Curation Update

Carolyn Grant and the ADS Team

ADS Users Group Meeting, 19-20 Nov. 2020

Search by Affiliation Update

- Background
  - Years of affiliation data cleanup, normalization, assigning IDs

A03543 Attila (Jozsef) Univ Szedged (Hungary)
A03543 Attila Jozsef University, Szeged
A03543 Department of Experimental Physics University of Szeged Szeged Hungary
A03543 Department of Experimental Physics, University of Szeged, Dóm tér 9, H-6723 Szeged, Hungary
A03543 Department of Experimental Physics, University of Szeged, Dóm tér 9, Szeged, Hungary
A03543 Department of Experimental Physics, University of Szeged, Szeged, Dóm tér 9, 6720 Hungary
A03543 Department of Optics, JATE University, Szeged, Hungary
A03543 Department of Theoretical Physics, University of Szeged, Tisza Lajos krt 84-86, Szeged 6720, Hungary and Department of Experimental Physics, University of Szeged, Dóm tér 9, Szeged 6720, Hungary
A03543 Departments of Theoretical and Experimental Physics, University of Szeged, Szeged, 6720 Dóm tér 9., Hungary
A03543 Dept Experimental Phys JATE Univ Szeged Dom ter H Szeged Hungary
A03543 Dept Experimental Phys Univ Szeged Szeged Dom ter Hungary
A03543 Dept Optics & Quantum Electronics Univ Szeged406 Szeged Hungary
A03543 Dept Optics Quantum Electronics Univ Szeged H Pf Szeged Hungary
A03543 Dept Optics Quantum Electronics Univ Szeged PO Box: H Szeged Hungary
A03543 Dept Phys Chem Materials Sciences Univ Szeged
A03543 Dept Theoretical Phys Univ Szeged
A03543 Dept Theoretical Phys Univ Szeged Hungary
A03543 Dept Theoretical Phys Univ Szeged Tisza L krt Szeged Hungary H
A03543 JATE Univ Research Group Laser Phys Hungary
A03543 Jozsef Attila Tudomanyegyetem
A03543 On leave from the Department of Optics and Quantum Electronics, University of Szeged, Dóm tér 9, Szeged, Hungary

Search by Affiliation Update

- Background
  - Years of affiliation data cleanup, normalization, assigning IDs
  - Filter by institutions enabled Jan. 2019

ads

QUICK FIELD: Author First Author Abstract Year Fulltext All Search Terms

Start New Search

inst:"U Toronto"

Your search returned **35,918** results

AUTHORS
COLLECTIONS
REFEREED
INSTITUTIONS

- U Toronto 35.9k
  - U Toronto 20.8k
  - Dep Phy 8.4k
  - CITA 2.9k
  - Dep Ast Astrop 2.6k
  - Dunlap Inst 884

more

Show highlights Show abstracts Hide Sidebars Go To Bottom

1 2021NIMPA.98564661A 2021/01 cited: 2
The MATHUSLA test stand
Alidra, Maf; Alpigiani, Cristiano; Ball, Austin *and 23 more*

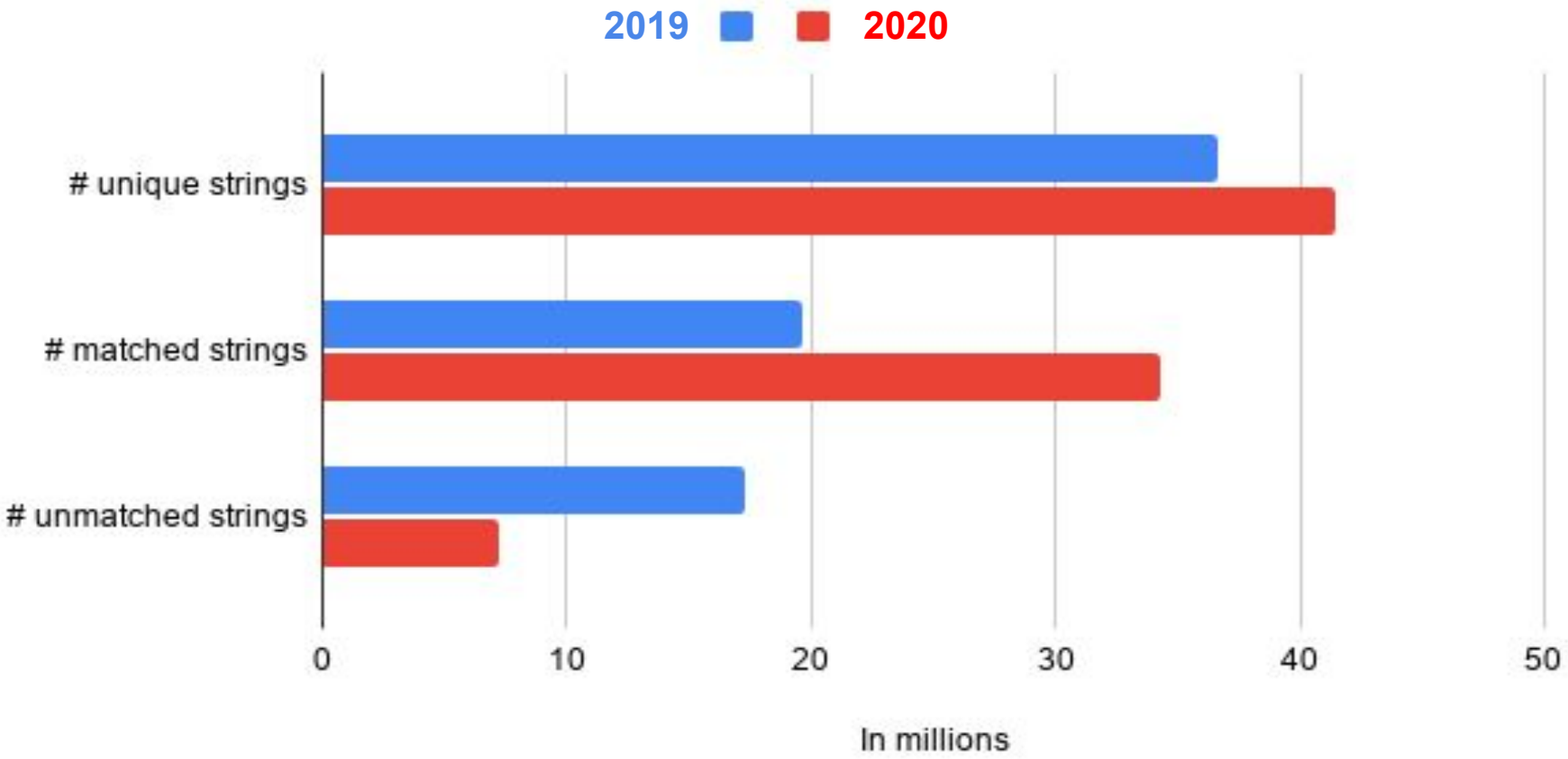
2 2021JDE...271..280E 2021/01
Solvability in the sense of sequences for some fourth order non-Fredholm operators
Efendiev, Messoud; Vougalter, Vitali

3 2020NucFu..60l6030D 2020/12
Manipulation of E×B drifts in a slot divertor with advanced shaping to optimize detachment
Du, Hailong; Guo, H. Y.; Stangeby, P. C. *and 4 more*

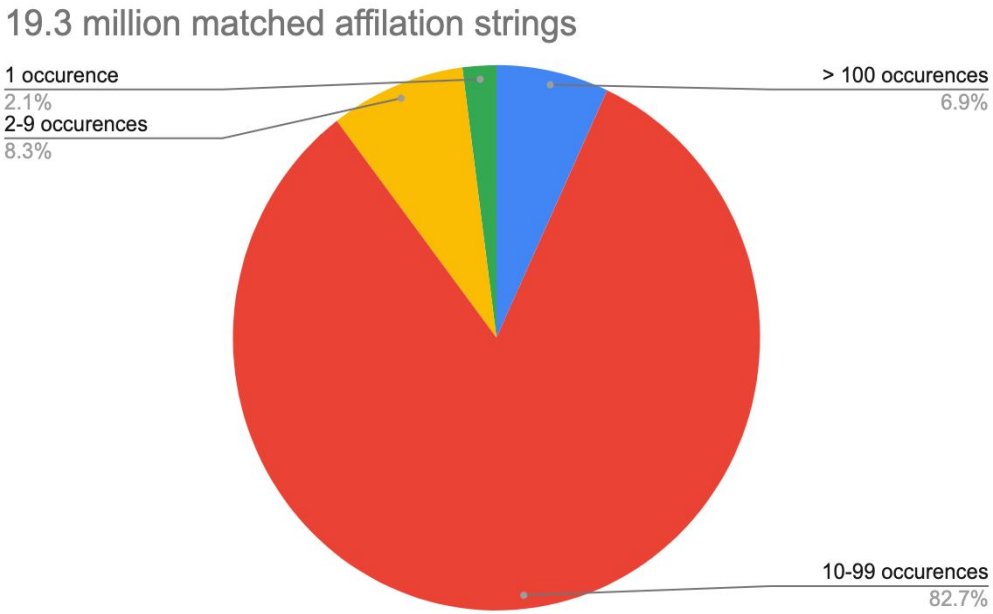
Search by Affiliation Update

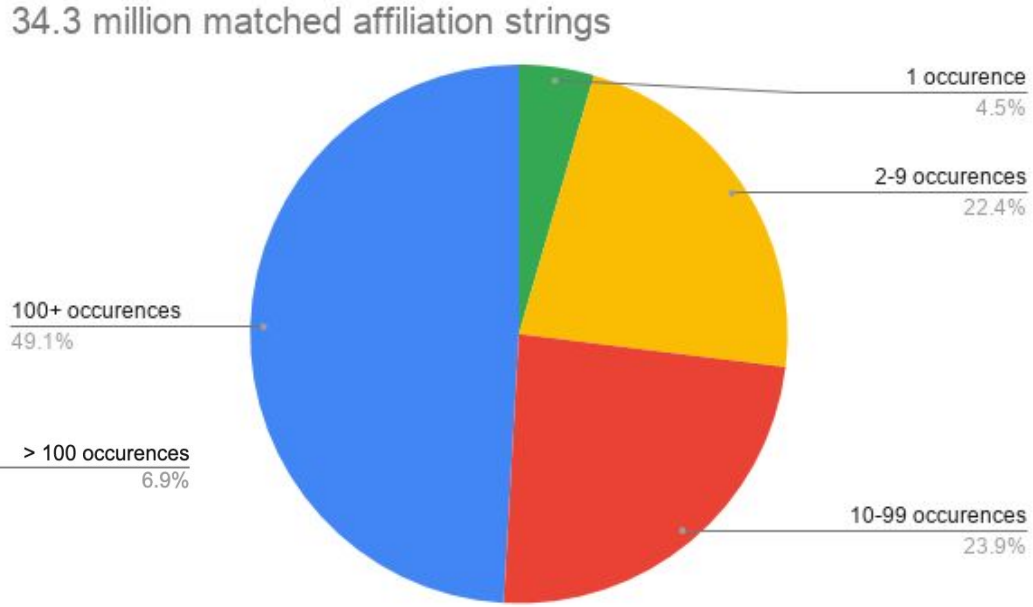
- Background
  - Years of affiliation data cleanup, normalization, assigning IDs
  - Filter by institution enabled Jan. 2019
- Update of coverage
  - 10% increase in identified institutions (6400 -> 7000)
  - 14% increase in # unique strings (36.5 million -> 41.5 million)
  - 32% increase in # matched strings! (53% -> 88%)

Affiliation Strings



2019

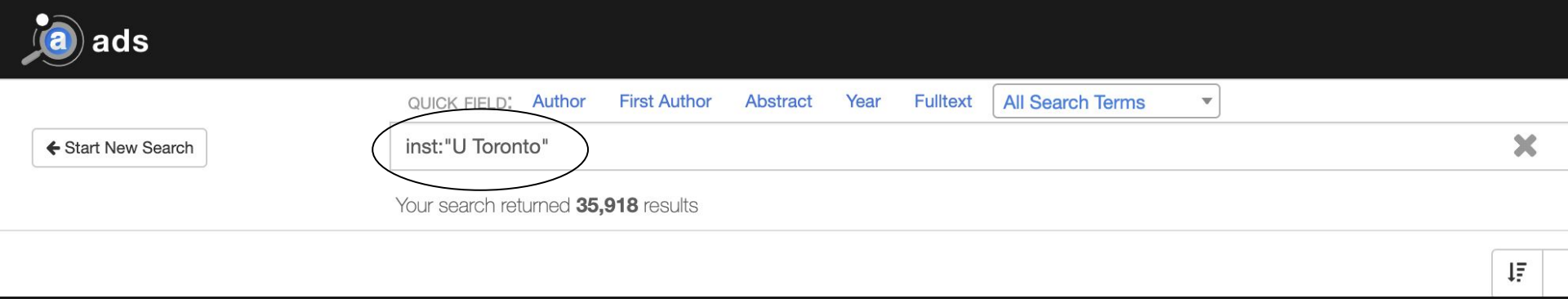

19.3 million matched affilation strings
1 occurence
2.1%
2-9 occurences
8.3%
> 100 occurences
6.9%
10-99 occurences
82.7%

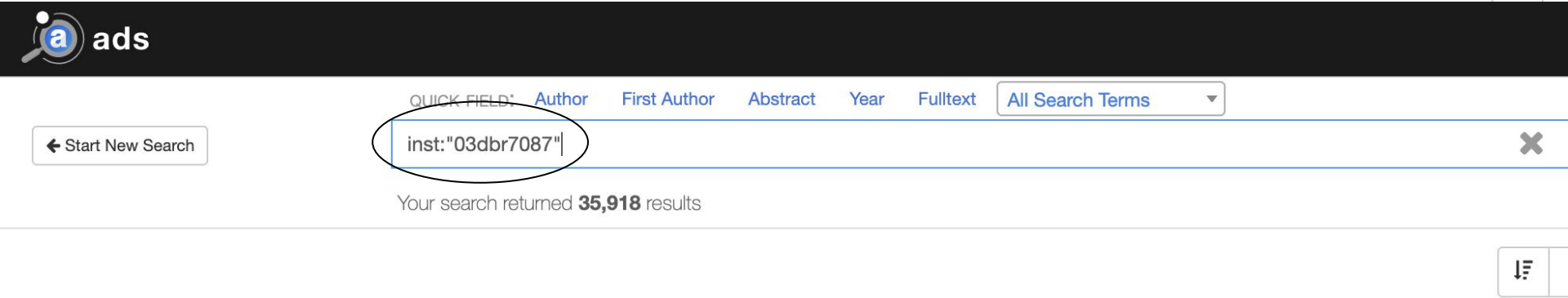

34.3 million matched affiliation strings
1 occurence
4.5%
2-9 occurences
22.4%
100+ occurences
49.1%
10-99 occurences
23.9%

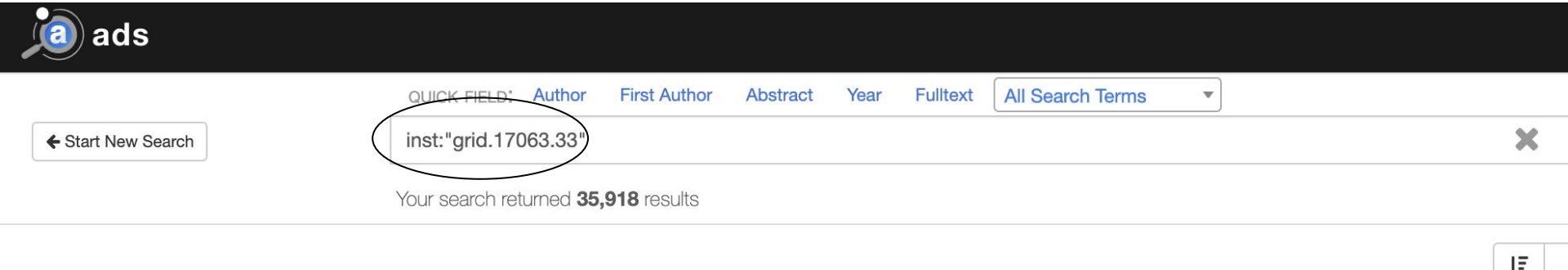
2020

Search by Affiliation Update

- Background
  - Years of affiliation data cleanup, normalization, assigning IDs
  - Filter by institution enabled Jan. 2019
- Update of coverage
  - 10% increase in identified institutions (6400 -> 7000)
  - 13% increase in # unique strings (36.9 million -> 41.5 million)
  - 30% increase in # matched strings! (53% -> 83%)
- Integration with ROR (Research Organization Registry)
  - Community-led effort for assigning unique IDs to every research org
  - Helping drive the effort to resolve to department-level
  - https://ucd-library.github.io/ror-extend-demo/details.html


ads
QUICK FIELD: Author First Author Abstract Year Fulltext All Search Terms
Start New Search
inst:"U Toronto"
Your search returned 35,918 results


ads
QUICK FIELD: Author First Author Abstract Year Fulltext All Search Terms
Start New Search
inst:"03dbr7087"
Your search returned 35,918 results


ads
QUICK FIELD: Author First Author Abstract Year Fulltext All Search Terms
Start New Search
inst:"grid.17063.33"
Your search returned 35,918 results

Breaking Dependencies on Classic

1. Data extraction

2. Data collection

3. Data indexing

4. Data storage

Breaking Dependencies on Classic

1. Data extraction

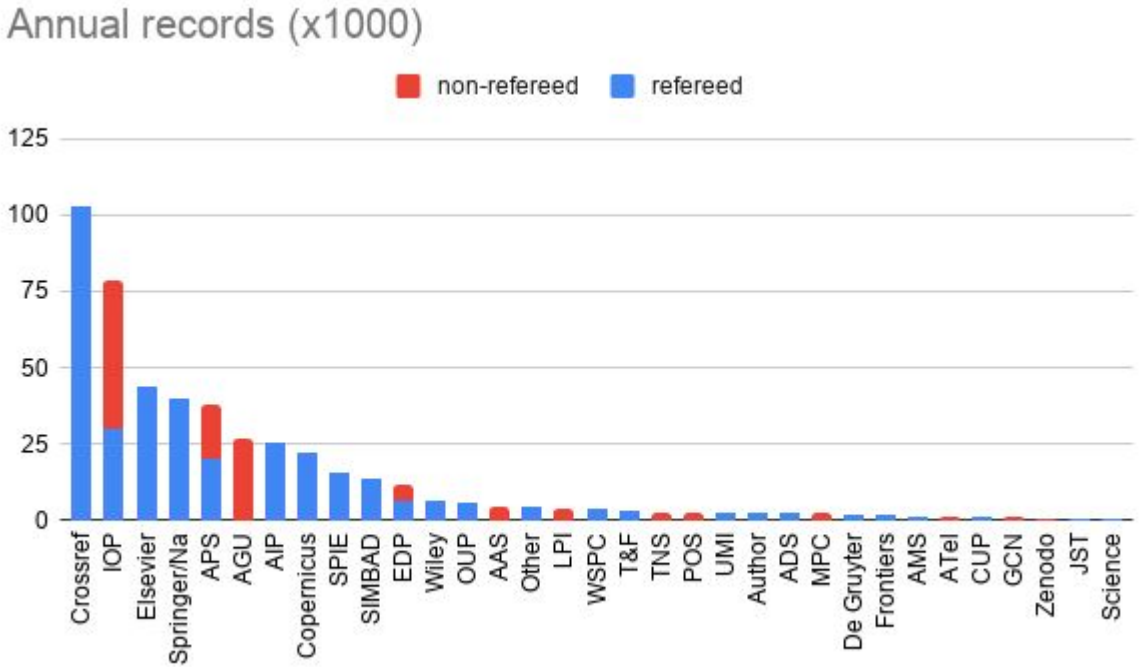
- **The data**
  - Annually ~500K records, 10-12 million citation pairs
  - Data from multiple sources
  - Need validation before input
  - Matching existing records (tmp -> pub; arxiv -> pub; SIMBAD)
  - Enabling smaller journals and conferences by keeping format simple
- **Porting legacy code from perl to python**
  - Approximately 40 versions of input formats
  - JATS - about 12, but not all JATS are equal (most already ported)
  - Springer/Elsevier/Wiley non-JATS
  - Crossref basic data, could possibly improve via API

Pyingest more configurable, reusable library of parsers

Annual records per publisher

There is not a direct relationship between the number of records per publisher and the difficulty of processing the data or the importance of including the journal. Other considerations include:

- Importance to the field
- Quality/depth of the data
- Amount of handwork



Breaking Dependencies on Classic

2. Data collection

- **Harvesting**
  - Weekly harvest from 8-10 sources
  - Dependence on adsftp problematic
- **External links**
  - Decoupled from the rest of the metadata
  - No validation
- **References and full text**
  - Decoupled from the rest of the metadata
  - Flat file storage tied to bibcodes

Breaking Dependencies on Classic

3. Data indexing

- **Tied to weekly curation workflow**
  - Higher cadences currently not supported
  - Preprint matching slows down physics update
- **Expand direct ingest**
  - Arxiv ~750/day
  - Astronomy ~300/day
  - Physics ~3000/day
- **Break dependency on bibcode**
  - Increase early content
  - Facilitate new content
  - Simplify including non-standard content

Breaking Dependencies on Classic

4. Data storage

- **Storage deeply entwined with classic architecture**
- **Define a new storage system**
  - Storing: bib data, ref data, full text data, usage data, link data
  - Updating: tracking changes, deletions, provenance. Auto vs. manual
  - Editing: single vs. batch changes; bib vs. non-bib data
- **Define a new data model**
  - Ease of validation with schema
  - No longer dealing with purely bibliographic data (e.g. software and data products)

First steps in Modernizing Curation Workflows

- Best Practices
  - Curation task force meeting regularly
  - Designing curation framework based on data requirements
  - Version Control for all curation code
- Feedback Forms
  - Old feedback forms disabled
  - New forms using API to generate content from the database
  - More robust and increased functionality (author ordering, ORCID)

Journals Database Update

- Summary
  - Central storage for ADS Journals holdings data
  - Specialized holdings information for ADS Curators and external users (e.g. librarians)
  - Reference sources, fulltext holdings, journal histories (publisher, volume/issue availability, etc)
  - Replacement for some Classic curation and indexing data
- Current Status
  - Working on record histories & update triggers
  - Pipeline deployment container in development
- To do:
  - Curator interface via external application (Google Sheets or equiv)
  - API for external queries

Thank you!

Carolyn Grant and the ADS Team

cgrant@cfa.harvard.edu