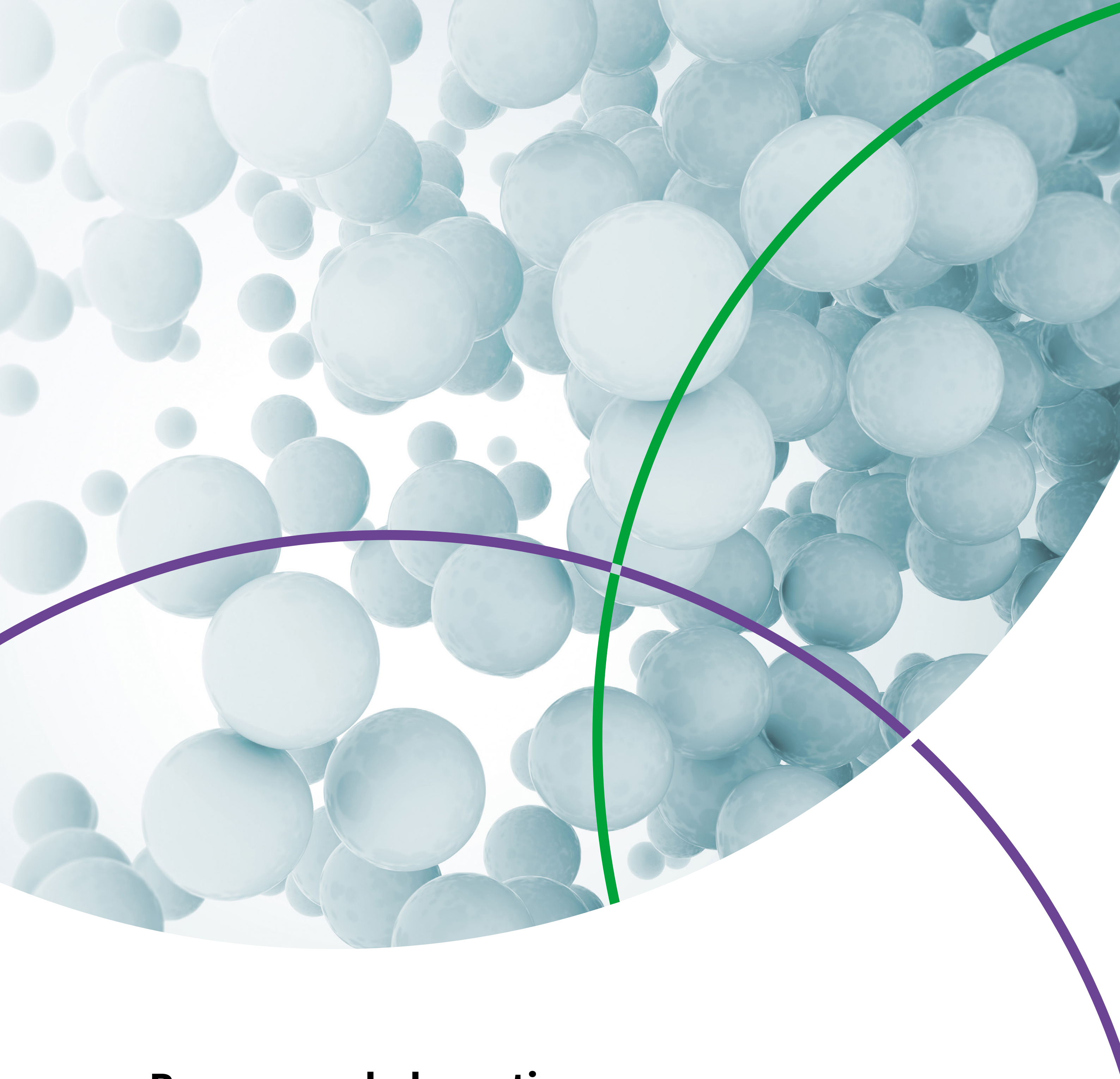

Recommended practices to promote scholarly data citation and tracking

The role of the Data Citation Index

Web of Science
Trust the difference

Clarivate
Analytics

“The Data Citation Index...aims to provide a clearer picture of the full impact of research output, as well as to act as a significant tool for data attribution and discovery.”

Introduction

The history of scholarly advancement is closely linked to data re-use. In the spheres of science, social science, and arts and literature, the work and ideas of early scientists and scholars have lead to new and important discoveries in the eras that followed. While in times past, the passing on of scholarly data might have consisted of an inherited laboratory notebook or astronomical observations, today the preservation and dissemination of data increasingly takes place in the digital realm. As the volume of available scholarly data continues to increase at an exponential rate, scholarly societies and academic, private, and government entities look for new ways to disseminate and interpret this vast reservoir of information[1]. Meanwhile, the variety of data used by different disciplines presents unique challenges as entities look to devise standards that accommodate the needs of particular stakeholder groups, as well as the data needs and conventions of scholars in specialized and established areas of study.

Concurrent with these developments has been an increased interest in methods to assess the full impact of scholarly research, including traditional published research products, such as journal publications, as well as nontraditional products such as datasets and software. In this context, the practice of data citation has gained widespread attention in the academic community as a solution to issues of discovery and attribution for non-traditional scholarly output.

Why Cite Data?

Formal data citation has many benefits for the scientific and scholarly community. While open data has been recommended as a means to better instigate scientific discoveries and ensure reproducible results, there have been few demonstrable rewards for data-gathering institutions and individuals to fund programs and facilities for long-term data preservation and access. In many cases scholars and organizations must put into practice requirements as described by governing bodies with an interest in open data. Formal citation of data allows for these research stakeholders to receive proper credit for their work. Researchers may also gain information regarding data reuse, including in cases where data is not necessarily deposited in conjunction with the publication of a journal article, such as with publicly funded research organizations. Also, new metrics on scholarly output may provide benefits for funding and tenure considerations. These desirable outcomes have led groups such as FORCE11 to develop principles of data citation that advocate data objects as unique citable entities[2].

A New Data Tool

The Data Citation Index was launched in 2012 by *Clarivate Analytics* as a part of *Web of Science* suite of resources. In this index, descriptive records are created for data objects and linked to literature articles in the *Web of Science*. As data citation practices increase, the resource aims to provide a clearer picture of the full impact of research output, as well as to act as a significant tool for data attribution and discovery.

The resource has been developed with attention to the data and metadata needs of various scholarly disciplines, as well as the requirements of publishers and funders in these areas. *Clarivate* has also entered into partnerships that promote the shared mission of increasing the acceptance of research data as citable contributions to the scholarly record. Through collaborations with data providers and organizations, the Data Citation Index looks to support stakeholders at every step in the data lifecycle, including researchers, data repositories/publishers, and administrators. Such support is enabled by best practices discussed here as they relate to these entities in the context of creation, deposition, and curation of metadata necessary for tracking data citation in the included data sources.

Why Cite Data?

- Enables research conclusions to be verified and validated
- Makes reproducibility of premises and results possible
- Exposes data findings and their value to a wider audience
- Ensures a mechanism for receiving credit for scholarly work and an opportunity for tracking/translating such attribution into rewards

Data provider collaborations

In order to better address the needs of the data community, *Clarivate* has partnered with individual data repositories as well as large-scale providers of data and metadata. These partnerships enable the creation of bibliographic records for content in the Data Citation Index.

Clarivate is forging a number of partnerships with individual repositories and databases to provide metadata for the creation of bibliographic records for data in the Data Citation Index. Descriptive, structural, and administrative metadata[3] are obtained using a variety of harvesting protocols (including the OAI-PMH XML-based standard exchange protocol). Through close liaison/content negotiation with its various data partners, the Data Citation Index builds upon the various common metadata standards employed to provide a cross-disciplinary data resource. *Clarivate* can provide support to create the necessary metadata via a simplified, discipline-agnostic XML schema.

Ways to improve citing data

Data Citation Examples: Recommended	Data Citation Examples: Not Recommended
Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. PANGAEA. http://dx.doi.org/10.1594/PANGAEA.726855	Irino & Tada (2009). Chemical and mineral compositions of sediments from ODP Site 127-797. Published by PANGAEA [www.pangaea.de]
Elliott, Joshua (2013): Simulated county- and state-level maize yields, 1979-2012. Version 1. Figshare. http://dx.doi.org/10.6084/m9.figshare.501263	Elliott's Maize Yield Data (2013). Data accessed from Figshare [June 15, 2015]
Uniprot Consortium (2014): P0DKE6.Uniprot Knowledgebase. http://www.uniprot.org/uniprot/P0DKE6	Uniprot Database. http://www.uniprot.org

> *When selecting a data repository for inclusion in a data management plan, authors may look to the journals in their discipline; in some cases, journals provide recommendations or require that data be deposited in a specific repository or in one of a list of recognized data centers.*

Clarivate has sought to expand the coverage of the Data Citation Index by partnering with two major data services: DataCite and the Australian National Data Service (ANDS). These organizations aim to encourage the dissemination, discoverability, and citability of research datasets from a diverse array of data repositories and data publishers through the provision of aggregated search and discovery facilities.

DataCite (https://www.datacite.org) is a leading global nonprofit organization dedicated to enabling people to find, share, use, and cite data. Formed in 2009, the aim is to provide reliable, easy-to-use persistent data identification services to their partners underpinned by the DOI system (Digital Object Identifier – http://www.doi.org). DataCite engages stakeholders, including researchers, scholars, data centers, libraries, publishers, and funders through advocacy, guidance, and services. *Clarivate*'s own recommendation on how to cite a data resource is aligned with DataCite recommendations.

The Australian National Data Service (ANDS) was established in 2008 and aims to enable Australia's research data to become a national strategic resource supporting better, more efficient research and improved policy input. ANDS is partnering with research institutions throughout Australia to help them manage their data more effectively. Their aggregated search and discovery facilities allow researchers on a global level to find and reuse Australian data as part of their ongoing work. Similar to DataCite, ANDS encourages the use of DOIs as persistent identifiers for data objects.

> *“Upon data submission, it is important for authors to consider future citation of the data, as well as creating enriched metadata that increase the discoverability of the data set.*

Clarivate has reached an agreement with both DataCite and ANDS to include metadata for their partner repositories in the Data Citation Index[4],[5], with editorial staff appraising DataCite and ANDS metadata to verify that they are within scope for inclusion in the resource. Through partnerships of this kind, metadata requirements from a diverse array of sources can be normalized across various platforms and data types to encourage the employment of a standard lexicon to address researcher, data provider, and funder requirements.

Creation and deposition of data and metadata

Publicly available research data has been associated with an increased rate of citation for data authors[6]. Scholarly publishers, funding bodies, and government entities increasingly require research authors to make their data publicly available, particularly through repository deposition. Organizations such as the National Science Foundation (NSF) require a discussion of future data management upon submission of grant proposals for project funding[7]. Broad policy requirements put forward by these groups may lack specificity; however the author may often refer to discipline- and data-type specific guidelines for sharing of research results.

Why Deposit Data?

When selecting a data repository for inclusion in a data management plan, authors may look to the journals in their discipline; in some cases, journals provide recommendations or require that data be deposited in a specific repository or in one of a list of recognized data centers. Repositories suggested by journals or publishers are often specialized for the subject area of the publication, as well as for the data and metadata requirements of scholars and scientists in that discipline. It is not uncommon for publishers and funders to recommend or require that the data repositories used be public and/or make the data freely available. Exceptions are made where privacy concerns exist, such as in cases of health data with information about human subjects, data referencing sensitive sites of scientific interest (habitats for endangered species, etc), and commercially sensitive data. Recommended data centers are usually well established, either in their own discipline or as a multidisciplinary data resource.

If the author is affiliated with an academic institution, that institution may have a data repository of its own, often through the institutional library; researchers may wish to consult with data librarians at their institution to discuss the size and format of data sets which are accommodated. Where little guidance for data deposition exists, or where no discipline-specific repository exists or meets criteria, authors are encouraged to explore the deposition of their research results in multidisciplinary repositories that meet their criteria for curation of data and metadata.

Clarivate offers a searchable list of data repositories and sources from across the world and across disciplines which have so far been selected for coverage by the Data Citation Index.

Upon data submission, it is important for authors to consider future citation of the data, as well as creating enriched metadata which increase the discoverability of the data set. Metadata elements such as keywords and discipline-specific indexing terms provide avenues for other researchers to discover and re-us e the data.

Author affiliations and grant and funding agency details provide important information regarding funded research output. This information is highly desirable; however, certain metadata elements are an absolute requirement for the accurate, formal data citation advocated by the Data Citation Index.

> *“Created data sets should cite previous data and traditional literature items (journal articles, books, etc.) where appropriate, and these publications should include data citations in reference sections or other specialist data sections of the paper...*

Elements of a Data Citation

Required elements follow the basic, discipline-agnostic data citation guidelines put forward by DataCite (https://www.datacite.org/services/cite-your-data.html). Metadata elements needed for data citation include:

Author/Creator	Individuals or organizations that created or contributed to the data set; this metadata element is vital to guarantee attribution and credit for data contributor, and to provide metrics for their nontraditional scholarly output
Year	The year of “publication” of the data; when it is made publicly available, such as through deposition in a repository
Title	The title of the data object, which may differ from the title of the parent research paper/project
Publisher	The data repository that houses the data and/or the governing organization responsible for publishing, (i.e., making available) the data
Version	Dynamic data sets or those where new editions may be issued (such as with error corrections or new values) must employ proper version control to guarantee accuracy and uniqueness in data citation
Permanent Identifier	A unique and persistent identifier should be assigned; for example, a Digital Object Identifier (DOI); in Data Citation Index citations, this bibliographic element may take the form of a unique URL, databank accession number, or other permanent identifier such as Handle (hdl) (http://www.handle.net/)

A number of data community organizations, including the Research Data Alliance (RDA), DataCite, and *Clarivate*, encourage authors to practice formal data citation in their work. In the absence of universal standards and guidance, each record in the Data Citation Index includes a recommended formal data citation to use for that data object. Created data sets should cite previous data and traditional literature items (journal articles, books, etc.) where appropriate, and these publications should include data citations in reference sections or other specialist data sections of the paper, as designated by the literature publisher.

The submission process should encompass metadata creation to include careful consideration of proper data attribution, where contributing authors and organizations are given credit through author and repository/publisher attribution, or through citation of contributing data and their individual or institutional authors. Elevation of data to an equal footing with citable publications encourages increased citation of researcher output. Providing data to an established data repository may demonstrate agreement with funder and publisher requirements for data access and preservation[8]. Visible output of previously funded research projects provides evidence to funders of positive outcome of funding, helpful in future grant applications, while evidence of data re-use through citation tracking will validate return on funding investment. Discoverability of data through best practices in data deposition and metadata provision increases the likelihood of re-use of the data, ability to reproduce the research, and hence onward data citation and credit for data set authors[9,10], which is gaining increased use in tenure and career decisions in research institutions.

The cross-disciplinary search capabilities of the Data Citation Index enable greater future re-use of research data and new research discoveries through synthesis of data sets from different research areas.

Data and metadata dissemination and curation

A number of standards for the accreditation of data repositories have recently been proposed or have come into use. These include The European Framework for Audit and Certification of Digital Repositories[11] and The ICSU World Data System (WDS) Criteria for Membership and Certification[12]. In a recent project, the Peer REview for Publication and Accreditation of Research Data in the Earth science (PREPARDE) group created guidelines for repository accreditation in the context of publishing and peer review of data papers submitted to data-specific journals13. The various guidelines put forward share certain common elements, yet may differ where their approach to repository accreditation is dependent upon the requirements and interests of certain stakeholders.

Where PREPARDE has created guidelines for accreditation with a view to repositories as data publication partners in the context of traditional journal publications and data journals, our recommendations here reinforce best practices toward straightforward and unambiguous data citation and discovery. These practices include the use of unique and persistent identifiers, clear attribution, rich metadata, and metadata curation, in addition to sufficient funding and other considerations that show evidence of a commitment to future data preservation and hence continued citation to an extant record or object. Citations and mentions of data repositories and their constituent data objects in published literature provide further evidence that the data are valuable to the academic community and other users of *Web of Science*.

Researcher/Data Author Best Practices

- Regard data equally with other citable research output such as journal publications
- Deposit data in an established data repository committed to long-term preservation and use of permanent IDs for data
- Contribute mandatory metadata with deposited data: authors, year, title, publisher, version
- Contribute further metadata to advance discovery: abstract, author affiliations, funding information, keywords, and data-specific information such as data type and methodology
- Practice detailed, formal data citation in data and publications; cite dataset permanent IDs

> “*Repositories … should provide or require sufficient metadata for deposited data to create a formal citation to an identifiable data object with a unique access point.*

The repositories included in the Data Citation Index should provide or require sufficient metadata for deposited data to create a formal citation to an identifiable data object with a unique access point. Where dynamic data are concerned, there should be clear practice in place to identify the data used through dates or versioning to assist reproducibility of results. Authors should be clearly defined, and the repository should be committed to providing these individuals and organizations with proper credit; to this end, we recommend the inclusion of the data author's institution as well as any funding information, such as funding organization or grant number. Metadata should be curated by the repository or publisher, with quality checks in place to determine whether required elements have been correctly and consistently supplied by submitting authors.

Further, the repository or publisher needs to have a system in place to issue permanent, unique identifiers to data sets to enable identification, citation, and future retrieval of the specific data object in question. New versions of data objects should be identified, with a clear distinction between version of the data and metadata. DOIs or other unique URIs that accompany metadata should resolve/link to descriptive landing pages where the data can be downloaded, or where the user can request access to the data. Metadata should be made available through a programmatic access point where possible, such as an OAI-PMH endpoint. Adding tags to indicate the type of media or data present aids in filtering data sets, code, and other resources in the range of nontraditional scholarly output and enables the identification and filtering of nontraditional repository-curated materials from traditional resources, such as journal publications and theses. Detailed information regarding new, updated, and deleted content records enables clarity and accuracy in future content updates. Through our collaborative partnership with DataCite, metadata submitted by data repositories for the purpose of obtaining DataCite DOIs can be routinely evaluated for harvest and for inclusion in the Data Citation Index.

Other considerations exist when building a repository for long-term data preservation at an academic institution. Currently, many institutional data repositories often consist largely of published literature by authors from that institution. A modern institutional repository that serves researchers from the sciences, social sciences, and arts and humanities will accommodate the need for long-term storage of data, software, images, videos, and more.

Repository/Publisher/Data Provider Best Practices

- Curate and validate metadata for completeness, accuracy, and consistency
- Issue permanent IDs for data objects
- Provide unique landing pages for data objects
- Maintain detailed update information and practice versioning
- Indicate data resource type in metadata
- Ensure clear attribution for data objects
- Document the repository mission and policies for inclusion

Benefits of inclusion for repositories include increased visibility due to the use of *Web of Science* in more than 6,500 institutions, as well as dedicated links to the repository for each data item. Increased citation to the repository or publisher aids in future funding requests and metrics for repository use. While currently around 350 data repositories have been included in the Data Citation Index, not all of these achieve all of the criteria described here. In practice, dedication to detail in metadata and data curation, approaches to data set versioning, and reliability and permanence of links to data objects vary widely in this still-emerging landscape. This is also true for many of the repository accreditation criteria put forward by other groups and stakeholders. As data citation and data publishing become more commonplace, these variations will coalesce into a more unified set of guidelines that accommodates the needs of the various stakeholders involved through the work of organizations such as *Clarivate*, DataCite, and RDA.

> *“Benefits of inclusion for repositories include increased visibility due to the use of Web of Science in more than 6,500 institutions, as well as dedicated links to the repository for each data item.*

Research funding, publishing, and assessment

Clarity and specificity in data deposition and publishing guidelines aid research authors as they develop data management plans. Funding organizations and journal publishers are encouraged to make detailed requirements for data sharing available, as well as to have in place well-defined mechanisms to check for compliance with regulations. Recently, a group of journals including *Science* and *Nature* put forward recommendations for publishing standards to further promote data in the academic community[14]. Other recent suggestions for author incentives include rewarding researchers found to be properly disseminating data; conversely researchers not considering issues related to research data dissemination and long-term preservation could be exposed to a lack of funding or inability to publish[15,16]. Journals should work with scientists, scholars, and funding agencies to establish benchmarks and develop enforcement methods toward fulfilling stated requirements. *Clarivate* also encourages publishers to adopt policies and procedures to accommodate and enforce the use of formal data citation and to consider the repositories they recommend for the use of best practices in data preservation and discovery.

The Data Citation Index seeks to assist publishers and funders in determining whether researchers are complying with their data requirements by enabling tracking of data re-use and citation in the research literature established in *Web of Science*, as well as through data discovery using author, institution, and funding information. Developments such as those described above will enable funders to better use the resource to view research output of previously funded research projects, while institutional administrators may better assess the output of academic departments and individual researchers.

Literature Publisher/Funding Organization Best Practices

- Create specific guidelines for data deposition
- Develop formal data citation policies
- Enforce requirements for data sharing and citation
- Establish metadata criteria to allow persistent and unique identification of data in citations

Important practices for Data Citation

- Cite papers that describe the data in addition to, not as a replacement for, citing the data themselves.
- Cite the data in the formal bibliography or in a specific data acknowledgements section, rather than in footnotes or the methods section.
- Formal citation enables secondary services such as *Clarivate* to more readily track the impact and value of the research (e.g., through citation counts). Thus, data can receive the same benefits of the management infrastructure for journal articles.
- Formal citation is the most appropriate way of providing the information necessary to locate and access the data.
- When citing data, use a recognised citation style – either as required by the publisher or a suggested standard (e.g., *Clarivate*, Datacite, etc.).
- Be specific. If there are several versions of the dataset, cite the exact version used in the research.
- Cite data at the finest level of granularity appropriate. Ideally this will be supported by a specific associated identifier, depending on how the data were produced/published. Where necessary, this can be supplemented in the text with more information on the specific subsets or features of the data used.
- Always include dataset identifiers where possible (e.g., DOI or Repository-assigned ID).
- Consider data as primary records of research – cite and be cited.

Conclusion

In order to enable increased citation, discovery, and preservation of scholarly data, further development is needed at each stage of the research data lifecycle. These developments will benefit stakeholder groups in the data community, as well as the scholarly community at large. Data authors, curators, disseminators, and funders with a stated commitment to data citation and access will demonstrate this commitment by enacting practices that promote discovery and accountability. Through our experience with research data and metadata from a variety of disciplines, as well as with a wide range of data repositories based upon a variety of different models, *Clarivate* puts forward these recommendations to increase clarity in the mechanisms needed to uniquely identify submitted data sets and enable new scientific discoveries for scientists and scholars throughout the world.

About the Data Citation Index

Data Citation Index on *Web of Science* provides a single point of access to quality research data from repositories across disciplines and around the world. Research data for this index include data studies, as well as data sets deposited in a recognized repository. Our evaluation process is always underway, with repositories added as often as weekly and existing coverage always under review.

Through linked content and summary information, this index provides researchers with critical perspective and context that is absent when data sets or repositories are viewed in isolation. Updated weekly, this index:

- Includes more than 7 million records from high-quality repositories worldwide
- Features records built from descriptive metadata to create bibliographic records and cited references for digital research
- Provides the scholarly community with standard citation formats for digital research
- Illuminates connections between primary data sets and their context, giving researchers a more complete picture of research output
- Helps users measure the contribution of digital research in specific disciplines and identify potential collaborators
- Enables researchers to discover and provide – or receive – credit for the creation of digital scholarly research data
- Provides data studies from 1900 to present

References

1. Stijn Hoorens, Jeff Rothenberg, Constantijn van Oranje, Martijn van der Mandele, Ruth Levitt (2007). RAND Europe Technical Report. Addressing the uncertain future of preserving the past Towards a robust strategy for digital archiving and preservation. RAND Corporation.

2. FORCE11 (2011). Joint Declaration of Data Citation Principles. [online document]. https://www.force11.org/datacitation/

3. Green, A., Macdonald, S., and Rice, R. (2009). Policy-making for Research Data in Repositories: A Guide. Version 1.2. Data Information Specialists Committee-UK. [online document]. http://www.disc-uk.org/docs/guide.pdf

4. Thomson Reuters Collaborates with DataCite to Expand Discovery of Research Data. [online press release]. http://news.clarivate.com/2014-08-27-Thomson-Reuters-Collaborates-with-DataCite-to-Expand-Discovery-of-Research-Data

5. Thomson Reuters Collaborates with Australian National Data Service to Raise the Profile of Research Data. [online press release]. http://news.clarivate.com/2013-11-05-Thomson-Reuters-Collaborates-with-Australian-National-Data-Service-to-Raise-the-Profile-of-Research-Data

6. Piwowar H.A., Day R.S., Fridsma D.B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

7. National Science Foundation. Press release 10-077 (2010). Scientists seeking NSF funding will soon be required to submit data management plans. http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928

8. ELIXIR et al. (2014). Principles of data management and sharing at European Research Infrastructures. Zenodo. doi:10.5281/zenodo.8304

9. Piwowar H.A, Vision T.J. (2013). Data reuse and the open data citation advantage. PeerJ 1:e175

10. Henneken, E.A., and Accomazzi, A. (2011). Linking to data – Effect on citation rates in astronomy. ASP Conference Series. arxiv.org. http://arxiv.org/pdf/1111.3618v1.pdf

11. European Framework for Audit and Certification of Digital Repositories. Trusted Digital Repository.eu. [online document]. (2010). http://www.trusteddigitalrepository.eu/Trusted%20Digital%20Repository.html

12. ICSU World Data System. Certification of WDS Members. Summary. 11th June 2011. [online document]. https://www.icsu-wds.org/files/wds-certification-summary-11-june-2012.pdf

13. Callaghan et al. Guidelines on Recommending Data Repositories as Partners in Publishing Research Data. International Journal of Digital Curation 2014, Vol. 9, Iss. 1, 152–163, doi:10.2218/ijdc.v9i1.309

14. McNutt, M., Journals unite for reproducibility. Science 7 November 2014: Vol. 346 no. 6210 p. 679, doi: 10.1126/science.aaa1724

15. Van den Eynden, V. and Bishop, L. (2014). Incentives and motivations for sharing research data, A researcher's perspective.

16. Kratz, John. (2015). Making Data Rain. Data Pub. http://datapub.cdlib.org/2015/01/08/make-data-rain/

Who we are

Clarivate Analytics accelerates the pace of innovation by providing trusted insights and analytics to customers around the world, enabling them to discover, protect and commercialize new ideas faster. We own and operate a collection of leading subscription-based services focused on scientific and academic research, patent analytics and regulatory standards, pharmaceutical and biotech intelligence, trademark protection, domain brand protection and intellectual property management. *Clarivate Analytics* is now an independent company with over 4,000 employees, operating in more than 100 countries and owns well-known brands that include *Web of Science*, *Cortellis*, *Derwent*, *CompuMark*, *MarkMonitor* and *Techstreet*, among others. For more information, visit clarivate.com

To learn more, visit:
clarivate.com

North America

Philadelphia:	+1 800 336 4474 +1 215 386 0100

Latin America

Brazil:	+55 11 8370 9845
Other countries:	+1 215 823 5674

Europe, Middle East and Africa

London:	+44 20 7433 4000

Asia Pacific

Singapore:	+65 6775 5088
Tokyo:	+81 3 5218 6500

S027314
10.2017

clarivate.com

Web of Science
Trust the difference

Clarivate Analytics