Platform Transition: Next Steps

Alberto Accomazzi & the ADS Team
aacomazzi@cfa.harvard.edu
@aacomazzi

ADS Users Group Meeting - 11/3/207
Completing the Transition

- Getting rid of the ADS Classic search engine is the first, big step
  - Migrate applications (API)
  - Allow crawlers to index new pages
- Other major components that require reimplementations:
  - The ingest and curation system (code and workflow)
  - The reference resolver (building the citations database)
  - The notification service (myADS personalized alerts)
  - The article archive (serving content digitized by ADS)
- Until all these components are replaced, dependencies on the ADS Classic still exist (even if just under the hood)
Ingest and Curation

- **Scope of work:**
  - Revamp data ingest system developed in the early days of the project, which relies on an obsolete technology stack
  - Manage current data holdings, consisting of just a bunch of files on disk
  - Recode / replace 200K lines of (mostly) PERL code supporting data harvesting, normalization, validation, and ingest in ADS classic

- **Goals:**
  - Use modern technology stack for harvesting, curation workflows, pipelines; ideally do this by re-using an existing DL framework
  - Update metadata model to accommodate expanded content types (data and software)
  - Make ingest more, not less, efficient than current ADS classic
Reference Resolver

- Machinery which is used to extract, parse, and identify references, build the citation database
  - Existing code consists of 50K lines of mostly python code, fully reliant on ADS classic backoffice data and original metadata model (bibcode)
  - Thanks to the work being done on the new platform, we now have a better metadata store to use for reference validation
  - Still, a lot of the “secret sauce” is found in data extraction and parsing heuristics, which needs a major rewrite

- Goals
  - Collaborate with other projects, use existing reference extraction frameworks to get a jump start on extraction and parsing rewrite
  - Replace heuristics with ML approaches, train on ADS core journals
Notifications and Recommendations

- ADS provides notifications via the myADS service
- Subscribers receive daily or weekly notifications based on their profile
- Stay up to date on topics, people, citations, trending articles
- Currently used by 16k accounts total, 6k astro, 4.1k physics, 5.3k weekly arXiv, 5.2k daily arXiv
- Will need to be updated to use new search engine, recommender system
Article Archive

- ADS provides open access archival copies of all journal and conference proceedings articles in astronomy
- Current holdings consist of 663k articles, 4.9M scanned pages, 2TB
- While our digitization program is essentially over, we occasionally add some unique material (e.g. ICRC)
- Access to this content is also based on ADS Classic
- Will need to find new framework to move this content to
Build out Plan

- Keep track of user searches, enable saving searches and push notifications based on them
- Leverage on personalizations, ADS libraries to provide recommendations
- Leverage on ORCID to provide claim notifications
- Enable sharing of collections (libraries) to support collaborations
- Integrate with third-party manuscript editing systems