EPJ Web of Conferences **186**, 12003 (2018) https://doi.org/10.1051/epjconf/201818612003
LISA VIII

Managing Institutional Bibliographies using the ADS API:

A new workflow using Google Sheets

James Damon[1,2,⋆], *Edwin* Henneken[1,3], and *Alberto* Accomazzi[1,4]

[1] *NASA ADS*
[2] *orcid.org/0000-0002-1069-2376*
[3] *orcid.org/0000-0003-4264-2450*
[4] *orcid.org/0000-0002-4110-3511*

Abstract. Curating institutional bibliographies with the ADS web interface is currently a manual process that scales with the number of search terms. Long author lists and institutions with multiple sub-organizations or name variations increase the workload. Review work is monotonous and can take significant time depending on the size of the institution and the frequency of reviews. Consequently, bibliographies generated in this way are costly and may suffer from human error. We propose a semi-automated workflow that uses an iterative approach to discovery with ADS's new search engine and a recently developed Google Sheets add on. First, affiliation strings from a user created spreadsheet are searched with the ADS API and for each result the matched affiliation and the paired author are retrieved. Next, each author name string is searched and items where that author is paired with an empty affiliation field are retrieved. The results from both queries are then compiled into output sheets with pertinent information for manual review. Finally, the selected items can be added to an ADS library from the Google Sheets interface. The tool can also use previously rejected affiliation strings to flag false positives in subsequent queries. Curators do not need to have extensive technical skills in order to use the workflow and they can help improve the ADS by opting to share ORCIDs, author synonyms, and affiliation synonyms.

1 Introduction

Managing the bibliography for the Harvard-Smithsonian Center for Astrophysics has always been based on an initial author query with high recall but low precision. Each record in the results set is manually reviewed in order to maximize precision in the final selection. Finally, the selected records are added to the bibliography in a batch.

Maintaining the possibility of a methodical hand-curated process was a key design consideration during the creation of the tool. The designers opted to provide larger results sets with many fields so curators could make more informed decisions while maintaining control of the process.

The reduction in time cost and the added utility stem from the use of Google Sheets and the ADS API.[1] Spreadsheet software is commonly used for data manipulation. By building this tool, we are allowing curators to leverage existing skills and reduce the time needed to effectively manage a

⋆e-mail: jdamon@cfa.harvard.edu ORCID: 0000-0002-1069-2376

bibliography. Additionally, the tool can be used at varying scales. It was designed with institutional level bibliographies in mind but it can also be used for labs or individual authors. Furthermore, it allows the work to be split up among multiple collaborators.

2 Using the Tool

While it is up to individual users to take advantage of the tool's features during the completion of tasks, the designers had certain workflows in mind during the creation of the tool. The underlying assumptions and an example workflow are outlined in the following section.

2.1 Results Sets

When performing an ADS query for a bibliography, the two types of text strings that are used in the search are author name and affiliation. While ORCID is the recommended format for author lists, its adoption was recent enough that most bibliographies should still include full author name queries in the workflow. ADS search includes results for author name queries from known author name synonyms so it may not be necessary to search for variations of an author's name. The affiliation text string used by authors from some institutions may not be consistent so unlike the author name search, it is best to identify the possible variations. Each individual variation may then be searched or alternatively, if the different versions of the string have some overlap that is not likely to cause a large increase in false positives, that portion of the string may be a good candidate for use during affiliation queries.

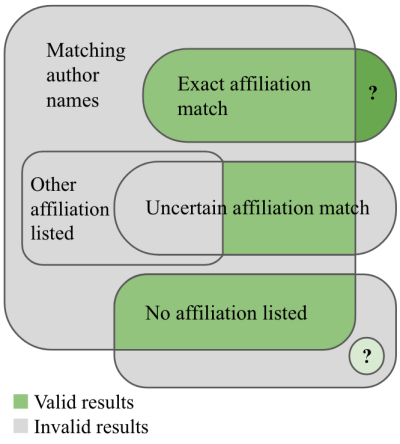

Figure 1: Results Sets

The ADS API is able to search author and affiliation strings in the same query however, it cannot currently search author and affiliation pairings.[2] In order to address this, for both affiliation and author name queries, the paired author name or affiliation is retrieved by the Google Sheets add on as well and included in the search results for the curator to review.

The results sets from affiliation and author name queries overlap as illustrated in figure 1. Exact affiliation matches and exact author name matches can be added automatically however other cases are up to the curator's discretion. Exact affiliation matches with unexpected authors may be the result of an out of date author list or authors falsely claiming an affiliation. Uncertain affiliation matches with exact author name matches may be the result of variations on common affiliation strings whereas uncertain affiliations paired with unexpected authors are more likely to be negative results that should not be included in the bibliography. Author name matches with no affiliation data require the curator to find and inspect the particular item in order to determine to validity of the record. Finally, there is a theoretical set of records that will not match on author names or affiliation due to typos or uncommon name/affiliation variations. For most bibliographies, these records would be ignored because of the extreme effort required to find them.

Table 1: Example affiliation variations

Original Affiliation String	Ambiguity	Example False Positive
Harvard-Smithsonian Center for Astrophysics	Exact	-
Harvard Smithsonian Center for Astrophysics	Exact	-
Harvard-Smithsonian Ctr. for Astrophysics	Exact	-
Harvard-Smithsonian CfA	Exact	-
Harvard College Observatory	Exact	-
HCO	Uncertain	H.C.Ã~. Universitetsparken
Harvard Astronomy Department	Exact	-
Smithsonian Astrophysical Observatory	Exact	-
Smithsonian Astr. Obs.	Exact	-
Smithsonian	Uncertain	Smithsonian Institution
SAO	Uncertain	SÃ£o Paulo

2.2 Example Workflow

Individual workflows will vary. The ideal workflow for the creation of a bibliography will be different from the workflow used to maintain a bibliography. Furthermore, the curator must determine an optimal level of diligence based on the time they are able to commit and the underlying purpose of the bibliography. Regardless of the specifics of the chosen workflow, the tool will be of most use for affiliation and author name queries. An example workflow follows and is represented by figure 2.

For new institutional bibliographies it may be useful to do an initial author query to search for variations on the name of the institution. Once the curator has confidence in the list of affiliation names that should be used to positively identify items for the bibliography, an affiliation query is run on each string and the results are viewable in the spreadsheet.

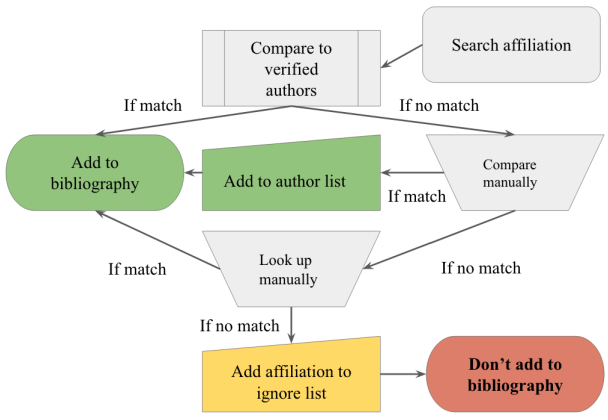

Figure 2: Example Workflow

When using an author list, the curator may compare the author names that are paired with the affiliation string of each result and add items that match both. Results from exact affiliation queries may be reviewed for unexpected items and then added as appropriate. Uncertain affiliation results should be reviewed more carefully and may include unusual affiliation name variations that the curator should be aware of.

If the curator has an author name list, they can perform an author query to make the bibliography more comprehensive. While some author names are unusual or unique, author queries tend to have higher recall and less precision overall due to frequent negative results from queries of common names. During the review of affiliation query results, many records would have been reviewed by the curator. In order to reduce the duplication of effort, the curator can use the spreadsheet to identify results from the author query that

were already reviewed during the affiliation query and ignore them. The curator may opt to review the affiliation strings that are paired with the author name results in order to find previously unknown affiliation strings or typos. For author name results that do not have an affiliation, stringent curators may decide to use the links to items on the ADS's web interface to look for affiliation information on the original paper.

Once all the desired papers have been identified by the curator, they may be used to generate an ADS library from within the Google Sheets interface.

3 Evaluation

In order to compare the time commitment needed to maintain a bibliography using the classic interface vs using the Google Sheets add on, student colleagues timed themselves while performing example tasks. Each task involved including or excluding records for a bibliography.

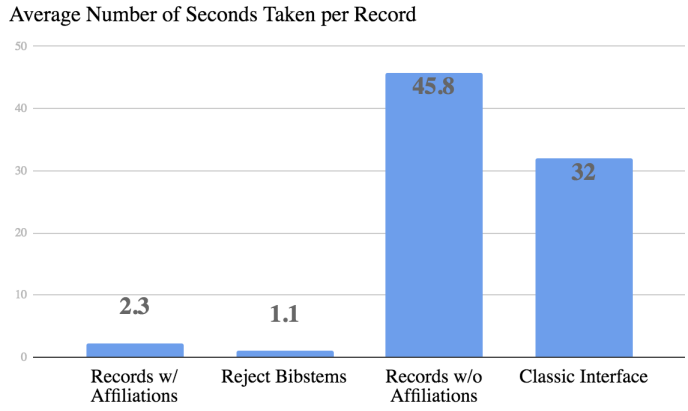

Figure 3

Task 1 used the new bibliography tool. Task 1a was selecting records from the set of results that had affiliation strings. Task 1b was rejecting a subset of records with certain bibstems from a results set . Task 1c was reviewing results that had empty or missing affiliation strings in the bibliography tool.

Task 2 was to use the ADS Classic interface to select records. The average time taken for each type of task can be seen in figure 3.

4 Results

By splitting review work into three discrete tasks and reducing the time needed for two out of three of the tasks, the tool offers an improvement over the previous method of bibliography generation as seen in the simulation results in figure 4.

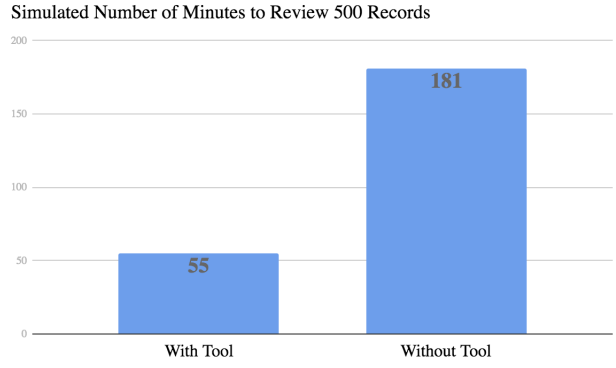

Figure 4

5 Future Work

Features currently under consideration or in process include bibgroup

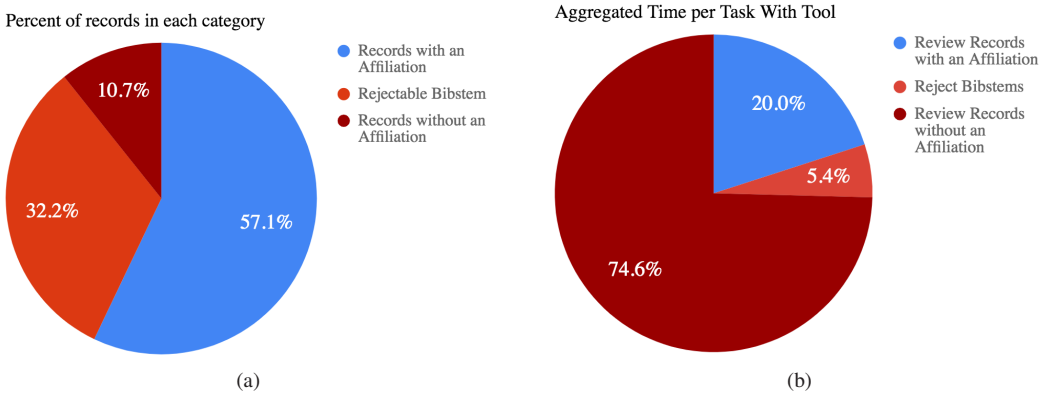

Figure 5

based queries, user defined searches, and formal synonym feedback mechanisms. Future development of fully featured spreadsheet based information retrieval interfaces may also be worth exploring.

Acknowledgements

NASA ADS is operated by the Smithsonian Astrophysical Observatory under NASA Cooperative Agreement NNX16AC86A

References

[1] A. Accomazzi, M. J. Kurtz, E. A. Henneken, et al., *ADS: The Next Generation Search Platform* in *Open Science at the Frontiers of Librarianship*, edited by András Holl, Soizick Lesteven, Dianne Dietrich, and Antonella Gasperini (2015), Vol. 492 of Astronomical Society of the Pacific Conference Series, p. 189–197

[2] C. S. Grant, D. M. Thompson, R. Chyla, et al., *Enabling Meaningful Affiliation Searches in the ADS* in *Open Science at the Frontiers of Librarianship*, edited by András Holl, Soizick Lesteven, Dianne Dietrich, and Antonella Gasperini (2015), Vol. 492 of Astronomical Society of the Pacific Conference Series, p. 208–211