Simple integration of data citation into research practice

Peter L. Pulsifer
peter.pulsifer@colorado.edu
Presentation Overview

- Simple Integration of data citation into research practice
- Challenges and considerations
- The polar community
- Education and training
- Promoting a culture of publication
- Integrating into systems:
  - Technology – data casting; data advertising
Simple Process

- Deposit/publish data
- Establish persistent identifier
- Cite using appropriate style
- Track statistics
- ... repeat...
- Get promoted!

(Simple!)

http://www.dcc.ac.uk/resources/how-guides/cite-datasets
Challenges and Considerations
Responsibilities of Data Creators

• Prepare and document data for public distribution
  • Implies knowing standards at some level of detail
• Submit data to repository or otherwise publish
• Cite prior or foundational data used in research
• Link to publication(s) (manual process in some cases)
• Consider and respond to feedback
Considerations

- **Granularity** – a single data set or multiple related data granules? Sub-setting?
- Change management and versioning
- Citation of physical objects
- Linking to creators - obtain unique creator identifier (e.g. ORCID)
The Polar Community
Complex, Multidisciplinary

- Community is based on geography rather than theme or discipline
- Examining phenomena ranging from molecular to cosmic scale; physical, life, social sciences etc.
- Many different kinds of actors (researchers, residents, policy makers etc.)
- Many disciplines and related journals and publication platforms
- A community of communities... intersecting with other communities
A closely related issue is that the intellectual work going into creating a good data set should be recognized and ideally cited in formal publications. Citation also aids reproducibility, which means that the citation must have a persistent locator or identifier to the data such as a digital object identifier. This in turn means the data need to be professionally stewarded over time.

Fair attribution is also fundamental to an ethically open data environment. Data creators and associated institutions deserve formally recognizable credit for their scientific contributions. IASC encourages this recognition through citation, appropriate co-authorship, and other means, and encourages IASC members to require appropriate credit and reference when serving as editors and reviewers.
Many Actors Working Together

Developing Standards for Data Citation & Attribution in Linguistics
Workshop 1 | Boulder, CO | September 18-20 2015
Educating The Polar Community
Broader Efforts

- **ARCTIC DATA COMMITTEE**

- Collection of national representatives and champions with an interest in integration and education
Confirmation of Key Priorities (WPs)

1. Documenting and understanding the Arctic data management ecosystem

2. Identifying and promoting common metadata elements

3. Engaging in data citation and publication movement

4. Promoting interoperability through action – interoperability experiment
Reference Existing Materials

Providing a citation for scientific data is a great way to...

The following web links provide much more information about data citation and have been picked to suit different levels of expertise:

1. A briefing paper created by the UK Digital Curation Centre and suitable for anyone wanting a general overview of data citation:
   https://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/data-citation-and-linking

2. A working guide on how to cite datasets created by the UK Digital Curation Centre and suitable for both researchers and data stewards.

Developing Standards for Data Citation & Attribution in Linguistics
Workshop 1 | Boulder, CO | September 18-20 2015
Data Centers
Promoting a Culture of Publication
Welcome to Polar Commons

Welcome to Polar Commons, an international non-profit organisation which is committed to providing information about a wide range of issues concerning the polar regions. We seek to raise public awareness about the threat of global climate change to the Earth’s high latitudes, promote scientific debate and encourage cooperation in the use and management of natural resources and at the same time, help preserve the environment and wildlife habitats.

Our mission is to inform and educate the public by providing access to facts, figures and data with a focus on the impact of human activities on the polar regions as well as offering access to the latest scientific discoveries, findings and insights into the changes in the Arctic and Antarctica. By pointing out the consequences of these changes not only for the environment and wildlife in the polar regions but for the entire world, we hope to attract the attention of the wider public and to prompt action. The end goal is to form an international movement that will create awareness as governments, corporations and other stakeholders with the power and influence of the oil, gas and mining industries, to act.
Ethics and Norms to Follow When Using PIC

PIC was developed with the goal to make information and data sharing as easy, effortless and time saving as possible, and encourage scientists, researchers, scholars, authors and other individuals and organisations to share their knowledge with both their colleagues and public. Each time new information is added to the already existing database, PIC users get a deeper insight and understanding of a wide range of polar-related subjects. At the same time, the format of our data sharing programme encourages interaction, collaboration and cooperation between the members of the scientific community as well as between scientists and the public on a wealth of issues concerning the polar regions.

In order for PIC to remain a valuable resource of polar data and information, continue to bring scientists together and promote public interest and understanding of the challenges the polar regions are facing, we have developed a set of ethics and norms for PIC users to follow when both contributing and retrieving data. The goal is to maintain the highest level of credibility, reliability and quality, and at the same time, provide PIC users with quality and up-to-date information. When signing up to PIC, users agree that they will:
Integrating into Systems
Data Centers

Developing Standards for Data Citation & Attribution in Linguistics
Workshop 1 | Boulder, CO | September 18-20 2015
Cite your data

Why is it so important to cite data? Books and journal articles have long benefited from an infrastructure that makes them easy to cite, a key element in the process of research and academic discourse. We believe that you should cite data in just the same way that you can cite other sources of information, such as articles and books. Data citation can help by:

- enabling easy reuse and verification of data
- allowing the impact of data to be tracked
- creating a scholarly structure that recognises and rewards data producers

Examples of data citation

We recognise that the challenges associated with data publication vary across disciplines, and we encourage research communities to develop citation systems that work well for them. Our recommended format for data citation is as follows:

- Creator (PublicationYear): Title. Publisher. Identifier
## Developing Standards for Data Citation & Attribution in Linguistics
### Workshop 1 | Boulder, CO | September 18-20 2015

<table>
<thead>
<tr>
<th>Allocator</th>
<th>DOI Registrations</th>
<th>Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>This Year</td>
</tr>
<tr>
<td>ANeS - Australian National Data Service</td>
<td>5,280</td>
<td>1,408</td>
</tr>
<tr>
<td>BL - The British Library</td>
<td>830</td>
<td>209</td>
</tr>
<tr>
<td>CDL - California Digital Library</td>
<td>2,223</td>
<td>833</td>
</tr>
<tr>
<td>CERN - CERN - European Organization for Nuclear Research</td>
<td>30,179</td>
<td>18,848</td>
</tr>
<tr>
<td>CICTI - National Research Council Canada</td>
<td>34,799</td>
<td>1,597</td>
</tr>
<tr>
<td>CRUI - CRUI2011</td>
<td>37,691</td>
<td>8,330</td>
</tr>
<tr>
<td>DATACITE - DataCite</td>
<td>25</td>
<td>12</td>
</tr>
<tr>
<td>DELFT - TU Delft Library</td>
<td>47,661</td>
<td>35,307</td>
</tr>
<tr>
<td>DK - Technical Information Center of Denmark</td>
<td>76,817</td>
<td>73,340</td>
</tr>
<tr>
<td>ESITDOH - TARTU University</td>
<td>487,676</td>
<td>497,549</td>
</tr>
<tr>
<td>ETHZ - ETH Zurich</td>
<td>918,396</td>
<td>85,217</td>
</tr>
<tr>
<td>GESIS - GESIS - Leibniz Institute for the Social Sciences</td>
<td>423,758</td>
<td>9,645</td>
</tr>
<tr>
<td>INIST - Institute for Scientific and Technical Information</td>
<td>5,925</td>
<td>3,173</td>
</tr>
<tr>
<td>JALC - Japan Link Center</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MTAKT - MTA Kvalitási</td>
<td>1,104</td>
<td>846</td>
</tr>
<tr>
<td>NECT - National Research Council of Thailand</td>
<td>34,928</td>
<td>18,690</td>
</tr>
<tr>
<td>OISTI - Office of Scientific and Technical Information (OISTI), US Department of Energy</td>
<td>19,042</td>
<td>16,785</td>
</tr>
<tr>
<td>PURDUE - Purdue University Library</td>
<td>7,333</td>
<td>1,850</td>
</tr>
<tr>
<td>ETHN - Swiss National Data Service</td>
<td>3,377</td>
<td>261</td>
</tr>
</tbody>
</table>
Developing Standards for Data Citation & Attribution in Linguistics
Workshop 1 | Boulder, CO | September 18-20 2015
Thomas Reuters DCI

START WHERE DISCOVERY BEGINS.

The Data Citation Index supports data discovery, reuse and interpretation that benefits everyone involved in the research lifecycle.

http://wokinfo.com/products_tools/multidisciplinary/dci/about/
The DLI (Data Literature Interlinking) Service

Introduction and vision
Challenges to realize the full potential of research data exist at different levels—from cultural aspects, such as proper rewards and incentives, to policy and funding, and to technology. The challenges are interconnected and impact a diversity of stakeholders in the research data landscape—including researchers, research organizations, funding bodies, data centres, and publishers. To make progress in overcoming barriers and building a stronger research data infrastructure, it is essential that the different stakeholders work together to address common issues and move forward on a common path. Alongside other organizations, the ICSU World Data System (ICSU-WDS), the Research Data Alliance (RDA), and OpenAIRE provide useful forums for such collaborations. In particular, they are today working in synergy on an initiative that brings together different parties in the research data landscape with the objective of creating the Data Literature Interlinking Service (DLI Service), namely, 'an open, freely accessible, web-based service that enables its users to identify datasets that are associated with a given article, and vice versa'. At the moment of writing, members of the initiative include: the ICSU-WDS–RDA DSP-WG, OpenAIRE, RDA, ICSU-WDS, STM, CrossRef, DataCite, ORCID, the Australian National Data Service, and the RMap project. The vision is that of moving away from several bilateral arrangements that characterizes the research ecosystem today, towards establishing common standards and tools that sit in the middle and interact with all parties (see Figure). Such a transition would facilitate interoperability between platforms and systems operated by the different parties, reduce systemic inefficiencies in the ecosystem, and ultimately enable new tools and functionalities to the benefit of researchers.
De-Centralized Indexing, Discovery & Linking

- What if data are not published in a major, formal repository?
- How do I increase the visibility of my data?
- Web-based documents (e.g. Web pages) can be “marked up” with embedded tags
- Tags are recognized and indexed by “crawlers”
- Generally related to Altmetrics concept and systems
Web Page Markup

```html
<body vocab="http://purl.org/linguistics/vocab/"> ...

<p resource="http://dx.doi.org/10.1234/data123">

<a href="http://dx.doi.org/10.123/goodpaper212"
property="providesDataFor">doi:10.123/goodpaper212</a>

</p> ...

</body>
```
Mass-market Search Engine

Domain-Specific Search Engine

Search Engine Crawler

Custom Application Crawler

finds data using

Data Creators / Users

creates

Data Set

Web Page Title

Web Page

includes

Marked-up HTML

Data Set Description

Web Page

describes

links to

Mass-market Search Engine

Data Creators / Users

creates

Data Set

Web Page Title

Mass-market Search Engine

Domain-Specific Search Engine

finds data using

Data Creators / Users

creates

Data Set

Web Page Title

Mass-market Search Engine

Domain-Specific Search Engine

finds data using

Data Creators / Users

creates

Data Set

Web Page Title
Variations

Data Ad

What is Data Ad?
Data Ad is a process that helps you advertise your data sets or collections. With Data Ad, data providers can help users and computers on the internet discover your data quickly and more efficiently. Data Ad uses Atom technology. This means that your Data Ad or feed can be viewed with any desktop or browser-based feed reader.

About the Data Ad Tool
You can use ATOM feeds so people and computer systems can find your data quickly and easily. For that purpose, NSIDC and the Libre project have created the Data Ad, a Web-based tool that helps you create ATOM feeds.

How to Create a Data Ad
Welcome to Schema.org

Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.

Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata and JSON-LD. These vocabularies cover entities, relationships between entities and actions, and can easily be extended through a well-documented extension model. Over 10 million sites use Schema.org to markup their web pages and email messages. Many applications from Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences.

Schema.org is sponsored by Google, Microsoft, Yahoo and Yandex. The vocabularies are developed by an open community process, using the public-schemaorg@w3.org mailing list and through GitHub.

A shared vocabulary makes it easier for webmasters and developers to decide on a schema and get the maximum benefit for their efforts. It is in this spirit that the sponsors, together with the larger community have come together, to provide a shared collection of schemas.

We invite you to get started!

View our blog at blog.schema.org or see release history.

Terms and conditions
System on Systems

- The de-Centralized approach has risks
- Persistence and long-term preservation are not guaranteed
- Model can be used as a step towards a full stewardship and curation solution
- Altmetrics can assist in evaluating data to justify more formal stewardship
Summary

- Data publication and citation is relatively easy to integrate into practice
- Some new considerations for researchers
- Education is key in diverse and complex polar community
- Many general tools available
- Decentralized approach provide foundation for integrating into applications to meet specific requirements (e.g. dealing with ‘long-tail’ data)
Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant Numbers:

- ARC 0856634 and ARC 1231638

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.