Title

A data citation roadmap for scholarly data repositories.

Journal

Scientific data, 6(1)

ISSN

2052-4463

Authors

Fenner, Martin

Crosas, Mercè

Grethe, Jeffrey S

et al.

Publication Date

2019-04-01

DOI

10.1038/s41597-019-0031-8

SCIENTIFIC DATA

OPEN

ARTICLE

A data citation roadmap for scholarly data repositories

Martin Fenner[1], Mercè Crosas[2], Jeffrey S. Grethe[3], David Kennedy[4], Henning Hermjakob[5], Phillippe Rocca-Serra[6], Gustavo Durand[2], Robin Berjon[7], Sebastian Karcher[8], Maryann Martone[3] & Tim Clark[9]

Received: 2 October 2017
Accepted: 12 March 2019
Published online: 10 April 2019

This article presents a practical roadmap for scholarly data repositories to implement data citation in accordance with the Joint Declaration of Data Citation Principles, a synopsis and harmonization of the recommendations of major science policy bodies. The roadmap was developed by the Repositories Expert Group, as part of the Data Citation Implementation Pilot (DCIP) project, an initiative of FORCE11.org and the NIH-funded BioCADDIE (https://biocaddie.org) project. The roadmap makes 11 specific recommendations, grouped into three phases of implementation: a) required steps needed to support the Joint Declaration of Data Citation Principles, b) recommended steps that facilitate article/data publication workflows, and c) optional steps that further improve data citation support provided by data repositories. We describe the early adoption of these recommendations 18 months after they have first been published, looking specifically at implementations of machine-readable metadata on dataset landing pages.

Introduction

The Joint Declaration of Data Citation Principles (JDDCP) published in 2014[1] and endorsed by a large number of scholarly and academic publishing organizations, lays out a set of principles on purpose, function and attributes of data citations. The first of these principles stresses that data should be considered legitimate, citable products of research[2]. The JDDCP condenses the results of substantial prior studies on science policy and practice[3–5].

The JDDCP intentionally focuses on data citation principles, as the implementation of these principles will differ across disciplines and communities. The roadmap presented here aims to provide practical guidance for repositories on implementing these data citation principles with a focus on life sciences, based on earlier work in this area, in particular Starr *et al.*[6] and Altman and Crosas[7], and are consistent with recent recommendations regarding data, code and workflows[8,9]. These recommendations for data repositories complement the DCIP project recommendations for publishers[10] and for globally unique resolution of Compact Identifiers[11]. While related recommendations might differ in implementation detail, we do not know of any conflicting recommendations that the reader should be aware of.

Data repositories play a central role in data citation, as they provide stewardship and discovery services to find data, give persistent access to the data being cited, and provide unique identifiers and metadata needed for data citation. For data citation, repositories need to work closely with a variety of stakeholders, including publishers, reference manager providers, data users, and of course researchers. Data citation practices and technologies supported by repositories will substantially assist development of new data discovery indexes such as DataMed[12] and Google Dataset Search (https://toolbox.google.com/datasetsearch).

[1]DataCite, Hannover, Germany. [2]Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA. [3]University of California San Diego, La Jolla, CA, USA. [4]University of Massachusetts Medical School, Worcester, MA, USA. [5]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Hinxton, Cambridgeshire, UK. [6]Oxford e-Research Centre, University of Oxford, Oxford, UK. [7]Standard Analytics, New York, NY, USA. [8]Qualitative Data Repository, Syracuse University, Syracuse, NY, USA. [9]University of Virginia School of Medicine, Charlottesville, VA, 22903, USA. These authors contributed equally: Martin Fenner and Mercè Crosas. Correspondence and requests for materials should be addressed to T.C. (email: twc8q@virginia.edu)

Level	#	Guideline
Required	1	All datasets intended for citation *must* have a globally unique persistent identifier that can be expressed as an unambiguous URL.
	2	Persistent identifiers for datasets *must* support multiple levels of granularity, where appropriate.
	3	The persistent identifier expressed as an URL *must* resolve to a landing page specific for that dataset, and that landing page must contain metadata describing the dataset.
	4	The persistent identifier *must* be embedded in the landing page in machine-readable format.
	5	The repository must provide documentation and support for data citation.
Recommended	6	The landing page *should* include metadata required for citation, and ideally also metadata facilitating discovery, in human-readable and machine-readable format.
	7	The machine-readable metadata *should* use schema.org markup in JSON-LD format.
	8	Metadata *should* be made available via HTML meta tags to facilitate use by reference managers.
	9	Metadata *should* be made available for download in BibTeX and/or another standard bibliographic format.
Optional	10	Content negotiation for schema.org/JSON-LD and other content types *may* be supported so that the persistent identifier expressed as URL resolves directly to machine-readable metadata.
	11	HTTP link headers *may* be supported to advertise content negotiation options

Table 1. Guidelines for Repositories.

Results

The guidelines are grouped into three phases: required, recommended and optional. Implementing these guidelines takes time and resources, it is therefore not only critical to provide specific guidelines, but also to give guidance on priorities: work needed to support the Joint Declaration of Data Citation Principles (required phase), additional work to facilitate article/data publishing workflows in collaboration with publishers (recommended phase), and extra work to support data citation that can be done by data repositories (optional phase). The Guidelines are summarized in Table 1, and are discussed in detail in the text following the table.

Details of each recommendation follow, with examples.

Persistent identifiers. A data citation *must* include a persistent method for identification that is machine actionable, globally unique, and widely used by a community (**JDDCP**, principle #4). The use of the persistent identifier should follow community best practices[6,13–16]. For implementation by data repositories, this means:

- **Persistent method for identification**. Unique identifiers, and metadata describing the data, and its disposition, *must* persist–even beyond the lifespan of the data they describe (**JDDCP**, principle #6). As an extension to this principle, data repositories should make provisions to keep unique identifiers and metadata available beyond the lifespan of the data or repository, ideally in a well-recognized and accepted standard metadata format.
- **Machine actionable**. The persistent identifier *must* be understood, and be resolvable, as an HTTP URI in accordance with IETF RFC 3986[16,17], including support for content negotiation[18].
- **Globally unique**. The identifier *must* use a prefix (namespace) if the identifier character string is only unique within a particular database, e.g. an accession number; and the prefix must be registered with a robust, institutionally stable global resolver such as the identifiers.org system at EMBL/EBI[11].
- **Widely used by a community**. The persistent identifier must be widely used in the community. For the life sciences this includes accession numbers, in combination with the database name for global uniqueness.

Persistent identifier granularity. Persistent identifiers for datasets must support multiple levels of granularity to support both the citation of a specific version and/or individual dataset, as well the citation of an unspecified version of a dataset and/or a collection of primary data. The levels of granularity supported by persistent identifiers must be documented.

In many domains, primary data is uniquely identified and cited as a collection of potentially many individual items. At the same time, these individual items need their own unique identifiers to support later reuse and recombination into different sets while maintaining the ability to cite the constituent data elements. An example is in the field of neuroimaging, where individual subject scans using a given imaging modality are the lowest level at which objects will be identified, while the primary publication will cite a collection level unique identifier. This imposes a requirement that lower-level identifiers need to be able to be grouped via a collection identifier and accessed as set elements from the overall collection landing page 18. Another example is the BioStudies database[19], which can provide storage for all the underlying data links and files for a publication.

Only in circumstances where multiple levels do not inherently exist in the data, i.e. no collections or other groupings exist, may this requirement be waived.

Landing pages. The persistent identifier expressed as HTTP URL must resolve to a specific landing page for that dataset or dataset collection. The persistent identifier expressed as HTTP URL must not resolve to the data itself 6, or to other representations of the metadata, unless special protocols such as content negotiation are used (see guideline 7 below). Relationships of the citation reference, repository landing page and underlying data are shown in Fig. 1.

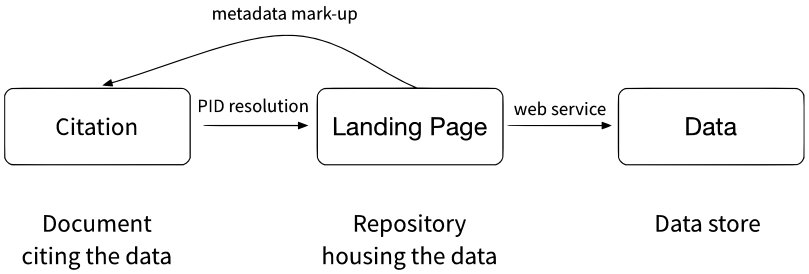

Fig. 1 Generic data citation - relationships of the citation reference, repository landing page and underlying data.

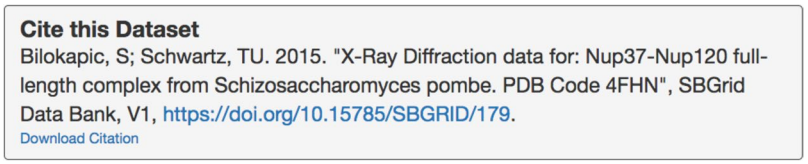

Fig. 2 Providing information about how a dataset should be cited, with download link for citation (in BibTex or other standard bibliographic reference manager format).

The landing pages *must* provide metadata with additional information about the dataset, and include links for accessing the dataset itself. The landing page *should* provide definitive information, including metadata, on how the dataset should be cited, other descriptive information about the dataset, as well as data accessibility and licensing information. Repositories should provide a landing page for every dataset or collection of datasets intended to be cited, which could be single entries, sets of entries, the entire repository or a curated database[6].

Reference to a statement describing the data and metadata persistence policies of the repository should also be provided at the landing page. Data persistence policies will vary by repository but should be clearly described, for example (using text template from[6]):

"[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]."

Figure 2 provides an example for how "Cite this Dataset" information can look in a landing page.

Persistent identifiers on landing pages. To verify that a persistent identifier resolves to a correct landing page, the persistent identifier *must* be embedded in the landing page in human-readable and machine-readable formats. This enables checks that the persistent identifier properly resolves to a landing page describing that identifier, and enables basic data citation by reference managers, and minimal validation by the publisher of persistent identifiers cited in documents. The persistent identifier should be found somewhere on the landing page, but is ideally embedded in schema.org markup and/or using HTML meta tags.

Example schema.org/JSON-LD

```
<application type="application/ld+json">
{"@id": "https://doi.org/10.5061/dryad.q447c/3"}
</application>
```

Example HTML meta tags

```
<meta name="DC.identifier" content="https://doi.org/10.5061/dryad.
q447c/3">
```

Documentation and author support. The repository *must* provide documentation about how data should be cited, how metadata can be obtained, and who to contact for more information. This documentation should follow the recommendations in this document, the DCIP Data Citation Primer[20], community recommendations provided by a number of organizations, but should also address the specifics of that particular data repository.

Metadata on landing pages. Landing pages should provide metadata required for data citation in both human- and machine-readable format, and should be accessible without requiring authentication. The landing page should show the citation metadata in human-readable form, e.g. formatted in one or more citation styles common to the community in a Cite this Dataset field and, possibly, provide means of copying/downloading

Citation Metadata	Dublin Core[a]	Schema.org[b]	DataCite[c]	DATS[d]
Dataset Identifier	identifier	@id*	identifier	identifier
Title	title	name	title	title
Creator**	creator	author	creator	creator
Data repository or archive	publisher	publisher	publisher	publisher
Publication Date	date	datePublished	publicationYear	date
Version	*not available*	version	version	version
Type	type	type	resourceTypeGeneral	type

Table 2. Citation metadata for Data Repositories. Key: [a]Dublin Core Metadata Element Set (https://dublincore.org/documents/dces/); [b]Dataset - Schema.org (https://schema.org/Dataset); [c]DataCite Metadata Working Group[21]; [d]Gonzalez-Beltran & Rocca-Serra[22,23]; *name of ID field depends on schema.org serialization format, it is **@id** for JSON-LD; **not all datasets will have "the main researchers involved in producing the data" (DataCite Schema), in which case the more generic "An entity primarily responsible for making the resource" from Dublin Core should be used, and this can also be an organization.

Discovery Metadata	Dublin Core	Schema.org	DataCite	DATS
Description	description	description	description	dataType dimension Material…*
Keywords	subject	keywords	subject	keywords
License	license	license	rights	license
Related Dataset**	isPartOf isVersionOf references	isPartOf citation	relatedIdentifier	isPartOf
Related Publication***	bibliographicCitation	citation	relatedIdentifier	publication

Table 3. Important discovery metadata for Data Repositories. Key: *DATS provides much more detailed metadata to describe a biomedical dataset; **related datasets can have part/whole relations (IsPartOf, etc.), version relations (IsVersionOf, etc.) or reference relations (references); ***related publications reference a dataset published previously, reference a dataset published in parallel with the publication, or otherwise document a dataset.

the citation as text. The landing page should also show all versions, or link to a page with version information. A visible link to machine-readable metadata should be provided.

The metadata elements needed for data citation are given in Table 2.

All metadata fields required for citation are part of Dublin Core (with the exception of *version*), the core schema.org specification, and by extension Bioschemas (https://bioschemas.org), as well as the DataCite and DATS metadata schemas[21–23].

In addition to the metadata required for citation, it is recommended to provide additional metadata on landing pages – again in human-readable and machine-readable formats – that help with data discovery, as shown in Table 3.

The metadata standards Dublin Core, schema.org and DataCite by their very nature of being generic only provide some metadata helpful for discovery, while DATS can provide much more detailed information about a biomedical dataset. Further information can be found in the DATS specification[24].

Information about related datasets should be provided where possible, as should information about related publications. They provide important information that can help with discovery. When a data repository knows about a publication citing a dataset, this information should be included in the metadata, complementing the information about the dataset found in the citing publication and enabling navigation between publication and dataset in both directions.

Metadata on landing pages using schema.org/JSON-LD. All dataset landing pages *should* provide machine-readable metadata using schema.org markup in JSON-LD format. JSON-LD is the easiest way to represent schema.org metadata, and is also used to represent DATS metadata in schema.org format[23,24]. The JSON-LD should be embedded in the HTML page using a `<script type="application/ld+json">` tag.

Examples

```
<script type="application/ld+json">
{"@context": "http://schema.org",
"@type": "Dataset",
"@id": "https://doi.org/10.3886/ICPSR08001.v2",
"name": "Cancer Surveillance and Epid
emiology in the
United States and Puerto Rico, 1973-1977 (ICPSR 8001)",
"author": "National Cancer Institute",
```

```
"publisher": "ICPSR - Interuniversity Consortium for Political and Social
Research",
"datePublished": "1984-05-03",
"dateModified": "2015-08-06T11:20:58Z",
"version": "v2",
"Description": "This dataset was produced as part of the Surveillance,
Epidemiology, and End Results (SEER) Program to monitor the incidence of
cancer and cancer survival rates in the United States, thus carrying out
the mandates of the National Cancer Act. The SEER Program had several objec-
tives: to estimate the annual cancer incidence in the United States, to
examine trends in cancer patient survival, to identify cancer etiologic
factors, and to monitor trends in the incidence of cancer in selected geo-
graphic areas with respect to demographic and social characteristics…"}
</script>
<script type="application/ld+json">
{"@context": "http://schema.org",
"@type": "Dataset",
"@id": "https://doi.org/10.2210/pdb5m95/pdb",
"name": "STAPHYLOCOCCUS CAPITIS DIVALENT METAL ION TRANSPORTER (DMT) IN
COMPLEX WITH MANGANESE",
"author": [
{"@type": "Person",
"givenName": " I.A.",
"familyName": "Ehrnstorfer"},
{"@type": "Person",
"givenName": " E.R.",
"familyName": " Geertsma"},
{"@type": "Person",
"givenName": "E.",
"familyName": " Pardon"},
{"@type": "Person",
"givenName": " J.",
"familyName": " Steyaert"},
{"@type": "Person",
"givenName": " R.",
"familyName": " Dutzler"}],
"datePublished": "2016-11-30",
"publisher": "Protein Data Bank, Rutgers University",
"citation": [
{
"@type": "ScholarlyArticle",
"@id": "https://doi.org/10.1038/nsmb.2904"
}]}</script>
```


For further examples please use DataCite Search (https://search.datacite.org/), which has embedded schema.org/JSON-LD metadata on every search result page for a single dataset for more than five million datasets.

Metadata via HTML Meta Tags. Data repositories *should* offer machine-readable metadata on landing pages using Highwire, PRISM[25], and/or Dublin Core HTML meta tags. These HTML meta tags are currently the preferred method of reference managers to extract the persistent identifier or full citation metadata from landing pages, as reference managers currently don't routinely support schema.org/JSON-LD metadata extraction.

Example


```
<meta name="DC.identifier" content="doi:10.1594/PANGAEA.727206"
scheme="DCTERMS.URI"/>
<meta name="DC.title" content="Landings of European lobster (Homarus gam-
marus) and edible crab (Cancer pagurus) from 1615 to 2009, Helgoland,
North Sea"/>
<meta name="DC.creator" content="Schmalenbach, Isabel"/>
<meta name="DC.creator" content="Mehrtens, Folke"/>
<meta name="DC.creator" content="Janke, Michael"/>
<meta name="DC.creator" content="Buchholz, Friedrich"/>
<meta name="DC.publisher" content="PANGAEA"/>
<meta name="DC.date" content="2011-01-28" scheme="DCTERMS.W3CDTF"/>
<meta name="DC.type" content="Dataset"/>
```


Metadata via downloadable file in standard bibliographic format. Repositories *should* provide a download link in a common bibliographic format – e.g. `bib` (BibTeX file format) and/or. `ris` (RIS file format) – on the landing page of the dataset. The file should include all metadata required for a data citation.

Example: BibTeX
```
@data{25240_2014,
author={Figueiredo, Dalson and Rocha, Enivado and Paranhos, Ranulfo and
Alexandre, José},
publisher={Harvard Dataverse},
title={How can soccer improve statistical learning?},
year={2014}, doi={10.7910/DVN/25240},
url={https://doi.org/10.7910/DVN/25240}}
```

Example: RIS
```
TY - DATAT1 - How can soccer improve statistical learning?
A1 - Figueiredo, Dalson
A1 - Rocha, Enivaldo
A1 - Paranhos, Ranulfo
A1 - Alexandre, José
Y1 - 2014
DO - 10.7910/DVN/25240
UR - https://doi.org/10.7910/DVN/25240
ER -
```

Content negotiation for machine-readable metadata. Persistent identifiers expressed as HTTP URI *must* by default resolve to the landing page for that dataset (see guideline #3). Data repositories and identifier service providers such as identifiers.org, N2T or DataCite in addition *may* implement HTTP content negotiation[26] for the persistent identifier expressed as HTTP URI, returning machine readable metadata in various formats. Content negotiation is for example supported by identifiers.org and DataCite and can return metadata in RDF-XML, BibTeX, schema.org and other metadata formats.

Example: Image Attribution Framework (IAF)
```
curl -H "Accept: application/xml"
http://iaf.virtualbrain.org/lp/10.18116/C6WC71
```
In addition, the HTML version of this page has a link to the XML (available without content negotiation at http://iaf.virtualbrain.org/lp/xml/10.18116/C6WC71).

Examples: DataCite
```
curl -LH "Accept: application/ld+json" http://doi.org/10.5061/DRYAD.8290N
curl -LH "Accept: application/vnd.citationstyles.csl+json"
http://doi.org/10.5061/DRYAD.8290N
```
Metadata in `application/vnd.citationstyles.csl + json` format are used as input by many reference managers, e.g. Zotero or Mendeley.

Support HTTP link headers. The persistent identifier (see guideline #2) and available content negotiation options (see guideline #9) *may* be provided in a HTTP link header[27]. This facilitates discovery of content negotiation options and makes it easier to fetch the identifier from large landing pages, as only a HTTP head request is needed).

Example
```
curl -I https://search.datacite.org/works/10.5061/dryad.q447c/3
HTTP/1.1 200 OK
Content-Type: text/html;charset=utf-8
Status: 200 OK
Link:<https://doi.org/10.5061/dryad.q447c/3>; rel="identifier",
<https://doi.org/10.5061/dryad.q447c/3>; rel="describedby";
type="application/vnd.datacite.datacite+xml",
<https://doi.org/10.5061/dryad.q447c/3>; rel="describedby";
type="application/ld+json",
<https://doi.org/10.5061/dryad.q447c/3>; rel="describedby";
type="application/vnd.citationstyles.csl+json",
<https://doi.org/10.5061/dryad.q447c/3>; rel="describedby";
type="application/x-bibtex"
```

Discussion

This document provides a roadmap for scholarly data repositories to implement support for data citation. Most if not all Required steps have already been implemented by many data repositories, and little if any work is needed by them to fully support the Joint Declaration of Data Citation Principles. More work is still needed to implement the Recommended steps, including support for schema.org/JSON-LD markup embedded into dataset landing pages. Data repositories that have implemented the required and recommended steps might be interested to look into the Optional steps for extra data citation support.

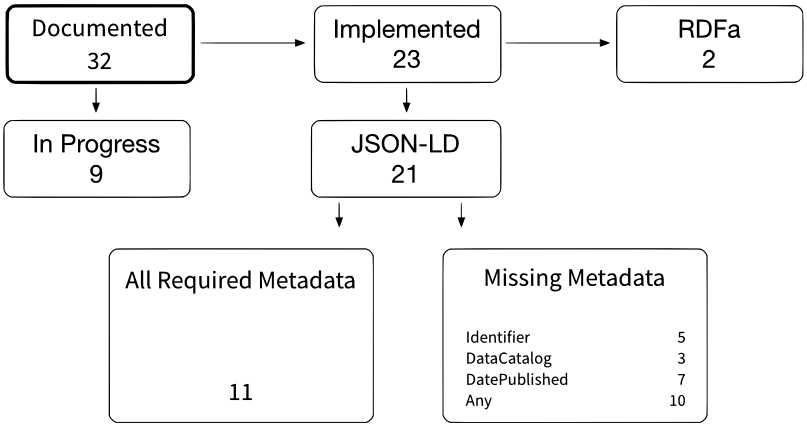

Fig. 3 Implementation status of Schema.org metadata in repository landing pages.

The Data Citation Implementation Pilot and this document focus on data citation support in scholarly data repositories. Using persistent identifiers, standard machine-readable metadata and landing pages of course not only supports data citation, but also facilitates data discovery. Data discovery requires more specific metadata than the metadata needed for data citation, and it is facilitated by a central index of all datasets. The NIH BD2K bioCADDIE project, of which the Data Citation Implementation Pilot is a small part, has developed standard metadata for biomedical data with DATS, and on a central index to search a large number of biomedical datasets with DataMed (https://datamed.org/). The European ELIXIR (https://www.elixir-europe.org/ project (https://www.elixir-europe.org/) in life sciences, and DataCite (all disciplines), are also working on standard metadata and a search index for data discovery. Both Elixir and DataCite are closely collaborating with bioCADDIE in these activities. The NIH Data Commons Pilot, which began in 2018, will further extend this work, and several of the authors of this document have participated in this project[28].

The data citation roadmap for scholarly data repositories described in this document is an important step towards full data citation support by data repositories. Going forward, a lot of work is still needed to fully implement these guidelines, and ongoing coordination amongst data repositories, publishers and other important stakeholders will be essential in this activity.

Methods

This roadmap was developed based on numerous discussions of the DCIP Repositories Early Adopters Expert Group, led by Martin Fenner and Mercè Crosas, including two in-person workshops in February (Boston) and June (San Diego) 2016, and in close coordination with the other DCIP expert groups. The resulting guidelines have been widely circulated since their first publication as a preprint on bioRXiv[29]. A course on the guidelines and how to implement them, was held at the FORCE11 Scholarly Communication Institute (FSCI) in August of 2017. The course instructors were Martin Fenner and Gustavo Durand, with guest speaker Natasha Noy from the schema.org initiative.

At the conclusion of the course, a hackathon was coordinated by Fenner and Durand, with Noy helping in schema.org metadata integrations. This hackathon was open to the course participants as well as other interested attendees at FSCI. Small teams that included staff from several data repositories were formed and each worked on implementing at least one of the ten guidelines for their respective data repositories. Overall, the hackathon focused on machine-readable metadata in landing pages, specifically in schema.org JSON-LD, and some repositories had implemented schema.org support by the end of the hackathon.

The course and hackathon provided valuable feedback regarding the guidelines; and served as both a propagation mechanism for the guidelines and a means of informal validation of current status with practitioners. Based on discussions at that time, with technologists from the sixteen repositories represented at our workshop, most of them had already implemented guidelines 1–6, and all had implemented guideline 1. Most had plans to implement all the guidelines, whether required, recommended, or optional. This led us to expect that many data repositories may already follow the *required* recommendations but need further work to implement the *recommended* or *optional* ones.

To follow up on the implementation of the guidelines, we looked at the adoption of guideline 8 six months after the above workshop and 12 months after the publication of the preprint. Guideline 8 recommends embedding machine-readable metadata in dataset landing pages, using the schema.org metadata standard. This particular guideline was clearly high on the priority list for implementation at the FSCI course, and its implementation was the main topic at the hackathon.

We reached out to the data sharing community using mailing lists, social media and personal communications starting in January 2018, and collected information about implementations using a CSV file hosted in a GitHub repository[30]. We found 32 data repositories embedding schema.org metadata as of May 2018, and information for 8 repositories was added by these repositories via GitHub pull request. We collected information about the inclusion of the metadata fields that were required or recommended in our repository recommendations, included

URLs for examples were available, and we checked whether all required metadata were included. These results are summarized in Fig. 3.

While the number of repositories in this sample is still small, we can see that a number of repositories not only are embedding schema.org metadata in their landing pages, but that half of them support all required metadata described in this document. The most frequently missing metadata elements are identifier and includedInDataCatalog/publisher and, surprisingly, publicationDate (which could also be the publication year). All these metadata elements can be easily added, but more work is probably needed to provide feedback to these early adopters. Two repositories implemented schema.org using RDFa. While this is an accepted serialization format for schema.org metadata, this document recommends standardization on JSON-LD to simplify tool development, e.g. reference manager support. We are also seeing a broad range of recommended metadata implemented, and that will help with data discovery, e.g. via Google Dataset Search. Recent software releases will also be helpful, including DataCite's new link checker[31]. We believe the development and release of such tools by major providers will further incentivize repositories to follow the guidelines in this article.

In addition to the implementations in repository landing pages noted earlier, we are also seeing implementations in supporting services for data repositories: the Dataverse repository platform added schema.org support in December 2017[32], and DataCite added support for direct DOI registration using schema.org metadata embedded in the dataset landing page in May 2018[33].

Data Availability

We compiled a dataset through community consultation which lists data repositories that embed schema.org metadata. The dataset is available as a CSV file within the Zenodo repository[30].

References

1. Data Citation Synthesis Group. Joint declaration of data citation principles. *FORCE11*, https://doi.org/10.25490/a97f-egyk (2014).
2. Altman, M., Borgman, C. & Crosas, M. An introduction to the joint principles for data citation. *Bull. Assoc. Info. Sci. Tech.* **41**, 43–45 (2015).
3. King, G. & Altman, M. A proposed standard for the scholarly citation of quantitative Data. *D-Lib Mag.* **13**, https://doi.org/10.1045/march2007-altman (2007).
4. Uhlir, P. F. (ed.) *For attribution: developing data attribution and citation practices and standards: summary of an international workshop*, https://www.nap.edu/read/13564/chapter/1 (National Academies, Washington DC, 2012).
5. CODATA-ICSTI Task Group on Data Citation Standards and Practice. Out of cite, out of mind: the current state of practice, policy, and technology for the citation of data. *Data Sci. J.* **12**, CIDCR1–CIDCR7, https://doi.org/10.2481/dsj.OSOM13-043 (2013).
6. Starr, J. *et al*. Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Comput. Sci.* **1**, e1, https://doi.org/10.7717/peerj-cs.1 (2015).
7. Altman, M. & Crosas, M. The evolution of data citation: from principles to implementation. *IASSIST Q.* **37**, 62–70 (2013).
8. Smith, A. M., Katz, D. S. & Niemeyer, K. E. Software citation principles. *PeerJ Comput. Sci.* **2**, e86, https://doi.org/10.7717/peerj-cs.86 (2016).
9. Stodden, V. *et al*. Enhancing reproducibility for computational methods. *Science* **354**, 1240–1241, https://doi.org/10.1126/science.aah6168 (2016).
10. Cousijn, H. *et al*. A data citation roadmap for scientific publishers. *Sci. Data* **5**, 180259, https://doi.org/10.1038/sdata.2018.259 (2018).
11. Wimalaratne, S. M. *et al*. Uniform resolution of compact identifiers for biomedical data. *Sci. Data* **5**, 180029, https://doi.org/10.1038/sdata.2018.29 (2018).
12. Chen, X. *et al*. DataMed – an open source discovery index for finding biomedical datasets. *J. Am. Med. Inform. Assoc.* **25**, 300–308, https://doi.org/10.1093/jamia/ocx121 (2018).
13. DataCite Metadata Working Group. DataCite metadata schema documentation for the publication and citation of research data, version 4.1. *Datacite e. V*, https://doi.org/10.5438/0014 (2017).
14. McMurry, J. A. *et al*. Identifiers for the 21st century: how to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLOS Biol.* **15**, e2001414, https://doi.org/10.1371/journal.pbio.2001414 (2017).
15. Fenner, M. *et al*. Thor: conceptual model of persistent identifier linking. *Zenodo*, https://doi.org/10.5281/zenodo.48705 (2016).
16. Berners-Lee, T., Fielding, R. & Masinter L. *Uniform Resource Identifier (URI): Generic Syntax, STD 66, RFC 3986*, https://doi.org/10.17487/RFC3986 (RFC Editor, 2005).
17. Treloar, A. Den Haag persistent object identifier – linked open data manifesto. *Zenodo*, https://doi.org/10.5281/zenodo.55666 (2011).
18. Honor, L. B., Haselgrove, C., Frazier, J. A. & Kennedy, D. N. Data citation in neuroimaging: proposed best practices for data identification and attribution. *Front. Neuroinformatics* **10**, 34, https://doi.org/10.3389/fninf.2016.00034 (2016).
19. McEntyre, J., Sarkans, U. & Brazma, A. The BioStudies database. *Mol. Syst. Biol.* **11**, 847, https://doi.org/10.15252/msb.20156658 (2015).
20. FORCE11. *Data citations: a primer*, https://force11.github.io/data-citation-primer/ (2016).
21. DataCite Metadata Working Group. DataCite metadata schema documentation for the publication and citation of research data, version 4.1. *DataCite e.V*, https://doi.org/10.5438/0014 (2017).
22. Gonzalez-Beltran, A. & Rocca-Serra, P. DataMed DATS specification v2.2 - NIH BD2K bioCADDIE. Zenodo, https://doi.org/10.5281/zenodo.438337 (2017).
23. Gonzalez-Beltran, A. N. *et al*. Data discovery with DATS, exemplar adoptions and lessons learned. *J. Am. Med. Inform. Assoc.* **25**, 13–16, https://doi.org/10.1093/jamia/ocx119 (2018).
24. Sansone, S.-A. *et al*. DATS, the data tag suite to enable discoverability of datasets. *Sci. Data* **4**, 170059, https://doi.org/10.1038/sdata.2017.59 (2017).
25. Hammond, T., Hannay, T. & Lund, B. *RDF site summary 1.0 modules: PRISM*, http://www.prismstandard.org/resources/mod_prism.html (2004).
26. Fielding, R. & Reschke, J. (eds) *Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content, RFC 7231*. 10.17487/RFC7231 (RFC Editor, 2014).
27. Van de Sompel, H. & Nelson, M. L. Reminiscing about 15 years of interoperability efforts. *D-Lib Mag.* **21**, https://doi.org/10.1045/november2015-vandesompel (2015).
28. NIH Common Fund. *NIH Data Commons Pilot Phase Consortium: awards made under Research Opportunity Announcement (ROA) RM-17-026*, https://commonfund.nih.gov/commons/awardees (2018).
29. Fenner, M. *et al*. A data citation roadmap for scholarly data repositories. Preprint at https://doi.org/10.1101/097196 (2017).

30. Fenner, M. *et al*. Listing of data repositories that embed schema.org metadata in dataset landing pages. *Zenodo*, https://doi.org/10.5281/zenodo.1263942 (2018).
31. Dasler, R. Link checker is here. *DataCite Blog*, https://doi.org/10.5438/vywf-6s91 (2018).
32. Dataverse Project. Dataverse 4.8.4 release adds support for schema.org. *Dataverse Project Blog*, https://dataverse.org/blog/latest-dataverse-update-adds-support-schemaorg (2017).
33. Dasler, R. DOI Fabrica 1.0 is here! *DataCite Blog*, https://doi.org/10.5438/0yk5-b755 (2018).

Acknowledgements

The Roadmap in this document resulted from meetings and discussions of the DCIP Expert Group, with input from data repositories, publishers, persistent identifier providers, reference manager specialists, and other experts on data citation. Implementation of the data citation principles involves many stakeholder groups, and the DCIP project has worked closely with them via several Expert Groups, and a coordinating steering group. The authors gratefully acknowledge the following members of the Data Citation Repositories Expert Group, who participated actively with the authors in workshops and/or telecons to develop this Roadmap: Cecilia Arighi (Protein Information Resource, University of Delaware); Ian Fore (National Cancer Institute, National Institutes of Health, Bethesda MD); Christian Haselgrove (University of Massachusetts Medical School); John Kunze (California Digital Library); Neil McKenna (Baylor College of Medicine); Pete Meyer, Harvard Medical School; Raman Prasad (IQSS, Harvard University); Peter Rose (University of California San Diego); Simone Sacchi (Trust-IT Services); Ryan Scherle (Dryad Digital Repository); Curtis Smith (EndNote, Thomson Reuters); Cathy Wu (Protein Information Resource, University of Delaware). Research reported in this publication was supported in part by the National Institutes of Health under award number U24HL126127 for the BioCADDIE project; and by the European Molecular Biology Laboratory (EMBL). Work on this project was coordinated by FORCE11 (https://force11.org), a not-for-profit community organization seeking to improve scholarly communication through digital technology. Finally, the authors wish to thank Stephanie Hagstrom of the University of California Library for her extremely helpful administrative work supporting the Data Citation Implementation Pilot, in organizing workshop and conference calls, and in coordinating website administration as well as logistics for workshop attendees.

Author Contributions

Martin Fenner and *Mercè Crosas* co-chaired the DCIP Publishers Expert Group which produced this article. They led regular telecons and organized the work of the group; including face-to-face meetings of participants (see Acknowledgements) in February (Boston) and June (San Diego) 2016. *Martin Fenner* and *Gustavo Durand* organized an additional workshop and tutorial where feedback was gathered from repository experts, at the FSCI 2017 meeting in San Diego. The listed authors (*Fenner, Crosas, Grethe, Kennedy, Hermjakob, Rocca-Serra, Durand, Berjon, Karcher, Martone and Clark*) developed and refined the Roadmap presented here, via the series of workshops and telecons described in the article; and reviewed and edited this text. The main text of this article was written by *Martin Fenner, Mercè Crosas*, and *Tim Clark*. *Tim Clark* coordinated the work of the Repositories Expert Group with the other DCIP participants (Publishers, Identifiers, JATS, and Primer/FAQ), and edited this article for publication. *Tim Clark* and *Maryann Martone* co-chaired the overall Data Citation Implementation Pilot, and supervised development of the Roadmaps.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.