Accessing, analyzing and visualizing research data metadata using DataCite and Jupyter Notebooks

DH Toolbox 16.02.2022

Anton Ninkov

Email: aninkov@uottawa.ca

Scholarly Communications Lab

École des sciences de l'information | School of Information Studies

University of Ottawa/Universitè d'Ottawa

uOttawa

L'Université canadienne
Canada's university

Outline

→ Part 1 – Meaningful Data Counts

→ Part 2 – DataCite

→ Part 3 – GraphQL API & Jupyter Notebook

→ Part 4 – Subject Classification Mapping

→ Part 5 - Survey

Part 1 - Meaningful Data Counts

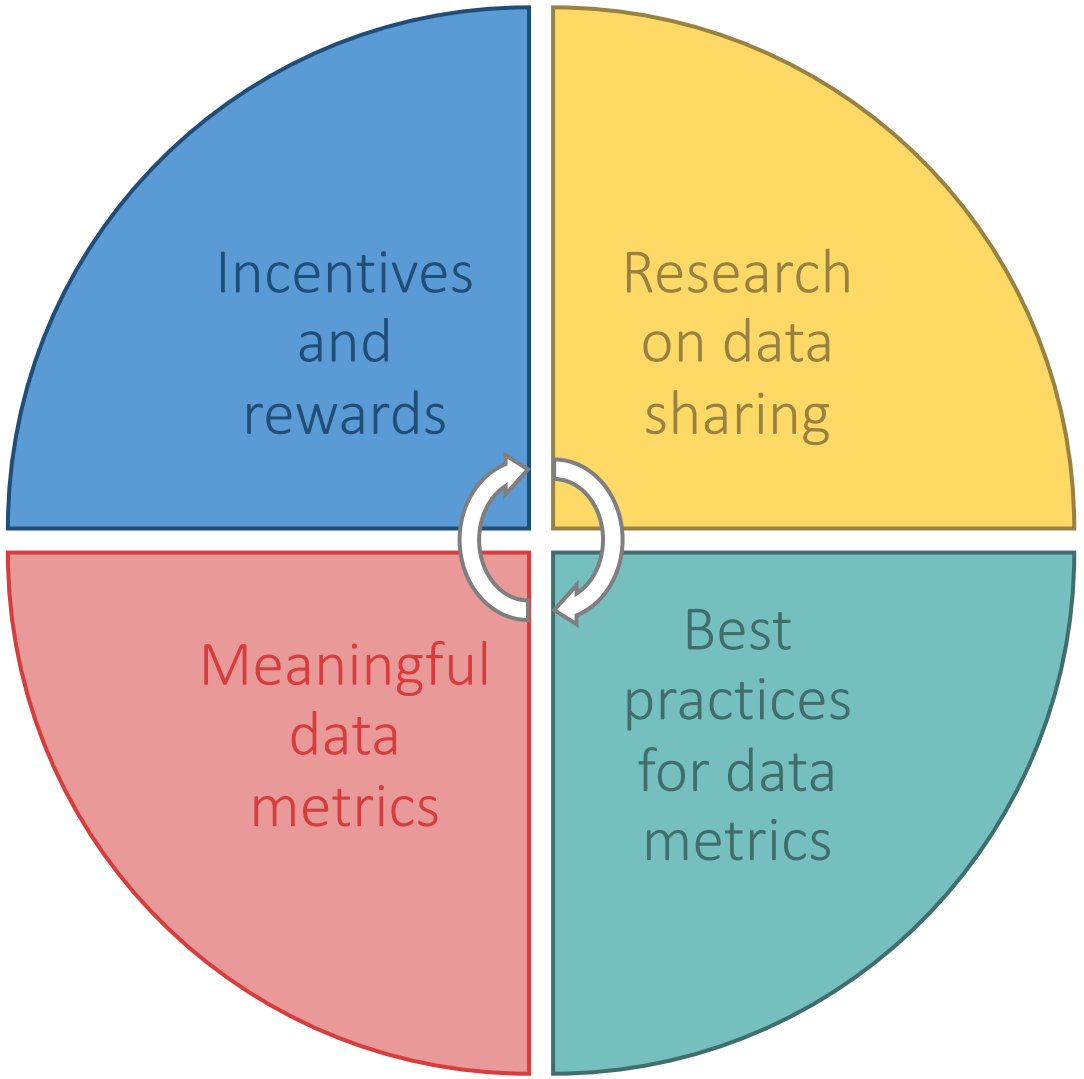
Scholarly Communications Lab Team

1. Dr. Stefanie Haustein
   → Principle Investigator, Associate Professor, University of Ottawa
2. Dr. Isabelle Peters
   → Principle Investigator, Professor, Kiel University and ZBW Leibniz-Information Center for Economics
3. Dr. Kathleen Gregory
   → Postdoctoral Fellow, University of Ottawa
4. Dr. Anton Ninkov
   → Postdoctoral Fellow, University of Ottawa

Part 1 - Meaningful Data Counts

Lack of data sharing and citations

1. Researchers do not share and cite datasets due to a lack of incentives and rewards in academia
2. Bibliometricians do not study research data as scholarly outputs because of a lack of evidence of data reuse and citations
3. Best practices for bibliometric studies on research data have not yet been developed, as use cases are missing
4. Meaningful data metrics are not developed and not available to incentivize open data practices

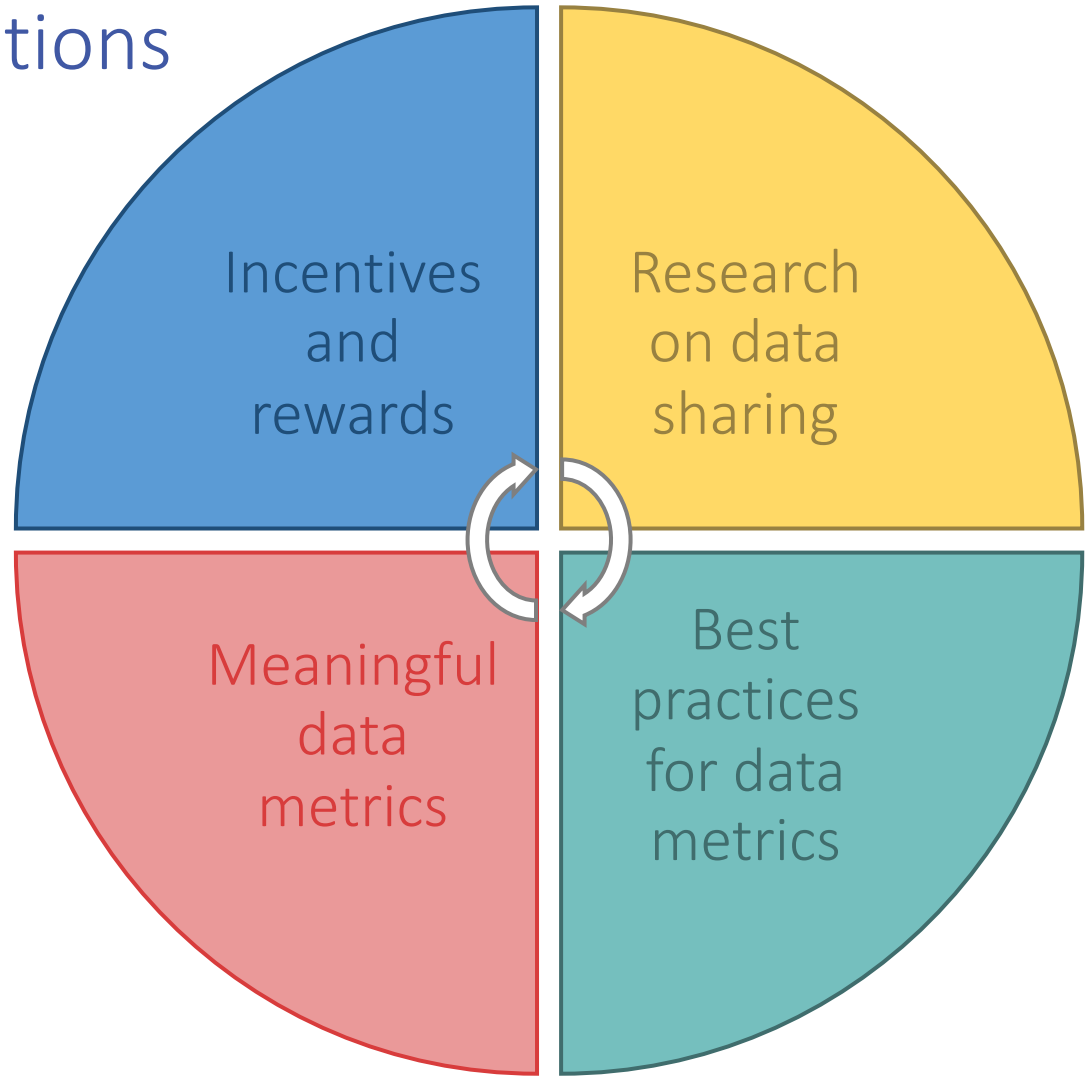


Morissette, E., Peters, I., & Haustein, S. (2020). Research data and the academic reward system. Poster presented at *OASPA Annual Conference*, 23-24 September 2020, online. https://doi.org/10.5281/zenodo.4034585

Part 1 - Meaningful Data Counts

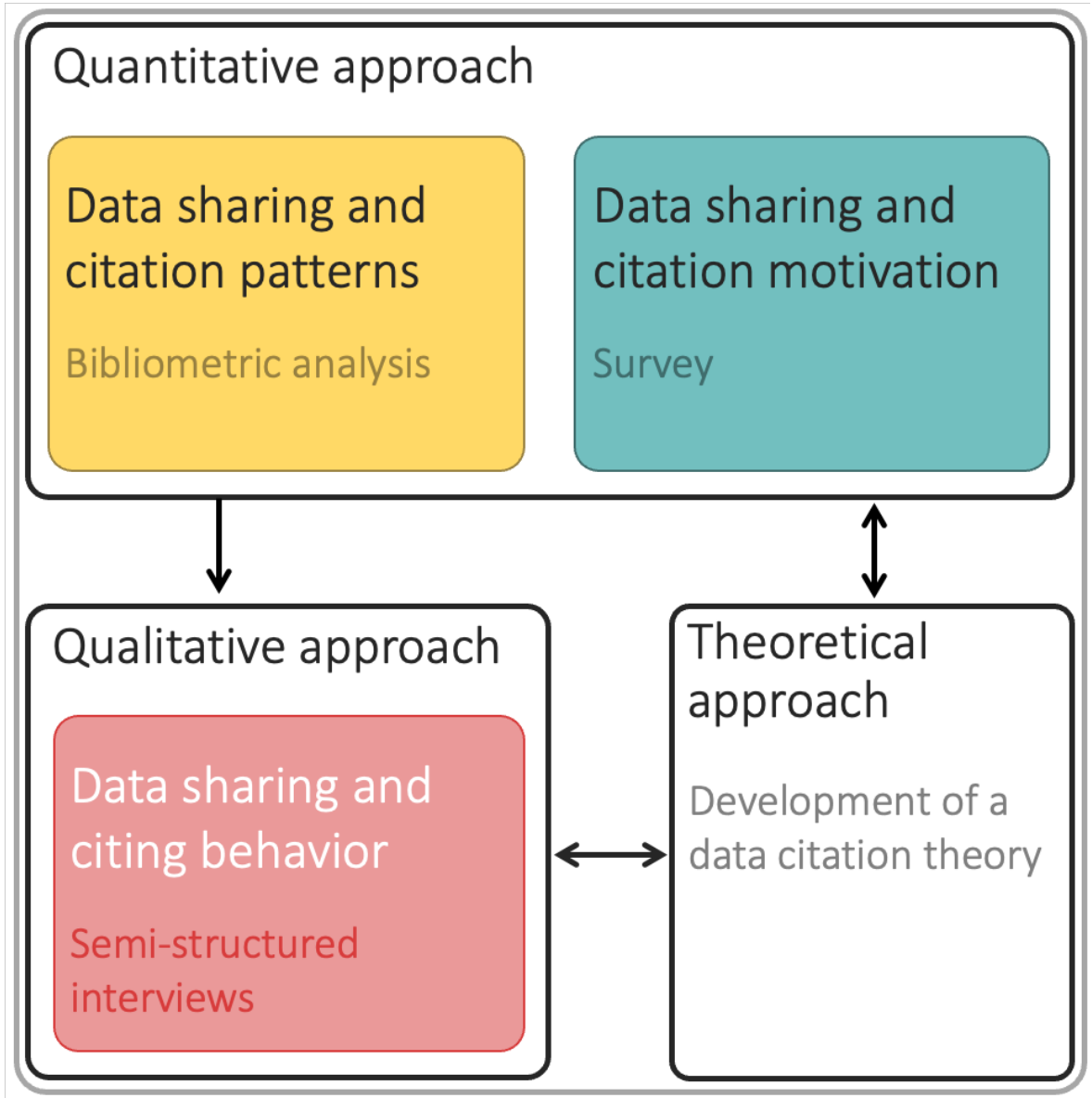
Incentivizing data sharing and citations

1. Researchers share and cite datasets due to incentives and rewards in academia
2. Bibliometricians study research data as scholarly outputs based on evidence of data reuse and citations
3. Best practices for bibliometric studies on research data are being developed, as use cases are shared
4. Meaningful data metrics are developed and available to incentivize open data practices



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

Part 1 - Meaningful Data Counts



Part 1 - Meaningful Data Counts

Data citation

→ Data citation and calls for standardization are not new matters of concern, milestones in standards development include:

→ Bermuda Principles (1996), CrossRef (1999), DataCite (2009)

→ Some work examines the state of data citation, highlighting:

→ Impermanent and untrackable nature

→ Data references often included only in the body of articles or acknowledgements rather than reference lists

→ Disciplinary differences are a recognized but unsolved problem

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.
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

Part 1 - Meaningful Data Counts

Need for disciplinary benchmarks

→Data sharing and reuse vary by discipline

→Determining disciplines of datasets

- →Intellectually (by user or archival staff)
- →Automated (based on dataset metadata)
- →By proxy (based on external metadata)

Borgman, C. L. (2015). Big data, little data, no data: Scholarship in the networked world.
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
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
Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLOS ONE, 10(8), e0134826. https://doi.org/10.1371/journal.pone.0134826
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

Part 2 - DataCite

→ DataCite is a non-profit organisation that provides persistent identifiers (DOIs specifically) for research data and other research outputs.

→ DOI names are provided by DataCite and other DOI registration agencies, coordinated by the International DOI Foundation (IDF).

→ DOI name can be assigned to any entity – physical, digital or abstract

Part 2 - DataCite

→DataCite members input data or DataCite collects it externally

Field	Description
Identifier	The Identifier is a unique string that identifies a resource.
Creator	The main researchers involved in producing the data, or the authors of the publication, in priority order.
Title	A name or title by which a resource is known.
Publisher	The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource.
Publication Year	The year when the data was or will be made publicly available.
Resource Type	A description of the resource.

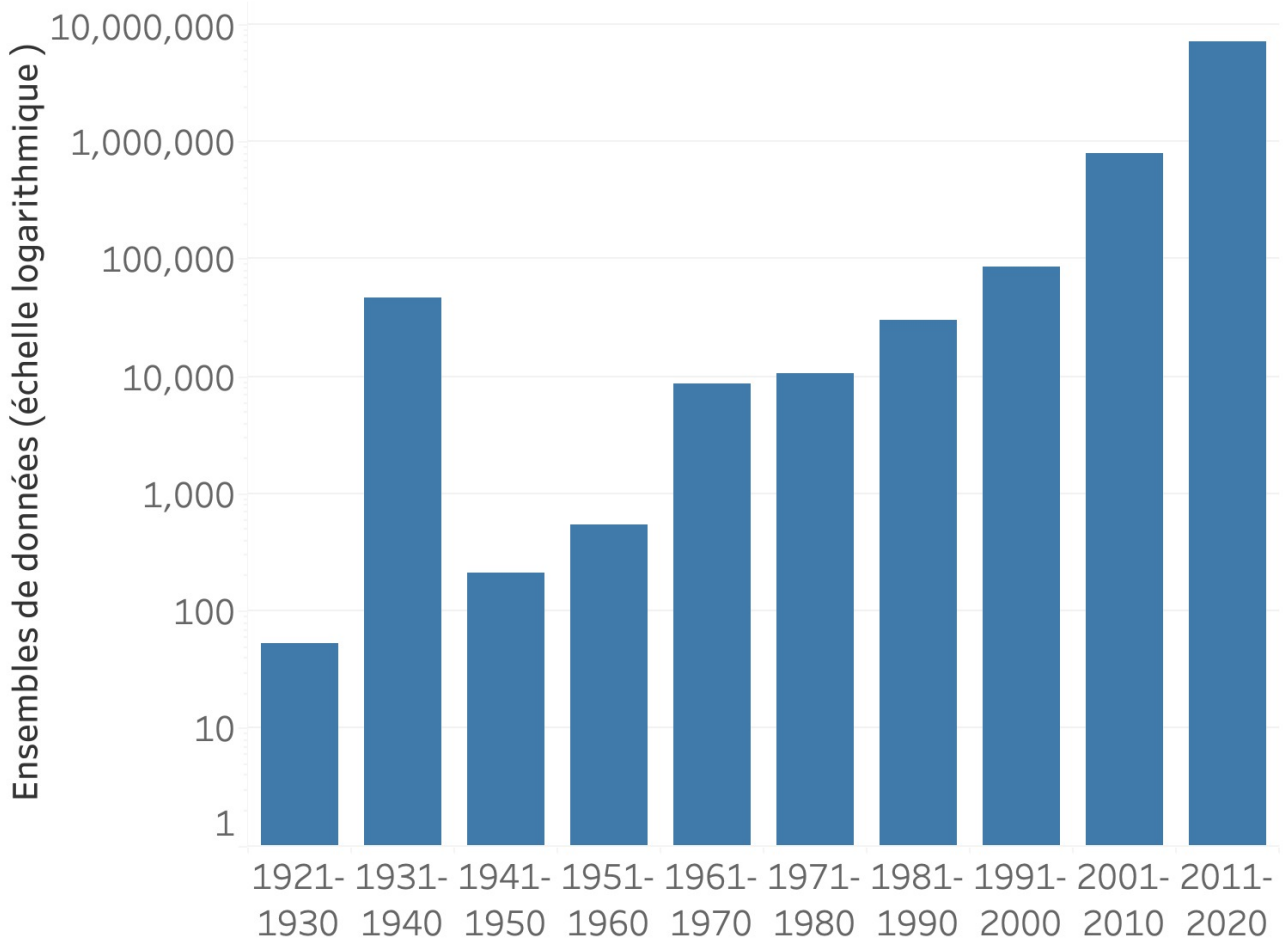
Part 2 - DataCite

Optional Fields

Field	Description
Subjects	Subject, keywords, classification codes, or key phrases describing the resource.
Contributors	The institution or person responsible for collecting, creating, or otherwise contributing to the development of the dataset.
Language	Primary language of the resource. Allowed values are taken from IETF BCP 47, ISO 639-1 language codes.
Funding	Information about financial support (funding) for the resource being registered.
relatedItems	Information about a resource related to the one being registered e.g. a journal or book of which the article or chapter is part.
Version	Version number of the resource. If the primary resource has changed the version number increases.

Part 2 - DataCite

Growth of Datasets on DataCite



Part 3 – GraphQL API & Jupyter Notebook

What is an API?

https://www.youtube.com/watch?v=s7wmiS2mSXY

Part 3 – GraphQL API & Jupyter Notebook

DataCite GraphQL API

→ GraphQL is a query language specifically for making queries across a graph.

→ With a “normal” REST API, you have to provide the query in a specified way in order to get the results the API provider wants to give you.

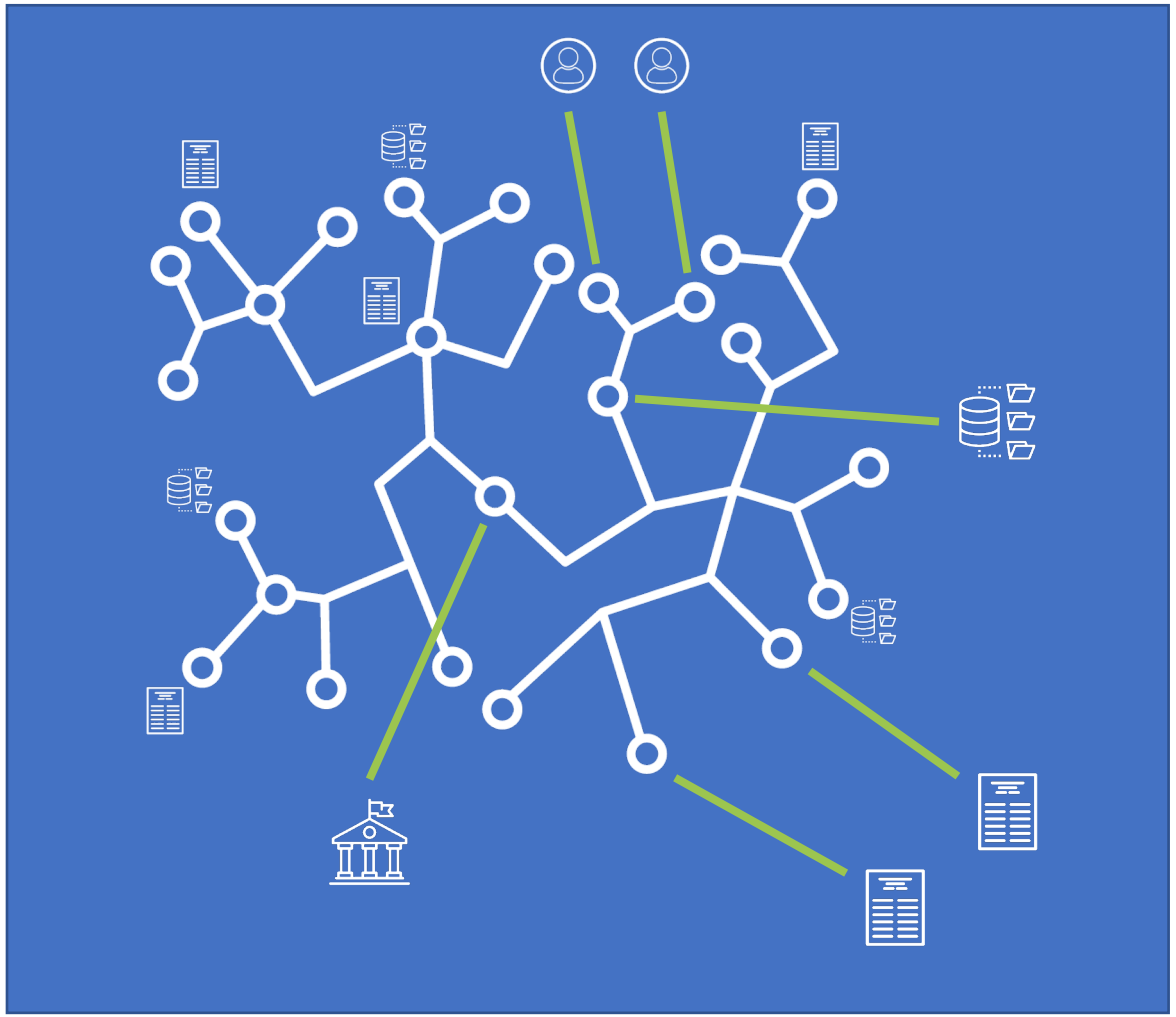
→ With GraphQL, you can better specify the information you want to receive.

Part 3 – GraphQL API & Jupyter Notebook

DataCite GraphQL API

Part 3 – GraphQL API & Jupyter Notebook

DataCite GraphQL API



Demonstration of GraphQL API

https://api.datacite.org/graphql

Part 3 – GraphQL API & Jupyter Notebook

Jupyter Notebook

→ Tool to develop open-source software, open-standards, and services for interactive computing across many programming languages.

→ Formerly iPython Notebooks

→ Programming languages: Julia, Python, and R

→ Creators can make code more easily understood and used

Part 3 – GraphQL API & Jupyter Notebook

Jupyter Notebook for Datasets on DataCite

→ Developed for ISSI2021 Conference

→ A tool to quickly see the metadata of all datasets on DataCite

→ Pulls data on command

Demonstration of Jupyter Notebook(s)

Part 3 – GraphQL API & Jupyter Notebook

Improving discipline information

→More discipline data

→ Revised OECD Fields of Science and Technology (FOS 2007)

→6 main fields, 40 subfields

→More metadata

→Enriching discipline metadata for disciplinary repositories

→Encourage repositories to make discipline information mandatory

→Explore automated approaches based on:

→Available dataset metadata

→External information

https://www.oecd.org/science/inno/38235147.pdf

Part 4 - Subject Classification Mapping

Case for Mapping

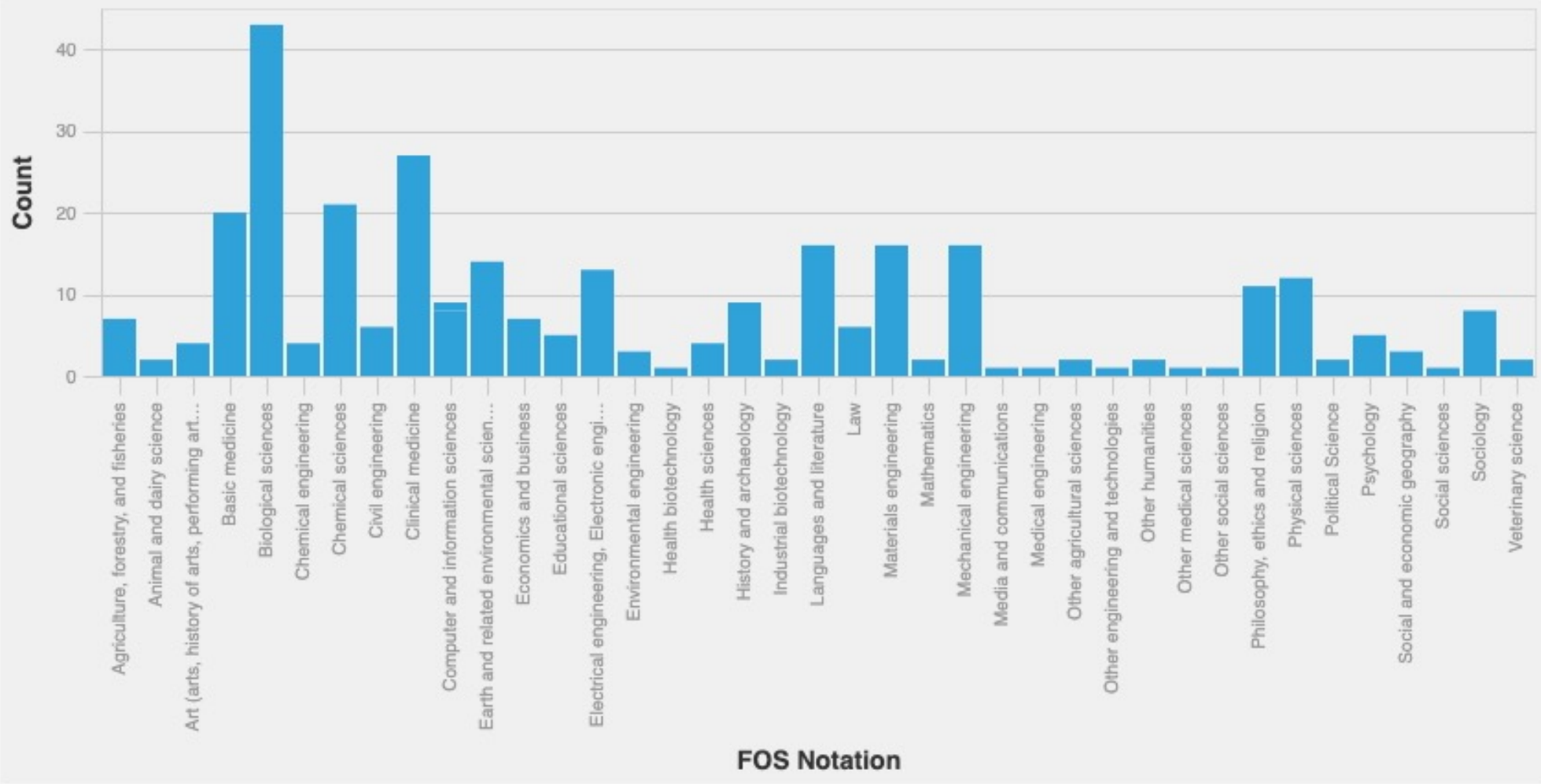
→Need more datasets on DataCite to have an OECD subject listed

→More repositories have subject classification listed using other systems

→Can we map systems to OECD?

Part 4 - Subject Classification Mapping

DFG -> OECD



 https://github.com/datacite/frascati-mappings

Part 4 - Subject Classification Mapping

Challenges

→DFG 10602 Asian Studies

→OECD 6.5 Other Humanities

→OECD 5.9 Other Sociology

→DFG 204 Microbiology, Virology and Immunology

→OECD 1.6 Biological Sciences

→OECD 3.1 Basic Medicine

Part 4 - Subject Classification Mapping

Challenges

307	Condensed Matter Physics	1.3	Physical sciences
30701	Experimental Condensed Matter Physics	1.3	Physical sciences
30702	Theoretical Condensed Matter Physics	1.3	Physical sciences
308	Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas	1.3	Physical sciences
30801	Optics, Quantum Optics, Atoms, Molecules, Plasmas	1.3	Physical sciences
309	Particles, Nuclei and Fields	1.3	Physical sciences
30901	Particles, Nuclei and Fields	1.3	Physical sciences
310	Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics	1.3	Physical sciences
31001	Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics	1.3	Physical sciences
311	Astrophysics and Astronomy	1.3	Physical sciences
31101	Astrophysics and Astronomy	1.3	Physical sciences

Part 5 - Survey

Overview

→Stratified sample by OECD classification

→6 millions authors with an email

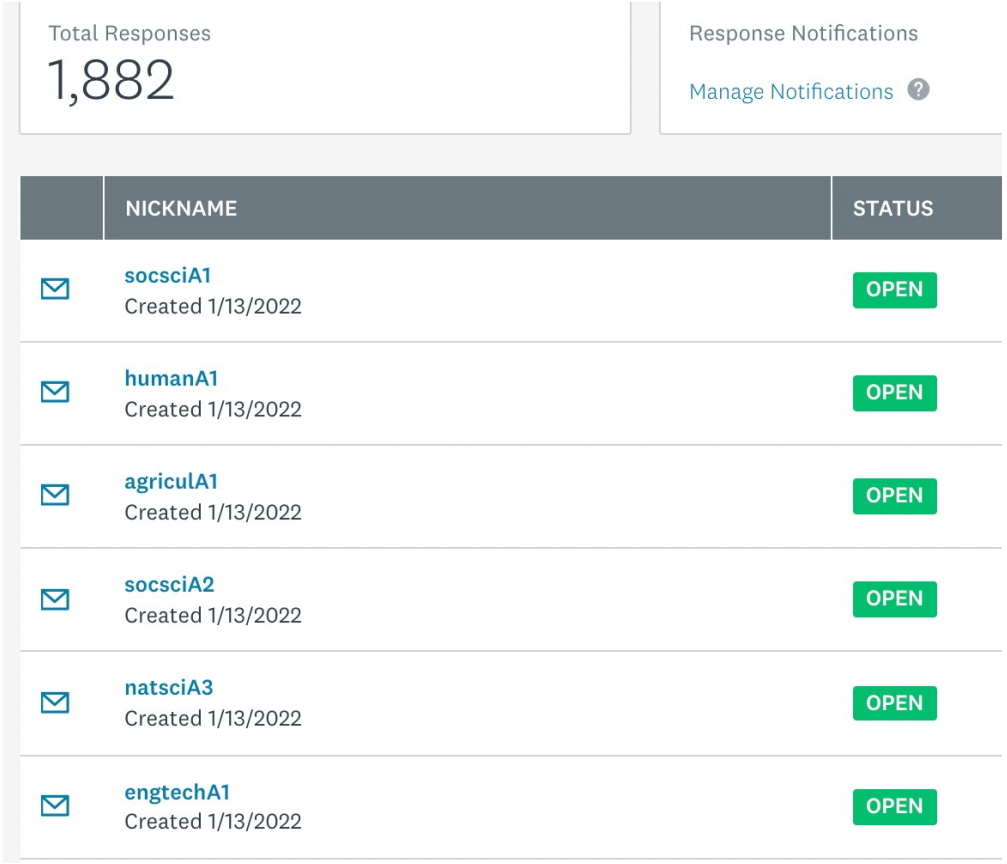
→150,000 authors contacted

→3 components

→Part 1: Reusing and citing data

→Part 2: Rewarding data management

→Part 3: Demographics



Part 5 - Survey

Part 1: Reusing and Citing Data

→ Have you ever reused data which other people have created, for any purpose?

→ How would you classify the type of data that you reuse?

→ How frequently do you reuse secondary data for the following purposes?

Part 5 - Survey

Part 2: Rewarding data management

→Have you ever shared your own research data?

→How important would it be for you to know the following information about when other people reuse your data?

→The number of citations your data have received

→Information about who has used your data

Part 5 - Survey

Part 3: Demographics

→How many years of professional experience do you have in your field?

→With which disciplinary domain do you most identify ?

→Which approach best describes the research methods which you use?

Conclusion

→ Meaningful Data Counts project is on going

→ DataCite has metadata available that is useful for our project

→ GraphQL API + Jupyter Notebook can make the data more accessible

→ Still subject classification problems that require mapping effort

→ Survey results will help explain data reuse behaviors

Thank you!

https://zenodo.org/communities/meaningfuldatacounts

Contact Anton at:

Email: aninkov@uottawa.ca

ORCID: 0000-0002-8276-7656

Twitter: @TheNinkov

GitHub: antonninkov