What is a DOI?

Martin Fenner

August 6, 2014

This Sunday Ian Mulvany and I will do a presentation on Open Scholarship Tools at *Wikimania 2014* in London. From the abstract:

> This presentation will give a broad overview of tools and standards that are helping with Open Scholarship today.

One of the four broad topics we have picked are *digital object identifiers (DOI)s*. We want to introduce them to people new to them, and we want to show some tricks and cool things to people who already now them. Along the way we will also try to debunk some myths about DOIs.

What a DOI looks like

DOIs - or better DOI names - start with a prefix in the format `10.x` where x is 4-5 digits. The suffix is determined by the organization registering the DOI, and there is no consistent pattern across organizations. The DOI name is typically expressed as a URL (see below). An example DOI would look like: http://dx.doi.org/10.5555/12345678. Something in the format **10/hvx** or http://doi.org/hvx is a shortDOI, and **1721.1/26698** or http://hdl.handle.net/1721.1/26698 is a handle. BTW, all DOIs names are also handles, so http://hdl.handle.net/10/hvx for the shortDOI example above will resolve correctly.

DOIs are persistent identifiers

Links to resources can change, particularly over long periods of time. Persistent identifiers are needed so that readers can still find the content we reference in a scholarly work (or anything else where persistent linking is important) 10 or 50 years later. There are many kinds of persistent identifiers, one of the key concepts - and a major difference to URLs - is to separate the identifier for the resource from its location. Persistent identifiers require technical infrastructure to resolve identifiers (DOIs use the Handle System) and to allow long-term archiving of resources. DOI registration agencies such as DataCite or CrossRef are required to provide that persistence. Other persistent identifier schemes besides DOIs include persistent uniform resource locators (PURLs) and Archival Resource Keys (ARKs).

DOIs have attached metadata

All DOIs have metadata attached to them. The metadata are supplied by the resource provider, e.g. publisher, and exposed in services run by registration agencies, for example metadata search and content negotiation (see below). There is a minimal set of required metadata for every DOI, but beyond that, different registration agencies will use different metadata schemata, and most metadata are optional. Metadata are important to build centralized discovery services, making it easier to describe a resource, e.g. journal article citing another article. Some of the more recent additions to metadata schemata include persistent identifiers for people (ORCID) and funding agencies (FundRef), and license information. The following API call will retrieve all publications registered with CrossRef that use a Creative Commons Attribution license (and where this information has been provided by the publisher):

```
http://api.crossref.org/funders/10.13039/100000001/works?filter=license.url:http://creativec
```

DOIs support link tracking

Links to other resources are an important part of the metadata, and describing all citations between a large number scholarly documents is a task that can only really be accomplished by a central resource. To solve this very problem DOIs were invented and the CrossRef organization started around 15 years ago.

Not every DOI is the same

The DOI system originated from an initiative by scholarly publishers (first announced at the Frankfurt Book Fair in 1997), with citation linking of journal articles its first application. This citation linking system is managed by CrossRef, a non-profit member organization of scholarly publishers, and more than half of the about 100 million DOIs that have been assigned to date are managed by them.

But many DOIs are assigned by one of the other 8 registration agencies. You probably know DataCite, but did you know that the Publications Office of the European Union (OP) and the Entertainment Identifier Registry (EIDR) also assign DOIs? The distinction is important, because some of the functionality is a service of the registration agency - metadata search for example is offered by CrossRef (http://search.crossref.org) and DataCite (http://search.datacite.org), but you can't search for a DataCite DOI in the CrossRef metadata search. There is an API to find out the registration agency behind a DOI so that you know what services to expect:

```
http://api.crossref.org/works/10.6084/m9.figshare.821213/agency
```

```
{
  "status": "ok",
  "message-type": "work-agency",
```

```
  "message-version": "1.0.0",
  "message": {
    "DOI": "10.6084/m9.figshare.821213",
    "agency": {
      "id": "datacite",
      "label": "DataCite"
    }
  }
}
```

DOIs are URLs

DOI names may be expressed as URLs (URIs) through a HTTP proxy server - e.g. http://dx.doi.org/10.5555/12345679, and this is how DOIs are typically resolved. For this reason the CrossRef DOI Display Guidelines recommend that *CrossRef DOIs should always be displayed as permanent URLs in the online environment*. Because DOIs can be expressed as URLs, they also have their features:

Special characters Because DOIs can be expressed as URLs, DOIs should only include characters allowed in URLs, something that wasn't always true in the past and can cause problems, e.g. when using SICIs (Serial Item and Contribution Identifier), an extension of the ISSN for journals:

`10.4567/0361-9230(1997)42:<0aEoSR>2.0.TX;2-B`

Content negotiation The DOI resolver at *doi.org* (or *dx.doi.org*) normally resolves to the resource location, e.g. a landing page at a publisher website. Requests that are not for content type `text/html` are redirected to the registration agency metadata service (currently for CrossRef, DataCite and mEDRA DOIs). Using content negotiation, we can ask the metadata service to send us the metadata in a format we specify (e.g. Citeproc JSON, bibtex or even a formatted citation in one of thousands of citation styles) instead of getting redirected to the resource. This is a great way to collect bibliographic information, e.g. to format citations for a manuscript. In theory we could also use content negotiation to get a particular representation of a resource, e.g. `application/pdf` for a PDF of a paper or `text/csv` for a dataset in CSV format. This is not widely support and I don't know the details of the implementation in the DOI resolver, but you can try this (content negotation is easier with the command line than with a browser):

```
curl -LH "Accept: application/pdf" http://dx.doi.org/10.7717/peerj.500 >peerj.500.pdf
```

This will save the PDF of the 500th PeerJ paper published last week.

Fragment identifiers As discussed in my last blog post, we can use frament identifiers to subsections of a document with DOIs, e.g. http://dx.doi.org/10.

1371/journal.pone.0103437#s2 or http://doi.org/10.5446/12780#t=00:20,00:27, just as we can with every other URL. This is a nice way to directly link to a specific document section, e.g. when discussing a paper on Twitter. Fragment identifiers are implemented by the client (typically web browser) and depend on the document type, but for DOIs that resolve to fulltext HTML documents they can add granularity to the DOI without much effort.

Queries URLs obviously support queries, but that is a feature I haven't yet seen with DOIs. Queries would allow interesting features, partly overlapping with what is possible with fragment identifiers and content negotiation, e.g. `http://dx.doi.org/10.7717/peerj.500?format=pdf`. II hope to find out more until Sunday.

Outlook

My biggest wish? Make DOIs more machine-readable. They are primarily intended for human users, enabling them to find the content associated with a DOI. But they sometimes don't work as well as they could with automated tools, one example are the challenges automatically resolving a DOI that I described in a blog post last year. Thinking about DOIs as URLs - and using them this way - is the right direction.