The AstroBibPile: Building a Dataset to Support AI-enabled Bibliography Curation efforts

Alberto Accomazzi

IVOA Interop | 22 May 2024

CENTER FOR ASTROPHYSICS HARVARD & SMITHSONIAN

SciX [SciXplorer.org]

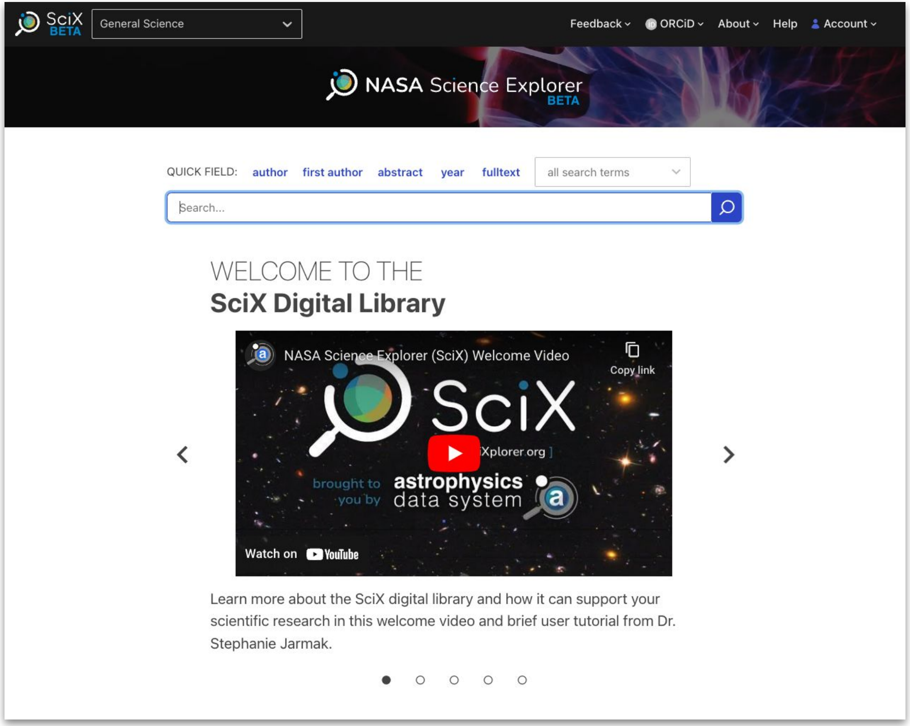
astrophysics data system

What is the NASA Science Explorer?

NASA SciX is a literature-based, open digital information system covering and unifying the research disciplines funded by the NASA Science Mission Directorate. It represents an extension of the NASA Astrophysics Data System to include all literature relevant to NASA Science research.

SciX supports NASA's Open Science efforts and enables interdisciplinary research and collaboration.

SciX currently indexes 20M articles, 260k data and software records, and provides links to almost 500k data products



https://SciXplorer.org

Context

The NASA Science Explorer (SciX) is primarily a literature database. SciX does not aim to be an index for all research data products, but rather **make relevant data products discoverable from the literature**, whenever feasible, through citation or data links.

Some types of data which are of most interest to SciX:

- Datasets “close” to publications, either as DBF, supporting archival links, or citations, as they **supplement** the science presented therein; examples include VizieR catalogs, text-mined Zenodo links, archival data links, data citations
- Reference catalogs, collections, and services, which are **highly used and cited**; examples include 2MASS, WISE, CSC, etc.
- Software records either mentioned or cited in scientific publications.

Two Strategies for Metadata Enrichment - Curation

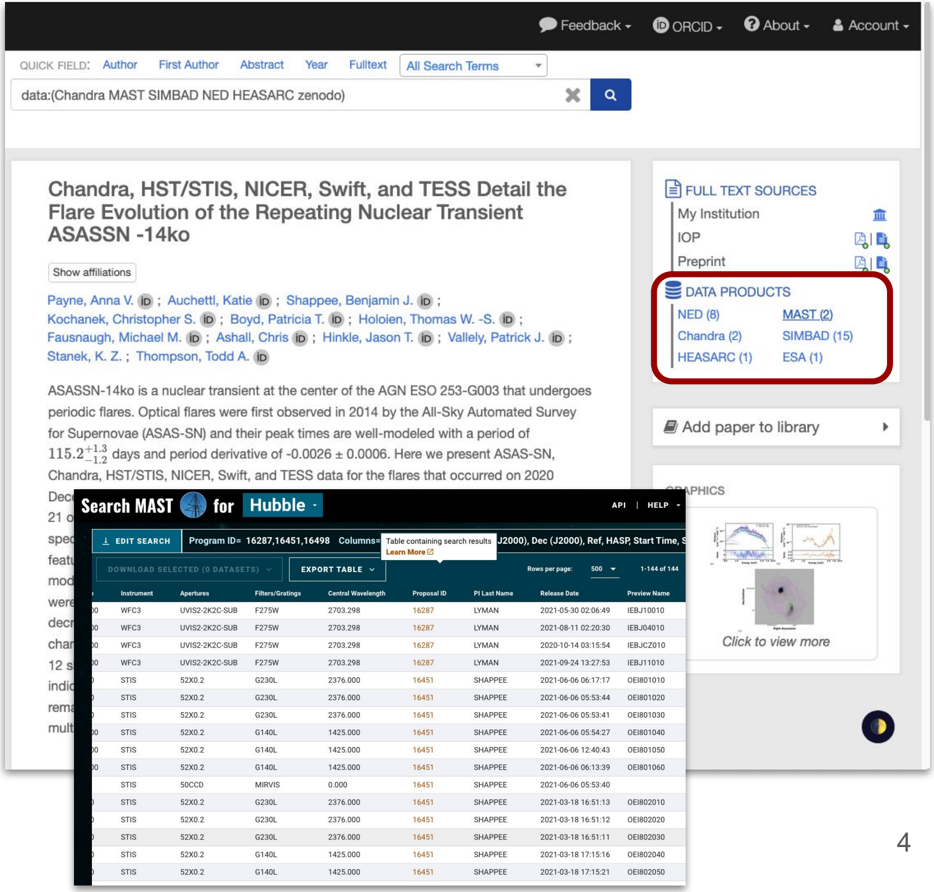
Curation of Bibliographies

ADS has been aggregating and exposing connections between bibliographic records and data products which are curated by librarians and archivists.

The largest contributors to this effort are projects in astrophysics which track astronomical objects (SIMBAD and NED), data catalogs (Vizier), and archives (Chandra, MAST, ESA, NOIRLab, etc.).

This provides a way to enrich new and existing bibliographic records whenever associated data is identified or entered in a knowledge base.

Thanks to librarians and archivists for enabling this capability!



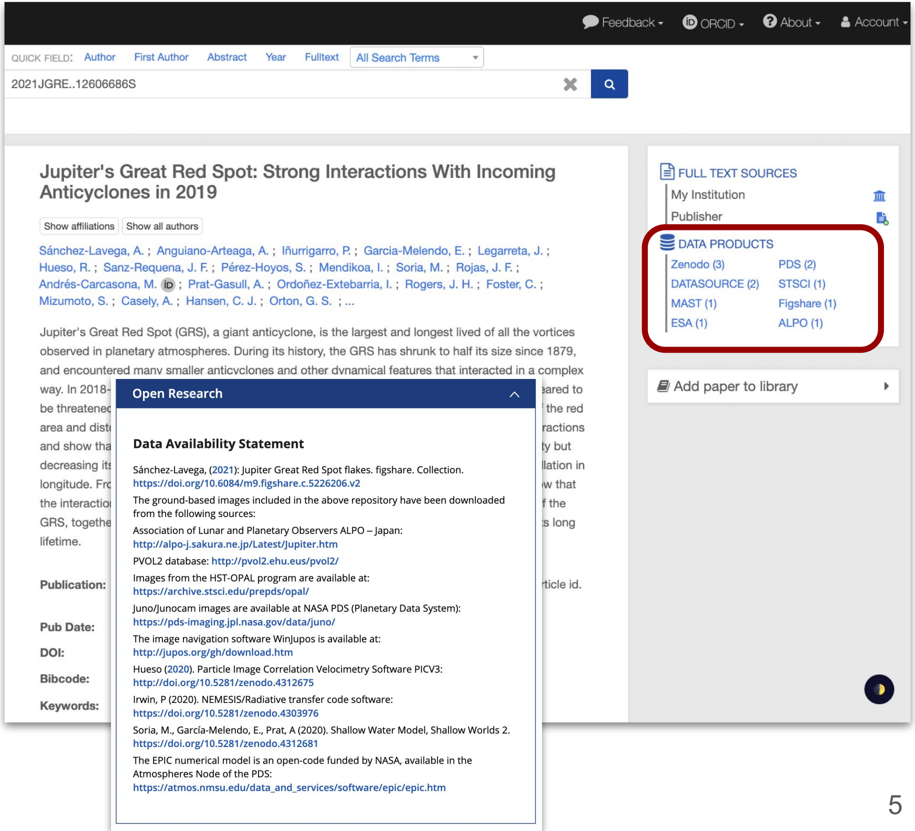
Two Strategies for Metadata Enrichment - TDM

Text Mining of the Literature

SciX obtains and processes the full-text of all papers in its database, maintains a citation database, and mines links to data products.

SciX detects the citation (in a reference list) or mention (in a data availability statement) of a software or data product, and records it in its database.

This helps, but doesn't replace, the curation work described earlier, which requires human evaluation of the content and context in which data products are mentioned in the literature.



Curation Workflow (Observatory Bibliographies)

Current Process:

1. Identifying candidate Publications through search of ADS/CrossRef
   a. Scope of journals being considered
   b. Refereed vs. non-refereed publications
2. Evaluation of Publications for Inclusion
   a. Science Papers ≅ use of data
   b. Engineering Papers ≅ instruments
   c. Non-science Papers ≅ mention of data

"There is tremendous diversity in the ways bibliographers track publications and maintain databases, due to parameters such as resources (personnel, time, budget, IT capabilities), type of observatory, historical practices, and reporting requirements to funders and outside agencies." (Observatory Bibliographers Collaboration 2024, arXiv:2401.00060)


arXiv:2401.00060v1 [astro-ph.IM] 29 Dec 2023

Draft version January 2, 2024
Typeset using LaTeX default style in AASTeX631

Assessing Your Observatory's Impact:
Best Practices in Establishing and Maintaining Observatory Bibliographies

Observatory Bibliographers Collaboration

Raffaele D'Abrusco,[1] Monique Gomez,[2] Uta Grothkopf,[3] Sharon Hunt,[4] Ruth Kneale,[5] Mika Konuma,[6] Jenny Novacescu,[7] Luisa Rebull,[8] Elena Scire,[9] Erin Scott,[1] Donna Thompson,[10] Lance Utley,[11] Christopher Wilkinson[7] And Sherry Winkelman[1]

[1] Chandra Data Archive, Chandra X-ray Center (CXC) / Center for Astrophysics, Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
[2] Instituto de Astrofísica de Canarias (IAC), Vía Láctea, s/n E-38205, La Laguna - Tenerife, Spain
[3] European Southern Observatory (ESO), Karl-Schwarzschild-Straße 2, 85748 Garching bei München, Germany
[4] NSF's NOIRLab, 950 N Cherry Ave, Tucson, AZ 85719, USA
[5] National Solar Observatory (NSO)/DKIST, retired, 3665 Discovery Drive, Boulder, CO 80303, USA
[6] National Astronomical Observatory of Japan (NAOJ), 2-chōme-21-1 Ōsawa, Mitaka, Tokyo 181-8588, Japan
[7] Space Telescope Science Institute (STScI), 3700 San Martin Dr., Baltimore, MD 21218, USA
[8] Infrared Science Archive (IRSA), IPAC, MS 100-22, Caltech, 1200 E. California Blvd, Pasadena, CA 91125, USA
[9] IPAC, Mail Code 314-6, Caltech, 1200 E. California Blvd. Pasadena, CA 91125, USA
[10] ADS / Center for Astrophysics, Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
[11] National Radio Astronomy Observatory (NRAO), 520 Edgemont Road Charlottesville, VA 22903-2475, USA

ABSTRACT

Observatories need to measure and evaluate the scientific output and overall impact of their facilities. An observatory bibliography consists of the papers published using that observatory's data, typically gathered by searching the major journals for relevant keywords. Recently, the volume of literature and methods by which the publications pool is evaluated have increased. Efficient and standardized procedures are necessary to assign meaningful metadata, enable user-friendly retrieval, and provide the opportunity to derive reports, statistics, and visualizations to impart a deeper understanding of the research output.

In 2021, a group of observatory bibliographers from around the world convened online to continue the discussions presented in Lagerstrom (2015). We worked to extract general guidelines from our experiences, techniques, and lessons learnt. This paper explores the development, application, and current status of telescope bibliographies and future trends. The paper briefly describes the methodologies employed in constructing the databases, along with the various bibliometric techniques used to analyze and interpret them. We explain reasons for non-standardization and why it is essential for each observatory to identify metadata and metrics that are meaningful for them; caution the (over-)use of comparisons among various facilities that are, ultimately, not comparable through bibliometrics; and highlight the benefits of telescope bibliographies, both for researchers within the astronomical community and for stakeholders beyond the specific observatories. There is tremendous diversity in the ways bibliographers track publications and maintain databases, due to parameters such as resources (personnel, time, budget, IT capabilities), type of observatory, historical practices, and reporting requirements to funders and outside agencies. However, there are also common sets of Best Practices. This paper describes some of our results from our collaborative discussions.

Corresponding author: Sharon Hunt
sharon.hunt@noirlab.edu

Curation Workflow (Observatory Bibliographies)

Future Trends:

"Efforts are underway to implement an automated paper classification system at STScI/MAST to identify science papers within a set of candidate papers; however, even if this product comes to fruition in the 2020s, it is expected that human intervention will be needed to extract additional information about the paper" (Observatory Bibliographers Collaboration 2024).

"It is worth noting that this ML approach does not completely remove human involvement in the process. Human expertise and learning are needed for marginal cases that are not resolved by existing capabilities. [...] There needs to be a continuing education program for retraining and updating the classifier models with new literature, which will require new labels identified by human experts periodically" (Chen et al. 2022).

Draft version January 2, 2024
Typeset using LaTeX default style in AASTeX631

arXiv:2401.00060v1 [astro-ph.IM] 29 Dec 2023

Assessing Your Observatory's Impact:
Best Practices in Establishing and Maintaining Observatory Bibliographies

Observatory Bibliographers Collaboration

Raffaele D'Abrusco,[1] Monique Gomez,[2] Uta Grothkopf,[3] Sharon Hunt,[4] Ruth Kneale,[5] Mika Konuma,[6] Jenny Novacescu,[7] Luisa Rebull,[8] Elena Scire,[9] Erin Scott,[1] Donna Thompson,[10] Lance Utley,[11] Christopher Wilkinson[7] And Sherry Winkelman[1]

[1] Chandra Data Archive, Chandra X-ray Center (CXC) / Center for Astrophysics, Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
[2] Instituto de Astrofísica de Canarias (IAC), Vía Láctea, s/n E-38205, La Laguna - Tenerife, Spain
[3] European Southern Observatory (ESO), Karl-Schwarzschild-Straße 2, 85748 Garching bei München, Germany
[4] NSF's NOIRLab, 950 N Cherry Ave, Tucson, AZ 85719, USA
[5] National Solar Observatory (NSO)/DKIST, retired, 3665 Discovery Drive, Boulder, CO 80303, USA
[6] National Astronomical Observatory of Japan (NAOJ), 2-chōme-21-1 Ōsawa, Mitaka, Tokyo 181-8588, Japan
[7] Space Telescope Science Institute (STScI), 3700 San Martin Dr., Baltimore, MD 21218, USA
[8] Infrared Science Archive (IRSA), IPAC, MS 100-22, Caltech, 1200 E. California Blvd, Pasadena, CA 91125, USA
[9] IPAC, Mail Code 314-6, Caltech, 1200 E. California Blvd. Pasadena, CA 91125, USA
[10] ADS / Center for Astrophysics, Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
[11] National Radio Astronomy Observatory (NRAO), 520 Edgemont Road Charlottesville, VA 22903-2475, USA

ABSTRACT

Observatories need to measure and evaluate the scientific output and overall impact of their facilities. An observatory bibliography consists of the papers published using that observatory's data, typically gathered by searching the major journals for relevant keywords. Recently, the volume of literature and methods by which the publications pool is evaluated have increased. Efficient and standardized procedures are necessary to assign meaningful metadata, enable user-friendly retrieval, and provide the opportunity to derive reports, statistics, and visualizations to impart a deeper understanding of the research output.

In 2021, a group of observatory bibliographers from around the world convened online to continue the discussions presented in Lagerstrom (2015). We worked to extract general guidelines from our experiences, techniques, and lessons learnt. This paper explores the development, application, and current status of telescope bibliographies and future trends. The paper briefly describes the methodologies employed in constructing the databases, along with the various bibliometric techniques used to analyze and interpret them. We explain reasons for non-standardization and why it is essential for each observatory to identify metadata and metrics that are meaningful for them; caution the (over-)use of comparisons among various facilities that are, ultimately, not comparable through bibliometrics; and highlight the benefits of telescope bibliographies, both for researchers within the astronomical community and for stakeholders beyond the specific observatories. There is tremendous diversity in the ways bibliographers track publications and maintain databases, due to parameters such as resources (personnel, time, budget, IT capabilities), type of observatory, historical practices, and reporting requirements to funders and outside agencies. However, there are also common sets of Best Practices. This paper describes some of our results from our collaborative discussions.

Corresponding author: Sharon Hunt
sharon.hunt@noirlab.edu

Human vs. Machine

Human Curation

✅ Process for identifying and evaluating relevant papers requires a subject matter expert driving the effort

✅ Librarians/archivists define principles behind the curation of each bibliography, based on the needs of each project

❌ Human involvement is expensive and is often a limiting factor in the curation process

❌ The involvement of a human in the loop makes the process somewhat subjective

Automated Text Mining

❌ Useful for finding documents which contain the required information, but additional analysis of sentiment and intent is difficult

❌ Difficult to capture the nuance behind mention of dataset in a paper or their relationship with other findings in the study

✅ Can be implemented at scale for all the records indexed in SciX

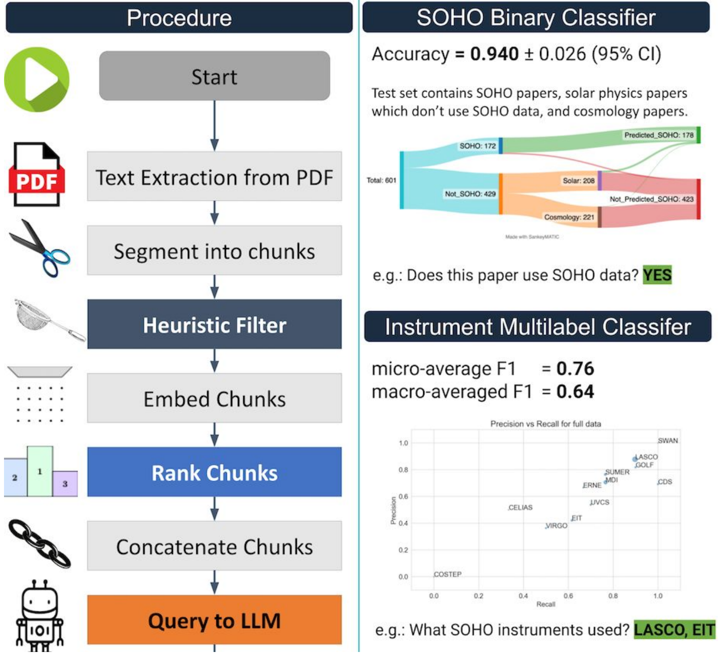
✅ Forces the curation process to become explicit and implementable, thus increasing its reproducibility

What *might* be Possible

Use NLP and AI to accelerate progress

- Named Entity Recognition: find and normalize mentions of missions, telescopes, instruments
- Knowledge Graphs: facilitate disambiguation and relevance of concepts in papers
- Large Language Models: use LLM's capabilities for reasoning and classification of data use vs. mention

Some of these techniques have been successfully applied to identification of papers using Heliophysics missions and Planetary Feature Names detection in the literature.



Buonomo et al, https://doi.org/10.5281/zenodo.8415073

Shapurian et al, arXiv:2312.08579

Figure 1: Pipeline for extracting planetary feature names from full text. The pipeline consists of candidate retrieval, false positive filtering, context analysis, KG matching, paper analysis, and language model querying stages.

What's Missing - Labeled Datasets and Methods

Datasets

- Each bibliography is curated by a different team using different criteria for inclusion
- The data used to create the bibliography (scientific papers) is not always accessible due to license restrictions
- The annotated bibliographies, and the set of metadata associated with them, are stored in different formats and different archives

We need uniform, open datasets that can be used to train tools in support of the curation effort

Methods

- Each observatory/archive has separately developed own methodology for obtaining fulltext and applying TDM techniques
- Criteria for evaluating "use" vs "mention" of data are non-uniform and have evolved over time, so making them explicit is useful
- No economy of scale when each observatory works on their own

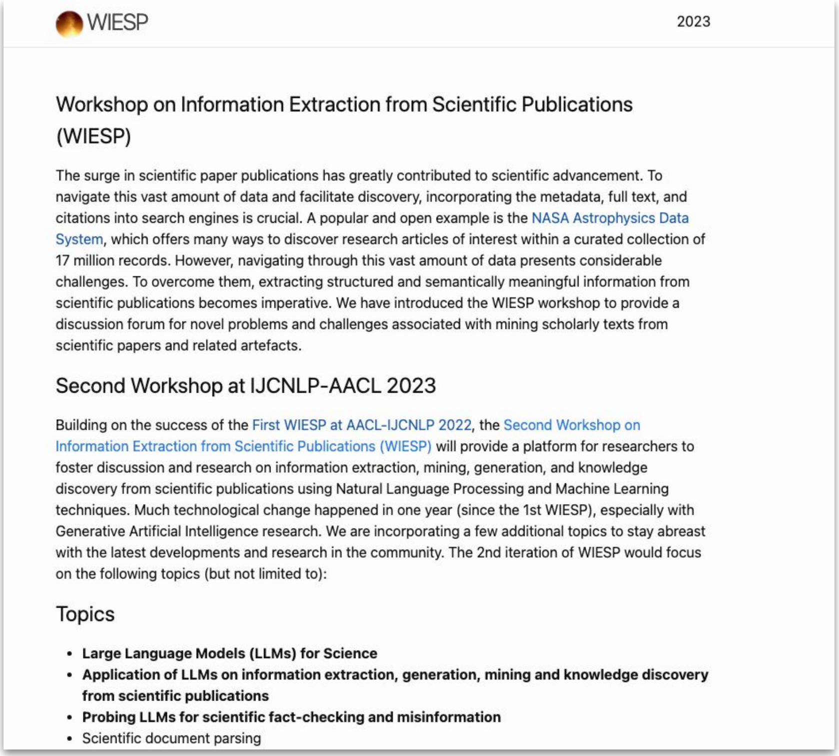
We need to make methodologies explicit so we can enable reproducibility and scalability of effort

A Proposal: The “AstroBibPile” Open Dataset

1. Collect the data and methodologies behind the major active bibliographies in Astronomy, publish it to HuggingFace along with a collection of OA papers that can be used for their analysis

2. Submit a proposal for a new WIESP workshop with this as a shared task at one of the 2025 ACL meetings: https://www.aclweb.org/portal/content/joint-call-workshops-proposals-eaclaclnaaclemnlp-2024

3. Extend to the rest of Space Sciences (and possibly Earth Sciences)



WIESP 2023

Workshop on Information Extraction from Scientific Publications (WIESP)

The surge in scientific paper publications has greatly contributed to scientific advancement. To navigate this vast amount of data and facilitate discovery, incorporating the metadata, full text, and citations into search engines is crucial. A popular and open example is the NASA Astrophysics Data System, which offers many ways to discover research articles of interest within a curated collection of 17 million records. However, navigating through this vast amount of data presents considerable challenges. To overcome them, extracting structured and semantically meaningful information from scientific publications becomes imperative. We have introduced the WIESP workshop to provide a discussion forum for novel problems and challenges associated with mining scholarly texts from scientific papers and related artefacts.

Second Workshop at IJCNLP-AACL 2023

Building on the success of the First WIESP at AACL-IJCNLP 2022, the Second Workshop on Information Extraction from Scientific Publications (WIESP) will provide a platform for researchers to foster discussion and research on information extraction, mining, generation, and knowledge discovery from scientific publications using Natural Language Processing and Machine Learning techniques. Much technological change happened in one year (since the 1st WIESP), especially with Generative Artificial Intelligence research. We are incorporating a few additional topics to stay abreast with the latest developments and research in the community. The 2nd iteration of WIESP would focus on the following topics (but not limited to):

Topics

- **Large Language Models (LLMs) for Science**
- **Application of LLMs on information extraction, generation, mining and knowledge discovery from scientific publications**
- **Probing LLMs for scientific fact-checking and misinformation**
- Scientific document parsing

AstroBibPile: Call for Contributors

- Do you have data that can be useful for this effort? Please consider contributing to the AstroBibPile.
- Are you interested in developing AI techniques to support the creation and maintenance of the bibliographies?
- Do you have additional use cases that could benefit from the AstroBibPile dataset?
- Would you like to know more?

Please get in touch!
aaccomazzi@cfa.harvard.edu

WIESP Shared Task 2025 - The “AstroBibPile”

AI Tools for Bibliography Curation

Alberto Accomazzi, Raffaele D’Abrusco, Jennifer Lynn Bartlett - Feb. 2024

Introduction and Motivation

A well-established way to assess the scientific impact of an observational facility in astronomy is the quantitative analysis of the studies published in the literature which have made use of the data taken by the facility. A requirement of such analysis is the creation of bibliographies which annotate and link data products with the literature, thus providing a way to use bibliometrics as an impact measure for the underlying data. Creating such links and bibliographies is a laborious process which involves specialists searching the literature for names, acronyms and identifiers, and then determining how observations were used in those publications, if at all (Observatory Bibliographers Collaboration, 2024).

The creation of such links represents more than just a useful way to generate metrics: doing science with archival data depends on being able to critically review prior studies and then locate the data used therein, a basic tenet behind the principle of scientific reproducibility. From the perspective of a research scientist, the data-literature connections provide a critical path to data discovery and access. Thus, by leveraging the efforts of librarians and archivists, we can make use of telescope bibliographies to support the scientific inquiry process. We wish to make the creation of such bibliographies simpler and more consistent by using AI technologies to support the efforts of data curators.

Typical Curation Process

While different groups use different approaches and criteria to the problem of bibliography creation and maintenance, the steps involved typically consist of the following:

1. Use a set of full-text queries to the ADS bibliographic database in order to find all possible relevant papers. This first step aims to identify articles that contain mention of the telescope/instrument of interest so that they can be further analyzed. For instance, the set of query terms used to find papers related to the Chandra X-Ray telescope may be “Chandra,” “CXC,” “CXO,” “AXAF,” etc.
2. Analyze the text containing mentions of the telescope/instrument and its variations in order to disambiguate the use of the terms of interest. For the Chandra example, this includes teasing apart the different entities associated with “Chandra,” which may correspond to a person, a ground-based telescope, or a space-based telescope.

https://docs.google.com/document/d/1zDW61dOvpYxaNi74U74F39vzmZzxQc6iP05rIsbLjZU/edit?usp=sharing

Thank You!