The NASA Astrophysics Data System: Content and Curation

Edwin Henneken, Carolyn Grant, Matthew Templeton, Donna Thompson, Steve McDonald

ADS Users Group Meeting
November 29, 2018
Who uses ADS?

Librarians

Researchers

Services (SIMBAD, NED, ...)

Integrations (API)

Scholarly Service

Public Service

"Search"

Primary Access

Secondary Access

"Linking"

Whatever the method of access, people access the ADS because of its content... so, what's in the ADS?
What’s in ADS?

As of 11/2018, ADS contains:

● 14.2 million records
● 110 million citation pairs
● 5.2 million scanned pages

Annually, we typically ingest:

● ~1 million records
● ~10 million new citation pairs

Division is approximately:

● 10:1 Physics:Astronomy
● 10:1 Journal article:Conference article (AST)
What users ask

- Typical content requests:
  - Do you have full text for a certain journal?
  - My library is downsizing, can we discard paper copies of journal x?
  - Paper x is not in the ADS, can you tell me why?

- Developed formal guidelines for inclusion in the ADS
Comprehensive Journal Database

- What journals are included?
- What is known about these journals?
  - Years
  - Open access
  - Full text available?
  - Completion
  - ISSN /e-ISSN and other standard identifiers (DOIs, etc.)
  - Change of publishers
  - Latest update
  - How does ADS get the records?
  - Refereed status
Search by Affiliation

One year ago...

Current efforts:

- Creation of canonical institution names, ids and abbreviations for facet (3300 inst + 2200 divisions) (3800 inst + 2800 departments)
- Assignment of raw strings to institutional ids (93% for astronomy, 60% for physics) 70% for physics)
- Development of python routine to match new and unmatched affiliations (90%) (99%)

Still to do:

- More physics identification ✓
- Development of user interface ✓
- Implementation of input pipeline ✓
- Development of curation tool
Affiliation Curation

Challenges:

● Sheer numbers:
  ADS contains ~34 million affiliations strings (4.5m AST, 29m PHY)
  Reduced tenfold once cleaned and uniqued
  2.25 of 3.4 million strings matched to an institution (93% AST, 70% PHY)

● Inconsistent naming:
  grid.473002.2, University of the State of Paraná, Paranavaí, Brazil
  grid.441795.a, Universidade Estadual do Norte do Paraná, Jacarezinho, Brazil
  grid.441662.3, State University of West Paraná, Cascavel, Brazil

● Multiple and/or changing names:
  State University of West Parana
  UNIOESTE
  Universidade Oeste do Paraná
  Universidade Estadual do Oeste do Paraná

● Nothing systematic from publishers (only 1 gives us institutional identifiers)
● Incomplete or ambiguous affiliations
● No existing standard or registry
Institutional Identifiers

- **Ringgold**
  - > 400,000 institutions (< 200,000 academic)
  - 165,000 parent (35,000 academic)
  - 500,000 correlated with ISNI

- **ISNI**
  - 500,000?

- **GRID**
  - 90,000 (18,000 academic)
  - 14,000 relationships

- **OrgRef**
  - 32,000

- **INSPIRE**
  - 11,000

- **ADS**
  - 6600 institutions
  - 3800 parents, 2800 children
  - ~1000 ADS-created
New Affiliation facet!

AUTHORS
- Kurtz, M 319
- Accomazzi, A 155
- Murray, S 151
- Grant, C 150
- Eichhorn, G 134

COLLECTIONS

REFEREED

AFFILIATIONS
- Si 187
- Harvard U 118
- Brown U 15
- U Turin 11
- CalTech 6

more

2 selected
- and
- or
- exclude

IPAC 3
JPL 2
Div Geo Plan Sci 1
Div Phy Math Ast 1
Affiliation Mapping and Augmentation

- Based on human-curated maps of affiliation strings and their canonical IDs (CSG)

- Two-step matching process
  - Lookup table of mapped strings
  - Text classification of unmatchable strings (scikit-learn)

- Augmentation
  - Identifying a text affiliation with a canonical ID
  - Creating searchable hierarchical facets for affiliations
Automated Matching

● Direct matching
  ○ Match a normalized affil string from metadata to kv-pairs of
    {normalized string: Affil ID}
  ○ If match exists, assign Affil ID
  ○ If no match exists, save for the text classifier

● Text classification
  ○ 6000+ classes (IDs), with many n-grams that characterize each class
  ○ Result: probability of best match for each of 6000+ classes
  ○ Curators review probability scores of [x] and higher
  ○ Good matches are added to the direct match dictionary
Direct Matching Data

A01400  60 Garden St Cambridge MA 02138 USA
A01400  60 Garden St., Cambridge, MA 02138, USA
A01400  60 Garden Street 02138 Cambridge MA USA
A01400  60 Garden Street Cambridge MA 02138 USA
...  
A01400  Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
A01400  Minor Planet Center, Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, United States
A01400  AMP Harvard Smithsonian CfA Cambridge MA
A01400  AMP Harvard Smithsonian CfA Cambridge MA USA
A01400  AMP Harvard-Smithsonian CfA Cambridge MA USA
A01400  AMP, Harvard-Smithsonian CfA, Cambridge, MA, USA
...  
A01400  Astrophysics Data System, Harvard-Smithsonian Center for Astrophysics 60 Garden Street Cambridge, MA 02138, USA
A01400  Astrophysics Data System, Harvard-Smithsonian Center for Astrophysics, 60 Garden St., Cambridge, MA 02138, USA
A01400  At Harvard Smithsonian Center Astrophys
A01400  At the Harvard Smithsonian Center for Astrophysics
A01400  At the Harvard-Smithsonian Center for Astrophysics
...
Attila (Jozsef) University, Szeged

Department of Experimental Physics, University of Szeged, Dóm tér 9, H-6723 Szeged, Hungary

Department of Experimental Physics, University of Szeged, Dóm tér 9, Szeged, Hungary

Department of Experimental Physics, University of Szeged, Szeged, Dóm tér 9, 6720 Hungary

Department of Optics, JATE University, Szeged, Hungary

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

Departments of Theoretical and Experimental Physics, University of Szeged, Szeged, 6720 Dóm tér 9., Hungary

Dept Experimental Phys JATE Univ Szeged Dom ter H Szeged Hungary

Dept Experimental Phys Univ Szeged Szeged Dom ter Hungary

Dept Optics & Quantum Electronics Univ Szeged406 Szeged Hungary

Dept Optics Quantum Electronics Univ Szeged H Pf Szeged Hungary

Dept Optics Quantum Electronics Univ Szeged PO Box: H Szeged Hungary

Dept Phys Chem Materials Sciences Univ Szeged

Dept Theoretical Phys Univ Szeged

Dept Theoretical Phys Univ Szeged Hungary

Dept Theoretical Phys Univ Szeged Tisza L krt Szeged Hungary H

JATE Univ Research Group Laser Phys Hungary

Jozsef Attila Tudomanyegyetem

On leave from the Department of Optics and Quantum Electronics, University of Szeged, Dóm tér 9, Szeged, Hungary
Text Classification Challenges

Curating affiliations required substantial human effort to identify *millions* of strings. The matching process will still require human curation of machine-learning results to keep our augmented metadata error-free.

- Ambiguity (e.g. right institution, wrong department; short, incomplete, or non-affil data [“deceased” “email:”])
- Metadata issues (e.g. multiple affils merged into one; misspellings; misparsed OCR)
- Computer power (memory intensive)
University of Texas, Arlington, Department of Physics
Department of Mathematics, University of Texas at Arlington, Arlington, Texas, 76019,

Rowan University
Department of Chemistry and Physics, Rowan University and Department of Physics, University of Maryland

Electronics and Telecommunications Research Institute, Korea
Electronics and Telecommunications Research Institute, Taejon 305-350, Korea

Babson College
Deceased. [Note: “Deceased” is a common affiliation!]
Curational Use of the ADS

The ADS is used for a lot more than just "search"

- **Grants**: output / metrics
  - Finding grant-publication links is a challenge
    (astronomy journals are not well represented in FundRef)
- **Bibliographies**: curation / metrics
  - Citation reports for large collections is a challenge
  - Exporting data into local management tools is hard for e.g. librarians with little to no scripting skills
- **Data/Software**: links / metrics
  - Ongoing discussions with PDS on improving data linking
  - Citation/usage metrics for data non-existent (learn from software citations?)
Software Citations - Update

GitHub

Make software citable

Asclepias Broker System

Metadata registration

DataCite

Metadata retrieval

Citation events

Zenodo

Author instructions

Article indexing

astrophysics data system
Software Citations - Update

Preliminary results

Distinct versions 979
Total citation number 1935

Top 5

"Lmfit: Non-Linear Least-Square Minimization And Curve-Fitting For Python" 103 citations
"Triangle.Py V0.1.1" 70
"Lasagne: First Release." 64
"Corner.Py: Corner.Py V1.0.2" 32
"Seaborn: V0.5.0 (November 2014)" 22

What about (citation) aggregation?
Ingest Pipeline: Operational Status

- Running For Over A Year
- Sustained Data Parity With Classic
- No Persistent Missing Data or Bibcodes
- No Major Outages
  - Rare 1-Day Server Issues
- Data Quality Issues
  - Direct arXiv Ingest In Production
  - From User Reports: Missing Appendix Text, Some Series Fields, etc.
  - Late Links To Arxiv (Several Hours)
Ingest Pipeline: Future Plans

- The Price Of Data Parity Is Eternal Vigilance
- Affiliation Facets And Citations In Production
- Refactor Metrics/Citations Data Pipeline
- Increase Overall Flexibility
  - More Welcoming To New Data
- Reduce Classic Dependencies
- Eventually, Consider Streaming Data Pipelines
  - Support More Frequent Updates
  - Perhaps Kafka
Questions?
The NASA Astrophysics Data System: Content and Curation

Edwin Henneken, Carolyn Grant, Matthew Templeton, Donna Thompson, Steve McDonald

ADS Users Group Meeting
November 29, 2018
Ingest Pipeline: Behind The Scenes

Software

- RabbitMQ, Postgres, Solr And 6 Pipelines
  - Scripted Deployments To Containers
- All Data Flows Through “Master”
  - Feeds Solr Replicator, Resolver Service and AWS-Based SQL
  - 1TB in Local Postgres
- Specific Time Windows For Updating Production Data Stores

People

- Daily Issue Identification
- MetaData Errors Fixed Immediately
@message: "2018A&C....25..213C"

<table>
<thead>
<tr>
<th>Group total</th>
<th>13M</th>
<th>12M</th>
<th>13M</th>
<th>10M</th>
<th>7M</th>
<th>6M</th>
<th>11M</th>
<th>19M</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fraction of public traffic (%)</strong></td>
<td><strong>88%</strong></td>
<td><strong>84%</strong></td>
<td><strong>82%</strong></td>
<td><strong>78%</strong></td>
<td><strong>69%</strong></td>
<td><strong>70%</strong></td>
<td><strong>80%</strong></td>
<td><strong>91%</strong></td>
</tr>
</tbody>
</table>

![Graph showing search traffic from 2010 to 2017](image1)

<table>
<thead>
<tr>
<th>Group total</th>
<th>0.5M</th>
<th>0.6M</th>
<th>0.7M</th>
<th>0.6M</th>
<th>0.6M</th>
<th>0.3M</th>
<th>0.5M</th>
<th>0.6M</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fraction of public traffic (%)</strong></td>
<td><strong>4%</strong></td>
<td><strong>4%</strong></td>
<td><strong>4%</strong></td>
<td><strong>5%</strong></td>
<td><strong>6%</strong></td>
<td><strong>4%</strong></td>
<td><strong>3%</strong></td>
<td><strong>3%</strong></td>
</tr>
</tbody>
</table>

![Graph showing social traffic from 2010 to 2017](image2)

"Search"

"Social"