Report of the ADS Users Group (ADSUG)
Nov 9-10, 2022

Introduction

The NASA Astrophysics Data System (ADS) continues to be the single most important tool in worldwide astrophysics, and contributes meaningfully to expanding our understanding of the Universe on a daily basis.

The ADSUG congratulates the ADS team on the unrelenting high standards they apply to this service: both the parts that users draw on every day, and all the hidden infrastructure and future developments that enable ADS to keep pace with demand and future developments. We particularly acknowledge the enormous amount of work that has taken place on maintaining / improving / expanding the core system, the extensive levels of user support, the successful expansion proposal submitted to NASA, and the ensuing work needed to prepare for the upcoming evolution into SciX.

In this report, the ADSUG provides commentary and recommendations on the following areas:

- Legacy components
- System development
- User support
- Staffing
- Expansion

A summary of the recommendations in these sections is provided toward the end of this report, along with an appendix containing a full list of relevant requests received from the community.

Legacy Components

The 2022 presentation to the ADS User Group identified four remaining pipelines with legacy dependencies:

- Import pipeline
- Data pipeline
- Full-text pipeline
- Citation capture

The ADS has already shown considerable forethought in addressing those legacy elements that will benefit from more immediate replacement, or in testing new software that will have the greatest return, for example, testing ¼ of the parsers that represent ⅔ of the literature at present.

The User Group's primary point of concern and advice concerning legacy pipelines is to prioritize those elements that will provide immediate return when ingesting, parsing, and counting references and citations for the expansion content areas in the order of expansion, and/or focusing on those improvements that are necessary for enabling ingest of expansion area content.

As a clear example, the Docmatcher tool meant to roll up arXiv preprint records with their refereed versions needs to be thoroughly developed and tested with the ESSOar and bioRxiv preprint processes to ensure that there is no risk in falling behind in those content areas as ingestion increases beyond astrophysics.

Reference extraction is one replacement project that will continue to pose exceptional challenges as the content areas and publisher/journal coverage increase with ADS's expansion.

The ADS Scan Explorer UI is a successful example of how ADS took advice from the 2021 ADSUG report to leverage the need for legacy component replacement to enhance the service. It also provides an outsourcing/contractor success story that ADS should refer to when identifying other projects that are of limited scope and which may be best addressed by external contractors. If internal ADS staff find that they are challenged to complete legacy replacement projects as the pressure of the content expansion increases and staffing challenges occur, the contracting of limited scope, one-time replacement projects, where gains in team knowledge are limited in value, may be an approach that ADS needs to repeat.

Ideally, any backend legacy pipeline replacements projects that are content agnostic will be completed before content ingestion expands to include Earth and Biological/Physical Sciences. Any legacy replacement projects that may be impacted by varying publisher formats must be carefully tested with examples of the literature from Earth Sciences and Biological & Physical Sciences. This will ensure that replacement services created to manage the existing Astrophysics, Heliophysics, and Planetary Sciences content do not require substantial changes or enhancements later on, as content expands further.

System Development

The ADS team has taken a thoughtful and responsible approach to deprecating legacy systems that remain critical to infrastructure, whilst exploring new technologies that will scale well. The deprecation plan for the legacy systems seems appropriate, allowing "ADS Classic" to transition into just one of the many ingestion sources for ADS. The proposed timescale for replacing the skeleton services of the back office system by the time of the 2023 ADSUG meeting seems appropriate. Once ADS Classic transitions to being just one ingestion source, there is a risk that this could prolong the final decision for when to shut down that system. This could balloon into significant technical debt by requiring the team to maintain a deprecated technology stack while simultaneously expanding in scope (scientific areas) and scale (sheer numbers). The ADSUG recognizes that it is realistic to describe the termination of the legacy system as a multi-year effort, but we advocate that the ADS team set a specific year and month, as specific dates are easier to work towards.

The astroBERT pilot has demonstrated success in context-aware entity recognition, and it is good to see the results (and data) being made publicly available through a workshop. There have been significant improvements in large language models in the last year, and the team seems up-to-date with these contributions, especially given their plans to consider improvements using conditional random fields, and semantic textual similarity. The ADSUG reiterates that given the rapid development in large language models, it is very possible that an internal tool created by the ADS team could be outstripped by a contribution by an external group. Similarly, it may require significant human time to achieve relatively small improvements in effectiveness.

Having an internal tool for context aware labelling still requires significant effort to roll out to users, or to integrate into applications where efficiency gains could be realized. Many of those choices will be independent of the underlying tool (e.g., astroBERT, or some successor). For these reasons, we suggest that the ADS team bring forward their plans to roll out the existing astroBERT tool. We propose that the team consider allowing (signed-in) ADS users to make label corrections to astroBERT predictions to reduce the long-term burden on the team. The ADSUG understands the ADS team's position that astroBERT is "good, but not production ready". However, given the rapid development in language models and the impending expansion requiring significant efficiencies (e.g., UAT Concept Extraction), it seems appropriate to consider rolling out an early implementation in which users could curate predictions. This kind of approach could also benefit the work in data enrichment, particularly in specialized sub-areas of astrophysics for which the ADS team may lack specific context. There are many ways that this could be implemented, and the ADSUG is confident that the ADS team is more than creative enough to implement some user-curated approach in a responsible and effective way.

Transitioning from scheduled (large) data ingestions to an event-driven system is a positive step that will likely help the ADS team scale into new areas. The choice of technologies (gRPC, Kafka) is appropriate, and reflects the team's commitment to be agile, modern, and responsive.

The hack week appears to have been a resounding success. The ADSUG finds it refreshing that future meetings are planned to be in-person; this is an effective way to keep the team connected. It's particularly impressive that the ADS team managed to develop a microservice (Docmatcher) to replace a critical part of the legacy infrastructure during that hack week.

The ADSUG recognizes the team's success in contracting astronomy-specific programmers to develop isolated systems (e.g., ADS Scan Explorer) that the ADS team has yet to be able to work on. The ADSUG feels that the ADS team has maintained the right approach (and balance) between in-house training and development, vs the outsourcing of specific work packages. We recognize that this will likely be an appropriate (and perhaps, even necessary) avenue for the development of isolated services in the future. Keeping projects where significant knowledge development occurs "in-house" would be appropriate.

The ADS team has made impressive updates to the user interface (Project Nectar), both in design and in the choice of technologies. The updated design thoughtfully captures the most common kinds of feedback from users, and remains both sleek and information-rich. The ADSUG concurs with the team's decision to limit any UI/UX changes on Bumblebee until the next major release.

Curating software is increasingly important in STEM-related fields, and the ADS team has made laudable improvements to ingesting and curating software associated with publications. Including the level of granularity down to the software version is particularly important for open science and reproducibility, and will have long-term benefits for users, code creators, and the field as a whole.

User Support

The ADS team continues to provide a large range of excellent support structures to its large user base. Many/most past issues noted by the community and/or raised in previous ADSUG reports have been satisfactorily addressed, and the community feedback provided to the ADSUG is dominated by a recurring theme of appreciation for a unique, responsive and critical service.

Substantive engagement of users continues to be a challenge: there has been only modest participation in Twitter “office hours”, and many feature requests submitted by the community to the ADSUG were for things that ADS can already do. The ADSUG recommends that the ADS team maintain their multi-stranded approach to user support (e.g., Twitter, blog, newsletter, help pages, conference booths), recognizing that (a) most users will continue to only make use of basic functionality of the service, and (b) there will be new challenges for support, communication and training as ADS expands into new disciplines. The ADSUG encourages the ADS team to develop bespoke strategies to engage and train each user community once ADS begins to enter a new field and develops an understanding of that field’s culture.

As in past years, we have sought feedback from the user community as to possible improvements to the ADS interface and capabilities. Requests that the ADSUG considers to have significant value or importance are as follows:

- Following on from the request in last year’s ADSUG report, the ADS has now implemented a basic print function so that people can easily produce a plain list of publications. This is very useful. However, this addition does not seem to be documented, nor is how to do this obviously available on a search results page. The ADSUG suggests that an explicit “print results” option be added to the results screen, either as a standalone button or in the “Export” pulldown menu.

- There is a growing prevalence of papers where there are two joint first authors (and also large-N author lists with corresponding authors who are not first authors). The ADSUG recommends that the ADS consider possible options to accommodate and index such practices, especially as ADS expands into other fields that may have different author conventions. The ADS team has highlighted some of the difficulties with introducing such a feature, including the lack of consistent data and meta-data structures, and the impact of joint authorship on bibliometric/citation statistics. However, we encourage the ADS to consider in more detail what options might be available to even partially implement this, even if a complete solution is not currently feasible.

- There are a variety of common reasons why some or all of an author’s papers might not show up in a search for that author. Examples include authors who are known both by a double-barrelled and abbreviated last name, or last names that are commonly misspelt. The ADS has noted the increased ambiguity that would result if a search returned additional names that are not an exact match to the search term entered. However, given that this is a significant issue that preferentially impacts those who are from non-Anglo/European backgrounds, the ADSUG encourages the ADS team to consider possible solutions. For example, suggestions or auto-complete options that appear as the user is entering a name could be extremely helpful.

- There are several fields in an ADS record, such as DOI, arXiv and bibcode, which a user often wants to copy and paste into another window or document. Highlighting these fields and then copying them, whether using a mouse or a mobile device, is not simple. It would be extremely helpful if a “copy to clipboard” button could be added next to each of these fields. (The bibcode field does seem to have this, but it is almost impossible to use it on a desktop/laptop: it appears when one hovers over the bibcode, but then disappears as one moves one’s mouse over the question mark button, unless one moves the mouse very slowly. It may be better just for the clipboard to be there all the time, and for all relevant

fields.)

- The help pages on the ADS site should be easier to find. At the moment, the help pages are accessed either from the “About” pull-down menu at the top of the page, or via “ADS Help” at the bottom of the page. The ADSUG recommends there be a one-click help link at the top of the page that can immediately take users to https://ui.adsabs.harvard.edu/help/. We also note that the Google search results offered from this page could be improved. For example, searching for “NSF” gives four paid advertisements, only producing the actual information needed (at https://ui.adsabs.harvard.edu/help/faq/) at the bottom of the search results.

Some of the items listed above were discussed in the ADSUG meeting and were noted as challenging. Nevertheless, we list them above because of their value to the community, and encourage the ADS to consider them as resources allow.

Staffing

Recruitment to support Expansion

The ADS will be doubling its workforce over the next 3.5 years. It is going to be challenging to find the personnel on such short time scales and to have the time and energy to go through the hiring process.

We recommend that the ADS team advertise the various job positions as broadly and aggressively as possible, even if any given venue has not historically yielded large numbers of applicants – the cost/energy of doing so is small compared to the potential reward. For example, advertising on the AAS job register may attract increasing numbers of astronomers graduating with strong development skills. Naturally, differentiated strategies will be required for the various different positions.

The ADSUG recommends that the ADS team receive additional support from CfA/HR, for both job advertisements and pre-screening (e.g.. support to allow job-opening postings on LinkedIn).

In the case of discipline-specific scientists, ADSUG recommends that the ADS team consider the possibility, as appropriate, of hiring remote scientists embedded in their academic research environments. Providing the opportunity for scientists to remain embedded in a research environment will enhance the appeal of these positions and will maximize the chances of hiring highly qualified candidates. Such embedded positions would assist researchers to stay abreast in fast-evolving sectors.

ADSUG also recommends that the ADS team considers funding multiple scientists, each at partial funding, to maximally cover the new disciplines (see also below).

As additional points, we recommend that it is clarified throughout all job-ad platforms that non-US citizens are welcome to apply (where appropriate).

Finally, we recommend to leverage the prominence of Harvard and the CfA by advertising ADS job openings at. e.g.. https://www.cfa.harvard.edu/opportunities.

In view of the future recruitment challenges, the ADSUG notes that the option for flexible remote working may not only be an attractive option, but an essential employment benefit for recruiting successful and competitive new team members, especially in sectors that have recently allowed ever more flexibility to employees. The possibility granted by the CfA of flexible teleworking for on-site employees seems to have been very beneficial to the ADS team. Therefore, the ADSUG recommends that the CfA supports the ADS team in keeping remote working at the current levels or at whatever levels the ADS team considers beneficial for the maximal functioning of the team, for the recruitment and for the retention of team members.

Onboarding and Retention

Effectively integrating new workers into hybrid teams is critical to achieving cohesive, long-lasting teams with high morale. The usual challenges are for existing managers to find time to onboard and train new staff while still completing their own assigned tasks. These challenges are exacerbated when the new hires represent a significant fraction of the total workforce and when a significant fraction of the workforce is remote. The User Group makes the following recommendations to address these issues:

- ADS should consider offering remote workers a substantial (e.g., 2–3 weeks) on-site onboarding experience.
- ADS should encourage or require on-site visits of remote workers at a regular cadence.
- ADS should explore the benefits of holding team retreats occurring more frequently than three times a year.
- ADS should explore the benefits of one-on-one mentorship opportunities for new hires, to help transmit the working culture that has led to ADS's long-term success.
- ADS should prepare the established staff for the inevitable changes that come from quickly doubling one's workforce, and should help them to feel included in any changes.
- ADS should continue to follow best practices for promoting an inclusive work environment to promote staff retention.

Expansion

Overall, the ADSUG agrees that the NASA SMD directorate made the right decision to leverage ADS's existing expertise in linking the literature within itself and to an increasingly open and complex set of data, software, and other research outputs. ADS's considerable experience with named entity recognition, full text parsing, reference extraction and resolution, and preprint matching are all services from which other SMD research areas will benefit as their content is added to the ADS. It is critical that NASA SMD understands that the expansion is the beginning of what should be considered a continual budget commitment, and that SMD appropriately request and earmark funding for the expanded ADS tool on a long term basis (assuming renewal of the long-standing cooperative agreement).

Essentially doubling the current FTE count from 16 to ~32 over the course of 4-5 years will be the most challenging aspect of the expansion, as addressed elsewhere in the Staffing portion of this report. While is it fair to assume that existing technical staff who specialize in data science, ML/AI, and other back end and front end development will be able to leverage past experiences and use economies of scale to manage the increase in content, the ADS should be careful to check its assumptions that the greatest amount of staffing will be required in the curator and scientists areas

with a lesser increase in technical staff. The challenges of managing backend development and legacy replacement projects – such as parsers and reference extractors – for the existing literature, while expanding in content areas that can influence the way in which backend tools operate, cannot be overstated. The same can be said for tangential projects, such as the Planetary Nomenclature project that will need to be mirrored and expanded as Earth Sciences and Biological & Physical Sciences content is added. All of these projects will require more of the technical staff as well as of the new content and science specialists.

In addition to the domain expertise and curation needed to ensure successful ingest and expansion of the indexed literature, outreach into the Earth Sciences and Biological & Physical Sciences communities should dominate the labor hours of future Project Scientists in the first 3-5 years of the expansion. An open source indexing service like the ADS does not have an equivalent in most other fields outside astrophysics, so educating these new content communities on the existence of the ADS service and how to use it is crucial to future use and success of the project (and therefore to continued funding beyond 2026).

ADS has shown considerable foresight in how to advertise its service – project scientists and supporting ADS staff must be properly funded to attend major conferences and events in their respective domains. Additionally, the ADSUG recommends that project scientists in content areas that are outside the scope of the Center for Astrophysics be permitted to work primarily from their host institutions where research in their domain (such as Earth Sciences) is conducted. This may result in a situation where ADS funds, say, two 0.5 FTE Earth Scientists, with their host research institutions hosting and paying for the other 0.5 FTE of each appointment. Continuing to work, collaborate with, and learn about new ADS user needs in these expanded communities may prove more challenging if these people are required to be physically located at the CfA.

The experimental SciX UI already shows immense promise, and the ADS team has appropriately identified aspects of expansion that could, if mismanaged, result in a degraded user experience for existing ADS users, e.g., the management of acronyms and automated synonym expansion.

The ADS team requested ADSUG feedback on the expansion advisory group vs. the traditional ADS User Group. For the expansion advisory board, it will be important for the ADS team to invite feedback from experts in the expanding content areas of Earth Sciences and Biological/Physical Sciences, while retaining one or more current users in astrophysics, heliophysics, and planetary sciences. “Current” ADS users who are already familiar with the service can both serve as a bridge between ADS and new content users, and can also identify ideas from other science areas that may improve the Astrophysics, Heliophysics, or Planetary Sciences collection and UI within SciX.

The ADSUG recommends that the ADS User Group continue to exist, gradually adding at least one and up to two current user representatives from each SMD content area as ADS becomes established as the predominant literature index in the new content areas. The intent is that in about 3-5 years time, there may no longer be a need for an expansion advisory group, but the ADSUG would remain.

Working with the primary science bodies of each content area, such as the AGU for Earth Sciences, would be the best way to recruit qualified individuals for the expansion advisory board and for future ADSUGs. For Biological & Physical Sciences, see https://spacemicrobes.com/eana/

Summary of Recommendations

1. Provide a specific schedule goal date for the termination of legacy systems.
2. Conduct legacy replacement projects in the context of expanding literature/differing publisher sources.
3. Make the ADS Help page more visible (currently not highly visible as identified by '?' symbol beside the word 'About')
4. Continue to provide support across a variety of channels to ensure a broad audience is reached.
5. Advertise recruitment as broadly as possible:
    - Publish in discipline specific forums (e.g., AAS Job Registry)
    - Work with the primary science bodies in each new content area to enhance reach of recruitment efforts.
    - Advertise on CfA's Harvard opportunities page.
    - Seek additional support from CfA/HR to match the increased recruitment and screening load.
6. Allow staff to work remotely to enable greater diversity and enhance the appeal of ADS positions relative to other employers. The team appears to function well in this mode.
7. Allow project scientists who work in content areas outside the scope of the CfA to contribute/work remotely so they can remain embedded within their research communities. These positions could be funded at partial levels to enhance diversity within a fixed budget.
8. As SciX is developed, expand, incrementally, the expertise within the ADSUG to allow expanded perspectives to be brought into the UG while maintaining an experienced UG membership who can provide experience driven feedback to ADS and mentorship to UG members from new disciplines.
9. Work towards integration of AI aware labelling tools within the search system even while the current astroBERT tool is considered immature. An early deployment could include an ability for users to curate predictions made by the AI tool.
10. Examine the feasibility of enabling a more diverse interpretation of the meaning of "First Author" and name synonyms.
11. Continue to use contract labor where appropriate or necessary to complete well defined high-priority projects that do not significantly build or enhance the team's knowledge of operational systems.

Conclusions

After several years of planning, the ADS is now about to move into a major new phase, in which restructuring and expansion will result in a new SciX project that will be very different from ADS. This will inevitably not be an entirely smooth process: it will be important that staff continue to feel valued, integrated and supported as the team expands, and that any problems or barriers associated with the expansion (whether technical, strategic or personnel-related) are identified and rectified quickly. The ADSUG will need to evolve to match the evolution of ADS into SciX, and looks forward to providing input and advice in new ways.

We conclude as we began, by noting the ridiculously outsized value that this small team adds to scientific inquiry, and by thanking them for the high standards and sustained commitment that they bring to this project. We particularly highlight the leadership of the ADS Principal Investigator, who has successfully led the team through the difficult years of the pandemic, and has now positioned the project for an exciting period of growth and change.