J Comput Aided Mol Des (2014) 28:1035–1041
DOI 10.1007/s10822-014-9776-5

DataCite and DOI names for research data

Janna Neumann · Jan Brase

Received: 31 March 2014 / Accepted: 7 July 2014 / Published online: 20 July 2014

Abstract The publication of research data is still not a widespread practice in many disciplines. The lack of acceptance of data as scientific output equal to scientific articles, and the lack of suitable infrastructures for the storage of data make it difficult to publish and cite data independently. The global consortium DataCite was established in 2009 to overcome the challenges of data citation. The aim of the consortium is to establish easy access to data, to increase the acceptance of data publication and to support data archiving. The use of Digital Object Identifiers (DOI) provides an easy method to access and re-use research data. The DOI facilitates the citation of data and therefore increases the availability and acknowledgement of research data.

Keywords Digital Object Identifier · Research data · Data citation · Global consortium

Introduction

> "To predict the outcome of novel chemical reactions, extensive information about the results of previous transformations must become available, notably those with unexpected and unsuccessful results. Perhaps it should go without saying that prediction will be impossible unless the data and information relating to previous transformations is satisfactorily curated" [1].

J. Neumann (✉) · J. Brase
German National Library of Science and Technology (TIB), Welfengarten 1B, 30167 Hannover, Germany
e-mail: Janna.Neumann@tib.uni-hannover.de

J. Brase
e-mail: Jan.Brase@tib.uni-hannover.de

This excerpt from Bird et al. 2013 summarizes the lack of information concerning results from research data. The availability of research data in today's global science is the key to more efficient and time-saving local research. Once data is managed, organized and provided with metadata as well as published and made citable other researchers gain the ability to find and re-use this data. Possibly the re-use of existing research data enables to save time and money, as costly experiments do not have to be repeated. However, in many disciplines the access to research data is still not resolved satisfactorily because of missing local infrastructure for the storage and provision of (digital) research data. The possibility of data publication provides several benefits for researchers. A major benefit is the increase of a researcher's reputation within the scientific community, especially if data publication is recognized as a result of scientific work in the same way as a research article is acknowledged in a journal. Even though this is only one argument why data publication should become more relevant, to achieve a good scientific reputation is still one of the most significant promoters for a scientific career in most disciplines. Moreover, a better visibility of research data can motivate scientists to conduct new research and avoid data duplication.

In this short paper we explain what a Digital Object Identifier (DOI) is and how it can be used to refer to research data. Furthermore we give a short introduction on DataCite and its services and benefits for research data citation.

Definitions

The term research data used in this article encompasses everything that can be the basis of scientific research. Next

to experimental data files and all kinds of graphic representations this includes audiovisual material and other non-textual material. The OAIS Reference Model defines research data as "a reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing" [2].

In this context research data management or data curation is defined as "the active management and appraisal of data over the lifecycle of scholarly and scientific interest" [3] by the Digital Curation Centre (DCC) in the UK [4].

To ensure the ability of referencing research data, the DOI is often used as one of several persistent identifiers. The DOI System is managed and administrated by the International DOI Foundation (IDF) that "provides a technical and social infrastructure for the registration and use of persistent interoperable identifiers for use on digital networks". [5] DOI is a registered trademark of the IDF.

DOI names for research data

If any kind of digital scientific content should be citable, the content as well as the citation needs to be persistent. The use of uniform resource locators (URLs) does not seem to be a good solution because of the numerous times pages or objects cannot be found anymore (*e. g. "Error 404 - Page not found"*) due to name or server changes. URLs refer to a specific location in the World Wide Web. Once this location has changed, the object is lost. By using persistent identifiers such as DOIs , the problem mentioned above can be solved as the DOI is a "digital identifier of an object" [6]. A DOI consists of a unique character string that identifies an entity in a digital environment—in other words it identifies the object itself and not the place where it is located. If the object is moved to another location (meaning the URL has changed) the only requirement is to update the URL in the underlying database. This ensures that the DOI persistently resolves to the location of the object [7].

The DOI system is based on the Handle System that is developed and maintained by the Corporation for National Research Initiatives (CNRI) [8]. The DOI system was started in 1998 by the IDF and was standardized as ISO 26324 in 2012 [6].

DOI names are traditionally used in publications of scientific findings. They are used as the core technology to refer to the electronic version of an article in a journal. The use of DOI in citations enables cross-linking between published articles and therefore provides better opportunities to share and access scientific findings across the internet [9]. Datasets are usually part of a traditional scientific publication in a scientific article and therefore cannot be cited independently [10]. As data is becoming more relevant for re-use and verification of research, the need for citing data has emerged as well. Research data as a citable contribution can not only account for one's scientific reputation but also avoid duplication of research as well as verify earlier results [11].

Therefore the use of DOI names for datasets enables the scientific community to move beyond journals and books and make more scientific and technical content visible, available and searchable. A sufficient description of datasets using standardized metadata enables the correct citation of this data. The citation of data allows "to detect, locate, obtain, and understand the data from prior research […]" [10] and to re-use data for new research.

Figure 1 displays an example for the publication of research data that is part of a scientific article. The data can be downloaded, re-used and cited separately by using the DataCite DOI.

Especially in the applied sciences where research progress often depends on the ability to proof experimental results and their "experimental reproducibility", the collection and the sharing of data is substantial [12].

However, the data sharing process requires an effective storage, description and organization of the collected data in order to enable fellow scientists and institutions to access, understand as well as re-use it [13, 14]. In addition to a standardized metadata description of research data, the use of DOI names also premises a stable infrastructure for data storage. The content has to be stored and made available persistently by an institution that commits to data management. In its "Principles and Guidelines for Access to Research Data" the Organization for Economic Co-operation and Development (OECD) lists several steps that should be considered when institutions are starting to deal with research data management [15]. The steps refer to legal aspects as well as quality control and sustainability of research data and include detailed supporting information. Next to the possibility to store research data in an institutional repository, an increasing number of discipline-specific data repositories (partly operated by data centers) have become available in recent years [16]. Additional information about these data infrastructures can be found e.g. via the online platform Databib [17] that lists over 600 different repositories [16] or via the online platform re3data [18]. Databib was initially developed by Purdue University in collaboration with Penn State University. It was funded by the Institute of Museum and Library Services (IMLS) in the United States and has been online since April 2012. Its editorial board identifies, catalogs, and curates a searchable index of research data repositories. The re3data - Registry of Research Data Repositories was developed by the Library and Information Services department (LIS) of the GFZ German Research Centre for Geosciences, the Computer and Media Service at the

Fig. 1 DataCite DOI 10.4125/pd0054th for research data attached to scientific article

Humboldt-Universität zu Berlin and the KIT Library at the Karlsruhe Institute of Technology (KIT). It is indexing data repositories since 2012 and is funded by the German Research Foundation DFG from 2012 to 2015 [18]. Recently Databib and re3data—Registry of Research Data Repositories announced their plan to merge the two projects into one service. The aim of this junction is to reduce duplication of the services. The infrastructure will be managed by DataCite by the end of 2015 [19].

Another recent development is the emerging of *"data journals"*, which sometimes are *"spin-offs"* from already existing journals. These journals allow publication of research data separately from the scientific article. In contrast to data that is part of the supplemented material included in a scientific article, data journals, such as *"Scientific Data"* launched by the Nature Publishing Group facilitate the discovery, re-use and citation of research data also because the citable research data publication is deposited and shared in accredited research data repositories [16].

DataCite

The international consortium DataCite was founded in December 2009 in London as a global not-for-profit organization. The managing office of DataCite is located at the German National Library of Science and Technology (TIB) [20] in Hannover, Germany.

The aim of DataCite is to:

- establish easier access to research data on the Internet,
- increase acceptance of research data as legitimate, citable contributions to the scholarly record, and

- support data archiving that will permit results to be verified and re-purposed for future study [21].

The global consortium acts as a global DOI registration agency for scientific content and is carried by its local member institutions. DataCite members work in close co-operation with the data centers responsible for the storage and accessibility of research data.

Before DataCite became an official DOI Registration Agency in 2009, TIB has been the world's first DOI registration agency for research data since 2005. TIB is the German National Library for all areas of engineering as well as architecture, chemistry, information technology, mathematics and physics and ranks as one of the world's largest specialised libraries [22].

Towards the end of 2008 the need for a better availability of research data and respective publication processes increased amongst the scientific community. Thus several information centers and libraries decided to establish a global DOI registration agency for scientific content. This new registration agency was meant to extend "the DOI model of TIB to a model of local agencies" following the approach of the publishers that use the central DOI registration infrastructure of CrossRef [23]. In March 2009, the institutions (among them TIB, British Library, Technical Information Center of Denmark, TU Delft, Canada Institute for Scientific and Technical Information (CISTI), California Digital Library and Purdue University) interested in the establishment of a global DOI registration agency signed a Memorandum of Understanding during the meeting of the International Council for Scientific and Technical Information (ICSTI) in Paris for the constitution of a partnership [23].

After the DataCite consortium was established in London with seven members, the association became an official member of the IDF and replaced TIB in its role of a DOI registration agency. Up to now, DataCite works together with 22 members and nine affiliated members. The members have registered over 3.3 million DOI names for datasets, gray literature and other non-textual materials. DataCite's definition of *"dataset"* is everything that can motivate new research. Therefore in this context *"dataset"* can be research data, images, videos, software etc. The assignment of DOI names to gray literature comes from the fact, that this literature is not part of the traditional scholarly workflow. However some disciplines consider even *"text"* as *"data"*. Hence, the mapping of any of these objects to DataCite's metadata field "resource type" is not as simple as it might seem. Apart form the fact that this is not a mandatory metadata field, the above mentioned is one of the reasons why many objects do not have defined the resource type. Therefore, the majority of the registered numbers of DOI names are assigned to research data and other non-textual materials. The annual growth of DOI names registered by DataCite has increased approximately from 83,000 to 700,000 per year in the past 5 years. In the time between 2009 and 2014, the technical infrastructure for the registration of DOI names as well as several additional services were established and expanded. The core element of the service infrastructure consists of the DataCite Metadata Store (MDS) [24] which archives the metadata of all registered objects in a database. To register an object, the metadata (in Extensible Markup Language (XML) format) must be uploaded (via browser interface or application programming interface (API)) to the MDS using the DataCite Metadata Schema [25]. The schema contains five mandatory fields, seven recommended and six optional fields. To enhance the prospects that metadata and therefore datasets will be found, cited and linked to original research, the provision of recommended metadata fields next to the mandatory set of properties is strongly advised by DataCite [26].

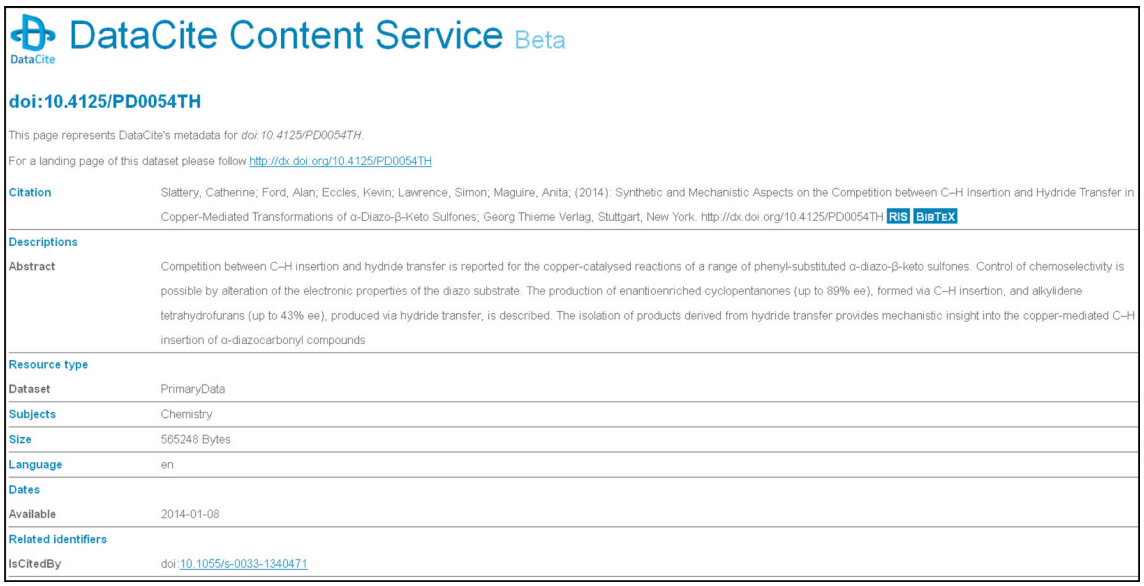
In addition to the MDS, DataCite provides a search tool (DataCite Search [27]) which indexes metadata from MDS and contains metadata for DOI names that were registered through DataCite.

Figure 2 illustrates the indexed metadata for DOI 10.4125/PD0054TH in DataCite and the link to the corresponding landing page of the dataset.

All indexed metadata are made available through DataCite's Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) service [28] which provides an API for metadata harvesting.

Furthermore, DataCite offers a detailed statistic portal [29], where stats and numbers of registered and resolved DOI names are displayed. In co-operation with CrossRef, a content negotiation service was also established. The content negotiation service [30] enables the user to persistently resolve all DOI names directly to their metadata in XML or RDF format [31]. Another implemented tool from this co-operation is the Citation formatter [32] which provides the citation of DataCite and CrossRef DOI names in various formatting styles.

The services and tools provided by DataCite described above aim to improve the scholarly infrastructure around data and other non-textual information. To emphasize this goal, in 2012 "DataCite and the International Association of Scientific, Technical and Medical Publishers (STM association [33]) signed a joined statement to encourage publishers and data centers to link articles and underlying data" [10]. Furthermore in 2012 Thomson Reuters started to build up their Data Citation Index that provides information about research data from various research data repositories [34]. One of the issues the Data Citation index is concerned with is the measurement of data citation. By indexing research data from data repositories the influence


DataCite Content Service Beta

DataCite

doi:10.4125/PD0054TH

This page represents DataCite's metadata for *doi:10.4125/PD0054TH*.

For a landing page of this dataset please follow http://dx.doi.org/10.4125/PD0054TH

Citation	Slattery, Catherine; Ford, Alan; Eccles, Kevin; Lawrence, Simon; Maguire, Anita; (2014): Synthetic and Mechanistic Aspects on the Competition between C–H Insertion and Hydride Transfer in Copper-Mediated Transformations of α-Diazo-β-Keto Sulfones; Georg Thieme Verlag, Stuttgart, New York. http://dx.doi.org/10.4125/PD0054TH RIS BibTeX
Descriptions	
Abstract	Competition between C–H insertion and hydride transfer is reported for the copper-catalysed reactions of a range of phenyl-substituted α-diazo-β-keto sulfones. Control of chemoselectivity is possible by alteration of the electronic properties of the diazo substrate. The production of enantioenriched cyclopentanones (up to 89% ee), formed via C–H insertion, and alkylidene tetrahydrofurans (up to 43% ee), produced via hydride transfer, is described. The isolation of products derived from hydride transfer provides mechanistic insight into the copper-mediated C–H insertion of α-diazocarbonyl compounds
Resource type	
Dataset	PrimaryData
Subjects	Chemistry
Size	565248 Bytes
Language	en
Dates	
Available	2014-01-08
Related identifiers	
IsCitedBy	doi:10.1055/s-0033-1340471

Fig. 2 DataCite Search Content Service for DOI: 10.4125/PD0054TH

of data publication is measured through the times cited [35]. DataCite and Thomson Reuters are in negotiation for the integration of quality research data from various accredited data repositories into the Data Citation index that are assigned with DataCite DOI names. This enables users to track the citation impact of research data assigned with DataCite DOI names.

In the next section we introduce the current DataCite members and affiliated members and refer to their respective duties and rights within the global consortium.

DataCite members

DataCite started with seven founding members in 2009. By 2014, the consortium consists of 22 members in 16 countries. This emphasizes the need of local representatives working together with a globally organized framework such as DataCite.

In Germany, four DataCite members are located: the German National Library of Science and Technology (TIB), the Leibniz Information Centre for Life Sciences (ZB MED), the Leibniz Information Centre for Economics (ZBW) and the Leibniz Institute for the Social Sciences (GESIS). These institutes are working mostly with discipline-specific data centers within the scope of their scientific area. Other European members include: The Library of the ETH Zürich in Switzerland, the Library of TU Delft, from the Netherlands, the Institut de l'Information Scientifique et Technique (INIST) from France, the Technical Information Center of Denmark, the British Library, the Swedish National Data Service (SND), the Conferenza dei Rettori delle Università Italiane (CRUI) from Italy, the Library and Information Centre of the Hungarian Academy of Sciences (MTA KIK), and the University of Tartu from Estonia. The representatives of North America are the California Digital Library, the Office of Scientific and Technical Information (OSTI), the Purdue University and the Canada Institute for Scientific and Technical Information (CISTI). DataCite members from Australia and Asia are the Australian National Data Service (ANDS), the National Research Council of Thailand (NRCT), and the Japan Link Center (JaLC) respectively. DataCite member from Africa is the South African Environmental Observation Network (SAEON) and last but not least the European Organisation for Nuclear Research (CERN) is an international member of DataCite.

The DataCite membership is open to all non-profit organizations who wish use the Registration Agency of DataCite for the allocation of DOI names for research data. Every member should be working with data centers for issuing DOI names. A member is allowed to take part in the working groups, has full voting rights on all decisions, and can register unlimited DOI names for themselves and their clients [36].

Next to the full members there are nine affiliated members. Affiliated members have an advisory function, are allowed to take part in the working groups and to attend the general assembly. Furthermore, affiliated members are interested in co-operation with DataCite on a superior level. In contrast to full members, associated members have restricted voting rights and do not act as allocators for DOI names with DataCite [36]. The affiliated members are: The Beijing Genomics Institute (BGI) in China, the

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) in Germany, the internationally operating ICSU World Data System (ICSU-WDS), the Korea Institute of Science and Technology Information (KISTI) in the Republic of Korea, the Digital Curation Center (DCC) in the United Kingdom, the Interuniversity Consortium for Political and Social Research (ICPSR), Microsoft Research, Harvard University Library and the Institute of Electrical and Electronics Engineers (IEEE) as the affiliated representatives of North America.

Conclusion

To ensure a maximum transparency and re-usability of data, a researcher first of all needs to decide what kind of research data should be shared with the scientific community. After that, questions about storage, organization, metadata description and citability have to be answered properly to guarantee re-use and citation of the shared data. In this second step, scientific institutions, data centers, libraries and further infrastructures such as DataCite assist in the publication process of research data. The assignment of DOI names for publicly available data sets is carried out by local institutions as a service for a local scientific community. Therefore if researchers want to share and publicize data using the DOI system, they should get in touch with their local institution or with a discipline-specific data repository that assigns DOI names for (open) research data.

References

1. Bird CL, Willoughby C, Coles SJ, Frey JG (2013) Data curation issues in the chemical sciences. Inf Stand Q 25(2):4–12. doi:10.3789/isqv25no3.2013.02
2. Reference Model for an Open Archival Information System (OAIS) (2012) Recommendation for Space Data System Practices. CCSDS 650.0-M-2. http://public.ccsds.org/publications/archive/650x0m2.pdf. Accessed 30 June 2014
3. Digital Curation Centre. JISC programmes. http://www.jisc.ac.uk/whatwedo/programmes/preservation/dcc.aspx. Accessed 30 June 2014
4. Digital Curation Centre (2014) JISC, United Kingdom. http://www.dcc.ac.uk. Accessed 30 June 2014
5. International DOI Foundation (2014) Digital Object Identifier System. http://www.doi.org/. Accessed 30 June 2014
6. International DOI Foundation (2012) The DOI system concept. In: DOI Handbook. doi: 10.1000/182
7. Paskin N (2002) Digital Object Identifiers. Proceedings of ICSTI Seminar: Digital Preservation of the Record of Science. ICSTI 14-15. Feb 2002. IOS Press, Oxford. http://www.doi.org/topics/020210_CSTI.pdf. Accessed 30 June 2014
8. Corporation for National Research Initiatives (CNRI). http://www.cnri.reston.va.us/. Accessed 30 June 2014
9. Paskin N (2004) Digital Object Identifiers for scientific data. In: 19th International CODATA Conference, Berlin, Germany. http://www.codata.org/04conf/abstracts/PublCitSciData/index.html. Accessed 30 June 2014
10. Brase J, Socha Y, Callaghan S, Borgman CL, Uhlir PF, Carroll B (2014) Data Citation—Principles and Practice. In: Ray JM (ed) Research data management—practical strategies for information professionals. Purdue University Press, West Lafayette, Indiana
11. Wallis JC, Rolando E, Borgman CL (2013) If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLoS One 8(7):e67332. doi:10.1371/journal.pone.0067332
12. Brooks BJ, Thorn AL, Smith M, Matthews P, Chen S, O'Steen B, Adams SE, Townsend JA, Murray-Rust P (2011) Ami—The chemist's amanuensis. J Cheminform 3:45. doi:10.1186/1758-2946-3-45
13. Costas R, Meijer I, Zahedi Z, Wouters P (2013) The Value of Research Data – Metrics for datasets from cultural and technical point of view. A Knowledge exchange Report, available from www.knowledge-exchange.info/datametrics. Accessed 30 June 2014
14. Dallmeier-Tiessen S, Darby R, Gitmans K, Lambert S, Suhonen J, Wilson M (2012) Compilation of results on drivers and barriers and new opportunities. Zenodo. doi:10.5281/zenodo.8309
15. Organisation for Economic Co-operation and Development (2007) OECD Principles and Guidelines for Access to Research Data from Public Funding. OECD Publishing. doi:10.1787/9789264034020-en-fr
16. Warr WA (2014) Data sharing matters. J Comput Aided Mol Des 28(1):1–4. doi:10.1007/s10822-013-9705-z
17. Databib (2014). http://databib.org/. Accessed 30 June 2014
18. re3data—Research Data Repository (2014). www.re3data.org. Accessed 30 June 2014
19. re3data—Research Data Repository (2014) DataCite, re3data.org, and Databib Announce Collaboration. http://www.re3data.org/2014/03/datacite-re3data-org-databib-collaboration/. Accessed 30 June 2014
20. German National Library of Science and Technology (2014). www.tib-hannover.de/en. Accessed 30 June 2014
21. DataCite (2014) What is DataCite. http://www.datacite.org/whatisdatacite. Accessed 30 June 2014
22. Brase J (2013) DataCite and linked data. Ital J Libr Inf Sci 4(1):365–373. doi:10.4403/jlis.it-5493
23. Brase J, Farquhar A, Gastl A, Gruttemeier H, Heijne M, Heller A, Piguet A, Rombouts J, Sandfaer M, Sens I (2009) Approach for a joint global registration agency for research data. Inf Serv Use 29(1):13–27. doi:10.3233/ISU-2009-0595
24. DataCite Metadata Store (MDS) (2014). https://mds.datacite.org/?lang=en. Accessed 30 June 2014
25. DataCite Metadata Schema Repository (2014). http://schema.datacite.org/. Accessed 30 June 2014
26. DataCite (2013) DataCite Metadata Schema for the Publication and Citation of Research Data. Version 3.0. doi: 10.5438/0008
27. DataCite Search (2014). http://search.datacite.org. Accessed 30 June 2014
28. DataCite's Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) service (2014). http://oai.datacite.org. Accessed 30 June 2014
29. DataCite Statistic Portal (2014). http://stats.datacite.org. Accessed 30 June 2014
30. DataCite Content Negotiation Service (2014). http://www.crosscite.org/cn. Accessed 30 June 2014
31. Brase J (2014) Making data citeable: DataCite. In: Bartling S, Friesike S (eds.) Opening Science. http://book.openingscience.org/. Accessed 20 June 2014. DOI: 10.1007/978-3-319-00026-8

32. DataCite and CrossRef Citation Formatter service (2014). http://crosscite.org/citeproc. Accessed 30 June 2014
33. International Association of Scientific, Technical and Medical Publishers (STM Association) (2014). http://www.stm-assoc.org. Accessed June 30 2014
34. Thomson Reuters Data Citation Index (2014). http://thomsonreuters.com/data-citation-index/. Accessed 30 June 2014
35. Thomson Reuters (2012) Collaborative Science - Solving the issues of discovery, attribution and measurement in data sharing. White Paper Report. http://thomsonreuters.com/products/ip-science/04_037/collaborative-science-essay.pdf. Accessed June 30 2014
36. DataCite Membership (2014). http://www.datacite.org/Membership. Accessed 30 June 2014