Dataset Identifiers in Astronomy & Beyond

Alberto Accomazzi

IVOA Interop Nara

Dec. 8, 2010

Identify what?

- Articles: bibcodes, DOIs
  http://adsabs.harvard.edu/abs/2008ApJ...685..919T
  http://dx.doi.org/10.1086/591019

- Astronomical Objects: SIMBAD, NED
  http://simbad.harvard.edu/simbad/sim-id?Name=NAME%20LMC&Ident=%40133169&submit=submit

- Services: IVOA identifiers
  ivo://CDS.VizieR

- Data Products: IVOA IDs, ADEC IDs, URIs, DOIs?
  ivo://CDS.VizieR/J/other/APh/26.282
  ADS/Sa.CXO#obs/123
  http://www.sdss.org/
  10.1086/317056/tab1

Data Products in Articles

- Reference to big surveys, projects is done via referencing "data" paper or website
- Reference to data product curated in archive is done via URI or (ideally) IVOA/ADEC ID
- Supplementary material for high-level data products published with article available
- Tables, spectra, images eventually incorporated in curated archives (Vizier, NED, SIMBAD)

ADEC Dataset IDs

- In 2004, ApJ introduces the capability to reference datasets in manuscripts
- Tagging and verification of datasets during editorial process
- Goal: create links to data products in HTML version of manuscript
- Article-dataset correlation to be propagated back to data archives

Nomenclature

- In 2003, the IVOA adopts a draft for the syntax of IVOA Identifiers:
  ivo://AuthorityID/ResourceKey
- In 2003, ADEC approves the definition of dataset identifiers:
  ivo://ADS/FacilityId#PrivateId
- Properties: unique, permanent, resolvable
- Broad range of granularity
- Both Static and Dynamic Data product support

Nomenclature: Examples

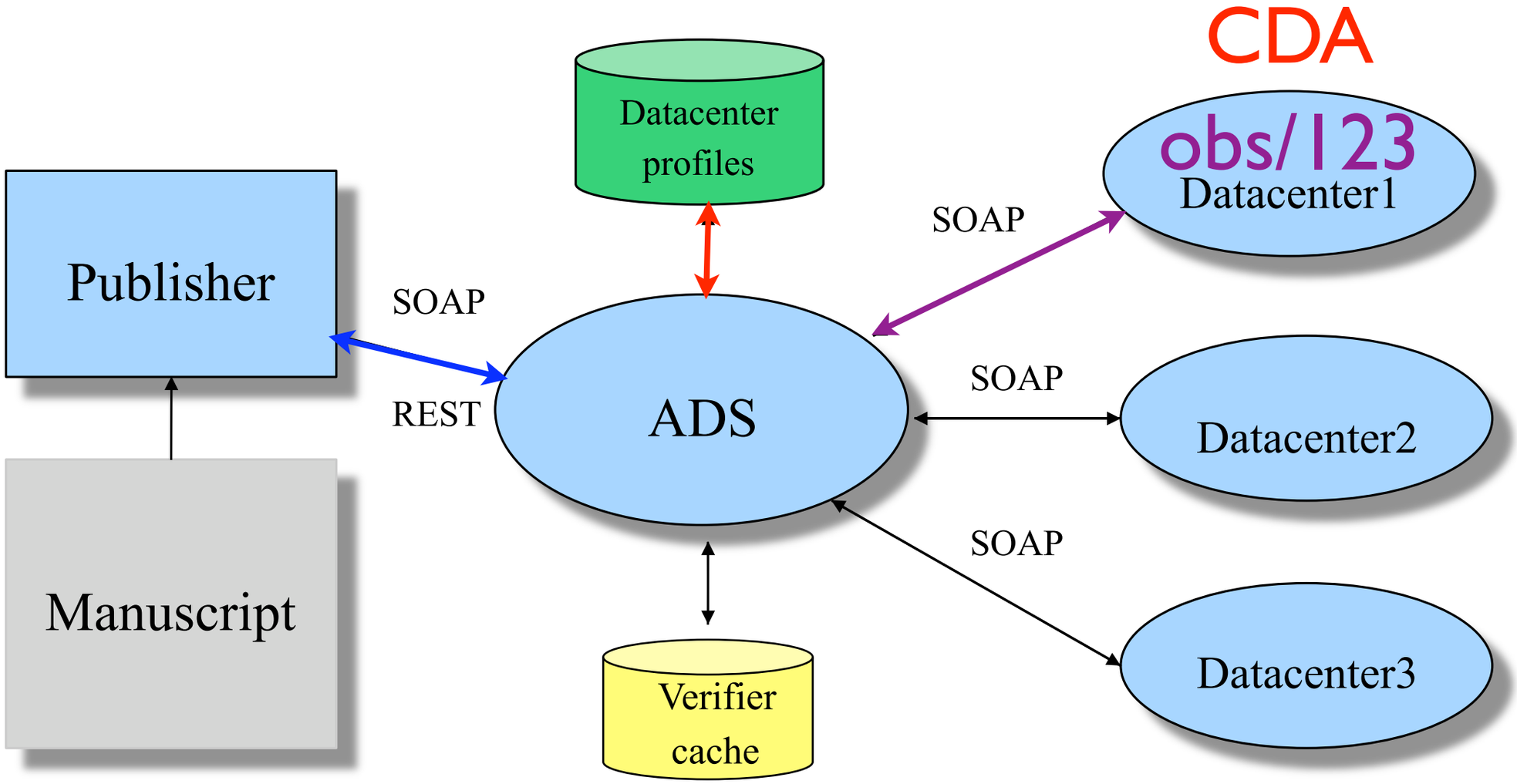
ivo://ADS/Sa.CXO#obs/123

ivo://ADS/Sa.CXO#DefSet/ChandraDeepFieldN1

ivo://ADS/Sa.CXO#Contrib/2007/MAUG1

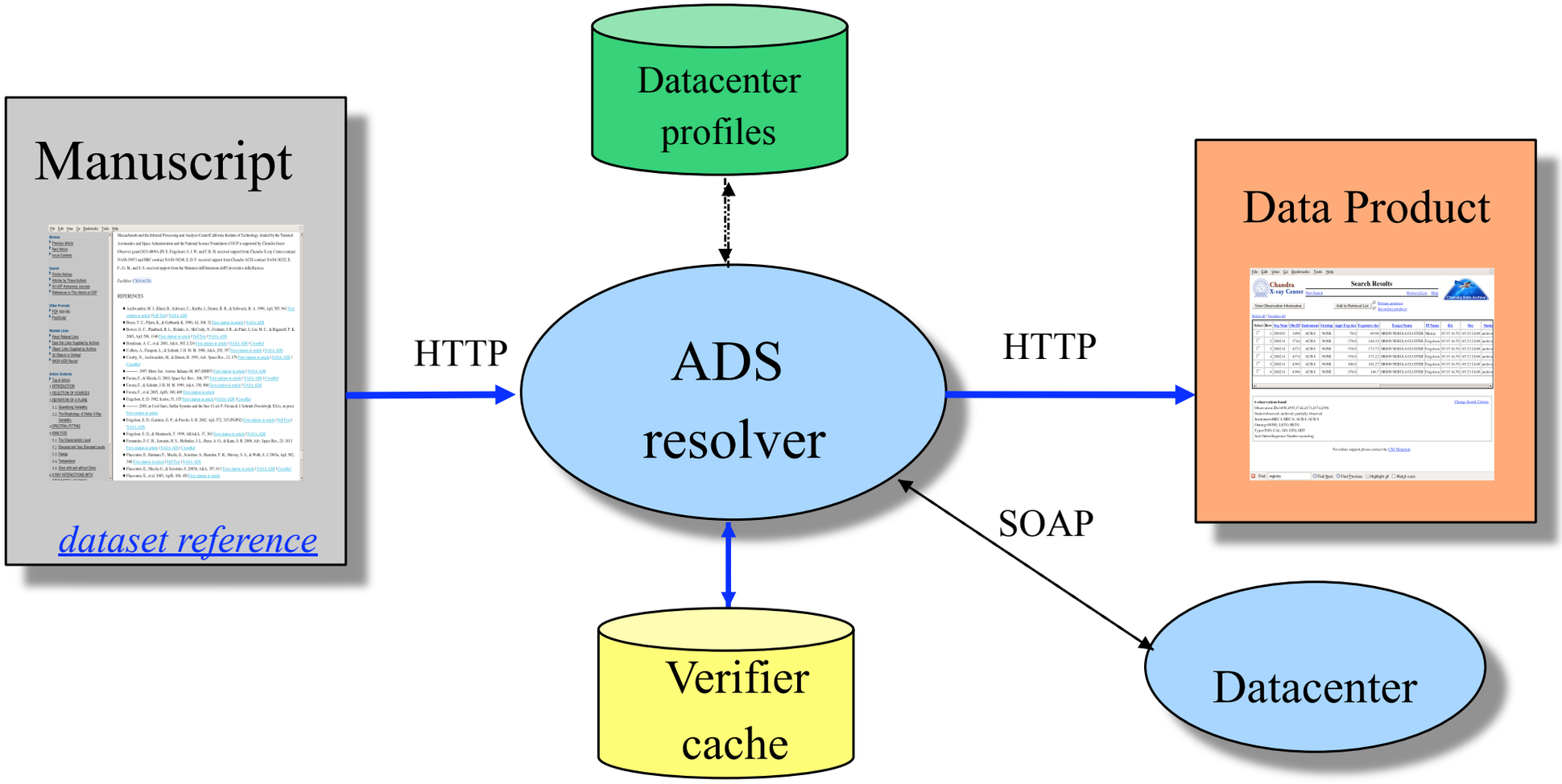
ivo://ADS/IRSA.Atlas#2006/0701/121559_24406

Implementation



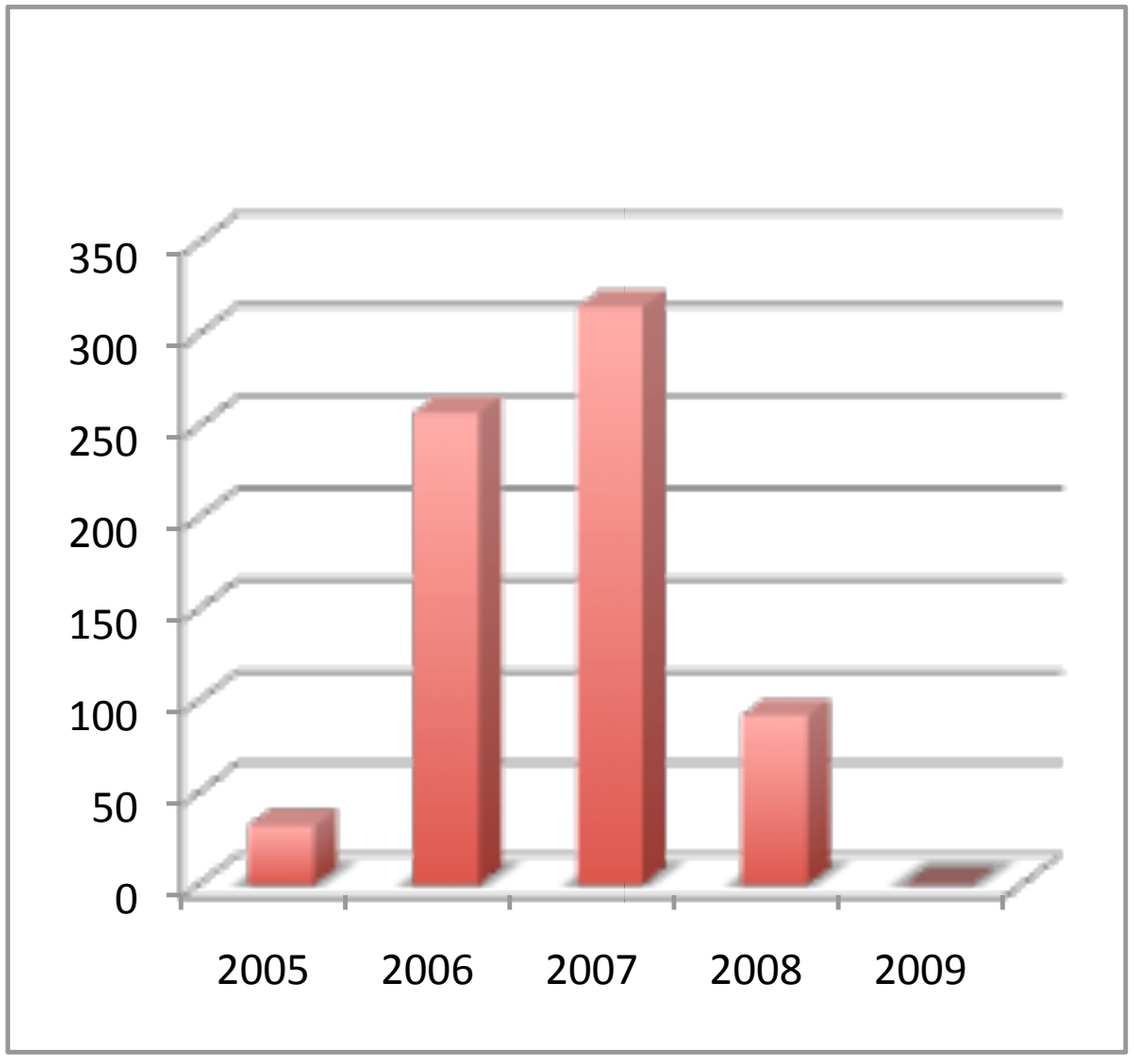
ivo://ADS/Sa.CXO#obs/123

Linking


Manuscript
dataset reference
Datacenter profiles
HTTP
ADS resolver
HTTP
Data Product
SOAP
Verifier cache
Datacenter

Problem: lack of adoption

- only NASA centers participating
- only 5 active archives, down from 6 in 2005
- only AJ and ApJ
- only "ADS" identifiers



Lessons Learned

- Requires commitment from archives
- Requires effort from authors, editors
- Not enough community buy-in
- No enforcement "stick" for data centers
- No reward "carrot" for authors, editors
- No silver bullet for curators
- Many parties involved, many points of failure

Issues

- Several standards gaining acceptance as identifiers for data products (both within articles and as stand-alone products). We will have to deal with this.
- Current IVOA Registry infrastructure provides one level of abstraction but does not deal with issues of persistence, multiplicity, may not scale to "data flood."
- Long-term persistence and naming major concern for Semantic Web applications, see Norman Gray's note: http://www.astro.gla.ac.uk/users/norman/ivoa/long-term-uris.html

Persistent IDs in DL world

- Permanent URLs (PURLs - OCLC)
- Handles (CNRI)
- Digital Object Identifiers (DOI)
- Archival Resource Keys (ARKs)
- EZIDs

A Way Forward

- Recast data linking in wider scope, the identification and linking of digital assets in astronomy: articles, objects, datasets
- Enable upload of high-level data products (plots, tables) to trusted, community-curated digital repository
- Provide infrastructure to enable minting of identifiers, tracking data products, exposing metadata, persistent link resolution
- Capture all identifiers and their inter-relationships during the peer review process
- Consider issues of branding, be realistic about goals

Actions

- Understand and capture data curation efforts before, during and after publishing
- Coordinated engagement with projects (ADS, SIMBAD, NED, Vizier), data archives and publishers critical
- VAO collaboration with NSF DataNet project will inform VAO DC&P efforts, with preservation as focus
- Splinter session at AAS meeting in Boston (May 2010) on data-literature links