DataCite
Linking Datasets to Publications

Matt Buys
June 27, 2024
CHORUS Forum

@datacite
@datacite@openbiblio.social
Connecting research, advancing knowledge

The how

The what
Great progress in data sharing over the last decade

National and funder policies in support of data sharing have dramatically raised awareness among communities and driven adoption.

Infrastructure available to support preservation, publication, curation and controlled access to meet a wide range of needs.

Over 15 million datasets registered with DataCite Commons commons.datacite.org
And still

Broad discoverability is not yet the norm for open data

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Description of the datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>7,182,036</td>
<td>Total profiles in ORCID</td>
</tr>
<tr>
<td>12,686</td>
<td>Total profiles in ORCID with at least one dataset</td>
</tr>
<tr>
<td>80,555</td>
<td>Total distinct datasets in ORCID</td>
</tr>
<tr>
<td>4,284</td>
<td>Total profiles in ORCID with at least one dataset indexed to Zenodo, Dryad, Figshare or DataCite</td>
</tr>
</tbody>
</table>

Research assessment frameworks focus heavily on publications & few include data

“I think one barrier to data sharing is the merit review process within institutions for tenure and promotion; things such as ‘how many people accessed your dataset’ are not valued”.


Exploring the relevance of ORCID as a source of study of data sharing activities at the individual-level: a methodological discussion. arXiv:2105.11825
Cultivating Change
Easy & Rewarding

Adoption of Open Research
by Hendricks & Buys (derived from Nosek: cos.io/blog/strategy-for-culture-change)
Researchers still perceive data sharing as a burden that brings little professional benefit.

We need to make it rewarding for researchers to share & cite data.

“I don’t have time. [...] It’s not incentivized either [...] so why would I spend some time on something where it really amounts to no recognition [...]. People who are in this community might think I’m a good person [...] and that’s fine. I can accept that. But I’m not going to track it, necessarily, because it doesn’t amount to anything that’s recognized.”

Professor, Anesthesiology, Canada

Data Citations

Provide a link between the dataset and another research object, pointing to the data being used or reused

A useful step toward evaluation of data usage:

- Valued by researchers
- Existing workflows available for repositories & publishers
- AI tools developed to identify data citations

Gregory et al., Tracking data: A survey investigating disciplinary differences in data citation. *Quantitative Science Studies* 2023; https://doi.org/10.1162/qss_a_00264
Make Data Count

**Build** open infrastructure and community-based standards

**Advocate** through outreach and adoption campaigns

**Contextualize** with evidence-based bibliometric studies
The Data Citation Corpus

Partnering across the community to support open and meaningful tools that provide insights on data usage

http://corpus.datacite.org/
Building on existing infrastructure
Immediate priorities

- We have to stop focusing on altering publisher behavior
- Repositories can and should add citation links to dataset metadata
- Text mining and NLP must be used to expedite surfacing data citation
- Leverage open metadata, persistent identifiers, aggregate linkages not exposed
IT’S TIME TO MAKE DATA COUNT TOGETHER.

Join us for the Make Data Count Summit on September 5–6, 2024 in London, UK

Live, in-person only event.