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

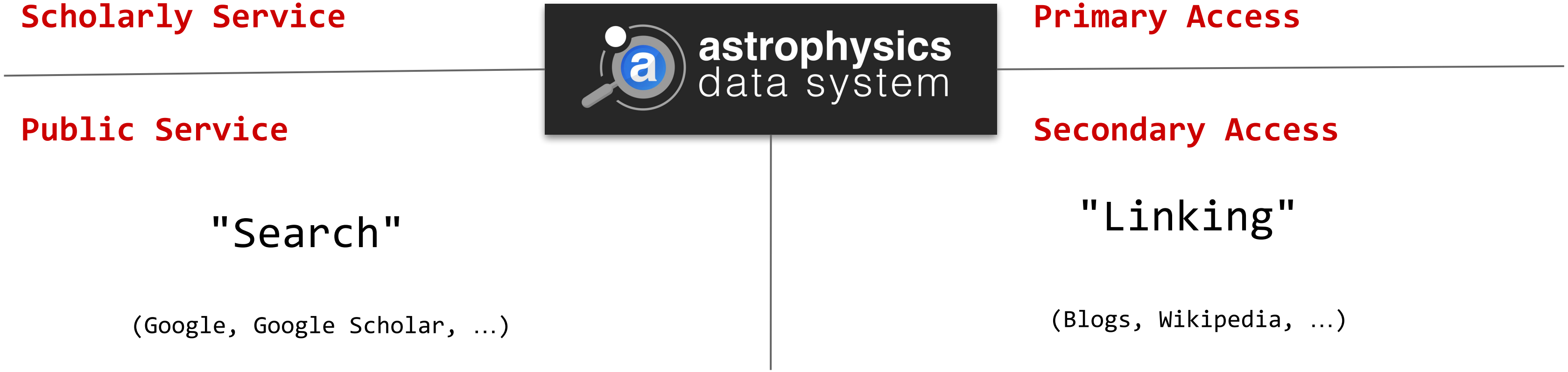
Who uses ADS?

Librarians

Services (SIMBAD, NED, …)

Researchers

Integrations (API)



Scholarly Service

Primary Access

Public Service

Secondary Access

"Search"

(Google, Google Scholar, …)

"Linking"

(Blogs, Wikipedia, …)

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

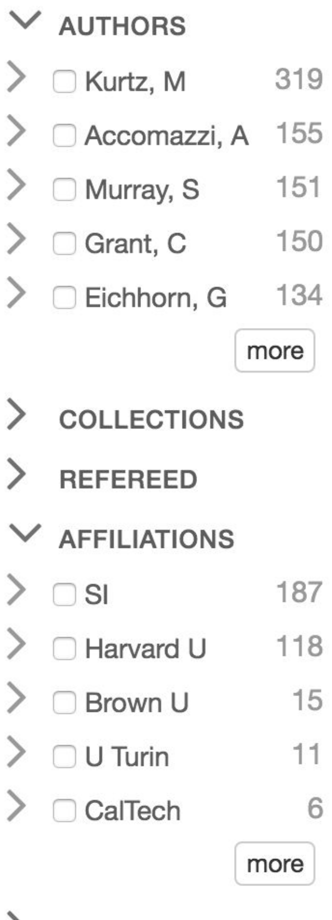
Challenges:

- 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

more

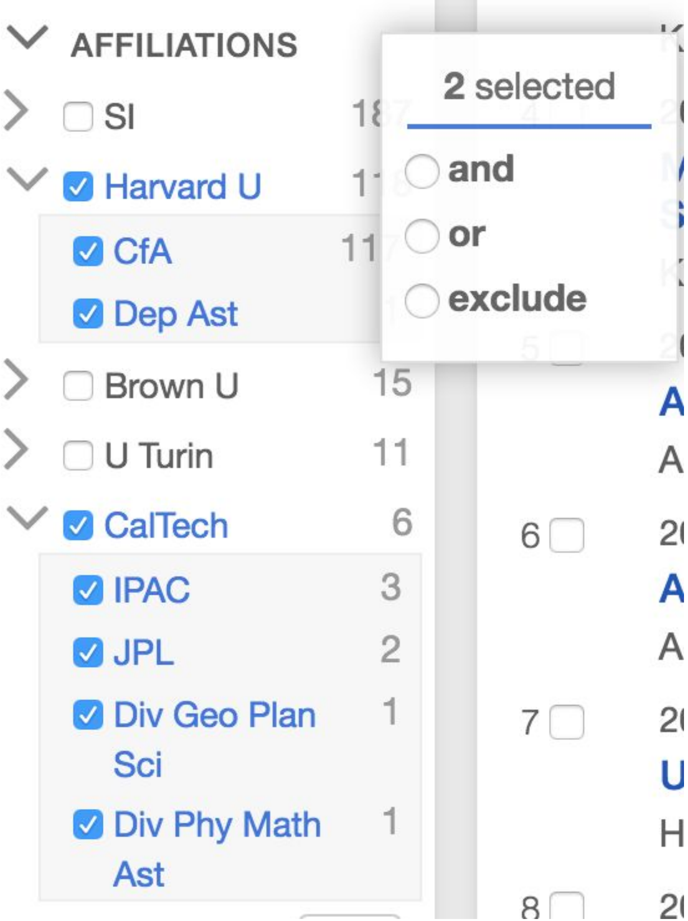
COLLECTIONS

REFEREED

AFFILIATIONS

□ SI	187
□ Harvard U	118
□ Brown U	15
□ U Turin	11
□ CalTech	6

more



AFFILIATIONS

□ SI	18
☑ Harvard U	1
☑ CfA	11
☑ Dep Ast	
□ Brown U	15
□ U Turin	11
☑ CalTech	6
☑ IPAC	3
☑ JPL	2
☑ Div Geo Plan Sci	1
☑ Div Phy Math Ast	1

2 selected

○ and

○ or

○ exclude

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
...
```

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, D&#243;m t&#233;r 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, D&#243;m t&#233;r 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, D&#243;m t&#233;r 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 &amp; 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

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 deptartment; short, incomplete, or non-affil data [“`deceased`” “`email:`”])
- Metadata issues (e.g. multiple affils merged into one; misspellings; misparsed OCR)
- Computer power (memory intensive)

0.966 University of Texas, Arlington, Department of Physics A11133

__Department of Mathematics__, University of Texas at Arlington, Arlington, Texas, 76019,

0.926 Rowan University A00607

Department of Chemistry and Physics, Rowan University __and Department of Physics, University of Maryland__

0.167 Electronics and Telecommunications Research Institute, Korea A04485

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

0.001 Babson College A00139

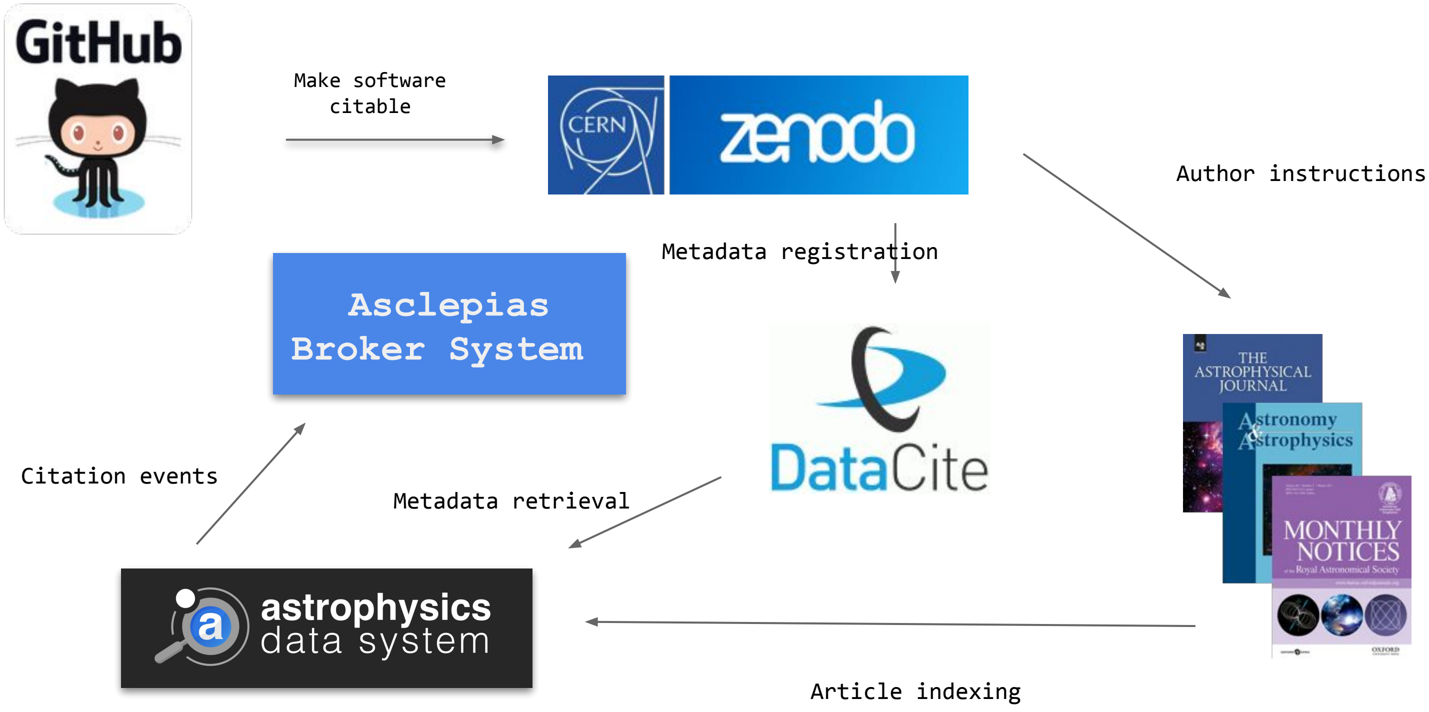
Deceased. **[Note: “Deceased” is a common affiliation!]**

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



Make software citable

Author instructions

Metadata registration

Asclepias Broker System

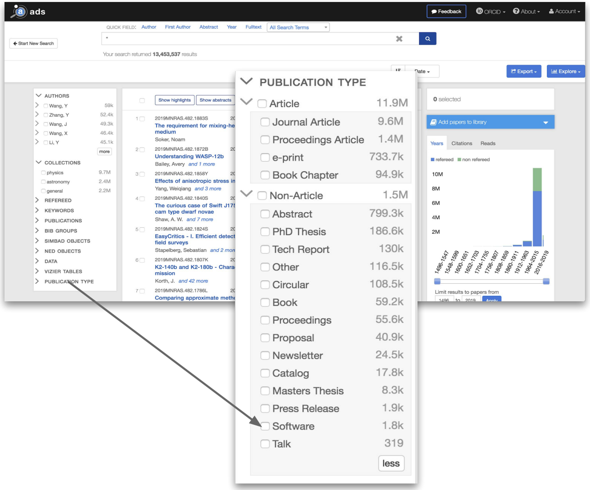
Citation events

Metadata retrieval

Article indexing

GitHub
CERN
zenodo
DataCite
THE ASTROPHYSICAL JOURNAL
Astronomy & Astrophysics
MONTHLY NOTICES of the Royal Astronomical Society
astrophysics data system

Software Citations - Update



ads

Feedback | ORCID | About | Account

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

Start New Search

Your search returned **13,453,537** results

Date | Export | Explore

AUTHORS
- □ Wang, Y 59k
- □ Zhang, Y 52.4k
- □ Wang, J 49.3k
- □ Wang, X 46.4k
- □ Li, Y 45.1k
- more

COLLECTIONS
- □ physics 9.7M
- □ astronomy 2.4M
- □ general 2.2M

REFEREED

KEYWORDS

PUBLICATIONS

BIB GROUPS

SIMBAD OBJECTS

NED OBJECTS

DATA

VIZIER TABLES

PUBLICATION TYPE

Show highlights | Show abstracts

1 □ 2019MNRAS.482.1883S
The requirement for mixing-he medium
Soker, Noam

2 □ 2019MNRAS.482.1872B
Understanding WASP-12b
Bailey, Avery *and 1 more*

3 □ 2019MNRAS.482.1858Y
Effects of anisotropic stress in
Yang, Weiqiang *and 3 more*

4 □ 2019MNRAS.482.1840S
The curious case of Swift J175 cam type dwarf novae
Shaw, A. W. *and 7 more*

5 □ 2019MNRAS.482.1824S
EasyCritics - I. Efficient detect field surveys
Stapelberg, Sebastian *and 2 more*

6 □ 2019MNRAS.482.1807K
K2-140b and K2-180b - Chara mission
Korth, J. *and 42 more*

7 □ 2019MNRAS.482.1786L
Comparing approximate meth

0 selected

Add papers to library

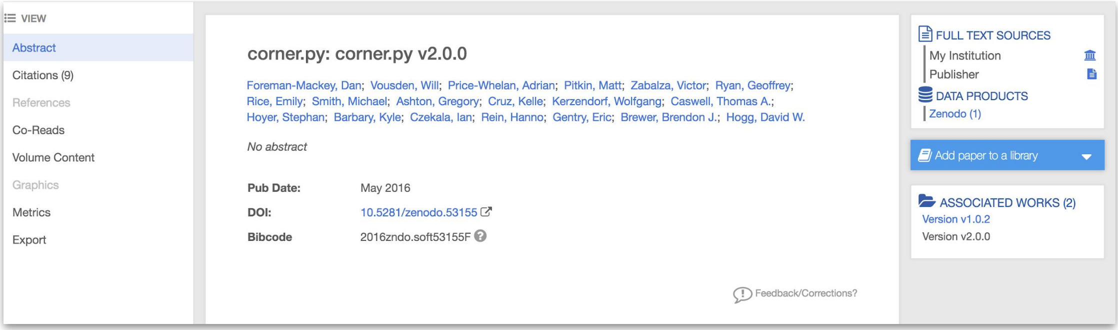
Years | Citations | Reads

Limit results to papers from 1496 to 2019 Apply

PUBLICATION TYPE

□ Article	11.9M
□ Journal Article	9.6M
□ Proceedings Article	1.4M
□ e-print	733.7k
□ Book Chapter	94.9k
□ Non-Article	1.5M
□ Abstract	799.3k
□ PhD Thesis	186.6k
□ Tech Report	130k
□ Other	116.5k
□ Circular	108.5k
□ Book	59.2k
□ Proceedings	55.6k
□ Proposal	40.9k
□ Newsletter	24.5k
□ Catalog	17.8k
□ Masters Thesis	8.3k
□ Press Release	1.9k
□ Software	1.8k
□ Talk	319

less



VIEW
- Abstract
- Citations (9)
- References
- Co-Reads
- Volume Content
- Graphics
- Metrics
- Export

corner.py: corner.py v2.0.0

Foreman-Mackey, Dan; Vousden, Will; Price-Whelan, Adrian; Pitkin, Matt; Zabalza, Victor; Ryan, Geoffrey; Rice, Emily; Smith, Michael; Ashton, Gregory; Cruz, Kelle; Kerzendorf, Wolfgang; Caswell, Thomas A.; Hoyer, Stephan; Barbary, Kyle; Czekala, Ian; Rein, Hanno; Gentry, Eric; Brewer, Brendon J.; Hogg, David W.

No abstract

Pub Date: May 2016

DOI: 10.5281/zenodo.53155

Bibcode 2016zndo.soft53155F

Feedback/Corrections?

FULL TEXT SOURCES
- My Institution
- Publisher

DATA PRODUCTS
- Zenodo (1)

Add paper to a library

ASSOCIATED WORKS (2)
- Version v1.0.2
- Version v2.0.0

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

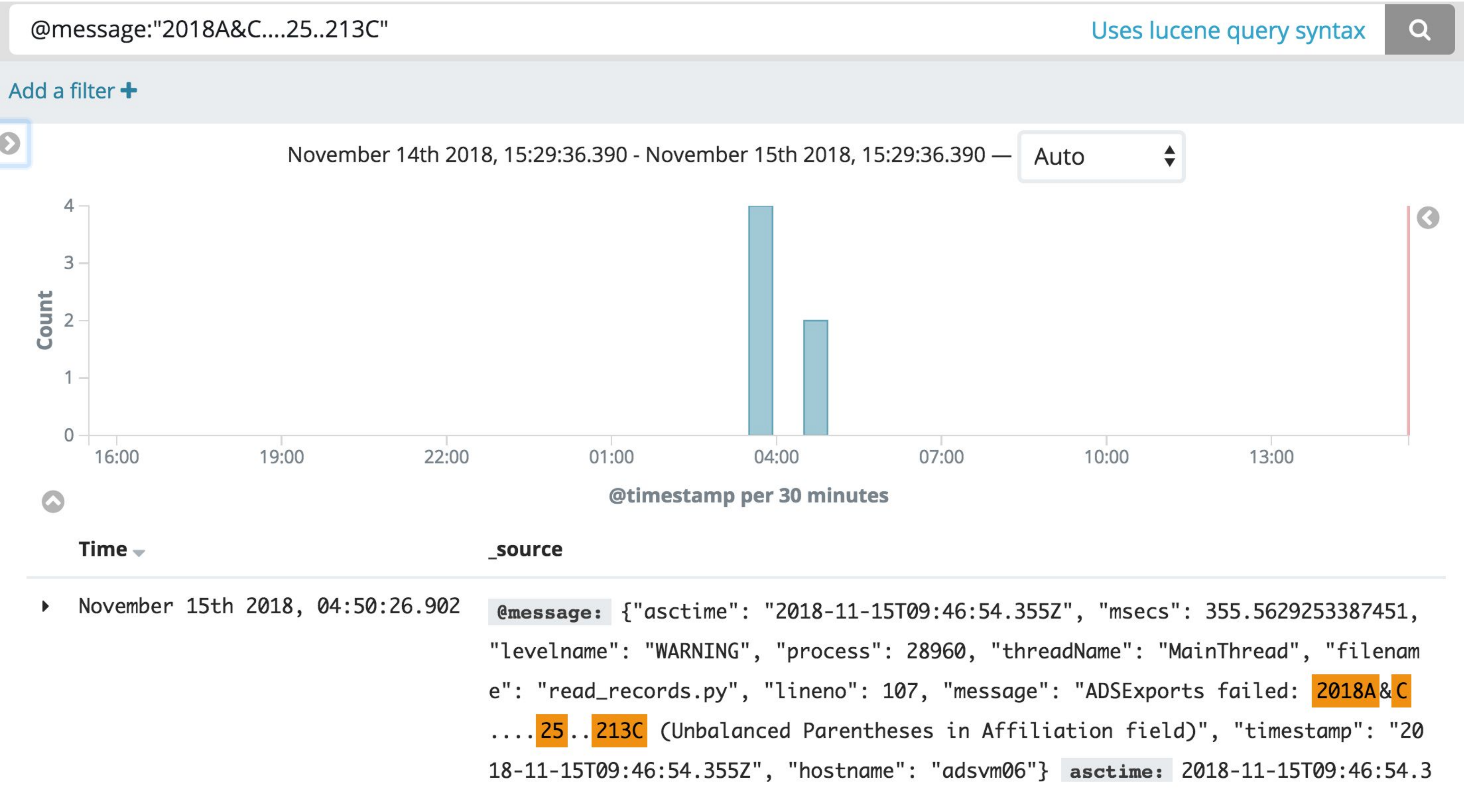
Ingest Pipeline: Behind The Scenes

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

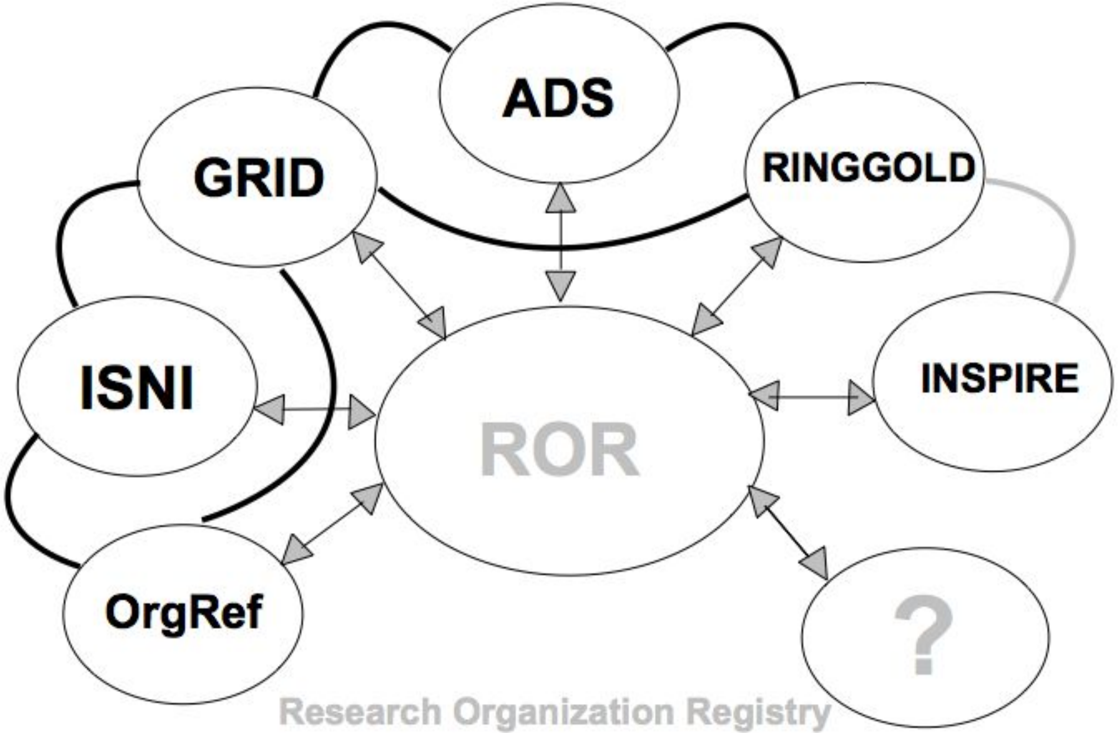


@message:"2018A&C....25..213C"

November 14th 2018, 15:29:36.390 - November 15th 2018, 15:29:36.390 — Auto

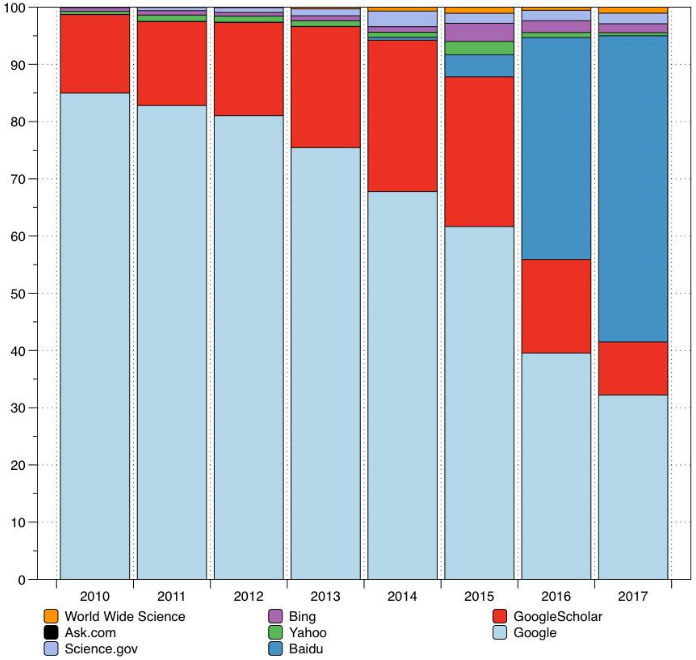
```
November 15th 2018, 04:50:26.902   @message: {"asctime": "2018-11-15T09:46:54.355Z", "msecs": 355.5629253387451,
"levelname": "WARNING", "process": 28960, "threadName": "MainThread", "filenam
e": "read_records.py", "lineno": 107, "message": "ADSExports failed: 2018A&C
....25..213C (Unbalanced Parentheses in Affiliation field)", "timestamp": "20
18-11-15T09:46:54.355Z", "hostname": "adsvm06"} asctime: 2018-11-15T09:46:54.3
```

Affiliation Identifiers Landscape

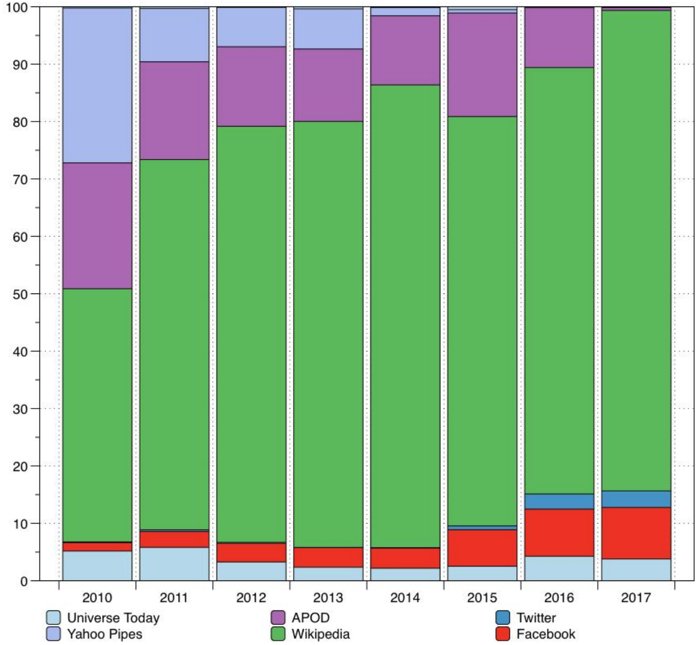


Group total	13M	12M	13M	10M	7M	6M	11M	19M
Fraction of public traffic (%)	88%	84%	82%	78%	69%	70%	80%	91%

0.5M	0.6M	0.7M	0.6M	0.6M	0.3M	0.5M	0.6M
4%	4%	4%	5%	6%	4%	3%	3%



"Search"



"Social"