Evaluating the possibilities of DataCite for developing ‘Open data metrics’ on the production and usage of datasets worldwide

Nicolas Robinson-Garcia, Phillipe Mongeon, Wei Jeng & Rodrigo Costas
Sharing and citing data

Promotion of **data sharing infrastructures**
- Data repositories
- Data Citation Index
- Persistent linkages (DOIs)

Promotion of **data sharing practices**
- Data sharing incentives
- Perceived benefits
Sharing and citing data

Promotion of data sharing infrastructures

- Data repositories
- Data Citation Index
- Persistent linkages (DOIs)

Maximizing investment

Promotion of data sharing practices

- Data sharing incentives
- Perceived benefits

Searching for evidences of data sharing
Aims of this study

1. Who shares data?
   - Which countries are sharing scientific data (in DataCite)?
   - Are there biases by discipline (in DataCite)?

2. Are there evidences of data reuse?
   - Are researchers using DOIs to link papers to datasets?
   - Are they mentioning datasets through social media?
From citing papers to citing data

Citations to publications
• Based on researchers’ communication patterns
• Influenced by research evaluation schemes
• Highly standardized and extended within the scientific practice

Citations to datasets
• Promoted by funding bodies
• Not embedded on scholarly communication patterns
• Heterogeneous forms of acknowledgement (paper, dataset, none...)
# The metadata schema of DataCite

## Mandatory fields

<table>
<thead>
<tr>
<th>ID</th>
<th>Property</th>
<th>Obligation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identifier (with mandatory type sub-property)</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>Creator (with optional name identifier and affiliation sub-properties)</td>
<td>M</td>
</tr>
<tr>
<td>3</td>
<td>Title (with optional type sub-properties)</td>
<td>M</td>
</tr>
<tr>
<td>4</td>
<td>Publisher</td>
<td>M</td>
</tr>
<tr>
<td>5</td>
<td>PublicationYear</td>
<td>M</td>
</tr>
<tr>
<td>10</td>
<td>ResourceType (with mandatory general type description sub-property)</td>
<td>M</td>
</tr>
</tbody>
</table>

**Source:** DataCite Metadata Working Group (2016).  
[http://doi.org/10.5438/0012](http://doi.org/10.5438/0012)
Citations to data

Recommended format

Creator (Publication Year):
Title. Publisher. Identifier

Source: DataCite Metadata Working Group (2016).
http://doi.org/10.5438/0012
Preliminary results

Types of data

- Dataset: 41.6%
- Other: 19.4%
- Text: 17.5%
- Collection: 6.8%
- Image: 14.3%
Preliminary results

Availability of Publisher information

- 80% Reported
- 20% Not reported
- Unknown
Preliminary results

What is a Publisher?

- Unassigned: 39.65%
- Thematic repository: 11.46%
- Institutional repository: 10.28%
- Research body: 5.49%
- Multidisciplinary repository: 2.01%
- Publisher: 0.55%
- National repository: 0.28%
- Private data producer: 0.26%
- Professional body: 0.25%
- Conference: 0.11%
- Author: 0.03%
- Educational body: 0.03%
### Preliminary results

- Citation and altmetric analysis
- Matches based only on DOIs!

<table>
<thead>
<tr>
<th>Description</th>
<th>Citations</th>
<th>Altmetric.com (Twitter)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DataCite records with DOI</td>
<td></td>
<td>6352875</td>
</tr>
<tr>
<td>Records with metrics (matched on DOI)</td>
<td>6432</td>
<td>14314</td>
</tr>
<tr>
<td>%records with metrics</td>
<td>0.10%</td>
<td>0.23%</td>
</tr>
<tr>
<td>Intensity (records with metrics/metric)</td>
<td>17.9</td>
<td>4.1</td>
</tr>
</tbody>
</table>
The no. 1s

The most cited dataset?

Thermal Transport Measurements of Individual Multiwalled Nanotubes
P. Kim, A. Majumdar, P. L. McEuen & L. Shi
Report published 2001 via The American Physical Society
Physical Review Letters

The most tweeted dataset?

Prevalence and use of Twitter among scholars
Jason Priem, Kaitlin Costello & Tyler Dzuba
Poster describing the results of a study examining Twitter use among a sample of 8,826 academics from 5 universities.
Preliminary results

Country of origin based on publisher info

<table>
<thead>
<tr>
<th>Country</th>
<th># Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK</td>
<td>1728428</td>
</tr>
<tr>
<td>Germany</td>
<td>966289</td>
</tr>
<tr>
<td>Switzerland</td>
<td>591062</td>
</tr>
<tr>
<td>USA</td>
<td>560799</td>
</tr>
<tr>
<td>Canada</td>
<td>81471</td>
</tr>
<tr>
<td>Spain</td>
<td>32795</td>
</tr>
<tr>
<td>Netherlands</td>
<td>26791</td>
</tr>
<tr>
<td>Italy</td>
<td>25241</td>
</tr>
<tr>
<td>Australia</td>
<td>21059</td>
</tr>
<tr>
<td>Ireland</td>
<td>19416</td>
</tr>
<tr>
<td>Austria</td>
<td>18571</td>
</tr>
<tr>
<td>USA, UK</td>
<td>12981</td>
</tr>
<tr>
<td>France</td>
<td>9443</td>
</tr>
<tr>
<td>Denmark</td>
<td>8804</td>
</tr>
<tr>
<td>BE, DE, IT, NL, ES</td>
<td>5366</td>
</tr>
<tr>
<td>Sweden</td>
<td>2816</td>
</tr>
<tr>
<td>Korea</td>
<td>2</td>
</tr>
</tbody>
</table>
Preliminary results

Country of origin based on publisher info

- UK
- Germany
- Switzerland
- USA
- Canada

Legend:
- Thematic repository
- Institutional repository
- Research body
- Multidisciplinary repository
- Publisher
- National repository
- Private data producer
- Professional body
- Conference
- Author
Bibliometric limitations

Technical

- Heterogeneity of sources
- Lack of basic data (affiliation)
- Lack of standard normalisation

Conceptual

- Publication vs. Data production patterns
- Data citations vs. Data reuse
- Conceptual heterogeneity
**Some examples**

**Publication author vs. Data producer distribution**

<table>
<thead>
<tr>
<th>Authors (non disambiguated)</th>
<th>WoS records</th>
<th>Creators</th>
<th>DataCite records</th>
</tr>
</thead>
<tbody>
<tr>
<td>WANG, Y</td>
<td>56596</td>
<td>Geml, József</td>
<td>487363</td>
</tr>
<tr>
<td>ZHANG, Y</td>
<td>54203</td>
<td>Ryberg, Martin</td>
<td>487351</td>
</tr>
<tr>
<td>WANG, J</td>
<td>49817</td>
<td>Lumbsch, H. Thorsten</td>
<td>487350</td>
</tr>
<tr>
<td>LIU, Y</td>
<td>46307</td>
<td>Tedersoo, Leho</td>
<td>487350</td>
</tr>
<tr>
<td>LI, Y</td>
<td>45773</td>
<td>Hampe, Felix</td>
<td>487350</td>
</tr>
</tbody>
</table>
Most productive data creator

József Geml
Dr. J. (József) Geml, Tenure tracker – Biodiversity Dynamics, Mycology

Contact
Email: jozsef.geml@naturalis.nl
Phone: +31 (0)71-58 87 265
Room number: Darwinweg 4

“We know so little about fungi that we unknowingly trample on hundreds of undiscovered species day by day.”
## Some examples

### Heterogeneity of Publisher information

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DOI numbers</td>
<td>JHEP 1311 (2013) 183</td>
</tr>
<tr>
<td>Copyright statements</td>
<td>ETH-Bibliothek Zürich, Bildarchiv / Fotograf: Unbekannt / Fel_027418-VE / Public Domain Mark (see the metadata of copyright: <a href="http://www.epics.ethz.ch/index/ethbib.bildarchiv/ETHBIB.Bildarchiv_Fel_008192-RE_257002.html">http://www.epics.ethz.ch/index/ethbib.bildarchiv/ETHBIB.Bildarchiv_Fel_008192-RE_257002.html</a>)</td>
</tr>
<tr>
<td>Hashtags???</td>
<td>&quot;#IHaveWrittenMyOwnOneNewScientificPaperOnTheGeographyAndMarsLife. #IHaveAlreadySuccessfullyOFFICIALYPublishedItOn #DiscoveryNews #ScienceAlert #DiscoveryChannelIndia #AndAlsoOn #DiscoveryCommunications #AndItsOFFICIALPublicationIsThereOnAllOfItSoPlea&quot;</td>
</tr>
</tbody>
</table>
Thank you!