CASRAI Research Dataset-Level Metrics

Alex Ball Thomas Ingram

2016-04-14

Rationale of the Interest Group

Two main concerns:

- Researchers should get career benefit from sharing data
  - How to demonstrate impact?
- Researchers can't tell if their data are reusable
  - How to get quality assurance?

Can we make the solutions more scalable with metrics?

Aims of the Interest Group

Define sets of **metrics** that can be used

- as a starting point for decisions about impact and quality
- as a motivator for doing the right thing

Potential **use cases**:

- a publisher displaying metrics for data underlying an article
- a repository giving depositors feedback about how their data are being used
- a funder wanting to evaluate the data outputs of funded research
- a university wanting to highlight impact of data outputs
- a university wanting to use data outputs in recruitment or promotion decisions

Related work: Making Data Count

- CDL/PLOS/DataONE project, Oct 2014 – Oct 2015
- Found researchers most interested in citations and downloads
- Repurposed Lagotto (PLOS Article Level Metrics) for data
  - Existing support for counting downloads, formal citations
  - Harvesting informal citations (links, IDs) from open access corpora
  - Monitoring mentions in social media, DataCite metadata, etc.
- Further reading
  - Lagotto on GitHub
  - 'Making data count', `http://doi.org/10.1038/sdata.2015.39`
  - 'When counting is hard', `https://blog.datacite.org/when-counting-is-hard/`

Related work: RDA/WDS Publishing Data Bibliometrics WG

- Oct 2014 – March 2016
- Found all stakeholder groups most interested in citations and downloads
- Draft recommendations:
  - Funders to mandate data deposit and citation
  - Stakeholders to agree on how to use repository statistics
  - Stakeholders to use multiple metrics, not rely on just one
- Further reading
  - `https://rd-alliance.org/groups/rdawds-publishing-data-bibliometrics-wg.html`

Related work: NISO Altmetrics Group B

- Two-phase project, June 2013 – Jan 2015 – Mar 2016
- Group B: ‘non-traditional research outputs and identifiers’
- Draft recommendations:
  - Implement FORCE11 data citation principles
  - Use COUNTER standards to count human-initiated downloads
  - Define new standards for counting (or not) machine interactions
- Further reading
  - `http://www.niso.org/topics/tl/altmetrics_initiative/`

Quality Dataset-Level Metrics for Repositories WG

Use case:

- A generalist data repository applies a set of metrics to a deposited dataset to determine whether it meets a given level of quality.

Aim:

- Define a clear set of metrics such repositories could use.

Activity:

- Two telcons
- Workshop at BioMed Bridges closing symposium
- Relaunch in 2016 for a three-month sprint

Quality considerations

- Does the submission have integrity?
- Was it collected according to an accepted methodology?
- Is the raw-to-result processing reproducible?
- What kinds of format(s) does it use?
- Is it properly structured and labelled?
- Does it include additional disciplinary-associated metadata?
- Has it been documented sufficiently for reuse?
- Has the dataset been appropriately licensed?
- Have appropriate access restrictions been applied?
- Has it already passed an expert review?
- Does it compare favourably with ‘known good’ datasets?

Thank you for your attention –

now get involved!

http://www.casrai.org/Quality_DLMs_for_Repositories