Data Citation Implementation

@ Dataverse

Mercè Crosas
Chief Data Science and Technology Officer,
IQSS, Harvard University
@mercecrosas

Workshop: Data Citation Pilot Project Kick-off
bioCADDIE supplemental project, NIH Big Data to Knowledge
Feb 3, 2016
Data Citation in Dataverse complies with the **Data Citation Principles**

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014

Altman, Crosas, The Evolution of Data Citation: From Principles to Implementation, IASSIST Quarterly; 2013
Increasing Food Security and Farming System Resilience in East Africa through Wide-Scale Adoption of Climate-Smart Agricultural Practices

Winowiecki, Leigh; Laderach, Peter; Mwongera, Caroline; Twyman, Jennifer; Mashisia, Kelvin; Okolo, Wendy; Eitzinger, Anton; Rodriguez, Beatriz, 2015, “Increasing Food Security and Farming System Resilience in East Africa through Wide-Scale Adoption of Climate-Smart Agricultural Practices”, http://dx.doi.org/10.7910/DVN/28703, Harvard Dataverse, V7

If you use these data, please add this citation to your scholarly resources. Learn about Data Citation Standards.

Description
The overall project goal is to improve food security and farming system resilience of smallholder mixed crop-livestock farmers in East Africa while mitigating climate change through wide-scale adoption of climate-smart agriculture (CSA). The project integrates interdisciplinary approaches, including participatory research, integrating a meta-analysis of CSA practices, real-time land and soil health assessments, crop suitability modelling, socio-economic appraisals and multi-dimensional trade-off analyses, as well as on-farm participatory evaluations of CSA to identify, test, implement, and upscale locally appropriate CSA practices.

Subject
Earth and Environmental Sciences; Medicine, Health and Life Sciences; Social Sciences; Other

Keyword
Climate Smart Agriculture, Food Security, Land Health, Soil, Socio-Economic, Adaptation, Mitigation
The Persistent Identifier applies to the entire Dataset, not to individual Files.
The same Persistent Identifier applies to All Versions of the Dataset

<table>
<thead>
<tr>
<th>Versions</th>
<th>Files</th>
<th>Citation Metadata</th>
<th>Author(s)</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.0</td>
<td>Added: 2; Removed: 1; Changed File Metadata: 5</td>
<td>Contact (1 Changed); Additional Citation Metadata: ()</td>
<td>Paola Camargo</td>
<td>February 2, 2016</td>
</tr>
<tr>
<td>6.1</td>
<td>Added: 2; Removed: 1</td>
<td>Terms of Use/Access Changed</td>
<td>Cathy Garlick</td>
<td>April 23, 2015</td>
</tr>
<tr>
<td>6.0</td>
<td>Added: 2; Removed: 1</td>
<td></td>
<td>Juliana Muriel Osorio</td>
<td>March 3, 2015</td>
</tr>
<tr>
<td>5.0</td>
<td>Changed File Metadata: 3</td>
<td></td>
<td>Juliana Muriel Osorio</td>
<td>February 18, 2015</td>
</tr>
<tr>
<td>4.0</td>
<td>Added: 1</td>
<td></td>
<td>Juliana Muriel Osorio</td>
<td>January 30, 2015</td>
</tr>
<tr>
<td>3.0</td>
<td>Citation Metadata: Author (2 Added)</td>
<td></td>
<td>Juliana Muriel Osorio</td>
<td>January 23, 2015</td>
</tr>
<tr>
<td>2.0</td>
<td>Added: 1</td>
<td></td>
<td>Juliana Muriel Osorio</td>
<td>January 22, 2015</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>This is the first published version.</td>
<td>Hector F. Tobon R.</td>
<td>January 16, 2015</td>
</tr>
</tbody>
</table>

Only major versions (not minor) appear in the generated data citation.
Citation for Quantitative (tabular) Data

Authors, Published Year, Dataset Title, **Persistent Identifier**, Repository Name, Version, **Universal Numerical Fingerprint (UNF)**, [File name], [var 1], [var 2], [var...]

Checksum independent of file format

Specify File in Dataset

Specify a subset of variables in Tabular Data File

Following: Altman, King, A Proposed Standard for the Scholarly Citation of Quantitative Data, D-Lib, 2007
Dataverse – DataCite Workflow

**EZID API**
1. Dataset Created in Dataverse
2. Mint DOI with status “reserved” in EZID, send citation metadata
3. Dataset published in Dataverse
4. Change status to “public” in EZID
5. New version of Dataset
6. Send updated citation metadata

**DataCite API**
1. Dataset Created in Dataverse
2. Reserve local DOI in Dataverse
3. Dataset published in Dataverse
4. Mint DOI in DataCite, send citation metadata
5. New version of Dataset
6. Send updated citation metadata
Additional Metadata in Dataverse

**Citation Metadata**
- Authors
- Title
- Description
- Dates
- Contact
- Subject
- ...

**Domain Metadata**
- Life Sciences: based on ISA-Tab (and OBI and NCBI taxonomy)
- Other domains (social science, astronomy)

**File Metadata**
- File header metadata
- File description, type
- Variable metadata
What’s Coming Next

SBGrid Data Repository, Biomedical Dataverse (Sliz HMS, Crosas IQSS)

Social Science Big Data (King, Crosas at IQSS)

Data Provenance (Seltzer SEAS, Crosas, King IQSS)

Privacy Tools to share sensitive data (SEAS, Berkman, Privacy Lab, IQSS, MIT)
Future Data Citation Extensions

• Provenance Metadata to be used in citation services
• Extended Domain Metadata (e.g., Life Sciences) to be used in citation services
• Support for Privacy, Sensitive Datasets:
  – A DataTag (blue, green, yellow, orange, red, crimson) assigned to each dataset that defines its sensitive level, with security and access requirements
• Support for Large (Streaming) Datasets:
  – Many files per Dataset. E.g., Primary Structure Dataset with thousands of images
  – Large Streaming Dataset. E.g., Geospatial Tweets

Sweeney, Crosas, Bar-Sinai Sharing Sensitive Data with Confidence: The DataTags System, JOTS, 2015
Citation for Big Data: Large, Streaming, or Sensitive Datasets

Authors, Published Year, Title, **Persistent Identifier**, Repository Name, Version, [Subset: Query or Variable], [DataTag]

- Be able to cite entire Big Data dataset (with one Persistent Identifier), as well as specify granularity when needed
- Should the query be a RESTful url?
- Should the subset be defined by variable/attributes metadata?
- Should the DataTag be part of the citation for sensitive data?