Introduction to DOIs

(and some thoughts on data citation)

Alberto Accomazzi

IVOA Interop

31 Oct 2015

DOIs are Persistent Identifiers

- Links to online resources change regularly, often on a timescale of a few years
- Persistent identifiers (PIDs) can be used to identify resources over long periods of time regardless of their location (URL)
- The price to pay for this is technical infrastructure providing registration and resolution services
- In the case of DOIs, registration is made via Registration Authorities (RAs) which are members of the DOI foundation (IDF). Most popular RAs:
  - CrossRef (used for scholarly publications)
  - DataCite (used by data providers, technical information centers)
  - mEDRA (used by some European Publishers)
- Other persistent identifiers exist, such as PURLs, ARKs

What DOIs look like

- DOIs (aka DOI names) follow the syntax:

  10.xxxx/<suffix>

  where xxxx represents a 4- or 5-digit number assigned to a resource provider (typically a publisher). The provider controls what goes in <suffix>
- Examples:
  - 10.1088/0004-637X/812/2/136 (registered by IOP, corresponds to 2015ApJ...812..136B)
  - 10.5479/ADS/bib/1901LicOB.1.58P (registered by ADS, corresponds to 1901LicOB...1...58P)
  - 10.5281/zenodo.10505 (registered by Zenodo, a poster presented at the LISA conference)
- DOIs also appear as:
  - doi:10.1088/0004-637X/812/2/136
  - http://dx.doi.org/10.1088/0004-637X/812/2/136
  - http://doi.org/10.1088/0004-637X/812/2/136

DOIs have registered metadata associated to them

- Resource provider deposits metadata when resource is registered and DOI minted
- Metadata for DOI follows a schema which is dependent on RA and resource:
  - Crossref deposit schema: http://help.crossref.org/deposit_schema
  - DataCite metadata schema: https://schema.datacite.org/meta/kernel-3/index.html
- Registration Authority provides services to harvest and search DOI metadata through one or more APIs
- But just as not all DOIs are “equal,” not all RAs are equal in terms of the services that they provide, and the checking they do

DOI have resolution services

- Metadata deposited when a DOI is registered allows the creation of discovery services and harvesting interfaces:
  - https://search.datacite.org/ui?&q=accomazzi
  - http://search.crossref.org/?q=accomazzi
  - http://oai.datacite.org/oai?verb=Identify
- CrossRef and DataCite support content negotiation:
  - Landing page (HTML): http://data.datacite.org/10.5284/1000418
  - XML: curl -LH "Accept: application/x-datacite+xml" http://data.datacite.org/10.5284/1000418
  - BibTeX: curl -LH "Accept: application/x-bibtex" http://data.datacite.org/10.5284/1000418
  - (this also works directly when querying the registry via http://doi.org/<doi>)
- Support for doc fragments: http://dx.doi.org/10.1371/journal.pone.0103437#s2

Some DataCite DOIs in action

- Software: https://zenodo.org/record/31760
- Dataset: https://zenodo.org/record/13297
- Thesis: https://zenodo.org/record/32163
- Presentation: https://zenodo.org/record/18107
- Poster: https://zenodo.org/record/10505
- Restricted access: https://zenodo.org/record/31293
- Plot: http://dx.doi.org/10.6084/m9.figshare.1284334

Versioning

“If I have assigned a DOI and I make a change to my material, should I assign a new DOI?”

> The IDF does not have any rules on this. Individual RAs adopt appropriate rules for their community and application. As a general rule, if the change is substantial and/or it is necessary to identify both the original and the changed material, assign a new DOI name. (DOI handbook)

In practice: if versioning is crucial, assign a DOI to each resource. You can also assign a DOI to the un-versioned resource record (leading to latest version). But beware of what this may do to counting citations to data products.

What should be assigned a DOI?

Stop and answer the following questions first:

- Why do I need a DOI for this resource? (Usually: so it can be cited)
- Can I generate the necessary metadata to register the DOI?
- Will I be able to maintain the resource available long-term and update the metadata as appropriate when it moves (or the access URL changes)?
- Do I need the DOI to track different versions of the dataset?

Note that having a DOI (or some other kind of PID) helps with tracking citations but does not necessarily solve the problem, which requires changes in publishing policies and a system to track references to this content (what ADS does for papers)

Some use cases

- High level data products associated with a paper (e.g. Vizier catalogs):
  Straightforward (authorship, metadata inherited from paper)
- Data Catalogs (e.g. 2MASS, Wise, etc) as a “whole:”
  Pretty straightforward, need to figure out proper authorship rules
- Individual data catalog tables, specific releases:
  Possibly useful, but maybe not strictly necessary
- Data collections (e.g. all observations from an archive analyzed in a paper or series of papers, currently an ongoing MAST prototype project):
  Useful, although requires infrastructure to capture collection and metadata
- Individual ObsIds (pointed observations): Useful, also requires infrastructure

Some final thoughts

The main reason why we are talking about this is to enable “Data Citation” and “Repeatability,” which are good things but which do not come for free. In order to find out whether your data is cited we need the following things:

- A persistence layer over your data products, with the capability to update the corresponding metadata with a DOI registration authority when things change
- Buy-in from publishers allowing DOIs to be listed as references
- An indexing system which identifies these citations and publishes their information (ADS does this for papers and high-level data products, but won’t be indexing arbitrary data collections or individual ObsIds, so an alternative mechanism is needed)