IJDC | *Conference Pre-print*

Inter-organizational coordination work in digital curation: The case of Eurobarometer

Kristin R. Eschenfelder
The Information School, University of Wisconsin-Madison

Kalpana Shankar
School of Information and Communication Studies, University College Dublin

Abstract

Open research is predicated upon seamless access to curated research data. Major national and European funding schemes, such as Horizon Europe, strongly encourage or require publicly funded data to be FAIR - that is, Findable, Accessible, Interoperable, Reusable (Wilkinson, 2016). What underpins such initiatives are the many data organizations and repositories working with their stakeholders and each other to establish policies and practices, implement them, and do the curatorial work to increase the available, discoverability, and accessibility of high quality research data. However, such work has often been invisible and underfunded, necessitating creative and collaborative solutions.

In this paper, we briefly describe how one such case from social science data: the processing of the Eurobarometer data set. Using content analysis of administrative documents and interviews, we detail how European data archives managed the tensions of curatorial work across borders and jurisdictions from the 1970s to the mid-2000s, the challenges that they faced in distributing work, and the solutions they found. In particular, we look at the interactions of the Council of European Social Science Data Archives (CESSDA) and social science data organizations (DO) like UKDA, ICPSR, and GESIS and the institutional and organizational collaborations that made Eurobarometer "too big to fail". We describe some of the invisible work that they underwent in the past in making data in Europe findable, accessible, interoperable, and conclude with implications for "frictionless" data access and reuse today.

Submitted 15 December 2019 ~ *Accepted* 19 February 2020

Correspondence should be addressed to Kristin R. Eschenfelder, School of Computer, Data & Information Sciences, University of Wisconsin-Madison, 4217 HC White Hall, 600 North Park Street, Madison, WI 53706. US. Email: eschenfelder@wisc.edu

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Open research is predicated upon seamless access to curated research data. Major national and European funding schemes, such as Horizon Europe, strongly encourage or require publicly funded data to be FAIR - that is, Findable, Accessible, Interoperable, Reusable (Wilkinson, 2016). What underpins such initiatives are the many data organizations and repositories working with their stakeholders and each other to establish policies and practices, implement them, and do the curatorial work to increase the available, discoverability, and accessibility of high quality research data. However, such work has often been invisible and underfunded, necessitating creative and collaborative solutions.

In this paper, we briefly describe how one such case from social science data: the processing of the Eurobarometer data set. Using content analysis of administrative documents and interviews, we detail how European data archives managed the tensions of curatorial work across borders and jurisdictions from the 1970s to the mid-2000s, the challenges that they faced in distributing work, and the solutions they found. In particular, we look at the interactions of the Council of European Social Science Data Archives (CESSDA) and social science data organizations (DO) like UKDA, ICPSR, and GESIS and the institutional and organizational collaborations that made Eurobarometer "too big to fail". We describe some of the invisible work that they underwent in the past in making data in Europe findable, accessible, interoperable, and conclude with implications for "frictionless" data access and reuse today.

Method

This paper is part of a larger project on the sustainability of social science data archives. For this paper, we obtained with permission 895 documents from the Council on European Social Science Data Archives (CESSDA) Archives (currently headquartered in Bergen, Norway) and other materials from individual social science data archives that interacted with CESSDA, including the UK Data Archives (UKDA) and the Inter-university Consortium for Political and Social Research (ICPSR) in Ann Arbor, Michigan. The documents include copies of correspondence between the chair and various members, minutes of most bi-annual meetings, emails, and secondary literature such as articles describing CESSDA or its projects written by current and former CESSDA leaders or others in the field. We also rely heavily on the meeting minutes to supplement the correspondence and draw on the secondary literature as well. These documents were coded by the authors for themes related to the project as well as topics that emerged from the data. After this document-level content analysis, intermediate "analytical memos" (Chenail, 2012) or narratives were developed.

We limit our discussion to sharing of data between and among data archives not researcher to researcher sharing, or even motivations for researchers to deposit data in data archives. We also limit discussion to a certain type of data – structured quantitative data stemming from social sciences research, predominately political science. While this places limits on our findings, the lessons learned from social science data archives are applicable to data archives of many different types that seek to share data across national boundaries. As we do not own the primary documents upon which we draw, we regretfully cannot open the data upon which this paper builds.

Review of the Literature

As data are moved across boundaries – from one spreadsheet to another, from one researcher to another, or across organizations or national boundaries, data meet resistances and potentially create new ones. In his book *A Vast Machine*, Paul Edwards described this concept as "data friction". He writes:

> "Every movement of data across an interface comes at some cost in time, energy, and human attention. Every interface between groups and organizations, as well as between machines, represents a point of resistance. … In social systems, data friction consumes energy and produces turbulence and heat – that is, conflicts, disagreements, and inexact, unruly processes." (Edwards et al., 2011)

As Bates (2018) and others (Beer, 2013; White, 2017) have argued, data frictions are also mobilizers of resources. As Bates (2018) explains, frictions bring actors together to create new, shared understandings and work alignments as they work out ways to make data move across boundaries.

> "Data friction influences what data are captured and how they are, or are not, made accessible and re-usable by different social actors, and ultimately how data movements are bringing social actors into new and complex forms of relation with one another." (Bates, 2018, p. 425).

In other words, rather than thinking about such frictions solely as impediments to inter-organizational data collaboration, new questions arise if we also consider data frictions as an opportunity for structuring relationships and activities. For example, in his study of shared infrastructures such as labs and large complex technologies such as planes, John Law identifies the strategies people draw on in developing infrastructure and its activities and keep it going (Law, 1994). Law draws attention to the work and effort needed to achieve and maintain organization and suggests major "styles" that actors draw on and mix up to organize including:

- Administrative: Rationalizing and managing through planning, reporting, assessment and adherence to rules.
- Entrepreneurial: Pragmatically shifting actions based on changing conditions, emphasizing accountability and responsibility to follow through.
- Vision, or using charisma, stories and shared goals to motivate.
- Vocation, or emphasizing human special skills and knowledge versus mechanized or automated processes. (Law, 1994; Law & Mol, 1995)

One can identify these "styles" in DO quite readily; they deploy vision and vocation to acquire data sets and "market" them, but the curatorial and organizational work must be done or the data they acquire has no use. These "styles", tailored to individual organizational actors, must be meshed if DO are to collaborate – as they often do, for numerous reasons. Relevant and useful datasets are typically distributed across multiple institutions and need to be coordinated across institutions rather than centralized within one institution (Bertot & Choi, 2013). Coordination has also resulted in the development of standards and policies for data archiving, preservation, administration, and discoverability.

Thus, these DO are not just organizations, but research infrastructures that support and even generate new kinds of research. Their workings comprise an ever-changing set of relationships with depositors, users, and other stakeholders – including each other. Development and ongoing management of infrastructure require significant coordination because many

infrastructures are shared between organizations (e.g., projects, labs, universities) and their data reside in dispersed geographical and institutional locations. Negotiations across disciplines and the professions that support infrastructure are necessary to bring data together including choosing underlying architectures, exchange and storage standards, curation and use policies, different stakeholder. The institutional dynamics embedded in these negotiations shape infrastructure design (Mayernik, 2016), maintenance, and repair (Ribes et al., 2013)

Infrastructures must adapt over time to changing local conditions, new components and changing stakeholders (Borgman, 2014; Borgman et al., 2015; Edwards et al., 2013). Infrastructures are typically intended for long term service and therefore must deal with issues unique to long-term temporal scales (Karasti, Baker, & Millerand, 2010; Paine & Lee, 2014; Jirotka, Lee, & Olson, 2013), Long term sustainability of infrastructure is a challenge due to: changes in the underlying subjects, objects, methods and fields of science, changes in expectation about temporal or geographical scale, managerial and financial challenges to the organizations that host data infrastructures, and funding and regulatory shifts (Ribes & Polk, 2014). "Infrastructure time" requires a different mindset that fosters innovation while also maintaining stability and backwards compatibility for ongoing users (Karasti, Baker, & Millerand, 2010; Ribes & Finholt, 2009)

Geography also matters. Where a data organization, or its components, are geographically located matters because location influences mission, host relations, regulations, and national political loyalties. For example, Bates, Lin, and Goodale (2016) chronicle the geographical and temporal movements of meteorological data in the UK from their inception in the national weather office, gaps filled in by individuals with local knowledge of the particular weather stations, and ultimately acted upon by financial markets and other "end users". To give another example of the importance of place, as of May 2017, a data repository that is located in a European country will be governed by the General Data Protection Regulation (GDPR) with respect to privacy and data ownership. A misuse or breach of data (accidental or otherwise) may happen by a user outside of the European Union but the liability and responsibility rests with the repository in Europe, which may put in place technical and policy tools to minimize risk of such breaches.

Data organizations exist in a field-level web of interlinked stakeholders and partners (Eschenfelder & Shankar, 2015), and it is at this level we focus the current study. Studies of inter-organizational collaboration have pointed to benefits for knowledge and resource sharing (Borgman, Wallis, & Maynerik, 2012; Ribes & Polk, 2014). For example, the activities of the National Center for Atmospheric Research (NCAR) point to a constantly evolving ecosystem and the establishment of internal coherence before intra-organizational data sharing and data discovery can be achieved (Baker et al., 2015).

The holy grail at all levels of data sharing is interoperability, or the seamless exchange, use, and re-use of data regardless of location. Data are not naturally interoperability, and interoperability requires a great deal of human labor and institutional support: People and organizations must develop and apply a "constellation of concepts, approaches, techniques and technologies" in order to "make heterogeneous data work with each other" (Ribes, 2017, p.1515). Seamlessness is never seamless because data and DO are material entities. But "flow" implies smoothness, and as Borgman (2015) argues, "data do not flow like oil". Instead, the movements of data are more accurately represented as starts, stops dead ends, compromises - that is, frictions (Edwards, 2010; Mayernik, 2016).

Early History

CESSDA, or the Consortium of European Social Science Data Archives, is a European Research Infrastructure based in Norway and composed of a board of representatives from member European nations. As described on their website, "CESSDA provides large-scale, integrated and sustainable data services to the social sciences. It brings together social science

data archives across Europe, with the aim of promoting the results of social science research and supporting national and international research and cooperation." As implied by the quotes use of the term "integrated" and "brings together," while CESSDA undertakes projects as an organization, major activities remain with member data organizations including the provision of data services, stewardship of data, and relationships with data contributors, with stewardship over their data.

The 1970s and 1980s saw the start of several large cross-national European polling/survey efforts such as the Eurobarometer, the European Values Survey, and the International Social Survey Programme (ISSP) among others, and CESSDA was involved in curating and disseminating these. While demographic data for nations had long been available, comparative data on values and opinions had not, and the creation of these data sets allowed for researchers to ask and answer exciting new research questions (Bréchon, 2009). These comparative studies ideally used the same questions and same methodologies in each nation, allowing comparative analysis and they were conducted regularly, allowing longitudinal analysis, and longitudinal/comparative combinations.

One of these, the Eurobarometer, or "EuroB" as it was called amongst data organizations, was started in 1974 for and by the European Commission (EC). The European Commission funded and continues to fund the creation, deployment and analysis of the Eurobarometer survey and has ownership of the Eurobarometer data. While the Eurobarometer has grown more complex over time, at its core it is a bi-annual multi-country tracking of economic and social issues via a structured orally delivered survey (Inglehart & Reif, 1991). As described by ICPSR – one of the data hosts of Eurobarometer data

> "The standard Eurobarometer surveys are designed to provide a regular monitoring of the social and political attitudes among the European publics, to obtain regular readings of support for European integration, public awareness of and attitudes toward European unification, the institutions of the European Communities, as of 1992 the European Union, and its policies in complementary fashion." (ICPSR Eurobarometer Survey Series)

Curating EuroB, or, Belling the Cat

While the project was funded and analysed by the European Commission, and delivered via EC contractors, funding did not include archival or user servicing work. While the EC retained the right to not distribute the data, it historically always made the data open access by providing free copies to data organizations (mostly GSEIS in Germany and ICPSR in the US).

With the EC as owner of the EuroBarometer data and various national data organizations both in and outside of Europe the custodians, the burden of curating these large and complex data sets created complex challenges for data organizations. The workload for curating these cross-national data sets was very high. It could take two years or longer for data organizations to get the data, clean it, and make it available. Some data within the Eurobarometer were embargoed for periods of time as they were politically sensitive or funded by special interests; such embargoing increases costs of curation even more. Moreover, once the data was deposited, archives complained that they were not being used because users couldn't find relevant data sets due to poor study descriptions, no cross-language indexing, and at the time, no computer searching. There were originally no standardised vocabularies for searching both for studies and also for questions.

However, Eurobarometer was too big and important a data set to ignore, as most other national opinion surveys at the time were not shared (Bréchon, 2009). As a result, many CESSDA member data organizations actively sought to be a part of the Eurobarometer effort by ensuring researcher access to usable Eurobarometer data and sought credit for their work curating and hosting Eurobarometer data. For example GESIS in Germany, one of the homes

of EuroB, even inserted itself into the citation for use of Eurobarometer for secondary analysis So while all CESSDA members agreed in theory on the importance of Eurobarometer required coordinated efforts, it was unclear how best to coordinate these efforts among European data archives.

Coordination Work: Processing EuroBarometer

Confronted by the new mission of curating and making available the data from the EuroB, and other cross national studies, and motivated to avoid duplication of curation work, European archives first considered organizing the work by survey, with each archive managing one major survey. This proved to be unworkable, so one of the member data archives volunteered to do all of the work. But at this stage, the amount of work involved in curating these data sets may not have been apparent, and it is clear that the member repositories weren't fully cognizant of all of the subtasks that came to be involved in curation.

While these efforts were going on (with the Commission complaining about delayed timetables and poor data quality), CESSDA members learned that ICPSR had independently received its own copies of Eurobarometer data from the EC and was proceeding with processing. This acquisition of European data by an American data organization flew in the face of the tacit assumption that Europeans ought to take care of European data. To add insult to injury, ICPSR restricted access to its data to its paying members, leaving open the possibility that European researchers might have to pay to use future Eurobarometer data (traditionally European archives gave each other access to European data for free or cost of reproduction). Accordingly, some CESSDA members expressed concern that Eurobarometer data therefore might not be as available to European researchers in the future and that efforts needed to be undertaken to insure availability. As one member of CESSDA argued, European data organizations should NOT do processing that ICPSR was already doing. But reliance on ICPSR for processing could mean that Europeans could have to pay for access to archival versions of European data if ICPSR did not give it to them.

In short order, EuroBs become too much work for ICPSR as well. By 1991, ICPSR reported that it could not continue to allocate as many resources towards Eurobarometer processing. Problems with ICPSR archival processing led to a re-emergence of calls to bring Eurobarometer processing back to Europe. But continued lack of resources for curation made taking on the Eurobarometer difficult, and CESSDA members, despite enthusiasm about European curation in theory, did not organize themselves to undertake collaborative curation.

By late 1991 and early 1992 resource constraints combined with better communication with European archives, led ICPSR to reach out to CESSDA to develop a collaboration to disperse the workload of curating all the Eurobarometer data. The ensuring years saw the coordinated development of a new archival standard (later became the DDI) to enable these reluctant partners to coordinate efforts (Williams, Shankar, & Eschenfelder 2017). The new relationship offered participants better European input into study coding decisions, recognition of their efforts on ICPSR materials, joint efforts to secure external funding, and European reuse of the processed data for CD-ROM products.

Discussion and Conclusion

Even this brief history points to some concerns that data archives continue to attend to. For one, data do not move seamlessly, and never have; geographical and material boundaries and distances matter. For much of this period in discussion it was difficult to physically move data (even with the advent of the TCP/IP protocol and FTP).

National and regional boundaries are obvious boundaries of note. Local and regional practice and law shape the curatorial process (consider a contemporary example: the 2017

introduction of the General Data Protection Rule in Europe). A related challenge faced by European data organizations in promoting data awareness and exchange was language. Scholars were not able to judge the value of materials described in another language even if the data itself were accessible. Ideally individual data organizations would translate all their study descriptions, study questions, and other study materials into other European languages. But a great deal of local material is created only in local languages and there have never been enough resources to translate all of it.

These challenges and tensions experienced by European and US archives in arranging the movement of data across borders in the past are still an issue today. While our findings are drawn from the social sciences field, the tensions from these themes are likely applicable to data curation in a variety of scientific fields. For example, Sands et al. (2012) examine "data flow" in astronomers' publications and McNally et al. (2011) describe similar work to examine several data-intensive disciplines (gene sequencing and sensor-based environmental science).

This history illustrates some of the many challenges and tensions experienced by European archives in arranging the movement of data across borders that are potentially still issue today even as they may play out slightly differently. Data territoriality, awareness of available datasets, credentialing for data creators and requesters, and the technical dimensions of curation are part of the whole lifecycle of data that are familiar to the staff and management of research data repositories. However, since funders may be unaware of the level of work and coordination needed and thus unwilling to pay for that "invisible" labour, data organizations have evolved flexible business models to get that work done and complex mechanisms of collaboration and coordination with other DO.

This paper's contribution is two-fold. First we explore data frictions at the inter-organizational and the supra-organizational levels. Specifically, we focus on the role of national boundaries and their profound influence on what data is acquired/managed. Secondly, we examine data friction both as resistance, but also as an opportunity for ordering. We identify the multiple resistances that influence the movement of data between DO and across national borders. We also identify how organizations structure their understandings and activities to support working across boundaries, and how these orderings become institutionalized in shared expectations, norms and supra-organizational governance documents. We show that new knowledge infrastructures arise and existing ones are re-configured because data organizations acknowledge their points of difference and position themselves to think "with" them.

Acknowledgements

The authors wish to acknowledge the financial support of the Alfred P. Sloan Foundation, Wisconsin Alumni Research Foundation, and the University of Wisconsin-Madison School of Library and Information Science Sarah M. Pritchard Faculty Support Fund.

References

Baker, K. S., Mayernik, M. S., Thompson, C. A., Nienhouse, E., Williams, S., & Worley, S. (2015). Poster presented at IDCC15 International Digital Curation Conference, London, 9-12 February 2015. https://www.dcc.ac.uk/events/idcc15/posters

Bates, J. (2018). The politics of data friction. Journal of Documentation, 74(2), 412-429.

Bates, J., Lin, Y. W., & Goodale, P. (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. Big Data & Society, 3(2), 2053951716654502.

Beer, D. (2013). Popular culture and new media: The politics of circulation. Basingstoke: Palgrave Macmillan.

Bertot, J. C., & Choi, H. (2013, June). Big data and e-government: issues, policies, and recommendations. In Proceedings of the 14th Annual International Conference on Digital Government Research (pp. 1-10). ACM.

Borgman, C. L., Wallis, J. C., & Mayernik, M. S. (2012). Who's got the data? Interdependencies in science and technology collaborations. Computer Supported Cooperative Work (CSCW), 21(6), 485-523.

Borgman, C. L., Scharnhorst, A., Van den Berg, H., Van de Sompel, H., & Treloar, A. (2015, November). Who uses the digital data archive?: an exploratory study of DANS. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (p. 96). American Society for Information Science.

Bréchon, P. (2009). A breakthrough in comparative social research: the ISSP compared with the Eurobarometer, EVS and ESS surveys. In The International Social Survey Programme 1984-2009 (pp. 50-66). Routledge.

Chenail, R. J. (2012). Conducting qualitative data analysis: Reading line-by-line, but analyzing by meaningful qualitative units. The Qualitative Report, 17(1), 266-269.

Edwards, P. N. (2010). A vast machine: Computer models, climate data, and the politics of global warming. The MIT Press.

Eschenfelder, K. R., Shankar, K., Downey, G., & Lin, R. (2015). Sustainability of social science data archives: A historical network perspective. IASSIST 2015.

Inglehart, R., & Reif, K. (Eds.). (2016). Eurobarometer: The Dynamics of European Public Opinion Essays in Honour of Jacques-rené Rabier. Springer.

Jirotka, M., Lee, C. P., & Olson, G. M. (2013). Supporting scientific collaboration: Methods, tools and concepts. Computer Supported Cooperative Work (CSCW), 22(4-6), 667-715.

Karasti, H., Baker, K. S., & Millerand, F. (2010). Infrastructure time: Long-term matters in collaborative development. Computer Supported Cooperative Work (CSCW), 19(3-4), 377-415.

Law, J., (1994). Organizing modernity. Oxford: Blackwell.

Mayernik, M. S. (2016). Research data and metadata curation as institutional issues. Journal of the Association for Information Science and Technology, 67(4), 973-993.

McNally, R., Mackenzie, A., Hui, A., & Tomomitsu, J. (2012). Understanding the 'intensive'in 'data intensive research': Data flows in next generation sequencing and environmental networked sensors. International Journal of Digital Curation, 7(1), 81-94.

Paine, D., & Lee, C. P. (2014, October). Producing Data, Producing Software: Developing a Radio Astronomy Research Infrastructure. In IEEE 10th International Conference on e-Science , Vol. 1, pp. 231-238. IEEE.

Ribes, D. (2017, February). Notes on the concept of data interoperability: Cases from an ecology of AIDS research infrastructures. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 1514-1526). ACM.

Ribes, D., & Finholt, T. A. (2009). The long now of infrastructure: Articulating tensions in development. Journal of the Association for Information Systems: Vol. 10: Iss. 5, Article 5. doi: 10.17705/1jais.00199

Ribes, D., Jackson, S., Geiger, S., Burton, M., & Finholt, T. (2013). Artifacts that organize: Delegation in the distributed organization. Information and Organization, 23(1), 1-14.

Ribes, D., & Polk, J. B. (2014). Flexibility relative to what? Change to research infrastructure. Journal of the Association for Information Systems, 15, 287.

Sands, A., Borgman, C. L., Wynholds, L., & Traweek, S. (2012). Follow the data: How astronomers use and reuse data. Proceedings of the American Society for Information Science and Technology, 49(1), 1-3.

Williams, R. D., Shankar, K., & Eschenfelder, K. R. (2017). Two views of the data documentation initiative: Stakeholders, collaboration and metadata standards creation. Proceedings of the Association for Information Science and Technology, 54(1), 455-462.

White, J. M. (2017). Following data threads. In R. Kitchin, T.P. Lauriault, & G. McArdle (Eds.), Data and the City. Abingdon: Routledge, pp. 85-97.

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. doi:10.1038/sdata.2016.18