A Digital Curation Centre ‘working level’ guide

How to Cite Datasets and Link to Publications

Alex Ball (DCC) and **Monica Duke** (DCC)

How to Cite Datasets and Link to Publications

Introduction

This guide will help you create links between your academic publications and the underlying datasets, so that anyone viewing the publication will be able to locate the dataset and vice versa. It provides a working knowledge of the issues and challenges involved, and of how current approaches seek to address them. This guide should interest researchers and principal investigators working on data-led research, as well as the data repositories with which they work.

Why cite datasets and link them to publications?

The motivation to cite datasets[1] arises from a recognition that data generated in the course of research are just as valuable to the ongoing academic discourse as papers and monographs. Scientific journals have traditionally supported research by disseminating knowledge in such detail that first, peer scientists could judge the strength of the conclusions based on the quality of the premises and research methods employed, and second, further investigations could be based upon it. In many disciplines, though, the paper alone is no longer sufficient for these purposes: the underlying data also need to be shared.[2,3,4]

As a medium, the journal paper owes its success in part to the control systems put in place around it: mechanisms allowing authors to be open about their research while still receiving due credit; metrics used to translate such attributions into rewards for authors and their institutions; and archives ensuring that the work is permanently available for reference and reuse.[5] If datasets are to be regarded as first-class records of research, as they need to be, a similar set of control systems needs to be constructed around them.

A major part of this work can be achieved using a robust citation mechanism for referencing datasets from within traditional publications. Provided the citation contains the name of a responsible agent, it can be used to assign due credit. By providing a globally unique identifier, it can be used to track the impact of a particular dataset. A citation is also an ideal place to provide the information needed to locate and access the dataset. In this way, datasets can take advantage of the infrastructure already in place to manage journal papers.

The rise of electronic journals has led to new and valuable services being layered over the top of papers, among them the provision of forward links to papers citing the current one. Such links help the reader to gauge the impact of the paper, place it within the literature and in some cases gain awareness of flaws or issues discovered by others. Forward links from datasets to the papers that cite them provide all the same benefits, as well as ensuring that documentation for the dataset can be found.

Ultimately, bibliographic links between datasets and papers are a necessary step if the culture of the scientific and research community as a whole is to shift towards data sharing, increasing the rapidity and transparency with which science advances.

[1] The term 'dataset' is used throughout this guide to mean a logically complete set of data; some systems or services prefer the terms 'data product' or 'data package'.

[2] Stodden, V. (2009). Enabling reproducible research: Open licensing for scientific innovation. *International Journal of Communications Law and Policy*, *13*, 1–25. Retrieved 2 September 2010, from `http://www.ijclp.net/files/ijclp_web-doc_1-13-2009.pdf`.

[3] *Open to all?: Case studies of openness in research*. (2010, September). Research Information Network and National Endowment for Science, Technology and the Arts. Retrieved 1 May 2011, from `http://www.rin.ac.uk/system/files/attachments/NESTA-RIN_Open_Science_V01_0.pdf`.

[4] Lynch, C. (2009). Jim Gray's fourth paradigm and the construction of the scientific record. In T. Hey, S. Tansley & K. Tolle (Eds.), *The fourth paradigm: Data-intensive scientific discovery* (pp. 177–183). Redmond, WA: Microsoft Research. Retrieved 14 July 2010, from `http://research.microsoft.com/en-us/collaboration/fourthparadigm/`.

[5] Mackenzie Owen, J. (2007). *The scientific article in the age of digitization* (ch. 2). Information Science and Knowledge Management. Dordrecht: Springer. `doi:10.1007/1-4020-5340-1`.

Requirements for data citations

The SageCite Project has identified a set of requirements for dataset citations and any services set up to support them.[6]

The citation itself must be able to identify uniquely the object cited, though different citations might use different methods or schemes to do so.

It must be able to identify subsets of the data as well as the whole dataset.

It must provide the reader with enough information to access the dataset; indeed, when expressed digitally it should provide a mechanism for accessing the dataset through the Web infrastructure.

It must be usable not only by humans but also by software tools, so that additional services may be built using these citations. In particular, there need to be services that use the citations in metrics to support the academic reward system, and services that can generate complete citations.

Elements of a data citation

The elements that would make up a complete citation are a matter of some debate. The following list is a superset taken from four different papers on the subject. [7,8,9,10]

Author. The creator of the dataset.[7,8,9,10]

Publication date. Whichever is the later of: the date the dataset was made available,[7] the date all quality assurance procedures were completed,[8,9] and the date the embargo period (if applicable) expired.[10]

Title. As well as the name of the cited resource itself,[7,10] this may also include the name of a facility[8] and the titles of the top collection and main parent sub-collection (if any) of which the dataset is a part.[9]

Edition. The level or stage of processing of the data, indicating how raw or refined the dataset is.[8]

Version. A number increased when the data changes, as the result of adding more data points or re-running a derivation process, for example.[10]

Feature name and URI. The name of an ISO 19101:2002[11] 'feature' (e.g. GridSeries, ProfileSeries) and the URI identifying its standard definition, used to pick out a subset of the data.[8]

Resource type. Examples: 'database',[9] 'dataset'.[10]

Publisher. The organisation either hosting the data[10] or performing quality assurance.[8]

Unique numeric fingerprint (UNF). A cryptographic hash of the data, used to ensure no changes have occurred since the citation.[7]

Identifier. An identifier for the data, according to a persistent scheme.[7,8,9,10]

Location. A persistent URL from which the dataset is available. Some identifier schemes provide these via an identifier resolver service.[7,8,9,10]

The most important of these elements – the ones that should be present in any citation – are the author, the title and date, and the location. These give due credit, allow the reader to judge the relevance of the data, and permit access the data, respectively. In theory, they should between them uniquely identify the dataset; in practice, a formal identifier is often needed. The most efficient solution is to give a location that consists of a resolver service and an identifier (for an example, see Figure 3 on page 4).

Note that the way in which these elements would be styled and combined together in the finished citation depends on the style in use for citations of textual publications. Figure 1 provides example data citations drawn from commonly used style manuals, while Figure 2 shows the citation formats suggested by three data repositories.

Digital Object Identifiers

There are several types of persistent identifier that could be used to identify datasets: examples include Handles, Archival Resource Keys (ARKs) and Persistent

[6] Duke, M. (2011, August 22). Requirements for data citation: The prequel [Blog post]. Retrieved 22 August 2011, from the SageCite blog: `http://blogs.ukoln.ac.uk/sagecite/2011/08/22/requirements-for-data-citation-the-prequel/`.

[7] Altman, M. & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*, *13*(3/4). `doi:10.1045/march2007-altman`

[8] Lawrence, B. N., Jones, C. M., Matthews, B. M. & Pepler, S. J. (2008, February 1). *Data publication* (Claddier Project Report No. 3). BADC. Retrieved 11 May 2011, from `http://purl.org/oai/oai:epubs.cclrc.ac.uk:work/43641`

[9] Green, T. (2010, February). *We need publishing standards for datasets and data tables*. OECD Publishing. `doi:10.1787/787355886123`

[10] Starr, J. & Gastl, A. (2011). isCitedBy: A metadata scheme for DataCite. *D-Lib Magazine*, *17*(1/2). `doi:10.1045/january2011-starr`

[11] ISO 19101. (2002). *Geographic information – Reference model*. 1st ed. International Organization for Standardization.

APA

Cool, H. E. M., & Bell, M. (2011). *Excavations at St Peter's Church, Barton-upon-Humber* [Data set]. doi: 10.5284/1000389

Chicago (notes)

2. H. E. M. Cool and Mark Bell, Excavations at St Peter's Church, Barton-upon-Humber (accessed May 1, 2011), doi:10.5284/1000389.

Cool, H. E. M., and Mark Bell. Excavations at St Peter's Church, Barton-upon-Humber (accessed May 1, 2011). doi:10.5284/1000389.

MLA

Cool, H. E. M., and Mark Bell. "Excavations at St Peter's Church, Barton-upon-Humber." Archaeology Data Service, 2001. Web. 1 May 2011. ⟨http://dx.doi.org/10.5284/1000389⟩.

Oxford

Cool, H. E. M. & Bell, M. (2011) *Excavations at St Peter's Church, Barton-upon-Humber* [data-set]. York: Archaeology Data Service [distributor] ⟨DOI 10.5284/1000389⟩

Figure 1: Data citations in common styles[12]

PANGAEA

Willmes, S et al. (2009): Onset dates of annual snowmelt on Antarctic sea ice in 2007/2008. `doi: 10.1594/PANGAEA.701380`

Dryad

Kingsolver JG, Hoekstra HE, Hoekstra JM, Berrigan D, Vignieri SN, Hill CE, Hoang A, Gibert P, Beerli P (2001) Data from: The strength of phenotypic selection in natural populations. Dryad Digital Repository. `doi: 10.5061/dryad.166`

Dataverse

Frederico Girosi; Gary King, 2006, 'Cause of Death Data', `http://hdl.handle.net/1902.1/UOVMCPSWOL` `UNF:3:9JU+SmVyHgwRhAKclQ85Cg==` IQSS Dataverse Network [Distributor] V3 [Version].

Figure 2: Data citation formats suggested by repositories

URLs (PURLs), all of which can be resolved to an Internet location. Arguably the scheme that is gaining most traction is the Digital Object Identifier (DOI).

The DOI System is an identifier scheme administered by the International DOI Foundation.[13] It is built on the Handle System but has its own conventions and an independent business model. The identifiers themselves have the standard Handle structure of prefix, slash, suffix (see Figure 3). All DOI prefixes begin with '`10.`' to mark them as such; the prefix may be further subdivided with dots, but otherwise the characters in a DOI have no special significance.

$$\underbrace{\text{http://dx.doi.org/}}_{\text{resolver service}}\underbrace{\text{10.5284}}_{\substack{\text{prefix}\\\text{(assigning}\\\text{body)}}}/\underbrace{\text{1000389}}_{\substack{\text{suffix}\\\text{(resource)}}}$$

Figure 3: Anatomy of a DOI

While there are several services available that can resolve a DOI to an Internet location,[14] the preferred one is `http://dx.doi.org/`. Appending a DOI to this URL creates a further URL that can be used to access the associated resource.

The task of managing the DOI registers is delegated to registration agencies that each specialise in a type of resource. For research datasets, the registration agency is the DataCite Consortium.[15] The consortium is made up of libraries and data centres from across the globe, led by the German National Library of Science and Technology (TIB). Among the services it provides are human and machine interfaces for simple end-user administration of DOI registrations. DataCite also collects metadata about each dataset it registers.[16] These metadata may be searched through a Web interface[17] or harvested using OAI-PMH.[18]

Individuals wishing to register a DOI for their dataset normally do so via their data repository, rather than directly through DataCite. Any *repository* wishing to register DOIs needs to obtain a username and password from DataCite to gain access to the registration service. Alternatively, the organisation can manage its DOIs through a third-party service such as EZID.[19] The username and password are not needed for the metadata search or OAI-PMH services.

[12] *Publication Manual of the American Psychological Association* (6th ed., p. 211). (2010). Washington, DC: American Psychological Association. *Chicago Manual of Style* (16th ed., p. 764). (2010). Chicago, IL: University of Chicago Press. Gibaldi, J. (2008). *MLA style manual and guide to scholarly publishing* (3rd ed., pp. 213–214, 238–239). New York: Modern Language Association of America. R. M. Ritter (Ed.). (2002). *Oxford Manual of Style* (p. 551). Oxford, UK: Oxford University Press.

[13] DOI System Website, URL: `http://www.doi.org/`.

[14] Some publishers provide resolvers for their own DOIs, while the Handle resolver `http://hdl.handle.net/` can be used for any DOI.

[15] DataCite Website, URL: `http://www.datacite.org/`.

[16] DataCite Metadata Schema Repository, URL: `http://schema.datacite.org/`. Version 2.0 of the DataCite metadata scheme is discussed by Starr & Gastl (2011).[10]

[17] DataCite Metadata Search service, URL: `http://search.datacite.org/`.

[18] DataCite OAI-PMH service, URL: `http://oai.datacite.org/`.

[19] EZID Website, URL: `http://www.cdlib.org/services/uc3/ezid/`.

While best practice has yet to emerge on some matters, (see 'Current issues and challenges' below), certain conventions are already becoming established.

- Authors should use the URL version of the DOI (i.e. including the resolver) wherever possible.
- When organisations register a DOI for a resource, they should not introduce semantic elements into the suffix, especially not metadata that might change over time (e.g. publisher, archive, owner).
- As DOIs are used to cite data as evidence, the dataset to which a DOI points should also remain unchanged, with any new version receiving a new DOI.

Example

Sage Bionetworks is a non-profit biomedical research organisation which is creating the Sage Commons,[20] an infrastructure for community-based modelling of large multi-contributor datasets.[21] The Commons already features a repository of curated datasets;[22] a new computational platform and repository front-end called Synapse will be added towards the end of 2011.

In this area of research, methods, tools and workflows are just as important as data. Taverna workflows, for example, provide a means of recording and documenting each step of the modelling process so that it can be shared with the scientific community. Furthermore, the workflows may be executed by Taverna Workbench, allowing the results from the pipeline to be reproduced. The SageCite Project worked with Sage Bionetworks to demonstrate both capturing workflows using Taverna, and making them citable resources using DataCite DOIs.[23]

Current issues and challenges

While the basics of data citation can be derived by analogy with the citation of textual publications, especially electronic ones, there are finer points such as issues of granularity, fine-grained and unambiguous credit and citation placement that merit special attention.

Granularity

With print publications, the issue of citing at different levels of granularity is relatively straightforward. The documents listed within a bibliography or reference section represent intellectual wholes: single-author monographs are referenced as whole books, but with journal issues, conference proceedings and edited collections the relevant papers are referenced individually. More granular references (to sections, pages, etc.) are made at the point of citation in the text, rather than in the bibliography.

Datasets are a little more complicated. A dataset may form part of a collection and be made up of several files, each containing several tables, each containing many data points. There are also more abstract subsets that can be used, such as features and parameters. At the other end of the scale, it is not always obvious what would constitute an intellectual whole: it can be argued, for example, that investigations should be the primary units of citation rather than individual datasets.[24] For authors, the pragmatic solution is to list datasets at whatever level of granularity has been chosen by the host repository for assigning identifiers. If a finer level of granularity is required, the in-text citation should provide the reader with the information needed to find the subset. As conventions for doing this have yet to be established, if the repository provides identifiers at several levels of granularity, the finest-grained level that meets the need of the citation should be used in the bibliography, to minimise the additional information needed.

Microattribution

Where a dataset is assembled from very many contributions, crediting each contributor individually becomes unfeasible using traditional techniques. Microattribution is a way of crediting contributors in a more compact fashion, to keep the operation manageable. It can also be used to credit people or organisations whose contributions don't fit the roles of creator or compiler: for example, those who implement or carry out intermediate data processing steps.

[20] Sage Bionetworks Commons Web page, URL: http://sagebase.org/commons/.

[21] Derry, J. M. J., Mangravite, L. M., Suver, C., Furia, M., Henderson, D., Schildwachter, X., … Friend, S. H. (2011, April 4). Developing predictive molecular maps of human disease through community-based modeling. *Nature Precedings*. doi:10.1038/npre.2011.5883.1

[22] Furia, M. & Sieberts, S. (2011, March 31). *Sage Bionetworks data curation guidelines*. Version 2.1. Sage Bionetworks. Retrieved 15 August 2011, from http://precedings.nature.com/documents/5883/version/1/files/npre20115883-1.pdf

[23] SageCite Project blog, URL: http://blogs.ukoln.ac.uk/sagecite/.

[24] Lawrence, B. (2011, January 7). Citation, Digital Object Identifiers, persistence, correction and metadata [Blog post]. Retrieved 12 May 2011, from http://home.badc.rl.ac.uk/lawrence/blog/2011/01/07/citation,_digital_object_identifiers,_persistence,_correction_and_metadata.

Instead of providing a traditional citation to the data collection paper associated with each contribution, a table is produced that lists each contribution and the agent responsible. Where possible, standard identifiers (for both contributions and contributors) are used to abbreviate the entries, and the table is included in the paper's supplementary data.

This technique is still relatively new: the first paper to use microattribution to encourage comprehensive sharing of genetic variation data in a defined system was published in 2011.[25] Once the technique is more established, repositories should consider making microattribution data available in machine-interpretable form, rather than as supplementary spreadsheets, to aid its use in metrics and other services.

Contributor identifiers

If contributors have a common name, or move between many different institutions, giving them an unambiguous credit is somewhat problematic. A possible solution is for each contributor to be given a unique identifier, to be used in connection with all their publications, data contributions, and so on. While several identifier schemes are already well established, most are arguably unsatisfactory because they are either too narrowly scoped, proprietary or focused on authentication rather than attribution. There are however two schemes being developed specifically for attribution.

The Open Researcher and Contributor Identifier (ORCID) is a scheme specifically aimed at academic authors.[26] It has gained support from over 200 organisations, including major academic publishers. The underlying infrastructure is still being developed as of mid-2011, but the intention is to maintain a registry of IDs, each associated with a researcher profile and a list of publications to which that researcher has contributed. The registry will also allow the profile to be linked to identifiers and profiles from other schemes such as Thomson Reuters' ResearcherID,[27] Scopus,[28] Scholar Universe,[29] and RePEc.[30]

The International Standard Name Identifier (ISNI) scheme is a draft ISO standard for registering 'Public Identities': people, pseudonyms, personas and legal entities involved in the creation or distribution of intellectual property.[31] It is thus a broader scheme than ORCID, allowing organisations to be identified as well as individuals. ISNIs take the form of a 16-digit number (though the last digit may be 'X'); each identifier is supported by a metadata record containing details such as name(s), date of birth, fields of endeavour and roles within them, titles of creations and a URI for further information.

As the primary utility for such identifiers will be to support software tools, they will probably be better placed in machine-readable metadata than written out for human inspection. Nevertheless, the ORCID Initiative envisages ORCID IDs being included in parentheses after author names in textual citations, as in Figure 4.

> Chaturvedi, V. (AAA-1019-2010). (2004). Editorial. *Mycopathologia*, *157*, iii–iv. Retrieved from `http://dx.doi.org/10.1023/B:MYCO.0000020677.89178.15`

Figure 4: Example citation using an ORCID ID

Placement of data citations

Treating datasets as first-class records of research implies placing citations to them in the bibliography, works cited or references section of a document. This is required by Pensoft journals, for example, which also specify that the in-text pointer to the full citation should occur in a dedicated 'Data Resources' section.[32]

There is, however, a special relationship between a dataset and the paper describing its collection (as opposed to subsequent papers that cite it); it could be argued that the way to mark this would be to include the (full) data citation elsewhere in the document.[33] The data publishing journal *Earth System Science Data*, for example, usually cites the collected data in a dedicated 'Data coverage and parameter measured' section. Alternatively, if the acknowledgements section

[25] Giardine, B., Borg, J., Higgs, D. R., Peterson, K. R., Philipsen, S., Maglott, D., ... Patrinos, G. P. (2011). Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. *Nature Genetics*, *43*, 295–301. `doi:10.1038/ng.785`.

[26] ORCID Initiative Website, URL: `http://www.orcid.org/`.

[27] ResearcherID Website, URL: `http://www.researcherid.com/`.

[28] Scopus Website, URL: `http://www.scopus.com/`.

[29] Scholar Universe, URL: `http://www.scholaruniverse.com/`.

[30] RePEc Author Service Website, URL: `http://authors.repec.org/`.

[31] ISO/DIS 27729. (n.d.). *Information and documentation – International standard name identifier (ISNI)*. Draft International Standard. International Organization for Standardization.

[32] Penev, L., Mietchen, D., Chavan, V., Hagedorn, G., Remsen, D., Smith, V. & Shotton, D. (2011, May 26). *Pensoft data publishing policies and guidelines for biodiversity data*. Pensoft. Retrieved 4 July 2011, from `http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf`.

[33] Piwowar, H. (2011, May 5). Links from the data collection article: Inline or in the bibliography? [Blog post]. Retrieved 3 June 2011, from the Research Remix blog: `http://researchremix.wordpress.com/2011/05/05/inline-or-biblio/`.

is already being mined for funder information, it may be appropriate to put the data citation there.[34]

On the other hand, there is value in citing datasets consistently across all papers, in terms of simplifying both editorial guidelines and author training. Bibliographies also tend to be better indexed and more freely available than the main texts of papers, and would therefore afford the citation greater visibility.

Summary for researchers

If you have generated/collected data to be used as evidence in an academic publication, you should deposit them with a suitable data archive or repository as soon as you are able. If they do not provide you with a persistent identifier or URL for your data, encourage them to do so.

When citing a dataset in a paper, use the citation style required by the editor/publisher. If no form is suggested for datasets, take a standard data citation style (e.g. DataCite's[10]) and adapt it to match the style for textual publications.

Give dataset identifiers in the form of a URL wherever possible, unless otherwise directed.

Include data citations alongside those for textual publications. Some reference management packages now include support for datasets, which should make this easier.

Cite datasets at the finest-grained level available that meets your need. 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.

If a dataset exists in several versions, be sure to cite the exact version you used.

When you publish a paper that cites a dataset, notify the repository that holds the dataset, so it can add a link from that dataset to your paper.

The remainder of this guide is aimed at those responsible for the supporting infrastructure, rather than researchers.

Building a citation infrastructure

This section provides an overview of some of the technologies available to support data citation.

Citation Notification Service

The TrackBack protocol is one of a family of linkback protocols that allow a blog article to list and link to later articles that mention or comment on it, allowing the reader to follow a debate across many blogs.[35] It works in the following way. On publication of an article, the blogging software looks up all the pages to which the article links, and scans them for embedded TrackBack URLs. Having found one, the software sends an HTTP POST request (as used by longer Web forms) to the TrackBack URL. At a minimum, the request contains a link to the article; it may also contain the article's title, the title of the blog, and an excerpt typically showing the link in context. The blog responsible for the TrackBack URL then sends back a brief XML acknowledgement to indicate either success or failure in understanding the request, known as a TrackBack 'ping'.

The CLADDIER Project defined an extended version of the TrackBack protocol for use as a Citation Notification Service in digital object repositories.[36] The main extensions were, at the sending end,[37,38]

- 'metadata' and 'metadataformat' fields for adding arbitrary metadata to the TrackBack ping;
- a 'type' field to allow the same protocol to be used for forward citations ('reverse TrackBacks') and republications;
- an 'action' field to allow existing TrackBacks to be removed (an 'anti-TrackBack') or edited;

and at the receiving end

- additional RDF metadata that could be embedded alongside the TrackBack URL, such as

[34] *Acknowledgement of funders in scholarly journal articles: Guidance for UK research funders, authors and publishers*. (2008, February). Research Information Network. Retrieved 3 June 2011, from `http://www.rin.ac.uk/our-work/research-funding-policy-and-guidance/acknowledgement-funders-journal-articles`.

[35] Six Apart. (2007). TrackBack manual. Retrieved 18 October 2011, from `http://www.movabletype.org/documentation/trackback_manual.html`.

[36] CLADDIER Project Website, URL: `http://claddier.badc.ac.uk/trac/`.

[37] Matthews, B., Portwin, K., Jones, C. & Lawrence, B. (2007, November 30). *Recommendations for data/publication linkage* (CLADDIER Project Report No. 3). STFC. Retrieved 18 May 2011, from `http://claddier.badc.ac.uk/trac/attachment/wiki/WikiStart/Report_III_RecommendationsForDataLinking-final.doc`.

[38] Matthews, B., Duncan, A., Jones, C., Neylon, C., Borkum, M., Coles, S. & Hunter, P. (2009, December). A protocol for exchanging scientific citations. *Fifth IEEE International Conference on e-Science (e-Science 2009)* (pp. 171–177). Los Alamitos, CA: IEEE Computer Society. `doi:10.1109/e-Science.2009.32`.

bibliographic information about the citable resource (to permit reverse TrackBacks) or an alternative URL to which to send anti-TrackBack pings;

- the option to use a whitelist of trusted senders to prevent spam.

As a demonstration, CLADDIER implemented the Citation Notification System in STFC's ePub repository and the BADC repository. The follow-on project StoreLink implemented the system as plugins for EPrints, DSpace and Fedora repository software.[39] StoreLink was itself followed by the Webtracks Project, which is generalising the system to form the Inter-Repository Communication (InteRCom) protocol and extending its usage beyond e-print repositories to STFC's ICAT data catalogue, open electronic notebooks and scientific publishers.[40]

Example

Knowledge Blog is an alternative scholarly publication platform based on WordPress blogging software.[41,42] It makes heavy use of linkbacks, for example as the mechanism for linking an article with its reviews, and could therefore be used together with the Citation Notification Service to provide bi-directional links to datasets. Its KCite plugin allows for the automatic generation of citations from just a DOI (using the metadata lookup API from the CrossRef registration agency) or a PubMed identifier,[43] though this has not yet been extended to work with DataCite DOIs.

Nanopublications

A nanopublication is, simply put, a statement and a set of annotations on it, the whole of which is citable in its own right.[44] The idea is that a scientific publication or dataset is broken down into individual statements, expressed as RDF triples: that is, in the form subject–predicate–object, e.g. malaria is-carried-by mosquitoes. Each of these statements is assigned a URI and then made the object of further statements (annotations) that say, for example, who made the statement, the document or dataset from which the statement was extracted, the date the statement was published. The set formed by the original statement and these annotations is itself given a URI and thus becomes a nanopublication.

The reason for doing this is to provide a robust mechanism for aggregating information and data into a knowledge base from which new inferences may be drawn. The robustness comes from the annotations, which provide a resource for assessing the reliability of the statement. A nanopublication of a statement is said to contribute to the 'S-Evidence' for that statement; if, on aggregating a large number of nanopublications, one ends up with two conflicting statements, one would compare the S-Evidence for each statement to decide which should be used for further inference.

In order to make this work, one needs to be able to identify unambiguously every concept and entity to which the nanopublications refer. Nanopublications are therefore best suited to disciplines which are already well supported by RDF-friendly ontologies. For concepts and entities that do not sit easily within a formal ontology, a more relaxed approach such as that provided by the Concept Wiki can be used.[45]

Citation Typing Ontology

The Citation Typing Ontology (CiTO) is a formal language for specifying why one resource cites another.[46] It contains several terms particularly relevant for data citation; additional terms can be found in the extension ontology CiTO4Data.[47]

- *Uses data from/provides data for.* These terms describe the relationship between a dataset and a paper describing work using that dataset.
- *Cites as data source/is cited as data source by.* These terms imply the above relationship but

[39] StoreLink Project summary Web page, URL: http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/storelink.aspx.

[40] Webtracks Project Web page, URL: http://www.stfc.ac.uk/e-Science/projects/medium-term/metadata/webtracks/22422.aspx.

[41] Knowledge Blog Website, URL: http://knowledgeblog.org/.

[42] WordPress Website, URL: http://wordpress.org/.

[43] Lord, P., Cockell, S., Swan, D. C. & Stevens, R. (2011, June 7). The Ontogenesis Knowledgeblog: Lightweight semantic publishing [Blog post]. Retrieved 13 July 2011, from Knowledge Blog: http://knowledgeblog.org/128

[44] Groth, P., Gibson, A. & Velterop, J. (2010, January). The anatomy of a nano-publication. *Information Services and Use*, *30*(1/2), 51–56. doi:10.3233/ISU-2010-0613.

[45] The Concept Wiki Website, URL: http://www.conceptwiki.org/.

[46] Shotton, D. (2010). CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics*, *1*(Suppl 1), S6. doi:10.1186/2041-1480-1-S1-S6.

[47] Shotton, D. & Peroni, S. (2011b, March 30). *CiTO, the Citation Typing Ontology*. Version 2.0. Retrieved 26 May 2011, from http://purl.org/spar/cito/. Shotton, D. & Peroni, S. (2011a, February 25). *CiTO4Data, the Citation Typing Ontology for Data*. Version 1.0. Retrieved 26 May 2011, from http://purl.org/spar/cito4data/.

also indicate that the paper formally cites the dataset.

- *Contains assertion from/provides assertion for.* These terms describe, for example, the relationship between a full dataset and a nanopublication based upon it.
- *Compiles/is compiled by.* These terms describe, for example, the relationship between a dataset and the software used to derive it.

Certain of the other terms may be useful in clarifying how datasets or nanopublications relate to one another, e.g. *confirms/is confirmed by*, *corrects/is corrected by*, *disagrees with/is disagreed with by*, *extends/is extended by*, *updates/is updated by*.

Data citation infrastructures

The following repositories and systems provide examples of data citation infrastructures in practice, both in terms of human workflows and software, that could be reused by other repositories. Sample citations provided by each of them can be found in Figure 2 on page 4.

PANGAEA

PANGAEA (Data Publisher for Earth and Environmental Science) is hosted by the Alfred Wegener Institute for Polar and Marine Research and the Center for Marine Environmental Sciences in Germany.[48] It is the data archive and distribution system for the World Data Centre for Marine Environmental Sciences (WDC-MARE) and the designated archive for the data publishing journal *Earth System Science Data*.

Throughout its history, PANGAEA has collaborated extensively with scientific publishers; it provides links from data holdings to the traditional publications that reference them, and wherever possible, those publications reference the holdings in PANGAEA. Initially datasets were cited using standard URLs, but now DOIs are used as the canonical identifier for all PANGAEA holdings.[49]

Once the author has uploaded the data and metadata, a curator checks the completeness of the metadata and consistency of the data, then imports the data into the archive. Having checked that the data are properly indexed by the system, the curator performs technical quality control tests, sets appropriate access conditions and refers the result back to the author for proofing. Once the author and curator are both satisfied, the data are published and assigned a DOI. Once this has happened, the metadata and data are both considered static.

The middleware component of PANGAEA, panFMP, has been released as open source software.[50] Some of the associated visualisation and conversion tools have been made available as freeware.[51]

Dryad

Dryad is a data repository specialising in evolutionary biology and ecology, developed by the National Evolutionary Synthesis Center and the University of North Carolina Metadata Research Center.[52] It is a preferred data archive for several journals including *The American Naturalist*, *Molecular Ecology*, *Molecular Biology and Evolution*, *Evolutionary Applications*, *Heredity* and *Nature*.

Dryad has now settled on DOIs to identify its datasets. As with PANGAEA, catalogue records for the data holdings in Dryad contain the citation of the accompanying publication as well as a sample citation for the data itself.

After the author has submitted the data and metadata to Dryad, a curator checks that the files contain the right sort of information before performing a series of quality control procedures. When these have been completed, a DOI is assigned to the data and sent to the author, and the catalogue record goes live in the repository. The record is updated with the citation of the data collection paper once it is published.[53]

Dryad is based on the DSpace digital repository;[54] the Dryad extensions have been released as open source software.[55]

[48] PANGAEA Website, URL: http://www.pangaea.de/.

[49] Diepenbroek, M., Schindler, U. & Grobe, H. (2008). PANGAEA: An ICSU World Data Center as a networked publication and library system for geoscientific data. *WEBIST 2008: Proceedings of the 4th International Conference on Web Information Systems and Technologies* (Vol. 2, pp. 149–154). Funchal, Madeira, Portugal. Institute for Systems and Technologies of Information, Control and Communication. Retrieved 23 May 2011, from http://hdl.handle.net/10013/epic.28613.

[50] PANGAEA Framework for Metadata Portals Website, URL: http://www.panfmp.org/.

[51] PANGAEA Software Web page, URL: http://www.pangaea.de/software/.

[52] Dryad Website, URL: http://datadryad.org/.

[53] Feinstein, E. (2010, December 2). What happens after you submit your data to Dryad? [Blog post]. Retrieved 24 May 2011, from the Dryad News and Views blog: http://blog.datadryad.org/2010/12/02/what-happens-after-you-submit-your-data-to-dryad/.

[54] DSpace Website, URL: http://www.dspace.org/.

[55] Dryad code repository, URL: http://dryad.googlecode.com/.

Dataverse

The Dataverse Network is a software application for building data repositories called dataverses.[56] It is developed by a community led by the Institute for Quantitative Social Science (IQSS) at Harvard University. As well as the original Dataverse Network at IQSS, there are also instances at the University of North Carolina and the University of the Thai Chamber of Commerce. Dataverses within the same Network may be cross-searched, and Dataverse Networks may also be linked to provide cross-searching facilities.

Authors may set up their own dataverse or contribute to an existing one. After filling out a metadata entry form and uploading the data files associated with a study, the author submits the data for review. The curator for the dataverse can then modify the metadata before releasing the study.[57]

Where data have been uploaded in SPSS, SATA or GraphML formats, a Unique Numeric Fingerprint is calculated for each data file and the study as a whole. In the IQSS Dataverse Network, studies are automatically assigned Handles. The catalogue page can display a citation for the corresponding data collection paper alongside a sample citation for the data.

Authors are welcome to upload data to the Henry A. Murray Research Archive at Harvard, or create their own dataverses in the IQSS Dataverse Network.[58] Alternatively, institutions can set up their own Dataverse Network using the open source software.[59]

Current implementation issues

Two current issues for repositories are how to cater for both manual and automatic uses of citations, and how to deal with dynamic datasets.

Manual and automatic use of citations

It is good practice for the URL in a data citation to lead to a *landing page* for the dataset, rather than to initiate a direct download. The landing page should enable readers to ensure they have located the right dataset, to (re-)familiarise themselves with the research context and supporting documentation, to consider licence terms prior to downloading and to switch to a more recent version (or otherwise-formatted representation) of the data if required. Landing pages also help to create a more even user experience between datasets available through direct access and those available through mediated access.

Since for the most part data are processed by software, it can help to accelerate progress if software tools are also able to retrieve data by means of the same URL. Software tools, like human readers, may wish to be selective with regard to versions and representations, to avoid data with an unsuitable licence, to download supporting documentation or data, or to select individual files or other subsets of the data. Such use cases require that the URL actually returns the machine-readable equivalent of a landing page. The technique used by the ACRID Project,[60] for example, is to provide an index of the data and metadata associated with a workflow in the form of an OAI-ORE Resource Map.[61]

Clearly humans and software have different requirements for the dataset landing page. One way to satisfy both would be to embed the metadata intended for software tools as RDF within the human-readable Web page. This can be done using either RDFa as in Figure 5,[62] or HTML5 microdata as in Figure 6.[63]

```
<html xmlns="http://www.w3.org/1999/xhtml"
      xmlns:cito="http://purl.org/spar/cito/"
      xmlns:dc="http://purl.org/dc/terms/"
      version="XHTML+RDFa 1.0">
...
<p>
  Supplement to: Author, A. (2011). ... <a
  about="http://dx.doi.org/10.9876/data123"
  rel="cito:providesDataFor" href="http://dx
  .doi.org/10.123/paper45">doi:10.123/paper45
  </a>
</p>
...
</html>
```

Figure 5: Example of using RDFa to embed a link to a publication within a dataset's Web page

[56] Dataverse Network Project Website, URL: http://thedata.org/.

[57] King, G. (2007). An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods and Research*, *36*(2). doi:10.1177/0049124107306660.

[58] Henry A. Murray Research Archive Website, URL: http://www.murray.harvard.edu/.

[59] Dataverse Network code repository, URL: http://sourceforge.net/projects/dvn/.

[60] ACRID Project Website, URL: http://www.cru.uea.ac.uk/cru/projects/acrid/.

[61] C. Lagoze, H. Van de Sompel, P. Johnston, M. Nelson, R. Sanderson & S. Warner (Eds.). (2008, October 17). *ORE user guide: Primer*. Version 1.0. Open Archives Initiative. Retrieved 1 June 2011, from http://www.openarchives.org/ore/1.0/primer.

[62] B. Adida & M. Birbeck (Eds.). (2008, October 14). *RDFa primer*. W3C Working Group Note. World Wide Web Consortium. Retrieved 1 June 2011, from http://www.w3.org/TR/xhtml-rdfa-primer/.

[63] I. Hickson (Ed.). (2011, May 25). *HTML microdata*. W3C Working Draft. World Wide Web Consortium. Retrieved 4 July 2011, from http://www.w3.org/TR/2011/WD-microdata-20110525/.

```
<p itemscope
   itemid="http://dx.doi.org/10.9876/data123">
  Supplement to: Author, A. (2011). ... <a
  href="http://dx.doi.org/10.123/paper45"
  itemprop="http://purl.org/spar/cito/
  providesDataFor">doi:10.123/paper45</a>
</p>
```

Figure 6: Example of using HTML5 microdata to embed a link to a publication within a dataset's Web page

An alternative method of serving both constituencies would be to use *content negotiation*. This is where the Web server keeps several different representations of a resource; when a Web client requests the resource, the server sends back the representation that best matches the client's preferred content type (as expressed by the 'Accept' HTTP header). In this case, the Web server would keep as the dataset landing page an HTML Web page for human readers and an RDF/XML document (say) for software tools.

While archives and repositories are broadly consistent in the information they provide to readers on their landing pages – descriptive metadata, a sample citation, a link to an accompanying paper, a link to the data files or instructions on how to access them, licence terms – they are still experimenting with the information they provide to software tools.

Example

Acta Crystallographica Section E (*Acta Cryst E*) is an online, open access data journal published by the International Union of Crystallographers.[64] It operates a workflow whereby data are submitted by authors at the same time as the data collection paper. The data are checked automatically for validity, and the validation report passed to reviewers. On publication, the data are made available for download from the page for the paper.

The XYZ Project has developed additional tools to support workflows like these.[65] It also explored, with *Acta Cryst E*, the possibility of embedding data directly within the Web pages of journal papers, using RDF and microformats in a profile of HTML known as Scholarly HTML.[66]

Versioning

One of the important features of the citation system is that a reader should be able to identify and retrieve the exact same resource that the author used when answering the research question. This is critical in the case of data as even typographical corrections may significantly change the conclusions drawn from a dataset. There is also the potential for many more versions from which to choose, since data may be made available in versions from different stages of processing,[67] as well as from different points in time. With this in mind, data repositories should ensure that different versions are independently citable (with their own identifiers).

The problem comes when repositories have to deal with rapidly changing datasets, and it is a slightly different problem depending on whether the dataset is frequently *revised*, that is, data points are continually improved or updated, or frequently *expanded*, such as sensor data maintained as a time series. Either way, to keep the versions manageable there are two possible approaches the data repository can take: time slices and snapshots.

With the *snapshot* approach, at regular intervals or at the request of a citing author, a snapshot is taken of the dataset and made citable. This is the better solution for revised datasets, as after retrieving the data the reader or author need not perform any additional operations to arrive at the required data. It is also better for expanding datasets where authors are concerned with the whole time series.

With the *time slice* approach, the citable entity becomes the set of updates made to a dataset during a particular time period rather than the full dataset itself (e.g. the 2008 data from a series running since 1950). This would be inappropriate for revised datasets, as the author or reader would need to assemble the data from a base file and several incremental change files. To a lesser extent, it would also be cumbersome for authors using a large proportion of an expanding dataset, as they would need to cite multiple time slices to build up the required range; but if an author is only concerned with data from a short period of time this approach is more suitable than a full snapshot.

Note that these discussions only concern how datasets are presented to users as citable resources. It does not affect how a repository might store the data,

[64] Acta Cryst E Website, URL: `http://journals.iucr.org/e/journalhomepage.html`.

[65] XYZ Project blog, URL: `http://projectxyz.wordpress.com/`.

[66] P. Sefton (Ed.). (2011, May 3). *Scholarly HTML core*. Retrieved 14 July 2011, from `http://scholarlyhtml.org/2011/05/03/scholarly-html-core-3/`

[67] Whether data from intermediate stages of processing should be made citable depends on the value added by processing, the reversibility of the technique and the utility of such data within the discipline.

so long as it can guarantee that the same identifier always returns the same data.

Summary for data repositories

Ensure that anyone wishing to cite a dataset you host can use a persistent identifier that you provide to do so. For this, choose an identifier scheme which allows the identifier to be resolved to a URL. This URL should belong to a landing page that contains descriptive information about the dataset, as well as links or instructions for accessing it.

Once an identifier has been assigned to a (version/snapshot of) a dataset, ensure that it and any explanatory metadata remain static over time. Ensure that the identifiers remain unique and associated with the correct versions.

Assign identifiers to static datasets only when no further changes or corrections are expected (i.e. after quality control checks are complete). For dynamic datasets, assign identifiers when new snapshots or time slices are created, whether this is on a regular basis or on demand.

Provide data depositors with a sample citation for their dataset, for use in academic publications.

Provide links from dataset landing pages to those published papers of which you are aware that cite the dataset. This may require collaboration with authors and publishers.

For more information about registering DOIs for datasets, contact your local DataCite member.[68] For more information about registering Archival Resource Keys,[69] contact the California Digital Library.

Acknowledgements

Thank you to Sarah Callaghan (STFC), Shirley Crompton (STFC), Michael Diepenbroek (WDC-MARE), Margaret Henty (ANDS), Catherine Jones (STFC), Sarah Jones (DCC), Florance Kennedy (DCC), Phillip Lord (Newcastle University) and Tom Pollard (BL) for helpful comments.

Further information

Two other DCC guides cover this topic:

Awareness Level: *Introduction to Curation: Data Citation and Linking* (2011) by Alex Ball and Monica Duke

Awareness Level: *Introduction to Curation: Persistent Identifiers* (2006) by Joy Davidson

Data citation. (2011, May 3). [Awareness Level Guide]. Retrieved 6 June 2011, from the Australian National Data Service: `http://www.ands.org.au/guides/data-citation-awareness.html`

Lane, M. A. (2008, September 10). *Data citation in the electronic environment*. Global Biodiversity Information Facility. Retrieved 2 September 2011, from `http://www.danbif.dk/Documents/gbif-documents/DataCitation-Lane2008.pdf`

Lawrence, B., Jones, C., Matthews, B., Pepler, S. & Callaghan, S. (2011). Citation and peer review of data: moving towards formal data publication. *International Journal of Digital Curation*, 6(2), 4–37. Retrieved 31 August 2011, from `http://www.ijdc.net/index.php/ijdc/article/view/181`

Newton, M. P., Mooney, H. & Witt, M. (2010). *A description of data citation instructions in style guides*. Poster presented at the 6th International Digital Curation Conference, Chicago, IL, 7–8 December 2010. Retrieved 24 August 2011, from `http://docs.lib.purdue.edu/lib_research/121/`

Page, R. (2009, April 20). Semantic publishing: Towards real integration by linking [Blog post]. Retrieved 11 May 2011, from the iPhylo blog: `http://iphylo.blogspot.com/2009/04/semantic-publishing-towards-real.html`

Why and how should I cite data? (2011). Retrieved 8 June 2011, from the Inter-University Consortium for Political and Social Research: `http://www.icpsr.umich.edu/icpsrweb/ICPSR/support/faqs/8707350211996719508`

Wilkinson, M. (2011a, July 28). So you want to cite your data: The consequences of data citation [Blog post]. Retrieved 16 August 2011, from the SageCite Knowledge Blog: `http://sagecite.knowledgeblog.org/2011/07/28/why-do-we-need-datacitation/`

Wilkinson, M. (2011b, July 28). Why do we need data citation: Take two [Blog post]. Retrieved 16 August 2011, from the SageCite Knowledge Blog: `http://sagecite.knowledgeblog.org/2011/07/28/why-do-we-need-data-citation-take-two/`

Please cite as: Ball, A. & Duke, M. (2011). 'How to Cite Datasets and Link to Publications'. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. Available online: `http://www.dcc.ac.uk/resources/how-guides`

Follow the DCC on Twitter: @digitalcuration, #ukdcc

[68] List of DataCite members, URL: `http://datacite.org/members`.

[69] Archival Resource Keys Website, URL: `https://confluence.ucop.edu/display/Curation/ARK`.