Royal Netherlands Academy of Arts and Sciences (KNAW)
KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN

Measuring research data archives

Scharnhorst, Andrea; Tykhonov, Vyacheslav; Indarto, Eko; Doorn, P.K.

2023

DOI (link to publisher)
10.5281/zenodo.10555758

document version
Publisher's PDF, also known as Version of record

document license
CC BY

Link to publication in KNAW Research Portal

citation for published version (APA)
Scharnhorst, A., Tykhonov, V., Indarto, E., & Doorn, P. K. (2023). *Measuring research data archives*. Paper presented at 86th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – 31, 2023 |London, United Kingdom, London, United Kingdom. https://doi.org/10.5281/zenodo.10555758

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the KNAW public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain.
• You may freely distribute the URL identifying the publication in the KNAW public portal.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
pure@knaw.nl

Download date: 12. Feb. 2025

Measuring research data archives

Scharnhorst, Andrea	DANS-KNAW, The Netherlands | andrea.scharnhorst@dans.knaw.nl
Tykhonov, Vyacheslav	DANS-KNAW, The Netherlands | vyacheslav.tykhonov@dans.knaw.nl
Indarto, Eko	DANS-KNAW, The Netherlands | eko.indarto@dans.knaw.nl
Doorn, Peter	DANS-KNAW, The Netherlands | peter.doorn@dans.knaw.nl

ABSTRACT

Data and data practices attract growing attention by those who measure scientific research. But, such as in scientometrics in general, many investigations look at data and data practices through the lens of output: for instance, in the form of formal and informal data citations (Gregory 2021). In this paper, we report about experiments and investigations done in a research data archive over a couple of years to gain insides in practices in and around research data archives. We discuss various methodologies, findings and reflect about opportunities and challenges when it comes to 'measure' how the data life cycle and the research life cycle interact. We are interested in raising awareness to the topic and to relate to similar initiatives in the field of quantitative studies of science.

KEYWORDS

research data management, research data repositories and archives, indicators, research evaluation

INTRODUCTION: Research Data as ore

Research Data has been labeled as the new oil or gold for research since more than a decade now. But currently, most of it still is as ore hidden somewhere, not yet fully brought to the daylight. Open science, reproducible science and FAIRness (Wilkinson et al, 2016) - all those science political paradigms - rely on making research data as **input** into research a better documented resource which then supposedly better can be shared, reused, and further exploited. As always in the history of scientometrics, with new ways of documentation come also the possibility to execute quantitative analysis with newly documented information. This is no other when it comes to data. However, compared with the century-long formalisation of scholarly communication, or the decade long further standardisation of citation indexing, data are still a relative new research asset. Despite of initiatives as Data Citation Indexes, data practices are still to a large extend informal, and not documented in a standardised way (Gregory et al., 2023).

Working at a research data archive, we came across various needs to 'measure the archive' expressed by various stakeholders. There is the inner working of the archive itself for which quantitative analysis might be useful. The data managers, the management team of the archive, and the funders of the archive's activities being the stakeholders of such an analysis. Quantitatively describing an archive also might benefit the users of an archive. Here, a quantitative fingerprint (maybe even visualised) might benefit the information retrieval of users, making them the stakeholder of such an analysis. Last but not least, data repositories are also in competition with each other. The digitisation and world-wide web created a multitude of ways to deposit data for further use. Specified research data search engines are an attempt to make data better visible in output but even more so as possible input material. By improving information retrieval across repositories also new ways to analyse them emerge. Data repositories also get certified[1] and have to document their processes - but there are no indicators (yet) coupled to such a certification process, and the certification process itself is so tailored and detailed in a way that it does not lend itself to a scaling up operation.

In this paper we report about two kinds of experiments we executed in the past years. First, we look at the whole landscape of research repositories using lenses of data portals and search engines. Second, we report about some technical possibilities to measure the collection (and its use) of one specific research data archive (or groups of them). Both experiments have been executed as part of various research projects. Conceptually, we also build on an analysis executed in the context of an ethnographic study (Borgman et al, 2015. We conclude with some lessons learned and potential future research directions.

(1) Research repositories - a landscape analysis

In the EOSC Synergy Project (2019-2022)[2] a Dashboard[3] has been developed to respond to the growing interest in

[1] https://www.coretrustseal.org

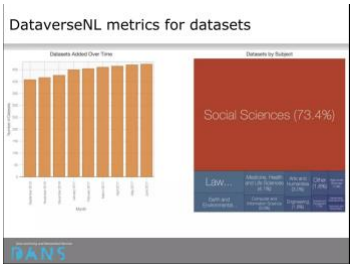
[2] https://cordis.europa.eu/project/id/857647

[3] https://bi-poc.dataverse.tk/superset/dashboard/6/?preselect_filters=%7B%7D&standalone=true&native_filters=%28%29 and Permalink: https://web.archive.org/web/20230710110219/https://bi-poc.dataverse.tk/superset/dashboard/6/?preselect_filters=%7B%7D&standalone=true&native_filters=%28%29

86th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – 31, 2023 |London, United Kingdom.

Author(s) retain copyright, but ASIS&T receives an exclusive publication license.

Open Science and to explore possible metrics on the landscape of data facilities and services (Doorn 2021). This web-based exploration focuses on the Dutch research data landscape, but has been informed and motivated by other (survey-based) European research data landscape analysis (Kruminas et al. 2022). For the *Dutch Open Science Dashboard* a limited number of quantitative indicators was selected. The Dashboard starts with an overview about research infrastructures as natural hosts of data services. It contains a disciplinary overview, compares outcomes from different sources, zooms into the use of so-called Persistent Identifiers (such as DOI) and contains more traditional quantitative indicators such as the number of datasets, and the content type of data deposited.



(2) Visual analytics of a research data archive – Miniverse of Dataverse instances

DANS as an institute of the Royal Netherlands Academy of Arts and Sciences hosts several Data Stations (research data repositories)[4] and a National platform for research institutions to channel research data archiving. All of those services run on the Dataverse platform[5] for which visual analytics based on the Apache Superset has been tested (Matela et al. 2021) With such an integration any of the metadata information as well as temporal developments of content of a Dataverse instance can be monitored.

CONCLUSION: Lessons learned and future work

Any quantitative analysis is only as good as the data collection on which it is run. The first example shows that there are platforms which can be harvested to compare data repositories. But, any automatic harvesting runs into problems when ingested data are not harmonised. The presented landscape analysis in the first example shows what might be possible but comparison of different sources also unveils where problems lie.

Means to enhance information retrieval are often tailored towards one repository with one or several designated communities - this means that metadata blocks are tailored towards domain needs and this makes the building of any union catalog or overview about various repositories' content challenging. But, the experiment in the second example also shows what in principle is possible when building web-interfaces based on API's and set up as micro services.

Still, public funded explorations are often project-based, hence not sustainable. On the other side: data search engines, data monitors, data citation indices are (also) developed by large information analytic companies (see Elsevier's Data Monitor[6]). Do we need a public concerted action to ensure that information to measure Open Science also stays open?

Maybe it is time for scientometrics to revive the area of input indicators: the number of repositories, their size, their staff are relevant to assess the availability of research data and their quality in different domains. This is a precondition to trace data citations later in formal scholarly communication. Could we gain an overview about what

[4] https://dans.knaw.nl/nl/ and Permalink: https://web.archive.org/web/20230604194031/https://dans.knaw.nl/nl/

[5] https://dataverse.harvard.edu/

[6] https://beta.elsevier.com/products/data-monitor?trial=true

is possible in the area of measuring research data practices based on scientometric expertise? Which level of granularity of indicators we need in this area? Do we need to go back to more high-level indicators (as the old STI), or do we indeed need more fine-grain measures? To which extent can semantic web technology and standardization in the description of research assets help here?

REFERENCES

Borgman, C. L., Van de Sompel, H., Scharnhorst, A., van den Berg, H., & Treloar, A. (2015). Who uses the digital data archive? An exploratory study of DANS. In Proceedings of the Association for Information Science and Technology (Vol. 52, pp. 1–4). http://doi.org/10.1002/pra2.2015.145052010096

Doorn, P. K. (2021). Dutch Open Science Dashboard 2020-2021. Website https://bi-poc.dataverse.tk/superset/dashboard/6/

Gregory, K. (2021). Findable and reusable? Data discovery practices in research. Maastricht University. https://doi.org/10.26481/dis.20210302kg

Gregory, K., Ninkov, A., Ripp, C., Roblin, E. Peters, I., Haustein, S. (2023). Tracing data: A survey investigating disciplinary differences in data citation. Zenodo. Preprint: https://doi.org/10.5281/zenodo.7555266

Krūminas, Pijus , Joy Davidson, Ingrid Dillo, Carmela Asero, Jonas Antanavičius, Peter Doorn, Aurinta Garbašauskaitė, Marjan Grootveld, Laurence Horton, Žilvinas Martinaitis, Adriana Rantcheva, Wilko Steinhoff, & Maaike Verburg. (2022). European Research Data Landscape Study Report (deliverables 3.2, 4.2, 5.2). Zenodo. https://doi.org/10.5281/zenodo.7351121

Matela, M. (Developer), Olejniczak, K. (Developer), Indarto, E. (Developer), Parkoła, T. (Author/Maker), & Tykhonov, V. (Developer). (2021). Dataverse Superset integration. Software https://github.com/SSHOC/dataverse-superset

Wilkinson, M.D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The fair guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018

ACKNOWLEDGEMENTS

This work has been supported by projects such as EOSC Future, SSHOC, CLARIAN.NL, FAIRsFAIR, and FAIRImpact.