How to Cite Datasets and Link to Publications

A Report of the Digital Curation Centre

Monica Duke Alex Ball

30 October 2012

Hello, my name is Alex Ball and I work for the Digital Curation Centre in the UK with my colleague Monica Duke. The Digital Curation Centre, or DCC, is a centre of expertise in digital curation and research data management funded by JISC, which is an agency that helps to develop and maintain information systems in the higher and further education sectors. For about five years now JISC has been pushing to improve research data management in the UK, and as part of that, we at the DCC are publishing a series of guidance documents based on themes set by JISC.

One of the themes is data citation, and at about this time last year we published both a Briefing Paper and a How-to Guide on the subject (*slide*). We have some copies here to give away and you can also download a copy (Figure 1) from our website or read it online.

http://www.dcc.ac.uk/resources/how-guides/cite-datasets

Figure 1: How-to Guide on data citation, on the Web

As I only have twenty minutes, I'm not going to be able to go through the whole document. Instead, I'll pick out some of the more interesting issues we came across when putting the guide together.

1 Motivation

I guess I don't need to convince anyone here about the need for data publication and citation, but to understand it we have to think about scholarly communications more generally. Journals are the big success story in this area, but what made them so popular (Figure 2)?

- Awareness raising
- Protection from plagiarism
- ~~Verification of results~~
- ~~Basis for future research~~
- Reward models
- Permanent access

Figure 2: What's great about journal papers?

They provided a way of communicating research results such that others could verify the results and build on them, while also ensuring authors received due credit, and in time rewards, for their work. Formal publication also meant formal archiving could take place. But as the process of conducting research has become more specialist and complicated, your average scientific journal paper can no longer contain all the information it needs to make the research reproducible (*transition*); we also need the underlying data. But we won't get data routinely shared until all these things apply to data as well as to journal papers. I would argue (Figure 3) that, given time, data citations are what will make it happen, because the citation model is well understood and trusted.

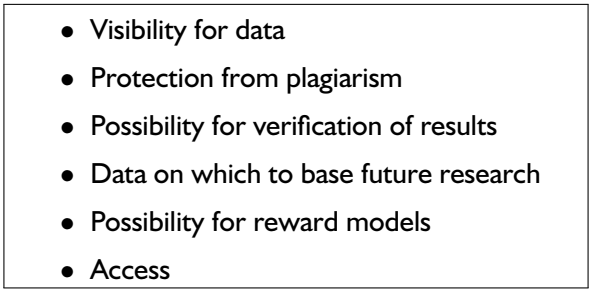

Figure 3: What data citations provide

What should data citations look like? Well, every journal has its own idea of what a citation should look like so the important point is what a citation should include (*slide*).

2 Elements of a data citation

Here are four standard citation styles I found in the literature: see the Guide for the full references. Which elements do they use?

Author, Publication date, Title, Version, Feature, Resource type, Publisher, Identifier, Location, Unique Numeric Fingerprint.

Altman and King (2007): Dataverse

- Sidney Verba. 1998. "U.S. and Russian Social and Political Participation Data," hdl:1902.4/00754 `UNF:3:ZNQRI14053UZq389x0Bffg?==` NORC [Producer]; data set [Type (DC)] ICPSR [Distributor].

Lawrence et al. (2008): BADC

- Iwi, A. and B. N. Lawrence (2004). A 500 year control run of HadCM3. [GridSeries, http://ndg.nerc.ac.uk/csml2/GridSeries] Version 1. BADC. urn:badc.nerc.ac.uk_coapec500yr [Available from http://badc.nerc.ac.uk/data/coapec500yr].

Green (2010): OECD

- OECD (2009), "Key short-term indicators", Main Economic Indicators (database). doi: 10.1787/data-00039-en http://dx.doi.org/10.1787/data-00039-en (Accessed on 14 September 2009)

Starr and Gastl (2011): DataCite

- Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. V.2. Geological Institute, University of Tokyo. Dataset. doi:10.1594/PANGAEA.726855. http://dx.doi.org/10.1594/PANGAEA.726855

There are five elements that occur in all four styles, four of which have a long pedigree in scholarly citation:

- Author
- Publication date
- Title
- Location (= identifier)

Despite the fact that we've had ISBNs since 1970 and online catalogues in widespread use since the mid-1980s, identifiers didn't really start to catch on in citations until the introduction of DOIs in the last five to ten years. I'd guess this is because with things like ISBNs there is no central register that allows you to look up the item; booksellers and libraries have had to build them up for themselves. So identifiers have tended to be used, if at all, more like checksums for making sure you had the right item, rather than as a way of accessing resources.

But the Web is changing all that. We now have ways of making locations persistent enough to be used as an identifier *(transition)*. While it's possible to do this by carefully managing URLs, it's more usual to achieve it by using a fake location, made up of a resolver service and the identifier, that redirects to the real location. DOIs are getting the most traction for datasets that are considered 'published', with Handles and ARKs being used more for ephemeral datasets.

- Publisher

Another change wrought by the Web is that we are now used to getting scholarly content direct from there rather than from library shelves *(transition)*. This makes the publisher more important than ever as both the host of information and the guarantor of its quality.

That might seem straightforward enough, but of course it's never as simple as that.

3 Issues and challenges

Take the author, for example. Authorship is a strange concept in the concept of a dataset. More natural roles might be a compiler, or a principal investigator, or a corporate owner. Furthermore, it is far easier to rack up a silly number of contributors with datasets than with textual publications. In such cases, a simple citation like this isn't going to cut the mustard. Most likely you'll need some sort of microattribution approach (*slide*).

This spreadsheet was submitted as part of the supplementary data for an article published in Nature Genetics last year. You'll see it attributes each genetic variation in the dataset to its contributor, as identified by a Thompson Reuter ResearcherID (other contributor ID schemes are available). This was very much a proof of concept. In future we might hope for this sort of information to be made available as linked data, preferably somewhere more accessible than supplementary data, like DataCite's metadata store.

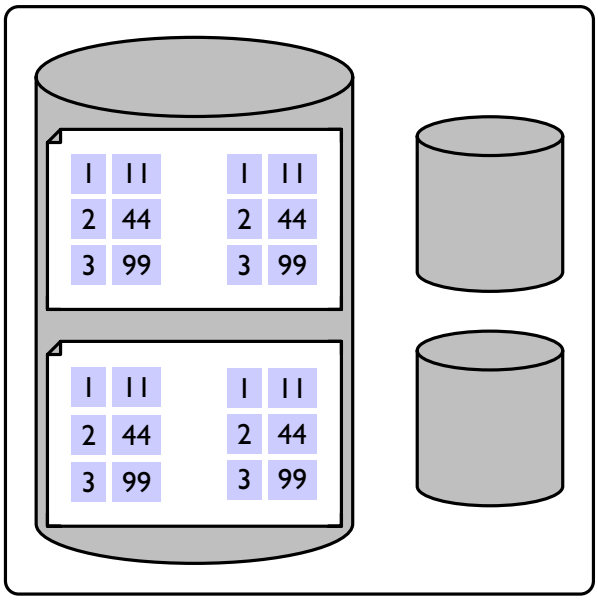

Figure 4: Granularity

Granularity can also be an issue. Just as you might cite only a sentence or a page of an article, with data you might find yourself citing only a single data point, or a table, or a file containing several tables, or dataset made up of many files. You might want to cite a more abstract subset of data such as one of the Features I mentioned earlier, or you might want to cite a whole collection of datasets.

The practical answer is:

- Cite datasets at the finest level that is appropriate and for which an identifier is provided.
- If that is not fine enough, provide details of the subset of data you are using at the point in the text where you make the citation.

So, now you have an in-text citation and a bibliographic reference. Where should that reference go?

- Special data resources section?
- Acknowledgements? These are already mined for funder information, so could be mined for data citations as well.
- Accession codes? In 2011, Nature published a data DOI for the first time (see http://dx.doi.org/10.1038/nbt.1992 – an article on the genomes of rhesus macaques), and later, in a paper on the recent outbreak of E-Coli in Germany, published the DOI for a dataset held by the Beijing Genomic Institute for the first time. In both cases the editors decided to put the citation in with accession codes rather than the reference list as the datasets hadn't been peer-reviewed.
- Reference list?

This is something that's still being worked out by the movers and shakers, but if data is to be thought of as a first-class research output, it really should be in the reference list. While we're on this topic, there's a related issue in the case of data reuse that if the data

citation is in the reference list, should it appear alongside or independently of a reference to the related article?

The data might well be useless without the kind of context that a journal article provides, but in print journals with a limit on the number of references, one could be consider it a waste of a slot to include citations to both the paper and the data. This is an area where pervasive forward linking would solve a lot of problems. If publishers can be sure that when a reader follows a link to a dataset, the landing page would forward them on to the data collection paper and any other papers using it, or even other high quality documentation, they might be more open to accepting a lone data citation where it is appropriate.

That is why we are recommending the following:

- Include the citation in the reference list – some reference management packages now include support for datasets, which should make this easier.
- When your data collection paper is published, notify the repository holding the dataset.
- When you publish a paper in which you reuse a prior dataset, notify the repository holding that dataset.

The other issue I want to talk about is dataset identifiers, and how they should be applied to dynamic datasets. There are two ways a dataset can be dynamic (Figure 5). The first *(animate)* is where the dataset is fairly stable in its extent, but points are revised every so often. A table of the masses of subatomic particles would fall under that category.

- Revised datasets

- Expanding datasets

Figure 5: Types of dynamic datasets *(Click on illustrations to animate them)*

The other, more common case *(animate)* is where a dataset is continually expanded with new data, such as with sensor data.

There are three ways of making such datasets citable.

1. Differentiate versions by access date rather than ID

2. Take time slices

3. Take snapshots

The first option I know is adopted by the National Snow and Ice Data Center in the US, because first, in the disciplines they serve the dataset itself is more important than the version, and second, the Federation of Earth Science Information Partners of which they are a part believe that the identifiers they assign aren't identifiers at all but locations, because you can resolve them to addresses.[1] It's not a view I share, and so I'm not keen on this option. The second approach really only makes sense with expanding datasets, and even then works best if the researchers tend to use one slice of the set at a time. Even so, it is possible to combine it with the first approach, or the third one which is the one I reckon is most generally suitable; if the rate of change is particularly frequent, it would probably be best to take these snapshots on demand rather than at predefined intervals. The apparent downside of the third option is that it seems to involve massive duplication of data, but there's nothing to stop the data backend generating these snapshots on the fly from a single master sequence.

There's plenty more I could go on to talk about, but time is pressing so instead I'll flash the headlines before your eyes.

4 Guidance for researchers

When publishing a paper...

- Deposit any data you have collected and used as evidence.
- Ask for a persistent ID/URL for your deposited data.
- When your data collection paper is published, notify the repository holding the dataset.

When citing a prior dataset...

- Use the data citation style required by the editor/publisher.
- If no style is specified, use a standard data citation style, adapted to match the style for textual publications.
- Default to writing IDs in the form of URLs if possible.
- Include the citation in the reference list – some reference management packages now include support for datasets, which should make this easier.
- Cite datasets at the finest level that is appropriate and for which an identifier is provided.
- If that is not fine enough, provide details of the subset of data you are using at the point in the text where you make the citation.

[1] http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines#Note_on_Versioning_and_Locators

Figure 6: The How-to Guide has been cited in the literature

- Cite the exact version of the dataset you need.
- When your paper is published, notify the repository holding the dataset you used.

5 Guidance for data repositories

- Provide persistent IDs for the datasets you host.
  - The ID should remain unique.
  - The ID should always point to the same version.
  - The ID should resolve to a URL.
  - The URL should locate the dataset's landing page. This URL should belong to a landing page that contains descriptive information about the dataset, as well as links or instructions for accessing it.
- The explanatory metadata should not change for a dataset with a persistent ID.
- IDs should only be assigned once no further changes are expected.
- With dynamic datasets, provide IDs for snapshots or time slices.
- Provide sample citations on dataset landing pages.
- Link from landing pages to publications citing the dataset. This may require collaboration with authors and publishers.

6 Putting it into practice

In the year since we published this guidance it has made quite an impression (Figure 6). It was mentioned earlier this year by Matthew Mayernik in the *Bulletin of the American*

Society for Information Science and Technology[2] in the same breath as the guidelines put out some months earlier by the Federation of Earth Science Information Partners (ESIP). Once they saw them, ESIP themselves called our guide 'the most useful guide' on data citation (*transition*).[3] The correspondence paper 'Adventures in Data Citation' by Edmunds, Pollard, Hole, and Basford uses the guide as shorthand for best practice in data citation,[4] as does the editorial in the inaugural issue of the data journal GigaScience.[5] Most recently (*transition*), Thomson Reuters refer to it in their essay on the selection policy for their new *Data Citation Index*.[6]

It is good to see the guidelines being used in practice, but the landscape is developing all the time. So we're keeping a watchful eye on the evolution of data citation practices, and hope to bring out an updated version of the guide in the first half of next year.

Monica Duke, Alex Ball. DCC/UKOLN, University of Bath. http://www.ukoln.ac.uk/ukoln/staff/

The DCC is funded by JISC.

For more information, please visit http://www.dcc.ac.uk/

[2] http://www.asis.org/Bulletin/Jun-12/JunJul12_MayernikDataCitation.html

[3] http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines#Introduction_and_Summary

[4] http://dx.doi.org/10.1186/1756-0500-5-223

[5] http://dx.doi.org/10.1186/2047-217X-1-11

[6] http://wokinfo.com/media/pdf/DCI_selection_essay.pdf