Bringing data in sight

Data citations in research

Forum Bibliometrie 2021

Kathleen Gregory | @gregory_km

Scholarly Communications Lab, Ottawa/Vancouver

Data Archiving and Networked Services, Netherlands

The rise of research data

Data as ‘first class research objects’

The First Principle: Importance (status)

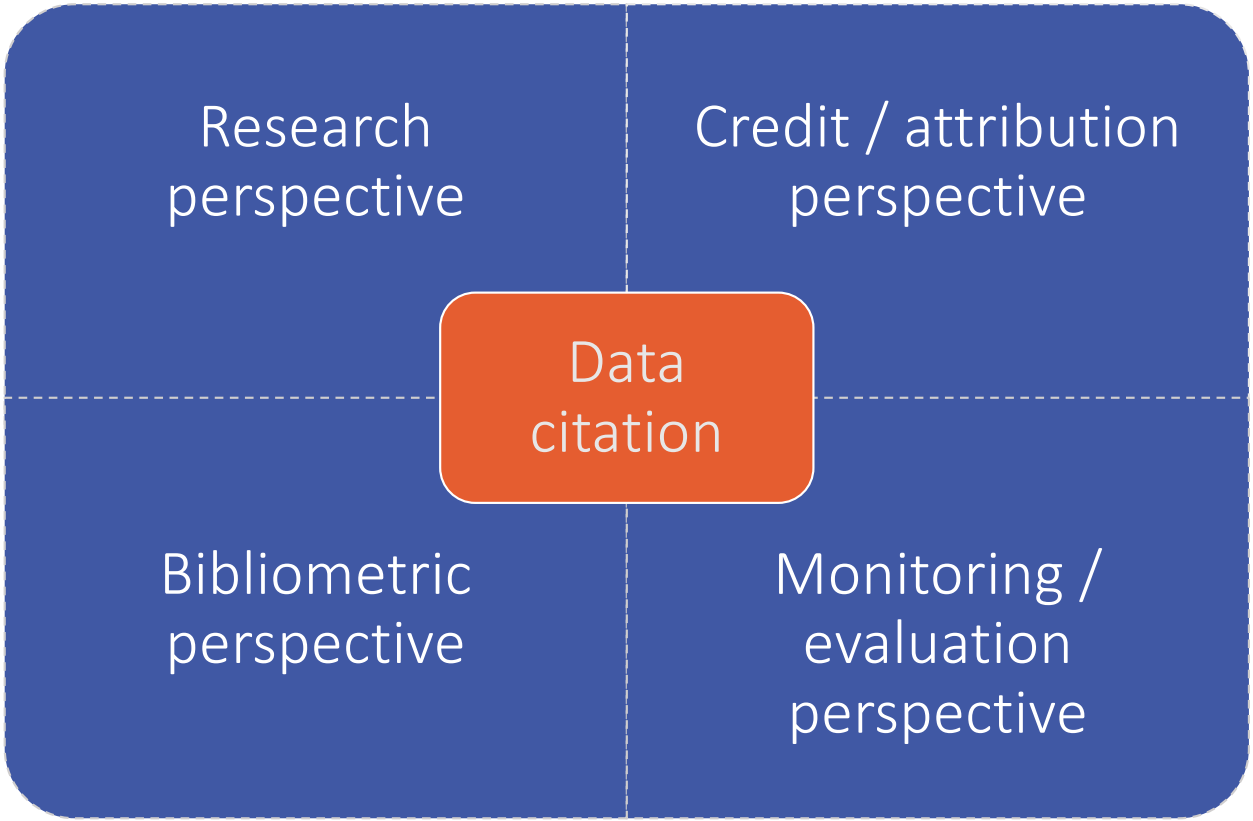
Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications

CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013.

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles, 2014.

The potentials and promises of data citation

Why is data citation important?



Outline

→ Why is data citation challenging?

→ Bringing data “in sight”

- Facilitating data citation
- Studying data citation practices

→ Bringing data insight

- Meaningful Data Counts project

Why is data citation challenging?

Challenges for data citation

Data are defined differently in different communities.

→ One person's signal is another person's noise.

→ Data do not stand alone.

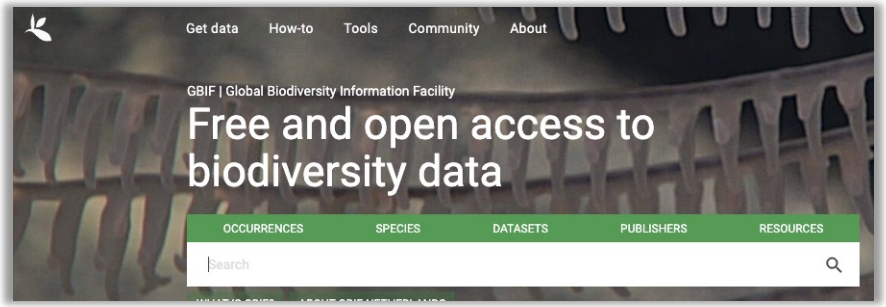
→ Disciplinary differences in data sharing and reuse

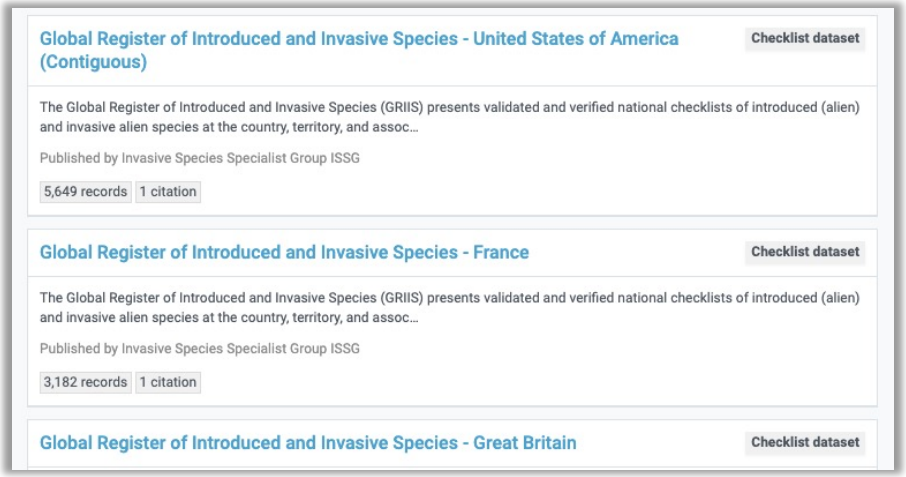
istockphoto.com by kotoffei

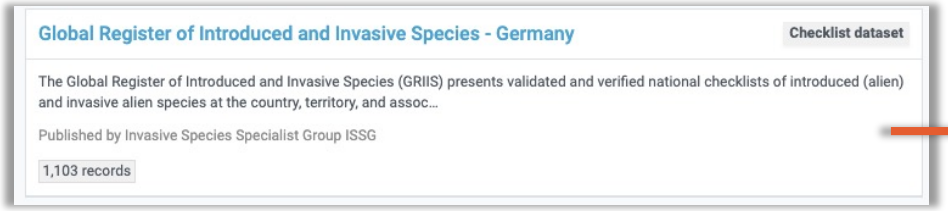
Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.

Challenges for data citation

Granularity







id	locationID	countryCode	occurrenceStatus	establishmentMeans
119000	Germany	DE	Present	Native|Alien
119001	Germany	DE	Present	Alien
119002	Germany	DE	Present	Alien

Borgman, C. L. (2016). Data citation as a bibliometric oxymoron. In C. R. Sugimoto (Ed.), Theories of informetrics and scholarly communication (pp. 93–116). De Gruyter. https://doi.org/10.1515/9783110308464-008

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6–20. https://doi.org/10.1002/asi.23917

Challenges for data citation

Fixity – dynamic data

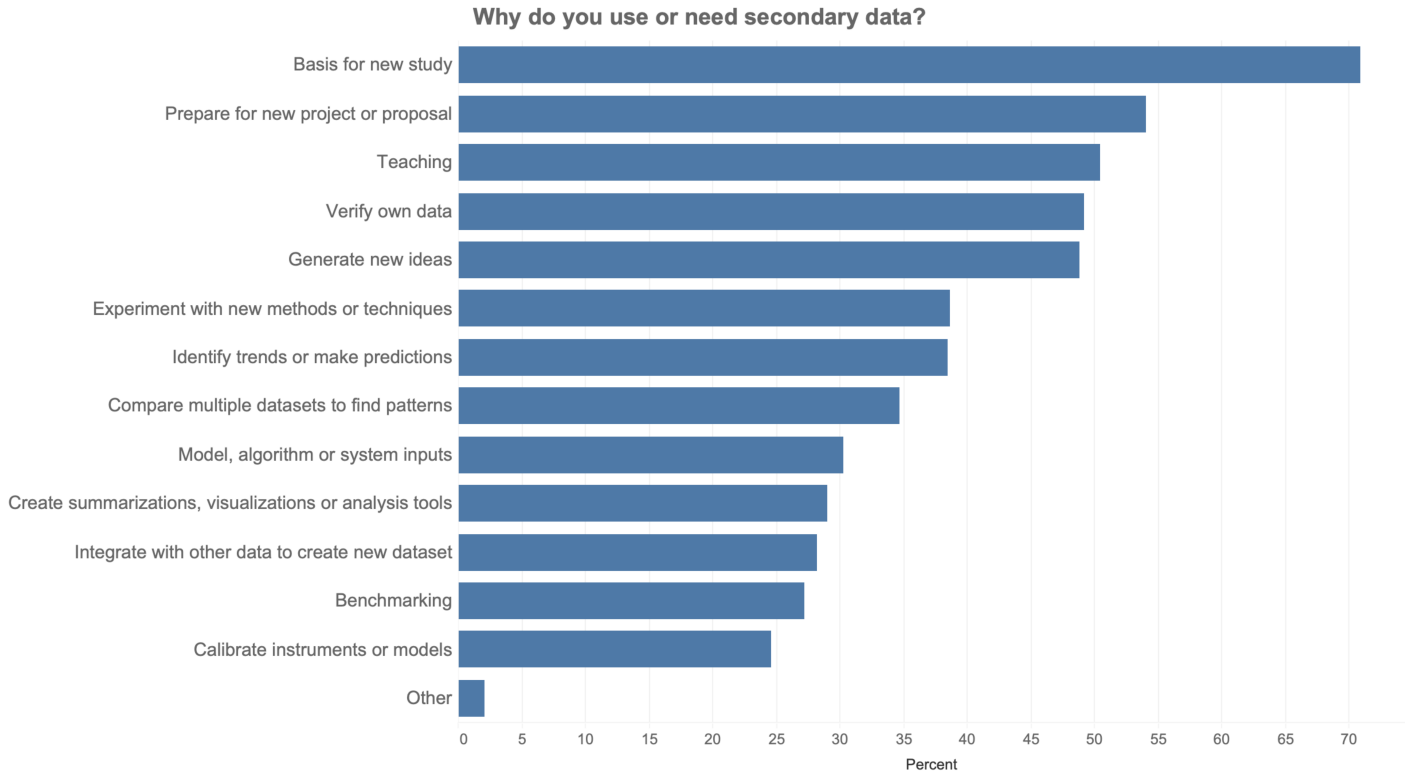
→ Data which are constantly collected, evolving

→ Example: High resolution climate change scenarios

https://www.rd-alliance.org/dynamic-data-citation-frequently-modifying-high-resolution-climate-data

Challenges for data citation

Not all uses may be cited.



Multiple responses possible. Percents are percent of respondents (n=1677).

Gregory, K., Groth, P. Scharnhorst, A., Wyatt, S. (2020). Lost or found? Discovering data needed for research. Harvard Data Science Review. https://doi.org/10.1162/99608f92.e38165eb

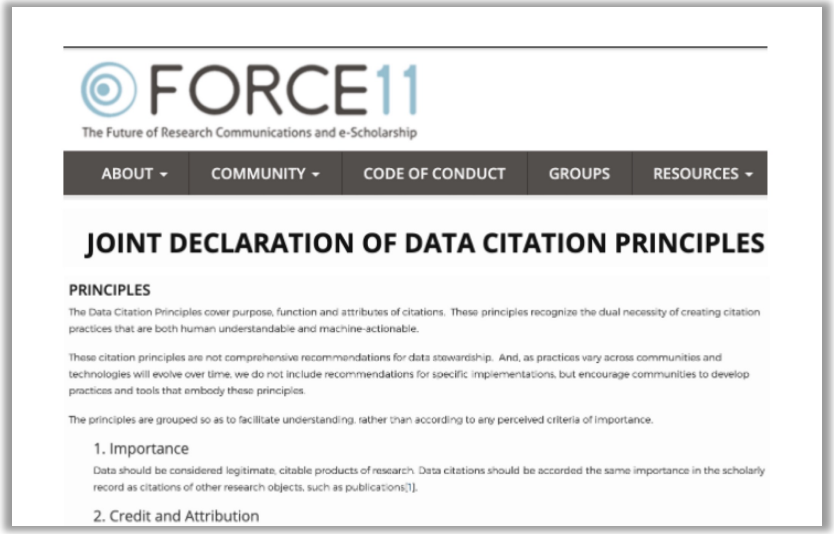
Bringing data “in sight”

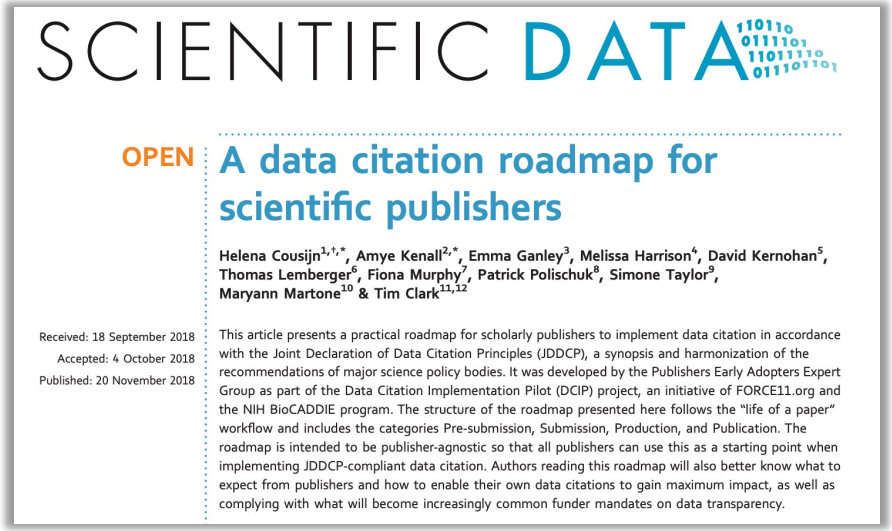
Facilitating data citation

Studying data citation practices

Facilitating data citation

Many principles, roadmaps, projects, working groups





Bibliometric data sources

Data Citation Index and (increasingly) DataCite

Data Citation Index

→ Clarivate (2012)

→ 380+ data repositories indexed

→ Selection criteria: subject, repository attributes, geographic scope

→ Document types: repository, data study, data set, software

→ WoS subject categories / research areas

https://clarivate.com/webofsciencegroup/solutions/webofscience-data-citation-index/

DataCite

→ Non-profit organization (2009)

→ Assigns DOIs to data, etc.

→ DataCite members input metadata according to DataCite schema

→ Or DataCite collects data externally

→ 6 required metadata fields

→ Subject not required

https://datacite.org/

Studying data citation practices

Lack of permanence, standardization, citations

→ Possibly impermanent citations: URLs, low incidence of DOIs, link rot
(Yoon et al., 2019; Peters et al., 2016; Pepe et al., 2014)

→ Informal references in text or acknowledgements
(Park et al., 2018)

→ Majority of data remain uncited; certain data types cited more (DCI)
(Robinson-Garcia et al., 2016; Peters et al., 2016)

Studying data citation practices

Disciplinary practices

→ Hard sciences (biomedical in particular) account for ~80% of records in Data Citation Index

(Torres-Salinas et al., 2014)

→ Disciplinary distribution of data citations varies by document type (DCI)
(Robinson-Garcia et al., 2016)

→ Tendency to use URLs rather than DOIs in natural & life sciences (DCI)

(Peters et al., 2016)

Studying data citation practices

Gaps and opportunities

→DataCite as a potential source for bibliometric research

→But…metadata incompleteness, lack of standardization, ambiguous definitions (Robinson-Garcia et al., 2017)

→Knowledge of the 'why' behind data citation practices (Mayernik, 2012)

→Opportunity to critically examine assumptions about data citation (Parsons & Fox, 2013; Borgman, 2016)

Bringing data *insight*

Meaningful Data Counts

Meaningful Data Counts Project

MDC research project – Bibliometricians + infrastructure

→Generating empirical evidence

→Data sharing and citation patterns

→ Bibliometric analysis

→ Theoretical approach

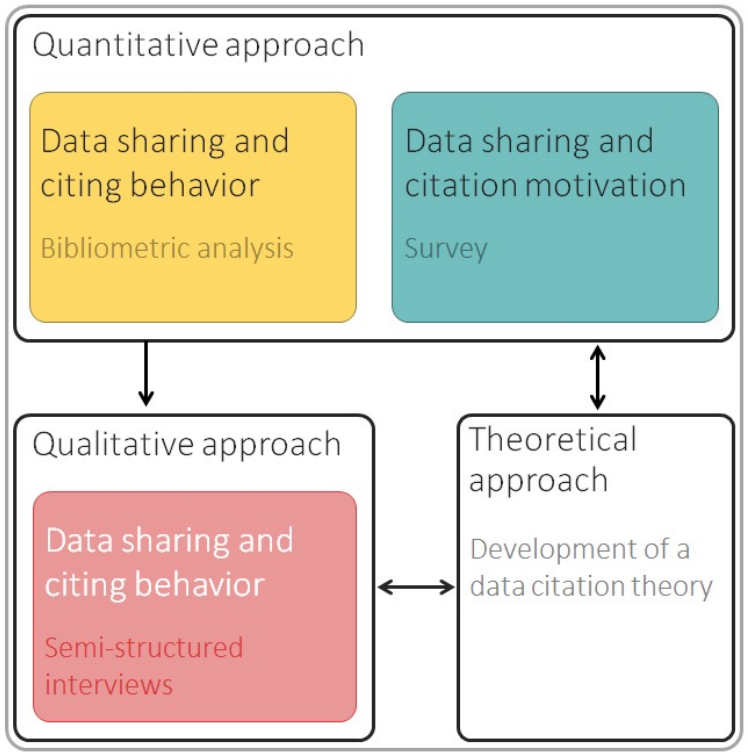
→Motivations to (not) share, use and cite data

→ Survey

→ Semi-structured interviews

→ Theoretical approach

→Creating infrastructure for bibliometricians



Morissette, E., Peters, I., & Haustein, S. (2020). Research data and the academic reward system. Poster presented at *OASPA Annual Conference*, 23-24 September 2020, online. https://doi.org/10.5281/zenodo.4034585

More questions than answers

Bringing insight takes work

→ Disciplining data
- →How are disciplinary fields assigned to data?
- →What do they represent – context of creation, aboutness, context of reuse?
- →How can metadata be enriched?

→ Data communities
- →Going beyond disciplines

→ Responsible evaluation of data practices
- →Situated within communities (Mustajoki et al; 2021)

Thank you!

https://zenodo.org/communities/meaningfuldatacounts

Kathleen Gregory | @gregory_km

kathleen.gregory@uottawa.ca

Thanks to the MDC project team:

Anton Ninkov, Stefanie Haustein, Isabella Peters

References

- Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.
- Borgman, C. L. (2016). Data citation as a bibliometric oxymoron. In C. R. Sugimoto (Ed.), Theories of informetrics and scholarly communication (pp. 93–116). De Gruyter. https://doi.org/10.1515/9783110308464-008
- CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*, *12*(0), CIDCR1–CIDCR75. https://doi.org/10.2481/dsj.OSOM13-043
- Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk
- Gregory, K., Groth, P. Scharnhorst, A., Wyatt, S. (2020). Lost or found? Discovering data needed for research. Harvard Data Science Review. https://doi.org/10.1162/99608f92.e38165eb
- Mayernik, M. S. (2012). Data citation initiatives and issues. *Bulletin of the American Society for Information Science and Technology*,
- *38*(5), 23–28. https://doi.org/10.1002/bult.2012.1720380508
- Morissette, E., Peters, I., & Haustein, S. (2020). Research data and the academic reward system. Poster presented at *OASPA Annual*
- *Conference*, 23-24 September 2020, online. https://doi.org/10.5281/zenodo.4034585
- Mustajoki, H., Pölönen, J., **Gregory, K.**, Ivanović, D., Brasse, V., Kesäniemi, J., Koivisto, E., & Pylvänäinen, E. (2021). *Making FAIReR assessments possible. Final report of EOSC Co-Creation projects: "European overview of career merit systems'' and "Vision for research data in research careers."* https://doi.org/10.5281/zenodo.4701375
- Park, H., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. Journal of the Association for Information Science and Technology, 69(11), 1346-1354

References

- Parsons, M., & Fox, P. (2013). Is Data Publication the Right Metaphor? *Data Science Journal*, *12*(0), WDS32–WDS46. https://doi.org/10.2481/dsj.WDS-042
- Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: An extended analysis of citations and altmetrics. *Scientometrics*, *107*(2), 723–744. https://doi.org/10.1007/s11192-016-1887-4
- Pepe, A., Goodman, A., Muench, A., Crosas, M., & Erdmann, C. (2014). How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers. *PLoS ONE*, *9*(8), e104798. https://doi.org/10.1371/journal.pone.0104798
- Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, *67*(12), 2964–2975. https://doi.org/10.1002/asi.23529
- Robinson-Garcia, N., Mongeon, P., Jeng, W., & Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. Journal of Informetrics, 11(3), 841–854. https://doi.org/10.1016/j.joi.2017.07.003
- Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6–20. https://doi.org/10.1002/asi.23917
- Torres-Salinas, D., Martín-Martín, A., & Fuente-Gutiérrez, E. (2014). Analysis of the coverage of the Data Citation Index – Thomson Reuters: Disciplines, document types and repositories. *Revista Española de Documentación Científica*, *37*(1), e036. https://doi.org/10.3989/redc.2014.1.1114
- Yoon, J., Chung, E., Lee, J. Y., & Kim, J. (2019). How research data is cited in scholarly literature: A case study of HINTS: How HINTS data is cited in scholarly literature. Learned Publishing, 32(3), 199–206. https://doi.org/10.1002/leap.1213