Incorporating Data Sharing activities in the Reward System of Science

Rodrigo Costas
Center for Science and Technology Studies (CWTS-Leiden University)

5th April 2017, Barcelona
Outline

• Data metrics and the reward system of science
  • *The value of Research Data* (Knowledge Exchange, 2013)

• Practical possibilities based on DataCite
  • *Discussion of issues*
  • *Recommendations for improvement*
## Data metrics and the reward system of Science

### Table 3: Metrics scenarios for the data publishing models (and comparison with the current scientific publication model)

<table>
<thead>
<tr>
<th>Types of metrics</th>
<th>Currently available tools with possibilities for “data metrics”</th>
<th>Metric dimensions</th>
<th>Scientific publication</th>
<th>Data Publication</th>
<th>Data Publication</th>
<th>Stand-alone data publications</th>
<th>Journal data publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data publication &amp; citation-based metrics</td>
<td>- Data Citation Index (Web of Science) - Google Scholar - Scopus - Microsoft Academic Search - DataCite</td>
<td>Size-dependent</td>
<td>Yes</td>
<td>Difficult (1)</td>
<td>Yes (4)</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Size-independent</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Direct average performance</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Source-based performance</td>
<td>Yes</td>
<td>No</td>
<td>Yes (3)</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Altmetrics-based metrics</td>
<td>- ImpactStory - Twitter, Facebook - Mendeley - CiteULike - Repositories - Data Journals</td>
<td>Social media indicators</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Readership counts</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Downloads &amp; views counts (DUI metrics)</td>
<td>Yes</td>
<td>Difficult (2)</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

Main challenges for data metrics

1) Lack of recognition for data sharing and data reuse
2) Reduction in research opportunities
3) Misuse of data (including lack of credit to the data creators)
4) Perceived potential decrease in “research quality” by reusing the same data
5) Disciplinary differences
The data sharing ‘vicious circle’

Scholars don’t share because is not rewarded and is not worth the time

Data sharing and reuse transactions are low, making data metrics difficult or impossible

Data sharing and data reuse cannot be measured, making it difficult to reward it
How to break the ‘vicious circle’?

- Development of reward systems that include data (alt)metrics
- Standardization, formalization and tracking of data publications/citations
- Reduction of costs
- Reduction negative perception
- Solution of technical & organizational problems
- Coordination of initiatives
- Development of standards & interoperability
So, let’s do it!

• How can we ‘practically’ measure open data sharing activities? What are the challenges?

• Robinson-García, N.; Mongeon, P.; Jeng, W.; Costas, R. (submitted). DataCite as a novel bibliometric source: coverage, strengths and limitations

• Simple approach: to analyze ‘data production’ with DataCite
• What and why?
  – International non-for-profit consortium (since 2009) of public research institutes, funding bodies and publishers worldwide
  – Mission: to promote open research data accessibility and tracking, advocating for the use of Digital Object Identifiers (DOI)
  – Largest data source on registered data publications

• How?

Other metadata:
- Contributor
- Resource type
- Related identifier
- Data Center
- And others…
Critical analysis: 7,440,415 records (Apr. 2016)

Figure 1. Distribution on metadata information by fields

* Non-standardized fields
### Data types

Sometimes texts are also here!

**Conceptual issues:**
- What is a ‘dataset’?
- What is data?

<table>
<thead>
<tr>
<th>Data type</th>
<th>Number of records</th>
<th>%</th>
<th>Most frequent subtypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>1,867,627</td>
<td>41.69</td>
<td>Dataset (63.5%), Metadata (5.8%), Data package (4.1%)</td>
</tr>
<tr>
<td>Text</td>
<td>786,882</td>
<td>17.56</td>
<td>Conference papers (15.5%), Journal articles (15.4%), Report (10.1%)</td>
</tr>
<tr>
<td>Image</td>
<td>641,404</td>
<td>14.32</td>
<td>Image (11.9%), Figure (11.2%), Plate (8.1%)</td>
</tr>
<tr>
<td>Collection</td>
<td>303,638</td>
<td>6.78</td>
<td>Collection (20.7%), Gaussian job archive (9.1%), Report (4.7%)</td>
</tr>
<tr>
<td>Software</td>
<td>12,340</td>
<td>0.03</td>
<td>Simulation tool (16.9%), Software (10.8%), Code (5.3%)</td>
</tr>
<tr>
<td>Audiovisual</td>
<td>4,470</td>
<td>0.10</td>
<td>Audiovisual (43.8%), Media (23.9%), Teaching material (8.5%)</td>
</tr>
<tr>
<td>Film</td>
<td>960</td>
<td>0.02</td>
<td>Experiment (5.4%), Video (0.4%), Animation (0.1%)</td>
</tr>
<tr>
<td>Physical Object</td>
<td>587</td>
<td>0.01</td>
<td>Archival object (63.9%), HIAPER-HAIS airborne sensor (2.4%), Physical object (0.9%)</td>
</tr>
<tr>
<td>Event</td>
<td>508</td>
<td>0.01</td>
<td>Conference presentation (73.4%), Presentation (9.6%), Event (1.6%)</td>
</tr>
<tr>
<td>Model</td>
<td>470</td>
<td>0.01</td>
<td>Model (2.8%), Ontology (0.9%), Shapefiles (0.2%)</td>
</tr>
<tr>
<td>Interactive Resources</td>
<td>287</td>
<td>0.01</td>
<td>Interactive resources (12.2%), Learning object (2.1%), Sites Web (0.3%)</td>
</tr>
<tr>
<td>Sound</td>
<td>234</td>
<td>0.01</td>
<td>Recording, oral (4.3%), Sound (0.4%), Conference (0.4%)</td>
</tr>
<tr>
<td>Workflow</td>
<td>209</td>
<td>&lt; 0.01</td>
<td>Taverna 2 workflow (7.2%), Workflow (1.0%), RapidMiner workflow (0.5%)</td>
</tr>
<tr>
<td>Service</td>
<td>18</td>
<td>&lt; 0.01</td>
<td>Service (88.9%), S-map (5.6%), Data provider (5.6%)</td>
</tr>
<tr>
<td>Other</td>
<td>871,549</td>
<td>19.45</td>
<td>Data sheet (98.2%), Oceanographic cruise (0.7%), Field expedition (0.7%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4,480,077</strong></td>
<td><strong>100</strong></td>
<td></td>
</tr>
</tbody>
</table>
**Dates**

Figure 6. Number of records per year using the publication year in DataCite, 1950-2020 period.

A. Contents of the record

- **A. E. Borell with wildcat**
  - Date: Oct 1929
  - Location: Ruby Lake, Nev. S. Whitlape, Col., Ikeme, Analone, Photograph: H. H. Sheldon

B. DataCite record

- **A. E. Borell with wildcat**
- **H. H. Sheldon**
- Work published 1929 via Museum of Vertebrate Zoology

[DOI] https://doi.org/10.7294/X79M656M

**Cite**
## Origin of the records / affiliations

<table>
<thead>
<tr>
<th>Type of publishers</th>
<th># records</th>
<th>% data records</th>
<th># publishers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thematic repository</td>
<td>2205204</td>
<td>67.78%</td>
<td>18</td>
</tr>
<tr>
<td>Institutional repository</td>
<td>852954</td>
<td>83.26%</td>
<td>35</td>
</tr>
<tr>
<td>Research body</td>
<td>764962</td>
<td>43.18%</td>
<td>33</td>
</tr>
<tr>
<td>Multidisciplinary repository</td>
<td>408355</td>
<td>88.45%</td>
<td>2</td>
</tr>
<tr>
<td>Publisher</td>
<td>149305</td>
<td>9.85%</td>
<td>24</td>
</tr>
<tr>
<td>National repository</td>
<td>40634</td>
<td>42.33%</td>
<td>5</td>
</tr>
<tr>
<td>Firm data producer</td>
<td>20704</td>
<td>9.67%</td>
<td>6</td>
</tr>
<tr>
<td>Professional body</td>
<td>19215</td>
<td>0.00%</td>
<td>5</td>
</tr>
<tr>
<td>Conference</td>
<td>18571</td>
<td>0.00%</td>
<td>1</td>
</tr>
<tr>
<td>Author</td>
<td>8025</td>
<td>0.00%</td>
<td>2</td>
</tr>
<tr>
<td>Educational body</td>
<td>2326</td>
<td>93.59%</td>
<td>1</td>
</tr>
<tr>
<td>Not found</td>
<td>621544</td>
<td>78.74%</td>
<td>24</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>5111799</strong></td>
<td><strong>72.47%</strong></td>
<td><strong>156</strong></td>
</tr>
</tbody>
</table>

---

### Dataset Bar Chart

- **Data-Planet**
- **TU Loodamusmuuseum**
- **Digital Science**
- **PANGAEA-Publishing Network for Geoscience Data**
- **E-Pics Bildarchiv**
- **Imperial College London**
- **ResearchGate**
- **Global Biodiversity Information Facility**
- **E-Periodica**
- **Pitt Quantum Repository**
- **Long Term Ecological Research Network**
- **Partnership for Interdisciplinary and E-Literature**
- **Leibniz-Institut für Astrophysik**
- **HEPData.net**
- **NRCT Data Center**
- **Dryad**
- **Universität Zürich, ZORA**
- **University of British Columbia**
- **German Medical Science**
- **ETH E-Collection**

---

### Number of Data Records by Datacenter and Publisher

- **USA**
- **United Kingdom**
- **Estonia**
- **Germany**
- **Switzerland**
- **Denmark**
- **Canada**
- **Thailand**
- **Australia**
- **Ireland**
- **Sweden**
- **Italy**
- **New Zealand**
- **Romania**
- **Czech Republic**
- **Peoples R China**
- **South Africa**
- **Belgium**
- **Ghana**
- **Spain**
- **Poland**
- **Hungary**
- **Tanzania**
- **Austria**
- **Ungary**

- **Dataset**
- **Text**
- **Collection**
- **Image**
- **Software**
- **Audiovisual**
- **Physical/Object**
- **Film**
- **InteractiveResource**
- **Sound**
- **Other**
Other issues

<table>
<thead>
<tr>
<th>Authors</th>
<th>WoS</th>
<th>Data Creators (not disambiguated)</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>
</tbody>
</table>

Need to reconsider “data production modes”
Aspects to improve for data metrics using DataCite

• Data quality improvement:
  – Structure and quality of records
    • Standardization of fields (e.g. data creators, publishers, etc.)
    • Refinement of data types (more detail)
  – Completeness (e.g. common metadata for all records)
  – Relevant missing elements:
    • Affiliations of data creators (institutional and geographic data)
  – Coverage analysis
    • What are we missing?

• Conceptual issues:
  – Operationalization of the concept of ‘data’
    • Typologies of data
  – Dates for data
  – Modes of data production
Thank you very much!

Questions? Comments?