Data Citation Quantity and Quality in Research Output of a Large-Scale Educational Panel Study

Nadine Mahrholz
Leibniz Institute for Educational Trajectories (LIfBi)
Wilhelmsplatz 3
96047 Bamberg
+49 951 863-3985
nadine.mahrholz@lifbi.de

Anke Reinhold
German Institute of International Educational Research (DIPF)
Schloßstraße 29
60486 Frankfurt
+49 69 24708-339
reinhold@dipf.de

Marc Rittberger
German Institute of International Educational Research (DIPF)
Schloßstraße 29
60486 Frankfurt
+49 69 24708-327
rittberger@dipf.de

ABSTRACT
In this paper, we report preliminary results of a small-scale case study about the data citation quantity and quality in research output of the National Educational Panel Study (NEPS), a longitudinal study analyzing educational processes in Germany across the lifespan. In order to collect research output based on NEPS data, we searched for and examined publications of a randomly selected sample of 72 NEPS data users. Altogether, we found 18 publications to be relevant for citation analysis. Compared to previous studies, the citation behavior in our sample can be assessed as better. However, publications often lack the inclusion of central data citation elements, such as a persistent identifier. The quality of data citations seems to vary across different types of research output. In a follow-up study, we plan to do a comprehensive sampling and analysis of NEPS related research output in order to verify our findings, and also to include further panel studies to compare citation behavior across different studies.

CCS Concepts
• **Information systems→Information retrieval**

Keywords
Bibliometric Analysis; Content Analysis; Scholarly Metrics; Data Citation

1. INTRODUCTION
Bibliometrics Analysis as "a science of science" [2] includes the identification of citation patterns with the principal aim of measuring performance in a given scientific domain. However, not only the citation of research findings documented in journal articles or other types of scientific output should be addressed in bibliometric analyses. In addition, the citation of data which was used to produce or to amend the scientific results and was not necessarily collected by the author itself must be evaluated likewise as specific requirements for the citation of data exist, as e.g. described by [1, 8, 12].

Open Science increasingly incorporates the openness of research data: Data sharing as the "release of research data for use by others" [3] has already been widely adopted by researchers in the environmental sciences, biology or physics. However, research data remain uncited to a large extent [6, 10] and data citation can still not be described as "a normative behavior in scholarly writing" due to the "multiplicity of data types" as well as "a lack of awareness regarding existing standards" [9] despite the fact that sharing research data leads to an increase in citation rates of a scholarly work independent of impact factor, release date or country of origin [11].

However, these attitudes towards data citation are most likely to be challenged in the near future as an increasing number of data repositories are being set up, providing researchers with large quantitative data sets, ready for reuse (e.g. as scientific use files or for remote as well as on-site access). In the case of panel studies, most providers (for example LIfBi, EU-SILC, ELFE[1]) obligate their users to cite the dataset used appropriately. Some providers even provide concrete guidelines for data citation, e.g. LIfBi and TREE (Transitions from Education to Employment).

According to our knowledge, only two studies have empirically analyzed whether data citation by researchers in the social sciences is carried out adequately in accordance to existing requirements [7, 9]. Mooney [7] identified a lack of data citation in a sample of 49 journal articles doing secondary analysis of datasets provided in a political and social sciences database: 61% of the articles failed to provide any type of citation. Moreover, 47% of the articles do not provide the citation of data-related publications. Mooney & Newton [9] developed categories for data citation on the basis of proposed data citation standards. The citation elements include Author, Title, Date, Publisher, Material Designator, Electronic Retrieval Location, and Persistent Identifier. However, there seems to be no work that differentiates between publication types in the analysis. Furthermore, in contrast to previous data citation studies, we focus on data only from educational science.

i-KNOW '15, October 21-23, 2015, Graz, Austria
 ISBN 978-1-4503-3721-2/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2809563.2809617

[1] Leibniz Institute for Educational Trajectories (LIfBi), European Union Statistics on Income and Living Conditions (EU-SILC), Étude Longitudinale Française depuis l'Enfance (ELFE).

Therefore, we take the largest educational study in Germany, the National Educational Panel Study (NEPS) as a use case and focus on identifying whether and how registered data users actually cite NEPS data across different publication types.

2. METHODS

In this section, we describe the methods used to retrieve relevant publications for data citation analysis. Furthermore, a preliminary scheme for data categorization is presented.

2.1 Publication Search and Collection

One particular provider of research data for the field of educational research is the Leibniz Institute for Educational Trajectories (LIfBi) in Germany. It offers – upon application – aggregated statistical data for scientific analysis to researchers in the fields of educational research, sociology, psychology, and others. The main study conducted by the LIfBi is the National Educational Panel Study (NEPS). Similar to other panel studies, data users have to sign a data use agreement upon the provision of NEPS research data. Users agree to cite NEPS data in publications according to specific requirements provided by the institution. Furthermore, they agree to notify the data provider about publications that are based on the analysis of NEPS data.

In order to create a collection of NEPS related publications, we searched for publications of a subset of NEPS data users and analyzed their publications for relevant citations. We decided to use this method to also find publications that are less easily retrievable, e.g. those publications that are not reported to the institution by the users, or publications that lack the inclusion of a persistent identifier.

So far, we have concentrated on the analysis of publications of a random sample of users who applied for data access in the year 2013, which was the first year of a remarkable number of users of NEPS. Altogether, 84 research projects with NEPS data were registered in 2013[2]. We took a random sample of 30 projects. After the exclusion of projects that were only registered for internal purposes (e.g. methodical projects for item development), 27 projects remained that were regarded for further analysis. Across all these projects, we identified 72 unique users from various disciplines, such as sociology, psychology, educational research, and economics.

The resources for the identification of NEPS related publications of the 72 users included the authors' reference lists on personal homepages (e.g. on institutional websites), the general bibliographical database Web of Science as well as – depending on the author's research background – domain-specific databases like ERIC for educational research, or SOLIS for sociology. In certain cases, additional material, such as project reports or conference abstracts, could be found by conducting Google and Google Scholar searches.

This procedure was chosen to ensure a high recall of NEPS related research output. The time span was defined from 2013 (the year in which all above mentioned projects were registered) to 2015. We performed author searches in all databases that were regarded as relevant. Retrieved documents were further examined to determine if they referenced NEPS data. This was done by scanning documents' abstracts and by performing keyword searches on the documents. As keywords we used different versions of the study title (e.g. in the case of English documents "NEPS", "National Educational Panel Study"). Beyond that, the keyword "data" was used in order to identify further panel studies with which authors have worked and which might serve as use cases in further data citation studies (see Section 4).

Bibliographical information for all documents that referenced NEPS data were saved in a database. For each document, information on the resources in which the document was found, authors' information (type of institution, discipline), as well as information on the document itself (publication type, language) were included in the database. Furthermore, documents' full texts were saved if available. In a next step, we cleaned the data set, meaning that we excluded duplicates (publications that were co-authored by two or more users from our sample and therefore retrieved more than once), publications in which NEPS data were not analyzed (e.g. publications simply containing an overview about the NEPS study and its methods), and research output for which no full text or other material was available (e.g. conference presentations).

2.2 Data Categorization

We performed a content analysis of the scholarly output that remained after the data cleaning process. First of all, we analyzed if the research output contained common elements that are suggested in data citation guidelines (e.g. [5, 12]). Referring to the Data Citation Synthesis Group [4], we divided citation elements into different categories, which are all provided and available for each data set at NEPS:

1. Data Provider
2. Study Title
3. Information on Data: starting cohort, wave
4. Data Version
5. Persistent Identifier: Digital Object Identifier (DOI)

Based on the citation elements classification scheme of the Data Citation Synthesis Group [4], elements (1) and (2) serve as "credit and attribution". Element (3) is an element to further describe the data used. The data version (4) serves to "specify and verify" the dataset. The inclusion of a persistent identifier (5) is a central element for the "unique and persistent identification" of a dataset.

We examined, if and which parts (title, text, footnotes, tables/figures, references; see also [9]) of the analyzed scholarly output contained any of the above mentioned data citation elements. We also assessed the quality of data citations by analyzing if users formally cited the research data according to the specific requirements given by the data provider. According to these requirements, users are obligated to cite a text element that contains the above mentioned data citation elements. In addition, they are also obligated to cite a basic reference that gives an overview of the NEPS study. There are no specifications on where to place the data citation elements in the publication.

3. PRELIMINARY FINDINGS

In this section we first present general findings of the publication search. After that, preliminary findings of the citation analysis are summarized.

3.1 General Findings (Search Strategy)

Altogether, we found NEPS related scholarly output for 27 out of the 72 data users. The following table shows the total and unique numbers of publications retrieved.

[2] A list of all NEPS projects and the related data users is publicly available at https://www.neps-data.de/en-us/datacenter/researchprojects.aspx.

Table 1. Total and unique numbers of retrieved NEPS related publications

	Total number of publications	Number of unique publications
NEPS related publications	76	48
NEPS related publications with a focus on data analysis	37	19

A large part of the research output was found on personal homepages (20 out of 76 publications), but also via Google Scholar (35 publications). The Web of Science (6 publications) and domain-specific databases (minimum 1 publication, maximum 14 publications) turned out be mediocre tools for search. It is to note, however, that our final sample of 76 publications also includes grey literature which is usually not recorded in literature databases. 28 publications (reports and conference abstracts) were only found via general searches performed on Google.

As Table 1 shows, after the exclusion of duplicates and publications that were not based on the analysis of research data, we got only 19 publications left. We assume that a large part of data users who registered for data use in the year 2013, have not yet published results on their projects. There might also be users who actually do not publish any of their project results. A further analysis of the publication behavior of different user groups (e.g. with regard to academic status or scientific discipline), or a user survey might provide further insights on these results.

The list of relevant research output mainly contained conference contributions (9 abstracts, 1 set of presentation slides, and 1 conference paper). Three contributions were classified as reports, and another three contributions as journal papers (2 published in peer-reviewed journals). Two contributions could not be specified. Altogether, most of the research output retrieved can be classified as grey literature.

3.2 Data Citation

In this section we report preliminary results of the data citation analysis that was performed for 18 publications. One of the above mentioned 19 relevant publications could not be included in the analysis because it was marked as being in press and no full text was available. Due to the small size, the following results just provide a rough overview of users' citation behavior. They cannot be seen as representative and must be validated in further studies. Altogether, only 7 out of 18 publications mentioned the data providing institution LIfBi. The study title (18) was mentioned in all publications. Also most publications (17) specified the data used in some way. Both elements were mainly included in the text section of the publications (see Table 2). A few times, they were also included in the references section as parts of NEPS related references, such as the study overview article or technical reports. Information to further specify the dataset used (data version, which is – in the case of NEPS data – also part of the digital object identifier) was only included in half (9) of the publications. The digital object identifier (DOI), which is a central element for the identification of NEPS related publications, was also only included in about a half (8) of all publications.

Table 2. Data citation elements in NEPS related research output

	Title	Text	Foot-notes	Figures/ Tables	Refer-ences	No. of Docu-ments
Data Provider	0	3	4	1	0	7
Study Title	2	18	4	5	8	18
Informa-tion on Data: starting cohort, wave	0	17	4	6	6	17
Data Version	0	2	5	1	3	9
Digital Object Identifier (DOI) incl. Data Version	0	2	5	1	1	8

With regard to citation specifications provided by the institution, results show that one third of the publications contained the text element which users are obligated to cite in publications. It is to note, however, that this element was mainly missing in conference contributions, such as presentation abstracts. This shows that the data citation quality seems different across the publication types. The result has to be further validated on a larger publication sample. The overview article, which data users are also obligated to cite, appeared in about two third of all publications, mainly in the text section (11 publications), and to a lesser extent in the references section (7 publications). The latter section was often missing in conference abstracts. Again, we found differences between different publication types, since users in our sample only failed to cite the overview article in conference abstracts and slides.

4. DISCUSSION

Compared to a similar study [9], the citation behavior in our small-scale sample can be assessed as better due to a higher percentage of articles providing any type of data citation as well as references to data-related publications. However, most publications in the sample of NEPS projects lack the inclusion of central data citation elements, such as persistent identifiers. The quality of data citations seems to vary across different publications types. Altogether, we found the citation quality to be lower in publications that were not formally published, such as conference abstracts. These findings have to be verified on a larger sample.

The search for publications turned out to be time intensive and only few publications could be retrieved. However, this method should ensure a high recall of data users' research output. It is to assume that there is a certain publication delay, meaning that for a part of projects registered in the year of 2013 no publications are yet available. Therefore, we plan to repeat the data analysis at a subsequent date. Furthermore, we searched for the publications of a random sample of data users. We plan to do a complete analysis of all 84 projects which applied for NEPS data in 2013 and also to include further panel studies to compare citation behavior across different studies in order to increase the validity of the findings.

One goal is to give concrete recommendations for data citation, not only with regard to journal articles but also to other scientific output like posters or slides.

5. REFERENCES

[1] Altman, M. and King, G. 2007. A proposed standard for the scholarly citation of quantitative data. *D-lib Magazine* 13, 3/4 (March/April 2007).

[2] Andrés, A. 2009. *Measuring academic research: How to undertake a bibliometric study*. Chandos Publishing, Oxford.

[3] Borgman, C. L. 2012. The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology* 63, 6, 1059-1078.

[4] Data Citation Synthesis Group. 2014. Joint Declaration of Data Citation Principles. M. Martone, Ed. FORCE11, San Diego, CA. URL: https://www.force11.org/group/joint-declaration-data-citation-principles-final (12.08.2015).

[5] DataCite International Data Citation Metadata Working Group. 2014. DataCite Metadata Schema for the Publication and Citation of Research Data. Version 3.1 October 2014. URL: http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf (12.08.2015).

[6] Robinson-García, N., Jiménez-Contreras, E., and Torres-Salinas, D. 2015. Analyzing data citation practices using the data citation index. *Journal of the American Society for Information Science and Technology* 66, 9.

[7] Mooney, H. 2011. Citing data sources in the social sciences: do authors do it? *Learned Publishing* 24, 2, 99-108.

[8] Mooney, H. 2013. A Practical Approach to Data Citation: The Special Interest Group on Data Citation and Development of the Quick Guide to Data Citation. *IASSIST Quarterly* 37, 1-4, 71-73.

[9] Mooney, H. and Newton, M. P. 2012. The Anatomy of a Data Citation: Discovery, Reuse, and Credit. *Journal of Librarianship and Scholarly Communication* 1, 1, eP1035.

[10] Peters, I., Kraker, P., Lex, E., Gumpenberger, C., and Gorraiz, J. 2015. Research Data Explored: Citations versus Altmetrics. In *Proceedings of the 15th International Conference on Scientometrics and Informetrics* (Istanbul, Turkey, June 29 – July 03, 2015). ISSI'15. Bogaziçi University Printhouse, 172-183.

[11] Piwowar, H. A., Day, R. S., and Fridsma, D. B. 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE* 2, 3, e308. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/ PMC1817752/ (12.08.2015).

[12] ZBW, GESIS & RatSWD. 2015. Auffinden, Zitieren, Dokumentieren: Forschungsdaten in den Sozial- und Wirtschaftswissenschaften. URL: http://auffinden-zitieren-dokumentieren.de/wp-content/uploads/2015/03/Forschungsdaten_DINA4_ONLINE_VER_02_06.pdf (12.08.2015).