A peer-reviewed version of this preprint was published in PeerJ on 27 May 2015.

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://doi.org/10.7717/peerj-cs.1

Achieving human and machine accessibility of cited data in scholarly publications

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

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

ABSTRACT

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.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

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

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. 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 (Alsheikh-Ali et al. (2011)). 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 and Ellis (2012); Prinz et al. (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 1965-78, 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 and Pop (2008); Shendure and 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 (Altman et al. (2013); Altman and King (2006); Uhlir (2012); Ball and Duke (2012)). Uniform citation of robustly archived, 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 (Altman et al. (2013); Altman and King (2006); Uhlir (2012); Ball and Duke (2012); Goodman et al. (2014); Borgman (2012); Parsons et al. (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 and 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 out-

[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.

[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.

lined 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 (Altman et al. (2013)).

Three common practices in academic publishing today block the systematic reuse of data. One 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 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); Altman et al. (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 (Altman et al. (2013); Uhlir (2012); Ball and Duke (2012); Altman and 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."

[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 (U.S.), 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).*

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

- ***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."

- ***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 archival 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 et al. (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

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

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.

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 and Taylor (2002); Richardson and 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 and 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 et al. (1998); Jacobs and Walsh (2004))[5].

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.).

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

[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.)

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

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 HTTP(s) 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 resolver at `http://doi.org` resolves the DOI `10.1098/rsos.140216` to the URI `http://rsos.royalsocietypublishing.org/content/1/3/140216`. And the identifiers.org resolution service, at `http://identifiers.org`, resolves the PubMed identifier 16333295 to `http://www.ncbi.nlm.nih.gov/pubmed/16333295`. 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.

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 and 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 (DOIs) 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 U.S. 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. 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)).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

Handle System (HDLs)

Handles (HDLs) 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 et al. (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, 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 et al. (1998); Jacobs and 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 (ARKs) 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 and Starr (2006); Kunze (2003); Kunze and Rodgers (2001); Janée et al. (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 Bibliothèque Nationale de France, and the U.S. National Library of Medicine (Kunze and Starr (2006); Peyrard et al. (2014); Kunze (2012)).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

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
HTTP(s) URI	HTTP or HTTP(s) 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

Table 1. Examples of identifier schemes meeting JDDCP criteria.

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

National Bibliography Number (NBNs)

National Bibliography Numbers (NBNs) 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 et al. (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. Landing pages should combine human-readable and machine-readable information on a selection of the following items.

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.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

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
HTTP(s) 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

Table 2. Identifier scheme persistence and object removal behavior

[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.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

- (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 and Newman (2002)).
- **Version**: The dataset version identifier (if applicable).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

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/).

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 et al. (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 and 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 and 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]."*

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

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

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*[7] and *F1000Research*[8].

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[9] Data Citation Implementation Group [10], during 2014, as a follow-on project to the successfully concluded Joint Data Citation Principles effort.

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

[7] http://www.nature.com/sdata/data-policies/repositories

[8] http://f1000research.com/for-authors/data-guidelines

[9] Force11.org (http://force11.org) is a community of scholars, librarians, archivists, publishers and research funders that has arisen organically to help facilitate the change toward improved knowledge creation and sharing. It is incorporated as a US 501(c)3 not-for-profit organization in California.

[10] (DCIG, https://www.force11.org/datacitationimplementation)

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

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.

ACKNOWLEDGMENTS

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.

REFERENCES

Accomazzi, A., Henneken, E., Erdmann, C., and Rots, A. (2012). Telescope bibliographies: an essential component of archival data management and operations. volume 8448 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*.

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., and Ioannidis, J. P. A. (2011). Public availability of published research data in high-impact journals. *PLoS ONE*, 6(9):e24357.

Altman, M., Arnaud, E., Borgman, C., Callaghan, S., Brase, J., Carpenter, T., Chavan, V., Cohen, D., Hahnel, M., Helly, J., Kishor, P., Li, J., Linares, F., McMahon, B., MorgenRoth, K., Muryama, Y., Murphy, F., Palnisamy, G., Parsons, M., Roug, S., Sagen, H., Smit, E., VanDeneter, M., Uhlir, P., Zettsu, K., and Socha, Y. E. (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.

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

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

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

Begley, C. G. and Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533.

Berners-Lee, T., Fielding, R., and Masinter, L. (1998). Rfc2396: Uniform resource identifiers (uri): Generic syntax. `https://www.ietf.org/rfc/rfc2396.txt`.

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

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

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F. (2008). Extensible markup language (xml) 1.0 (fifth edition): W3c recommendation 26 november 2008. `http://www.w3.org/TR/REC-xml/`.

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

CNRI (2014). Handle system: Unique and persistent identifiers for internet resources. http://www.w3.org/TR/webarch/#identification.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, 1(3).

Data Citation Synthesis Group (2014). Joint declaration of data citation principles. http://force11.org/datacitation.

Data Documentation Initiative (2012). Data documentation initiative specification. 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. 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. 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. http://dublincore.org/documents/dces/.

Dyson, E. (2003). Online registries: The dns and beyond... http://doi.contentdirections.com/reprints/dyson_excerpt.pdf.

ECMA (2013). Ecma-404: The json data interchange format. http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf.

Editors (2010). Credit where credit is due. *Nature*, 462(7275):825.

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

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

Gao, S., Sperberg-McQueen, C. M., and Thompson, H. S. (2012). W3c xml schema definition language (xsd) 1.1 part 1: Structures: W3c recommendation 5 april 2012. http://www.w3.org/TR/xmlschema11-1/.

GitHub Guides (2014). Making your code citable. https://guides.github.com/activities/citable-code/.

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

Gray, A., Dumontier, M., Marshall, M., Baran, J., Ansell, P., Bader, G., Bando, A., Callahan, A., Cruz-toledo, J., Gombocz, E., Gonzalez-Beltran, A., Groth, P., Haendel, M., Ito, M., Jupp, S., Katayama, T., Krishnaswami, K., Lin, S., Mungall, C., Le Novere, N., Laibe, C., Juty, N., Malone, J., and Rietveld, L. (2014). Data catalog vocabulary (dcat): W3c recommendation, 16 january 2014. http://www.w3.org/2001/sw/hcls/notes/hcls-dataset/.

Greenberg, S. A. (2009). How citation distortions create unfounded authority: analysis of a citation network. *BMJ*, 339.

Gudgin, M., Hadley, M., Mendelsohn, N., Moreau, J.-J., Nielsen, H. F., Karmarkar, A., and Lafon, Y. (2007). Soap version 1.2 part 1: Messaging framework (second edition): W3c recommendation 27 april 2007. http://www.w3.org/TR/soap12-part1/.

Haas, H. and Brown, A. (2004). Web services glossary: W3c working group note 11 february 2004. http://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/#webservice.

Hakala, J. (2001). Rfc3188: Using national bibliography numbers as uniform resource names. https://tools.ietf.org/html/rfc3188.

Hilse, H.-W. and Kothe, J. (2006). Implementing persistent identifiers. http://xml.coverpages.org/ECPA-PersistentIdentifiers.pdf.

Holtzman, K. and Mutz, A. (1998). Rfc2295: Transparent content negotiation in http. https://www.ietf.org/rfc/rfc2295.txt.

Hourclé, J., Chang, W., Linares, F., Palanisamy, G., and Wilson, B. (2012). Linking Articles to Data.

International DOI Foundation (2014). Doi handbook. http://www.doi.org/hb.html.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, 2(8):e124.

ISO/TC 211 (2014). Iso 19115-1:2014: Geographic information metadata, part 1:fundamen-

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

tals. http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53798.
Jacobs, I. and Walsh, N. (2004). Architecture of the world wide web, volume one w3c recommendation 15 december 2004. http://www.w3.org/TR/webarch/#identification.
Janée, G., Kunze, J., and Starr, J. (2009). Identifiers made easy. http://ezid.cdlib.org/.
Juty, N., Le Novère, N., and Laibe, C. (2012). Identifiers.org and miriam registry: community resources to provide persistent identification. *Nucleic Acids Research*, 40(D1):D580–D586.
Klyne, G. and Newman, C. (2002). Rfc 3339: Date and time on the internet: Timestamps. http://www.ietf.org/rfc/rfc3339.txt.
Kunze, J. (2003). Towards electronic persistence using ark identifiers. *Proceedings of the 3rd ECDL Workshop on Web Archives. August 21st, 2003, Trondheim, Norway*.
Kunze, J. (2012). The ark identifier scheme at ten years old. Technical report, California Digital Library.
Kunze, J. and Rodgers, R. (2001). The ark identifier scheme. Technical report, Internet Engineering Task Force.
Kunze, J. and Starr, J. (2006). Ark (archival resource key) identifiers. http://www.cdlib.org/inside/diglib/ark/arkcdl.pdf.
Lagoze, C. and Van de Sompel, H. (2007). Compound Information Objects: The OAI-ORE Perspective.
Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R., and Warner, S. (2008). Ore user guide - resource map discovery. http://www.openarchives.org/ore/1.0/discovery.
Library of Congress (1997). The relationship between urns, handles, and purls. http://memory.loc.gov/ammem/award/docs/PURL-handle.html.
Mali, F., Erickson, J., and Archer, P. (2014). Data catalog vocabulary (dcat): W3c recommendation, 16 january 2014. http://www.w3.org/TR/vocab-dcat/.
Maunsell, J. H. (2014). Credit where credit is due. *The Journal of Neuroscience*, 34(21):7043.
Moats, R. (1997). Rfc2141: Uniform resource name syntax. https://tools.ietf.org/html/rfc2141.
National Center for Biotechnology Information (2014). http://jats.nlm.nih.gov/publishing/tag-library/1.1d2/index.html.
Nottingham, M. (2010). Rfc5988: Web linking. https://www.ietf.org/rfc/rfc5988.txt.
OCLC ((accessed January 2, 2015)). Purl help. https://purl.org/docs/help.html.
Parsons, M. A., Duerr, R., and Minster, J.-B. (2010). Data citation and peer review. http://dx.doi.org/10.1029/2010EO340001.
Peterson, D., Gao, S., Malhotra, A., Sperberg-McQueen, C. M., and Thompson, H. S. (2012). W3c xml schema definition language (xsd) 1.1 part 2: Datatypes: W3c recommendation 5 april 2012. http://www.w3.org/TR/xmlschema11-1/.
Peyrard, S., Kunze, J., and Tramoni, J.-P. (2014). The ark identifier scheme: Lessons learnt at the bnf. *DC-2014 – The Austin Proceedings*.
Prinz, F., Schlange, T., and 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.
Rans, J., Day, M., Duke, M., and Ball, A. (2013). Enabling the citation of datasets generated through public health research. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp051762.PDF.
Rekdal, O. B. (2014). Academic urban legends. *Social Studies of Science*, 44(4):638–654.
Richardson, L. and Ruby, S. (2011). *RESTful Web Services*. O'Reilly, Sebastopol CA.
Salzberg, S. L. and Pop, M. (2008). Bioinformatics challenges of new sequencing technology. *Trends in Genetics*, 24:142–149.
Shendure, J. and Ji, H. (2008). Next-generation dna sequencing. *Nature Biotechnology*, 26:1135–1145.
Shepherd, Fiumara, Walters, Stanton, Swisher, Lu, Teoli, Kantor, and Smith (2014). Content negotiation - http. 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.
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.
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.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

Vasilevsky, N. A., Brush, M. H., Paddock, H., Ponting, L., Tripathy, S. J., LaRocca, G. M., and Haendel, M. A. (2013). On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ*, 1:e148.

APPENDIX

Serving landing pages: implementation details

Ideally, all versions of the landing page would be resolvable from a single URI through content negotiation (Holtzman and 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.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015

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.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.697v4 | CC-BY 4.0 Open Access | rec: 11 Feb 2015, publ: 11 Feb 2015