FORCE11

The Future of Research Communications and e-Scholarship

WORKSHOP

Data Citation Pilot Project Kick-off

February 3, 2016
Boston, MA

This workshop is a supplemental project of bioCADDIE
through the NIH Big Data to Knowledge, Grant 1U24AI117966-01

bioCADDIE

NIH National Institutes of Health
Turning Discovery Into Health

DATA CITATION PILOT PROJECT KICK-OFF WORKSHOP ATTENDEE LIST

First Name	Last Name	Organization
Jeff	Beck	NCBI / NLM / NIH
Robin	Berjon	Wiley
Geoffrey	Bilder	Crossref
John	Chodacki	California Digital Library, UC3
Tim	Clark	Massachusetts General Hospital
Merce	Crosas	Harvard University
Patricia	Cruse	DataCite
Martin	Fenner	DataCite
Ian	Fore	NCI / NIH
Carole	Goble	ELIXIR, The University of Manchester
Florian	Graef	EMBL-EBI
Jeffrey	Grethe	bioCADDIE, UC San Diego
William	Gunn	Mendeley / Elsevier
Stephanie	Hagstrom	FORCE11, UC San Diego
Sandra	Hausmann	Frontiers
Henning	Hermjakob	EMBL-EBI
Leah	Honor	University of Massachusetts
Chun-Nan	Hsu	UC San Diego
Sebastian	Karcher	Northwestern University, Citation Style Language
David	Kennedy	University of Massachusetts, Worcester
Debbie	Lapeyre	Mulberry Technologies, Inc.
Dawei	Lin	NIH
Maryann	Martone	Hypothesis / UC San Diego
Meredith	McCusker	Thomson Reuters, EndNote
Neil	McKenna	Nuclear Receptor Signaling Atlas, Baylor College of Medicine
Daniel	Mietchen	NLM / NIH
Lyubomir	Penev	Pensoft Publishers
Matias	Piipari	Springer, Manuscripts.app Limited
Patrick	Polischuk	PLOS
Simone	Sacchi	Columbia University
Joan	Starr	California Digital Library, DataCite
Mike	Taylor	Elsevier
Dan	Valen	figshare
Richard	Wynne	Aries Systems Corporation
Raymond	Yee	Gluejar, Inc.

JOINT DECLARATION OF DATA CITATION PRINCIPLES

PREAMBLE

Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse.

In support of this assertion, and to encourage good practice, we offer a set of guiding principles for data within scholarly literature, another dataset, or any other research object.

These principles are the synthesis of work by a number of groups (https://www.force11.org/datacitation/workinggroup). As we move into the next phase, we welcome your participation and endorsement of these principles.

PRINCIPLES

The Data Citation Principles cover purpose, function and attributes of citations. These principles recognize the dual necessity of creating citation practices that are both human understandable and machine-actionable.

These citation principles are not comprehensive recommendations for data stewardship. And, as practices vary across communities and technologies will evolve over time, we do not include recommendations for specific implementations, but encourage communities to develop practices and tools that embody these principles.

The principles are grouped so as to facilitate understanding, rather than according to any perceived criteria of importance.

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 [1].

2. 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 [2].

3. Evidence

   In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited [3].

4. Unique Identification

   A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community [4].

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 [5].

6. Persistence

   Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe [6].

7. 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 verifying that the specific timeslice, version and/or granular portion of data retrieved subsequently, is the same as was originally cited [7].

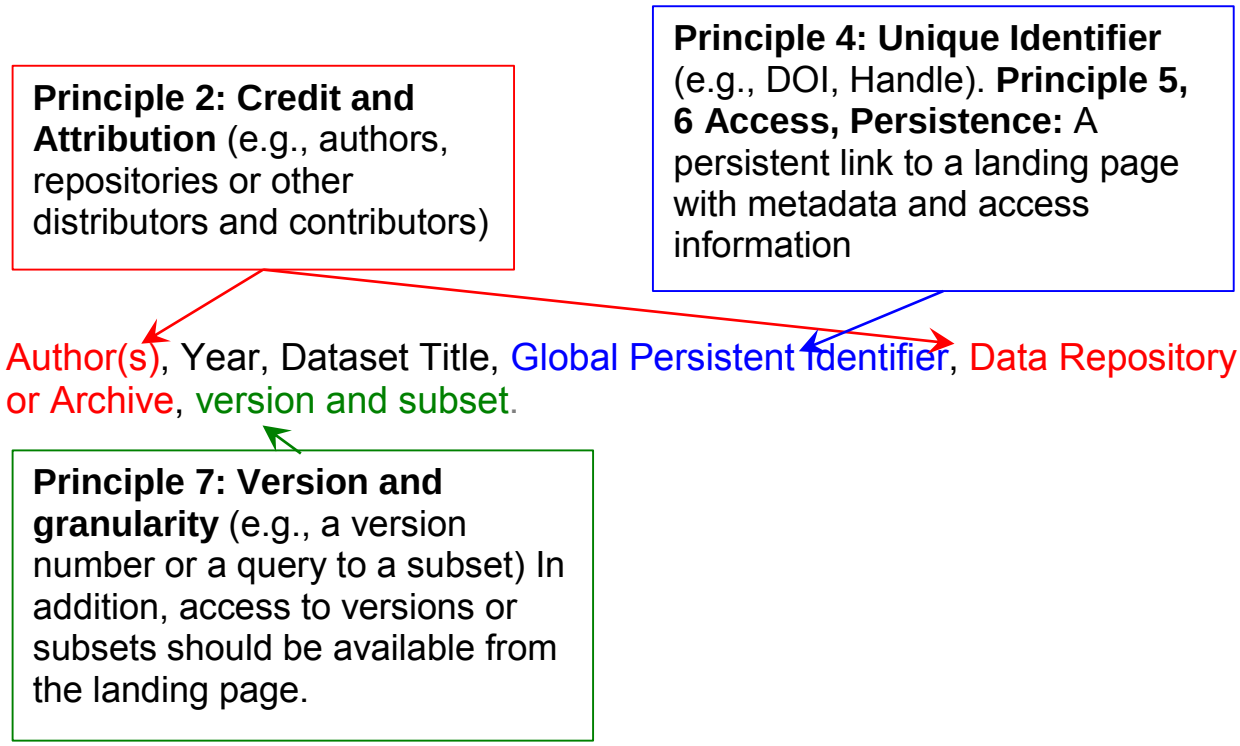
8. Interoperability and Flexibility

   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 [8].

When citing this document please use:
Data Citation Synthesis Group: **Joint Declaration of Data Citation Principles**. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final].

DATA CITATION GENERIC EXAMPLE

This is an example of a data citation as it would appear in a reference list. Note that the format is not intended to be defined with this example, as formats will vary across publishers and communities [Principle 8: Interoperability and flexibility].



PLACEMENT OF CITATIONS

Intra-work:

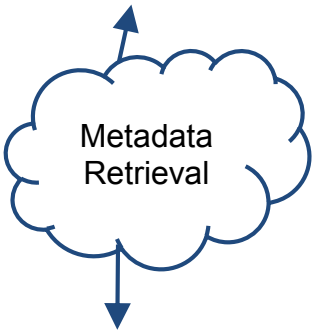
Provides enough information to identify cited data reference in included reference list. May include additional subset information identifying subset specific to that claim.
***Example**:* The plots shown in Figure X show the distribution of selected measures from the main data [Author(s), Year, subset-for-figx].

Full Citation:

Citation may vary in style but should be included in full reference list along with citations to other types or works.
Example:
References Section
Author(s), Year, Journal, Title, Journal, Publisher, DOI.
Author(s), Year, Dataset Title, Global Persistent Identifier, Data Repository or Archive, version and subset,
Author(s), Year, Book Title, Book, Publisher, ISBN.

CITATION METADATA

Author(s), Year, Dataset Title, Global Persistent Identifier, Data Repository or Archive, version and subset.



<contributor role="" id="">Name</contributor>
…
<fixity type="MD5">XXXX</fixity>
<fixity type="UNF">UNF:XXXX</fixity>

<content type>data</content type>
<format>HDF5</format>

Principle 4:
Requires Unique **Machine-Resolvable** Identifier

This could be implemented through DOIs, handles, or other well-recognized machine resolvable standards.

Principle 5:
Requires citation to support access to associated **metadata**

This could be implemented through metadata kept by the ID resolver system, in a standard disciplinary index, such as CrossRef or DataCite, or through machine readable information embedded in a surrogate object (e.g., landing page) to which the id resolves.

Principle 2:
Facilitate attribution & credit for **all contributors**

Principle 7:
Provide **sufficient detail to verifiably link citation and data**

For example, within the DOI system, *principleAgent* fields can be used to document contributors; *referent Identifiers* could be used for fixity information; and *creationType* can be used to identify cited object as data. Under the richard DataCite scheme richer format and contributor role information could be represented.

Principle 5:
Metadata must enable humans and **machines** to make informed use of the data

DATA CITATION PRINCIPLES GLOSSARY

ATTRIBUTION:

(First used in principle 2)
Specification of terms of use of data, usually in the form of a license. Legal attribution is founded on intellectual property rights and licenses as well as on strong normative values in the research community, and the data citations concern individual rights and norms of credit and publicity. Legal attribution is therefore distinguished in these principles from normative (scholarly) attribution, which is concerned with the incentives and systems of scholarly credit and evaluation (adapted from CoData 2013: https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article).

CITATION:

(First used in preamble)
A formal structured reference to another scholarly published or unpublished work (adapted from https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_pdf).

In traditional print publishing, a "bibliographic citation" refers to a formal structured reference to another scholarly published or unpublished work. (This is in contrast to formal bibliometric terminology in which references are made, and citations received.) Typically, intra-document citation pointers to these structured references are marked and abbreviated. These are accompanied by the full bibliographic references to the work appearing in the bibliography or reference list, often following the end of the main text, and is called a "reference" or "bibliographic reference." Traditional print citations include "pinpointing" information, typically in the form of a page range that identifies which part of the cited work is being referenced.

The terminology commonly used for digital citation has come to differ from this older print usage. We adopt the more current usage in which "citation" is used to refer to the full bibliographic reference information for the object. The current usage leaves open the issue of the terminology used to describe the more granular references to data, including subsets of observations, variables, or other components and subsets of a larger data set. These granular references are often necessary in-text to describe the precise evidential support for a data table, figure, or analysis and are analogous to the "pin citation" used in the legal profession or the "page reference" used in citing journal articles. The term "deep citation" has been applied to granular citation to subsets of data.

DATA:

(First used in preamble)
Any record which can be used to support a scholarly research argument, even if it may not be considered valid evidence in all disciplines. In the social sciences, data may include survey responses, interviews and historical documents. Source: modified from http://vso1.nascom.nasa.gov/vso/misc/vocab_2p3.pdf.

The term "data" as used in this document is meant to be broadly inclusive. In addition to digital manifestations of literature (including text, sound, still images, moving images, models, games, and simulations), digital data refers as well to forms of data and databases that are not self-describing -- that generally require the assistance of metadata, computational machinery and/or software in order to be useful, such as various types of laboratory data including spectrographic, genomic sequencing, and electron microscopy data; observational data, such as remote sensing, geospatial, and socio-economic data; and other forms of data either generated or compiled by humans or machines (adapted from CoData Report, 2013: https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article).

DATASET:

(First used in preamble)
Recorded information, regardless of the form or medium on which it may be recorded including writings, films, sound recordings, pictorial reproductions, drawings, designs, or other graphic representations, procedural manuals, forms, diagrams, work flow, charts, equipment descriptions, data files, data processing or computer programs (software), statistical records, and other research data." (from the U.S. National Institutes of Health (NIH) Grants Policy Statement via DataCite's Best Practice Guide for Data Citation). - From DataCite Business Models Principles
http://www.datacite.org/sites/default/files/Business_Models_Principles_v1.0.pdf

IDENTIFIER AND PERSISTENT IDENTIFIER:

(First use in principle 6)
An identifier is an association between a character string and an object. Objects can be files, parts of files, names of persons or organizations, abstractions, etc. Objects can be online or offline. Character strings include URLs, serial numbers, names addresses, etc. A "persistent identifier" is an identifier that is available and managed over time; it will not change if the item is moved or renamed. This means that an item can be reliably referenced for future access by humans and software (from http://n2t.net/ezid/home/understanding).

INTEROPERABILITY:

(First used in principle 8)
The ability of making systems and organizations work together (adapted from Wikipedia: https://en.wikipedia.org/wiki/Interoperability). Access to research data, as facilitated by data citations, requires technological infrastructure that is appropriately designed and based on interoperability best practices that include data quality control, security, and authorizations. Currently, interoperability at both the semantic and the infrastructure levels is important to ensure that data citations facilitate access to research data. However, organizations working to develop improved infrastructures that foster interoperability should widely communicate the standards, guidelines, and best practices that are being implemented; adopt standards for data documentation (such as metadata) and dissemination (data citations, including bidirectional links from data to publications and vice versa); and maintain an up-to-date knowledge of the evolution of not only the technologies implemented but also the best practices efforts being executed by the community of practice (adapted from CoData Report, 2013, Ch. 5: https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article).

MACHINE-ACTIONABLE:

(First used in introduction to principles)
Content that can be used and manipulated by computers (http://www.libraries.psu.edu/tas/jca/ccda/docs/tf-MRData3.pdf).

METADATA:

(First used in preamble)
Information about the data being tracked within a data system. Metadata typically conforms to a metadata information model. Metadata may include, for example, the name of the sensor used to collect the data or person who collected the data, where the data was collected, information about the units and dimensionality of the data, and other notes recorded by the investigator about how the data has been processed. Source: modified from http://vso1.nascom.nasa.gov/vso/misc/vocab_2p3.pdf.

Metadata is information (data) about the object and its disposition, such as the name of the object's creator, the date of creation, the target URL, the version of the object, its title, and so on (from http://n2t.net/ezid/home/understanding).

RESEARCH OBJECT:

(First used in preamble)
Sharable, reusable digital objects that enable research to be recorded and reused (adapted from Wikipedia: http://wiki.myexperiment.org/index.php/Research_Objects).

SCHOLARSHIP:

(First used in preamble)
Serious formal study or research of a subject (adapted from Merriam-Webster Dictionary).

VERIFICATION, PROVENANCE AND FIXITY:

(First used in principle 7)
Verification means to reliably establish the relationship between the cited object of a original citation and a current object -- verification enables one to confirm that the data retrieved is the data cited. This is separate from persistence, which remains the responsibility of the archive, not the citation.

Types of verification information include fixity -- which can be used directly to assess the integrity of specific content, and provenance, which provides information about parts of the chain of custody and/or processing to which the content was subject. Specific forms of citation verification include, but are not limited to: embedding fixity information in the citation itself; associating the citation with a surrogate (such as a landing page) where additional metadata, such as the data form, fixity, and final stage of provenance, are given explicitly; or associating such metadata with the DOI, handle, or other persistent identifier persistent identifier itself directly, through the persistent identifier's resolution or index service (adapted from CoData, 2013).

VERSION:

(First used in principle 7)
A modified dataset based on a single designated dataset -- roughly equivalent to an "edition" in FRBR terms. [1]

This is often denoted with a number that is increased when the data changes, and can also be described by a "timeslice" or access date where a formal version is unavailable, for example [2].

[1] http://archive.ifla.org/VII/s13/frbr/frbr2.htm
[2] Starr, J., & Gastl, A. (2011). Is CitedBy: A metadata scheme for DataCite. D-Lib Magazine, 17(1/2). doi:10.1045/january2011-starr.

REFERENCES

1. 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: 1-75., <http://dx.doi.org/10.2481/dsj.OSOM13-043> sec 3.2.1; Uhlir (ed.) 2012,*Developing Data Attribution and Citation Practices and Standards.* National Academies. <http://www.nap.edu/download.php?record_id=13564>, ch. 14.; Altman, Micah, and Gary King. 2007. "A proposed standard for the scholarly citation of quantitative data." *D-lib Magazine* 13.3/4. <http://www.dlib.org/dlib/march07/altman/03altman.html>
2. CODATA 2013, Sec 3.2; 7.2.3; . Uhlir (ed.), 2012. ,ch. 14
3. CODATA 2013, Sec 3.1; 7.2.3; Uhlir (ed.) 2012, ch. 14
4. Altman-King 2007; CODATA 2013, Sec 3.2.3, Ch. 5; Ball, A., Duke, M. (2012). ‘Data Citation and Linking’. DCC Briefing Papers. Edinburgh: Digital Curation Centre. <http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/data-citation-and-linking>
5. CODATA 2013, Sec 3.2.4, 3.2.5, 3.2.8
6. Altman-King 2007; Ball & Duke 2012; CODATA 2013, Sec 3.2.2
7. Altman-King 2007; CODATA 2013, Sec 3.2.7, 3.2.8
8. CODATA 2013, Sec 3.2.10

PeerJ Computer Science

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, United States of America
[2] Institute of Quantitative Social Sciences, Harvard University, Cambridge, MA, United States of America
[3] Stanford University School of Medicine, Stanford, CA, United States of America
[4] Center for International Earth Science Information Network (CIESIN), Columbia University, Palisades, NY, United States of America
[5] National Snow and Ice Data Center, Boulder, CO, United States of America
[6] ORCID, Inc., Bethesda, MD, United States of America
[7] Oregon Health and Science University, Portland, OR, United States of America
[8] World Wide Web Consortium (W3C)/Centrum Wiskunde en Informatica (CWI), Amsterdam, Netherlands
[9] ICSU Committee on Data for Science and Technology (CODATA), Paris, France
[10] Solar Data Analysis Center, NASA Goddard Space Flight Center, Greenbelt, MD, United States of America
[11] Public Library of Science, San Francisco, CA, United States of America
[12] European Organization for Nuclear Research (CERN), Geneva, Switzerland
[13] Columbia University Libraries/Information Services, New York, NY, United States of America
[14] SBA Research, Vienna, Austria
[15] Institute of Software Technology and Interactive Systems, Vienna University of Technology/TU Wien, Austria
[16] American Physical Society, Ridge, NY, United States of America
[17] Elsevier, Oxford, United Kingdom
[18] Harvard Medical School, Boston, MA, United States of America

Submitted 15 December 2014
Accepted 5 February 2015
Published 27 May 2015

Corresponding author
Tim Clark, tim_clark@harvard.edu

Academic editor
Harry Hochheiser

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj-cs.1**

ABSTRACT

Reproducibility and reusability of research results is an important concern in scientific communication and science policy. A foundational element of reproducibility and reusability is the open and persistently available presentation of research data. However, many common approaches for primary data publication in use today do not achieve sufficient long-term robustness, openness, accessibility or uniformity. Nor do they permit comprehensive exploitation by modern Web technologies. This has led to several authoritative studies recommending uniform direct citation of data archived in persistent repositories. Data are to be considered as first-class scholarly objects, and treated similarly in many ways to cited and archived scientific and scholarly literature. Here we briefly review the most current and widely agreed set of principle-based recommendations for scholarly data citation, the Joint Declaration of Data Citation Principles (JDDCP). We then present a framework for operationalizing the JDDCP; and a set of initial recommendations on identifier schemes, identifier

resolution behavior, required metadata elements, and best practices for realizing programmatic machine actionability of cited data. The main target audience for the common implementation guidelines in this article consists of publishers, scholarly organizations, and persistent data repositories, including technical staff members in these organizations. But ordinary researchers can also benefit from these recommendations. The guidance provided here is intended to help achieve widespread, uniform human and machine accessibility of deposited data, in support of significantly improved verification, validation, reproducibility and re-use of scholarly/scientific data.

Subjects Human–Computer Interaction, Data Science, Digital Libraries, World Wide Web and Web Science
Keywords Data citation, Machine accessibility, Data archiving, Data accessibility

INTRODUCTION

Background

An underlying requirement for verification, reproducibility, and reusability of scholarship is the accurate, open, robust, and uniform presentation of research data. This should be an integral part of the scholarly publication process.[1] However, *Alsheikh-Ali et al. (2011)* found that a large proportion of research articles in high-impact journals either weren't subject to or didn't adhere to any data availability policies at all. We note as well that such policies are not currently standardized across journals, nor are they typically optimized for data reuse. This finding reinforces significant concerns recently expressed in the scientific literature about reproducibility and whether many false positives are being reported as fact (*Colquhoun, 2014*; *Rekdal, 2014*; *Begley & Ellis, 2012*; *Prinz, Schlange & Asadullah, 2011*; *Greenberg, 2009*; *Ioannidis, 2005*).

Data transparency and open presentation, while central notions of the scientific method along with their complement, reproducibility, have met increasing challenges as dataset sizes grow far beyond the capacity of printed tables in articles. An extreme example is the case of DNA sequencing data. This was one of the first classes of data, along with crystallographic data, for which academic publishers began to require database accession numbers as a condition of publishing, as early as the 1990's. At that time sequence data could actually still be published as text in journal articles. The Atlas of Protein Sequence and Structure, published from 1965 to 78, was the original form in which protein sequence data was compiled: a book, which could be cited (*Strasser, 2010*). Today the data volumes involved are absurdly large (*Salzberg & Pop, 2008*; *Shendure & Ji, 2008*; *Stein, 2010*). Similar transitions from printed tabular data to digitized data on the web have taken place across disciplines.

Reports from leading scholarly organizations have now recommended a uniform approach to treating research data as first-class research objects, similarly to the way textual publications are archived, indexed, and cited (*CODATA-ICSTI Task Group , 2013*; *Altman & King, 2006*; *Uhlir, 2012*; *Ball & Duke, 2012*). Uniform citation of robustly archived,

[1] Robust citation of archived methods and materials—particularly highly variable materials such as cell lines, engineered animal models, etc.—and software—are important questions not dealt with here. See *Vasilevsky et al. (2013)* for an excellent discussion of this topic for biological reagents.

described, and identified data in persistent digital repositories is proposed as an important step towards significantly improving the discoverability, documentation, validation, reproducibility, and reuse of scholarly data (*CODATA-ICSTI Task Group , 2013*; *Altman & King, 2006*; *Uhlir, 2012*; *Ball & Duke, 2012*; *Goodman et al., 2014*; *Borgman, 2012*; *Parsons, Duerr & Minster, 2010*).

The Joint Declaration of Data Citation Principles (JDDCP) (*Data Citation Synthesis Group, 2014*) is a set of top-level guidelines developed by several stakeholder organizations as a formal synthesis of current best-practice recommendations for common approaches to data citation. It is based on significant study by participating groups and independent scholars.[2] The work of this group was hosted by the FORCE11 (http://force11.org) community, an open forum for discussion and action on important issues related to the future of research communication and e-Scholarship.

The JDDCP 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 & Crosas, 2013*).

The purpose of this document is to outline a set of common guidelines to operationalize JDDCP-compliant data citation, archiving, and programmatic machine accessibility in a way that is as uniform as possible across conforming repositories and associated data citations. The recommendations outlined here were developed as part of a community process by participants representing a wide variety of scholarly organizations, hosted by the FORCE11 Data Citation Implementation Group (DCIG) (https://www.force11.org/datacitationimplementation). This work was conducted over a period of approximately one year beginning in early 2014 as a follow-on activity to the completed JDDCP.

Why cite data?

Data citation is intended to help guard the integrity of scholarly conclusions and provides a basis for integrating exponentially growing datasets into new forms of scholarly publishing. Both of these goals require the systematic availability of primary data in both machine- and human-tractable forms for re-use. A systematic review of current approaches is provided in *CODATA-ICSTI Task Group (2013)*.

Three common practices in academic publishing today block the systematic reuse of data. The first is the citation of primary research data in footnotes, typically either of the form, "data is available from the authors upon request", or "data is to be found on the authors' laboratory website, http://example.com". The second is publication of datasets as "Supplementary File" or "Supplementary Data" PDFs where data is given in widely varying formats, often as graphical tables, and which in the best case must be laboriously screen-scraped for re-use. The third is simply failure in one way or another to make the data available at all.

Integrity of conclusions (and assertions generally) can be guarded by tying individual assertions in text to the data supporting them. This is done already, after a fashion, for image data in molecular biology publications where assertions based on primary data contained in images typically directly cite a supporting figure within the text

[2] Individuals representing the following organizations participated in the JDDCP development effort: Biomed Central; California Digital Library; CODATA-ICSTI Task Group on Data Citation Standards and Practices; Columbia University; Creative Commons; DataCite; Digital Science; Elsevier; European Molecular Biology Laboratories/European Bioinformatics Institute; European Organization for Nuclear Research (CERN); Federation of Earth Science Information Partners (ESIP); FORCE11.org; Harvard Institute for 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.

containing the image. Several publishers (e.g., PLoS, Nature Publications, and Faculty of 1000) already partner with data archives such as FigShare (http://figshare.com), Dryad (http://datadryad.org/), Dataverse (http://dataverse.org/), and others to archive images and other research data.

Citing data also helps to establish the value of the data's contribution to research. Moving to a cross-discipline standard for acknowledging the data allows researchers to justify continued funding for their data collection efforts (*Uhlir, 2012*; *CODATA-ICSTI Task Group , 2013*). Well defined standards allow bibliometric tools to find unanticipated uses of the data. Current analysis of data use is a laborious process and rarely performed for disciplines outside of the disciplines considered the data's core audience (*Accomazzi et al., 2012*).

The eight core Principles of data citation

The eight Principles below have been endorsed by 87 scholarly societies, publishers and other institutions.[3] Such a wide endorsement by influential groups reflects, in our view, the meticulous work involved in preparing the key supporting studies (by CODATA, the National Academies, and others (*CODATA-ICSTI Task Group , 2013*; *Uhlir, 2012*; *Ball & Duke, 2012*; *Altman & King, 2006*) and in harmonizing the Principles; and supports the validity of these Principles as foundational requirements for improving the scholarly publication ecosystem.

- ***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."
- ***Principle 2—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."
- ***Principle 3—Evidence:*** "In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited."
- ***Principle 4—Unique Identification:*** "A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community."
- ***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."
- ***Principle 7—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 verifying that the specific time slice, version and/or granular portion of data retrieved subsequently is the same as was originally cited."

[3] These organizations include the *American Physical Society, Association of Research Libraries, Biomed Central, CODATA, CrossRef, DataCite, DataONE, Data Registration Agency for Social and Economic Data, ELIXIR, Elsevier, European Molecular Biology Laboratories/European Bioinformatics Institute, Leibniz Institute for the Social Sciences, Inter-University Consortium for Political and Social Research, International Association of STM Publishers, International Union of Biochemistry and Molecular Biology, International Union of Crystallography, International Union of Geodesy and Geophysics, National Information Standards Organization (US), Nature Publishing Group, OpenAIRE, PLoS (Public Library of Science), Research Data Alliance, Royal Society of Chemistry, Swiss Institute of Bioinformatics, Cambridge Crystallographic Data Centre, Thomson Reuters, and the University of California Curation Center (California Digital Library).*

- ***Principle 8—Interoperability and Flexibility:*** "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."

These Principles are meant to be adopted at an institutional or discipline-wide scale. The main target audience for the common implementation guidelines in this article consists of publishers, scholarly organizations, and persistent data repositories. Individual researchers are not meant to set up their own data archives. In fact this is contrary to one goal of data citation as we see it—which is to get away from inherently unstable citations via researcher footnotes indicating data availability at some intermittently supported laboratory website. However individual researchers can contribute to and benefit from adoption of these Principles by ensuring that primary research data is prepared for archival deposition at or before publication. We also note that often a researcher will want to go back to earlier primary data from their own lab—robust archiving positively ensures it will remain available for their own use in future, whatever the vicissitudes of local storage and lab personnel turnover.

Implementation questions arising from the JDDCP

The JDDCP were presented by their authors as Principles. Implementation questions were left unaddressed. This was meant to keep the focus on harmonizing top-level and basically goal-oriented recommendations without incurring implementation-level distractions. Therefore we organized a follow-on activity to produce a set of implementation guidelines intended to promote rapid, successful, and uniform JDDCP adoption. We began by seeking to understand just what questions would arise naturally to an organization that wished to implement the JDDCP. We then grouped the questions into four topic areas, to be addressed by individuals with special expertise in each area.

1. Document Data Model—How should publishers adapt their document data models to support direct citation of data?
2. Publishing Workflows—How should publishers change their editorial workflows to support data citation? What do publisher data deposition and citation workflows look like where data is being cited today, such as in *Nature Scientific Data* or *GigaScience*?
3. Common Repository Application Program Interfaces (APIs)—Are there any approaches that can provide standard programmatic access to data repositories for data deposition, search and retrieval?
4. Identifiers, Metadata, and Machine Accessibility—What identifier schemes, identifier resolution patterns, standard metadata, and recommended machine programmatic accessibility patterns are recommended for directly cited data?

The **Document Data Model** group noted that publishers use a variety of XML schemas (*Bray et al., 2008*; *Gao, Sperberg-McQueen & Thompson, 2012*; *Peterson et al., 2012*) to model scholarly articles. However, there is a relevant National Information Standards

Organization (NISO) specification, NISO Z39.96-2012, which is increasingly used by publishers, and is the archival form for biomedical publications in PubMed Central. [4] This group therefore developed a proposal for revision of the NISO Journal Article Tag Suite to support direct data citation. NISO-JATS version 1.1d2 (*National Center for Biotechnology Information, 2014*), a revision based on this proposal, was released on December 29, 2014, by the JATS Standing Committee, and is considered a stable release, although it is not yet an official revision of the NISO Z39.96-2012 standard.

[4] NISO Z39.96-2012 is derived from the former "NLM-DTD" model originally developed by the US National Library of Medicine.

The **Publishing Workflows** group met jointly with the Research Data Alliance's Publishing Data Workflows Working Group to collect and document exemplar publishing workflows. An article on this topic is in preparation, reviewing basic requirements and exemplar workflows from *Nature Scientific Data*, *GigaScience (Biomed Central)*, *F1000Research*, and *Geoscience Data Journal (Wiley)*.

The **Common Repository APIs** group is currently planning a pilot activity for a common API model for data repositories. Recommendations will be published at the conclusion of the pilot. This work is being undertaken jointly with the ELIXIR (http://www.elixir-europe.org/) Fairport working group.

The **Identifiers, Metadata, and Machine Accessibility** group's recommendations are presented in the remainder of this article. These recommendations cover:

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

RECOMMENDATIONS FOR ACHIEVING MACHINE ACCESSIBILITY

What is machine accessibility?

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

Web services are methods of program-to-program communication using Web protocols. The World Wide Web Consortium (W3C, http://www.w3.org) defines them as "software system[s] designed to support interoperable machine-to-machine interaction over a network" (*Haas & Brown, 2004*).

Web services are always "on" and function essentially as utilities, providing services such as computation and data lookup, at *web service endpoints*. These are well-known Web addresses, or Uniform Resource Identifiers (URIs) (*Berners-Lee, Fielding & Masinter, 1998*; *Jacobs & Walsh, 2004*).[5]

[5] URIs are very similar in concept to the more widely understood Uniform Resource Locators (URL, or "Web address"), but URIs do not specify the location of an object or service—they only identify it. URIs specify *abstract* resources on the Web. The associated server is responsible for resolving a URI to a specific physical resource—if the resource is resolvable. (URIs may also be used to identify physical things such as books in a library, which are not directly resolvable resources on the Web.)

RESTful Web services follow the REST (Representational State Transfer) architecture developed by Fielding and others (*Fielding, 2000*). They support a standard set of operations such as "get" (retrieve), "post" (create), and "put" (create or update) and are highly useful in building hypermedia applications by combining services from many programs distributed on various Web servers.

Machine accessibility and particularly RESTful Web service accessibility is highly desirable because it enables construction of "Lego block" style programs built up from various service calls distributed across the Web, which need not be replicated locally. RESTful Web services are recommended over the other major Web service approach, SOAP interfaces (*Gudgin et al., 2007*), due to our focus on the documents being served and their content. REST also allows multiple data formats such as JSON (JavaScript Object Notation) (*ECMA, 2013*), and provides better support for mobile applications (e.g., caching, reduced bandwidth, etc.).

Clearly, "machine accessibility" is also an underlying prerequisite to human accessibility, as browser (client) access to remote data is always mediated by machine-to-machine communication. But for flexibility in construction of new programs and services, it needs to be independently available apart from access to data generated from the direct browser calls.

Unique identification

Unique identification in a manner that is machine-resolvable on the Web and demonstrates a long-term commitment to persistence is fundamental to providing access to cited data and its associated metadata. There are several identifier schemes on the Web that meet these two criteria. The best identifiers for data citation in a particular community of practice will be those that meet these criteria and are 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 demonstrated a long-term commitment to persistence. Best practice, given the preceding, is to choose a scheme that is also cross-discipline. *Machine actionable* in this context means resolvable on the Web by Web services.

There are basically two kinds of identifier schemes available: (a) the native HTTP and HTTPS schemes where URIs are the identifiers and address resolution occurs natively; and (b) schemes requiring a resolving authority, like Digital Object Identifiers (DOIs).

Resolving authorities reside at well-known web addresses. They issue and keep track of identifiers in their scheme and *resolve* them by translating them to URIs which are then natively resolved by the Web. For example, the DOI 10.1098/rsos.140216 when appended to the DOI resolver at http://doi.org, resolves to the URI http://rsos.royalsocietypublishing.org/content/1/3/140216. Similarly, the biosample identifier SAMEG120702, when appended as ("biosample/SAMEG120702") to the identifiers.org resolver at http://identifiers.org, resolves to the landing page www.ebi.ac.uk/biosamples/group/SAMEG120702. However resolved, a cited identifier should continue to resolve to an intermediary *landing page* (see below) even if the underlying data has been de-accessioned or is otherwise unavailable.

Table 1 Examples of identifier schemes meeting JDDCP criteria.

Identifier scheme	Full name	Authority	Resolution URI
DataCite DOI (as URI)	DataCite-assigned Digital Object Identifier	DataCite	http://dx.doi.org
CrossRef DOI (as URI)	CrossRef-assigned Digital Object Identifier	CrossRef	http://dx.doi.org
Identifiers.org URI	Identifiers.org-assigned Uniform Resource Identifier	Identifiers.org	http://identifiers.org
HTTPS URI	HTTP or HTTPS Uniform Resource Identifier	Domain name owner	n/a
PURL	Persistent Uniform Resource Locator	Online Computer Library Center (OCLC)	http://purl.org
Handle (HDL)	Handle System HDL	Corporation for National Research Initiatives (CNRI)	http://handle.net
ARK	Archival Resource Key	Name Assigning or Mapping Authorities (various)[a]	http://n2t.net; Name Mapping Authorities
NBN	National Bibliographic Number	Various	Various

Notes.

[a] Registries maintained at California Digital Library, Bibliothèque National de France and National Library of Medicine.

By a commitment to persistence, we mean that (a) if a resolving authority is required that authority has demonstrated a reasonable chance to be present and functional in the future; (b) the owner of the domain or the resolving authority has made a credible commitment to ensure that its identifiers will always resolve. A useful survey of persistent identifier schemes appears in *Hilse & Kothe (2006)*.

Examples of identifier schemes meeting JDDCP criteria for robustly accessible data citation are shown in Table 1 and described below. This is not a comprehensive list and the criteria above should govern. Table 2 summarizes the approaches to achieving and enforcing persistence, and actions on object (data) removal from the archive, of each of the schemes.

The subsections below briefly describe the exemplar identifier schemes shown in Tables 1 and 2.

Digital Object Identifiers (DOIs)

Digital Object Identifiers are an identification system originally developed by trade associations in the publishing industry for digital content over the Internet. They were developed in partnership with the Corporation for National Research Initiatives (CNRI), and built upon CNRI's *Handle System* as an underlying network component. However, DOIs may identify digital objects *of any type*—certainly including data (*International DOI Foundation, 2014*).

DOI syntax is defined as a US National Information Standards Organization standard, ANSI/NISO Z39.84-2010. DOIs may be expressed as URIs by prefixing the DOI with a resolution address: `http://dx.doi.org/<doi>`. DOI Registration Agencies provide services for registering DOIs along with descriptive metadata on the object being identified. The DOI system Proxy Server allows programmatic access to DOI name resolution using HTTP (*International DOI Foundation, 2014*).

DataCite and **CrossRef** are the two DOI Registration Agencies of special relevance to data citation. They provide services for registering and resolving identifiers for cited data.

Table 2 **Identifier scheme persistence and object removal behavior.**

Identifier scheme	Achieving persistence	Enforcing persistence	Action on object removal
DataCite DOI	Registration with contract[a]	Link checking	DataCite contacts owners; metadata should persist
CrossRef DOI	Registration with contract[b]	Link checking	CrossRef contacts owners per policy[c]; metadata should persist
Identifiers.org URI	Registration	Link checking	Metadata should persist
HTTPS URI	Domain owner responsibility	None	Domain owner responsibility
PURL URI	Registration	None	Domain owner responsibility
Handle (HDL)	Registration	None	Identifier should persist
ARK	User-defined policies	Hosting server	Host-dependent; metadata should persist[d]
NBN	IETF RFC3188	Domain resolver	Metadata should persist

Notes.

[a] 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."

[b] The CrossRef persistence contract language reads in part: "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…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…"

[c] 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."

[d] For example, the French National Library has rigorous internal checks for the 20 million ARKs that it manages via its own resolver.

Both require persistence commitments of their registrants and take active steps to monitor compliance. DataCite is specifically designed—as its name would indicate—to support data citation.

A recent collaboration between the software archive GitHub, the Zenodo repository system at CERN, FigShare, and Mozilla Science Lab, now makes it possible to cite software, giving DOIs to GitHub-committed code (*GitHub Guides, 2014*).

Handle System (HDLs)

Handles are identifiers in a general-purpose global name service designed for securely resolving names over the Internet, compatible with but not requiring the Domain Name Service. Handles are location independent and persistent. The system was developed by Bob Kahn at the Corporation for National Research Initiatives, and currently supports, on average, 68 million resolution requests per month—the largest single user being the Digital Object Identifier (DOI) system. Handles can be expressed as URIs (*CNRI, 2014*; *Dyson, 2003*).

Identifiers.org Uniform Resource Identifiers (URIs)

Many common identifiers used in the life sciences, such as PubMed or Protein Data Bank IDs, are not natively Web-resolvable. Identifiers.org associates such database-dependent identifiers with persistent URIs and resolvable physical URLs. Identifiers.org was developed and is maintained at the European Bioinformatics Institute, and was built on top of the MIRIAM registry (*Juty, Le Novére & Laibe, 2012*).

Identifiers.org URIs are constructed using the syntax `http://identifiers.org/<data resource name>/<native identifier>`, where `<data resource name>` designates a particular database, and `<native identifier>` is the ID used within that database to retrieve the record. The Identifiers.org resolver supports multiple

alternative locations (which may or may not be mirrors) for data it identifies. It supports programmatic access to data.

PURLs

PURLs are "Persistent Uniform Resource Locators", a system originally developed by the Online Computer Library Center (OCLC). They act as intermediaries between potentially changing locations of digital resources, to which the PURL name resolves. PURLs are registered and resolved at http://purl.org, http://purl.access.gpo.gov, http://purl.bioontology.org and various other resolvers. PURLs are implemented as an HTTP redirection service and depend on the survival of their host domain name (*OCLC, 2015*; *Library of Congress, 1997*). PURLs fail to resolve upon object removal. Handling this behavior through a metadata landing page (see below) is the responsibility of the owner of the cited object.

HTTP URIs

URIs (Uniform Resource Identifiers) are strings of characters used to identify resources. They are the identifier system for the Web. URIs begin with a *scheme name*, such as `http` or `ftp` or `mailto`, followed by a colon, and then a scheme-specific part. HTTP URIs will be quite familiar as they are typed every day into browser address bars, and begin with `http:`. Their scheme-specific part is next, beginning with "`//`", followed by an identifier, which often but not always is resolvable to a specific resource on the Web. URIs by themselves have no mechanism for storing metadata about any objects to which they are supposed to resolve, nor do they have any particular associated persistence policy. However, other identifier schemes with such properties, such as DOIs, are often represented as URIs for convenience (*Berners-Lee, Fielding & Masinter, 1998*; *Jacobs & Walsh, 2004*).

Like PURLs, native HTTP URIs fail to resolve upon object removal. Handling this behavior through a metadata landing page (see below) is the responsibility of the owner of the cited object.

Archival Resource Key (ARKs)

Archival Resource Keys are unique identifiers designed to support long-term persistence of information objects. An ARK is essentially a URL (Uniform Resource Locator) with some additional rules. For example, hostnames are excluded when comparing ARKs in order to prevent current hosting arrangements from affecting identity. The maintenance agency is the California Digital Library, which offers a hosted service for ARKs and DOIs (*Kunze & Starr, 2006*; *Kunze, 2003*; *Kunze & Rodgers, 2013*; *Janée, Kunze & Starr, 2009*).

ARKs provide access to three things—an information object; related metadata; and the provider's persistence commitment. ARKs propose inflections (changing the end of an identifier) as a way to retrieve machine-readable metadata without requiring (or prohibiting) content negotiation for linked data applications. Unlike, for example, DOIs, there are no fees to assign ARKs, which can be hosted on an organization's own web server if desired. They are globally resolvable via the identifier-scheme-agnostic N2T (Name-To-Thing, http://n2t.net) resolver. The ARK registry is replicated at the California

Digital Library, the Bibliothéque Nationale de France, and the US National Library of Medicine (*Kunze & Starr, 2006*; *Peyrard, Kunze & Tramoni, 2014*; *Kunze, 2012*).

National Bibliography Number (NBNs)

National Bibliography Numbers are a set of related publication identifier systems with country-specific formats and resolvers, utilized by national library systems in some countries. They are used by, for example, Germany, Sweden, Finland and Italy, for publications in national archives without publisher-assigned identifiers such as ISBNs. There is a URN namespace for NBNs that includes the country code; expressed as a URN, NBNs become globally unique (*Hakala, 2001*; *Moats, 1997*).

Landing pages

The identifier included in a citation should point to a landing page or set of pages rather than to the data itself (*Hourclé et al., 2012*; *Rans et al., 2013*; *Clark, Evans & Strollo, 2014*). And the landing page should persist even if the data is no longer accessible. By "landing page(s)" we mean a set of information about the data via both structured metadata and unstructured text and other information.

There are three main reasons to resolve identifiers to landing pages rather than directly to data. First, as proposed in the JDDCP, the metadata and the data may have different lifespans, the metadata potentially surviving the data. This is true because data storage imposes costs on the hosting organization. Just as printed volumes in a library may be de-accessioned from time to time, based on considerations of their value and timeliness, so will datasets. The JDDCP proposes that metadata, essentially cataloging information on the data, should still remain a citable part of the scholarly record even when the dataset may no longer be available.

Second, the cited data may not be legally available to all, even when initially accessioned, for reasons of licensing or confidentiality (e.g. Protected Health Information). The landing page provides a method to host metadata even if the data is 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 that may be available.

Landing pages should contain the following information. Items marked "conditional" are recommended if the conditions described are present, e.g., access controls are required to be implemented if required by licensing or PHI considerations; multiple versions are required to be described if they are available; etc.

- (recommended) **Dataset descriptions:** The landing page must provide descriptions of the datasets available, and information on how to programmatically retrieve data where a user or device is so authorized. (See ***Dataset description*** for formats.)
- (conditional) **Versions:** What versions of the data are available, if there is more than one version that may be accessed.
- (optional) **Explanatory or contextual information:** Provide explanations, contextual guidance, caveats, and/or documentation for data use, as appropriate.

- (conditional) **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 they are required.
- (recommended) **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).
- (recommended) **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.)
- (conditional) **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.
- (optional) **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).

Content encoding on landing pages

Landing pages should provide both human-readable and machine-readable content.

- **HTML**; that is, the native browser-interpretable format used to generate a graphical and/or language-based 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.

Best practices for dataset description

Minimally the following metadata elements should be present in dataset descriptions:

- **Dataset Identifier**: A machine-actionable identifier resolvable on the Web to the dataset.
- **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**: The organization and/or contact who published the dataset and is responsible for its persistence.
- **PublicationDate/Year/ReleaseDate**: ISO 8601 standard dates are preferred (*Klyne & Newman, 2002*).
- **Version**: The dataset version identifier (if applicable).

Additional recommended metadata elements in dataset descriptions are:

- **Creator Identifier(s)**: ORCiD[6] or other unique identifier of the individual creator(s).
- **License**: The license or waiver under which access to the content is provided (preferably a link to standard license/waiver text (e.g. https://creativecommons.org/publicdomain/zero/1.0/).

[6] ORCiD IDs are numbers identifying individual researchers issued by a consortium of prominent academic publishers and others (*Editors, 2010*; *Maunsell, 2014*).

When multiple datasets are available on one landing page, licensing information may be grouped for all relevant datasets.

A World Wide Web Consortium (http://www.w3.org) standard for machine-accessible dataset description on the Web is the W3C Data Catalog Vocabulary (DCAT, *Mali, Erickson & Archer, 2014*). It was developed at the Digital Enterprise Research Institute and later standardized by the W3C eGovernment Working Group, with broad participation, and underlies some other data interoperability models such as (*DCAT Application Profile Working Group, 2013*) and (*Gray et al., 2014*).

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 that we suggest for provisional use.

Data in the described datasets might also be described using other formats depending on the application area. Other possible approaches for dataset description include DataCite metadata (*DataCite Metadata Working Group, 2014*), Dublin Core (*Dublin Core Metadata Initiative, 2012*), the Data Documentation Initiative (DDI) (*Data Documentation Initiative, 2012*) for social sciences, or ISO19115 (*ISO/TC 211, 2014*) for Geographic information. Where any of these formats are used they should support at least the minimal set of recommended metadata elements described above.

Serving the landing pages

The URIs used as identifiers for citation should resolve to HTML landing pages with the appropriate metadata in a human readable form. To enable automated agents to extract the metadata these landing pages should include an HTML `<link>` element specifying a machine readable form of the page as an alternative. For those that are capable of doing so, we recommend also using Web Linking (*Nottingham, 2010*) to provide this information from all of the alternative formats.

Should content management systems be developed specifically for maintaining and serving landing pages, we recommend both of these solutions plus the use of content negotiation (*Holtzman & Mutz, 1998*).

A more detailed discussion of these techniques and our justification for using multiple solutions is included in the Appendix. Note that in all of these cases, the alternates are other forms of the landing page. Access to the data itself should be indicated through the DCAT fields `accessURL` or `downloadURL` as appropriate for the data. Data that is spread across multiple files can be indicated by linking to an ORE resource map (*Lagoze & Van de Sompel, 2007*).

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. It is closely related to the question of persistent identifiers, that is, the identifiers must always resolve *somewhere*, and as noted above, this should be to a landing page.

But in the widest sense, persistence is a matter of service guarantees. Organizations providing trusted repositories for citable data need to detail their persistence policies transparently to users. 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.

IMPLEMENTATION: STAKEHOLDER RESPONSIBILITIES

Research communications are made possible by an ecosystem of stakeholders who prepare, edit, publish, archive, fund, and consume them. Each stakeholder group endorsing the JDDCP has, we believe, certain responsibilities regarding implementation of these recommendations. They will not all be implemented at once, or homogeneously. But careful adherence to these guidelines and responsibilities will provide a basis for achieving the goals of uniform scholarly data citation.

1. Archives and repositories: (a) Identifiers, (b) resolution behavior, (c) landing page metadata elements, (d) dataset description and (e) data access methods, should all conform to the technical recommendations in this article.
2. Registries: Registries of data repositories such as databib (http://databib.org) and r3data (http://www.re3data.org) should document repository conformance to these recommendations as part of their registration process, and should make this information readily available to researchers and the public. This also applies to lists of "recommended" repositories maintained by publishers, such as those maintained by *Nature Scientific Data* (http://www.nature.com/sdata/data-policies/repositories) and *F1000Research* (http://f1000research.com/for-authors/data-guidelines).
3. Researchers: Researchers should treat their original data as first-class research objects. They should ensure it is deposited in an archive that adheres to the practices described

here. We also encourage authors to publish preferentially with journals which implement these practices.

4. Funding agencies: Agencies and philanthropies funding research should require that recipients of funding follow the guidelines applicable to them.
5. Scholarly societies: Scholarly societies should strongly encourage adoption of these practices by their members and by publications that they oversee.
6. Academic institutions: Academic institutions should strongly encourage adoption of these practices by researchers appointed to them and should ensure that any institutional repositories they support also apply the practices relevant to them.

CONCLUSION

These guidelines, together with the NISO JATS 1.1d2 XML schema for article publishing (*National Center for Biotechnology Information, 2014*), provide a working technical basis for implementing the Joint Data Citation Principles. They were developed by a cross-disciplinary group hosted by the Force11.org digital scholarship community. [7] Data Citation Implementation Group (DCIG, https://www.force11.org/datacitationimplementation), during 2014, as a follow-on project to the successfully concluded Joint Data Citation Principles effort.

[7] 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.

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 ongoing note of repository compliance to these guidelines, and provide compliance checklists.

We are aware that some journals are already citing data in persistent public repositories, and yet not all of these repositories currently meet the guidelines we present here. Compliance will be an incremental improvement task.

Other deliverables from the DCIG are planned for release in early 2015, including a review of selected data-citation workflows from early-adopter publishers (Nature, Biomed Central, Wiley and Faculty of 1000). The NISO-JATS version 1.1d2 revision is now considered a stable release by the JATS Standing Committee, and is under final review by the National Information Standards Organization (NISO) for approval as the updated ANSI/NISO Z39.96-2012 standard. We believe it is safe for publishers to use the 1.1d2 revision for data citation now. A forthcoming article in this series will describe the JATS revisions in detail.

We 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.

Integrity of scholarly data is not a private matter, but is fundamental to the validity of published research. If data are not robustly preserved and accessible, the foundations of published research claims based upon them are not verifiable. As these practices and guidelines are increasingly adopted, it will no longer be acceptable to credibly assert any

claims whatsoever that are not based upon robustly archived, identified, searchable and accessible data.

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

ACKNOWLEDGEMENTS

We are particularly grateful to PeerJ Academic Editor Harry Hochheiser (University of Pittsburgh), reviewer Tim Vines (University of British Columbia), and two anonymous reviewers, for their careful, very helpful, and exceptionally timely comments on the first version of this article. Many thanks as well to Virginia Clark (Université Paul Sabatier), John Kunze (California Digital Library) and Maryann Martone (University of California at San Diego) for their thoughtful suggestions on content and presentation.

APPENDIX

Serving landing pages: implementation details

Ideally, all versions of the landing page would be resolvable from a single URI through content negotiation (*Holtzman & Mutz, 1998*), serving an HTML representation for humans and the appropriate form for automated agents. In its simplest form, content negotiation uses the HTTP `Accept` and/or `Accept-Language` headers to vary the content returned based on media type (a.k.a. MIME type) and language. ARK-style inflections propose an alternate way to retrieve machine-readable metadata without requiring content negotiation.

Some web servers have provision to serve alternate documents by using file names that only vary by extension; when the document is requested without an extension, the web server returns the file highest rated by the request's `Accept` header. Enabling this feature typically requires the intervention of the web server administrator and thus may not be available to all publishers.

The content negotiation standard also allows servers to assign arbitrary tags to documents and for user agents to request documents that match a given tag using the `Accept-Features` header. This could allow for selection between documents that use the same media type but use different metadata standards.

Although we believe that content negotiation is the best long-term solution to make it easier to provide for automated agents, this may require building systems to manage landing page content or adapting existing content management systems (CMS). For a near-term solution, we recommend web linking (*Nottingham, 2010*).

Web linking requires assigning a separate resolvable URI for each variant representation of the landing page. As each alternative has a URI, the documents can be cached reliably without requiring additional requests to the server hosting the landing pages. Web linking also allows additional relationships to be defined, so that it can also be used to direct automated agents to landing pages for related data as well as alternatives. Web linking also allows for a title to be assigned to each link, should they be presented to a human:

```
Link: “uri-to-an-alternate” rel=“alternate”
      media=“application/xml” title=“title”
```

We recommend including in the title the common names of the metadata schema(s) used, such as DataCite or DCAT, to allow automated agents to select the appropriate alternative.

As an additional fallback, we also recommend using HTML <link> elements to duplicate the linking information in the HTML version of the landing page:

```
<link href=“uri-to-an-alternate”;rel=“alternate”;
      media=“application/xml”;title=“title”>
```

Embedding the information in the HTML has the added benefit of keeping the alternate information attached if the landing page is downloaded from a standard web browser. This is not the case for web linking through HTTP headers, nor for content negotiation. In addition, content negotiation may not send back the full list of alternatives without the user agent sending a `Negotiate: vlist` header (*Shepherd et al., 2014*).

As each of the three techniques have points where they have advantages over the others we recommend a combination of the three approaches for maximum benefit, but acknowledge that some may take more effort to implement.

Serving landing pages: linking to the data

Note that the content being negotiated is the metadata description of the research data. The data being described should not be served via this description URI. Instead, the landing page data descriptions should reference the data.

If the data is available from a single file, directly available on the internet, use the DCAT `downloadURL` to indicate the location of the data.

If the data is available as a relatively small number of files, either as parts of the whole collection, mirrored at multiple locations, or as multiple packaged forms, link to an ORE resource map (*Lagoze et al., 2008*) to describe the relationships between the files.

If the data requires authentication to access, use the DCAT `accessURL` to indicate a page with instructions on how to request access to the data. This technique can also be used to describe the procedures on accessing physical samples or other non-digital data.

If the data is available online but is excessive in volume, use the DCAT `accessURL` to link to the appropriate search system to access the data.

For data systems that are available either as bulk downloads or through sub-setting services, include both `accessURL` and `downloadURL` on the landing page.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was funded in part by generous grants from the US National Institutes of Health and National Aeronautics and Space Administration, the Alfred P. Sloan Foundation, and the European Union (FP7). Support from the National Institutes of Health (NIH)

was provided via grant # NIH 1U54AI117925-01 in the Big Data to Knowledge program, supporting the Center for Expanded Data Annotation and Retrieval (CEDAR). Support from the National Aeronautics and Space Administration (NASA) was provided under Contract NNG13HQ04C for the Continued Operation of the Socioeconomic Data and Applications Center (SEDAC). Support from The Alfred P. Sloan Foundation was provided under two grants: a. Grant # 2012-3-23 to the Harvard Institute for Quantitative Social Sciences, "Helping Journals to Upgrade Data Publication for Reusable Research"; and b. a grant to the California Digital Library, "CLIR/DLF Postdoctoral Fellowship in Data Curation for the Sciences and Social Sciences". The European Union partially supported this work under the FP7 contracts #269977 supporting the Alliance for Permanent Access and #269940 supporting Digital Preservation for Timeless Business Processes and Services. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Institutes of Health (NIH): # NIH 1U54AI117925-01.
Alfred P. Sloan Foundation: #2012-3-23.
European Union (FP7): #269977, #269940.
National Aeronautics and Space Administration (NASA): NNG13HQ04C.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Joan Starr and Tim Clark conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, performed the computation work, reviewed drafts of the paper.
- Eleni Castro, Mercè Crosas, Michel Dumontier, Robert R. Downs, Ruth Duerr, Laurel L. Haak, Melissa Haendel, Ivan Herman, Simon Hodson, Joe Hourclé, John Ernest Kratz, Jennifer Lin, Lars Holm Nielsen, Amy Nurnberger, Stefan Proell, Andreas Rauber, Simone Sacchi, Arthur Smith and Mike Taylor performed the experiments, analyzed the data, performed the computation work, reviewed drafts of the paper.

REFERENCES

Accomazzi A, Henneken E, Erdmann C, Rots A. 2012. Telescope bibliographies: an essential component of archival data management and operations. In: *Society of Photo-Optical Instrumentation Engineers (SPIE) conference series*. vol. 8448. Article id 84480K, 10 pp DOI 10.1117/12.927262.

Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA. 2011. Public availability of published research data in high-impact journals. *PLoS ONE* **6(9)**:e24357 DOI 10.1371/journal.pone.0024357.

Altman M, Crosas M. 2013. The evolution of data citation: from principles to implementation. *IAssist Quarterly* **(Spring)**:62–70. *Available at http://www.iassistdata.org/iq/evolution-data-citation-principles-implementation*.

Altman M, King G. 2006. A proposed standard for the scholarly citation of quantitative data. *DLib Magazine* **13(3/4)**. *Available at http://www.dlib.org/dlib/march07/altman/03altman.html*.

Ball A, Duke M. 2012. How to cite datasets and link to publications. Technical report. DataCite. *Available at http://www.dcc.ac.uk/resources/how-guides*.

Begley CG, Ellis LM. 2012. Drug development: raise standards for preclinical cancer research. *Nature* **483(7391)**:531–533 DOI 10.1038/483531a.

Berners-Lee T, Fielding R, Masinter L. 1998. RFC2396: Uniform resource identifiers (URI): generic syntax. *Available at https://www.ietf.org/rfc/rfc2396.txt*.

Booth D, Haas H, McCabe F, Newcomer E, Champion M, Ferris C, Orchard D. 2004. Web services architecture: W3C working group note 11 February 2004. Technical Report. World Wide Web Consortium. *Available at http://www.w3.org/TR/ws-arch/*.

Borgman C. 2012. Why are the attribution and citation of scientific data important? In: Uhlir P, ed. *For attribution—Developing Data Attribution and Citation Practices and Standards. Summary of an international Workshop*. Washington D.C.: National Academies Press.

Bray T, Paoli J, Sperberg-McQueen CM, Maler E, Yergeau F. 2008. Extensible markup language (XML) 1.0 (fifth edition): W3C recommendation 26 November 2008. *Available at http://www.w3.org/TR/REC-xml/*.

Clark A, Evans P, Strollo A. 2014. FDSN recommendations for seismic network DOIs and related FDSN services, version 1.0. Technical report. International Federation of Digital Seismograph Networks. *Available at http://www.fdsn.org/wgIII/V1.0-21Jul2014-DOIFDSN.pdf*.

CNRI. 2014. Handle system: unique and persistent identifiers for internet resources. *Available at http://www.w3.org/TR/webarch/#identification*.

CODATA-ICSTI Task Group on Data Citation Standards and Practices. 2013. Out of cite, out of mind: the current state of practice, policy and technology for data citation. *Data Science Journal* **12(September)**:1–75 DOI 10.2481/dsj.OSOM13-043.

Colquhoun D. 2014. An investigation of the false discovery rate and the misinterpretation of *p*-values. *Royal Society Open Science* **1(3)**:140216 DOI 10.1098/rsos.140216.

Data Citation Synthesis Group. 2014. Joint declaration of data citation principles. *Available at http://force11.org/datacitation*.

Data Documentation Initiative. 2012. Data documentation initiative specification. *Available at http://www.ddialliance.org/Specification/*.

DataCite Metadata Working Group. 2014. Datacite metadata schema for the publication and citation of research data, version 3.1 October 2014. *Available at http://schema.datacite.org/meta/kernel-3.1/doc/DataCite-MetadataKernel_v3.1.pdf*.

DCAT Application Profile Working Group. 2013. DCAT application profile for data portals in Europe. *Available at https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-application-profile-data-portals-europe-final*.

Dublin Core Metadata Initiative. 2012. Dublin core metadata element set, version 1.1. *Available at http://dublincore.org/documents/dces/*.

Dyson E. 2003. Online registries: the DNS and beyond. *Available at http://doi.contentdirections.com/reprints/dyson_excerpt.pdf*.

ECMA. 2013. ECMA-404: the JSON data interchange format. *Available at http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf*.

Editors. 2010. Credit where credit is due. *Nature* **462(7275)**:825 DOI 10.1038/462825a.

Fielding RT. 2000. Architectural styles and the design of network-based software architectures. Doctoral dissertation, University of California at Irvine. *Available at https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm*.

Fielding RT, Taylor RN. 2002. Principled design of the modern web architecture. *ACM Transactions on Internet Technology* **2(2)**:115–150 DOI 10.1145/514183.514185.

Gao S, Sperberg-McQueen CM, Thompson HS. 2012. W3C XML schema definition language (XSD) 1.1 part 1: structures: W3C recommendation 5 April 2012. *Available at http://www.w3.org/TR/xmlschema11-1/*.

GitHub Guides. 2014. Making your code citable. *Available at https://guides.github.com/activities/citable-code/*.

Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, Di Stefano R, Gil Y, Groth P, Hedstrom M, Hogg DW, Kashyap V, Mahabal A, Siemiginowska A, Slavkovic A. 2014. Ten simple rules for the care and feeding of scientific data. *PLoS Computational Biology* **10(4)**:e1003542 DOI 10.1371/journal.pcbi.1003542.

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, Rietveld L. 2014. Data catalog vocabulary (DCAT): W3C recommendation 16 January 2014. *Available at http://www.w3.org/2001/sw/hcls/notes/hcls-dataset/*.

Greenberg SA. 2009. How citation distortions create unfounded authority: analysis of a citation network. *BMJ* **339**:b2680 DOI 10.1136/bmj.b2680.

Gudgin M, Hadley M, Mendelsohn N, Moreau J-J, Nielsen HF, Karmarkar A, Lafon Y. 2007. SOAP version 1.2 part 1: messaging framework (second edition): W3C recommendation 27 April 2007. *Available at http://www.w3.org/TR/soap12-part1/*.

Haas H, Brown A. 2004. Web services glossary: W3C working group note 11 February 2004. *Available at http://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/#webservice*.

Hakala J. 2001. RFC3188: using national bibliography numbers as uniform resource names. *Available at https://tools.ietf.org/html/rfc3188*.

Hilse H-W, Kothe J. 2006. Implementing persistent identifiers. *Available at http://xml.coverpages.org/ECPA-PersistentIdentifiers.pdf*.

Holtzman K, Mutz A. 1998. RFC2295: transparent content negotiation in HTTP. *Available at https://www.ietf.org/rfc/rfc2295.txt*.

Hourclé J, Chang W, Linares F, Palanisamy G, Wilson B. 2012. Linking articles to data. In: *3rd ASIS&T Summit on Research Data Access & Preservation (RDAP) New Orleans, LA, USA*. *Available at http://vso1.nascom.nasa.gov/rdap/RDAP2012_landingpages_handout.pdf*.

International DOI Foundation. 2014. DOI handbook. *Available at http://www.doi.org/hb.html*.

Ioannidis JPA. 2005. Why most published research findings are false. *PLoS Medicine* **2(8)**:e124 DOI 10.1371/journal.pmed.0020124.

ISO/TC 211. 2014. ISO 19115-1:2014: geographic information metadata, part 1: fundamentals. *Available at http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53798*.

Jacobs I, Walsh N. 2004. Architecture of the world wide web, volume one W3C recommendation 15 December 2004. *Available at http://www.w3.org/TR/webarch/#identification*.

Janée G, Kunze J, Starr J. 2009. Identifiers made easy. *Available at http://ezid.cdlib.org/*.

Juty N, Le Novére N, Laibe C. 2012. Identifiers.org and MIRIAM registry: community resources to provide persistent identification. *Nucleic Acids Research* **40(D1)**:D580–D586 DOI 10.1093/nar/gkr1097.

Klyne G, Newman C. 2002. RFC3339: date and time on the internet: timestamps. *Available at http://www.ietf.org/rfc/rfc3339.txt*.

Kunze J. 2003. Towards electronic persistence using ARK identifiers. In: *Proceedings of the 3rd ECDL workshop on web archives*. Trondheim, Norway, *Available at https://confluence.ucop.edu/download/attachments/16744455/arkcdl.pdf*.

Kunze J. 2012. The ARK identifier scheme at ten years old. In: *Workshop on metadata and persistent identifiers for social and economic data, Berlin. Available at http://www.slideshare.net/jakkbl/the-ark-identifier-scheme-at-ten-years-old*.

Kunze J, Rodgers R. 2013. The ARK identifier scheme. Technical report. Internet Engineering Task Force. *Available at https://tools.ietf.org/html/draft-kunze-ark-18*.

Kunze J, Starr J. 2006. ARK (archival resource key) identifiers. *Available at http://www.cdlib.org/inside/diglib/ark/arkcdl.pdf*.

Lagoze C, Van de Sompel H. 2007. Compound information objects: the OAI-ORE perspective. Open Archives Initiative – Object Reuse and Exchange. *Available at http://www.openarchives.org/ore/documents/CompoundObjects-200705.html*.

Lagoze C, Van de Sompel H, Johnston P, Nelson M, Sanderson R, Warner S. 2008. ORE user guide—resource map discovery. *Available at http://www.openarchives.org/ore/1.0/discovery*.

Library of Congress. 1997. The relationship between URNs, Handles, and PURLs. *Available at http://memory.loc.gov/ammem/award/docs/PURL-handle.html*.

Mali F, Erickson J, Archer P. 2014. Data catalog vocabulary (dcat): W3C recommendation 16 January 2014. *Available at http://www.w3.org/TR/vocab-dcat/*.

Maunsell JH. 2014. Unique identifiers for authors. *The Journal of Neuroscience* **34(21)**:7043 DOI 10.1523/JNEUROSCI.1670-14.2014.

Moats R. 1997. RFC2141: uniform resource name syntax. *Available at https://tools.ietf.org/html/rfc2141*.

National Center for Biotechnology Information. 2014. *Available at http://jats.nlm.nih.gov/publishing/tag-library/1.1d2/index.html*.

Nottingham M. 2010. RFC5988: web linking. *Available at https://www.ietf.org/rfc/rfc5988.txt*.

OCLC. 2015. Purl help. *Available at https://purl.org/docs/help.html* (accessed 2 January 2015).

Parsons MA, Duerr R, Minster J-B. 2010. Data citation and peer review. *Available at http://dx.doi.org/10.1029/2010EO340001*.

Peterson D, Gao S, Malhotra A, Sperberg-McQueen CM, Thompson HS. 2012. W3C XML schema definition language (XSD) 1.1 part 2: datatypes: W3C recommendation 5 April 2012. *Available at http://www.w3.org/TR/xmlschema11-1/*.

Peyrard S, Kunze J, Tramoni J-P. 2014. The ARK identifier scheme: lessons learnt at the BNF. In: *Proceedings of the international conference on Dublin core and metadata applications 2014. Available at http://dcpapers.dublincore.org/pubs/article/view/3704/1927*.

Prinz F, Schlange T, Asadullah K. 2011. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* **10(9)**:712–713 DOI 10.1038/nrd3439-c1.

Rans J, Day M, Duke M, Ball A. 2013. Enabling the citation of datasets generated through public health research. *Available at http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp051762.PDF*.

Rekdal OB. 2014. Academic urban legends. *Social Studies of Science* **44(4)**:638–654 DOI 10.1177/0306312714535679.

Richardson L, Ruby S. 2011. *RESTful web services*. Sebastopol CA: O'Reilly.

Salzberg SL, Pop M. 2008. Bioinformatics challenges of new sequencing technology. *Trends in Genetics* **24**:142–149 DOI 10.1016/j.tig.2007.12.006.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nature Biotechnology* **26**:1135–1145 DOI 10.1038/nbt1486.

Shepherd, Fiumara, Walters, Stanton, Swisher, Lu, Teoli, Kantor, Smith. 2014. Content negotiation. Mozilla developer network. *Available at https://developer.mozilla.org/docs/Web/HTTP/Content_negotiation*.

Stein L. 2010. The case for cloud computing in genome informatics. *Genome Biology* **11(5)**:207–213 DOI 10.1186/gb-2010-11-5-207.

Strasser B. 2010. Collecting, comparing, and computing sequences: the making of Margaret O. Dayhoff's atlas of protein sequence and structure, 1954–1965. *Journal of the History of Biology* **43(4)**:623–660 DOI 10.1007/s10739-009-9221-0.

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. *Available at http://www.nap.edu/openbook.php?record_id=13564*.

Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, LaRocca GM, Haendel MA. 2013. On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ* **1**:e148 DOI 10.7717/peerj.148.

Citing Data in Journal Articles using JATS

Deborah Aleyne Lapeyre
Mulberry Technologies, Inc.
17 West Jefferson Street, Suite 207
Rockville, MD 20850
Phone: 301/315-9631
Fax: 301/315-8285
info@mulberrytech.com
http://www.mulberrytech.com

Version 1.0 (June 2015)

Citing Data in Journal Articles using JATS

Appendixes

slide 1

JATS: The Journal Article Tag Suite

- The article publishing piece of the data citing story
- JATS enables publishers to cite data sources *in journal articles*
- Tagging allows:
  - human readability
  - machine discoverability
  - flexibility to express different types of data citations

slide 2

JATS (ANSI/NISO Z39-96-2012) is

- XML for tagging journal articles
- Used by:
  - STM journal publishers (production tag set and/or interchange)
    (US, England, Japan, Korea, Australia, Canada, Brazil, China, Germany, Norway, Sweden, Switzerland, France, Croatia, Russia, Belgium, Egypt, Oman, United Arab Emirates, etc.)
  - National Libraries (US, UK, Australia)
  - Archives (PubMed Central, JSTORE/ITHAKA)
  - Aggregators and web-hosts (Highwire, Silverchair, Atypon)
  - Standards bodies to produce standards (ISO, IEEE)

slide 3

JATS Names XML Elements for Publishing

- JATS available in DTD, XSD, and RNG XML model formats
- The Tag Set names and describes the content of:
  - metadata elements (contributor, surname, abstract)
  - textual elements (paragraph, figure, verse)
  - tables (XHTML and OASIS models)
  - elements for math (MathML 2.0 or 3.0)
  - bibliographic reference elements (article title, publisher, publication year)

slide 4

How Publishers Cite Data

- In the narrative text
- In the bibliography (references list)
- In an additional reference list just for data

Force11 recommends tagging them as references are tagged

slide 5

How JATS Tags References

- Bibliographic reference lists (<ref-list>) are in the back of:
  - articles
  - sections
  - boxed-text
- Reference lists contain references (<ref>)
- References contain citations (<mixed-citation>)
  each of which contains the description of one cited source

slide 6

Force11 Recommends JATS Mixed Citation

<mixed-citation> is

- a bag-of-text with all punctuation and spacing preserved
- some elements inside can be tagged
- how much tagging is up to the publisher

Lapeyre, Deborah Aleyne, *Poodles of the World*. Journal of Big Dogs, 2015 vol: 13, pages: 2525-2535 DOI: 10.1165/JCM.02419-05

```
<ref id="B45">
<mixed-citation publication-type="journal">
<string-name>
  <surname>Lapeyre</surname>,
  <given-names>Deborah Aleyne</given-names>
</string-name>,
<article-title>Poodles of the World</article-title>.
<source>Journal of Big Dogs</source>,
<year>2015</year> vol: <volume>13</volume>,
pages: <fpage>2525</fpage>-<lpage>2535</lpage> DOI:
<pub-id pub-id-type="doi">10.1165/JCM.02419-05</pub-id>
</mixed-citation>
</ref>
```

slide 7

What is needed to Cite Data?

- Best practices for dataset description
  (what an archive should keep)
- Data citing recommendations from style guides, publishers, archives, researchers, consortia

Dataset Description Metadata

(for deposit to an archive)

Force11 minimum elements that should be present in a dataset description:

1. Dataset Identifier
2. Title of the dataset
3. Creator
4. Publisher/Contact
5. Publication Date/ Year / Release Date
6. Version of the dataset
7. *Description (longer explanation than the title)*

Items 1-6 can/should be part of a bibliographic citation

slide 9

How Publishers Want Data Cited

- Over 55 sources were polled on what data fields to use to cite data
- Here are the top 10 (mentioned by most, mandatory in many)

1. Persistent global dataset Identifier
2. Title/Name of the dataset
3. Author/Creator
4. Publisher/Distributor/Repository
5. Publication Date / Year / Release Date
6. Version of the dataset
7. Resource Type
8. Location of publisher/distributor
9. Access date and time
10. Additional URI/location/bridge service

slide 10

New JATS Elements Requested by Force11

- <data-title>
  - the formal title or name of a cited data source (or a component of a cited data source)
  - equivalent to <article-title>
  - may be used with <source> for hierarchical relationships</source>
- <version>
  - full version statement (maybe only a number) for cited data or software
  - @designator attribute can hold the simple version number:

```
<version designator="16.2">16th version, second release</version>
```

JATS Elements for Citing Data (1)

1. Persistent Global Identifier
   - <pub-id pub-id-type='doi'>
2. Title/Name of the dataset
   - <data-title> (similar to <article-title>)
   - <source>
3. Author/Creator
   - <name> or <string-name>
   - <collab>
4. Publisher/Distributor/Repository
   - <publisher>
5. Publication Date / Year / Release Date
   - <date>
   - <year>

(See Appendix 2 for more complete mappings)

JATS Elements for Citing Data (2)

1. Version of the dataset
   - <version>
   - <edition>
   - <date-in-citation content-type="update">
2. Resource Type
   - @publication-format
     (print, electronic, video, audio, ebook, online-only)
3. Location of publisher/distributor
   - <publisher-loc>
4. Access date and time
   - <date-in-citation content-type="access-date">
   - <year>
5. Additional URI/location/bridge service
   - <ext-link>
   - <uri>

slide 13

New Attributes Values (1)

- @publication-type on citations
  - typically “`book`”, “`journal`”, “`standard`”
  - new value “`data`”
  - defined as “a dataset or other research collection such as a spreadsheet”
- @person-group-type on <person-group>
  - typically “`author`”, “`editor`”, “`compiler`”
  - new value “`curator`”
  - used for citing datasets and art

slide 14

New Attributes and Values for @pub-id

- New attribute @assigning-authority
  - says who assigned the ID (such as an ARK or DOI)
  - values are organizations such as “crossref”, “figshare”, “pdb”, “genbank”, “pubmed”
- The @pub-id-type (“doi”, “archive”, “isbn”) gets new values for citing data:
  - “`accession`” (Bioinformatics: a unique identifier given to a DNA or protein sequence record for tracking the sequence record and the associated sequence over time in a data repository.)
  - “`ark`” (Archival Resource Key: a Uniform Resource Locator (URL) containing the word "ark" that is a multi-purpose identifier for information objects of any type)
  - “`handle`” (HDL: Handle identifier, part of the Handle System for assigning, managing, and resolving persistent identifiers for digital objects and other resources on the Internet)

slide 15

Machine Resolvable Problem Not Solved

- JATS enables; it does ***not enforce***
- JATS was designed for interchange among:
  - publishers and their partners
  - archives and libraries
  - aggregators and hosting services
- There is no one right way to cite data
  - different publishers different styles
  - *how much* to record is a buiness decision

slide 16

What Else is Needed?

Data miners and machine resolvers need

- As much uniformity as possible
- Common agreements
- Best practices

Force11 and JATS4R (JATS for Reuse: `http://jats4r.org`)

slide 17

Data Citation Examples

As we have time and desire to geek

With thanks to Daniel Mietchen, Johanna McEntyre, Jeff Beck, Chris Maloney, and the Force11 Data Citation Implementation Group

slide 18

Dryad Digital Repository, referenced through a DOI

Dubuis JO, Samanta R, Gregor T (2013). Data from: *Accurate measurements of dynamics and reproducibility in small genetic networks*. Dryad Digital Repository doi:10.5061/dryad.35h8v

```
<mixed-citation publication-type="data">Dubuis JO, Samanta R,
Gregor T (<year iso-8601-date="2013">2013</year>).  Data from:
<data-title>Accurate measurements of dynamics and reproducibility
in small genetic networks</data-title>. <source>Dryad Digital
Repository</source> doi:<pub-id pub-id-type="doi">10.5061/dryad.35h8v</pub-id>
</mixed-citation>
```

slide 19

GenBank Protein

Homo sapiens cAMP responsive element binding protein 1 (CREB1), transcript variant A, mRNA. GenBank NM_004379.3.

```
<mixed-citation publication-type="data">
<data-title>Homo sapiens cAMP responsive element binding protein 1
(CREB1), transcript variant A, mRNA</data-title>. <source>GenBank</source>
<ext-link ext-link-type="genbank" xlink:href="NM_004379.3">NM_004379.3</ext-
link>.
</mixed-citation>
```

slide 20

RNA Sequence

Xu, J. et al. *Cross-platform ultradeep transcriptomic profiling of human reference RNA samples by RNA-Seq*. Sci. Data 1:140020 doi: 10.1038/sdata.2014.20 (2014).

```
<mixed-citation publication-type="data">Xu, J. <etal/>
<data-title>Cross-platform ultradeep transcriptomic profiling
of human reference RNA samples by RNA-Seq</data-title>.
<source>Sci. Data</source> <volume>1</volume>:
<elocation-id>140020</elocation-id>
doi: <pub-id pub-id-type="doi">10.1038/sdata.2014.20</pub-id>
(<year iso-8601-date="2014">2014</year>).
</mixed-citation>
```

slide 21

Protein Data Bank in Europe sample

Kollman JM, Charles EJ, Hansen JM, 2014, *Cryo-EM structure of the CTP synthetase filament*, http://www.ebi.ac.uk/pdbe/entry/EMD-2700, Publicly available from The Electron Microscopy Data Bank (EMDB).

```
<mixed-citation publication-type="data">Kollman JM, Charles EJ, Hansen JM,
<year iso-8601-date="2014">2014</year>, <data-title>Cryo-EM structure of
the CTP synthetase filament</data-title>, <ext-link ext-link-type="uri"
xlink:href="http://www.ebi.ac.uk/pdbe/entry/EMD-2700">
http://www.ebi.ac.uk/pdbe/entry/EMD-2700</ext-link>, Publicly available
from <source>The Electron Microscopy Data Bank (EMDB)</source>.
</mixed-citation>
```

Data in figshare, referenced through a DOI

Mulvany, Ian, *citing-dataset-elements*. FigShare, 2014/06/30, 10.6084/m9.figshare.1088363.

```
<mixed-citation publication-type="data">
<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'>
<year>2014</year>/<month>06</month>/<day>30</day></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>
```

Di Stefano B, Collombet S, Graf T. Figshare http://dx.doi.org/10.6084/m9.figshare.939408 (2014).

```
<mixed-citation publication-type="data">Di Stefano B, Collombet S,
Graf T. <source>Figshare</source> <ext-link ext-link-type="uri"
xlink:href="http://dx.doi.org/10.6084/m9.figshare.939408">
http://dx.doi.org/10.6084/m9.figshare.939408</ext-link>
(<year iso-8601-date="2014">2014</year>).
</mixed-citation>
```

Data Curator

The value “curator” was added to the list of suggested values for the @person-group-type attribute. Here is an example of how the “curator” value might be used for @person-group-type:

Frankis, Michael, curator. "Mountain bluebird." *Encyclopedia of Life*, available from `http://eol.org/pages/1177542`. Accessed 30 Mar 2015.

```
<mixed-citation publication-type="data">
<person-group person-group-type='curator'>
<name><surname>Frankis</surname><given-names>Michael</given-names></name>
</person-group>, curator. "<data-title>Mountain bluebird</data-title>."
<source>Encyclopedia of Life</source>, available from
<ext-link ext-link-type='uri' xlink:href='http://eol.org/pages/1177542'>
http://eol.org/pages/1177542</ext-link>. Accessed
<date-in-citation  content-type="access-date"
iso-8601-date="2015-03-30">30 Mar 201</date-in-citation>.
</mixed-citation>
```

Assigning Authority

A new attribute @assigning-authority was added to the elements <ext-link> and <pub-id>. The existing attribute @pub-id-type should now only be used to state how the element content is to be interpreted as an identifier. For example, a "DOI" would have the @pub-id-type attribute value of "doi", and the @assigning-authority attribute value might be "crossref" or "figshare". (Note that values are in lowercase for both attributes!) Another example from the life sciences would be: @pub-id-type value of "accession", @assigning-authority of "uniprot".

Mulvany, Ian, *citing-dataset-elements*. Figshare, 2014/06/30, `10.6084/m9.figshare.1088363`.

```
<mixed-citation publication-type="data">
<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'>
<year>2014</year>/<month>06</month>/<day>30</day></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>
```

slide 25

New @pub-id-type Values

New values for the @pub-id-type attribute ("accession", "ark", and "handle") were added to JATS for tagging data sources.

Heinz D.W., Baase W.A., et al. *How amino-acid insertions are allowed in an alpha-helix of T4 lysozyme*. RCSB Protein Data Bank, accession 102l. 10.2210/pdb102l/pdb

```
<mixed-citation publication-type='data'>
<name><surname>Heinz</surname><given-names>D.W.</given-names></name>,
<name><surname>Baase</surname><given-names>W.A.</given-names></name>,
<etal>et al.</etal> <data-title>How amino-acid insertions are allowed in
an alpha-helix of T4 lysozyme</data-title>.
<source>RCSB Protein Data Bank</source>, accession
<pub-id pub-id-type='accession' assigning-authority='pdb'
xlink:href='http://www.rcsb.org/pdb/explore/explore.do?structureId=102l'>102l</
pub-id>.
<pub-id pub-id-type='doi' xlink:href='http://dx.doi.org/10.2210/pdb102l/pdb'>
10.2210/pdb102l/pdb</pub-id>
</mixed-citation>
```

slide 26

External Media: Database on CD-ROM, DVD, or Disk

Walker MM, Keith LH. EPA's Clean Air Act air toxics database [disk]. Boca Raton (FL): Lewis Publishers; 1992-1993. 4 computer disks: 3 1/2 in.

```
<mixed-citation publication-type="data" publication-format="disk">
<name><surname>Walker</surname><given-names>MM</given-names></name>,
<name><surname>Keith</surname><given-names>LH</given-names></name>.
<data-title>EPA's Clean Air Act air toxics database</data-title> [disk].
<publisher-loc>Boca Raton (FL)</publisher-loc>: <publisher-name>Lewis Publish-
ers</publisher-name>;
<date-in-citation content-type="copyright-year"
iso-8601-date="1992">1992-1993</date-in-citation>.
4 computer disks: 3 1/2 in.</mixed-citation>
```

slide 27

Record from a Web Data Repository

Benz, Michael; Braband, Henrik; Schmutz, Paul; Halter, Jonathan; Alberto, Roger. *C21 H49 Al Cl7 N7 O7 Tc*, version 130981. From Crystallography Open Database, accession 1517518.

```
<mixed-citation publication-type='data'>
<name><surname>Benz</surname><given-names>Michael</given-names></name>;
<name><surname>Braband</surname><given-names>Henrik</given-names></name>;
<name><surname>Schmutz</surname><given-names>Paul</given-names></name>;
<name><surname>Halter</surname><given-names>Jonathan</given-names></name>;
<name><surname>Alberto</surname><given-names>Roger</given-names></name>.
<data-title>C21 H49 Al Cl7 N7 O7 Tc</data-title>,
version <version>130981</version>.
From <source>Crystallography Open Database</source>, accession
<pub-id pub-id-type='accession'
 assigning-authority='crystallography open database'
 xlink:href='http://www.crystallography.net/cod/1517518.html'>1517518</pub-id>.
</mixed-citation>
```

slide 28

Add Health Data Set

Harris, Kathleen Mullan. 2009. *The National Longitudinal Study of Adolescent to Adult Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002; Wave IV, 2007-2009* [machine-readable data file and documentation]. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill. DOI: 10.3886/ICPSR27021.v9

```
<mixed-citation publication-type="data">
<name><surname>Harris</surname><given-names>Kathleen Mullan</given-names></
name>.
<date-in-citation content-type="pub-date"><year>2009</year></date-in-citation>.
<data-title>The National Longitudinal Study of Adolescent to Adult
Health (Add Health), Waves I &amp;amp; II, 1994–1996; Wave III,
2001–2002; Wave IV, 2007-2009</data-title>
[machine-readable data file and documentation]. <publisher-loc>Chapel Hill,
NC</publisher-loc>: <publisher-name>Carolina Population Center, University of
North Carolina at Chapel Hill</publisher-name>. DOI: <pub-id pub-id-type='doi'
xlink:href='http://dx.doi.org/10.3886/ICPSR27021.v9'>10.3886/ICPSR27021.v9</pub-
id>
</mixed-citation>
```

slide 29

GigaScience Sample

Zheng LY, Guo XS, He B, Sun LJ, Pi CM, Jing H-C: Genome data from [http://dx.doi.org/10.5524/100012] GigaScience 2011.

```
<mixed-citation publication-type="data">Zheng LY,
Guo XS, He B, Sun LJ, Pi CM, Jing H-C: Genome data from
[<ext-link ext-link-type="uri" xlink:href="http://dx.doi.org/10.5524/100012">
http://dx.doi.org/10.5524/100012</ext-link>] <source>GigaScience</source>
<year iso-8601-date="2011">2011</year>.
</mixed-citation>
```

slide 30

Colophon

- Slides and handouts created from a single XML source
- Projected in HTML (created from XML by XSLT)
- Handouts distributed in PDF
  - source XML transformed to XHTML + CSS
  - PDF from that
  - all lights out; no pagination or tables adjusted

Appendix A

Possible Elements in a JATS <mixed-citation>

A <mixed-citation> element is a bag-of-text that may contain, intermixed with the text (letters, numbers, or special characters), the following elements:

Any combination of:

- <inline-supplementary-material> Inline Supplementary Material Metadata
- Related Material Elements
  - <related-article> Related Article Information
  - <related-object> Related Object Information
- <hr> Horizontal Rule
- <string-date> Date as a String
- Emphasis Elements
  - <bold> Bold
  - <fixed-case> Fixed Case
  - <italic> Italic
  - <monospace> Monospace Text (Typewriter Text)
  - <overline> Overline
  - <overline-start> Overline Start
  - <overline-end> Overline End
  - <roman> Roman
  - <sans-serif> Sans Serif
  - <sc> Small Caps
  - <strike> Strike Through
  - <underline> Underline
  - <underline-start> Underline Start
  - <underline-end> Underline End
  - <ruby> Ruby Annotation Wrapper
- <alternatives> Alternatives For Processing
- Inline Display Elements
  - <inline-graphic> Graphic, Inline
  - <private-char> Private Character (Custom or Unicode)
- <chem-struct> Chemical Structure (Display)

- <inline-formula> Formula, Inline
- <label> Label (of an Equation, Figure, Reference, etc.)
- Math Elements
  - <tex-math> TeX Math Equation
  - <mml:math> Math (MathML Tag Set)
- Other Inline Elements
  - <abbrev> Abbreviation or Acronym
  - <milestone-end> Milestone End
  - <milestone-start> Milestone Start
  - <named-content> Named Special (Subject) Content
  - <styled-content> Styled Special (Subject) Content
- <annotation> Annotation in a Citation
- <article-title> Article Title
- <chapter-title> Chapter Title in a Citation
- <collab> Collaborative (Group) Author
- <collab-alternatives> Collaboration Alternatives
- <comment> Comment in a Citation
- <conf-acronym> Conference Acronym
- <conf-date> Conference Date
- <conf-loc> Conference Location
- <conf-name> Conference Name
- <conf-sponsor> Conference Sponsor
- <data-title> Data Title
- <date> Date
- <date-in-citation> Date within a Citation
- <day> Day
- <edition> Edition Statement, Cited
- Linking Elements
  - <email> Email Address
  - <ext-link> External Link
  - <uri> Uniform Resource Identifier (URI)
- <elocation-id> Electronic Location Identifier
- <etal> Et Al.
- <fpage> First Page

- <gov> Government Report, Cited
- <institution> Institution Name: in an Address
- <institution-wrap> Institution Wrapper
- <isbn> ISBN
- <issn> ISSN
- <issn-l> ISSN-L (Linking ISSN)
- <issue> Issue Number
- <issue-id> Issue Identifier
- <issue-part> Issue Part
- <issue-title> Issue Title
- <lpage> Last Page
- <month> Month
- <name> Name of Person
- <name-alternatives> Name Alternatives
- <object-id> Object Identifier
- <page-range> Page Ranges
- <part-title> Part Title in a Citation
- <patent> Patent Number, Cited
- <person-group> Person Group for a Cited Publication
- <pub-id> Publication Identifier for a Cited Publication
- <publisher-loc> Publisher's Location
- <publisher-name> Publisher's Name
- <role> Role or Function Title of Contributor
- <season> Season
- <series> Series
- <size> Size
- <source> Source
- <std> Standard, Cited
- <string-name> Name of Person (Unstructured)
- <supplement> Supplement Information
- <trans-source> Translated Source
- <trans-title> Translated Title
- <version> Version Statement
- <volume> Volume Number

- <volume-id> Volume Identifier
- <volume-series> Volume Series
- <year> Year
- <fn> Footnote
- <target> Target of an Internal Link
- <xref> X (cross) Reference
- Baseline Change Elements
  - <sub> Subscript
  - <sup> Superscript
- <x> X - Generated Text and Punctuation

Appendix B

Mapping Data Citing Components to JATS Elements

Prior to the June 2014 Force11 meeting, over 55 primary data sources (style guides, Archive submission guidelines, publisher's websites, schemas such as the DataCite Schema, articles on citing data by thought leaders, etc.) were reviewed to see what data fields were recommended for citing data such as genomic datasets. While dozens of data items were mentioned, most of the sources agreed on some variation of the top ten, with many making these mandatory. In the following pages, these requested data fields have been mapped to the JATS elements from JATS Committee Draft 1.1d3.

In the pages that follow:

- A numbered heading gives the data field name or names (as found in multiple sources).
- The paragraph below it will give an approximate definition. (Many definitions have been taken from ESIP Data Citation Guidelines [Ruth Duerr 2012] and the DataCite Schema documentation.)
- The bulleted item(s) that follow show JATS elements that could be used to represent this data within a citation. A tagged sample of each element is given.

1. Persistent Global Dataset Identifier / Locator / DOI / URL

Possibly a URL, but ideally a persistent identifier (DOI, PURL, Handle, ARK). The HTTP form of the DOI is preferred by some sources.

- <pub-id> with @pub-id-type

  ```
  <pub-id pub-id-type="doi">10.1128/JCM.02410-08</pub-id>
  ```

  ```
  <pub-id pub-id-type="doi">10.1099/ijs.0.039248-0</pub-id>
  ```

  Linking attributes can be added to make the non-URL-DOI a live link:

  ```
  <pub-id="doi" xlink:href="http://dx.doi.org/http://dx.doi.org/10.6070/
  H4WM1BBQ">
  10.6070/H4WM1BBQ</pub-id>
  ```

- <ext-link> with @ext-link-type

```
<ext-link-type="uri"
xlink:href="http://dx.doi.org/http://dx.doi.org/10.6070/H4WM1BBQ">
http://dx.doi.org/http://dx.doi.org/10.6070/H4WM1BBQ</ext-link>
```

- <uri>

```
<uri xlink:href="http://www.biomedcentral.com/1471-2180/13/198"/>
```

2. Title/Name of the Dataset

Formal title of the dataset (may include applicable dates). Similar to an article title in its role in the citation. Because a dataset located in a repository or inside a portion of a repository, there are two elements available. The <source> can be used to name repository levels.

- <data-title>

```
<data-title>Monitoring the Future: A Continuing Study of American Youth (12th
Grade Survey)</data-title>
```

- <source> `<source>figshare</source>` or

```
<source>Dryad Digital Repository</source>
```

3. Creator/Author/Rightsholder/Primary Responsibility

Data creators. People or organizations responsible for developing (intellectual work) the dataset. Primary Responsibility

Potential JATS Equivalents:

- ***<name>***

```
<name>
<surname>Edelstein</surname>
<given-names>PH</given-names>
</name>
```

- <string-name>

```
<string-name>
<surname>Edelstein</surname>,
<given-names>PH</given-names>
</string-name>
```

- person-group/name

```
<person-group person-group-type="author">
<name>
<surname>Edelstein</surname>
<given-names>PH</given-names>
</name>
</person-group>
```

- person-group/collab

```
<person-group person-group-type="author">
<collab collab-type="compilers">The BAC Resource Consortium</collab>
</person-group>
```

- <institution>

```
<institution content-type="university">Boston University</institution>
```

- <institution-wrap>

```
<institution-wrap>
<institution-id institution-id-type="Ringgold">1812</institution-id>
<institution content-type="university">Harvard University</institution>
</institution-wrap>

<institution-wrap>
<institution-id institution-id-type="Ringgold">1846</institution-id>
<institution-id
  institution-id-type="ISNI">0000 0001 2170 1429</institution-id>
<institution content-type="university">Boston University</institution>
</institution-wrap>
```

4. Publisher/Distributor/ Repository/ Data Center / Archive

The organization distributing and curating the data (responsible for its persistence, ideally over the long term) such as a Data Center or Archive

- <publisher-name>

```
<publisher-name>Lewis Publishers</publisher-name>
```

or

```
<publisher-name>Carolina Population Center, University of
North Carolina at Chapel Hill</publisher-name>
```

or

```
<publisher-name>Public Library of Science</publisher-name>
```

5. Publication Date/ Year / Release Date

When this version of the dataset was made available for citation. May be only a year. Some sources place this inside the dataset title/name.

- <date>

```
<date iso-8601-date="2015-06">
<month>June</month><year>2015</year>
</date>
```

- <year>

```
<year iso-8601-date="2015-06">2015</year>
```

6. Version

The precise version number of the data used.

- <version>

```
<version>16.2.1</version>
```

or

```
<version designator="16.2">16th version, second release</version>
```

7. Resource Type

Material designator; medium; general type description The only way current JATS has to record this is @publication-format/@publication-type.

```
<mixed-citation publication-type="data"
  publication-format="online">...</mixed-citation>
<mixed-citation publication-type="data"
  publication-format="spreadsheet">...</mixed-citation>
```

8. Location of Publisher/Distributor

Location of the party publishing the data; may include such as city, state, country.

- <publisher-loc>

```
<publisher-loc>San Francisco, USA</publisher-loc>
```

9. Access Date and Time

Exactly when the online data was accessed

- <date-in-citation>

```
<date-in-citation content-type="access-date"
  iso-8601-date="2014-06-13:10:00">
Accessed on: <year>2014</year>, <month>June</month>,
<day>13</day> at 10:00am
</date-in-citation>
```

10. Additional URI/ Location / Bridge Service

Additional URI, location, bridge service, secondary distributor, reflector, or other institutional role such as funding. Typically holds a URL in addition to the regular DOI

- <ext-link> with @ext-link-type attribute

```
<ext-link ext-link-type="uri" xlink:href="http://
r-forge.r-project.org/projects/splits">
http://r-forge.r-project.org/projects/splits</ext-link>
```

- <uri>

```
<uri xlink:href="http://www.biomedcentral.com/1471-2180/13/198"
www.biomedcentral.com/1471-2180/13/198</uri>
```

Key components of data publishing: Using current best practices to develop a reference model for data publishing

Authors: Claire C. Austin[1,2,3,] ,Theodora Bloom*[,4], Sünje Dallmeier-Tiessen[§,5], Varsha K. Khodiyar[6], Fiona Murphy[§,7], Amy Nurnberger[8], Lisa Raymond[9], Martina Stockhause[10], Jonathan Tedds[11], Mary Vardigan[12] , Angus Whyte[13]

[§]Corresponding authors: Sünje Dallmeier-Tiessen, Fiona Murphy

Contributors: Timothy Clark[14], Eleni Castro[15], Elizabeth Newbold[16], Samuel Moore[17], Brian Hole[18]
Author affiliations: [1]Environment Canada, [2]Research Data Canada, [3]Carleton University, [4]BMJ, [5]CERN, [6]Nature Publishing Group,[7]University of Reading, [8]Columbia University, [9]Woods Hole Oceanographic Institution, [10]German Climate Computing Centre (DKRZ) , [11]University of Leicester, [12]University of Michigan/ICPSR, [13]Digital Curation Centre - Edinburgh.

Contributor affiliations: [14]Massachusetts General Hospital/Harvard Medical School, [15]Harvard University/IQSS, [16]The British Library, [17]Ubiquity Press.

Author statement: All authors affirm that they have no undeclared conflicts of interest. Opinions expressed in this paper are those of the authors and do not necessarily reflect the policies of the organizations with which they are affiliated.
Authors contributed to the writing of the article itself and significantly to the analysis. Contributors shared their workflows with the group (for the analysis). Authors are listed in alphabetical order.

*Theodora Bloom is a member of the Board of Dryad Digital Repository, and works for BMJ, which publishes medical research and has policies around data sharing.

Abstract

Purpose:
Availability of workflows for data publishing could have an enormous impact on researchers, research practices and publishing paradigms, as well as on funding strategies and career and research evaluations. We present the generic components of such workflows in order to provide a reference model for these stakeholders.
Methods:
The RDA-WDS Data Publishing Workflows group set out to study the current data publishing workflow landscape across disciplines and institutions. A diverse set of workflows were examined to identify common components and standard practices, including basic self-publishing services, institutional data repositories, long term projects, curated data repositories, and joint data journal and repository arrangements.
Results:
The results of this examination have been used to derive a data publishing reference model comprised of generic components. From an assessment of the current data publishing landscape, we highlight important gaps and challenges to consider, especially when dealing with more complex workflows and their integration into wider community frameworks.
Conclusions:
It is clear that the data publishing landscape is varied and dynamic, and that there are important gaps and challenges. The different components of a data publishing system need to work, to the greatest extent possible, in a seamless and integrated way. We therefore advocate the implementation of existing standards for repositories and all parts of the data publishing

process, and the development of new standards where necessary. Effective and trustworthy data publishing should be embedded in documented workflows. As more research communities seek to publish the data associated with their research, they can build on one or more of the components identified in this reference model.

CONTENTS

Data availability

Data from the analysis presented in this article are available:
Bloom, T., Dallmeier-Tiessen, S., Murphy, F., Khodiyar, V.K., Austin, C.C., Whyte, A., Tedds, J., Nurnberger, A., Raymond, L., Stockhause, M., Vardigan, M. *Zenodo* doi: 10.5281/zenodo.33899 (2015)

Introduction

Various data publishing workflows have emerged in recent years to allow researchers to publish data through repositories and dedicated journals. While some disciplines, such as the social sciences, genomics, astronomy, geosciences, and multidisciplinary fields such as Polar science, have established cultures of sharing research data[1] via repositories[2], it has traditionally not been common

[1] When we use the term 'research data' we mean data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results. All digital and non-digital outputs of a research project have the potential to become research data. Research data may be experimental, observational, operational, data from a third party, from the public sector, monitoring data, processed data, or repurposed data (Research Data Canada, 2015, Glossary of terms and definitions, http://dictionary.casrai.org/Category:Research_Data_Domain).

[2] A repository (also referred to as a data repository or digital data repository) is a searchable and queryable interfacing entity that is able to store, manage, maintain and curate Data/Digital Objects. A repository is a managed location (destination, directory or 'bucket') where digital data objects are registered, permanently stored, made accessible and retrievable, and curated (Research Data Alliance, Data Foundations and Terminology Working Group. http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page). Repositories preserve, manage, and provide access to many types of digital material in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse. There must be sufficient control for the digital material to be authentic, reliable, accessible and usable on a continuing basis (Research Data Canada, 2015, Glossary of terms and definitions, http://dictionary.casrai.org/Category:Research_Data_Domain). Similarly, 'data services' assist organizations in the capture, storage, curation, long-term preservation, discovery, access, retrieval, aggregation, analysis, and/or visualization of scientific data, as well as in the associated legal frameworks, to support disciplinary and multidisciplinary scientific research.

practice in all fields for researchers to deposit data for discovery and reuse by others. Typically, data sharing has only taken place when a community has committed itself towards open sharing (e.g. Bermuda Principles and Fort Lauderdale meeting agreements for genomic data[3]), or there is a legal[4] requirement to do so, or where large research communities have access to discipline-specific facilities, instrumentation or archives.

A significant barrier to moving forward is the wide variation in best practices and standards between and within disciplines. Examples of good practice include standardized data archiving in the geosciences, astronomy and genomics. Archiving for many other kinds of data are only just beginning to emerge or are non-existent [1]. A major disincentive for sharing data via repositories is the amount of time required to prepare data for publishing, time that may be perceived as being better spent on activities for which researchers receive credit (such as traditional research publications, obtaining funding, etc.). Unfortunately, when data are sequestered by researchers and their institutions, the likelihood of retrieval declines rapidly over time [2].

The advent of publisher and funding agency mandates to make accessible the data underlying publications is shifting the conversation from "Should researchers publish their data?" to "How can we publish data in a reliable manner?". We now see requirements for openness and transparency, and a drive towards regarding data as a first-class research output. Data publishing can provide significant incentives for researchers to share their data by providing measurable and citable output, thereby accelerating an emerging paradigm shift. Data release is not yet considered in a comprehensive manner in research evaluations and promotions, but enhancements and initiatives are under way within various funding and other research spaces to make such evaluations more comprehensive [3]. While there is still a prevailing sense that data carry less weight than published journal articles in the context of tenure and promotion decisions, recent studies demonstrate that when data are publicly available, a higher number of publications results [4,5].

The rationale for sharing data is based on assumptions of reuse - if data are shared, then users will come. However, the ability to share, reuse and repurpose data depends upon the availability of appropriate knowledge infrastructures. Unfortunately, many attempts to build infrastructure have failed because they are too difficult to adopt. The solution may be to enable infrastructure to develop around the way scientists and scholars actually work, rather than expecting them to work in ways that the data center, organisational managers, publishers or funders would wish them to [6]. Some surveys have found that researchers' use of repositories ranks a distant third - after responding to individual requests and posting data on local websites [7].

Traditionally, independent replication of published research findings has been a cornerstone of scientific validation. However, there is increasing concern surrounding the reproducibility of published research, i.e. that a researcher's published results can be reproduced using the data, code, and methods employed by the researcher [8-10]. Here too, a profound culture change is needed if reproducibility is to be integrated into the research process [11-13]. Data availability is key to reproducible research and essential to safeguarding trust in science.

As a result of the move toward increased data availability, a community conversation has begun about the standards, workflows, and quality assurance practices used by data repositories and data journals. Discussions and potential solutions are primarily concerned with how best to handle the vast amounts of data and associated metadata in all their various formats. Standards at various levels are being developed by stakeholder groups and endorsed through international bodies such as the Research Data Alliance (RDA), the World Data System of the International Council for Science (ICSU-WDS), and within disciplinary communities. For example, in astronomy there has been a long process of developing metadata standards through the International Virtual Observatory Alliance

[3] http://www.genome.gov/10506376

[4] For example, the Antarctic Treaty Article III states that "scientific observations and results from Antarctica shall be exchanged and made freely available." http://www.ats.aq/e/ats_science.htm

(IVOA)[5], while in the climate sciences the netCDF/CF convention was developed as a standard format including metadata for gridded data. Even in highly diverse fields such as the life sciences, the BioSharing[6] initiative is attempting to coordinate community use of standards. Increasingly there is a new understanding that data publishing ensures long-term data preservation, and hence produces reliable scholarship, demonstrates reproducible research, facilitates new findings, enables repurposing, and hence realises benefits and maximises returns on research investments.

But what exactly is data publishing? Parsons and Fox [14] question whether publishing is the correct term when dealing with digital information. They suggest that the notion of data publishing can be limiting and simplistic and they recommend that we explore alternative paradigms such as the models for software release and refinement, rather than one-time publication [14]. Certainly, version control[7] does need to be an integral part of data publishing, and this can distinguish it from the traditional journal article. Dynamic data citation is an important feature of many research datasets which will evolve over time, e.g. monitoring data and longitudinal studies [15]. The data journal *Earth System Science Data* is addressing this challenge with its approach to 'living data'[8]. The RDA Dynamic Citation Working group has also developed a comprehensive specification for citing everything from a subset of a data set to data generated dynamically, 'on-the-fly' [16]. International scientific facilities typically plan periodic scientifically processed data releases through the lifetime of a mission (e.g. XMM-Newton X-ray Telescope source catalogue, [17], in addition to making underlying datasets available through archives according to embargo policies.

In 2011, Lawrence et al. [18] defined the act of 'publishing data,' as: "*to make data as permanently available as possible on the Internet*." Published data will have been through a process guaranteeing easily digestible information as to its trustworthiness, reliability, format and content. Callaghan et al. [19] elaborate on this idea, arguing that formal publication of data provides a service over and above the simple act of posting a dataset on a website, in that it includes a series of checks on the dataset of either a technical (format, metadata) or a more content-based nature (e.g. are the data accurate?). Formal data publication also provides the data user with associated metadata, assurances about data persistence, and a platform for the dataset to be found and evaluated – all of which are essential to data reuse. An important consideration for our study is that support for 'normal' curation falls short of best practice standards. For example, having conducted a survey of 32 international online data platforms [20], the Standards & Interoperability Committee of Research Data Canada (RDC)[9] concluded that there is still a great deal of work to be done to ensure that online data platforms meet minimum standards for reliable curation and sharing of data, and developed guidelines for the deposit and preservation aspects of publishing research data.

With the present study, a first step is taken towards a reference model comprising generic components for data publishing - which should help in establishing standards across disciplines. We describe selected data publishing solutions, the roles of repositories and data journals, and characterize workflows currently in use. Our analysis involved the identification and description of a diverse set of workflows, including basic self-publishing services, long-term projects, curated data repositories, and joint data journal and repository arrangements. Key common components and standard practices were identified as part of a reference model for data publishing. These could help with standardising data publishing activities in the future (while leaving enough room for disciplinary or institutional

[5] http://www.ivoa.net

[6] https://biosharing.org

[7] Version control (also known as 'revision control' or 'versioning') is control over a time period of changes to data, computer code, software, and documents that allows for the ability to revert to a previous revision, which is critical for data traceability, tracking edits, and correcting errors. TeD-T: Term definition tool. Research Data Alliance, Data Foundations and Terminology Working Group. http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page .

[8] http://www.earth-system-science-data.net/living_data_process.html

[9] Research Data Canada (RDC) is an organisational member of Research Data Alliance (RDA) and from the beginning has worked very closely with RDA. See: "Guidelines for the deposit and preservation of research data in Canada, http://www.rdc-drc.ca/wp-content/uploads/Guidelines-for-Deposit-of-Research-Data-in-Canada-2015.pdf and, "Research Data Repository Requirements and Features Review," http://hdl.handle.net/10864/10892

practices). It is worth noting that there is continued discussion about many of the key definitions. The working group is presenting core data publishing terms based on the analysis.

Methods and materials

The RDA-WDS Publishing Data Workflows Working Group (WG) was formed to provide an analysis of a reasonably representative range of existing and emerging workflows and standards for data publishing, including deposit and citation, and to provide components of reference models and implementations for application in new workflows. The present work was specifically focused on articulating a draft reference model comprising generic components for data publishing workflows that others can build upon. We also recognize the need for the reference model to promote workflows that researchers find usable and attractive.

To achieve this, the working group followed the OASIS definition of a reference model as:
"…an abstract framework for understanding significant relationships among the entities of some environment, and for the development of consistent standards or specifications supporting that environment. A reference model is based on a small number of unifying concepts and may be used as a basis for education and explaining standards to a non-specialist. A reference model is not directly tied to any standards, technologies or other concrete implementation details, but it does seek to provide a common semantics that can be used unambiguously across and between different implementations". [10]

A particularly relevant example is the OAIS Reference Model for an Open Archival Information System.[11] This model has shaped the Trusted Digital Repository (TDR) standards which frame repository best practice for ingesting, managing and accessing archived digital objects. These have recently been exemplified by the DSA-WDS Catalogue of Requirements [12] and are particularly relevant for their emphasis on making workflows explicit.
Our specific concerns in the working group build on such standards, to guide implementation of quality assurance and peer review of research data objects, their citation, and linking with other digital objects in the research and scholarly communication environment.
A case study approach was in keeping with this aim. Case studies explore phenomena in their context, and generalise to theory rather to populations [21]. Similarly, drafting a conceptual model does not require us to make generalisable claims to the repository population as a whole, but it does commit us to testing its relevance to repositories, and other stakeholders, through community review and amendment.

As the membership of the RDA-WDS Publishing Data Workflows WG was reasonably diverse in terms of disciplinary and stakeholder participation, we drew upon that group's knowledge and contacts, and issued calls to participate under the auspices of the RDA and WDS, in collaboration with the Force11 Implementation Group[13] to identify best practices and case studies in data publishing workflows. Presentations and workshops at RDA plenary meetings were used to validate the approach and progress. With this iterative approach, we identified an initial set of repositories, projects and publishing platforms which were thought to be reasonably representative of institutional affiliation and domain-specific or cross-disciplinary focus. These workflows served as a case study

[10] Source: OASIS, *https://www.oasis-open.org/committees/soa-rm/faq.php*

[11] "Recommendation for Space Data System Practices: Reference Model for an Opean Archival Information System (OAIS), CCSDS 650.0-M-2." http://public.ccsds.org/publications/archive/650x0m2.pdf DataCite (2015). "DataCite Metadata Schema for the Publication and Citation of Research Data." http://dx.doi.org/10.5438/0010

[12] Draft available at: https://rd-alliance.org/group/repository-audit-and-certification-dsa%E2%80%93wds-partnership-wg/outcomes/dsa-wds-partership

[13] Force11 (2015). Future Of Research Communications and e-Scholarship https://www.force11.org/group/data-citation-implementation-group

for the analysis to identify likely examples of 'data publishing' from repositories, projects and publishing platforms, whether institutional, domain-specific, or cross-disciplinary.

Publicly available information was used to describe the workflows on a common set of terms. In addition, repository representatives were invited to present and discuss their workflows via videoconference and face-to-face meetings. Emphasis was given to workflows facilitating data citation and the provision of 'metrics' for data was added as a consideration. Information was organized into a comparison matrix and circulated to the group for review, whereupon a number of annotations and corrections were made. Empty fields were populated, where possible, and terms were cross-checked and harmonized across the overall matrix. Twenty-six examples were used for comparison of characteristics and workflows. However, one workflow (Arkivum) was judged not to qualify for the definition of 'data publishing' as it emerged in the course of the research, so the final table consists of twenty-five entities (Table 1).

Table 1. Repositories, projects and publishing platforms selected for analysis of workflows and other characteristics

Workflow provider name	**Workflow provider type**	**Workflow provider specialist research area, if any**	**Deposit initiator**
ENVRI reference model	Guidelines	Environmental sciences	Project-led
PREPARDE	Guidelines	Earth sciences	Researcher led (for *Geoscience Data Journal*)
Ocean Data Publication Cookbook	Guidelines	Marine sciences	Researcher-led
Scientific Data, Nature Publishing Group	Journal		Researcher- (author) led
F1000Research	Journal	Life sciences	Researcher led; editorial team does a check
Ubiquity Press OHDJ	Journal	life, health and social sciences	Researcher-led
GigaScience	Journal	Life and biomedical sciences	Researcher- (author) led
Data in Brief	Journal		Author-led
Earth System Science Data Journal, Copernicus Publications	Journal	Earth sciences	Researcher-led for data article. Researcher-led for data submission to repository
Science and Technology Facilities Council Data Centre	Repository	Physics and space sciences	Researcher-led as part of project deliverables
National Snow and Ice Data Center	Repository	Polar Sciences	Project- or researcher-led
INSPIRE Digital library	Repository	High energy Physics	Researcher-led

UK Data Archive (ODIN)	Repository	Social sciences	Researcher-led
PURR Institutional Repository	Repository		Researcher- /Librarian-led
ICPSR	Repository	Social and behavioural sciences	Researcher-, acquisitions officer-, and funder-led
Edinburgh Datashare	Repository		Researcher-led, librarian assists
PANGAEA	Repository	Earth sciences	Researcher-led
WDC Climate	Repository	Earth sciences	Researcher- or project-led
CMIP/IPCC-DDC	Repository	Climate sciences	Project-led[14]
Dryad Digital Repository	Repository	Life sciences	Researcher-led
Stanford Digital Repository	Repository		Researcher-led
Academic Commons Columbia	Repository		Researcher and repository staff
Data Repository for the University of Minnesota (DRUM)	Repository		Researchers from institution
ARKIVUM and Figshare	Repository		Researcher-led
OJS/ Dataverse data repository, all disciplines	Repository		Researcher-led; part of journal article publication process

Workflows were characterized in terms of the discipline, function, data formats, and roles involved. We also described the extent to which each exhibited the following 10 characteristics associated with data publishing:

- The assignment of persistent identifiers (PIDs) to datasets, and the PID type used -- e.g. DOI, ARK, etc.
- Peer review of data (e.g. by researcher and by editorial review)
- Curatorial review of metadata (e.g. by institutional or subject repository)
- Technical review and checks (e.g. for data integrity at repository/data centre on ingest)
- Discoverability: was there indexing of the data, and if so, where?
- Links to additional data products (data paper; review; other journal articles) or "stand-alone" product
- Links to grant information, where relevant, and usage of author PIDs
- Facilitation of data citation
- Reference to a data life cycle model
- Standards compliance

The detailed information and categorization can be found in the analysis dataset comprising the comparison matrix [22].

[14] Data Citation concept for CMIP6/AR6 is available as draft at: http://www.earthsystemcog.org/projects/wip/resources/

Results and analysis

Towards a reference model in data publishing

Definitions for data publishing workflows and outputs

The review of the comparison matrix of data publishing workflows produced by the RDA-WDS Publishing Data Workflows WG [22] revealed a need for standardization of terminology. We therefore propose definitions for six key terms: research data publishing, research data publishing workflows, data journal, data article, data review, and data repository entry.

Research data publishing

"Research data publishing is the release of research data, associated metadata, accompanying documentation, and software code (in cases where the raw data have been processed or manipulated) for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way. Data publishing occurs via dedicated data repositories and/or (data) journals which ensure that the published research objects are well documented, curated, archived for the long term, interoperable, citable, quality assured and discoverable – all aspects of data publishing that are important for future reuse of data by third party end-users."

This definition applies also to the publication of confidential and sensitive data with the appropriate safeguards and accessible metadata. A concrete example of such a workflow may be a published journal article that includes discoverability and citation of a dataset by identifying access criteria for reuse[15]. Harvard University is currently developing a tool that will eventually be integrated with Dataverse to share and use confidential and sensitive data in a responsible manner[16].

Research data publishing workflows

Research data publishing workflows are activities and processes that lead to the publication of research data, associated metadata and accompanying documentation and software code on the Web. In contrast to interim or final published products, workflows are the means to curate, document, and review, and thus ensure and enhance the value of the published product. Workflows can involve both humans and machines and often humans are supported by technology as they perform steps in the workflow. Similar workflows may vary in their details, depending on the research discipline, data publishing product and/or the host institution of the workflow (e.g. individual publisher/journal, institutional repository, discipline-specific repository).

Data journal

A data journal is a journal (invariably Open Access) that publishes data articles. The data journal usually provides templates for data description and offers researchers guidance on where to deposit and how to describe and present their data. Depending on the journal, such templates can be generic or discipline focused. Some journals or their publishers maintain their own repositories. As well as supporting bi-directional linking between a data article and its corresponding dataset(s), and facilitating persistent identification practices, data journals provide workflows for quality assurance(i.e. data peer review), and should also provide editorial guidelines on data quality assessment.

Data article

A data article is a 'data publishing' product, also known as a 'data descriptor', that may appear in a data journal or any other journal. When publishers refer to 'data publishing' they

[15] indirect linkage or restricted access - see e.g. Open Health Data Journal, http://openhealthdata.metajnl.com

[16] http://privacytools.seas.harvard.edu/datatags

usually mean a data article rather than the underlying dataset. Data articles focus on making data discoverable, interpretable and reusable rather than testing hypotheses or presenting new interpretations (by contrast with traditional journal articles). Whether linked to a dataset in a separate repository, or submitted in tandem with the data, the aim of the data article is to provide a formal route to data-sharing. The parent journal may choose whether or how standards of curation, formating, availability, persistence or peer review of the dataset are described. By definition, the data article provides a vehicle to describe these qualities, as well as some incentive to do so. The length of such articles can vary from micro papers (focused on one table or plot) to very detailed presentation of complex datasets.

Data review

Data review comprises a broad range of quality assessment workflows, which may extend from a technical review of metadata accuracy to a double-blind peer review of the adequacy of data files and documentation and accuracy of calculations and analyses. Multiple variations of review processes exist and are dependant upon factors such as publisher requirements, researcher expectations, or data sensitivity. Some workflows may be similar to traditional journal workflows, in which specific roles and responsibilities are assigned to editors and reviewers to assess and ensure the quality of a data publication. The data review process may therefore encompass a peer review that is conducted by invited domain experts external to the data journal or the repository, a technical data review conducted by repository curation experts to ensure data are suitable for preservation, and/or a content review by repository subject domain experts.

Data repository entry

A data repository entry is the basic component of data publishing consisting of a persistent, unique identifier pointing to a landing page that contains a data description and details regarding data availability and the means to access the actual data [22]

Key components of data publishing

Analysis of workflows by the RDA-WDS data publishing WG identified the components that contribute to a generic reference model for data publishing. We distinguish basic and add-on services. The basic set of services consists of entries in a trusted data repository, including a persistent identifier, standardized metadata, and basic curation (Figure 1).

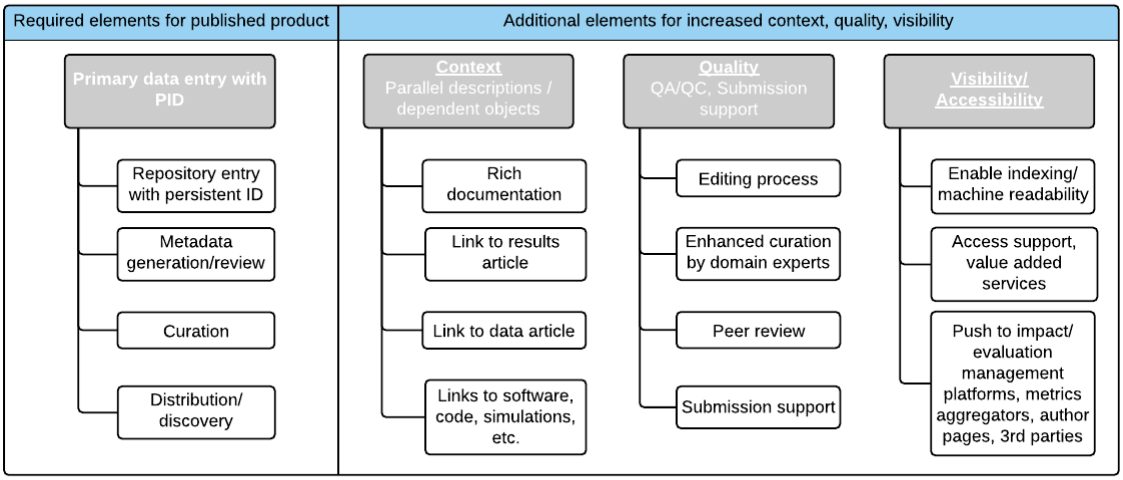


Figure 1. Data publishing key components. Elements that are required to constitute data publication are shown in the left panel, and optional services and functions in the right panel.

Optional add-ons could include components such as contextualisation through additional embedding into data papers or links to traditional papers. Some authors and solutions make a distinction between

metadata publication and data publication. We would argue that data and their associated metadata must at least be bi-directionally linked in a persistent manner, and that they need to be published together and viewed as a package, since metadata are essential to the correct use, understanding, and interpretation of the data.

Important add-ons are quality assurance/quality control (QA/QC)[17] and peer review services. Different variations of such services exist, ranging from author-led, editor-driven, librarian-supported solutions, to (open) peer review. Such components are crucial enablers of future data reuse and reproducible research. Our analysis found that many services offer or are considering offering such services. The third group of add-ons aims to improve visibility, as shown on the right panel of Figure 1. This set of services is not currently well established, and this hampers data reuse. Other emerging services include connection of data publishing workflow with indexing services, research information services (CRIS), or metrics aggregators.

To ensure the possibility of data reuse, data publishing should contain at least the basic elements of curation, QA/QC, and referencing, plus additional elements appropriate for the use case (Figure 1). Depending on the use case, however, it might be appropriate to select a specific set of elements from the key components (following some best practices). In the light of future reuse, we would argue that the basic elements of curation, QA/QC, and referencing should always be included.

Detailed workflows and dependencies

We present a traditional article publication workflow (Fig. 2-1), a reproducible research workflow (Fig. 2-2), and a data publication workflow (Fig. 2-3).
The workflow comparison found that it is usually the researcher who initiates the publication process once data have been collected and are in a suitable state for publication, or meet the repository requirements for submission. Datasets may be published in a repository with or without an associated data article. However, there are examples for which there is a direct 'pipe' from a data production 'machine' to a data repository (genome sequencing is one such example). Depending on the data repository, there are both scientific and technical [18,23] quality assurance activities regarding dataset content, description, format, and metadata quality before data are archived for the long term. The typical data repository creates an entry for a specific dataset or a collection thereof. Most repositories invest in standardized dissemination for datasets, i.e. a landing page for each published item, as recommended by the Force11 Data Citation Implementation Group[18] [24]. Some repositories facilitate third-party access for discoverability or metrics services.

[17] Quality assurance: The process or set of processes used to measure and assure the quality of a product. Quality control: The process of meeting products and services to consumer expectations (Research Data Canada, 2015, Glossary of terms and definitions, http://dictionary.casrai.org/Category:Research_Data_Domain)

[18] https://www.force11.org/datacitationimplementation

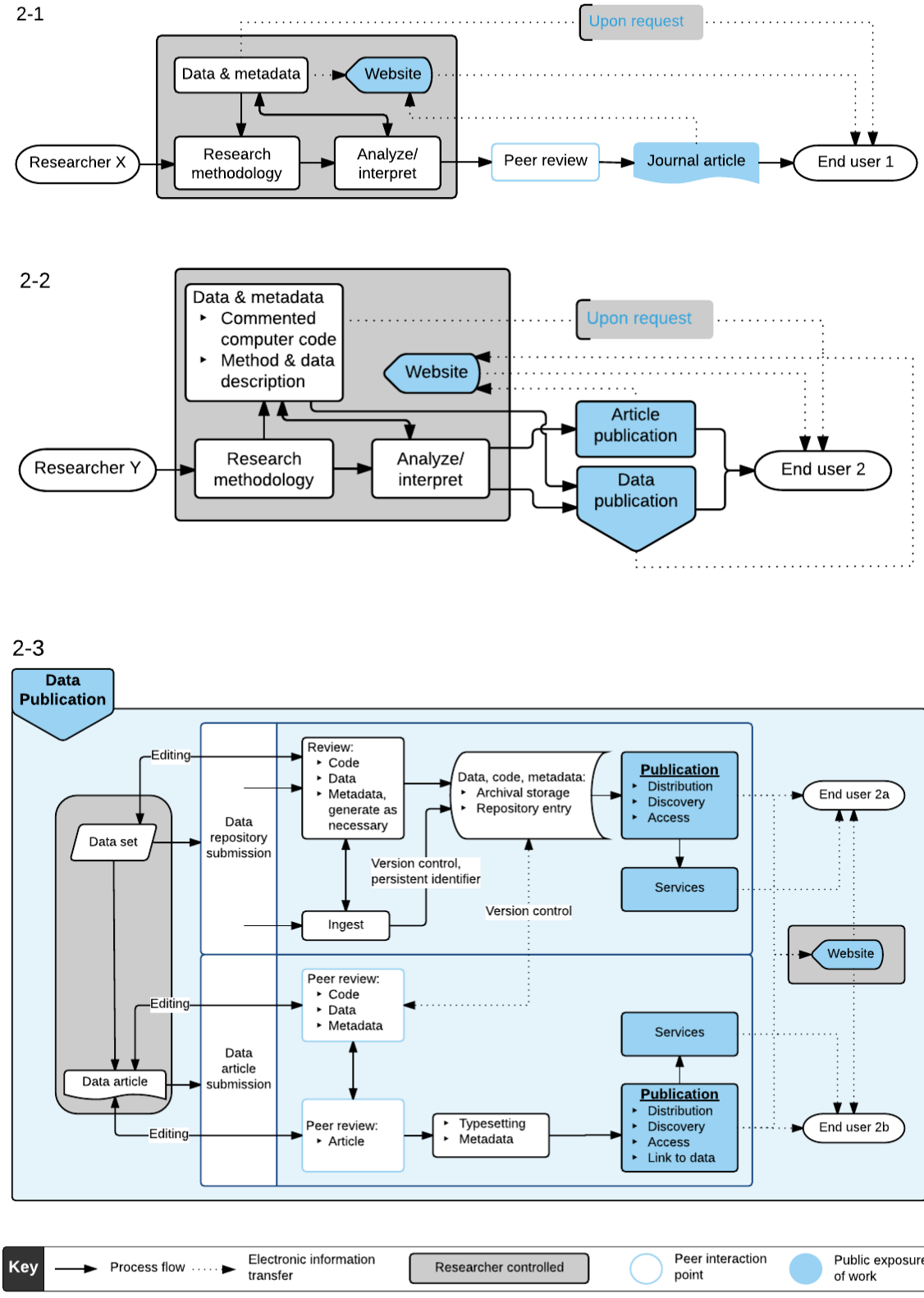


Figure 2. Research data publication workflows. We present a traditional article publication workflow (Fig. 2-1), a reproducible research workflow (Fig. 2-2), and a data publication workflow (Fig. 2-3).

As shown in Figure 2, researchers can and do follow a number of different pathways to communicate about their data. Traditionally, research results are published in journals and readers (end user 1) interested in the data would need to contact authors to access underlying data, or attempt to access it

from a researcher-supported website (Figure 2-1). Emerging processes supporting greater reproducibility in research include some form of data publication (Figure 2-2). This includes the special case of standalone[19] data publications with no direct connection to a paper. These are common in multiple domain areas (e.g. the large climate data intercomparison study CMIP[20]). Figure 2-3 illustrates the two predominantly emerging data publication workflows emerging from our analysis: (a) submission of a dataset to a repository; and, (b) submission of a data article to a data journal. Both workflows require that datasets are submitted to a data repository.

The data publication process shown in Figure 2-3 may be initiated at any time during research once the data are sufficiently complete and documented, and may follow a variety of paths. A repository will typically provide specific templates for metadata and additional documentation (e.g. methodology or code specific metadata). The submission may then be reviewed from a variety of perspectives depending on the policies and practices of the repository. These review processes may include formatting issues, content, metadata or other technical details. Some repositories may also require version control of the data set. There is a great deal of variability between repositories in the type of data accepted, available resources, the extent of services offered, and workflows. Figure 2-3 illustrates the elements common to the workflows of the data repositories selected for the present study (Figure 2-3) are consistent with those shown in Figure 1.

A researcher may also choose to initiate the data publication process by submitting a data article for publication in a data journal. This workflow is also illustrated in Figure 2-3, and while it is in part dependent on data repositories (data journals typically identify approved repositories[21]), the data article publication process has the opportunity to more consistently provide some of the advantages of data publication as represented in the 'Additional elements' of Figure 1. Data journals are similar to the traditional research journal (Figure 2-1) in that their core processes consist of peer review and dissemination of the datasets. Naturally, reviewers must have pre-publication access to the dataset in a data repository, and there needs to be version control solutions for datasets and data papers. Whether publishing data via a data article or a data repository, both workflows have the potential to be incorporated into the current system of academic assessment and reward in an evolutionary process rather than a disruptive departure from previous systems.

Data publication workflows supporting reproducible research give end-users access to managed and curated data, code and supporting metadata that have been reviewed and uploaded to a trusted repository (Fig. 2, end-user 2a). If an associated data article is published, end users will also have further contextual information (Fig. 2, end-user 2b). The traditional journal article may be published as usual, and may be linked to the published data and/or data article as well. There are some hard-wired automated workflows for data publishing (e.g. with the Open Journal Systems-Dataverse integration [25), or there can be alternate automated or manual workflows in place to support the researcher (e.g. Dryad).

Data deposit

We found that a majority of data deposit mechanisms underlying data publishing workflows are initiated by researchers, but their involvement beyond the initial step of deposition varied across repositories and journals. Platform purpose (e.g. data journal vs. repository) and the ultimate perceived purpose and motivation of the depositor of the data all affect the process. For example, a subject-specialist repository, such as is found at Science and Technology Facilities Council (STFC) or the National Snow and Ice Data Center (NSIDC), screens submissions and assesses the levels of metadata and support required. Data journals, however, typically adopt a 'hands-off' approach: the journal is the 'publication' outlet, but the data are housed elsewhere. Hence the journal publishing

[19] Defined in e.g. [18]

[20] Program for Climate Model Diagnosis and Intercomparison. (n.d.). Coupled Model Intercomparison Project (CMIP). Retrieved November 11, 2015, from http://www-pcmdi.llnl.gov/projects/cmip/

[21] Approved by the data journal

team often relies on external parties – repository managers and the research community in general[22] – to manage data deposit and to assess whether basic standards are met for data deposition or if quality standards are met (see details below).

Ingest

We found that discipline-specific repositories had the most rigorous ingest and review processes and that more general repositories, e.g. institutional repositories (IRs) or Dryad, had a lighter touch given the greater diversity of use cases and practice around data from diverse disciplines. Some discipline-specific repositories have multiple-stage processes including several QA/QC processes and workflows based on OAIS. Many IRs have adopted a broader approach to ingest necessitated by their missions, which involves archiving research products generated across their campuses, especially those found in the long-tail of research data, including historical data that may have been managed in diverse ways. As data standards are developed and implemented and as researchers are provided with the tools, training, and incentives needed to engage in modern data management practices, ingest practices will no doubt improve.

When data journals rely on external data repositories to handle the actual data curation, there needs to be a strong collaboration between the journal and repository staff beyond trust that the repository will pursue data management and ingestion according to acceptable standard procedures. Data journals and data repositories are encouraged to make public and transparent any such agreements (e.g. Service Level Agreements). Ultimately, however, this level of one-to-one interaction is not scalable and automated procedures and repository standards will be needed.

Quality assurance (QA) and quality control (QC)

We found that QA/QC typically occurs at three points during the data publishing workflow: (1) during data collection and data processing, prior to submission of the data to a repository; (2) during submission and archiving of the data; and, (3) during a review or editorial procedure. We distinguish between traditionally understood peer review and the internal reviews that repositories and journals also generally conduct (Fig. 2), which may touch on content, format, description, documentation, metadata, or other technical details.
QA/QC procedures vary widely and may involve authors/reviewers for QA of the content and documentation, and data managers/curators, librarians and editors for technical QA. Quality criteria can include checks on data, metadata and documentation against repository, discipline[23] and project standards.

Most repositories and all of the data journals that we reviewed had some QA/QC workflows, but the level and type of services varied. Established data repositories (e.g. ICPSR or Dataverse [22] tended to have dedicated data curation personnel to help in standardising and reviewing data upon submission and ingestion, especially in the area of metadata. Some domain repositories (e.g. ICPSR) go farther and conduct in-depth quality control checks on the data, revising the data if necessary in consultation with the original investigator. Other repositories responsible for the long-term archiving of project data (e.g. the IPCC-DDC[24]) document their QA results. Some data repositories rely on researchers for the QA/QC workflows to validate the scientific aspects of data, metadata and documentation. Technical support, data validation or QA/QC was also done by some repositories, but the level of engagement varied with the service and the individual institutions: some checked file integrity, while others offered more complex preservation actions, such as on-the-fly data format

[22] Post-publication peer review is becoming more prevalent and may ultimately strengthen the Parsons-Fox continual release paradigm. See, for instance, F1000 Research and Earth System Science Data and the latter journal's website: http://www.earth-system-science-data.net/peer_review/interactive_review_process.html.

[23] An example for a discipline standard is the format and metadata standard NetCDF/CF used in Earth System Sciences: http://cfconventions.org/

[24] Intergovernmental Panel on Climate Change Data Distribution Centre (IPCC-DDC): http://ipcc-data.org

conversions. Some multi-purpose repositories provided support to researchers for QA/QC workflows, but this was not a standard practice. Overall, QA/QC in data publishing is a 'hot-button' topic and is debated heavily and continuously within the community. Mayernik et al. describe a range of practice in technical and academic peer review for publishing data [26].

The journal workflows we examined typically straddled the dual processes of reviewing the dataset itself and the data papers, which were carried out separately and then checked to ensure that the relationship between the two was valid. Such QA/QC workflows for data journals demand a strong collaboration with the research community and their peer reviewers but also between publisher and data repository in workflow co-ordination, versioning and consistency.

Given the wide range of QA/QC services currently offered, future recommendations should consider the following:

- Repositories which put significant effort into high levels of QA/QC benefit researchers whose materials match the repository's portfolio by making sure their materials are fit for reuse. This also simplifies the peer review process for associated data journals and lowers barriers to uptake by researchers.
- General research data repositories which must accommodate a wide variety of data may have some limitations in QA/QC workflows and these should be made explicit.
- Information about quality level definitions and quality assessment procedures and results should be explicit readily available to users (and also possibly to third parties, such as aggregators or metric services).

There appears to be a trend toward data being shared earlier in the research workflow, at a stage where the data are still dynamic (see for example Meehl et al., [27]). There is a need, therefore, for QA/QC procedures that can handle dynamic data.

Data administration and long-term archiving

Data administration and curation activities may include dealing with a variety of file types and formats, creation of access level restrictions, the establishment and implementation of embargo procedures, and assignment of identifiers. We found an assortment of practices in each of these areas. These vary from providing file format guidelines alone to active file conversions; from supporting access restrictions to supporting only open access; administering flexible or standardized embargo periods; and employing different types of identifiers. Several discipline-specific repositories already have a long track record of preserving data and have detailed workflows for archival preservation. Other repositories are fairly new to this discussion and continue to explore potential solutions.

Most repositories in our sample have indicated a commitment to persistence and the use of standards. The adoption of best practices and standards would increase the likelihood that published data will be maintained over time and lead to interoperable and sustainable data publishing. Repository certification systems have been gaining momentum in recent years and could help facilitate data publishing through collaboration with data publishing partners such as funders, publishers and data repositories. The range of certification schemes[25] includes those being implemented by organizations such as the Data Seal of Approval (DSA)[26] and the World Data System (ICSU-WDS)[27]. Improved adoption of such standards would have a big impact on interoperable and sustainable data publishing.

[25] Data Seal of Approval (DSA); Network of Expertise in long-term Storage and Accessibility of Digital Resources in Germany (NESTOR) seal / German Institute for Standardization (DIN) standard 31644; Trustworthy Repositories Audit and Certification (TRAC) criteria / International Organization for Standardization (ISO) standard 16363; and the International Council for Science World Data System (ICSU-WDS) certification.

[26] Data Seal of Approval: http://datasealofapproval.org/en/

[27] World Data System certification https://www.icsu-wds.org/files/wds-certification-summary-11-june-2012.pdf

Dissemination, access and citation

Data packages in most repositories we analyzed were summarized on a single landing page that generally offered some basic or enriched (if not quality assured) metadata. This usually included a DOI and sometimes another unique identifier as well or instead. We found widespread use of persistent identifiers and a recognition that data must be citable if it is to be optimally useful.[28] It should be noted that dissemination of data publishing products was, in some cases, enhanced through linking and exposure (e.g. embedded visualization) in traditional journals. This is important, especially given the culture shift needed within research communities to make data publishing the norm.

Dissemination practices varied widely. Many repositories supported publicly accessible data, but diverged in how optimally they were indexed for discovery. As would be expected, data journals tended to be connected with search engines, and with abstracting and indexing services. However, these often (if not always) related to the data article rather than to the dataset per se. The launch of the Data Citation Index[29] by Thomson Reuters and projects such as the *Data Discovery Index*[30] are working on addressing the important challenge of data discovery and could serve as an accelerator to a paradigm shift for establishing data publishing within research communities.

One example of such a paradigm shift occurred in 2014 when the Resource Identifier Initiative (RII) launched a new registry within the biomedical literature. The project covered antibodies, model organisms (mice, zebrafish, flies), and tools (i.e. software and databases), providing a fairly comprehensive combination of data, metadata and platforms to work with. Eighteen months later the project was able to report both a cultural shift in researcher behaviour and a significant increase in the potential reproducibility of relevant research. As discussed in Bandrowski et al [28], the critical factor in this initiative's success in gaining acceptance and uptake was the integrated way in which it was rolled out. A group of stakeholders including researchers, journal editors, subject community leaders and publishers - within a specific discipline, neuroscience - worked together to ensure a consistent message. This provided a compelling rationale, coherent journal policies (which necessitated compliance in order for would-be authors to publish), and a specific workflow for the registration process (complete with skilled, human support if required). Further work is needed to determine exactly how this use case can be leveraged across the wider gamut of subjects, communities and other players.

FAIR principles[31] and other policy documents [10] explicitly mention that data should be accessible. Data publishing solutions ensure that this is the case, but some workflows allow only specific users to access sensitive data. An example is survey data containing information that could lead to the identification of respondents. In such cases, a prospective data user could access the detailed survey metadata to determine if meets his/her research needs, but a data use agreement would need to be signed before access to the dataset would be granted. The metadata, data article or descriptor could be published openly, perhaps with a Creative Commons license, but the underlying dataset would be unavailable except via registration or other authorization processes. In such a case the data paper would allow contributing researchers to gain due credit, and it would facilitate data discovery and reuse[32].

Citation policies and practice also vary by community and culture. Increasingly, journals and publishers are including data citation guidelines in their author support services. In terms of a best

[28] Among the analyzed workflows it was generally understood that data citation which properly attributes datasets to originating researchers can be an incentive for deposit of data in a form that makes the data accessible and reusable, a key to changing the culture around scholarly credit for research data.

[29] http://wokinfo.com/products_tools/multidisciplinary/dci/

[30] http://grants.nih.gov/grants/guide/rfa-files/RFA-HL-14-031.html

[31] https://www.force11.org/group/fairgroup/fairprinciples

[32] See e.g. Open Health Data journal http://openhealthdata.metajnl.com/

practice or standard, the *Joint Declaration of Data Citation Principles*[33] is gathering critical mass, and becoming generally recognized and endorsed. Discussions concerning more detailed community practices are emerging; for example, whether or not publishing datasets and data papers – which can then be cited separately from related primary research papers – is a fair practice in a system that rewards higher citation rates. However, sensible practices can be formulated.[34]

Other potential value-added services, and metrics

Many repository or journal providers look beyond workflows that gather information about the research data, and also want to make this information visible to other information providers in the field. This can add value to the data being published. If the information is exposed in a standardized fashion, data can be indexed and be made discoverable by third-party providers, e.g. data aggregators (Figure 1). Considering that such data aggregators often work beyond the original data provider's subject or institutional focus, some data providers enrich their metadata (e.g. with data-publication links, keywords or more granular subject matter) to enable better cross-disciplinary retrieval. Ideally, information about how others download or use the data would be fed back to the researcher. In addition, services such as ORCID[35] are being integrated to allow researchers to connect their materials across platforms. This gives more visibility to the data through the different registries, and allows for global author disambiguation. The latter is particularly important for establishing author metrics. During our investigation, many data repository and data journal providers expressed an interest in new metrics for datasets and related objects. Tracking usage, impact and reuse of the shared materials can enrich the content on the original platforms and encourage users to engage in further data sharing or curation activities. Such information is certainly of interest to infrastructure and research funders[36].

Diversity in workflows

While workflows may appear to be fairly straightforward and somewhat similar to traditional static publication procedures, the underlying processes are, in fact, quite complex and diverse. The diversity was most striking in the area of curation. Repositories that offered self-publishing options without curation had abridged procedures, requiring fewer resources but also potentially providing less contextual information and fewer assurances of quality. Disciplinary repositories that performed extensive curation and QA had more complex workflows with additional steps, possibly consecutive. They might facilitate more collaborative work at the beginning of the process, or include standardized preservation steps.

There was metadata heterogeneity across discipline-specific repositories. Highly specialized repositories frequently focused on specific metadata schemas and pursued curation accordingly. Some disciplines have established metadata standards, similar to the social sciences' use of the Data Documentation Initiative standard[37]. In contrast, more general repositories tended to converge on domain-agnostic metadata schemas with fields common across disciplines, e.g. the mandatory DataCite fields[38].

Data journals are similar in overall workflows, but differ in terms of levels of support, review and curation. As with repositories, the more specialized the journal (e.g. a discipline in the earth sciences with pre-established data sharing practices), the more prescriptive the author guidelines and the more

---

[33] Data Citation Synthesis Group, 2014. Accessed 17 November 2015: https://www.force11.org/group/joint-declaration-data-citation-principles-final

[34] See Sarah Callaghan's blogpost: Cite what you use, 24 January 2014. Accessed 24 June 2015: http://citingbytes.blogspot.co.uk/2014/01/cite-what-you-use.html

[35] http://orcid.org/

[36] Funders have an interest in tracking Return on Investment to assess which researchers/projects/fields are effective and whether proposed new projects consist of new or repeated work.

[37] Accessed 17 November 2015: http://www.ddialliance.org

[38] Accessed 17 November 2015: https://schema.datacite.org

specialized the review and QA processes. With the rise of open or post-publication peer review, some data journals are also inviting the wider community to participate in the publication process.

The broader research community and some discipline-based communities are currently developing criteria and practices for standardized release of research data. The services supporting these efforts, whether repositories or journals, also generally show signs of being works in progress or proof-of-concept exercises rather than finished products. This is reflected in our analysis dataset [22]. Depending partly on their state of progress during our review period (1 February - 30 June 2015), and also on the specificity of the subject area, some workflow entries were rather vague.

Discussion and conclusions

Although the results of our analysis show wide diversity in data publishing workflows, key components were fairly similar across providers. The common components were grouped and charted in a reference model for data publishing. Given the rapid developments in this field and in light of the disciplinary differences, diversity of workflows might be expected to grow even further. Through the RDA Working Group we will seek further community review and endorsement of the generic reference model components, and carry out further analyses of such disciplinary variations. However, the results of our study suggest that new solutions (e.g. for underrepresented disciplines) could build on the identified key components that best match their use case. Some evident gaps and challenges (described below) hinder global interoperability and adoption of a common model.

Gaps and challenges

While there are still many disciplines for which no specific domain repositories exist, we are seeing a greater number of repositories of different types (re3data.org indexes over 1,200 repositories). In addition to the disciplinary repositories, there are many new repositories designed to house broader collections, e.g. Zenodo, Figshare, Dryad, Dataverse, and the institutional repositories at colleges and universities. "Staging" repositories are also being established that extend traditional workflows into the collaborative working space -- e.g. Open Science Framework[39] which has a publishing workflow with Dataverse. Another example is the SEAD[40] (Sustainable Environment Actionable Data) project, which provides project spaces in which scientists manage, find, and share data, and which also connects researchers to repositories that will provide long-term access to, and preservation of data.

Despite much recent data publishing activity, our analysis of the case studies found that challenges remain, in particular when considering more complex workflows. These include:

- Bi-directional linking. How do we link data and publications persistently in an automated way? Several organizations, including RDA and WDS[41], are now working on this problem. A related issue is the persistence of links themselves.[42]
- Software management. Solutions are needed to manage, preserve, publish and cite software. Basic workflows exist (involving code sharing platforms, repositories and aggregators), but much more work is needed to establish a wider framework, including community updating and initiatives involving linking to associated data .
- Version control. In general, we found that repositories handle version control in different ways, which is potentially confusing. While some version control solutions might be tailored to discipline-specific challenges, there is a need to standardize. This issue also applies to provenance information.

[39] https://osf.io/

[40] http://sead-data.net/

[41] RDA/WDS Publishing Data Services WG: https://rd-alliance.org/groups/rdawds-publishing-data-services-wg.html and https://www.icsu-wds.org/community/working-groups/data-publication/services

[42] See the hiberlink Project for information on this problem and work being done to solve it: http://hiberlink.org/dissemination.html

- Sharing restricted-use data. Repositories and journals are generally not yet equipped to handle confidential data. It is important that the mechanism for data sharing be appropriate to the level of sensitivity of the data. The time is ripe for the exchange of expertize in this area.
- Role clarity. Data publishing relies on collaboration. For better user guidance and greater confidence in the services, an improved understanding of roles, responsibilities, and collaboration is needed. Documentation of 'who does what' in the current, mid- and long-term would ensure a smoother provision of service.
- Business models. There is strong interest in establishing the value and sustainability of repositories. Beagrie and Houghton[43] produced a synthesis of data centre studies combining quantitative and qualitative approaches in order to quantify value in economic terms and present other, non-economic, impacts and benefits. A recent Sloan-funded meeting of 22 data repositories led to a white paper on Sustaining Domain Repositories for Digital Data[44]. However, much more work is needed to understand viable financial models for publishing data[45] and to distinguish trustworthy collaborations.
- Data citation support. Although there appears to be widespread awareness, there is only partial implementation of the practices and procedures recommended by the Data Citation Implementation Group. There is a wide range of PIDs emerging, including ORCID, DOI, FunderRef, RRID, IGSN, ARK and many more. Clarity and ease of use need to be brought to this landscape.[46]
- Metrics. Creators of data and their institutions and funders need to know how, and how often, their data are being reused.
- Incentives. Data publishing offers potential incentives to researchers, e.g. a citable data product, persistent data documentation, and information about the impact of the research. Also, many repositories offer support for data submission. Benefits of data publishing need to be better communicated to researchers. In addition, stakeholders should disseminate the fact that formal data archiving results in greater numbers of papers and thus more science, as Piwowar and Vision, and Pienta et al. [4,5] have shown. There should also be increased clarity with respect to institutional and funder recognition of the impact of research data.

The challenges of more complex data – in particular, big data and dynamic data – need also to be addressed. Whereas processes from the past 10 years focus on irrevocable, fully documented data for unrestricted (research) use. data publishing needs to be 'future-proof' (Brase et al. [29]). There is a requirement from research communities[47] to cite data before it has reached an overall irrevocable state and before it has been archived. This particularly holds true for communities with high volume data (e.g. high-energy physics; climate sciences), and for data citation entities including multiple individual datasets for which the time needed to reach an overall stable data collection is long. Even though our case study analysis found that data citation workflows are implemented or considered by many stakeholder groups involved in data publishing, dynamic data citation challenges have not been widely addressed. Version control and keeping a good provenance record[48] of datasets are also critical for citation of such data collections and are indispensable parts of the data publishing workflow.

With respect to gaps and challenges, we recognize that the case studies we analyzed are limited in scope. This relates to an overall challenge we encountered during the project: it is difficult to find clear and consistent human-readable workflow representations for repositories. The trust standards

[43] http://blog.beagrie.com/2014/04/02/new-research-the-value-and-impact-of-data-curation-and-sharing/

[44] http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf

[45] RDA/WDS Publishing Data Costs IG addresses this topic: https://rd-alliance.org/groups/rdawds-publishing-data-ig.html

[46] http://project-thor.eu/

[47] For example, in genomics, there is the idea of numbered "releases" of, for example, a particular animal genome, so that while refinement is ongoing it is also possible to refer to a reference data set.

[48] For scientific communities with high volume data, the storage of every dataset version is often too expensive. Versioning and keeping a good provenance record of the datasets are crucial for citations of such data collections. Technical solutions are being developed, e.g. by the European Persistent Identifier Consortium (EPIC).

(e.g. Data Seal of Approval[49], Nestor, ISO 16363 and World Data System) require that repositories document their processes, so this may change in the future, but we would add our recommendation that repositories publish their workflows in a standard way for greater transparency. This would bolster confidence in repositories, and also increase user engagement.
The diversity we found is not surprising, nor is it necessarily undesirable. Case studies and ethnographies of data practices have found that workflows for dealing with data 'upstream' of repositories are highly diverse. Data sharing practices vary considerably at the sub-disciplinary level in many cases (e.g. Cragin et al [30]), so there is likely to be continued need to support diverse approaches and informed choice rather than unified or monolithic models (Pryor, [31]). Our analysis shows that a variety of workflows has evolved, and more are emerging, so researchers may be able to choose their best fit on the basis of guidance that distinguishes relevant features, such as QA/QC and different service or support levels.

Best practice recommendations and conclusions

Based on selected case studies, key components in data publishing have been identified, leading to a reference model in data publishing. The analysis, and in particular the conversations with the key stakeholders involved in data publishing workflows, highlighted best practices which might be helpful as recommendations for organizations establishing new workflows and to those seeking to transform or standardize existing procedures:

- Start small and build components one by one in a modular way with a good understanding of how each building block fits into the overall workflow and what the final objective is. These building blocks should be open source/shareable components.
- Follow standards whenever available to facilitate interoperability and to permit extensions based on the work of others using the same standards. For example, Dublin Core is a widely used metadata standard, making it relatively easy to share metadata with other systems. Use disciplinary standards where/when applicable.
- It is especially important to implement and adhere to standards for data citation, including the use of persistent identifiers (PIDs). Linkages between data and publications can be automatically harvested if DOIs for data are used routinely in papers. The use of researcher PIDs such as ORCID can also establish connections between data and papers or other research entities such as software. The use of PIDs can also enable linked open data functionality[50].
- Document roles, workflows and services. A key difficulty we had in conducting the analysis of the workflows was the lack of complete, standardized and up-to-date information about the processes and services provided by the platforms themselves. This impacts potential users of the services as well. Part of the trusted repository reputation development should include a system to clarify ingest support levels, long-term sustainability guarantees, subject expertize resource, and so forth.

In summary, following the idea of the presented reference model and the best practices, we would like to see a workflow that results in all scholarly objects being connected, linked, citable, and persistent to allow researchers to navigate smoothly and to enable reproducible research. This includes linkages between documentation, code, data, and journal articles in an integrated environment. Furthermore, in the ideal workflow, all of these objects need to be well documented to enable other researchers (or citizen scientists etc) to reuse the data for new discoveries. We would like to see information standardized and exposed via APIs and other mechanisms so that metrics on data usage can be captured. We note, however, that biases in funding and academic reward systems need to value data-

[49] http://datasealofapproval.org

[50] At time of writing, CrossRef had recently announced the concept and approximate launch date for a 'DOI Event Tracker', which could also have considerable implications for the perceived value of data publishing as well as for the issues around the associated metrics. (Reference: http://crosstech.crossref.org/2015/03/crossrefs-doi-event-tracker-pilot.html by Geoffrey Bilder, accessed 26 October 2015.)

driven secondary analysis and reuse of existing data, as well as data publishing as a first class object. More attention (i.e. more perceived value) from funders will be key to changing this paradigm.

One big challenge is that there is a need to collaborate more intensively among the stakeholder groups. For example, repositories and higher education institutions (holding a critical mass of research data) and the large journal publishers (hosting the critical mass of discoverable, published research) have not yet fully engaged with each other. Although new journal formats are being developed that link data to papers and enrich the reading experience, progress is still being impeded by cultural, technical and business model issues.

We have demonstrated that the different components of a data publishing system need to work, where possible, in a seamless fashion and in an integrated environment. We therefore advocate the implementation of standards, and the development of new standards where necessary, for repositories and all parts of the data publishing process. Data publishing should be embedded in documented workflows, to help establish collaborations with potential partners and to guide researchers, enabling and encouraging the deposit of reusable research data that will be persistent while preserving provenance.

REFERENCES

1. Schmidt B, Gemeinholzer B, Treloar A (2015). Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey. https://docs.google.com/document/d/1jRM5ZlJ9o4KWIP1GaW3vOzVkXjIIBYONFcd985qTeXE/ed
2. Vines TH, Albert AYK, Andrew RL, DeBarre F, Bock DG, Franklin MT, Kimberly J. Gilbert KJ, Moore JS, Renaut S,Rennison DJ (2014) The availability of research data declines rapidly with article age. Current Biology 24(1): 94-97
3. Hicks D, Wouters P, Waltman L, De Rijcke S, Rafols,I. (2015) Bibliometrics: The Leiden Manifesto for research metrics. Nature, 520(7548).http://www.nature.com/news/bibliometrics-the-leiden-manifesto-for-research-metrics-1.17351 . Accessed 10 November 2015
4. Piwowar H, Vision T. (2013) Data reuse and the open data citation advantage. PeerJ Computer Science. https://peerj.com/articles/175/. Accessed 10 November 2015
5. Pienta AM, Alter GC, & Lyle JA (2010).The enduring value of social science research: The use and reuse of primary research data. http://hdl.handle.net/2027.42/78307. Accessed 10 November 2015
6. Borgman, CL (2015) Big data, little data, no data: Scholarship in the networked world.MIT Press, Cambridge MA
7. 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.org/10.1371/journal.pone.0067332
8. Peng RD (2011) Reproducible research in computational science. Science, 334(6060): 1226-1227.
9. Thayer KA, Wolfe MS, Rooney AA, Boyles AL, Bucher JR, and Birnbaum LS (2014) Intersection of systematic review methodology with the NIH reproducibility initiative. Environmental Health Perspectives. 122: A176–A177. http://ehp.niehs.nih.gov/wp-content/uploads/122/7/ehp.1408671.pdf. Accessed 10 November 2015
10. George BJ, Sobus JR, Phelps LP, Rashleigh B, Simmons JE, Hines RN (2015) Raising the bar for reproducible science at the U.S. Environmental Protection Agency Office of Research and Development. Toxicological Sciences, 145(1):16–22. http://toxsci.oxfordjournals.org/content/145/1/16.full.pdf+html

11. Boulton G. et al (2012) Science as an open enterprise. Royal Society, London. , https://royalsociety.org/policy/projects/science-public-enterprise/Report/ Accessed 10 November 2015
12. Stodden V, Bailey DH, Borwein J, LeVeque RJ, Rider W, and Stein W (2013) Setting the default to reproducible. Reproducibility in computational and experimental mathematics (2013) , Institute for Computational and Experimental Research in Mathematics. https://icerm.brown.edu/tw12-5-rcem/icerm_report.pdf. Workshop report accessed 10 November 2015.
13. Whyte A, Tedds J (2011) Making the case for research data management. DCC briefing papers.Digital Curation Centre, Edinburgh .\ http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm. Accessed 10 November 2015
14. Parsons M, and Fox P (2013) Is data publication the right metaphor? Data Science Journal 12. http://doi.org/10.2481/dsj.WDS-042 -. Accessed 10 November 2015
15. Rauber A, Pröll, S (2015) Scalable dynamic data citation approaches, reference architectures and applications RDA WG Data Citation position paper. Draft version. https://rd-alliance.org/groups/data-citation-wg/wiki/scalable-dynamic-data-citation-rda-wg-dc-position-paper.html. Accessed 13 November 2015
16. Rauber A, Asmi A, van Uytvanck D, Pröll, S (2015) Data citation of evolving data: recommendations of the Working Group on Data Citation (WGDC) Draft – Request for comments. Revision of September 24th 2015. https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_150924.pdf Accessed 6 November 2015
17. Watson et al (2009) The XMM-Newton serendipitous survey. V. The Second XMM-Newton serendipitous source catalogue, Astronomy and Astrophysics, Volume 493, Issue 1, 2009, pp.339-373, DOI: 10.1051/0004-6361:200810534
18. Lawrence B, Jones C, Matthews B, Pepler S, Callaghan S (2011) Citation and peer review of data: Moving toward formal data publication. The International Journal of Digital Curation. doi:10.2218/ijdc.v6i2.20r
19. Callaghan S, Murphy F, Tedds J, Allan R, Kunze J, Lawrence R, Mayernik MS, Whyte A (2013) Processes and procedures for data publication: A case study in the geosciences.The International Journal of Digital Curation, 8(1). doi:10.2218/ijdc.v8i1.253
20. Austin CC, Brown S, Fong N, Humphrey C, Leahey L, Webster P (2015). Research data repositories: Review of current features, gap analysis, and recommendations for minimum requirements. Presented at the IASSIST Annual Conference, Minneapolis MN, June 2-5; IASSIST Quarterly Preprint. International Association for Social Science, Information Services, and Technology. https://drive.google.com/file/d/0B_SRWahCB9rpRF96RkhsUnh1a00/view. Accessed 13 November 2015
21. Yin, R (2003) Case study research: Design and methods. Fifth edition. Sage Publications, Thousand Oaks, CA
22. Murphy F, Bloom T, Dallmeier-Tiessen S, Austin CC, Whyte A, Tedds J, Nurnberger A, Raymond L, Stockhause M, Vardigan M (2015). WDS-RDA-F11 Publishing Data Workflows WG Synthesis FINAL CORRECTED. Zenodo. 10.5281/zenodo.33899. Accessed 17 November 2015.
23. Stockhause M, Höck H, Toussaint F, Lautenschlager M (2012) Quality assessment concept of the World Data Center for Climate and its application to the CMIP5 data. Geoscientific Model Development 5(4):1023-1032. doi:10.5194/gmd-5-1023-2012
24. Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, Hourclé J, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T. (2015) Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science 1(e1) https://dx.doi.org/10.7717/peerj-cs.1
25. Castro E., Garnett A. (2014) Building a bridge between journal articles and research data: The PKP-Dataverse Integration Project. International Journal of Digital Curation, 9(1):176-184. doi:10.2218/ijdc.v9i1.311

26. Mayernik MS., Callaghan S, Leigh R, Tedds JA, Worley S.(2015) Peer Review of datasets: When, why, and how. Bulletin of the American Meteorological Society, 96(2): 191–201. http://dx.doi.org/10.1175/BAMS-D-13-00083.1
27. Meehl, GA, Moss R, Taylor KE, Eyring V Stouffer RJ, Bony S and Stevens B (2014). Climate Model Intercomparisons: Preparing for the Next Phase, Eos Trans. AGU, 95(9), 77. doi:10.1002/2014EO090001
28. Bandrowski A, Brush M, Grethe JS, Haendel MA, Kennedy DN, Hill S, Hof PR, Martone ME, Pols M, Tan S, Washington N, Zudilova-Seinstra E, Vasilevsky N. The Resource Identification Initiative: A cultural shift in publishing [version 1; referees: 2 approved] *F1000Research*2015, **4**:134 (doi: 10.12688/f1000research.6555.1).
29. Brase J; Lautenschlager M; Sens I (2015). "The Tenth Anniversary of Assigning DOI Names to Scientific Data and a Five Year History of DataCite." D-Lib Magazine. 21(1/2)r. doi: 10.1045/january2015-brase
30. Cragin MH,Palmer CL, Carlson JR, Witt M (2010) Data sharing, small science and institutional repositories.Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368(1926):4023 –4038.
31. Pryor (2009) Multi-scale data sharing in the life sciences: Some lessons for policy makers.International Journal of Digital Curation 4(3):, 71-82. doi:10.2218/ijdc.v4i3.115