Reinhold, Anke; Rittberger, Marc; Mahrholz, Nadine

Data citation policies of data providers within the scope of longitudinal studies in life course research

Ràfols, Ismael [Hrsg.]; Molas-Gallart, Jordi [Hrsg.]; Castro-Martínez, Elena [Hrsg.]; Woolley, Richard [Hrsg.]: Proceedings of the 21st International Conference on Science and Technology Indicators, València, September 14-16, 2016. Valencia : Editorial Universitat Politècnica de València 2016, S. 115-121

21ST international conference on science and technology indicators

STI Conference
2016 · València

Peripheries, frontiers and beyond

14 ·16 September 2016
Universitat Politècnica de València

BOOK OF PROCEEDINGS

www.sti2016.org

Data Citation Policies of Data Providers within the scope of Longitudinal Studies in Life Course Research[1]

Anke Reinhold*, Marc Rittberger** and Nadine Mahrholz***

*reinhold@dipf.de
German Institute for International Educational Research (DIPF), Schloßstraße 29, Frankfurt, 60486 (Germany)

** rittberger@dipf.de
German Institute for International Educational Research (DIPF), Schloßstraße 29, Frankfurt, 60486 (Germany)

*** nadine.mahrholz@uni-hildesheim.de
University of Hildesheim, Universitätsplatz 1, Hildesheim, 31141 (Germany)

ABSTRACT
In this article, a small-scale case study analyzing the nature of data citation policies within the scope of longitudinal studies in life course research is presented. The sample consists of eight data providers from Europe, North-America and Australia and was evaluated with regard to eight criteria which potentially affect data citation behavior of researchers in the field, for example the wording of data citation obligations or sanctions for not citing research data in accordance to given requirements. The study demonstrates that research data providers follow a wide range of approaches to data citation, especially in terms of data citation location within a publication as well as disposal obligations for data-related publications. However, this diversity might lead to inconsistency in data citation behaviour and also to a general lack of comparability of data citation quantity and quality as relevant factors in research evaluation.

INTRODUCTION
In order to meet the requirements of funding organisations or policy makers, the scientific output of researchers, research groups, institutions and even countries is regularly tracked by indicators that measure for example citation rates or citation impact. With the rise of altmetrics, attention in research monitoring has also shifted towards research activities that are – exclusively or complementarily – visible on the social web. However, citation analysis is still mainly focused on publication-related research output and so far only a few works have discussed the distinctiveness of research data as a considerable factor in citation analysis and research evaluation. For example, quantitative analyses of the Data Citation Index (DCI) (Thomas Reuters) (e.g. Peters, Kraker, Lex, Gumpenberger & Gorraiz, 2015; Robinson-García, Jiménez-Contreras & Torres-Salinas, 2015) as well as subject-specific publication depositories (Mooney, 2011; Mooney & Newton, 2012) have shown a general uncitedness of research data in the social sciences and the humanities, despite the fact that sharing research data can be associated with higher citation rates (Piwowar, Day & Fridsma, 2007).

Studies analysing the quality of data citation behaviour also uncovered that data citation is not carried out adequately with regard to existing requirements of academic journals (Mooney & Newton, 2012) or research data providers (Mahrholz, Reinhold & Rittberger, 2015). Additionally, as argued by Robinson-García et al. (2015), the citedness of research data

[1] This work was supported by the Leibniz Institute for Educational Trajectories (LIfBi).

heavily depends on the quality of data-related information provided by data repositories and varies across disciplines. Furthermore, data citation policies of scientific journals tend to be slightly stricter in the natural sciences than in the social sciences (cf. Blahous et al., 2015). A case study analysing data citation and sharing policies in the environmental sciences also demonstrates that "an overwhelming majority of funding agencies, repositories and journals fail to provide explicit directions for sharing and citing data" (Weber et al., 2011, p. 1). Obviously, making research data accessible and usable is a time-consuming and cost-intensive task. As a consequence, these activities should be appreciated by the scientific community and moreover demand for the inclusion of data citation indicators as a relevant factor in research monitoring. However, in order to make valid statements about data citation quantity and quality, it is necessary to thoroughly analyse the nature of data citation policies within a certain domain. In this paper, data citation policies of eight research data providers in Europe, the United States and Australia within the scope of longitudinal studies in life course research are being evaluated, e.g. with regard to citation principles and sanctions for data users who do not cite adequately. The aim of the study is to outline the different approaches followed by data providers or data repositories in terms of data citation policies which might influence data usage and citation behaviour of researchers in the domain.

Life course research is currently a very dynamic field of research in the social sciences. It provides stakeholders in politics and education with extensive and reliable data about life paths, transitions and decisions in private as well as professional lives. Furthermore, societal changes over extended timeframes of several years or even decades are being monitored. Longitudinal studies in live course research are generally characterized by large sample sizes, different cohorts of participants and various waves of surveys. There is also a strong demand for protecting sensitive personal information, e.g. about performance in school or the parent-child relationship, which are retrieved in these studies at a large scale. As a result, data providers in life course research generally dispose of high data security standards and offer a variety of data access modes, different type of data formats and data granularity. Users generally have to commit to data use agreements and are obliged to cite the research data used according to specific requirements. These data citation policies include aspects of contractual obligations of data citation, concrete requirements of including data citation elements (e.g. a persistent identifier) (cf. Mooney & Newton, 2012) or the position of the data citation within a publication (e.g. in the abstract or the references section) as well as disposal obligations for publications based on the research data provided.

DATA CITATION POLICIES OF DATA PROVIDERS IN LIFE COURSE RESEARCH – A CASE STUDY

For the case study a sample of eight longitudinal studies across the life course in Europe, North-America and Australia was identified by means of six criteria to ensure comparability: 1) thematic focus on educational and personal transitions, 2) ongoing research project, 3) at least a national or international perspective, 4) elaborated data access technologies (e.g. via a data center), 5) data use agreements as a prerequisite for data usage of sensitive data, 6) mention of data citation requirements.[2] Based on these criteria the following longitudinal studies were selected:

[2] The criteria were applied to the result set of an extensive web search which retrieved overall 19 longitudinal studies across the life course in Europe, North-America and Australia. The starting point for the web search was a list of longitudinal studies in the social sciences issued by Mallock, Riege & Stahl (2016, p. 146-148).

Table 1. Sample of longitudinal studies across the life course.

Study name	Research topics	Country	Start in year
Étude Longitudinale Française depuis l'Enfance (ELFE)	Impact of family circumstances, living conditions and environment on the physical and psychological development, health and socialization of children.	France	2011
Millennium Cohort Study (MCS)	Influence of early family context on child development and outcomes throughout childhood, adolescence and adulthood.	UK	2000
Negotiating the Life Course	Changing life courses and decision-making processes of men and women as the family and society move from male breadwinner orientation in the direction of higher levels of gender equity.	Australia	1997
National Educational Panel Study (NEPS)	Educational processes from early childhood to late adulthood.	Germany	2009
Panel Analysis of Intimate Relationships and Family Dynamics (pairfam)	Partnership and family dynamics in Germany.	Germany	2008
Socio-Economic Panel (SOEP)	Objective living conditions, values, willingness to take risks, current social changes, and the relationships and interdependencies among these areas.	Germany	1984
Transitions from Education to Employment (TREE)	Post-compulsory educational and labour market pathways of school leavers.	Switzerland	2001
Panel Study of Income Dynamics (PSID)	Employment, income, wealth, expenditures, health, marriage, childbearing, child development, philanthropy, education, and numerous other topics.	US	1968

For each of the research data providers in the domain of longitudinal studies across the life course, the following eight factors were documented by thoroughly eliciting regulatory and user service information on the data providers web sites[3]: 1) *wording of obligations with regard to data citation*, 2) *requirements for obligatory data citation elements*, 3) *requirements for data citation location within a publication*, 4) *availability of concrete examples for data citation*, *5) obligation to report data-related publications*, 6) *period of notification for data-related publications,* 7) *disposal obligation*[4] *for data-related publications* and 8) *sanctions for*

[3] For the analysis, different information sources on the providers' websites were reviewed, e.g. the data use agreements or the specific data citation section. The URLs of the homepages of all data providers in the sample are mentioned in the reference section.

[4] Publications which are based on a specific dataset are to be submitted to the data provider as a paper-based or digital version according to an agreement of use.

not citing research data in accordance to the requirements. From the point of view that the citedness of research data heavily depends on the quality of data-related information provided by data repositories (cf. Robinson-García et al., 2015), it is legitimate to assume that all of these factors might affect data citation behaviour of researchers in the field.

FINDINGS AND DISCUSSION

All eight data providers issue data use agreements that oblige their users to cite research data. The wording of these obligations (1) in the data use agreements differs significantly, ranging from very concrete citation specifications to rather general requests to cite in accordance to "academic conventions". Furthermore, all providers name obligatory data citation elements (2): Seventy-five percent of the data providers in the sample demand for including a distinct data version, 50% for including a Digital Object Identifier (DOI)[5] and 37.5% for naming a specific reference article which outlines the original study design. All eight data providers ask for the inclusion of an acknowledgement phrase indicating either the name of the study or the data center involved. These findings clearly indicate that data providers in life course research generally follow a top-down approach to prevent uncitedness of research data. It is also noticeable that again only 37.5% of providers in the sample provide guidelines for data citation location within a publication (3), e.g. for citing the study as the originator of the data in the title, the abstract or the reference section. This is surprising as it can be assumed that these recommendations are not only useful for guiding data users in the writing process. The recommendations might also foster awareness amongst researchers about the "quality" of a data citation within a document. For example, a data citation in the title or in the abstract can possibly be assessed as more valuable than a data citation in the caption of table or a figure. Interestingly, the data use agreement of the French ELFE study already indicates that users are obliged to cite the study in the title *and* the body of the text if the article is exclusively or primarily based on ELFE data (ELFE, 2014).

Apart from one, all data providers publish concrete examples for data citation on their websites which for example include the names of the authors (of a reference article), the name of the study and the DOI (4). Of course, researchers can already refer to more general data citation guidelines (cf. DataCite, 2014; ESRC, 2016; ZBW, GESIS & RatSWD, 2015[6]). Precise citation examples which relate to the actual study in use might nevertheless be even more important for supporting researchers and help them to prevent citation errors. Seventy-five percent of the providers in the sample insist on the obligation to report data-related publications (5) with only one provider, the Leibniz Institute for Educational Trajectories (LIfBi) for the NEPS data, calling for a period of notification for data-related publications of four weeks before publishing (6). And 50% of the data providers even issue a disposal obligation for publications using research data (7). Surprisingly, only one data provider – again LIfBi – calls for sanctions if data users do not cite in accordance to the data use agreement (cf. LIfBi, 2015) (8)[7]. In summary, it might be assumed that research data providers have already identified a need for action with regard to data citation misbehaviour. However, it still needs to be verified whether the data citation policies described here are

[5] One data provider has just recently added the obligatory inclusion of a DOI in his citation recommendations – this might be an indicator that the DOI becomes more widely accepted within the domain.
[6] This publication is not available in English yet.
[7] Although it is not explicitly stated that non-citations cause a breach of contract, citing the study name and the dataset used for analysis can be interpreted as "essential obligations" of the data use agreement.

appropriate measures for achieving high citation rates and citation quality of research data issued by providers of longitudinal data in life course research.

As stated in the introduction, the main goal of the study was to outline the variety of data citation policies within life course research and to discuss possible implications for data use and citation behaviour in the field. It could be demonstrated that data providers follow differing approaches in terms of data citation requirements. This involves data versions, identifiers and reference articles describing the original study design. In addition, data providers differ substantially with regard to recommendations for data citation location as well as disposal obligations for data-related publications. This might lead to a high diversity in data citation behaviour of data users in the field and potentially to non-comparable results in data citation analysis. It is therefore reasonable to argue that data providers should pursue the harmonisation of data citation specifications – in close cooperation with journals and research institutions involved in life course research. Furthermore, policy makers should strongly encourage the development of domain-specific data citation indicator sets for the valid representation of scientific output, allowing for an improved comparability and traceability of research.

LIMITATIONS OF THE STUDY AND OUTLOOK

We are aware that our research has some limitations. First, the study consists of a small sample which is not representative for data usage and citation within the social sciences in general. Second, there might be other longitudinal studies in life course research that meet the selected criteria presented above. Third, there is a predominance of European longitudinal studies in the sample. Finally, the study does not investigate the influence of data citation policies on the actual data citation behaviour of researchers in the field. A consecutive study, analysing data citation quantity and quality in a large sample of data-related publications in the social sciences might substantially enhance our understanding of data citation behaviour. Despite these limitations we believe our work has highlighted the importance of critically examining data citation policies beforehand as one milestone of coherent and comparable data citation analysis.

REFERENCES

Blahous, B., Gorraiz, J., Gumpenberger, C., Lehner, O., Stein, B. & Ulrych, U. (2015). Forschungsdatenpolicies in wissenschaftlichen Zeitschriften – Eine empirische Untersuchung. In *ZfBB*, 62, 12-24.

DataCite International Data Citation Metadata Working Group (2014). DataCite Metadata Schema for the Publication and Citation of Research Data. Version 3.1 October 2014. Retrieved March 16, 2016 from: http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf.

Étude Longitudinale Française depuis l'Enfance (ELFE) (2014): Charte d'accès aux données Elfe. Retrieved March 16, 2016 from: http://www.elfe-france.fr/images/documents/charte_acces_donnees_plateforme.pdf.

Economic and Social Research Council (ESRC) (2016). Data Citation. What you need to know. Retrieved March 16, 2016 from: https://www.ukdataservice.ac.uk/media/104397/data_citation_online.pdf.

Leibniz-Institute of Educational Trajectories (LIfBi) (2015). NEPS. Datennutzungsvertrag. Retrieved March 16, 2016 from: https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Datenzugangswege/Vertraege/NEPS_Datennutzungsvertrag_DE.pdf.

Mahrholz, N., Reinhold, A. & Rittberger, M. (2015). Data Citation Quantity and Quality in Research Output of a Large Educational Panel Study. In *Proceedings of the 15th International Conference of Knowledge Technologies and Data-Driven Business* (I-Know '15). Graz: ACM.

Mallock, W., Riege, U., Stahl, M. (2016). *Informationsressourcen für die Sozialwissenschaften. Datenbanken, Längsschnittuntersuchungen, Portale, Institutionen*. Wiesbaden: Springer.

Millennium Cohort Study (MCS). Retrieved March 16, 2016 from: http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=851&sitesectiontitle=Welcome+to+the+Millennium+Cohort+Study.

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

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

Negotiating the Life Course. Retrieved March 16, 2016 from: http://lifecourse.anu.edu.au/.

Panel Analysis of Intimate Relationships and Family Dynamics (pairfam). Retrieved March 16, 2016 from: http://www.pairfam.de/en/.

Panel Study of Income Dynamics (PSID). Retrieved March 16, 2016 from: https://psidonline.isr.umich.edu/.

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

Piwowar, H.A., Carlson, J.D. & Vision, T.J. (2011). Beginning to track 1000 datasets from public repositories into the published literature. In *Proceedings of the American Society for Information Science and Technology*, 48, 1, 1-4.

Piwowar, H.A., Day, R.S. & Fridsma, D.B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate, *PLoS ONE* 2, 3, e308. Retrieved March 16, 2016 from: http://www.ncbi.nlm.nih.gov/pmc/articles/ PMC1817752/.

Robinson-García, N., Jiménez-Contreras, E. & Torres-Salinas, D. (2015). Analyzing data citation practices using the data citation index. In *Journal of the Association for Information Science and Technology*, DOI: http://dx.doi.org/10.1002/asi.23529 (Preprint).

Socio-Economic Panel (SOEP). Retrieved March 16, 2016 from: http://www.diw.de/en/soep.

Transitions from Education to Employment (TREE). Retrieved March 16, 2016 from http://www.tree.unibe.ch/index_eng.html.

Weber, N.M., Piwowar, H.A. & Vision, T.J. (2010). Evaluating Data Citation and Sharing Policies in the Environmental Sciences. In *Proceedings of the American Society for Information Science and Technology*, 47, 1, 1-2.

ZBW, GESIS & RatSWD (2015). Auffinden, Zitieren, Dokumentieren: Forschungsdaten in den Sozial- und Wirtschaftswissenschaften. Retrieved 16 March, 2016 from: http://auffinden-zitieren-dokumentieren.de/wpcontent/uploads/2015/03/Forschungsdaten_DINA4_ONLINE_VER_02_06.pdf.