Hello, my name is Alex Ball and I work for the DCC. Now, how many people here have been at the conference over the last couple of days? Well then, you’ll all be aware that the DCC has been publishing a new set of guidance documents based on themes derived from JISC’s Managing Research Data Programme.

One of the themes is data citation, and we’ve recently published both a Briefing Paper and a How-to Guide on the subject (slide). Those of you with conference packs will have a copy of the How-to Guide and I have some spare copies here. You can also download a copy from our website (Figure 1) 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

The guide was co-written by myself and my colleague Monica Duke. I will be honest with you that out of the two of us, Monica is the expert and I’m a relative novice, but I learnt a lot while writing this guide. 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 I came across when putting the guide together.

1 Motivation

I guess I don’t need to convince anyone here about the need for data citation and linking from data to associated publications, but for completeness here’s a few thoughts on the matter. First, what made journal publication 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?

It 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. Now, if you were here on Tuesday you’ll have heard Victoria Stodden talking about reproducible research, and the fact that journal papers really aren’t enough any more (transition); we also need the underlying data. But we won’t get data routinely shared until all these things apply to data as well. Given time, I would argue (Figure 3) that data citations are what will make it happen.

- Visibility for data
- Protection from plagiarism
- Possibility for verification of results
- Data on which to base future research
- Possibility for reward models
- Access

Figure 3: What data citations provide

As this is a workshop about linking I should say just a few words about linking from datasets to the papers that cite them (slide). As with any resource, they are important for gauging the impact of the dataset, placing it in context within the literature and if there are issues with it, pointing readers to a critique.

But back to data citations. What should they 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.

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


Lawrence et al. (2008): BADC


Green (2010): OECD

Starr and Gastl (2011): DataCite


Of these elements, I reckon the most important ones to include are the ones you would normally put in any citation to an online resource:

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

The location should be persistent enough to be an identifier, or put another, the location should be expressed in terms of an identifier and a resolver service. DOIs are getting the most traction for datasets that are considered ‘published’, with Handles and ARKs being used more for ephemeral datasets.

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 earlier this 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.

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? Earlier this year, 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 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.

There’s just one other issue I want to talk about and that’s how repositories should make dynamic datasets citable. There are two ways a dataset can be dynamic (Figure 5): please excuse the lack of animations. The first 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. The other, more common case is where a dataset is continually expanded with new data, such as with sensor data.

- Revised datasets
- Expanding datasets

There are three ways of making such datasets citable.

1. Differentiate versions by access date rather than ID; example: National Snow and Ice Data Center
   A

2. Take time slices
   A
   B
   C

3. Take snapshots
   A
   B
   C

I really don’t recommend the first option, even if it is true in some disciplines that the dataset itself is more important than the version. The second approach really only makes sense with expanding datasets, and even then works best if the researchers tend to use small portions of the set at a time. The third one 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 in front of 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.
- 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.

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