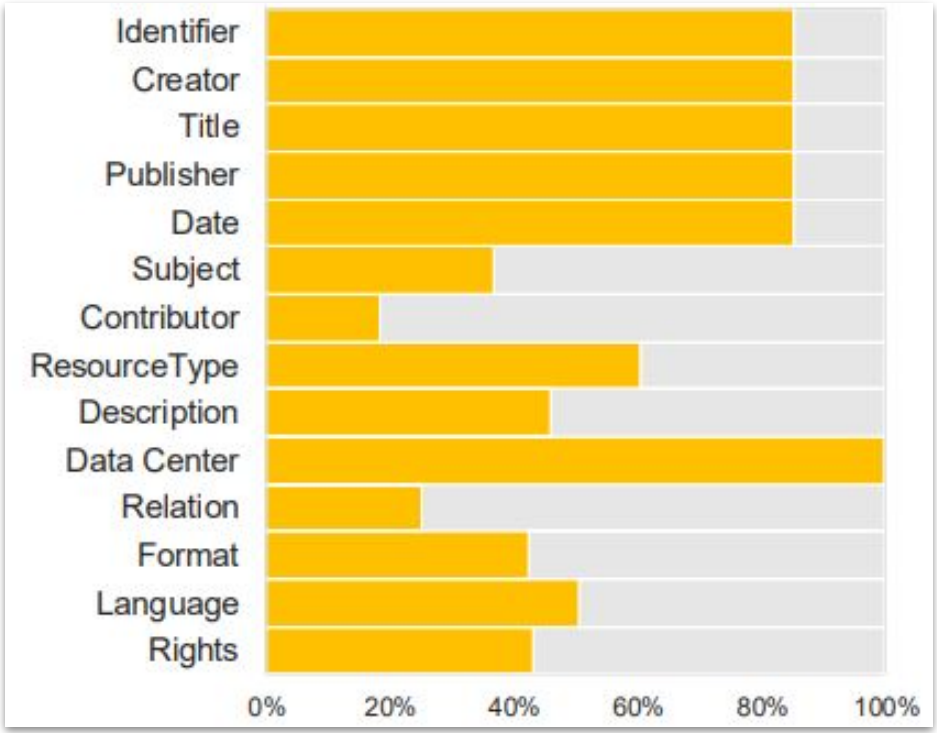
Metadata completeness and subject information

DataCite Open Hours
2022-05-18

Dorothea Strecker

Metadata completeness in DataCite (by element)

In 2017:



Availability of metadata elements for DataCite metadata records
(Robinson-Garcia et al., 2017)

Metadata completeness of subject information

Today ...

- ➔ only about 6 % of metadata records **of datasets** in DataCite included subject information (Ninkov et al., 2021)
- ➔ most notations (FOS) were from the natural sciences
- ➔ subjectScheme is rarely specified - standardisation issues remain (Habermann, 2020)

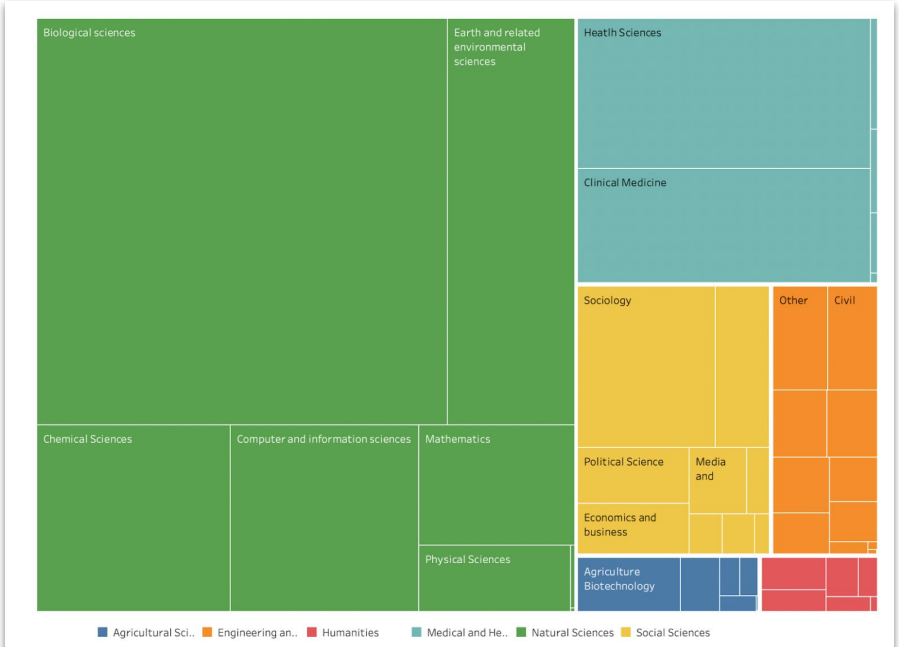


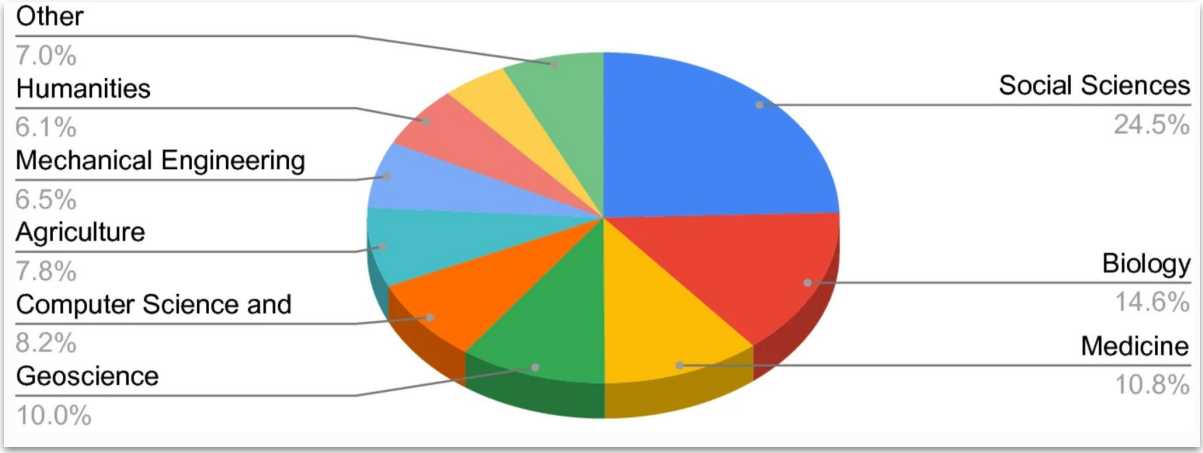
Figure 2. Data with metadata describing disciplinary domain, according to the OECD Fields of Science classification, retrieved on July 9th, 2021. For an interactive version of this tree map, with the most current data, please see our Jupyter Notebook [3]

Source: Make Data Count

Why subject information?

Subject information places data in a disciplinary context …

- ➔ facilitating data journeys
- ➔ improving research data infrastructures and services
- ➔ enabling scientometric studies



Discipline distribution of datasets displayed in Google Data Search results over 2 weeks in May 2020 (Benjelloun et al., 2020)

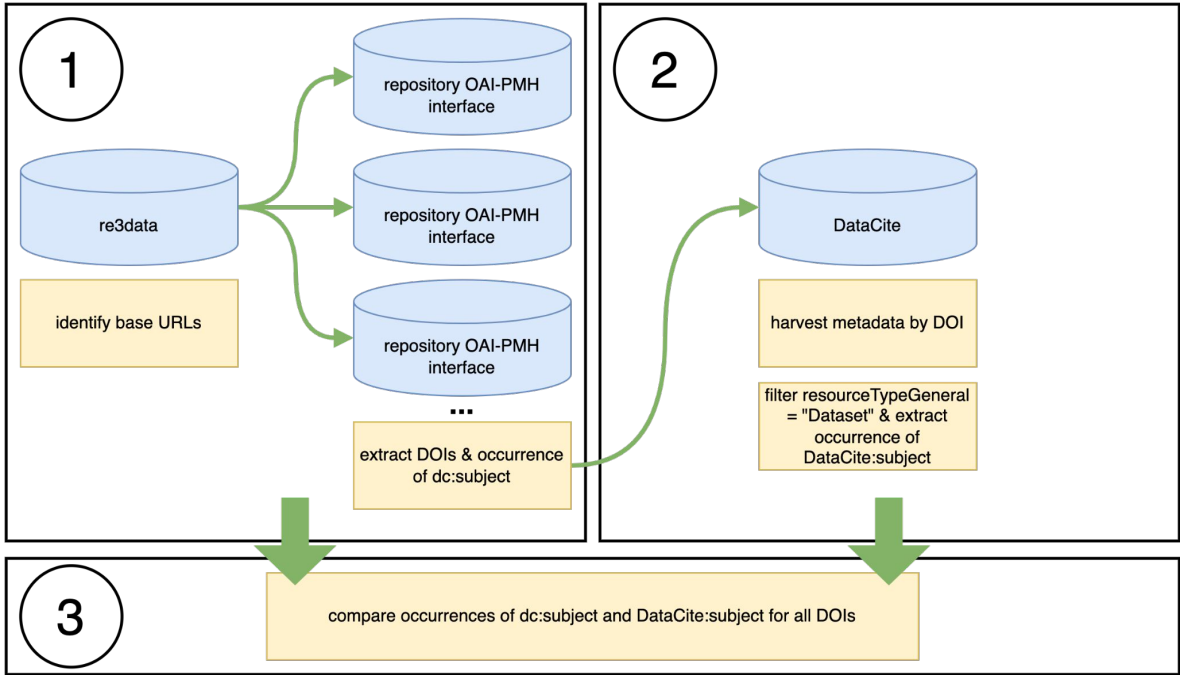
Improving the availability of subject information in DataCite?

- **Classify metadata records with supervised machine learning** based on title, description and keywords (Weber et al., 2020)

- **Using existing subject information from other data sources:**
  a. Using alternative versions of metadata records: metadata records at repositories
  b. Using "container metadata": re3data metadata

Metadata records at repositories

Do repositories deliver all subject information to DataCite?

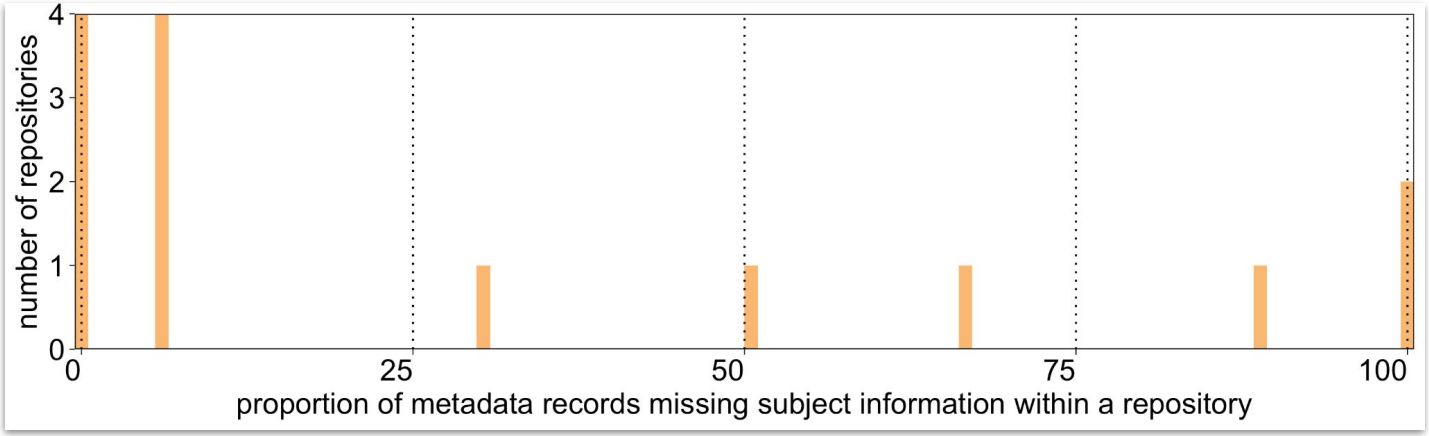


8.499 DOIs from 14 repositories

Metadata records at repositories

In this limited sample, …

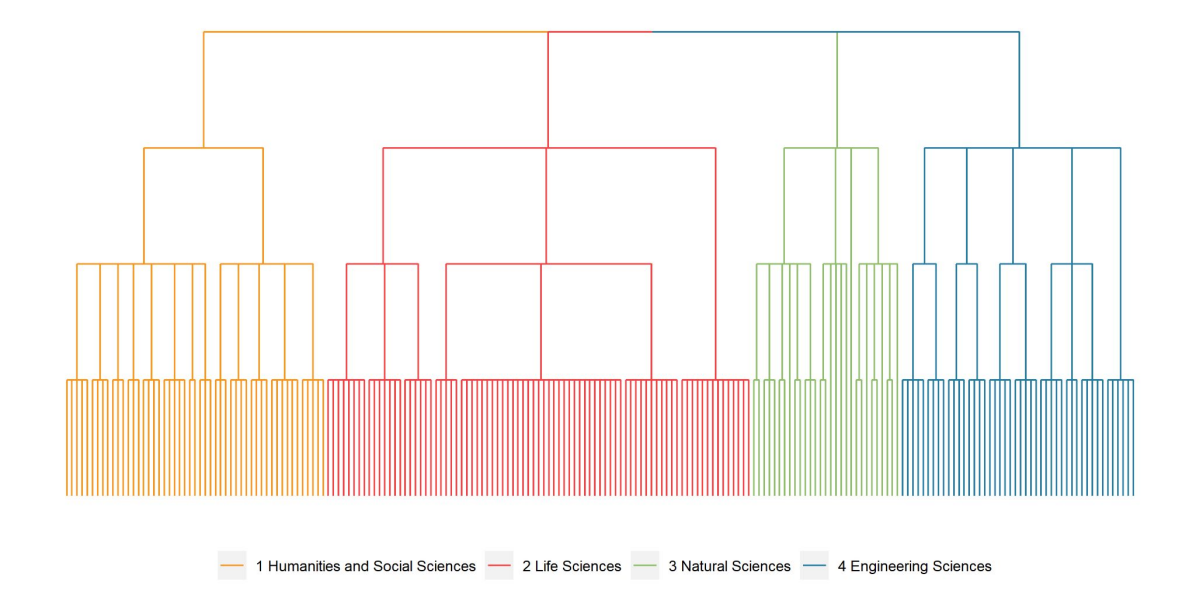
- ➔ only **9.08 %** of metadata records appear to be missing subject information in DataCite that is available via the OAI-PMH interface
- ➔ many repositories provide subject information for (almost) all of their metadata records, but some don't at all



re3data metadata

Could re3data subject information be used to enrich DataCite metadata?

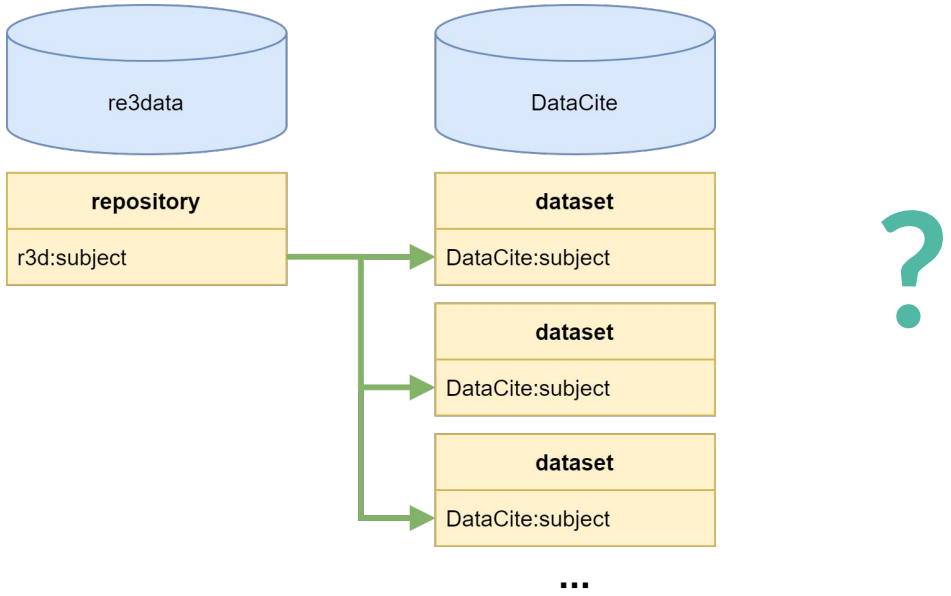
Blog post providing a detailed analysis of subject information in re3data



Subject classification used in re3data: DFG Subject Areas (2014)

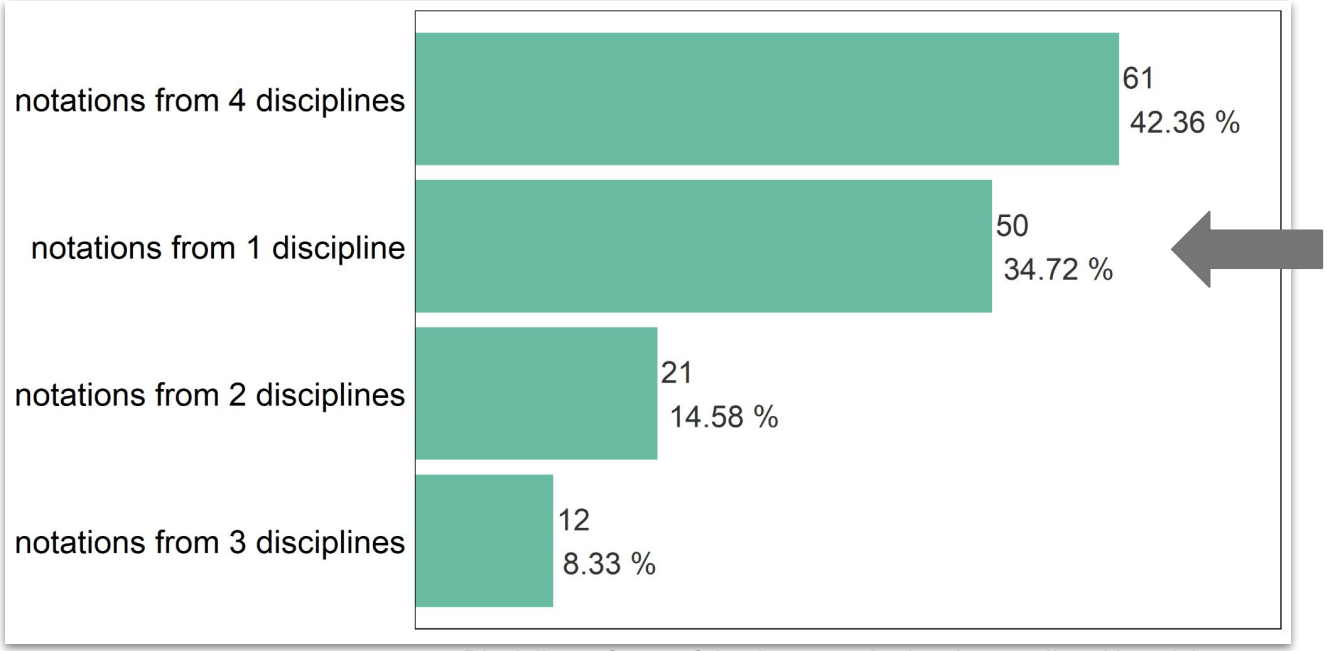
re3data metadata

Could re3data subject information be used to enrich DataCite metadata?



re3data metadata

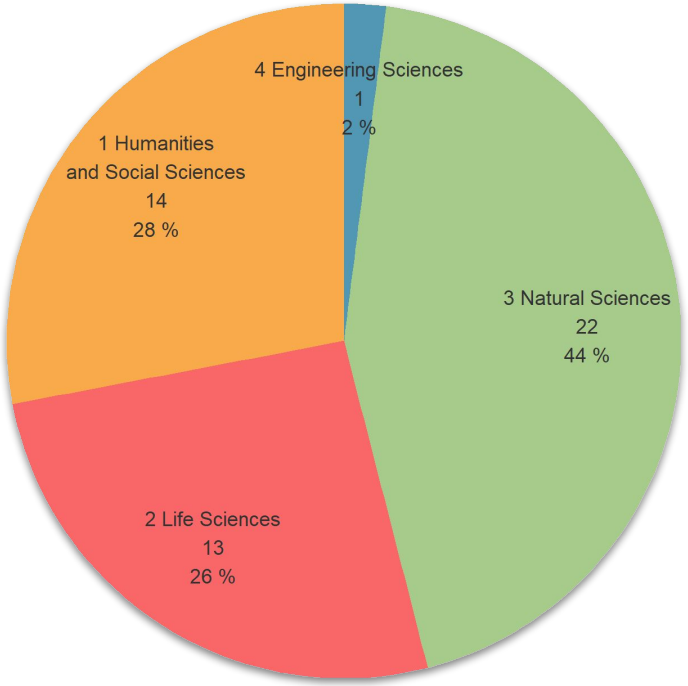
Can re3data subject information be used to enrich DataCite metadata?



Disciplinary focus of the 144 repositories that are listed in re3data and linked to a DataCite Client ID

re3data metadata

Subject distribution of these 50 repositories with a narrow disciplinary focus (highest level of abstraction)



Conclusion

- ➔ MakeDataCount / DataCite are working on improving the availability of subject information in DataCite.

- ➔ Both approaches explored here for using existing metadata to enrich DataCite subject information are limited, but could yield results.

- ➔ There is some indication that many repositories submit (almost) all subject information they have recorded to DataCite, but some don't submit subject information to DataCite at all.

- ➔ **What could help repositories with submitting the subject information they have?**

Thank you!

Contact:
dorothea.strecker@hu-berlin.de
@dorothearrr

References

Benjelloun, O., Chen, S., & Noy, N. (2020). Google Dataset Search by the Numbers. In J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, & L. Kagal (Eds.), The Semantic Web – ISWC 2020 (pp. 667–682). Springer. https://doi.org/10.1007/978-3-030-62466-8_41

Habermann, T. (2020, July 13). DataCite Subject Metadata. Metadata Game Changers. https://metadatagamechangers.com/blog/2020/7/13/datacite-subject-metadata

Ninkov, A., Gregory, K., Peters, I., & Haustein, S. (2021). Datasets on DataCite—An Initial Bibliometric Investigation. International Conference on Scientometrics & Informetrics (ISSI 2021), Leuven, Belgium (Virtual). Zenodo. https://doi.org/10.5281/zenodo.4730857

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

Weber, T., Kranzlmüller, D., Fromm, M., & de Sousa, N. T. (2020). Using supervised learning to classify metadata of research data by field of study. Quantitative Science Studies, 1(2), 525–550. https://doi.org/10.1162/qss_a_00049