Achieving human and machine accessibility of cited data in scholarly publications

Joan Starr, Eleni Castro, Mercè Crosas, Michel Dumontier, Robert R. Downs, Ruth Duerr, Laurel Haak, Melissa Haendel, Ivan Herman, Simon Hodson, Joe Hourclé, John Ernest Kratz, Jennifer Lin, Lars Holm Nielsen, Amy L Nurnberger, Stefan Proell, Andreas Rauber, Simone Sacchi, Arthur P. Smith, Michael R Taylor, Tim Clark

This short article provides operational guidance on implementing scholarly data citation and data deposition, in conformance with the Joint Declaration of Data Citation Principles (JDDCP, http://force11.org/datacitation) to help achieve widespread, uniform human and machine accessibility of deposited data. The JDDCP is the outcome of a cross-domain effort to establish core principles around cited data in scholarly publications. It deals with important issues in identification, deposition, description, accessibility, persistence, and evidential status of cited data. Eighty-five scholarly, governmental, and funding institutions have now endorsed the JDDCP. The purpose of this article is to provide the necessary guidance for JDDCP-endorsing organizations to implement these principles and to achieve their widespread adoption.

Achieving human and machine accessibility of cited data in scholarly publications

Joan Starr[1], Eleni Castro[2], Mercè Crosas[2], Michel Dumontier[3], Robert R. Downs[4], Ruth Duerr[5], Laurel L. Haak[6], Melissa Haendel[7], Ivan Herman[8], Simon Hodson[9], Joe Hourclé[10], John Ernest Kratz[1], Jennifer Lin[11], Lars Holm Nielsen[12], Amy Nurnberger[13], Stefan Proell[14], Andreas Rauber[15], Simone Sacchi[13], Arthur Smith[16], Mike Taylor[17], and Tim Clark[18]

[1]**California Digital Library, Oakland CA US**
[2]**Harvard University, Institute of Quantitative Social Sciences, Cambridge MA US**
[3]**Stanford University School of Medicine, Palo Alto CA US**
[4]**Center for International Earth Science Information Network (CIESIN), Columbia University, Palisades, New York US**
[5]**National Snow and Ice Data Center, Boulder CO US**
[6]**ORCID, Inc., Bethesda MD US**
[7]**Oregon Health and Science University, Portland OR US**
[8]**W3C/CWI, Amsterdam, the Netherlands**
[9]**CODATA (ICSU Committee on Data for Science and Technology), Paris FR**
[10]**Solar Data Analysis Center, NASA Goddard Space Flight Center, Greenbelt MD US**
[11]**Public Library of Science, San Francisco CA US**
[12]**European Organization for Nuclear Research (CERN), Geneva CH**
[13]**Columbia University Libraries/Information Services, New York NY US**
[14]**SBA Research, Vienna AT**
[15]**Institute of Software Technology and Interactive Systems, Vienna University of Technology / TU Wien, AT**
[16]**American Physical Society, Ridge NY US**
[17]**Elsevier, Oxford UK**
[18]**Harvard Medical School, Boston MA US**

ABSTRACT

This brief article provides operational guidance on implementing scholarly data citation and data deposition, in conformance with the Joint Declaration of Data Citation Principles (JDDCP, http://force11.org/datacitation) to help achieve widespread, uniform human and machine accessibility of deposited data.

The JDDCP is the outcome of a cross-domain effort to establish core principles around cited data in scholarly publications. It deals with important issues in identification, deposition, description, accessibility, persistence, and evidential status of cited data. Eighty-five scholarly, governmental, and funding institutions have now endorsed the JDDCP.

The purpose of this article is to provide the necessary guidance for JDDCP-endorsing organizations to implement these principles and to achieve their widespread adoption.

Keywords: data accessibility, data citation, machine accessibility, data archiving, scholarly communication, scientific communication

INTRODUCTION

Citation of robustly maintained, described and identified data in persistent digital repositories is an important step towards significantly improving the discoverability, provenance documentation, validation, and reuse of scholarly data; and in validating the robustness of assertions based upon particular data

(CODATA (2013); Altman and King (2006); Uhlir (2012); Ball and Duke (2012); Goodman et al. (2014); Borgman (2012)). It can help reduce the rate of false positives that persist in scholarly literature, and will be transformative in improving the robustness and reproducibility of research findings.

The Joint Declaration of Data Citation Principles (JDDCP) (https://www.force11.org/datacitation) outlines core principles for citing data, based on significant study by participating groups (1) and independent scholars (CODATA (2013); Altman and King (2006); Uhlir (2012); Ball and Duke (2012). It is the latest development in a collective process, reaching back to at least 1977, to raise the importance of data as an independent scholarly product, and to make data ~~transparently~~ available for verification and ~~reproducibility~~ (Altman and Crosas (2013)). However, the JDDCP deliberately did not provide implementation ~~guidelines, which~~ are addressed in this and forthcoming articles.

The purpose of this document is to outline a set of common guidelines to operationalize JDDCP-compliant machine accessibility. Its goal is to do this in a way that is as uniform as possible across conforming repositories and the associated citations of the data they contain. The recommendations outlined here were developed as part of a community process by members of the FORCE11 Data Citation Implementation Group (https://www.force11.org/datacitationimplementation), over a period of approximately one year.

Accessibility to machines and humans is fundamental to providing the required Web access to stable repositories of cited scholarly data and associated metadata, which may have differing lifecycles. This notion is implied by ***all eight*** of the JDDCP principles, beginning with

- ***Principle 1 - 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."

And it is particularly strongly endorsed in the following:

- ***Principle 4 - Unique Identification:*** "A data citation should include a persistent method for identification that is machine actionable. . . "
- ***Principle 5 - 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"
- ***Principle 6 - Persistence:*** "Unique identifiers, and metadata describing the data, and its disposition, should persist – even beyond the lifespan of the data they describe."

The methods proposed below cover:

- identifiers and identifier schemes;
- landing pages;
- minimum acceptable information on landing pages;
- best practices for dataset description; and
- recommended data access methods.

Additional sets of DCIG recommendations on other implementation issues for JDDCP endoresers will be provided in future articles.

WHAT IS MACHINE ACCESSIBILITY?

Machine accessibility of cited data, in the context of this document and the JDDCP, means access to data and metadata stored in a robust repository, by Web services (Booth et al. (2004) - preferably RESTful Web services (Fielding (2000); Fielding and Taylor (2002); Richardson and Ruby (2011)) - independently of integrated browser access by humans.

Clearly, "machine accessibility" is also an underlying prerequisite to human accessibility, as browser access to remote data is always mediated by machine-to-machine communication.

We call out machine accessibility separately here, as in the JDDCP, to emphasize the importance of program-to-program retrieval of data as an integrated services model component.

FIVE RECOMMENDATIONS FOR ACHIEVING MACHINE ACCESSIBILITY

Unique Identification

Unique identification in a manner that is machine-resolvable on the Web, and has a long term demonstrated commitment to persistence, is fundamental to providing access to cited data and its associated metadata. There are several identifier schemes on the Web, which meet these two criteria. The best scheme of identifiers for data citation in a particular community of practice, will be one which meets these criteria and is widely used in that community.

Our general recommendation, based on the JDDCP, is to use any currently-available identifier scheme that is machine actionable, globally unique, and widely (and currently) used by a community; and that has a long term commitment to persistence. Best practice is to choose a scheme that is cross-discipline.

Examples of identifier schemes, meeting JDDCP criteria for robustly-accessible data citation, are shown in Table 1 below.

Identifier	Resolution services	Achieving persistence	Enforcing persistence	Action on object removal
DataCite DOI	datacite.org	registration with contract (2)	link checking	DataCite contacts owners; metadata should persist
CrossRef DOI	crossref.org	registration with contract (3)	link checking	CrossRef contacts owners per policy (4); metadata should persist
Identifiers.org URI	identifiers.org	registration	link checking	metadata should persist
HTTP(s) URI	HTTP	domain owner responsibility	none	fail
PURL URI	purl.org	registration	none	fail
Handle (HDL)	handle.net	registration	none	identifier should persist
ARK	local host web server	user-defined policies	hosting server	host-dependent
N2T ARK	n2t.net	registration	link checking	n2t contacts provider; metadata should persist
NBN	nbn-resolving.org, (DE and CH), urn.fi (FI), various	IETF RFC3188 (Hakala (2001))	domain resolver	metadata should persist

Table 1. Examples of identifer schemes meeting JDDCP criteria.

Landing pages

The identifier included in a citation should point to a landing page or set of pages rather than to the data itself (Rans et al. (2013); Clark et al. (2014)). This is strongly implied by three considerations. First, as mentioned in the JDDCP, the metadata and the data may have different lifespans, the metadata potentially surviving the data. Second, the cited data may not be legally available to all, for reasons of licensing or confidentiality (e.g. Protected Health Information). The landing page provides a method to vend metadata even if the data are no longer present. And it also provides a convenient place where access credentials can be validated. Third, resolution to a landing page allows for an access point that is independent from any multiple encodings of the data which may be available.

By "landing page(s)" we mean a set of representations and presentations of information about the data via both structured metadata and unstructured text and other information. Landing pages should combine human-readable and machine-readable information on a selection of the following items.

- Tools/software: What tools and software may be associated or useful with the datasets, and how to obtain them (certain datasets are not readily usable without specific software).

- Versions: What versions of the data are available, if there are more than one.
- Explanatory or contextual information: Provide explanations, contextual guidance, caveats, and/or documentation for data use, as appropriate.
- Access controls: Access controls based on content licensing, Protected Health Information (PHI) status, Institutional Review Board (IRB) authorization, embargo, or other restrictions, should be implemented here if appropriate.
- Licensing information: Information regarding licensing should be provided, with links to the relevant licensing or waiver documents as required (e.g., Creative Commons CC0 waiver description (https://creativecommons.org/publicdomain/zero/1.0/), or other relevant material).
- Dataset descriptions. The landing page must provide information to programmatically retrieve data where a user or device is so authorized. (See ***Dataset description*** for formats);
- Persistence statement. Reference to a statement describing the data and metadata persistence policies of the repository should be provided at the landing page. Data persistence policies will vary by repository but should be clearly described. (See ***Persistence guarantee*** for recommended language.).
- Data availability and disposition: The landing page should provide information on the availability of the data if it is restricted, or has been de-accessioned (i.e. removed from the archive). As stated in the JDDCP, metadata should persist beyond de-accessioning.

Minimum acceptable information on landing pages

To provide a minimum acceptable level of information on landing pages, there are three guidelines.

1. Minimum content encoding formats for landing pages:
    - HTML (for humans); that is, native browser-interpretable format used to generate a graphical display in a browser window, for human reading and understanding.
    - At least one non-proprietary machine readable format; that is, a content format with a fully specified syntax capable of being parsed by software without ambiguity, at a data element level. Options: XML, JSON/JSON-LD, RDF (Turtle, RDF-XML, N-Triples, N-Quads), microformats, microdata, RDFa.
2. Minimum metadata content
    - Dataset Identifier: A machine-actionable identifier resolvable on the Web to the datase
    - Title: The title of the dataset.
    - Description: A description of the dataset, with more information than the title.
    - Creator: The person(s) and/or organizations who generated the dataset and are responsible for its integrity.
    - Publisher/Contact
    - PublicationDate/Year/ReleaseDate (ISO standard data preferred)
    - Version.
3. Additional suggested metadata content
    - Creator Identifier(s): ORCiD ID(s) of the individual creator(s
    - License: The license under which access to the content is provided (preferably a link to standard license text (e.g. https://creativecommons.org/publicdomain/zero/1.0/).

Best practices for dataset description

The World Wide Web Consortium (http://w3.org) standard for dataset description on the Web is the W3C Data Catalog Vocabulary (Mali et al. (2014)). This is a strongly endorsed best practice for dataset description, common across domains, and widely used. It is a settled standard that can be recommended without qualification.

The W3C Health Care and Life Sciences Dataset Description specification (Gray et al. (2014)), currently in editor's draft status, provides capability to add additional useful metadata beyond the DCAT vocabulary. This is an evolving standard which we recommend for provisional use.

Data might also be presented in other formats, depending on the application area, in which case, content negotiation would be desirable for the URI as well as the landing page URI.

Data access methods

The following are the recommended best approaches for serving content. These can and should be used together for maximum flexibility and accessibility.

1. Use HTTP Accept headers to serve different content based on the request.
    - Also known as "content negotiation"
    - Commonly used in REST web services to serve XML, JSON, HTML, or an RDF serialization.
    - Requires webmaster
    - Generic–works for any kind of 'alternate' type relationships
2. Use HTTP links to direct non-human agents to alternate representations.
    - HTTP response headers, when returning content, should contain entries like:

      ```
      Link:  <uri-to-an-alternate >; rel="alternate";
      media="application/xml"
      ```

    - Requires webmaster
    - Generic–works in any kind of served content
3. Using link elements in HTML to connect to associated content in other formats
    - Example: OAI-ORE to explain how files are inter-related or linking to a file with the DataCite XML.
    - Like "b" but doesn't require webmaster intervention
    - only works in HTML docs

Persistence guarantees

The topic of persistence guarantees is important from the standpoint of what repository owners and managers should provide to support JDDCP-compliant citable persistent data. We recommend that all organizations endorsing the JDDCP adopt a Persistence Guarantee for data and metadata based on the following template:

> *"[Organization/Institution Name] is committed to maintaining persistent identifiers in [Repository Name] so that they will continue to resolve to a landing page providing metadata describing the data, including elements of stewardship, provenance, and availability.*
> *[Organization/Institution Name] has made the following plan for organizational persistence and succession: [plan]."*

As noted in the ***Landing pages*** section, when data is de-accessioned, the landing page should remain online, continuing to provide persistent metadata and other information, including a notation on data de-accessioning. Authors and scholarly article publishers will decide on which repositories meet their persistence and stewardship requirements, based on the guarantees provided, and their overall experience in using various repositories. Guarantees need to be supported by operational practice.

CONCLUSION

These guidelines provide a working basis for implementing Principles 4, 5 and 6 of the Joint Data Citation Principles. They were developed in the Force11.org(9) Data Citation Implementation Group (DCIG, https://www.force11.org/datacitationimplementation), IDMETA subtask, during 2014, as a follow-on project to the successfully concluded Joint Data Citation Principles effort.

Registries of data repositories such as r3data (http://r3data.org) and publishers' lists of "recommended" repositories for cited data, such as those maintained by Nature Publications (http://www.nature.com/sdata/data-policies/repositories), should take note of repository compliance to these guidelines, and provide compliance checklists.

Other deliverables from the DCIG are planned for release in early 2015, including a revision to the NISO-JATS XML schema for document publication and archiving (NISO (2014)), specifically designed to support data citation; and a review of selected data-citation workflows from early-adopter publishers (Nature, Biomed Central, Wiley and Faculty of 1000). The NISO-JATS revision is currently under review by the National Information Standards Organization (NISO) as a draft of NISO JATS version 1.1d2.

It is our hope that publishing this document and others in the series will accelerate the adoption of data citation on a wide scale in the scholarly literature, to support open validation and reuse of results.

We welcome comments and questions, which should be addressed to the forcnet@googlegroups.com open discussion forum.

ACKNOWLEDGMENTS

Many thanks to Maryann Martone for her comments on this document.

ENDNOTES

1. Individuals representing the following organizations participated in the JDDCP development effort: Biomed Central; California Digital Library; CODATA-ICSTI; Columbia University; Creative Commons; DataCite; Digital Science; Elsevier; European Moelcular Biology Laboratories / European Bioinformatics Institute; European Organization for Nuclear Research (CERN); Federation of Earth Science Information Partners (ESIP); FORCE11.org; Harvard Institute of Quantitative Social Sciences; ICSU World Data System; International Association of STM Publishers; Library of Congress (US); Massachusetts General Hospital; MIT Libraries; NASA Solar Data Analysis Center; The National Academies (US); OpenAIRE; Rensselaer Polytechnic Institute; Research Data Alliance; Science Exchange; National Snow and Ice Data Center (US); Natural Environment Research Council (UK); National Academy of Sciences (US); SBA Research (AT);National Information Standards Organization (US); University of California, San Diego; University of Leuven / KU Leuven (NL); University of Oxford; VU University Amsterdam; World Wide Web Consortium (Digital Publishing Activity). See https://www.force11.org/datacitation/workinggroup for details.
2. The DataCite persistence contract language reads: "Objects assigned DOIs are stored and managed such that persistent access to them can be provided as appropriate and maintain all URLs associated with the DOI."
3. The CrossRef persistence contract language reads: "Member must maintain each Digital Identifier assigned to it or for which it is otherwise responsible such that said Digital Identifier continuously resolves to a response page ("Response Page") containing no less than complete bibliographic information about the corresponding Original Work (including without limitation the Digital Identifier), visible on the initial page, with reasonably sufficient information detailing how the Original Work can be acquired and/or a hyperlink leading to the Original Works itself (collectively, "Accessibility Standards")."
4. CrossRef identifier policy reads: "The ... Member shall use the Digital Identifier as the permanent URL link to the Response Page. The... Member shall register the URL for the Response Page with CrossRef, shall keep it up-to-date and active, and shall promptly correct any errors or variances noted by CrossRef."
5. "The Handle System includes an open set of protocols, a namespace, and a reference implementation of the protocols. The protocols enable a distributed computer system to store identifiers, known as handles, of arbitrary resources and resolve those handles into the information necessary to locate, access, contact, authenticate, or otherwise make use of the resources."

6. "This information can be changed as needed to reflect the current state of the identified resource without changing its identifier, thus allowing the name of the item to persist over changes of location and other related state information."
7. For example, the French National Library has rigorous internal checks for the 20 million ARKs that it manages via its own resolver.
8. In most cases the national libraries archive also the content itself (in addition to the content holder) to be preserved with the NBN.
9. Force11.org (http://force11.org) is a community of scholars, librarians, archivists, publishers and research funders that has arisen organically to help facilitate the change toward improved knowledge creation and sharing. It is incorporated as a US 501(c)3 not-for-profit organization in California.

REFERENCES

Altman, M. and Crosas, M. (2013). The evolution of data citation: From principles to implementation. *IAssist Quarterly*, Spring:62–70.

Altman, M. and King, G. (2006). A proposed standard for the scholarly citation of quantitative data. *DLib Magazine*, 13(3/4):march2007–altman.

Ball, A. and Duke, M. (2012). How to cite datasets and link to publications. Technical report, DataCite.

Booth, D., Haas, H., McCabe, F., Newcomer, E., Champion, M., Ferris, C., and Orchard, D. (2004). Web services architecture: W3c working group note 11 february 2004. Technical report, World Wide Web Consortium.

Borgman, C. (2012). *Why are the attribution and citation of scientific data important?* National Academy of Sciences' Board on Research Data and Information. National Academies Press., Washington DC.

Clark, A., Evans, P., and Strollo, A. (2014). Fdsn recommendations for seismic network dois and related fdsn services, version 1.0. Technical report, International Federation of Digital Seismograph Networks.

CODATA (2013). Out of cite, out of mind: The current state of practice, policy and technology for data citation. *Data Science*, 12:1–75.

Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures*. Doctoral dissertation.

Fielding, R. T. and Taylor, R. N. (2002). Principled design of the modern web architecture. *ACM Transactions on Internet Technology*, 2(2):115–150.

Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M., Hogg, D. W., Kashyap, V., Mahabal, A., Siemiginowska, A., and Slavkovic, A. (2014). Ten simple rules for the care and feeding of scientific data. *PLoS Comput Biol*, 10(4):e1003542.

Gray, A., Dumontier, M., Marshall, M., Baram, J., Ansell, P., Bader, G., Bando, A., Callahan, A., Cruz-toledo, J., Gombocz, E., Gonzalez-Beltran, A., Groth, P., Haendel, M., Ito, M., Jupp, S., Katayama, T., Krishnaswami, K., Lin, S., Mungall, C., Le Novere, N., Laibe, C., Juty, N., Malone, J., and Rietveld, L. (2014). Data catalog vocabulary (dcat): W3c recommendation, 16 january 2014. Technical report, World Wide Web Consortium.

Hakala, J. (2001). Rfc3188 - using national bibliography numbers as uniform resource names. Technical report, Internet Engineering Task Force (IETF).

Mali, F., Erickson, J., and Archer, P. (2014). Data catalog vocabulary (dcat): W3c recommendation, 16 january 2014. Technical report, World Wide Web Consortium.

NISO (2014). Jats: Journal article tag suite. ansi/niso z39.96-2012.

Rans, J., Day, M., Duke, M., and Ball, A. (2013). Enabling the citation of datasets generated through public health research. Technical report, Wellcome Trust.

Richardson, L. and Ruby, S. (2011). *RESTful Web Services*. O'Reilly, Sebastopol CA.

Uhlir, P. (2012). For attribution - developing data attribution and citation practices and standards: Summary of an international workshop (2012). Technical report, The National Academies Press.