Affiliations and Collaborations:
How can we improve author searches?

Carolyn Grant & the ADS Team
cgrant@cfa.harvard.edu

ADS Users Group Meeting - 11/2/2017

Affiliations

The data problem:

- Affiliations in ADS are free-form (> 30 million unique strings)
- Although fairly complete for recent refereed journals, most are missing for the grey literature and older material.
- Authors with multiple affiliations
- Inherent ambiguity (IoA)

Institute of Astronomy, Cambridge, UK
Institute of Astronomy, UNAM, Mexico
Institute of Astronomy, University of Tokyo
Institute of Astronomy, Moscow
Institute of Astronomy, Taiwan
Institute of Astronomy, University of Vienna
Institute of Astronomy, Bulgarian Academy of Sciences
Institute of Astronomy, University of Leuven

Affiliations

The usability problem:

- Searches by affiliation will inherently be incomplete - how to convey that to users?
- How best to integrate names, affiliations and ORCID ids in the user interface?
- Librarians and scientists potentially have different needs

Affiliations

Current efforts:

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

Affiliations (potentially messy UI)

Authors [show Affiliations]
[show ORCIDs]

- Smith, P (1000)
  - Smith, P (1000)
  - Smith, P A (100)
  - Smith, P N (10)
  - Smith, Peter (10)

Affiliations

- Harvard U (600)
  - Harvard U Dep Ast (300)
  - Harvard U Dep Phy (200)
  - CfA (80)
  - Harvard U Med Sch (40)

ORCID

- Smith, P (400)
  - Smith, P A [0000-0001-0001-0001] (300)
  - Smith, P B [0000-0001-0002-0003] (200)
  - Smith, Peter [0000-0001-0002-0003] (10)

Authors [hide Affiliations]
[show ORCIDs]

- Smith, P (1000)
  - Smith, P (1000)
    - Harvard U (60)
    - Yale U (40)
    - CfA (15)
  - Smith, P A (100)
    - U Mich (80)
  - Smith, P N (10)
    - Stanford U (8)
    - NASA (6)
  - Smith, Peter (10)
    - Harvard U (8)
    - CfA (2)

Authors [hide Affiliations]
[hide ORCIDs]

- Smith, P (1000)
  - Smith, P [0000-0001-0001-0001] (500)
    - U Mich (80)
    - Yale U (40)
    - NASA (15)
  - Smith, P [0000-0001-0002-0003] (500)
    - Harvard U (60)
    - CfA (15)
  - Smith, P A (100)
    - U Mich (80)
  - Smith, P N (10)
    - Stanford U (8)
    - NASA (6)
  - Smith, Peter [0000-0001-0002-0003] (10)
    - Harvard U (8)
    - CfA (2)

Affiliations

Still to do:

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

Collaborations

The data problem:

- Different publishers view collaborations differently, sometimes in author lists, sometimes in affiliations, sometimes in appendices, sometimes first in author list, sometimes last, etc.
- There may be many different names for the same collaboration (“The”, capitalization differences, abbreviations, greek characters, mixed case)
- arXiv match often fails
- Collaborations change over time

Multi-messenger Observations of a Binary Neutron Star Merger

LIGO Scientific Collaboration, Virgo Collaboration, Fermi GBM, INTEGRAL, IceCube Collaboration, AstroSat Cadmium Zinc Telluride Imager Team, IPN Collaboration, The Insight-Hxmt Collaboration, ANTARES Collaboration, The Swift Collaboration, AGILE Team, The 1M2H Team, The Dark Energy Camera GW-EM Collaboration, the DES Collaboration, The DLT40 Collaboration, GRAWITA: GRAvitational Wave Inaf TeAm, The Fermi Large Area Telescope Collaboration, ATCA: Australia Telescope Compact Array, ASKAP: Australian SKA Pathfinder, Las Cumbres Observatory Group, OzGrav, DWF (Deeper, Wider, Faster Program), AST3, CAASTRO Collaborations, The VINROUGE Collaboration, MASTER Collaboration, J-GEM, GROWTH, JAGWAR, Caltech-NRAO, TTU-NRAO, NuSTAR Collaborations, Pan-STARRS, The MAXI Team, TZAC Consortium, KU Collaboration, et al. (26 additional authors not shown)

Collaborations

The usability problem:

- Authors want to be able to find papers in which they are part of the collaboration
- Some authors may not want to pollute their results lists with papers where a given name is 1/3000th of the author list
- Authors may want to be able to search by collaboration

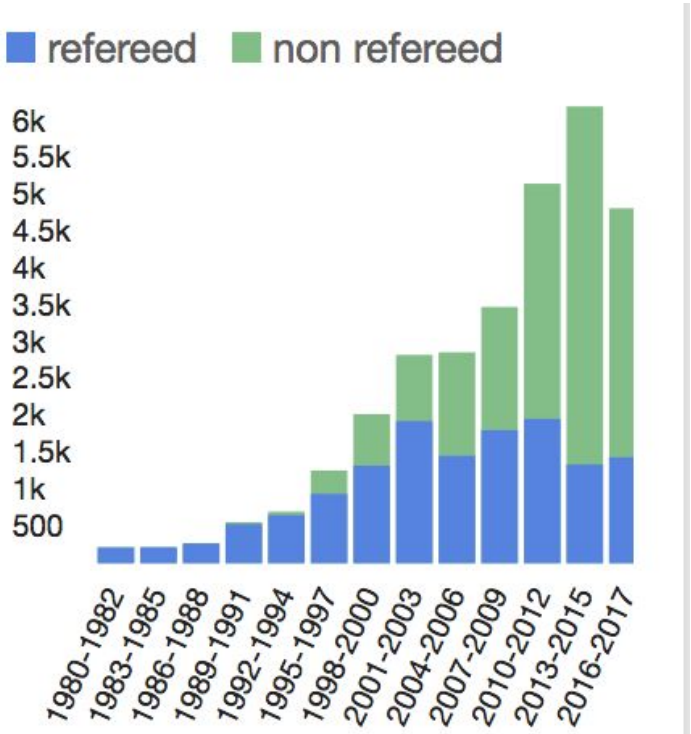
LIGO Scientific Collaboration
LIGO Collaboration
Ligo Scientific Collaboration
LIGO Science Collaboration
LIGO-Virgo Collaboration
Joint LIGO/Virgo working Group
LIGO-Virgo Scientific Collaboration
LIGO Team
LIGO/Virgo Collaboration
LIGO Virgo Scientific Collaboration

Fermi LAT Collaboration
Fermi-LAT Collaboration
Fermi Large Area Telescope Collaboration
Fermi Collaboration
Fermi GBM Collaboration
Fermi/LAT Collaboration
Fermi-Lat Collaboration
Fermi GBM Team
Fermi/Lat Collaboration
Fermi-LAT Team

Collaborations

Currently in ADS:

- Collaborations considered to be an author
- Minimal data cleanup to standardize
- Nomenclature stripped (different meanings? on behalf of, for the, and the)
- Names often added by hand (e.g. from Science appendices)
- 28K ast records, 38K phy records with collaboration in author list, comprising approximately 8K unique ast collaborations, 12K unique phy collaborations (18K total)



Collaborations

Scope of solutions (in increasing difficulty):

- Do nothing beyond some data cleanup
- Implement basic synonyms, akin to author synonyms
- Create “collaboration” concept
  - “Is in” collab, based on existing content (where we already have author lists)
  - Allow exclusion
  - Flat structure vs. multiple granularity? Is it enough to have “SDSS” or do we need a further breakdown (hopefully not)\
- Or ...

Collaborations

- Implement system like Inspire/Arxiv
  - Maintain lists of collaborations, members, dates
  - Collaborations provide author.xml (requires collaboration buy-in)
  - Are there multiple classes of things to track (consortium vs. team vs. collaboration)?
  - What about those that don't provide? Or legacy data?

Inspire has been doing it since the early 1970's and still needs 1 FTE for ongoing curation. ADS has more collaborations by an order of magnitude. Inspire attempts to disambiguate every author. They allow exclusion of collaborations only by limiting number of authors.