astrophysics data system

Backoffice and DevOps Updates

Fernanda de Macedo Alves, Jean-Claude Paquin, Mugdha Polimera, Thomas Allen, Taylor Jacovich and the ADS Team

ADS Users Group Meeting, 05-06 Dec. 2024

Smithsonian Institution

CENTER FOR ASTROPHYSICS HARVARD & SMITHSONIAN

SciX

NASA Partner

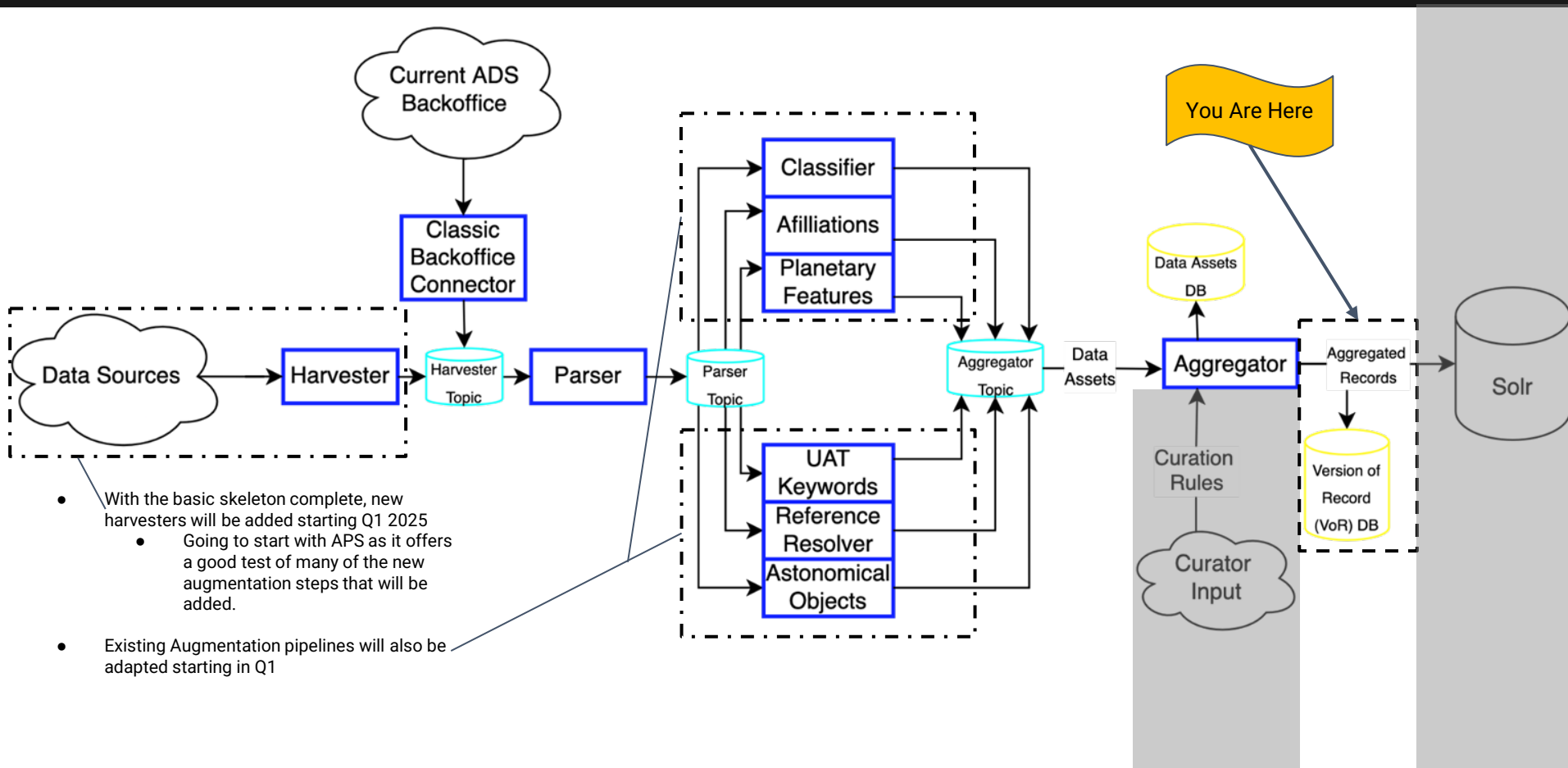
Introduction

- Honeycomb
  - State of Honeycomb
  - Software Updates
- The Move to ITS
- New Hardware
- ADSIngestParser
- ADSClassifierPipeline
- Library Annotations
- Search
- Additional Updates

The State of Honeycomb

- **Honeycomb is our next generation data ingestion architecture**
  - Important part of transitioning away from legacy systems
  - Modernizes the backoffice
  - Allows us to
    - Harvest more data more efficiently
    - Do more with the data we harvest
- **A test pipeline is currently deployed to the ADS Brain cluster**
  - Containerized deployment running on k3s kubernetes distribution
  - It contains all pipeline pieces through the Aggregator
    - Currently works to harvest ArXiv metadata
- **Need harvesters**
  - Currently have parsers covering a large portion of data sources
  - Need to add harvester to match
- **Transition Augmentation Pipelines**
  - Currently have several pipelines to transition
    - Some have been built with Honeycomb in mind
    - Some will need additional work
- **Curation Workflow**
  - Currently discussing with curators the best way to allow them to correct and otherwise modify ADS holdings

The State of Honeycomb



- With the basic skeleton complete, new harvesters will be added starting Q1 2025
  - Going to start with APS as it offers a good test of many of the new augmentation steps that will be added.
- Existing Augmentation pipelines will also be adapted starting in Q1

ADSIngestParser

- **Overview**
  - New Python framework for high-throughput parsing of input content
  - Improve handling of data over ADS classic perl modules
  - `ADSIngestParser` and `ingest_data_model` are building blocks for new aggregator (e.g., arXiv records will be parsed using the new `adsingestp.parsers.dubcore` parser)
- **Progress:**
  - Approximately ⅔ of all data providers being parsed with new parsers in production, comprising approximately 75% of records
  - Once Crossref parsing is fully transitioned, that will climb to over 90%
  - All major publishers in production: IOP, AIP, MDPI, COPERNICUS, VERSITA, IUCR, SPRINGER, APS, EDP, Science, T&F
  - Custom parsers: DublinCore (ArXiv), Elsevier, Crossref, Wiley
- **Enhancements since ADSUG 2023:**
  - Native language handling for author names, improved tag/markup handling for text blocks, tracking publisher-provided UAT keywords, tracking institution IDs (GRID+ISNI), improved affiliation and collaboration parsing

Classifier Pipeline

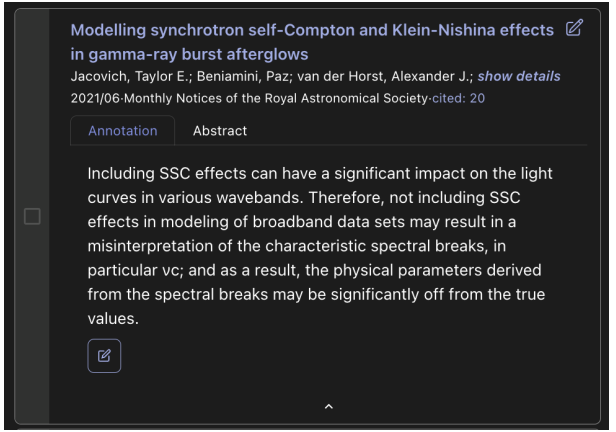
- The Classifier Pipeline will automate the placement of records into appropriate collections
  - Astronomy, Planetary Sciences, Heliophysics, Earth Science, Physics, Other
  - Collections allow relevancy boosts in domain specific searches
  - Designed to operate with or without a curator in the loop

- Development in Three phases
  - Phase 1: Use a fine-tuned language model to assign collections
  - Phase 2: Include journal based heuristics
  - Phase 3: Include citation graph information

- Development of phase 1 is nearing completion
  - Will have functional version in development environment this year (Q4 2024)
  - Stand-alone classification script currently operating in production environment
  - Current design allows the Classifier Pipeline to interface with existing and upcoming architecture

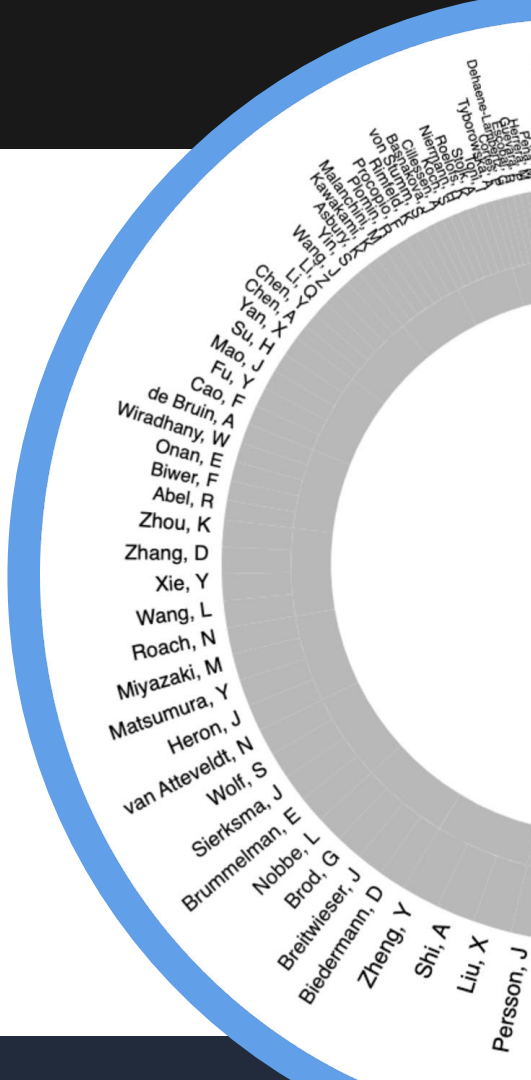
Library Annotations

- Collaboration tool that allows users to write and share notes inside libraries
  - This is a feature that was requested by users
  - It's only accessible on **scixplorer.org**
  - Each paper inside the library can have one corresponding note
  - Notes can be **read by anyone who has access to the library**
  - Only users with writing permissions can **write** notes
  - API documentation:
    - How to get all notes inside a library: https://scixplorer.org/help/api/api-docs.html#get-/biblib/libraries/-library_id-
    - Adding, deleting, getting and updating notes: https://scixplorer.org/help/api/api-docs.html#post-/biblib/notes/-library_id-/-document_id-



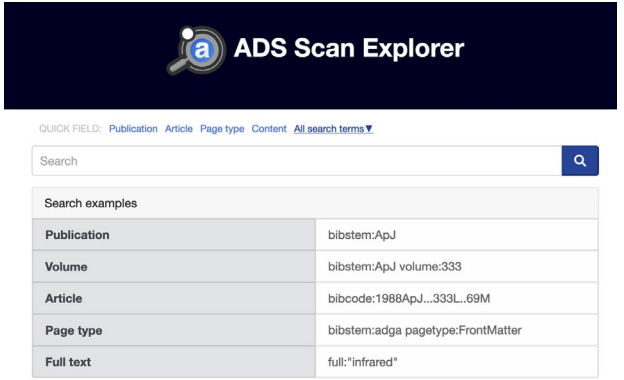
Search Updates

- Planetary feature names field
  - Users can now search for references to craters and other features
- Solr 9 upgrade
  - We've fully upgraded our search indexes from Solr 7 to Solr 9, decreasing query times & index size
  - Solr 9 also comes with a suite of features that will enable faster bug fixes and improved service reliability
- Testing against user queries
  - We now test against a sample of ~13k user queries to ensure search updates produce no major regressions in functionality



Additional Backend Updates

- API-Gateway
  - Winterway began work to replace the existing python 2.7 based API Gateway in 2023
    - It is a complete ground up rewrite of the Gateway that will allow to offer enhanced metrics and more fine-grained control over user access tokens and rate limits
  - Their latest version is currently in the ADS development environment
    - It is slated to go to production by 12/9/2024
- ADSScanExplorerService
  - The python 3 revision of the ScanExplorerService has received significant upgrades on its journey to production.
    - Pdf download efficiency has been increased by the use of pre-generated files for articles
    - On-the-fly generation of collection pdfs has been streamlined to allow for 100 pages to be downloaded at a time.
    - Several bugs have been squashed and the UI has also received updates to match SciX.



Changes to the Software Stack

- WEKA is being dropped in favor of one of Rook/Ceph or Quobyte
  - Ran into unresolvable hardware incompatibilities
- Rook/Ceph is Open Source Software while Quobyte is proprietary and requires a license

- Both Alternatives are built to run directly in kubernetes
  - Support for provisioning storage to running pods direct from Filesystem
  - Both have native object storage (removes the need for minIO, previous selected object gateway)
  - Lower CPU and memory overhead than WEKA
  - No need for anything to be installed in host OS

New Hardware

- Acquiring two major pieces of hardware
  - 2 GPU servers
    - Each has 8 Nvidia L40S GPUs (vs. 2 V100s in adsnlp)
    - Will be able to leverage storage on ADS Brain cluster
    - Allows the construction of production ready Machine Learning and AI backoffice pipelines
  - An additional upscaled, high memory node for hosting high consumption applications
    - Will slot into existing ADS Brain cluster
    - Offers sufficient resources to run backoffice database as well as backoffice search engine
    - Expands available raw storage space on cluster by ~15 TB.

The Move to ITS

- We have begun the transition of our backoffice infrastructure from being managed by the Syshelp team, to being managed by the CfA IT Services (ITS) team.
  - Catalyzed by CfA internal reorganization
  - ITS took over management of the ADS Brain cluster prior to the beginning of Honeycomb testing
  - ITS has taken over management of the Netapp file server that archives most backoffice data
  - ITS will begin transitioning the remaining ADS servers starting in 2025
- ITS has been involved in new hardware acquisitions
  - Helps us confirm new hardware will work within their management framework as well as their power and cooling budgets.
- ITS has started migrating ADS staff user accounts
  - Users still retain their current syshelp managed accounts during the transition period.

A Post-mortem of the September Incident

- During the period of approximately September 27th-October 8th ADS search became unreachable to the majority of users
  - This was caused by an ADS service generating extraneous calls to second-order search operators (ie. similar() or citations())
    - The high volume of second order operators overwhelmed the searchers and caused them to crash
    - Diagnosis was complicated by the fact that these requests were generated as a byproduct of legitimate user requests and with legitimate access tokens
  - Once the issue was identified, the service was reverted to a version that did not have the issue until a permanent patch could be deployed.
- To prevent recurrence of this issue we have added or will add the following mitigation strategies
  - We have added additional alert mechanisms to our monitoring software specifically aimed at identifying these types of issues.
  - We have modified the readiness probe on the searchers to better identify when a searcher is bogging down due to a high request volume.
  - We are working to implement a more sophisticated rate limiting scheme that should limit high volumes of computationally expensive requests.