A Data Citation Roadmap for Scholarly Data Repositories

Martin Fenner[a]*, Merce Crosas[b]*, Jeffrey S. Grethe[c], David Kennedy[d], Henning Hermjakob[e], Phillippe Rocca-Serra[f], Robin Berjon[g], Sebastian Karcher[h], Maryann Martone[i] and Tim Clark[j,k]

[a] DataCite, Hannover, Germany
[b] Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA
[c] Center for Research in Biological Systems (CRBS), University of California, San Diego, CA, USA
[d] Department of Psychiatry, University of Massachusetts Medical School, Worcester, MA, USA
[e] European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[f] Oxford e-Research Centre, University of Oxford, Oxford, UK
[g] Standard Analytics, New York, NY, USA
[h] Qualitative Data Repository, Syracuse University, Syracuse, NY, USA
[i] Department of Neurosciences, University of California, San Diego, CA, USA
[j] Massachusetts General Hospital, Boston, MA, USA
[k] Harvard Medical School, Boston, MA, USA

* These authors contributed equally to this work.

Published on behalf of the DCIP Repositories Early Adopters Expert Group. A full list of members appears in Appendix A.

Abstract

This article presents a practical roadmap for scholarly data repositories to implement data citation in accordance with the Joint Declaration of Data Citation Principles (Data Citation Synthesis Group, 2014), a synopsis and harmonization of the recommendations of major science policy bodies. The roadmap was developed by the Repositories Early Adopters Expert Group, part of the Data Citation Implementation Pilot (DCIP) project (FORCE11, 2015), an initiative of FORCE11.org and the NIH BioCADDIE (2016) program. 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.

Introduction

The Joint Declaration of Data Citation Principles (JDDCP) published in 2014 (Data Citation Synthesis Group, 2014) 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, starting with stressing that data should be considered legitimate, citable products of research (M Altman, Borgman, & Crosas, 2015). The JDDCP condense the results of substantial prior studies on science policy and practice (King, Gary & Altman, Micah, 2007; Uhlir, 2012; CODATA-ICSTI Task Group on Data Citation Standards and Practice, 2013).

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. (2015) and Altman and Crosas (2013), and are consistent with recent recommendations regarding data, code and workflows (Smith, Katz, & Niemeyer, 2016; Stodden et al., 2016).

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, and of course researchers. Data citation practices and technologies supported by repositories will substantially assist development of new data discovery indexes such as BioCADDIE.

This roadmap was developed based on numerous discussions of the DCIP Repositories Early Adopters Expert Group, including two in-person workshops in February (Boston) and June (San Diego) 2016, and in close coordination with the other DCIP expert groups.

The recommendations are grouped into three phases: required, recommended and optional. Implementing these recommendations takes time and resources, it is therefore not only critical to provide specific recommendations, 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). While a formal analysis has not yet been done, we expect that at this point in time many data repositories already follow all required recommendations, but that most of them need to do more work for the recommendations that are recommended or optional.

Recommendations

Required

1. All datasets intended for citation *must* have a globally unique persistent identifier that can be expressed as unambiguous URL.
2. Persistent identifiers for datasets *must* support multiple levels of granularity, where appropriate.
3. This persistent identifier expressed as URL *must* resolve to a landing page specific for that 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 helping with 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.

Optional

9. 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.
10. HTTP link headers *may* be supported to advertise content negotiation options
11. Metadata *may* be made available for download in Bibtex or other standard bibliographic format.

1. 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). 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 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 the RFC 3986 ("RFC 3986 - Uniform Resource Identifier (URI): Generic Syntax," 2005), including support for content negotiation (Treloar, 2011).
- **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.
  For data repositories that are not using globally unique identifiers, the DCIP EG2 Identifiers Expert Group is working on a bridging solution using common prefixes ("Prefix Commons," 2016) and resolver services (identifiers.org and n2t.net).
- **Widely used by a community.** Accession numbers, in combination with the database name for global uniqueness, are the most widely used identifiers in the life sciences.

2. 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.

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 (Honor, Haselgrove, Frazier, & Kennedy, 2016). Another example is the BioStudies database (McEntyre, Sarkans, & Brazma, 2015), which can provide storage for all the underlying data links and files for a publication.

3. 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 (Starr et al., 2015), or to other representations of the metadata, unless special protocols such as content negotiation are used (see recommendation #7 below).

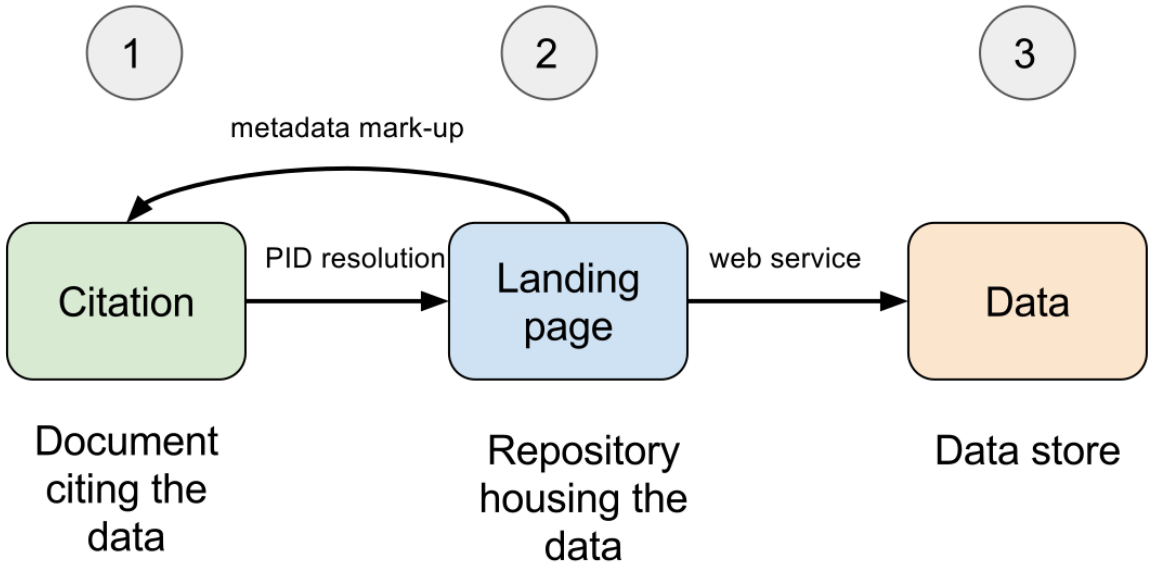

Fig. 1. Generic data citation - relationships of the citation reference, repository landing page and underlying data, taken with permission from ("Data Citations: A Primer," 2016).

Landing pages provide definitive information (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 (Starr et al., 2015).

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 (Starr et al., 2015).

Example

Cite this Dataset
Bilokapic, S; Schwartz, TU. 2015. "X-Ray Diffraction data for: Nup37-Nup120 full-length complex from Schizosaccharomyces pombe. PDB Code 4FHN", SBGrid Data Bank, V1, http://dx.doi.org/10.15785/SBGRID/179.
Download Citation

Fig. 2. Providing information about how a dataset should be cited, with Bibtex download link.

4. 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 basic data citation by reference managers, and enables 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">
```

5. 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. The DCIP FAQ Expert Group provides example documentation for data repositories.

6. Metadata on landing pages

Landing pages *should* provide metadata required for data citation in both human- and machine-readable format. The latter includes that access to the metadata should not require javascript, cookies or login. 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 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:

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

Tab. 1: Citation metadata for Data Repositories

[a] Dublin Core Metadata Element Set ("Dublin Core Metadata Element Set, Version 1.1," 2012)
[b] Dataset - Schema.org ("Dataset - schema.org," n.d.)
[c] DataCite Metadata Working Group (2016)
[d] Gonzalez-Beltran & Rocca-Serra (2016)

* 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.

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 ("BioSchemas," 2016), as well as the DataCite and DATS metadata schemata.

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, in particular:

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

Tab. 2: Important discovery metadata for Data Repositories

* 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 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 DataMed DATS specification.

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.

7. 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 (Gonzalez-Beltran & Rocca-Serra, 2016). 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 Epidemiology 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
    objectives: 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 geographic 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 (“DataCite Search,” 2016), which has embedded schema.org/JSON-LD metadata on every search result page for close to three million datasets.

8. Metadata via HTML Meta Tags

Data repositories *should* offer machine-readable metadata on landing pages using Highwire, PRISM (Hammond, Hannay, & Lund, 2004), 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
gammarus) 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" />
```

9. Content negotiation for machine-readable metadata

Persistent identifiers expressed as HTTP URI *must* by default resolve to the landing page for that dataset (see recommendation #3). Data repositories and identifier service providers such as identifiers.org or DataCite in addition *may* implement content negotiation 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 XML, RDF, Bibtex 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.

10. Support HTTP link headers

The persistent identifier (see recommendation #2) and available content negotiation options (see recommendation #9) *may* be provided in a HTTP link header (Van de Sompel & Nelson, 2015). 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"
```

11. Metadata via downloadable file in standard bibliographic format

Repositories *may* 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  -
```

Conclusions

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, and in particular support for schema.org/JSON-LD markup for metadata is still at an early stage. Data repositories that have implemented the required and recommended steps might be interested to look into the optional steps for extra data citation support.

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, is working on standard metadata for biomedical data with DATS, and on a central index to search a large number of biomedical datasets with DataMed (“DataMed | bioCADDIE DDI,” 2016). The European ELIXIR (“ELIXIR Data for life,” 2016) project (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 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 needed to implement these recommendations, and ongoing coordination amongst data repositories, and with publishers and other important stakeholders will be essential in this activity.

Acknowledgements

This document was generated by the DCIP Repositories Early Adopters 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 is working closely with them via a number of expert groups (FAQ, Identifiers, Publisher Early Adopters, Repository Early Adopters, and JATS, see Appendix C), and a coordinating steering group. The DCIP project was funded by NIH BD2K bioCADDIE (2016), which is developing a data discovery index prototype for the biomedical sciences.

References

Altman, M., Borgman, C., & Crosas, M. (2015). An introduction to the joint principles for data citation. *Bulletin of the American \ldots*, *41*(3), 43–45. https://doi.org/10.1002/bult.2015.1720410313

Altman, M., & Crosas, M. (2013). The Evolution of Data Citation: From Principles to Implementation. *IASSIST Quarterly*. Retrieved from http://scholar.harvard.edu/mercecrosas/publications/evolution-data-citation-principles-implementation

bioCADDIE. (2016). bioCADDIE | biomedical and healthCAre Data Discovery and Indexing Ecosystem. Retrieved December 22, 2016, from https://biocaddie.org/

BioSchemas. (2016). Retrieved December 22, 2016, from http://bioschemas.org/

CODATA-ICSTI Task Group on Data Citation Standards and Practice. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*, *12*(0). https://doi.org/10.2481/dsj.OSOM13-043

Data Citation Synthesis Group. (2014). Joint Declaration of Data Citation Principles. Retrieved December 22, 2016, from http://www.force11.org/datacitation

Data Citations: A Primer. (2016). Retrieved December 22, 2016, from http://force11.github.io/data-citation-primer/

DataCite Metadata Working. (2016). DataCite Metadata Schema for the Publication and Citation of Research Data v4.0. DataCite. Retrieved from http://search.datacite.org/works/10.5438/0013

DataCite Search. (2016). Retrieved December 22, 2016, from https://search.datacite.org/

DataMed | bioCADDIE DDI. (2016). Retrieved December 22, 2016, from https://datamed.org/

Dataset - schema.org. (2016). Retrieved December 12, 2016, from https://schema.org/Dataset

Dublin Core Metadata Element Set, Version 1.1. (2012). Retrieved December 22, 2016, from

http://dublincore.org/documents/2012/06/14/dces/

ELIXIR Data for life. (2016). Retrieved December 22, 2016, from https://www.elixir-europe.org/

FORCE11. (2015). Data Citation Implementation Pilot (DCIP). Retrieved December 22, 2016, from https://www.force11.org/group/dcip

Gonzalez-Beltran, A., & Rocca-Serra, P. (2016). DataMed DATS specification v2.1 - NIH BD2K bioCADDIE. *Zenodo*. https://doi.org/10.5281/zenodo.62024

Hammond, T., Hannay, T., & Lund, B. (2004). *RDF Site Summary 1.0 Modules: PRISM*. Publishing Requirements for Industry Standard Metadata (PRISM). Retrieved from http://purl.org/rss/1.0/modules/prism/

Honor, L. B., Haselgrove, C., Frazier, J. A., & Kennedy, D. N. (2016). Data Citation in Neuroimaging: Proposed Best Practices for Data Identification and Attribution. *Frontiers in Neuroinformatics*, *10*. https://doi.org/10.3389/fninf.2016.00034

King, Gary, & Altman, Micah. (2007). A Proposed Standard for the Scholarly Citation of Quantitative Data. https://doi.org/10.1045/march2007-altman

McEntyre, J., Sarkans, U., & Brazma, A. (2015). The BioStudies database. *Molecular Systems Biology*, *11*(12), 847. https://doi.org/10.15252/msb.20156658

Prefix Commons. (2016). Retrieved December 22, 2016, from https://github.com/prefixcommons/prefixes

RFC 3986 - Uniform Resource Identifier (URI): Generic Syntax. (2005). Retrieved December 22, 2016, from http://www.rfc-base.org/rfc-3986.html

Smith, A. M., Katz, D. S., & Niemeyer, K. E. (2016). Software citation principles. *PeerJ Computer Science*, *2*, e86. https://doi.org/10.7717/peerj-cs.86

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R., … 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

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., … Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, *354*(6317), 1240–1241. https://doi.org/10.1126/science.aah6168

Treloar, A. (2011). Den Haag Persistent Object Identifier – Linked Open Data Manifesto. *Zenodo*. https://doi.org/10.5281/zenodo.55666

Uhlir, P. E. (Ed.). (2012). *For Attribution: Developing Data Attribution and Citation Practices*

and Standards: Summary of an International Workshop. Retrieved from https://www.nap.edu/catalog/13564/for-attribution-developing-data-attribution-and-citation-practices-and-standards

Van de Sompel, H., & Nelson, M. L. (2015). Reminiscing About 15 Years of Interoperability Efforts. *D-Lib Magazine*, *21*(11/12). https://doi.org/10.1045/november2015-vandesompel

Appendix A:

DCIP Repositories Early Adopters Expert Group Membership

- Cecilia Arighi, Protein Information Resource, University of Delaware
- Robin Berjon, Standard Analytics
- Tim Clark, Massachusetts General Hospital & Harvard Medical School
- Mercé Crosas, IQSS, Harvard University (Co-Chair)
- Gustavo Durand, IQSS, Harvard University
- Martin Fenner, DataCite (Co-Chair)
- Ian Fore, NIH
- Jeffrey Grethe, University of California, San Diego
- Stephanie Hagstrom, University of California, San Diego
- Christian Haselgrove, University of Massachusetts Medical School
- Henning Hermjakob, European Bioinformatics Institute
- Sebastian Karcher, Qualitative Data Repository
- David Kennedy, 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
- Philippe Rocca-Serra, University of Oxford
- Peter Rose, University of California, San Diego
- Simone Sacchi, Columbia University
- Ryan Scherle, Dryad Digital Repository
- Curtis Smith, EndNote, Thomson Reuters
- Cathy Wu, Protein Information Resource, University of Delaware

Appendix B: Glossary

Dataset identifier

The Identifier is a unique string that identifies a resource.

Title

A name or title by which a resource is known.

Creator

The main researchers involved in producing the data.

Data repository/Publisher

The publisher is usually the data repository which hosts and guarantees the persistence of the landing page.

Publication date

The year when the data was or will be made publicly available.

Version

The version number of the resource.

Type

The general type of a resource.

Appendix C: Other DCIP Expert Groups

EG 1 FAQ

https://www.force11.org/group/data-citation-implementation-pilot-dcip/data-citation-faqs-repositories

EG 2 Identifiers

https://www.force11.org/group/dcip/eg2identifiers

EG 3 Publishers Early Adopters

https://www.force11.org/group/dcip/eg3publisherearlyadopters

EG 5 JATS

https://www.force11.org/group/dcip/eg5-jats