EPJ Web of Conferences **186**, 12011 (2018) https://doi.org/10.1051/epjconf/201818612011
LISA VIII

A Model for Using DataCite DOIs in Observatory Bibliographies

Arnold Rots[1,⋆], *Raffaele* D'Abrusco[1], and *Sherry* Winkelman[1]

[1] *Harvard-Smithsonian Center for Astrophysics*

Abstract. The use of DOI persistent identifiers has become an attractive mechanism for citing datasets in articles. However, taking the users' interests into account requires careful consideration of the way in which we apply these identifiers.
Our objective is the application of DOIs in such a way that individual datasets are properly cited while presenting the citations to the reader in a user-friendly manner.
This is achieved by making judicious use of the metadata structure provided by DataCite.

1 Introduction

The Chandra Data Archive's bibliographic database has been using IVOA persistent identifiers (PID), issued under the authority of the ADS, to link its datasets to articles in the literature. This type of identifier was established in 2002 by the NASA astrophysics data centers and has served us well, as it had enabled us to provide complete literature-data linking at a fine-grained level, resulting in the most complete observatory bibliography currently in existence. However, at this point it is time to transition to international standard identifiers that have come into existence after we started developing our bibliography. The obvious choice is the DataCite DOI. However, it is not an entirely trivial conversion. Just like each publication has its own unique PID, it is essential that each dataset receive its unique PID. The problem that arises is the aggregation of PIDs for any particular article. Currently, the ADS performs the aggregation function for the IVOA-type identifiers, but that will not be feasible for the DataCite DOIs. Aggregation is necessary, since dataset PIDs need to refer to a landing page and separate landing pages for each dataset would be extremely user-unfriendly. The solution, adopted by some, to mint a single PID for each article that holds all the datasets is not acceptable, since it violates the requirement that each dataset have a unique PID. The solution is to mint two types of PIDs, one for articles, the other for datasets, that refer to each other using standard DataCite DOI metadata elements. A related initiative is the Scholix project[1], an initiative under the auspices of the Research Data Alliance (RDA) that has the potential to provide comprehensive data-literature linking services and should be closely monitored.

2 Requirements

In order to satisfy the interests of the dataset providers and the readers, we identify three simple requirements.

⋆e-mail: arnold@rots.net ORCID: 0000-0003-2377-2356
[1]http://www.scholix.org/home

2.1 Each Dataset Needs its own DOI

This is a logical requirement, akin to requiring that each article in a given journal volume have its own DOI, so that its citations and other statistics can be tracked. And it is (obviously) simple to implement.

2.2 A Single Landing Page per Datatset

It shall be possible to reach a single page that lists (and provides links to) all articles that cite a given dataset. This is simple to implement, too, by including the references to the articles' DOIs in the metadata of the dataset's DOI (see Requirement 1).

2.3 A Single Landing Page per Repository per Article

It shall be possible to reach a single page that lists (and provides links to) all datasets that are cited in a given article. This is where it gets tricky. If we only have the DOIs specified by Requirement 1, we need an aggregation service that has the capability to aggregate all datasets cited in an article, collect their links, and pass that list on to a landing page. Since this service cannot be provided by the DOI repository, we need an independent mechanism. The solution is found in minting a separate series of DOIs, one per article, that contains in its metadata all the links to a given repository's cited DOIs. Note that sole reliance on this type of DOI does not allow for quick retrieval of dataset-based statistics, as datasets would then lack a DOI-based identity.

3 Implementation

We propose an implementation model satisfying the requirements above through the combination of three types of DOIs, the first of which is the DOI assigned to the article by its publisher.

3.1 Article

We assume each article has been issued a CrossRef DOI (or a BibCode; they should be interchangeable) by its publisher.

3.2 Article-based List of Cited Datasets

This is a DOI that contains references to the article and to each of the datasets cited in the article, allowing a single landing page to be provided for all cited datasets from a single repository. It provides necessary information for constructing an article-based landing page of cited datasets through:

- A single metadata element `IsCitedBy` points to the publisher's DOI
- One or more metadata elements `HasPart` point to individual datasets DOIs.

3.3 Dataset

This is a DOI that contains references to the article and to the list of datasets cited in the article, allowing a direct way of collecting linkage and statistics by dataset. It provides necessary information for constructing a dataset-based landing page of articles citing the dataset through:

- One or more metadata elements `IsPartOf` that point to the article based DOIs defined above that reference this dataset; these metadata elements are the reciprocal elements of the `HasPart` metadata elements in those DOIs.
- One or more metadata elements `IsCitedBy` that point to the publisher's DOIs corresponding to the ones cited in the article-based DOIs it is part of.

4 User Interface

The user/reader can view the citation information in two ways:

4.1 From the Article

The reader is directed from the article to a landing page providing access to all datasets from a specific repository, based on the `HasPart` metadata associated with the DOI specified in Section 3.2.

4.2 From the Dataset

From the dataset in the repository the user is directed to a landing page providing access to all articles that cite that particular dataset, based on the `IsPartOf` metadata associated with the dataset's DOI as specified in Section 3.3 and the `IsCitedBy` metadata in the article-based DOI, or just by the `IsCitedBy` metadata element in the dataset's DOI.

4.3 Observatory Bibliometrics

The observatory can gather its publication statistics from the dataset-based DOIs' metadata, although one would expect that information also to be available in the observatory's databases.

This work has been supported by NASA under contract NAS 8-03060 to the Smithsonian Astrophysical Observatory for operation of the Chandra X-ray Center.