The NASA Astrophysics Data System:
Data Ingest and Curation

Carolyn S. Grant & Donna M. Thompson

ADS Users Group Meeting
January 17, 2017

Overview (1/2)

- Input data sources and formats
    - Formats and sources
    - Metadata, fulltext, references, links
    - Delivery methods and schedules
- Parsing, ingest and normalization
    - Content is more complex these days (e.g. ORCIDs, collaborations)
    - Author and affiliation normalization
- Merging content, managing interoperability
    - arXiv, publisher, SIMBAD, NED
    - Keeping up with bibcodes

Overview (2/2)

- Curation policies
    - Content selection and criteria
    - Refereed / Non-refereed
    - Gray literature and predatory OA
- Getting the job done
    - Fixing references / citations
    - Back-filling content (important historical publications)
    - Help pages
- User Feedback

Scope of Data

Abstracts (12.5 million):

- Astronomy: 2.2 million, 100K annually, weekly updates
- Physics: 8.3 million, 500K annually, monthly+ updates
- ArXiv: 1.2 million, 75K annually, daily update
- General: 1.2 million, 50K annually, infrequent

Scope of Data, cont.

References (6.3 million articles):

- 100 million recognized references, 10 million annually

Full Text (4.6 million articles):

- ~650 journals

Outside Links (11.9 million):

- DOIs (8 million)
- pdf, html, associated*, comments*, data, library, preprint, multimedia*, SIMBAD, NED, SPIRES*, TOC*

Data Properties (refereed, open access, article types)

Sources of Data

- ~25 major publishers, ~650 journals
- ~75 minor publishers, mostly singleton journals
- SIMBAD, VizieR catalogs, proposals, author-submitted
- CrossRef feed, ~650 journals
- Book Series, books, e-books

Data Formats

- Mostly XML (several types)
- ADS tagged (%T, %A, etc.)
- tex/latex stragglers
- Free-form

- References a mix of XML, plaintext, pairs, etc.

- Full text a mix of XML, SGML, plaintext, tex, pdf, OCRd

Data Delivery

- ftp push
- ftp pull
- automated emails
- independent emails
- online submission
- automated web harvesting
- shared websites
- website scraping

Content from multiple sources

NASA/STI vs. Publisher vs. SIMBAD vs. NED

Publisher vs. SIMBAD

ArXiv vs. Published

Online Early vs. Published

- Hierarchy of quality/trust (field-dependent)
- Respect version of record
- Track changes
- Maximize content (e.g. arXiv abstracts/references)

Data Parsing

ADS has always been inclusive, not requiring a given format.

- Perl scripts, typically one per publisher
- Convert to common (tagged) format and create text files

Typical problems: character encodings, changing formats, inconsistent/incorrect data, “et al” and collaborations

Data Normalization

Currently:

- Mix titles
- Format authors
- Author synonyms
- Format keywords

Moving Forward:

- Institutions
- Non-western author names
- Collaborations

Data Ingestion

- Collect new data
- Parse files
- Load parsed files to database
- Add reference and links data

Moving forward

- Instant updates
- Port perl scripts to python
- Develop/reuse libraries for text extraction/mining

Bibcodes then and now

- Originally worked well for printed journals and for enabling interoperability with publishers and other data providers
- Bibcode model has been extended to cover content which is more complex
- Model no longer working for current content
- Mapping bibcodes to DOIs mostly works, but need to remove dependency on bibcodes moving forward and not all content has DOIs

Curation Policies

Selection and criteria:

> “If it isn’t in ADS, it doesn’t exist!”

Curation Policies: Content evaluation

- Be relevant to astronomy
- Be of a quality and scope of interest to an international audience
- Have an ISSN
- Be registered with CrossRef
- Regularly published
- Contain metadata (abstracts, titles, keywords, author names)
- Include articles written by professionals in their field

Curation Policies: Content Evaluation

- Relevance
  - Scientific study of astronomy
  - Content in any of the subfields of astronomy or physics
  - Non-astronomy content of interest to astronomers in different subfields
  - Content in related fields that has astronomical relevance

A full listing can be found at:

http://doc.adsabs.harvard.edu/abs_doc/faq.html#addjournal

Curation Policies: Recurring issues

Refereed vs non-refereed

Many look to us for guidance on this.

Curation Policies: Recurring Issues

How do you determine if a publication is refereed or not?

ADS considers articles refereed when they appear in journals that participate in peer review. Peer review is the process of having articles reviewed by experts in the field before publication. ADS staff check journal websites and author instructions to verify the peer review status. Sometimes periodical directories are consulted, such as Ulrich's International Periodicals Directory.

Non-refereed materials, such as conference proceedings, circulars and bulletin entries, are screened only by an editor and not sent out for peer review. The ADS does not consider materials that are verified only by an editor as refereed.

Occasionally there may be refereed articles in a non-refereed journal or non-refereed articles in a refereed journal (e.g. announcements or conference abstracts). ADS staff will mark the individual articles as refereed or non-refereed once this difference is known.

Curation Policies: Gray literature

What is it? Changing with time

- PhD theses
- Online-only conferences
- Reports, bulletins, white papers, bibliographies, proposals, catalogs
- Software code and data sets

Curation Policies: Predatory Publishers

Predatory publishers and Open Access

Typical request

- Is it real or not?
- Do we want this in ADS?

Curation Policies: Predatory Publishers

- Articles published elsewhere
- Editors not known in field of astronomy
- Reprints

- How do we figure this out?
    - Look at the website
    - See how publishers are handling the publications
    - Refer to our list of guidelines for acceptance
    - Refer to Beall's list

Getting the Job Done

- Fixing references
  - Determining the problem
  - Taking steps to resolve the issues

- Updating help pages
  - Keeping policy pages updated
  - Developing new content as changes occur
  - Identifying help documentation that needs to be ported from Classic to ADS Beta

Getting the Job Done

Filling in the gaps:

Historical content

Missing conferences

Observatory publications

Getting the Job Done

User Feedback

How and what we do

- Monitor user feedback and respond rapidly
- Bring issues to the appropriate staff member
- Discuss issues at staff meeting
- Respond by action!

Backup Material

Bibcodes now

PhD theses, conferences, e-conferences, books, e-books, telegrams, circulars, proposals, catalogues, posters, electronic article ids, software, datasets, press releases

```
YYYYabcd.ttttnnnnnl
 year abbr  type  id# init
```

Working (mostly) **for most** (but not all) **cases**

Bibcodes moving forward

5+ digit volumes (GCN, SPIE)

6+ digit ids (APS, IOP, Hindawi)

27+ issues (IOP)

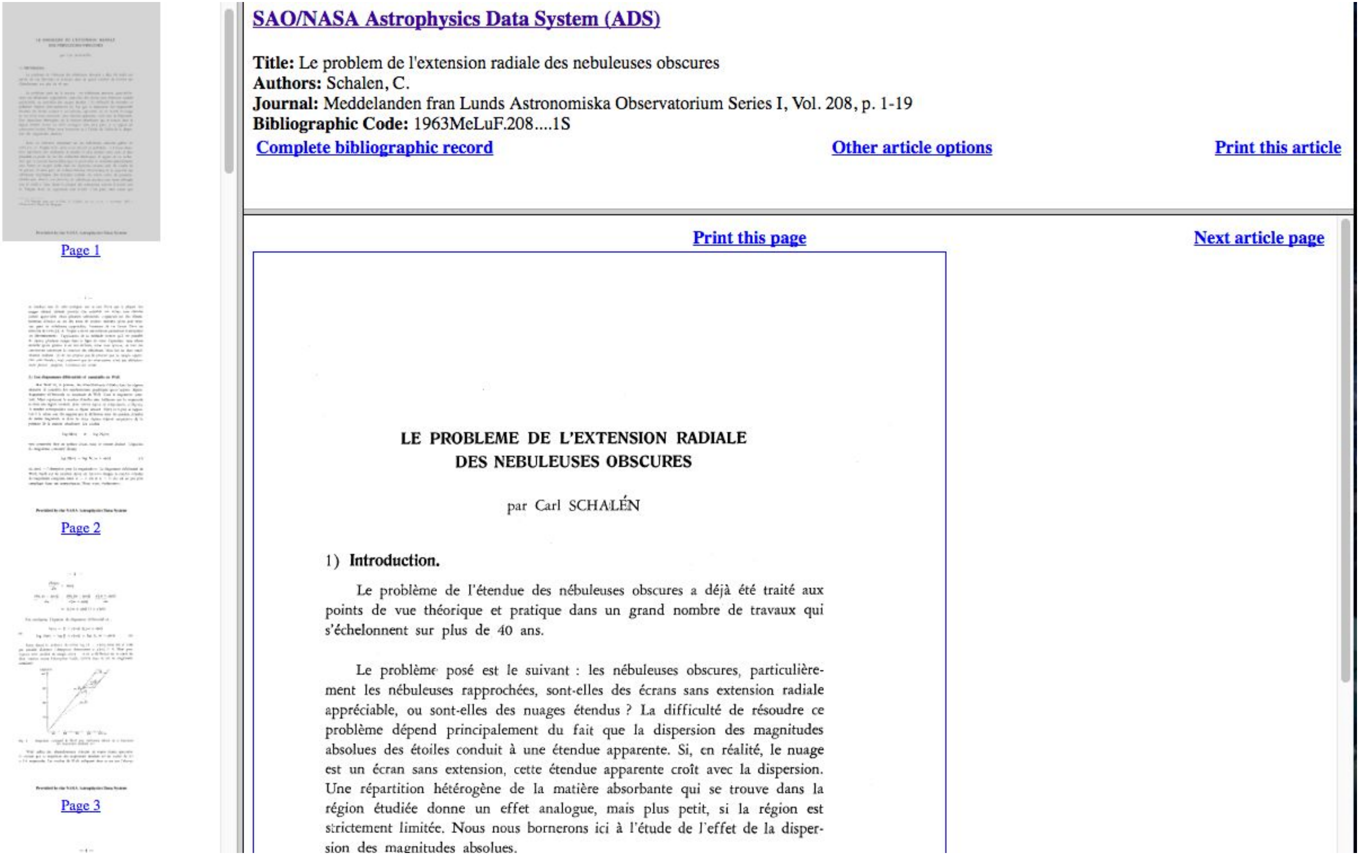
case sensitivity

pubyear discrepancies (especially Dec/Jan)

Extending bibcode breaks interoperability with others.

Need to make it optional.

Historical Scans: Observatory Publications



Title: Le problem de l'extension radiale des nebuleuses obscures
Authors: Schalen, C.
Journal: Meddelanden fran Lunds Astronomiska Observatorium Series I, Vol. 208, p. 1-19
Bibliographic Code: 1963MeLuF.208....1S

LE PROBLEME DE L'EXTENSION RADIALE
DES NEBULEUSES OBSCURES

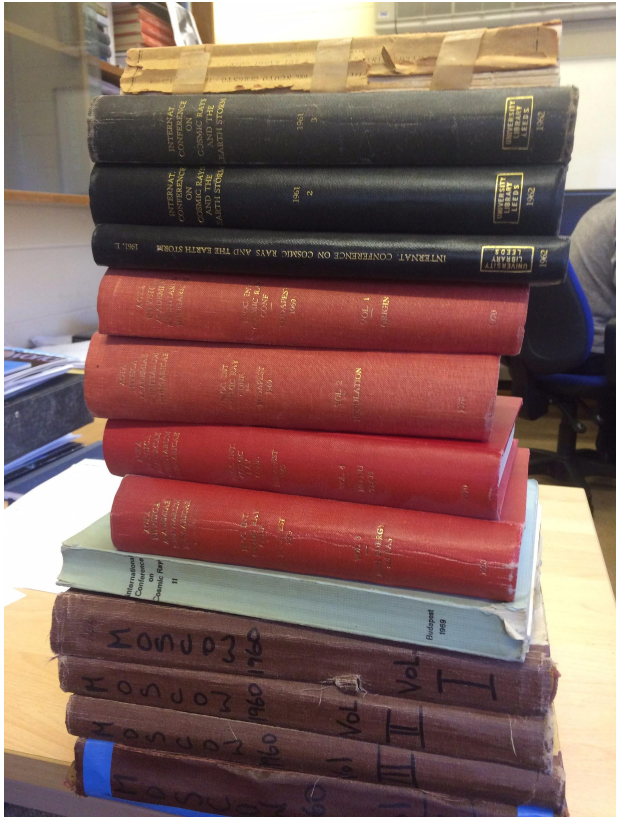
par Carl SCHALÉN

1) **Introduction.**

Le problème de l'étendue des nébuleuses obscures a déjà été traité aux points de vue théorique et pratique dans un grand nombre de travaux qui s'échelonnent sur plus de 40 ans.

Le problème posé est le suivant : les nébuleuses obscures, particulièrement les nébuleuses rapprochées, sont-elles des écrans sans extension radiale appréciable, ou sont-elles des nuages étendus ? La difficulté de résoudre ce problème dépend principalement du fait que la dispersion des magnitudes absolues des étoiles conduit à une étendue apparente. Si, en réalité, le nuage est un écran sans extension, cette étendue apparente croît avec la dispersion. Une répartition hétérogène de la matière absorbante qui se trouve dans la région étudiée donne un effet analogue, mais plus petit, si la région est strictement limitée. Nous nous bornerons ici à l'étude de l'effet de la dispersion des magnitudes absolues.

Conference Series


INTERNAT. CONFERENCE ON COSMIC RAYS AND THE EARTH STORM 1961 3
UNIVERSITY LIBRARY LEEDS 1962
INTERNAT. CONFERENCE ON COSMIC RAYS AND THE EARTH STORM 1961 2
UNIVERSITY LIBRARY LEEDS 1962
INTERNAT. CONFERENCE ON COSMIC RAYS AND THE EARTH STORM 1961. I.
UNIVERSITY LIBRARY LEEDS 1962
BUDAPEST 1969
VOL. 1 ORIGIN
VOL. 2
VOL. 3
International Conference on Cosmic Rays 11
Budapest 1969
MOSCOW 1960 VOL I
MOSCOW 1960 VOL II
MOSCOW 1960 Vol III
MOSCOW 1960