Work Ahead

Alberto Accomazzi and the ADS Team

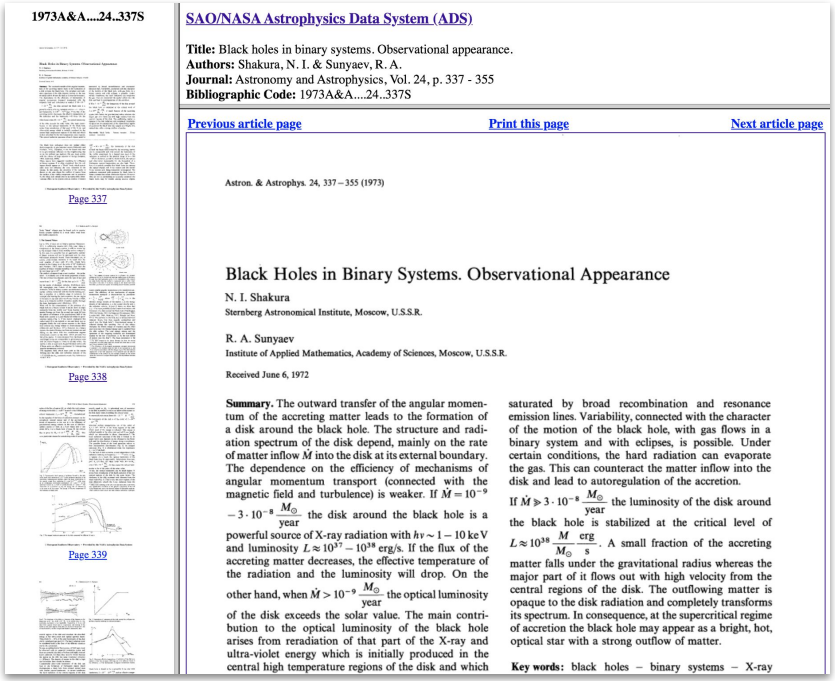
ADS Users Group Meeting, 20-21 Nov. 2019

Myth 1: We're done with Classic Interface

Reality: while users have been transitioned off the ADS Classic search engine, legacy components still exist

ADS Article Archive

- 25yo system used to process and visualize digitized scans of historical literature
- Current holdings consist of 670k articles, 5M scanned pages, accounting for ⅓ of the refereed literature in astronomy
- Ingest pipeline involves rasterization, document bundling, OCR, fulltext extraction
- Scope of work: replace 50K lines of C/PERL, decouple from ADS classic, build into UI

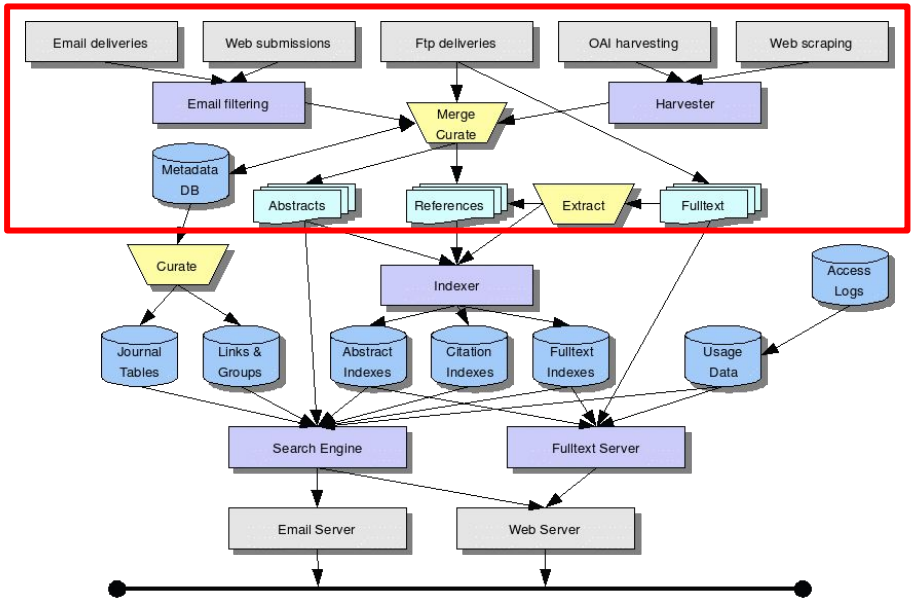


Myth 2: We're done with Classic Backend

Reality: while front-end has been re-engineered, back-office curation and ingest processes need to be updated

Harvesting, Ingest and Curation

- Data harvesting relies on PERL code developed over the past 20 years
- Data extraction now uses a mixture of PERL and python (since 2018), flat files
- New ingest pipeline should allow indexing at a higher cadence (daily updates from publishers)
- Curation framework needs to be as efficient as existing one, while yielding better outcome (e.g. richer data model, unicode cleanliness, supervised metadata enrichment)



Why we need to be efficient

Events:

- Page views
- Curation activities
- Search

Myth 3: We’re done with the UI

Reality: the UI now consists of a mixture of JavaScript legacy code and React-based elements, needs updating

- New components are being developed or designed as we speak
  - myADS notifications
  - Libraries ownership management and sharing
  - Recommendations
- Code should be refactored and improved
  - Remove legacy code and develop under single framework
  - Build in functionality which will allow us to perform experiments (A/B testing, personalizations)
  - Improve speed and user experience
  - Improve mobile experience
  - Improve accessibility

Program for 2021-25

- Complete system transition, update back-end components
  - Both a “problem” and an “opportunity,” so make the best of it
  - Will take most of ADS’s oxygen for the next couple of years
- Curation and Ingest Improvements
  - Continuous improvement of infrastructure focusing on reliability, currency, scalability
  - Continuous improvement of content, indexing fulltext, cited literature, software, data products
  - Implement community practices for content updates and versioning (record aggregation)
  - Support metadata enrichment through supervised machine learning (e.g. affiliations, keywords)
- New Efforts and Initiatives
  - Expand data linking content and framework to use industry standards (DataCite, Scholix)
  - Expand use of machine learning to support recommendations, personalization, visualizations
  - Improve text mining capabilities (entity detection, anchor text embedding, topic analysis)
  - Build dashboard functionality for several use cases (single researcher, institutional metrics, etc)
  - Support collaborative authoring and research environments, e-publishing platforms