The trouble with DOIs

Martin Fenner, Gobbledygook

October 9, 2011

ScienceCard is a new service that I started last month with the simple idea to automatically track all journal articles of a given author, and to collect the article-level metrics (citations, bookmarks, etc.) for these papers. ScienceCard requires unique identifiers for articles and authors to work. Unique identifiers for authors is a difficult topic regularly discussed in this blog. But I thought that using digital object identifiers (DOI) for journal articles would be easy. The system managed by CrossRef was started 10 years ago, and almost all journal publishers now use DOIs – there were 49,350,542 registered CrossRef DOI links as of today.

Figure 1: Number buttons by fragmented on Flickr

The first problem I encountered is that many bibliographic databases don't fully support DOIs. Most of them store DOIs, but not all of them allow queries using DOIs, and very few services allow linking to them using DOIs. In the end I had to store various other article identifiers in ScienceCard (currently PubMed ID, PubMed Central ID, Microsoft Academic Search ID, Mendeley UUID, Scopus ID). One side effect of this proliferation of identifiers is that (in very rare cases) DOIs are not unique in these bibliographic services. And it makes it more complicated than necessary to build tools based on DOIs. The members of CrossRef are publishers, the other service providers (whether public or private) seem to be reluctant to fully support a service where they have no direct influence.

The second problem with DOIs is that they are often not web-friendly. DOIs are really permanent URLs, and CrossRef has recently changed the display guidelines for DOIs to reflect this. Instead of **doi: 10.1371/journal.pcbi.0010057** we are supposed to show DOIs as **http://doi.org/10.1371/journal.pcbi.0010057**. The problem is that DOIs can contain characters such as "+", "(", "." or "/" that need to be escaped when used as URLs. Some ScienceCard examples include the following:

1. http://doi.org/10.1016/S0959-8049(05)80357-0
2. http://doi.org/10.1093/bioinformatics/12.4.357
3. http://doi.org/10.1021/bi980175+
4. http://doi.org/10.1642/0004-8038(2002)119[0088:SSCPEO]2.0.CO;2

These special characters can create problems when DOIs are used in software programs. ScienceCard for example wants to create links to articles in the format **http://sciencecard.org/10.1642/0004-8038(2002)119[0088:SSCPEO]2.0.CO;2.xml**, but this function is currently broken.

One possible solution are shortDOIs. Article (3) would for example become http://doi.org/dcp, whereas article (4) is rejected as invalid DOI. I would love to use shortDOIs in ScienceCard and other places (e.g. Twitter), but haven't found an API yet that automatically returns shortDOIs for DOIs.

Component DOIs directly link to a figure or table of a paper. This is an underused, but very useful feature, and is for example provided by the PLoS journals. Unfortunately component DOIs can confuse bibliographic databases and make it more difficult to track all the links to a given article. I had to write a little routine to detect component DOIs imported into ScienceCard.

Articles are sometimes updated or corrected, and many publishers will use a different DOI for the updated article. This is a problem when you want to track all references to this particular article. http://doi.org/10.1371/journal.pcbi.0020121 and http://doi.org/10.1371/journal.pcbi.0020181 are for example DOIs for the same PLoS Computational Biology article (the latter is the corrected version). Nature Precedings uses a format that is easier to understand for computers - http://doi.org/10.1038/npre.2011.4479.3 is for example a link to the third

version of this particular manuscript. CrossMark is a new CrossRef service that will make it easier to track the different versions of a manuscript, including retractions.

ScienceCard should of course not be limited to journal articles. I'm also interested in other scholarly content, e.g. preprints from **ArXiV** or research datasets from **DataCite**. But I want to first solve the problems with DOIs for journal articles, before I tackle the much bigger problems with uniquely identifying and tracking other scholarly contributions. Science blog posts are a good example. It would be wonderful to track them in ScienceCard, but I don't see how we can do that before we have a system in place that assigns unique and persistent identifiers to blog posts. For this and other reasons I really want unique identifiers for science blog posts, and we should also think about using DOIs for this purpose.

Update October 9: A ScienceCard example of multiple identifiers for the same paper:

- DOI: 10.1007/s10654-011-9572-7
- PubMed ID: 21461943
- PubMed Central ID: 3115050
- Microsoft Academic Search: 48849734
- Mendeley: 5b0023f0-609e-11e0-8f54-0024e8453de6
- Mendeley URL: http://www.mendeley.com/research/informativeness-indices-blood-pressure-obesity-serum-lipids-relation-ischaemic-heart-disease-mortality-huntii-study/
- Scopus: 79959714408