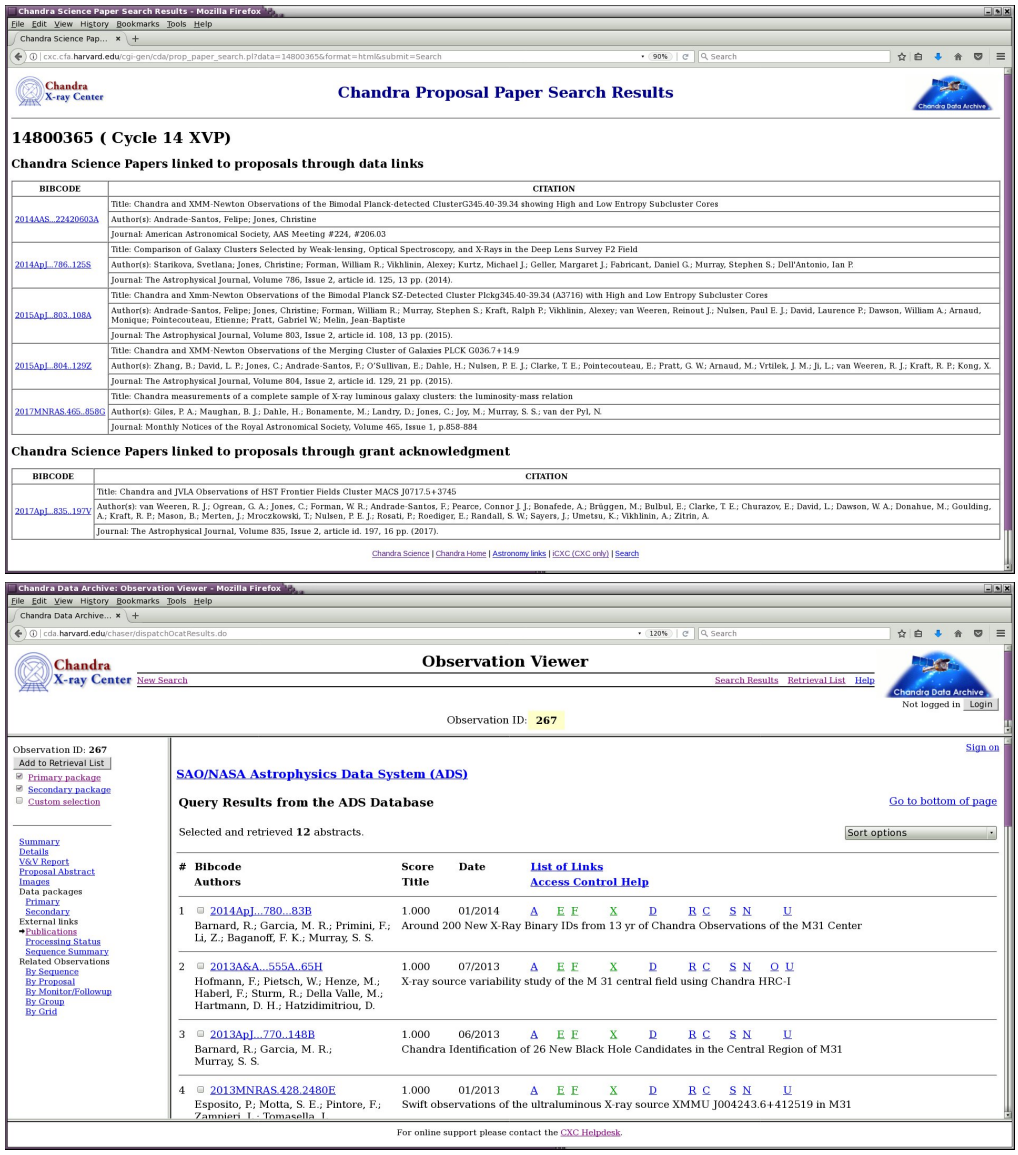

Figure 4. Screenshots of interfaces which access the Chandra Bibliography. The top panel shows search results from the Chandra Proposal Search pages for the XVP proposal for the Chandra COSMOS Legacy Program. The bottom panel displays the publications results within ChaSer for a search on M31 and then selecting ObsId 267.

upon for these assessments and in at least one instance the results lead to new metadata being added to the bibliography.

The first major assessment was with regards to the productivity of archive and theory awards granted by the CXC. These programs are money awards only, so data links were of limited use here. Our initial research focused on locating papers by searching ADS for Chandra archive and theory

grants. However, we were eventually given access to all grant numbers associated with all Chandra proposals. As a result, grant links are now a part of the bibliography and we can use this information for determining whether a paper is associated with an observing proposal. We have also found severe limitations in using grant information which are detailed in reference [4].

In observing cycles 13-16, the Chandra Call for Proposals solicited proposals for X-ray Visionary Projects (XVP) which described major coherent science programs to address key, high-impact scientific questions in current astrophysics. The time allocated to XVPs largely came from an increase in the total time available to the observatory due to changes in Chandra's orbit, so other time allocations were not affected. Future calls for XVPs will require a change in allocation to other observing categories. Because the XVPs are too young to have a reliable publication history, the CDA was asked to assess the science impact of XVP-proxies to provide insight into offering another XVP round. Our approach to the problem was to define aggregated observing programs from the archive to create pseudo-XVPs and using the Chandra bibliography to apply publication statistics to the aggregated sets. Details of the study are in reference [5].

5 What Next?

Expectations are that the Chandra Observatory will remain in operation for the next 10-15 years. As the archive grows, the needs of the astronomy community for accessing archival Chandra data will likely expand. As the observatory ages, program changes will likely be implemented. The Chandra bibliography will play a role in both of those developments. In anticipation of future needs, the CDA has a few projects planned to lead us into the next decade of Chandra science:

- Continue backfill efforts to migrate to Version 3 to take full advantage of the new metadata
- Add SIMBAD descriptions to the Observation Catalog to provide another set of flags for searching for Chandra data and Chandra papers
- Transition from Dataset Identifiers to DOIs [2] to aid in citing Chandra data
- Update Chandra Bibliography Search application to include expanded flags
- Add new flags for High Impact Papers [3]
- Explore crowd sourcing as a means of engaging citizen scientists to add flags to papers
- Migrate database infrastructure and applications to the Archive Development Team

6 Conclusion

In summary, it is possible to create and maintain a very rich bibliography for an observatory with <=1 FTE, but a diverse set of skills and tools are required. Processes should be automated where ever possible and as early as possible; clear documentation of metadata and classification procedures is essential; and feedback loops are necessary to fit the needs of classifiers.

As a bibliography grows in size and complexity, the uses of the bibliography will also expand. Interfaces for accessing the archive should link into the bibliography, making the bibliography an integral part of the observatory. As the observatory matures, the bibliography will be seen a resource for managing the observatory, so plan to handle more complex questions from upper management.

Once a bibliography is considered an integral part of an observatory, new projects will present themselves, sometimes at a far too rapid clip but that is part of the fun.

This research was supported by NASA contract NAS8-03060 (CXC) and relies heavily on the services provided by the Astrophysics Data System (ADS). We gratefully acknowledge the work by all past and current members

of the Chandra Data Archive Operations Team, who have helped build and maintain our bibliographic database. Without their efforts there would be no bibliography.

References

[1] A. Rots, S. Winkelman, G. Becker, Publications of the Astronomical Society of Pacific **124**, 391–399 (2012).

[2] A. Rots, R. D'Abrusco, S. Winkelman, *A Model for Using DataCite DOIs in Observatory Bibliographies*, These Proceedings (2018)

[3] S. Winkelman, A. Rots, R. D'Abrusco, *Evaluating High Impact Papers: Are We Missing Something?*, These Proceedings (2018)

[4] S. Winkelman, A. Rots, *Usefulness and dangers of relying on grant acknowledgments in an observatory bibliography* in *Observatory Operations: Strategies, Processes, and Systems VI* (2016) Vol. 9910, p. 99101W

[5] S. Winkelman, A. Rots, *Observatory bibliographies: a vital resource in operating an observatory* in *Observatory Operations: Strategies, Processes, and Systems VI* (2016) Vol. 9910, p.991006