Managing Institutional Bibliographies using the ADS API: A new workflow using Google Sheets

James Damon, Edwin Henneken, Alberto Accomazzi (NASA ADS)

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.

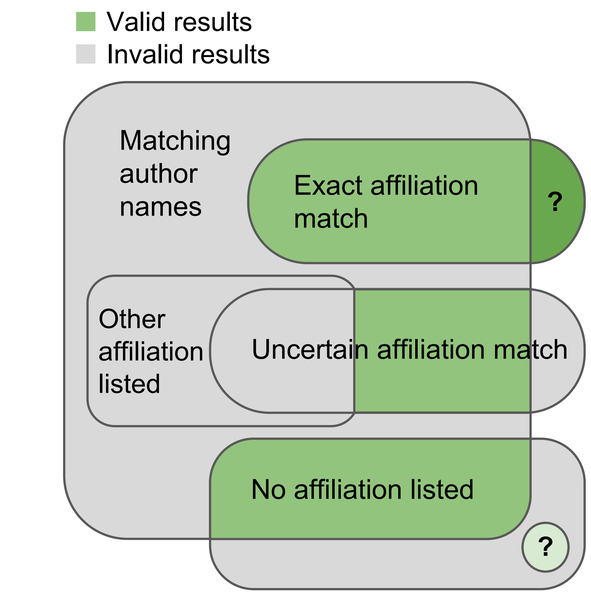
Background

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. The results are then manually reviewed in order to maximize precision in the final selection.

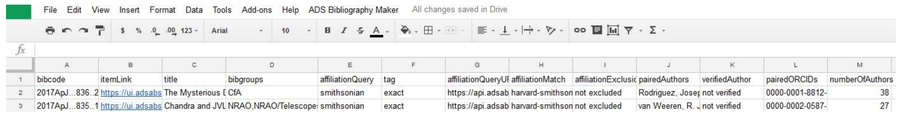
In designing the tool we did not want to eliminate the possibility of a methodical hand-curated process. We opted to provide larger results sets but to include fields so curators could make more informed decisions while maintaining control of the process.

The reduction in time cost and the added utility stem from the use of Google Sheets. 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 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.

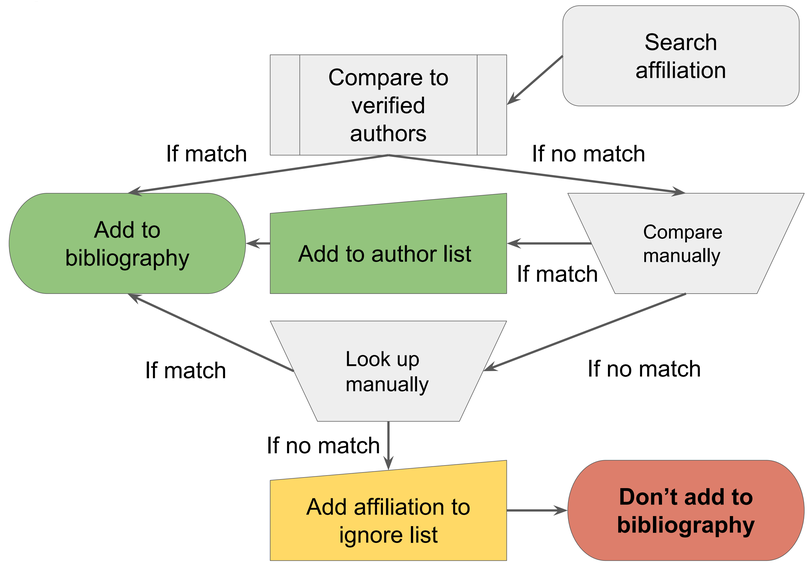
Bibliography Search Result Sets



The Tool Interface



Example Workflow



Evaluation

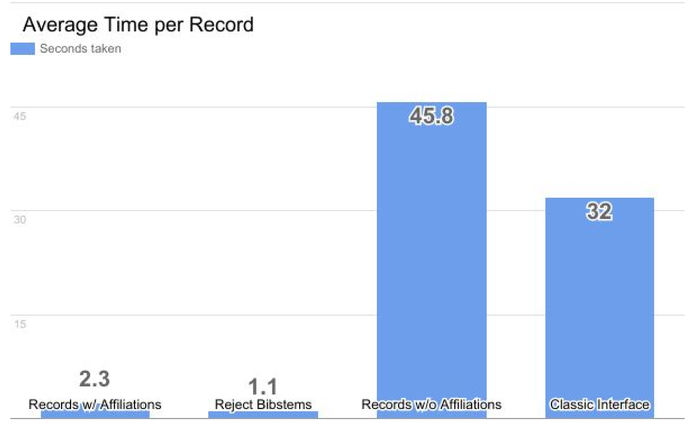
In order to compare the time commitment needed to maintain a bibliography using the classic interface vs using the Google Sheets add on, student colleagues timed themselves while performing example tasks. Each task involved including or excluding records for a bibliography.

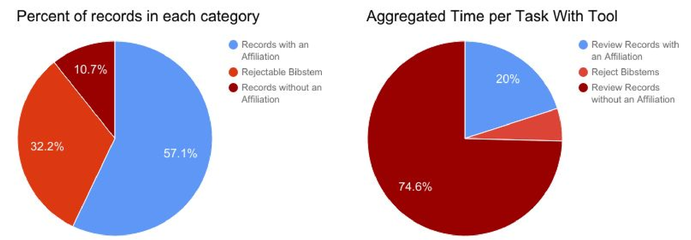
Task 1 was to use the ADS Classic interface to select records.

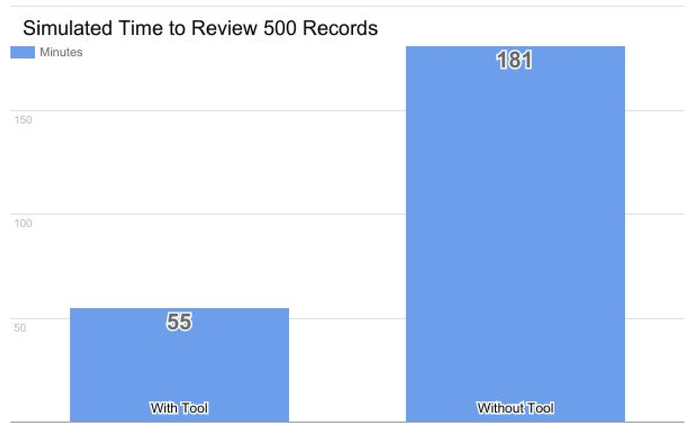
Task 2 used the new bibliography tool. Task 2a was selecting records from the set of results that had affiliation strings. Task 2b was rejecting a subset of records with certain bibstems from a results set . Task 2c was reviewing results that had empty or missing affiliation strings in the bibliography tool.

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.







Future Work

Features currently under consideration or in process include bibgroup based queries, user defined searches, and formal synonym feedback mechanisms.

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