Fragment Identifiers and DOIs

Martin Fenner, Gobbledygook

August 2, 2014

Before all our content turned digital, we already used **page numbers** to describe a specific section of a book or longer document, with older manuscripts using the folio before that. Page numbers have transitioned to electronic books with readers such as the Kindle supporting them eventually.

Image by Al Silonov from Wikimedia Commons. This file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license.

For content on the web we can use the **#** fragment identifier, e.g. https://en.wikipedia.org/wiki/Fragment_identifier#Proposals to navigate to a specific section of a web page. How the linking to this fragment is handled, depends on the **MIME** type of the document, and will for example be done differently for a text page than a video - YouTube understands minutes and seconds into a video as fragment identifier, e.g. https://www.youtube.com/watch?v=0UNRZEsLxKc#t=54m52s. Fragment identifiers are not only helpful to link to a subsection of a document, but of course also for navigation within a document.

All this is of course very relevant to scholarly content, which is usually much more structured, with most journal articles following the IMRAD - introduction, methods, results, and discussion - format, usually with additional sections such as abstract, references, etc. One approach to link to figures and tables within a scholarly articles is using component DOIs, e.g. specific DOIs for parts of a larger document. The publisher **PLOS** has been using them for a long time, and the number of component DOIs is rising, but most scholarly journal articles don't use component DOIs. And whereas component DOIs are a great concept for content such as figures (allowing us to describe the MIME type and other relevant metadata), they are probably not the best tool to link to a section or paragraph of a scholarly document.

As it turns out, we already have a tool for that, as the DOI proxy server gracefully forwards fragment identifiers (how did I miss this?). We can therefore use a DOI with a fragment identifier to

- Results section: http://doi.org/10.1371/journal.pone.0103437#s2
- Specific reference: http://doi.org/10.12688/f1000research.4263.1#ref-7
- Decision letter: http://doi.org/10.7554/eLife.00471#decision-letter

Obviously this only works if the DOI is resolved to the full-text of a resource, and not a landing page. And how the fragment identifiers are named and implemented is up to the publisher, and the DOI resolver has no information about them. These specific links are particularly nice for discussions of a paper, whether it is on Twitter or in a discussion forum. It appears that at least the Twitter link shortener keeps the fragment identifier, the link to the eLife decision letter is shortened to http://t.co/URWaYmGHnY. This kind of linking works particularly well if the publisher is using a fine-grained system of fragment identifiers, the publisher PeerJ for example allows links to a specific paragraph - e.g. http://doi.org/10.7717/peerj.500#p-15 - and allows users to ask a question right next to that section.

The examples above all use MIME type `text/html`, as this is what the example DOIs resolve to by default. I don't if and how publishers have implemented fragment identifiers for other formats such as PDF or ePub, and what happens if you combine fragment identifiers with content negotiation. The shortDOI service works with fragment identifiers as well: http://doi.org/pxd#decision-letter. Another interesting question would be how fragment identifiers are handled for datasets. Typically separate DOIs are assigned for multiple related datasets, but there could also be a place for fragment identifiers as well, e.g. to specify a subset via a date range. The solution depends again on the content type, and the popular `text/csv` is unfortunately not well suited for this, whereas JSON – using JSON Pointer – would work well.

Update 8/2/14: Leigh Dodds points out that handling the fragment identifier is up to the client and the fragment identifier is not sent to the server. Acrobat reader for example supports the `#page=` fragment identifier. He also mentions that there is a RFC7111 for fragment identifiers for the text/csv media type - browsers in the future might support something like `http://example.com/data.csv#row=5-7`.