

Persistent Dataset Identifiers

Arnold Rots
Alberto Accomazzi
SAO





Motivation

- This is about ***Astronomical Digital Objects***:
 - Publications – in a wide sense
 - Datasets – in a wide variety of places
 - Information on physical objects – as in NED and SIMBAD
- In order to make them useful they need to be:
 - ***Linked***
 - ***Preserved***
- Linking allows:
 - Searches – Discovery – Analysis
 - and requires:
 - Identifiers



Identifiers

- **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/simid?Name=NAME%20LMC&Ident=%403133169&submit=submit>
- **Services: IVOA identifiers**
 - <ivo://CDS.VizieR>
- **Data Products: IVOA IDs, ADEC IDs, URIs, DOIs?**
 - <ivo://CDS.VizieR/J/other/APh/26.282>
 - [Ivo://ADS/Sa.CXO#obs/123](ivo://ADS/Sa.CXO#obs/123)
 - <http://www.sdss.org/10.1086/317056/tab1>



Dataset Identifier Specification

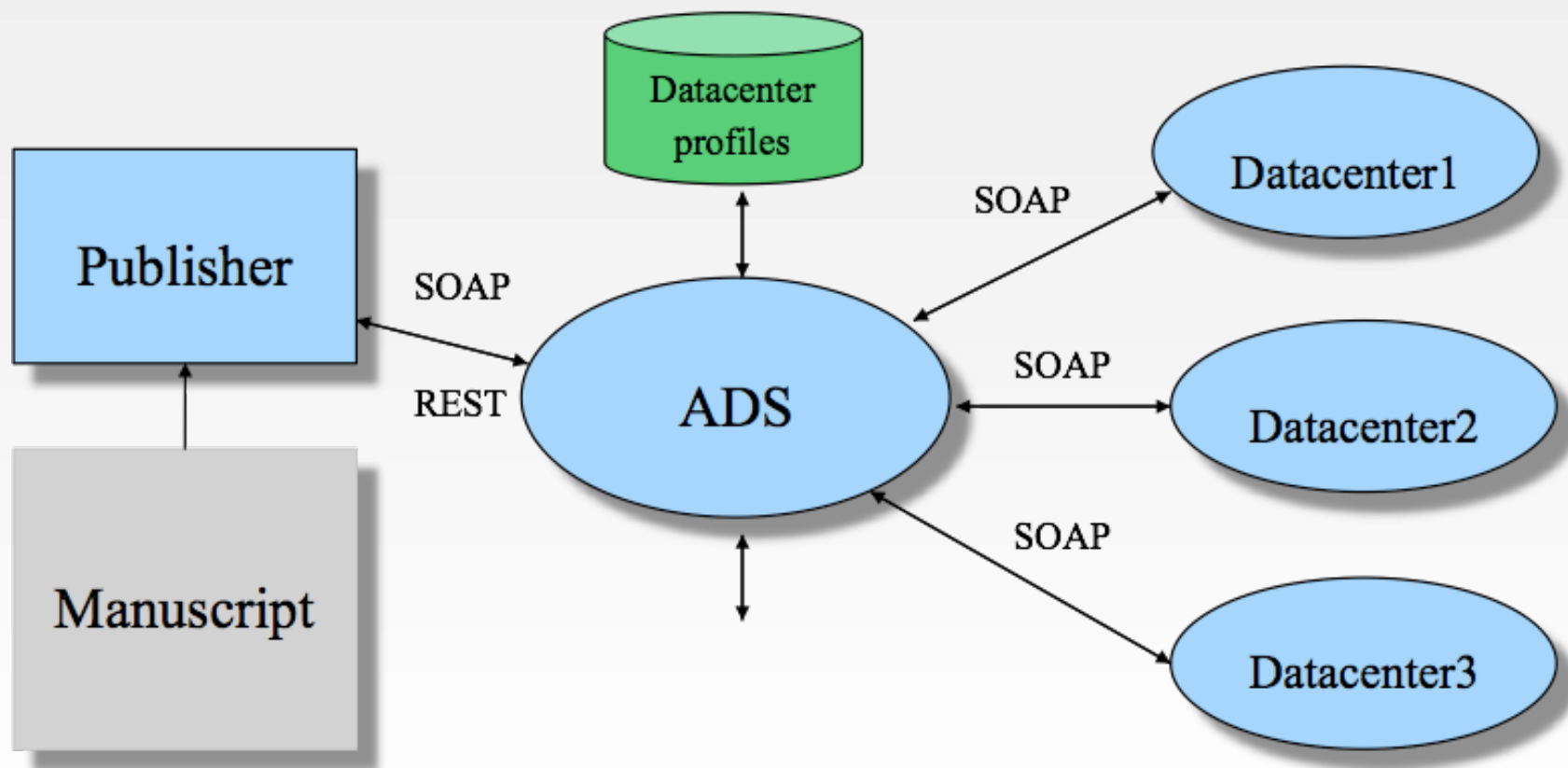
- In 2003, the IVOA adopted a draft for the syntax of IVOA Identifiers:
 - ivo://AuthorityID/ResourceKey**
 - Both Static and Dynamic Data product support
- Also in 2003, ADEC approved the definition of dataset identifiers, with ADS as naming authority:
 - ivo://ADS/FacilityId#PrivateId**
 - Properties: unique, permanent, resolvable, verifiable
 - Broad range of granularity (at facility's discretion)

Dataset Identifiers: Examples and Use

- Observation:
 - `ivo://ADS/Sa.CXO#obs/123`
- Predefined collection of observations:
 - `ivo://ADS/Sa.CXO#DefSet/ChandraDeepFieldN1`
- Contributed dataset:
 - `ivo://ADS/Sa.CXO#Contrib/2007/MAUG1`
- Atlas:
 - `ivo://ADS/IRSA.Atlas#2006/0701/121559_24406`
- Usage:
 - In 2004, ApJ introduces the capability to reference datasets in manuscripts
 - Tagging and verification of datasets during editorial process

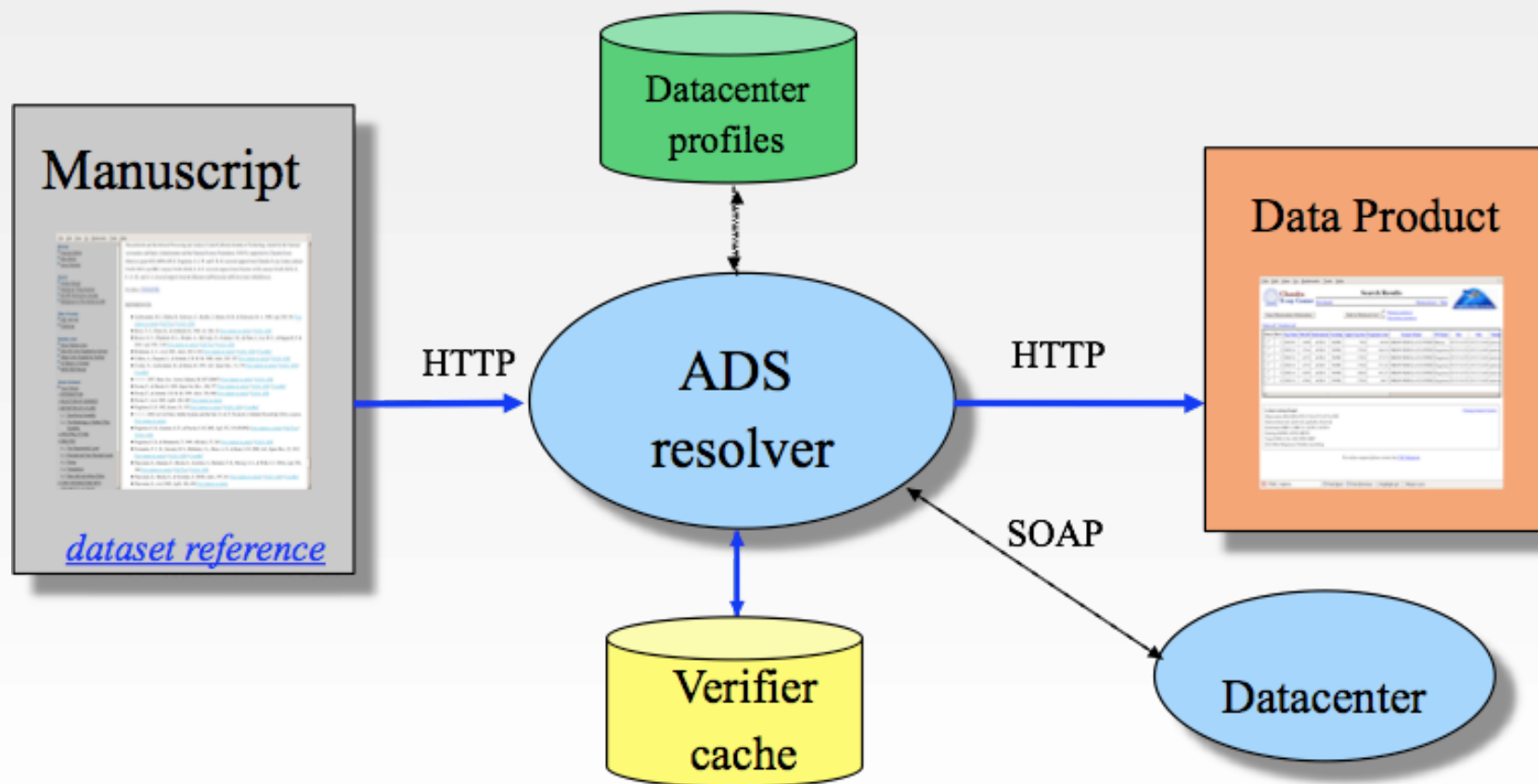


Current Implementation





Linking





Lessons Learned

- It works and works well when used
- It has been crucial in a prototype semantic browser
- But only NASA data centers participate
- Issues:
 - Requires commitment from archives
 - Requires effort from authors, editors, and archives
 - No stick for data centers
 - No carrot for authors and editors
 - Not enough community buy-in
 - No silver bullet for curators – still a lot of manual labor

Persistent IDs in Digital Library World

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

Why “ADS” Dataset Identifiers?

- We need something that works now
- And guarantees these properties:
 - Unique
 - Verifiable
 - Persistent – in perpetuity
 - Covers all types of research items:
data products, articles, objects, services
 - Leads unambiguously to dataset
 - Allows facilities flexibility in the definition of its private keys
 - Does not require version specification
- The *ADS* Naming Authority in
`ivo://ADS/<facilityID>/<privateID>`
per definition implies these requirements



Next Steps

- A repository for data products that need to be preserved
 - Data from projects and facilities that have a limited lifespan
 - Data behind plots, images, tables in articles that are not being preserved elsewhere
 - Anything else that is quoted in the literature or worth to be preserved
- A registry specifically for Dataset Identifiers
 - ADS serves *de facto* as such a registry, but it should be designed and implemented properly



Securing Long-term Stability

- This does not preclude a future change to, e.g., DOIs (bibcodes and DOIs coexist peacefully)
- A registry that can resolve Dataset Identifiers can also translate them
- The important issue is to design and implement the role of Dataset Identifiers such that all essential requirements are met; that will safeguard future development
- The ADEC Dataset Identifier specification (including its requirements on the facilities) satisfies this requirement