Gabriela Molina León, Gabriella Skitalinskaya, Nils Düpont, Jonas Klaff, Anton Schlegel, Hendrik Heuer, Andreas Breiter (2022): Co-Creating a Research Data Infrastructure with Social Policy Researchers. In: Proceedings of the 20th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies - Posters and Demos, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.48340/ecscw2022_p03

Co-Creating a Research Data Infrastructure with Social Policy Researchers

Gabriela Molina León, Gabriella Skitalinskaya, Nils Düpont, Jonas Klaff, Anton Schlegel, Hendrik Heuer, Andreas Breiter
University of Bremen, Germany
Contact Author: molina@uni-bremen.de

Abstract. We present a case study on co-creating a research data infrastructure together with social policy researchers. Over three years, we investigated how the social scientists worked with data, and designed a collaborative system to support them in the harmonization, validation, exploration, and sharing of research data. We conducted several co-creation workshops, interviews, surveys, and user studies not only to co-design the system but also to assess the benefits and limitations of our participatory approach for this interdisciplinary collaboration. The evaluation uncovered that the researchers were satisfied with the processes and tools that we developed, and that the system was successfully adopted. We found that when working in a large interdisciplinary project, especially in the context of social policy research, it is critical to assess the status of the data early on, and to discuss how the group and individual goals connect with each other, to ensure long-term engagement and commitment.

Introduction

In the last decade, we have witnessed a rapid increase in the quantity of data available in science. Accordingly, CSCW researchers have been studying how experts work with data in diverse domains to find out how technology can support cooperative scientific work (Velden et al., 2014). Vertesi and Dourish (2011) studied how the way planetary scientists produce data is a key factor in how they share data. Neang et al. (2021) investigated the social and organizational concerns surrounding data integration in oceanography. Overall, the scientific culture and practices of the disciplines play a critical role in how computer-support systems can facilitate scientific work (Jirotka et al., 2013). This is what Lee et al. (2006) call the *human infrastructure* of cyberinfrastructure.

Tenopir et al. (2015) found that the norms of data sharing vary highly between disciplines. While astronomy and biodiversity researchers have a culture of data sharing, medicine and social sciences researchers are less likely to share. According to Savage and Vickers (2009), researchers rarely create appropriate metadata early enough, which later leads to not releasing the data because of the associated workload.

Given the need for more efforts to support sharing in the social sciences, we sought to co-design a research data infrastructure together with social science researchers. Over three years, we collaborated with social policy experts in a multidisciplinary project aimed at analyzing and explaining social policy dynamics worldwide. We supported them on the harmonization, validation, exploration, and sharing of their datasets. Accordingly, we present a case study tackling the following research question:

RQ What to consider when applying co-creation as a design methodology to create a data infrastructure system for social policy researchers?

We present our insights on how social policy researchers organize their data work, and how we co-designed a data infrastructure to support them. According to the evaluation, the system was successfully adopted. We share our recommendations for data infrastructure projects based on our co-creation study.

Motivation and methods

Our case study is based on a multidisciplinary research project on global social policy involving 29 researchers from political science, sociology, geography, and computer science (CRC 1342: Global Dynamics of Social Policy, 2022). We report our insights from the first three years of our on-going collaboration.

The main goal of the project is to collect data on social policies worldwide. The data involves not only social policy indicators (i.e. variables) created by the researchers, but also indicators collected by institutions such as the World Bank. We designed an information system to harmonize, share, and explore said data.

We applied co-creation as a design methodology (co-design). Co-creation is based on conducting regular workshops with the stakeholders to not only design a solution *for* them, but also *with* them (Sanders, 2008). In the workshops, we used well-known methods for creative work such as wishful thinking (Kerzner et al., 2019), paper prototyping (Snyder, 2003), and reflective discussions (Molina León and Breiter, 2020).

To learn more about their work, we conducted contextualized interviews with researchers of different project roles, and collected artifacts such as data files, papers, and data analysis scripts. All the interviews and discussions were recorded and analyzed through open coding according to grounded theory. To evaluate the collaboration and the system, we conducted a survey and two user studies whose results we present in the Evaluation section.

The Information System

Through the workshops and interviews, we elicited and iteratively refined the following design requirements for the system:

R1 Support data harmonization. The researchers collected time series data from various sources in different formats (e.g. books, CSV files). They required support on combining the datasets together and preparing them for analysis.

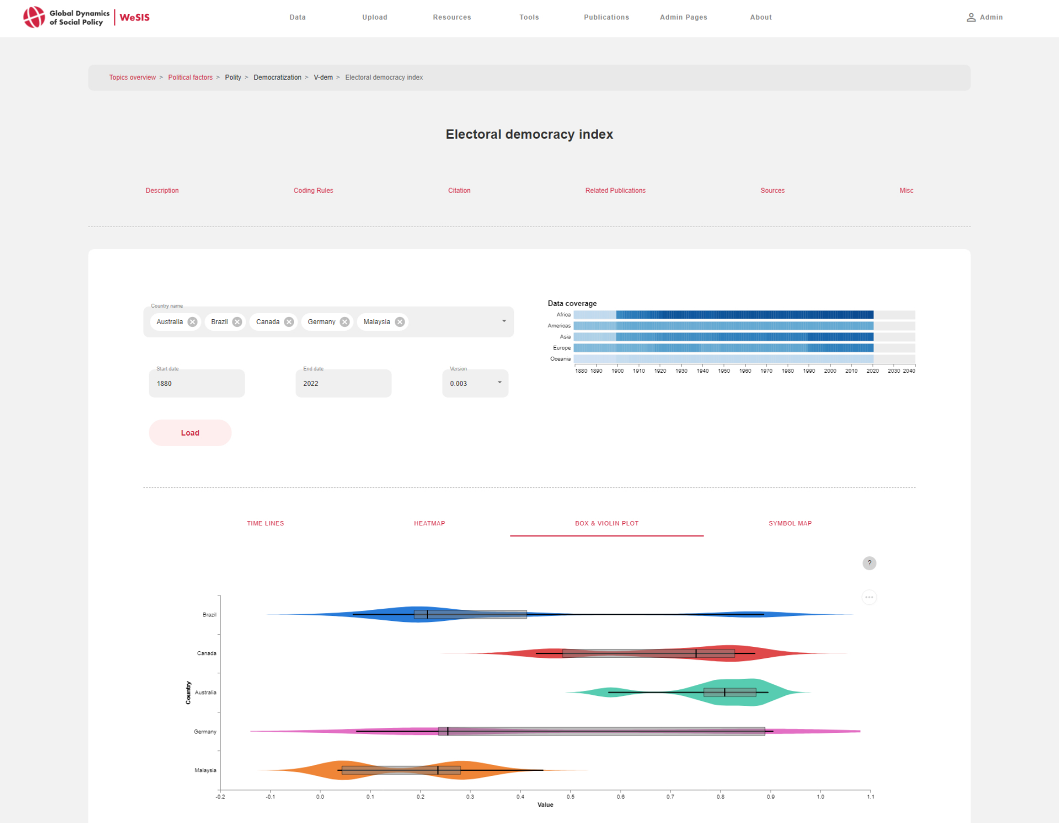
R2 Support data validation. The data standards agreed on needed to be validated systematically. The researchers wished for support on checking the data, e.g. verifying country names.

R3 Enable interactive data exploration. Once the data was in the system, the social scientists wished for tools to search and filter the indicators according to their research interests.

R4 Allow flexible sharing of data and resources. Sharing was a priority to collaborate with other researchers. Sharing tools would help ensure transparency, reproducibility, and reuse of their research and data.

To support data harmonization (**R1**), we established *Data and metadata standards* as guidelines for the data collection and merging processes. We created a dedicated wiki to document the standards and the data itself, ensuring a high level of documentation quality and transparency. Furthermore, we co-developed a universal dataset template. The template covered all necessary attributes for each data point and metadata. We also harmonized existing practices in data coding and established coding rules. These rules described the requirements for each template item, such as country codes, naming guidelines, etc.

For the data validation (**R2**), we implemented a validation pipeline, which thoroughly checked if the uploaded data fulfilled the standards and gave detailed feedback otherwise. For data exploration (**R3**), we designed three interfaces that present the data in different ways:



Topics overview > Political factors > Polity > Democratization > V-dem > Electoral democracy index

Electoral democracy index

Country name: Australia, Brazil, Canada, Germany, Malaysia

Start date: 1880

End date: 2022

Version: 0.003

Figure 1. The *Electoral democracy index* indicator page (upper part of the interface).

1. **Indicator page**. This page presents all the information about a particular indicator, covering its coding rules, sources, and more. Since the researchers wished to discover and analyze spatio-temporal data patterns, the page supports exploration through a coverage visualization, interactive search and filtering options, and a wide range of visualizations tailored for each data type (see Figure 1).

2. **Country profile**. Many theories and explanations in social policy research focus on countries as the focal unit of analysis. Thus, we co-designed profiles that zoom in on a specific country and shift the focus to the development *within* it. As such, the profile is a valuable tool to inform area studies, providing easy access to a set of key indicators.

3. **Data Explorer**. Here, we focus on supporting the analysis of multiple indicators simultaneously by providing basic correlation insights and visualizations tailored to different combinations of indicator types. While correlation is not causation, it helps uncovering possible relationships that can be further inspected and may inform inductive reasoning.

To support data sharing (**R4**), all pages provide various exporting options with version control and all visualizations are downloadable. While the system is still being prepared for general public access, registered users can compile indicators into so-called “datasets” and share them with non-registered users via token-based urls. For script sharing, we created the Community Notebooks page, where researchers can upload computational notebooks to reproduce and replicate results.

Evaluation

After the first five workshops, we conducted a survey to investigate how the researchers perceived the collaboration so far. Eight researchers participated. Despite the small sample, the results provided relevant insights. Paper prototyping and group discussions were the most preferred activities as they allowed the experts to concretize their ideas and refine them by discussing them with their peers. While researchers with high attendance were more positive about how their participation influenced the outcome, half of the participants did not find such regular meetings helpful for their work but noted that the workshops were the place where they learned most about the research of their colleagues.

A few months later, the first version of the system was almost ready to be released within the project. Before doing so, we conducted a small user study to evaluate the interface design and to further assess the benefits and limitations of our participatory approach. The researchers performed three navigation tasks focused on the data visualizations, and participated in an interview. We had six participants. That was the first time they could interact with the system, and four participants reported to be impressed because it offered more options than other systems they knew. This led to more positive answers about our collaboration being helpful for their work. In the interviews, the most mentioned issue was that not everyone was attending the workshops. Initially, we invited all researchers to encourage openness and diversity, but only a few attended regularly.

Shortly after releasing the system, we conducted a second study with 12 researchers to evaluate the validation and exploration features. The study consisted of five tasks. The first and second tasks required uploading a dataset, with and without errors. The other tasks involved searching and exploring a given indicator, interacting with a *Country profile*, and exploring indicator relationships in the *Data Explorer*. After each task, participants rated its difficulty, and shared any problems they had. Figure 2 presents the difficulty answer rates.

All but one participant completed the validation tasks successfully and everyone finished the exploration tasks successfully. Overall, the outcome was positive because most participants found all tasks easy to perform. The researchers found the validation tests especially helpful for verifying the data. However, this required additional work to adjust the data according to the established standards — in contrast to their previous manual approach. They especially appreciated the option to combine indicators in the Data Explorer, missing in other systems.

Regarding the co-creation process, the evaluation showed that the system fulfilled the requirements and that the participants felt that their ideas were included. However, the diversity of goals among the researchers, combined with

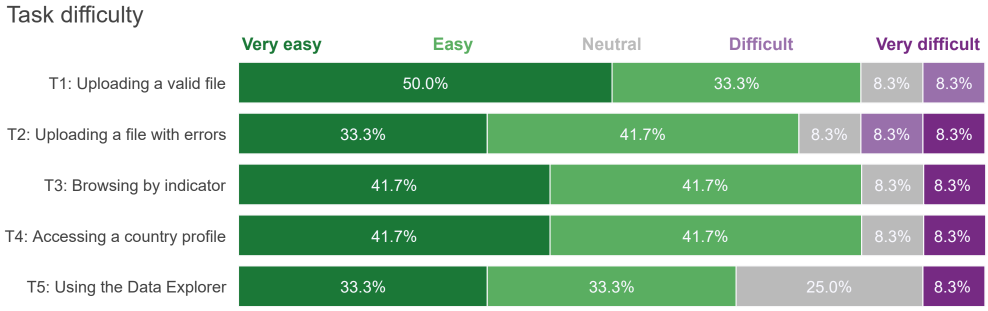


Figure 2. Difficulty rank per task in the second user study, evaluating the data validation and exploration features.

the inconsistent attendance, made it challenging to design custom features. Moreover, the researchers saw the benefit of a systematized workflow for future colleagues but considered that co-creating increased their workload.

Recommendations for data infrastructure projects

Based on our case study, we propose the following recommendations for researchers and practitioners who plan to co-create a data infrastructure:

1. *Ensure a limited yet representative group of participants actively involved in the process*. Initially, we invited all researchers. We noticed that too many people were involved, some attended rarely, and power structures influenced who voiced their opinion (e.g. doctoral students hesitated before disagreeing with their supervisors). Overtime, we decided to invite only two persons per research group and to organize teams mixing different groups and roles.

2. *Assess the status and amount of data available early on*. We planned to use example datasets for designing the system early on, yet such datasets were not ready. Thus, the design and development had to happen in parallel to the data collection, which is not rare for research data management systems.

3. *Connect individual and group goals, working in short iterations*. Long-term projects struggle with keeping participants engaged. Discuss the individual goals of every participant and how they connect to the project goal, prioritizing a balance between both. Short work iterations lead to less repetition and facilitate including the input of the participants in every step.

4. *Define the roles and tasks of the participants early on*. The expectations of the social scientists about the computer scientists, and viceversa, were different because each group overestimated the work speed of the other. This illustrates how misconceptions can easily occur in multidisciplinary projects. Although participatory methods are favored to get everyone's voice heard, it

is also important to clearly define the tasks and commitment needed for the collaboration to succeed.

Discussion and conclusions

Tenopir et al. (2015) suggest that creating a sound data infrastructure is a solution to impulse data sharing among researchers. However, designing for reproducibility has multiple constraints and challenges (Feger et al., 2020). Our study shows that designing such a system is a long-term process that requires a close and exhaustive collaboration. In the workshops, we found that some researchers did not identify themselves as users because it would take a long time for the system to reach a state where it could provide immediate benefits. This reflects one of the challenges of developing groupware applications reported by Grudin (1994): the disparity between work and (immediate) benefit.

Promoting collaboration among the researchers was another positive outcome beyond the system adoption. Participants developed a shared understanding of their collaborative research in the workshops. This confirms the findings of Neang et al. (2021) with oceanographers. Overall, our case study presents insights on how to co-create a data infrastructure for social policy research. Accordingly, we provide our recommendations for similar endeavors. Our work contributes to the open science efforts within the scientific community.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)– Projektnummer 374666841 – SFB 1342.

References

CRC 1342: Global Dynamics of Social Policy (2022): ‘About the CRC 1342’.

Feger, S. S., P. W. Wozniak, L. Lischke, and A. Schmidt (2020): ‘‘Yes, I Comply!’: Motivations and Practices around Research Data Management and Reuse across Scientific Fields’. *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW2.

Grudin, J. (1994): ‘Groupware and Social Dynamics: Eight Challenges for Developers’. *Commun. ACM*, vol. 37, no. 1, pp. 92–105.

Jirotka, M., C. P. Lee, and G. M. Olson (2013): ‘Supporting scientific collaboration: Methods, tools and concepts’. *Computer Supported Cooperative Work (CSCW)*, vol. 22, no. 4, pp. 667–715.

Kerzner, E., S. Goodwin, J. Dykes, S. Jones, and M. Meyer (2019): ‘A Framework for Creative Visualization-Opportunities Workshops’. *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 748–758.

Lee, C. P., P. Dourish, and G. Mark (2006): 'The Human Infrastructure of Cyberinfrastructure'. In: *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*. New York, NY, USA, p. 483–492, Association for Computing Machinery.

Molina León, G. and A. Breiter (2020): 'Co-creating Visualizations: A First Evaluation with Social Science Researchers'. *Computer Graphics Forum*, vol. 39, no. 3, pp. 291–302.

Neang, A. B., W. Sutherland, M. W. Beach, and C. P. Lee (2021): 'Data Integration as Coordination: The Articulation of Data Work in an Ocean Science Collaboration'. *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW3.

Sanders, E. (2008): 'An evolving map of design practice and design research'. *interactions*, vol. 15, no. 6, pp. 13–17.

Savage, C. J. and A. J. Vickers (2009): 'Empirical Study of Data Sharing by Authors Publishing in PLoS Journals'. *PLOS ONE*, vol. 4, no. 9, pp. 1–3.

Snyder, C. (2003): *Paper Prototyping: The Fast and Easy Way to Design and Refine User Interfaces*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Tenopir, C., E. D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch, D. Pollock, and K. Dorsett (2015): 'Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide'. *PLOS ONE*, vol. 10, no. 8, pp. 1–24.

Velden, T., M. J. Bietz, E. I. Diamant, J. D. Herbsleb, J. Howison, D. Ribes, and S. B. Steinhardt (2014): 'Sharing, Re-Use and Circulation of Resources in Cooperative Scientific Work'. In: *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*. New York, NY, USA, p. 347–350, Association for Computing Machinery.

Vertesi, J. and P. Dourish (2011): 'The Value of Data: Considering the Context of Production in Data Economies'. In: *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*. New York, NY, USA, p. 533–542, Association for Computing Machinery.