A Digital Curation Centre 'working level' guide

How to Cite Datasets and Link to Publications

Alex Ball (DCC) and Monica Duke (DCC)

Please cite as: Ball, A., & Duke, M. (2015). '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

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

[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`.

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.

Principles of data citation

The FORCE11 Data Citation Synthesis Group – whose members include representatives of the Research Data Alliance, the ICSU World Data System, and a range of projects[6] – has published a set of data citation principles.[7] The principles build on earlier work in this area, most notably by CODATA,[8] the (US) National Academies of Sciences, Engineering, and Medicine,[9] the DCC,[10] and the Institute for Quantitative Social Science, Harvard University,[11] and have been widely endorsed.[12]

Importance
Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.

Credit and Attribution
Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

Evidence
In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.

Unique Identification
A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

Access
Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

Persistence
Unique identifiers, and metadata describing the data, and its disposition, should persist – even beyond the lifespan of the data they describe.

Specificity and Verifiability
Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

Interoperability and Flexibility
Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

[6] FORCE11 Data Citation Synthesis Group, URL: https://www.force11.org/datacitation/workinggroup.
[7] FORCE11, Data Citation Synthesis Group. (2014). Joint declaration of data citation principles. Retrieved from https://www.force11.org/datacitation.
[8] CODATA/ITSCI Task Force on Data Citation. (2013). Out of cite, out of mind: The current state of practice, policy and technology for data citation. *Data Science Journal*, *12*, CIDCR1–CIDCR75. doi:10.2481/dsj.OSOM13-043.
[9] Uhlir, P. F. (Ed.). (2012). *For attribution – Developing data attribution and citation practices and standards: Summary of an international workshop*. Washington, D.C.: National Academies Press. Retrieved from http://www.nap.edu/openbook.php?record_id=13564.
[10] Ball, A. & Duke, M. (2012). *Data citation and linking*. Edinburgh, UK: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/data-citation-and-linking.
[11] 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
[12] 'Endorse the data citation principles', URL: https://www.force11.org/datacitation/endorsements.

Data citation for authors

The first half of this guide is aimed predominantly at researchers. It discusses the practical business of citing datasets, such as how to construct a data citation and use it in a research paper.

Ways of referencing data

The usual way of referencing the data directly underlying a publication is by means of a *data access statement*. For open data, this statement should say what is available from which repository, and provide a URL, identifier or accession code to help access the data. For restricted data, the statement should indicate the legal or ethical reason for the restriction, and provide a

link to a permanent record explaining the conditions of access.

While a simple statement of this sort fulfils the basic need to reference data, it falls short in several respects:

- if there is a typographical error in the identifier or URL, there is no additional information to locate the data among the repository's holdings;
- authors may be tempted to give the URL of the repository, rather than one specific to the dataset;
- it does not give due credit to the creators of the dataset – an especially important point if these are different from the authors of the publication;
- it does not treat data as a first-class record of research.

All of these issues may be resolved by enhancing the statement with a data citation. As with other citations, this involves providing an in-text pointer to an entry in the reference list.

If the publisher is not willing to accept a data citation, it is sometimes possible to work around this by citing a *data paper* instead. This is a paper that describes the dataset and its collection without drawing any scientific conclusions from it. Such papers may be published in a special section of a regular journal, or in a dedicated *data journal* such as *Earth System Science Data*.[13]

The placement of the data access statement/in-text citation varies between journals. Some, including those published by PLoS and Pensoft, specify the use of a dedicated section, for example 'Data resources' or 'Data access and terms of use'.[14] Others encourage authors to place it at the end of the abstract. Where no specific advice is given, it is usually recommended that authors put the statement in the acknowledgements section; this is because, as with the acknowledgement of the grant, the statement is often a condition of funding and locating the two together simplifies the task of checking compliance.[15] More detailed guidance on data access statements is given by some institutions.[16]

It is also possible to reference data from non-textual outputs, such as other datasets. Indeed, doing so may help satisfy the licensing conditions of the earlier datasets and encourage data sharing by supporting transitive credit models.[17] One straightforward way of doing this is to include with the dataset a table that lists the source datasets and indicates the subset that each one contributed. An early example of this was published as part of the supplementary data of a 2011 paper on microattribution for genetic variation data.[18] Another solution is to see if the repository holding the data will record the information among the metadata it holds for the dataset.

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. [11,19,20,21]

Author. The creator of the dataset.[11,19,20,21]

Publication date. Whichever is the later of: the date the dataset was made available,[11] the date all quality assurance procedures were completed,[19,20] and the date the embargo period (if applicable) expired.[21]

Title. As well as the name of the cited resource itself,[11,21] this may also include the name of a facility[19] and the titles of the top collection and main parent sub-collection (if any) of which the dataset is a part.[20]

[13] *Earth System Science Data*, URL: http://www.earth-syst-sci-data.net/.

[14] 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.

[15] *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.

[16] See, for example, guidance given by the universities of Bath (http://www.bath.ac.uk/research/data/sharing-reuse/data-access-statement.html) and Bristol (http://data.bris.ac.uk/research/using-data/).

[17] Katz, D. S. (2014). Transitive credit as a means to address social and technological concerns stemming from citation and attribution of digital products. *Journal of Open Research Software*, *2*(1), e20. doi:10.5334/jors.be.

[18] 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.

[19] 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. doi:10.2218/ijdc.v6i2.205

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

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

Edition. The level or stage of processing of the data, indicating how raw or refined the dataset is.[19]

Version. A number increased when the data changes, as the result of adding more data points or re-running a derivation process, for example.[21]

Feature name and URI. The name of an ISO 19101:2002[22] 'feature' (e.g. GridSeries, ProfileSeries) and the URI identifying its standard definition, used to pick out a subset of the data.[19]

Resource type. Examples: 'database',[20] 'dataset'.[21]

Publisher. The organisation either hosting the data[21] or performing quality assurance.[19]

Unique numeric fingerprint (UNF). A cryptographic hash of the data, used to ensure no changes have occurred since the citation.[11]

Identifier. An identifier for the data, according to a persistent scheme.[11,19,20,21]

Location. A persistent URL from which the dataset is available. Some identifier schemes provide these via an identifier resolver service.[11,19,20,21]

The most important of these elements – the ones that should be present in any citation – are the author, the title and date, the location, and the publisher. These give due credit, allow the reader to judge the relevance of the dataset, permit access to it, and give reassurances about its quality or persistence, 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 below).

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,[23,24,25,26] while Figure 2 shows the citation formats suggested by three data repositories.

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

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

(Bibliography) 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. and Bell, M. (2011), *Excavations at St Peter's Church, Barton-upon-Humber* [dataset] (York: Archaeology Data Service), doi: 10.5284/1000389

Figure 1: Data citations in common styles

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

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 URLs (PURLs), all of which can be resolved to an

[22] ISO 19101. (2002). *Geographic information – Reference model*. 1st ed. International Organization for Standardization.
[23] *Publication Manual of the American Psychological Association* (6th ed., p. 211). (2010). Washington, DC: American Psychological Association.
[24] *Chicago Manual of Style* (16th ed., p. 764). (2010). Chicago, IL: University of Chicago Press.
[25] 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.
[26] Ritter, R. M. (2002). *Oxford Manual of Style* (p. 551). Oxford, UK: Oxford University Press. Waddingham, A. (Ed.). (2014). *New Hart's rules: The Oxford style guide* (2nd ed., pp. 368-376). Oxford, UK: Oxford University Press.

Internet location. 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.[27] 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.

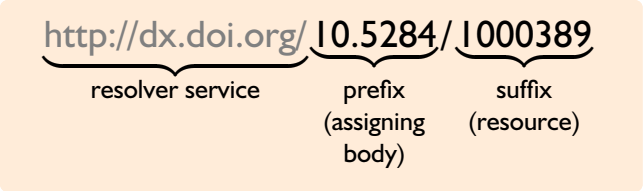

Figure 3: Anatomy of a DOI

While there are several services available that can resolve a DOI to an Internet location,[28] 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. Authors are encouraged to use the URL version of the DOI wherever possible, though some publishers prefer to print the bare DOI and embed the URL form as a hyperlink in digital versions.

Individuals wishing to register a DOI for their dataset would normally do so via their disciplinary data archive, institutional data repository, or a data sharing service such as figshare[29] or Synapse.[30]

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.[31] It has gained support from over 300 organisations, including major academic publishers, and been integrated into numerous research systems. Researchers can associate with their ORCID profiles a list of works to which they have contributed, as well as grants received and their educational and employment history. ORCID profiles can also be linked to identifiers and profiles from other schemes such as Thomson Reuters' ResearcherID,[32] Scopus,[33] Scholar Universe,[34] and RePEc.[35]

The International Standard Name Identifier (ISNI) scheme is an ISO standard for registering 'Public Identities': people, pseudonyms, personas and legal entities involved in the creation or distribution of intellectual property.[36] 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 is to support software tools, they are better placed in machine-readable metadata than written out for human inspection. It is therefore recommended that authors do not attempt to include ORCIDs or similar in their reference lists, but rather ensure they supply their own ORCID to publishers and repositories at the point of making their submission.

Granularity

With print publications, the issue of citing at different levels of granularity is relatively straightforward. The documents listed within a bibliography or reference list 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

[27] DOI System, URL: `http://www.doi.org/`.
[28] Some publishers provide resolvers for their own DOIs, while the Handle resolver `http://hdl.handle.net/` can be used for any DOI.
[29] Figshare, URL: `http://figshare.com/`.
[30] Synapse, URL: `https://www.synapse.org/`.
[31] ORCID, URL: `http://orcid.org/`.
[32] ResearcherID, URL: `http://www.researcherid.com/`.
[33] Scopus, URL: `http://www.scopus.com/`.
[34] Scholar Universe, URL: `http://www.scholaruniverse.com/`.
[35] RePEc Author Service, URL: `http://authors.repec.org/`.
[36] ISO 27729. (2012). *Information and documentation – International standard name identifier (ISNI)*. International Organization for Standardization.

at the point of citation in the text, rather than in the reference list.

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.[37] 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 reference list, to minimise the additional information needed.

Citing unreleased data

If citing a dataset that is not yet released, the rule of thumb is to provide in the reference as much information about it as is already known. At a miniumum, this should include the creator and title of the dataset. If the dataset has not yet been deposited, the date of collection should be included. If the dataset has been deposited but an online record is not yet available, the date can be given as 'in press' and the repository given in the publisher position. Once an online record is available, the full citation can be given. The full details of the status of the dataset – whether deposited, embargoed, restricted or openly available – should be explained in the data access statement.

As with references to manuscripts that have not yet been published, authors should revisit references to unreleased data prior to publication to ensure the information is as up to date as possible.

Citing physical data

There is no difference in principle between how one should cite physical data, such as samples or materials, and digital data. Physical data is often less reproducible or shareable than digital data, but the majority of issues that apply to it also apply to digital data that are too sensitive or voluminous to be transported over the Internet. In practice, the issue most likely to cause confusion is how and whether to provide a URL for the physical data.

If the physical data has an identifier in a scheme with a resolver service, this should be used as the URL. For example, the International Geo Sample Number (IGSN) is associated with a catalogue whose records can be accessed by appending the number to the resolver service URL: `http://www.geosamples.org/profile?igsn=`.

If the physical data has an identifier that cannot be resolved, it should be quoted elsewhere in the reference. The URL, meanwhile, should point to a page explaining how to gain access to the data, if applicable.

Summary for researchers

- If you have generated/collected data to be used as evidence in an academic publication, you should deposit it 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 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.

[37] 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`.

Data citation for repositories

The remainder of this guide is aimed at data repositories. It looks at the underlying infrastructure that supports data citation, and suggests ways in which repositories might participate in and build on existing activity.

Tools and services

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

DataCite DOIs

The task of managing 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.[38] 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.[39] These metadata may be searched through a Web interface[40] or harvested using OAI-PMH.[41]

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.[42] The username and password are not needed for the metadata search or OAI-PMH services.

While best practice has yet to emerge on some matters, certain conventions are already becoming established.

- 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.

[38] DataCite, URL: http://www.datacite.org/.
[39] DataCite Metadata Schema Repository, URL: http://schema.datacite.org/.
[40] DataCite Metadata Search service, URL: http://search.datacite.org/.
[41] DataCite OAI-PMH service, URL: http://oai.datacite.org/.
[42] EZID, URL: http://ezid.cdlib.org/.

Various organisations have shared their experiences of working with DataCite:

- Archaeology Data Service; University of Southampton;[43]
- Australian Antarctic Division; Australian National University; Dryad;[44]
- ForestPlots.net, University of Leeds;[45]
- Griffith University;[46]
- UK Data Archive;[47]
- University of Bristol.[48]

Notification Services

The CLADDIER Project developed a prototype Citation Notification Service for use by digital object repositories, based on the TrackBack protocol.[49] 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.[50] CLADDIER extended the protocol to permit richer metadata to be communicated each way between the citing and cited systems,[51] allow previous TrackBacks to be updated or deleted, and reduce the likelihood of spam

[43] British Library. (2013). *Working with the British Library and DataCite: Institutional case studies*. Retrieved from http://www.bl.uk/aboutus/stratpolprog/digi/datasets/DataCiteCaseStudies_2013.pdf.
[44] Australian National Data Service. (n.d.). Data citation [YouTube playlist]. Retrieved from https://www.youtube.com/playlist?list=PLG25fMbdLRa4peWpeZslW0cLSPYNjcbc1.
[45] British Library. (2015). *Datacite case study: ForestPlots.net at the University of Leeds*. Retrieved from http://www.bl.uk/aboutus/stratpolprog/digi/datasets/ForestPlot_CaseStudy_ForBL.pdf.
[46] Simons, N., Visser, K. & Searle, S. (2013). Growing institutional support for data citation: Results of a partnership between Griffith University and the Australian National Data Service. *D-Lib Magazine*, *19*(11/12). doi:10.1045/november2013-simons.
[47] Jisc. (2012, April). Data identifiers: How to ensure your data is properly cited [Webinar]. Retrieved from https://www.jisc.ac.uk/events/data-identifiers-how-to-ensure-your-data-is-properly-cited-11-apr-2012.
[48] Duke, M. & Gray, S. (2014). *Assigning Digital Object Identifiers to research data at the University of Bristol*. Edinburgh, UK: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/persistent-identifiers.
[49] CLADDIER Project page, URL: http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2005/claddier.
[50] Six Apart. (2007). TrackBack manual. Retrieved 18 October 2011, from http://www.movabletype.org/documentation/trackback_manual.html.
[51] Matthews, B., Portwin, K., Jones, C. & Lawrence, B. (2007, November 30). *Recommendations for data/publication linkage* (CLADDIER Project Report No. 3). STFC. retrieved 20 June 2012, from http://ie-repository.jisc.ac.uk/221/.

TrackBacks.[52]

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.[53] StoreLink was itself followed by the Webtracks Project, which generalised the system to form the Inter-Repository Communication (InteRCom) protocol and extend its usage beyond e-print repositories to STFC's ICAT data catalogue, open electronic notebooks and scientific publishers.[54]

Example

Knowledge Blog[55] was developed as an alternative scholarly publication platform based on WordPress[56] blogging software. 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 and DataCite registration agencies) or a PubMed identifier.[57]

The US-based SHARE initiative is implementing a different notification system that, while not directly related to citation, may assist with setting up links between systems.[58] The SHARE Notify service collects metadata from publishers and repositories about events such as data or pre-prints being deposited, or papers being published. This metadata is indexed in a database and made available through JSON and Atom feeds. If a repository is aware that a manuscript related to a dataset it holds is about to be published, it could monitor the feeds to discover when publication occurs. Similarly, by contributing to the SHARE Notify service, the repository could enable a publisher to discover when the dataset underlying a manuscript has been released.

Citation tracking services

One of the benefits of using formal data citations is that it should make it easier to assemble evidence that a dataset has had impact. As of mid-2015 it is quite hard to do this due to the variety of ways in which datasets are referenced in the literature. Nevertheless there are some services available that index these references.

The Thomson Reuters Data Citation Index was launched in October 2012.[59] It tracks citations and less structured references to data at four levels of granularity: nanopublications (see below), datasets, research studies, and data repositories. It relies not only on access to the full text of publications but also on an index of available datasets, the information for which is drawn from data repositories, data discovery services and DataCite.

Europe PubMed Central routinely text-mines its archive of full text articles for data citations. The resulting information is used by the PLoS Article Level Metrics service,[60] and is also available through the Europe PubMed Central RESTful Web service.[61]

For more information about tracking the impact of datasets, please see the DCC guide 'How to Track the Impact of Research Data with Metrics'.[62]

Network tracking services

The DLI (Data Literature Interlinking) Service is a result of a collaboration between the ICSU-WDS/RDA Data Publishing Services Working Group and the OpenAIRE initiative.[63] The aim of the service is to create a centrally curated graph of links between publications and datasets. Instead of relying on many bilateral arrangements between organisations, the idea is that publishers contribute to the graph links relating to articles they have published, while repositories contribute links relating to datasets they hold. Any

[52] 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.

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

[54] Webtracks Project blog, URL: http://webtracks.jiscinvolve.org/.

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

[56] WordPress, URL: https://wordpress.org/.

[57] KCite WordPress plugin, URL: https://wordpress.org/plugins/kcite/.

[58] SHARE initiative, URL: http://www.share-research.org/.

[59] Thomson Reuters Data Citation Index product page, URL: http://wokinfo.com/products_tools/multidisciplinary/dci/.

[60] Lin, J. & Fenner, M. (2013, December 5). Research findings: Going deeper than the article [Web log post]. Retrieved from PLoS Tech Blog: http://blogs.plos.org/tech/research-findings-going-deeper-than-the-article/.

[61] Europe PubMed Central RESTful Web service, URL: http://europepmc.org/RestfulWebService.

[62] Ball, A. & Duke, M. (2015). *How to track the impact of research data with metrics*. Edinburgh, UK: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/how-guides/track-data-impact-metrics.

[63] DLI Service, URL: http://dliservice.research-infrastructures.eu/.

interested party can then look up a resource in the graph and see which other resources are related to it, perhaps to enhance a record for that resource or perform bibliometric analysis. The graph may be queried either through a graphical Web interface or via an API.

The RMap Project is an initiative undertaken by the Data Conservancy, Portico and IEEE, with funding from the Alfred P. Sloan Foundation.[64] It has very similar aims to the DLI Service and is working closely with it. The scope of RMap is however a little broader, as it also records links with agents (authors, publishers, repositories, etc.) and software. The RMap graph may be queried via an API.

Nanopublications

A nanopublication is a statement and a set of annotations on it, the whole of which is citable in its own right.[65] 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.[66]

Citation Typing Ontology

The Citation Typing Ontology (CiTO) is a formal language for specifying why one resource cites another.[67] It contains several terms particularly relevant for data citation; additional terms can be found in the extension ontology CiTO4Data.[68,69]

- *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 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*.

Example implementations

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 above.

[64] RMap Project, URL: http://rmap-project.info/rmap/.
[65] 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.
[66] Concept Wiki, URL: http://www.conceptwiki.org/.
[67] Shotton, D. (2010). CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics*, *1*(Suppl 1), S6. doi:10.1186/2041-1480-1-S1-S6.
[68] Shotton, D. & Peroni, S. (2011a, March 30). *CiTO, the Citation Typing Ontology*. Version 2.0. Retrieved 26 May 2011, from http://purl.org/spar/cito/.
[69] Shotton, D. & Peroni, S. (2011b, February 25). *CiTO4Data, the Citation Typing Ontology for Data*. Version 1.0. Retrieved 26 May 2011, from http://purl.org/spar/cito4data/.

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.[70] It is the data archive and distribution system for the World Data Center 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.[71]

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.[72] Some of the associated visualisation and conversion tools have been made available as freeware.[73]

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.[74] 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.[75]

Dryad is based on the DSpace digital repository;[76] the Dryad extensions have been released as open source software.[77]

Dataverse

The Dataverse Network is a software application for building data repositories called dataverses.[78] 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.[79]

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

[70] PANGAEA, URL: `http://www.pangaea.de/`.

[71] 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*, 4th–7th May 2008 (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`.

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

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

[74] Dryad, URL: `http://datadryad.org/`.

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

[76] DSpace, URL: `http://www.dspace.org/`.

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

[78] Dataverse Network Project, URL: `http://thedata.org/`.

[79] 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`.

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.[80] Alternatively, institutions can set up their own Dataverse Network using the open source software.[81]

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 need 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,[82] 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.[83]

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 RDFa Lite as in Figure 4.

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

Figure 4: Example of using RDFa Lite to embed a link to a publication within a dataset's Web page[84]

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.[85] 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 Web page for the paper.

The XYZ Project has developed additional tools to support workflows like these.[86] 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.[87]

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

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

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

[83] Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. & Warner, S. (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.

[84] Sporny, M. (Ed.). (2015, March 17). *RDFa Lite 1.1*. 2nd ed. W3C Recommendation. World Wide Web Consortium. Retrieved from http://www.w3.org/TR/rdfa-lite/.

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

```
<mixed-citation>
  <name><surname>Mulvany</surname><given-names>Ian</given-names></name>,
  <data-title>citing-dataset-elements</data-title>.
  <source>FigShare</source>,
  <date-in-citation content-type='pub-date' iso-8601-date='2014-06-30'>
    <day>30</day><month>06</month><year>2014</year></date-in-citation>,
  <pub-id pub-id-type='doi' xlink:href='http://dx.doi.org/10.6084/m9.figshare.1088363'
    assigning-authority='figshare'>10.6084/m9.figshare.1088363</pub-id>.
</mixed-citation>
```

Figure 5: Example of using JATS to encode a data citation.

On the publisher side, the Journal Article Tag Suite (JATS) is a NISO standard for representing journal articles as XML.[88] It is based on the earlier National Library of Medicine (NLM) Archiving and Interchange Tag Suite. Following a proposal from the FORCE11 Data Citation Implementation Group, support for data citations was added to draft version 1.1d2 of JATS.[89] An example using it to mark up a citation of a data record in FigShare is shown in Figure 5.

Granularity

It is the responsibility of the repository to ensure that datasets are made citable and identifiable at an appropriate level of granularity. There are no hard and fast rules for this, as much depends on custom and practice within the discipline. As a rule of thumb, however, it is recommended that repositories assign identifiers at the finest level of granularity at which the data can be said to form an intellectual whole. Examples include a single genome, or the output of a single sensor during a mission.

Where many such datasets are likely to be cited at once, repositories may also wish to assign identifiers for collections. In this case, the collection should also represent an intellectual whole, such as the data collected by a particular study or activity. The hierarchical relationships between the collection and its constituent parts should be recorded in the metadata for the collection and for each part.

Versioning and dynamic data

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,[90] 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, repositories can present versions in three ways: time slices, full snapshots and partial snapshots.

With the *time slice* approach, the citable entity is the set of updates made to a dataset during a particular time period. This would be rather cumbersome for revisions to datasets, but may be appropriate for expanding datasets if researchers are only likely to need one or two of the time slices (e.g. weather data for a given year).

With the *full snapshot* approach, at regular intervals or at the request of a citing author, a snapshot is taken of the entire dataset and made citable. This is a 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

[86] XYZ Project blog, URL: https://projectxyz.wordpress.com/.

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

[88] Journal Article Tag Suite, URL: http://jats.nlm.nih.gov/.

[89] Mietchen, D., McEntyre, J., Beck, J., Maloney, C. & FORCE11 Data Citation Implementation Group. (2015). Adapting JATS to support data citation. *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2015*. Bethesda, MD: National Center for Biotechnology Information. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK280240/.

[90] 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.

is also better for expanding datasets where authors are concerned with the whole time series.

The recommendation of the Research Data Alliance Working Group on Data Citation (RDA WGDC) is however to use *partial snapshots*. This is where the citable entity is the result of a particular query on a dataset run at a certain time. If the query retrieves the full dataset, this approach reduces to the full snapshot approach, but otherwise it has the advantage of making the snapshot more precisely relevant to the research in question.

Note that these discussions only concern how datasets are presented to users as citable resources; they may be stored quite differently by repositories. The RDA WGDC recommends that datasets are versioned and operations on them (i.e. additions, modifications and deletions) are logged and timestamped so that the state of a set at any given time may be recreated. To support partial snapshots, the WGDC recommends the following steps:

1. When a user requests a citable snapshot, the query used is normalised, logged, timestamped, and compared against previous queries.

2. If the query has been used before to create a snapshot, and the result set has not changed in the meantime (as determined by checksums), the previous snapshot identifier/landing page is reused. Otherwise a new one is created.

3. The landing page for the query should contain details of the full dataset, a sample citation including the persistent identifier for the query, and a machine actionable link for retrieving the result set that the query returned at the time it was made.

For more detailed guidance on accomplishing this, see the full text of the recommendations.[91]

[91] Rauber, A., Asmi, A., van Uytvanck, D. & Pröll, S. (2015, June 9). *Data citation of evolving data: Recommendations of the Working Group on Data Citation*. Retrieved from Research Data Alliance website: `https://www.rd-alliance.org/system/files/documents/RDA-DC-Recommendations_150609.pdf`.

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.[92] For more information about registering Archival Resource Keys,[93] 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 on the first version of this guide, and to DCC colleagues for their comments on subsequent versions.

[92] List of DataCite members, `URL: https://www.datacite.org/about-datacite/members`.

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

Further information

Three 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
- **Working Level:** *How to Track the Impact of Research Data with Metrics* (2015) by Alex Ball and Monica Duke

The following may also be of interest:

- Altman, M. & Crosas, M. (2013). The evolution of data citation: From principles to implementation. *IASSIST Quarterly*, *37*(1-4), 62–70. Retrieved from http://www.iassistdata.org/iq/evolution-data-citation-principles-implementation
- Australian National Data Service. (n.d.). Data citation [YouTube playlist]. Retrieved from https://www.youtube.com/playlist?list=PLG25fMbdLRa4peWpeZslW0cLSPYNjcbc1
- Data Citation [Awareness Level Guide]. (2011, May 3). Retrieved 6 June 2011, from http://www.ands.org.au/guides/data-citation-awareness.html
- Economic and Social Research Council. (2012). Data citation: What you need to know. Retrieved from http://www.esrc.ac.uk/_images/Data_citation_booklet_tcm8-21453.pdf
- EZID. (2014, September 25). Conversations about data citation [Webinar recording]. Retrieved from California Digital Library website: http://ezid.cdlib.org/home/outreach
- Lane, M. A. (2008, September 10). *Data citation in the electronic environment*. Global Biodiversity Information Facility. Retrieved 2 September 2011, from http://www.gbif.org/orc/?doc_id=4884
- Mayernik, M. S. (n.d.). *Bridging data lifecycles: Tracking data use via data citations workshop report* (Technical Note No. NCAR/TN-494+PROC). National Center for Atmospheric Research. doi:10.5065/D6PZ56TX
- Mooney, H. & Newton, M. P. (2012). The anatomy of a data citation: Discovery, reuse, and credit. *Journal of Librarianship and Scholarly Communication*, *1*(1). doi:10.7710/2162-3309.1035
- Page, R. (2009, April 20). Semantic publishing: Towards real integration by linking [Blog post]. Retrieved 11 May 2011, from http://iphylo.blogspot.com/2009/04/semantic-publishing-towards-real.html
- Why and How Should I Cite Data? (2009, June 23). Retrieved 8 June 2011, from http://icpsr-support.blogspot.com/2008/10/why-and-how-should-i-cite-data.html
- UK Data Service. (2015, July 13). Citing data. Retrieved from http://ukdataservice.ac.uk/use-data/citing-data
- University of Bristol Research Data Service. (2015, February 1). *Citing your research data*. Retrieved from University of Bristol website: http://data.bris.ac.uk/files/2014/02/Citing-research-data.pdf
- Walton, D., Lowry, R. & Callaghan, S. (2012). Data citation and publication by NERC's Environmental Data Centres. *Ariadne*, *68*. Retrieved from http://www.ariadne.ac.uk/issue68/callaghan-et-al
- Wilkinson, M. (2011a, July 28). So you want to cite your data: The consequences of data citation [Blog post]. Retrieved 16 August 2011, from 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 http://sagecite.knowledgeblog.org/2011/07/28/why-do-we-need-data-citation-take-two/

The Digital Curation Centre (DCC) is a consortium of the Universities of Edinburgh, Glasgow and Bath, and receives funding from Jisc.

Follow the DCC on Twitter: @digitalcuration, #ukdcc

Published: 30 July 2015