Services to Make Sense of Data

Patricia Cruse, Executive Director, DataCite

Council of Science Editors
San Diego – May 2017
JoRD (Journal Research Data Policy Bank) sheds light the policies devised by academic publishers to promote linkage between journal articles and underlying research… (Jisc funded, ended in 2014)

How many journals make data sharing a requirement of publication?

<table>
<thead>
<tr>
<th>Results of Journal Survey</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total no. of Journals surveyed</td>
<td>371</td>
</tr>
<tr>
<td>Total no. of Journals with data sharing policies</td>
<td>162</td>
</tr>
<tr>
<td>Total no. of Journals that make sharing a requirement of publication</td>
<td>31</td>
</tr>
<tr>
<td>Total no. of Journals that enforce the policies</td>
<td>27</td>
</tr>
<tr>
<td>Total no. of Journals that state consequences for non compliance</td>
<td>7</td>
</tr>
</tbody>
</table>

https://jordproject.wordpress.com/2013/07/05-going-back-to-basics-reusing-data/
Back in the day…

- Figure 1. Two pages (scan) from Galilei's Sidereus Nuncius (“The Starry Messenger” or “The Herald of the Stars”), Venice, 1610.

**data** = drawings of Jupiter and its moons

**metadata** = timing of each observation

**text** = descriptions of methods, analysis, conclusions


Summing up the problem

We need reliable and unambiguous access to data!

- attribution
- collaboration and reuse
- reproducibility
- faster (and efficient) progress
- feed future researchers
But wait, there’s more

• Get credit and attribution (the village)

• Comply with publishers’ data sharing policies

• Meet funder mandates

• Meet institutional requirements

• Respond to community norms and practices
And still more…

- Researchers & Contributors
- Data and Software
- Grants and Projects
- Publications
- Funders
- Institutions
DataCite’s approach

Provide technical infrastructure:

• Create DOIs for research data
• Build and adopt services that promote data sharing
• Integrate with other community services

Provide community infrastructure:

• Advocate & communicate about the importance of data sharing
Welcome to DataCite

Locate, identify, and cite research data with the leading global provider of DOIs for research data.

Find what you're looking for by searching millions of records with extensive, reliable metadata.

Share your data and reuse the data of others to create the highest impact in the research community.

Cite your research sources with confidence, and receive proper credit when your work is reused.

Connect your research – publications, datasets, software, authors, institutions, and funding data all in one place.

Not-for-profit global initiative – Member organization – Community driven – over 900 data centers – over 9 million DOIs
The centrality of a DOI

DOI = Digital Object Identifier

an alphanumerical string created to:

• uniquely *identify/name* digital content
• serve as a *stable, persistent link* to that content’s location on the web
Down in the weeds

1. Take a dataset

2. Describe it

   - Title
   - Authors
   - Year
   - Description
   - And others...

3. Assign a DOI

   10.1234/exampledata

4. Reuse and reference!

   ATLAS Collaboration, “Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC: $H \rightarrow \gamma \gamma$,”
   http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44

5. Enjoy the benefits

   - Findability
   - Track citations
   - Reusability
   - Measure impact

✅ Unique
✅ Persistent
Joint Declaration of Data Citation Principles

1. **Importance:** legitimate, citable products of research

2. **Credit and Attribution:**
scholarly credit and normative and legal attribution to all contributors to the data

3. **Evidence:** In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited

4. **Unique Identification:** include a persistent method for identification

5. **Access:** data citations should facilitate access to the data themselves

6. **Persistence:** identifiers, and metadata should persist

7. **Specificity and Verifiability:**
data citations should facilitate identification of, access to, and verification of data

8. **Interoperability and Flexibility:**
data citation methods should be sufficiently flexible

---

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final].
DataCite’s Service integration = building bridges

make it easier
integrate with related content
provide impact
give credit
Putting the pieces together

- Researchers & Contributors
- Data and Software
- Grants and Projects
- Publications
- Funders
- Institutions
Linking Data with Data

Why it matters

- provide a complete picture of the data environment
- multiple versions of the same dataset
- subsets of larger datasets or heterogeneous collections
- dynamic data
- software
- workflows
Linking data to data example

Dataset

References, supplements

1 Related Work

Raw Data - “Biomineralization control related to population density under ocean acidification”
Stefano Goffredo, Fiorella Prada, Erik Caroselli, Bruno Capaccioni, Francesco Zaccanti, Luca Pasquini, Paola Fantazzini, Simona Ferrmani, Michela Reggi, Oren Levy, Katharina Elisabeth Fabricius, Zvy Dubinsky & Giuseppe Falini
Fileset published 2014 via Figshare

https://doi.org/10.6084/M9.FIGSHARE.941061

Cite  Add to ORCID record
Linking Data with Researchers & Contributors

Why it matters
- Credit and attribution
- Answers who, what, when, where
- Link one or more contributors to research output
Seamless integration with ORCID

**Researchers:** (1) use ORCID iD when submitting dataset (2) authorize DataCite to update your ORCID record.

**Data centers:** (1) collect ORCID identifiers during submission (2) embed iD in the work and include the iD when submitting to DataCite.

**DataCite:** Upon receipt of data from a data center with a valid identifier, DataCite automatically pushes information to the researcher’s ORCID record.
Linking Data with Articles

Why it matters

• Increase visibility and discovery of research data and articles
• Place research data in the right context to enable reuse
• Support credit attribution

Challenges:

• Data underlying findings are not always fully available
• Data underlying findings described in a are made available, but hidden in supplementary information
• Data underlying the findings are available, but not properly linked to/from article
Linking Data with Articles: Follow FAIR Data Principles

Force11: Data Sharing Principles

FAIR data

- Findable
  - Describe your data in a data repository
  - Apply a persistent identifier

- Accessible
  - Consider what will be shared
  - Obtain participant consent & perform risk management

- Interoperable
  - Use open formats
  - Consistent vocabulary
  - Common metadata standards

- Reusable
  - Consider permitted use
  - Apply appropriate licence

http://slideshare.net/lshtm/preparing-data-for-sharing-the-fair-principles
Example 1: One article links to five datasets

Journal article

Temperature-Induced Syntheses, Iodine Elimination, Enantiomers Resolution, and Single-Crystal-to-Single-Crystal Transformation of Imidazole-Co(II) Coordination Polymers with Amino-isophthalic Acid as Co-Ligand
Journal article published June 9, 2016

Related data

CCDC 1414538: Experimental Crystal Structure Determination
Hui-Fang Zhou, Tian He, Ke-Fen Yue, Yong-Liang Liu, Chun-Sheng Zhou, Ni Yan & Yao-Yu Wang
Work published 2016 via Cambridge Crystallographic Data Centre

CCDC 1059747: Experimental Crystal Structure Determination
Hui-Fang Zhou, Tian He, Ke-Fen Yue, Yong-Liang Liu, Chun-Sheng Zhou, Ni Yan & Yao-Yu Wang
Work published 2016 via Cambridge Crystallographic Data Centre

CCDC 1484151: Experimental Crystal Structure Determination
Hui-Fang Zhou, Tian He, Ke-Fen Yue, Yong-Liang Liu, Chun-Sheng Zhou, Ni Yan & Yao-Yu Wang
Work published 2016 via Cambridge Crystallographic Data Centre
Example 2: Software described in Journal of Open Source Software

**Armadillo: a template-based C++ library for linear algebra**
Conrad Sanderson & Ryan Curtin
Journal article published June 10, 2016 via JOSS

[DataCite (Crossref)](http://doi.org/10.21105/JOSS.00026)

**Armadillo C++ Linear Algebra Library**
Conrad Sanderson
Work published 2016 via Zenodo

[DataCite (Crossref)](http://doi.org/10.5281/ZENODO.55251)
Example 3: PLOS articles linked with at least one DataCite DOI

- **Public Library of Science (PLoS)**
  - 340 Works
  - **Seasonality and Locality Affect the Diversity of Anopheles gambiae and Anopheles coluzzii Midgut Microbiota from Ghana**
    - Jeweina Akerløf, Mathilde Gendrin, Nana Adjoe P. Pels, Dorothy Yeboah-Manu, George K. Christophides & Michael D. Wilson
    - Journal article published June 20, 2016 via PLOS ONE
    - DataCite (Crossref) 1
    - http://doi.org/10.1371/JOURNAL.PONE.0157529
  - **Genetic Diversification and Dispersal of Taro (Colocasia esculenta (L.) Schott)**
    - Journal article published June 17, 2016 via PLOS ONE
    - DataCite (Crossref) 1
    - http://doi.org/10.1371/JOURNAL.PONE.0157712
  - **Mosquito Saliva Increases Endothelial Permeability in the Skin, Immune Cell Migration, and Dengue Pathogenesis during Antibody-Dependent Enhancement**
    - Michael A. Schmid, Dustin R. Glasner, Sanjana Shah, Daniela Michlmayr, Laura D. Kramer & Eva Harris
    - Journal article published June 16, 2016 via PLoS Pathog
    - DataCite (Crossref) 1
    - http://doi.org/10.1371/JOURNAL.PPAT.1005678

- **Publisher**: Public Library of Science (PLoS)
- **No. of Works**: 542
- **Related Content**:
  - Seasonality and Locality Affect the Diversity of Anopheles gambiae and Anopheles coluzzii Midgut Microbiota from Ghana
  - Genetic Diversification and Dispersal of Taro (Colocasia esculenta (L.) Schott)
  - Mosquito Saliva Increases Endothelial Permeability in the Skin, Immune Cell Migration, and Dengue Pathogenesis during Antibody-Dependent Enhancement
Linking data to funders and organizations

Why it matters

- The research environment is complicated (everyone wants credit)
- Bring together scholarly output with all of the stakeholders
  - Researchers
  - Funders
  - Organizations
  - Grants
  - Projects
Linking to organization identifiers

a documented need for a comprehensive, open, and accessible organization identifier infrastructure

Content identifiers:
DataCite, Crossref

Contributor identifiers:
ORCID

Organization Identifiers?
Organization Identifier Working Group

Summary

The Organization Identifier (OrgID) Working Group was established in January 2017 to refine the structure, principles, and technology specifications for an open, independent, non-profit organization identifier registry to facilitate the disambiguation of researcher affiliations. The scope of work includes three separate but interdependent areas: Governance, Registry Product Definition, and Business Model & Funding. The goal of the Working Group is to create an implementation plan by the end of 2017.
Data, a first-class research output

LEARN MORE!
Journal articles v/s data

Primacy of journal articles
• currency of research
• sophisticated methods to gauge impact (citations, page views, downloads)
• understand relative impact and identify relationships

What about data?
• 1st class scholarly object
• lagging infrastructure
• broader role in research process
• has its own use & reuse profile
• Community of best practices are lacking
Data sets are different!

- Atomicity
- Versioning
- Non-congruency
- Composability

Thank you Matt Jones, NCEAS
For example - usage tracking: downloads

How do we count downloads? Sum, average, maximum, whole package?
A path forward

**Make Data Count: data usage stats and data citations**

1. provide formal recommendation for measuring data usage

2. further develop *Data Level Metrics (DLM)* Hub and services

3. expose exemplars to drive adoption

4. engage in community building
Thank you!

patricia.cruse@datacite.org
https://www.datacite.org
Twitter: @datacite