2023

Infrastructuring Open Science

Exploring RDM challenges and solutions for qualitative and ethnographic data

Dissertation zur Erlangung des Grades eines

Dr. rer. pol. an der Fakultät III

Wirtschaftswissenschaften, Wirtschaftsinformatik und Wirtschaftsrecht

der Universität Siegen

Vorgelegt durch

MA, Gaia Mosconi

Köln, Deutschland

Erstprüfer: Prof. Dr. Volkmar Pipek

Zweitprüfer: Prof. Dr. Aparecido Fabiano Pinatti de Carvalho

Dekan der Fakultät III: Prof. Dr. Marc Hassenzahl

- 2023 -

Acknowledgments

I would like to express my deep appreciation to my advisors, Volkmar Pipek and Aparecido Fabiano Pinatti de Carvalho, for their invaluable guidance, support, and encouragement throughout the process of writing this thesis. Their insightful feedback, constructive criticism, and expertise have been instrumental in shaping my research and enhancing its quality. I would also like to thank my family and friends for their unwavering love, encouragement, and support throughout my academic journey. Their understanding, patience, and belief in me have been a constant source of motivation and inspiration. A special thanks to Timo Kaerlein a close friend and a scholar who proofread a couple of my manuscripts, helped me push forward the work in the INF project, and supported me in the darkest times. Furthermore, I am grateful to my colleagues from Locating Media, the Collaborative Research Centre and the Department of Socio-informatics for their intellectual stimulation, and helpful discussions. Finally, I would like to extend another special thanks to my tutors, Dave Randall and Helena Karasti, for their exceptional guidance and mentorship throughout this research project. Their feedback and suggestions have been immensely valuable in shaping my ideas and enhancing the overall quality of my work; their wisdom, patience, and kindness have been invaluable, and I feel fortunate to have had the opportunity to work with them who I consider not only inspiring scholars but also good friends. Once again, I am deeply grateful to all those who have supported me in completing this thesis, and I acknowledge that their contributions have been essential to its success.

Abstract

In the last two decades, research data became to be recognized as an independent product in its own right and incrementally became more visible among policy makers, funding agencies and various academic stakeholders. In fact, driven by the Open Science agenda which aims "at making scientific research and *data* accessible to all", Open Research Data has become an important and desirable outcome of publicly funded research. This is proven by the increasing attention and specific funding schemes worldwide targeting the establishment of Research Data Management (RDM) policies and Research Data Infrastructures, to be developed according to the FAIR (Findability, Accessibility, Interoperability, and Re-use) data principles.

This dissertation takes these institutional and infrastructural developments as a point of departure and presents a long-term ethnographic account of the socio-technical challenges involved in translating the Open Science *grand vision* and related Research Data Management policies into practices. Since 2016, I have participated and carried out research in an information infrastructure project (INF) connected to a Collaborative Research Centre (CRC) composed by 14 interdisciplinary projects and funded by the German Research Foundation (DFG). The DFG expects all its funded projects, from all disciplines and research fields, to follow RDM policies and guidelines. Therefore, the aim of the INF project is to support the development of RDM practices, infrastructural solutions, and concepts which all together should lead to the curation, long-term preservation, sharing, and potential reuse of research data in our CRC. The focus of my study targeted specifically interdisciplinary research projects who apply mainly qualitative and ethnographic methods as data collection, being the majority in our context. For these types of methodological approaches, mainly applied by Humanities and Social Sciences (HSS) disciplines, the requirements for data management are relatively new, and only few technological aids and infrastructures have been developed thus far to specifically support the management of these sensitive and personal data characterized by additional epistemological, methodological and ethical challenges.

With my research, I went beyond the institutionalisation of research infrastructure and rather investigated scientific research practices 'on the ground'. By following an *infrastructuring* approach in synergies with previous work, my research proposes a shift from designing systems as fixed artefacts, to designing them as ongoing infrastructures, as a way of building socio-technical processes able to relate different contexts (institutional and practical) and create new (social-technical) relationships 'from within'. At the centre of our infrastructuring work, I

locate a socio-technical platform called ‘Research-hub’, established to customize, test, and study new RDM concepts and workflows expected to be implemented by INF in the long-term. Research-hub represents the socio-technical anchor point of the infrastructuring work undertaken but also stands as an example of a small scale and local research data infrastructure ‘in the making’.

The thesis outlines a vision to achieve RDM practices and workflows with a specific attention to curation and sharing practices in the CRC’s, an interdisciplinary and ethnography-driven research context, and reports on how we started to promote a bottom-up collaborative sharing culture essential to putting into practice the Open Science agenda and the practical implementation of RDM policies. For this purpose, a design concept, called Data Story, has been designed and iteratively evaluated through what I call 'embedded evaluation’ meaning that evaluation opportunities spontaneously emerged from my double role and my ongoing engagement in the field. In fact, since 2016 I have been members of the CRC myself, so I was part of the context I was called to design for (and with). Therefore, the thesis presents ‘embedded evaluation’ a methodological approach that can be fruitful to the CSCW and HCI communities, specifically for those projects engaged in infrastructuring, where the researcher carries multiple roles in the field (member of the community, researcher, designer) and where it is not possible to draw demarcations between investigative, design and evaluative work. Conceptually, I contribute to the expansion of the term ‘articulation work’ by illustrating how designing for RDM imply to support the work for future cooperation not yet known. In fact, the overall premises of the OS are based on this I ties assumption: researchers need to curate and open the data for other people to use and maybe even collaborate with them but there is no a predefined and clearly demarcated audience, and researchers don’t even know yet if sharing will happen at all. We demonstrate how the design concept Data Story supports this particular kind of articulation work, one that we called ‘anticipatory’, which is essential to develop coordination mechanics for the future cooperation that might occur. To conclude, my contribution is to provide approach for RDM and for new collaborative research data practices, capable to negotiate between top-down policies and bottom-up practices that can be sustainable and evolve over time. Qualitative research, as I will argue, is well suited to an understanding of emergent phenomena of this kind.

Table of Content

Part III: Discussion and conclusion

List of publications

No.	Publication	Venue	Status
P1	Mosconi, Gaia, Qinyu Li, Dave Randall, Helena Karasti, Peter Tolmie, Jana Barutzky, Matthias Korn, and Volkmar Pipek. "Three gaps in opening science." *Computer Supported Cooperative Work (CSCW)* 28 (2019): 749-789.	**JCSCW**	**Published**
P2	Mosconi, Gaia, Dave Randall, Helena Karasti, Saja Aljuneidi, Tong Yu, Peter Tolmie, and Volkmar Pipek. "Designing a Data Story: A Storytelling Approach to Curation, Sharing and Data Reuse in Support of Ethnographically-driven Research." *Proceedings of the ACM on Human-Computer Interaction* 6, no. CSCW2 (2022): 1-23.	**American CSCW**	**Published**
P3	Mosconi, Gaia, Helena Karasti, Dave Randall, and Volkmar Pipek. "Designing a Data Story: An Innovative Approach for the Selective Care of Qualitative and Ethnographic Data." *Media in Action| Volume 3* (2022): 207.	**Book Chapter**	**Published**
P4	Mosconi, Gaia, Aparecido Fabiano Pinatti de Carvalho, Hussain Abid Syed, Dave Randall, Helena Karasti, Volkmar Pipek. "Fostering Research Data Management in Collaborative Research Contexts: Lessons learnt from an 'Embedded' Evaluation of 'Data Story'." *Computer Supported Cooperative Work (CSCW), 1-39.*	**JCSCW**	**Published**

Table 1: list of publications included in this thesis.

Part I Foundations

The first part of my dissertation includes the structural and conceptual foundations. Chapter 1 (Introduction) introduces to the research field and presents the overall research questions and main contributions. Chapter 2 (Related Work) presents the theoretical foundations and situates the research objectives of this thesis within the relevant research discourse. Chapter 3 (Research Approach) outlines the research approach and methodology I conducted within this thesis. Chapter 4 (Platform development) outlines the vision for RDM that my colleagues and I developed around the collaborative platform Research-hub and explains different modules and design concepts built over the years.

Introduction

The development of the World Wide Web initiated several revolutions impacting all aspects of our society. The Sciences and the overall of field of production of knowledge are also experiencing rapid and drastic changes intensified with the digitalization. Over the past few years, new terms were coined to define the emergent aspects of the contemporary science: Cyberinfrastructure, eScience, eResearch, Science 2.0, Digital Humanities, Mode2, Open Science or Open Research, all umbrella terms that emphasize various aspects of the second scientific revolution (Fecher and Friesike 2014). Open Science, more than the others terms, is shaping the debate related to the future of academia and knowledge creation; in fact, Open Science is not only a term but it is an active movement that aims "at making scientific research and *data* accessible to all". Since the beginning of the year 1990, the movement promoted in academia the adoption of practices to make science more transparent and accessible during the whole research process. The Open Science movement elaborated historical statements and principles[1] explicitly addressing EU commissions, local governments, and research funders with the intention of influencing the political debate and trying to regularize licenses and disclosure for scientific literature (Open Access) and above all scientific data (Open Research Data). In particular, Open Research Data is considered crucial to meet the OS agenda and it is promoted for achieving three major goals: (1) to promote the reuse of data in new interdisciplinary contexts, (2) to ensure verifiability and good scientific practice, and (3) to provide greater returns from the public investment in research (OECD 2007; Christine L. Borgman 2015). In order to promote Open Data in academic contexts, Research Data Management (RDM) has been established as a pragmatic field which aims at studying the movement of data throughout their life cycle in order to ensure their long-term preservation, shareability and reuse. RDM, in itself, is a complex and long-term endeavor spanning the entire research lifecycle and beyond, requiring attention to the specifics of data creation, curation, storage, sharing and reusability (Treloar and Harboe-Ree 2008; Whyte and Tedds 2011) which are different practices but at the same time intertwined.

In the last twenty years, libraries, data centres and other institutions started to increasingly collaborate, build partnerships, and define policies and build up information infrastructures to support scientists in the handling of their data and to promote the development of RDM

[1] the Budapest Open Access Initiative in 2001, the Panton Principles in 2009, the Amsterdam Call for Action on Open Science presented to the Dutch Presidency of the Council of the European Union in May 2016. (Searched on date 22.09.2018)

practices (Pampel and Dallmeier-Tiessen 2014; Reilly 2012; Corti 2013). At the same time, many funding bodies started to mandate the creation of data management plans and the open access publication of the research data gathered in their funded projects. Knowing how to create a data management plan and how to efficiently manage data has become a sine qua non condition for receiving research funding from prestigious funding agencies both at national and European level. Quite recently in 2021, the United Nations has also publicly joined the support with a recommendation to implement OS worldwide (Leonelli 2022). In response to these institutional demands, we have seen the emergence of numerous general-purpose data repositories, at scales ranging from institutional (for example, a single university), to open globally-scoped repositories[2]. Examples of emerging data registration systems include Dryad, Dataverse, openICSPR, and Figshare. In 2016, stakeholders from academia, industry, publishers, and funding agencies published a concise and measurable set of principles called the FAIR Data Principles (Findable, Accessible, Interoperable and Re-usable) which should be respected when developing research data infrastructure. To highlight the importance of keeping data FAIR, the European Commission adopted the FAIR Data Principles and released new Guidelines on FAIR Data Management in Horizon 2020 (Commission 2016). The EC guidelines include several important changes that aim to improve the quality of project results, achieve greater efficiency, and achieve progress and growth of a transparent scientific process. Despite all these political efforts in pushing forward polices, developing standards, building new infrastructures, and sustaining a cultural change in academia, many disciplines are far from achieving the OS goals and the implementation of RDM practices are still an unsolved issue. For example, in Humanities and Social Sciences (HSS), collaborative and data-intensive research endeavours, the plurality of research methods, standards and traditions, ethical and legal implications, and heterogeneous practices in storing, processing, sharing and analysing data indicate higher barriers to the implementation of OS initiatives (Eberhard and Kraus 2018; Korn et al. 2017; Mosconi et al. 2019). Another layer of complexity in RDM is added by the overhead (additional work, time and costs) implied in the appropriation of data curation and the sharing practices which require researchers to engage in systematic organization of data (i.e. metadata creation, contextualization and structuring the storage of data) in on-going research projects and in anticipation of reuse, verifiability, and collaboration. The overall

[2] Dataverse, FigShare (http://figshare.com), Dryad, Mendeley Data (https://data.mendeley.com/), Zenodo (http:// zenodo.org/), DataHub (http://datahub.io), DANS (http://www.dans.knaw.nl/), and EUDat. The following digital repository systems are used by social science data archives and may be implemented locally, though they are not open source and may involve payment. They offer a range of data management and online data analysis features.

premises of the OS, in fact, are based on this assumption: researchers are required to curate and open their data for other people to use, but there is no clear audience, and primary researchers do not know in advance what kinds of re-use (or a potential collaboration) might take place, or indeed will happen at all.

In response to some of the issues mentioned above, this thesis investigates how to promote RDM practices in collaborative research contexts in which researchers apply mainly qualitative and ethnographic research methods. The interest of my work lies in understanding how Open Science objectives can become *a practice* acknowledging the needs of researchers in their everyday activities. Since 2016 I have been a member of a Collaborative Research Centre and worked in an Information Infrastructure project called INF funded by the German Research Foundations (DFG). Through this 'embedded' engagement, I investigated research data practices on the ground with attention to how data is organized and transformed along the research process. I soon realized that I became the medium through which meanings emerged and negotiations between institutional points of view and actual practices took place. In parallel, I have established a collaborative platform in which to develop solutions to support the ongoing management, curation and sharing of research data based on the insights gathered in the field. Specifically, a design concept called Data Story has been developed which aimed at supporting the ongoing curation, sharing and potentially reuse of qualitative and ethnographic research data in interdisciplinary contexts. Through the evaluation of this design concept, I outline in this thesis the theoretical concept of 'anticipatory articulation work' which characterize RDM (data curation and sharing work) in particular.

1.4 Research Goals and Research Questions

I advocate for 'infrastructuring' Research Data Management (with specific focus on data curation and sharing practices): which means to apply a socio-technical approach to this field and consider it as an ongoing iterative effort which needs to be negotiated and understood 'in the wild' in constant dialog with researchers in the field. The aim of my work is not to provide and study post-hoc solutions, like a repository or a database, where at the end of the process researchers are expected to upload research data but I wished to directly support data curation, sharing and long-term preservation practices and integrate them within the research process. Put it differently, the aim is to investigate how to promote a bottom-up collaborative sharing culture. Overall, this research aims to investigate and design a new research data infrastructure for research collaboration, in particular to support the development of data curation and data sharing practices in interdisciplinary research contexts. The approach I followed was to integrate bottom-up and top-down strategies in the development of research infrastructures in order to identify design principles that might facilitate the realization of the Open Science agenda in a participatory and sustainable way especially for the researchers impacted by it.

This research is guided by the following questions:

RQ: 1) What are the socio-technical challenges for the development and appropriation of RDM practices (preservation, curation and data sharing) in qualitative ethnographically-driven research contexts?

RQ: 2) How can we design tools and infrastructures to support the establishment of RDM practices in qualitative and ethnographically-driven research contexts?

RQ: 3) In what ways can infrastructuring support the development of new data practices (first and foremost curation and sharing) and eventually lead to data re-use across different disciplines?

To answer to these research questions, I have been an 'embedded researcher' since 2016. I performed qualitative interviews, (participatory) ethnographic observations with fellow colleagues from the Collaborative Research Centre. In the first year, I set up an offline space called "Research Tech Lab" for face-to-face events, where interdisciplinary scholars were invited to participate in open discussions about research data practices. Simultaneously, an online space has been deployed – Research-hub – a platform (in ongoing development) that support research collaboration and bring the offline discussion into a digital and distributed space. The platform ambition is to eventually support RDM practices across different research communities and it represents the socio-technological anchor point of the research infrastructure 'in the making'.

1.5 Summary of the contribution

In answering to the above research questions, the thesis makes three substantial contributions:

1. **Empirical contribution:** I present a long-term ethnographic account of the challenges that researchers face when confronted for the first time with Research Data Management policies which require them to engage with long-term preservation, curation and data sharing practices. My focus is a research context mainly composed

by researchers working in interdisciplinary projects and who apply qualitative and ethnographic methods and with no previous knowledge or experience in RDM.

2. **Conceptual and theoretical contribution:** designing for RDM imply to support a kind of articulation work one that I call 'anticipatory' meaning supporting not only articulation work in respect of current cooperation, but also the work for future cooperation not yet known.
3. **Design contribution:** through my long-term engagement I developed a vision for RDM workflow based on interconnected tools grounded in an Open-Source platform called Research-hub. Research-hub represents the socio-technical anchor point of the infrastructuring work undertaken but also stands as an example of a small scale and local research data infrastructure 'in the making'. In particular, the thesis explores a design concept called 'Data Story' which offers a means of enhancing and naturalizing curation practices through storytelling. I demonstrate how the Data Story concepts allows to negotiate between top-down policies and bottom-up practices, to support 'reflective' learning opportunities - with and about data - of many kinds and to develop coordination mechanics for the cooperation that will be.

1.3 Structure of the thesis

The thesis is structured as follows. In the Chapter 2 I will discuss existing related work starting with literature on Open Science and RDM institutional approaches. I will then move on to highlight practical challenges researchers face when dealing with RDM especially from previous literature in CSCW. Chapter 3 presents briefly the context, the methodology underlying this work, my positionality and two major conceptual influences connected to my work: articulation work and infrastructuring. Chapter 4 presents the platform development and the RDM vision which emerged from the customization of Research-hub platform. Chapters 5-9 are the core of the dissertation. Each chapter represents a publication, providing insights into the challenges involved in the RDM for qualitative and ethnographic data and suggest possible solution for these challenges. Chapter 11 closes the dissertation and present the discussion and conclusion.

Related work

This chapter introduces the related work connected to my thesis. I start by introducing the Open Science grand vision as a political ambition which has been widely promoted and whose realization is connected to the practical implementation of Open Research Data in all disciplines and fields as well as, crucially, in interdisciplinary contexts. I move on by introducing institutional models of Research Data Management (RDM) where RDM is presented as a normative and prescriptive field which aims at translating the OS and Open Research Data political ambitions into reality. CSCW has been one of the few areas that has pointed to some of major challenges researchers face when confronted with RDM, and hence I follow this with an examination of that literature. Some of these challenges exist regardless of the particular research domain under consideration, while others are specific of the methodological and epistemological characteristics connected to qualitative and ethnographic data which is the focus of my work. Then I present some challenges connected to the lack of tools and infrastructures that can support researchers in appropriating RDM practices. Finally, I conclude the chapter with literature on information infrastructure and infrastructuring which highlight a processual and relational perspective on infrastructures development grounded on the observation of local practices and ongoing design activities performed with direct involvement of research participants. An infrastructuring approach is needed in order to support new RDM practices not yet in place.

2.1 Open Science and Open Research Data

The digitalization of information at scale has had profound consequences for the conduct of scientific activity. It has been suggested that we are experiencing the emergence of the 4th paradigm in science, based on data-intensive scientific discoveries (Hey, Tansley, and Tolle 2009). The organizational, cultural, and infrastructural transformations happening within the academic landscape and guided by governments, funders, and research institutions worldwide have been characterised in a variety of ways. Among these are Science 2.0, Cyberinfrastructure and eScience. Recently however, 'Open Science' has become the preferred term, chosen after a public consultation by the European Commission, to address this putative transformation of scientific practices (practices which, in fact, have been underway in some form since the 1990s). Open Science does not have a fixed definition, but it is rather an "an umbrella term that encompasses almost any dispute about the future of knowledge creation and dissemination" (Fecher and Friesike, 2014, p.17). Nevertheless, increased efficiency, impact, transparency, verifiability, and accessibility of knowledge are core values and driving forces of Open Science (Fecher an Friesike, 2014).

In 2016, the European Commission defined Open Science as "a new approach to the scientific process based on cooperative work and new ways of diffusing knowledge by using digital technologies and new collaborative tools" (Commission 2016). From this point of view, OS is portrayed as a positive development for the academic landscape, closely linked to the digital transformation of scientific work. It is presented as an innovation that strictly depends on the availability of information, communication technologies and cooperative tools. OS, as a global phenomenon, aims at giving anyone interested in research unrestricted access to any investigation materials (publications, research data, or software) and at any point in time of an investigation, regardless of where they are and whether the interested person is a professional researcher or not. OS is therefore promoted to facilitate equity, sharing, and inclusion in the production and consumption of research by making previously inaccessible resources available to anyone interested in participating in research. A belief supporting this view is that principles such as collaboration, transparency, reproducibility, and openness are constitutive of good scientific practice (Burgelman et al. 2019).

Open Science policies are increasingly being implemented by research institutions, research funders, governments, publishers and by the European Commission, which recognizes Open Science a key priority to be pursued by all its funded research projects (Commission 2016). Open Science policies are guiding massive infrastructural investments and political initiatives. In this regard, the European Commission has invested around €250 million in an initiative called European Open Science Cloud (EOSC) which aims at providing "European researchers, innovators, companies and citizens with a federated and open multi-disciplinary environment where they can publish, find and reuse data, tools and services for research, innovation and educational purposes[3]". This initiative started in 2015 but was officially launched in 2018 where all the developed services were made available through the EOSC portal (https://eosc-portal.eu/) which is expected to be used by all research projects funded by the EU. At the political and organisational level, of relevance the European Strategy Forum on Research Infrastructures (ESFRI) established in 2002, with a mandate from the EU Council to support coherent strategies on research infrastructures in Europe, "and to facilitate multilateral initiatives leading to the better use and development of research infrastructures, at EU and

[3] https://eosc-portal.eu/about/eosc

international level[4]”. While at national level, for example in Germany, where the research reported on below is based, the NFDI[5] (National Research Data Infrastructure) represents the biggest infrastructural funding scheme, promoted by the German Research Foundation (DFG). In October 2020 the DFG funded eighteen consortia targeting a variety of disciplines with the goal of systematically managing scientific and research data, providing long-term data storage, backup and accessibility, networking the data both nationally and internationally and providing science-driven data services to research communities. Besides these recent infrastructural developments, which are trying to centralize data-driven services and infrastructures, in the last twenty years we have seen the proliferation of data centres and numerous general-purpose data repositories, at scales ranging from the institutional (e.g., a single university), to community-driven repositories, to the globally scoped.

In fact, a central role for achieving the OS ambitions is given to Open Research Data. Open Data are generally defined as “data freely available on the public Internet permitting any user to download, copy, analyse, re-process, pass them to software or use them for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself” (Panton Principle, Open Research Glossary). The central arguments in support of Open Research Data are: a) the possibility to reuse data in innovative ways across disciplines; b) the verifiability it guarantees for ensuring good scientific practice; and c) to provide greater returns from the public investment in research (OECD 2007; Christine L. Borgman 2015). In 2016, a group of various stakeholders from academia, industry, publishers and funding agencies even published a concise and measurable set of principles called the FAIR Data Principles, where research data are expected to be Findable, Accessible, Interoperable and Re-usable. The FAIR principles are intended to be implemented as a process with the ultimate goal of reuse in mind:

- **Findable**: Data need to be found by anyone interested in them. To do that metadata need to be created in human and machine-readable forms essential for automatic discovery of datasets and services. To achieve that (meta)data need to be assigned a globally unique and persistent identifier and registered or indexed in a searchable resource.
- **Accessible**: Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation. Here (meta)data

[4] ESFRI website: (https://www.esfri.eu/esfri-roadmap). Accessed 20 November 2022.
[5] NFDI: https://www.dfg.de/en/research_funding/programmes/nfdi/index.html

are retrievable by their identifier using a standardised communications protocol which is open, free, and implementable.

- **Interoperable**: The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing. Therefore, (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- **Reusable**: To be reused metadata and data should be well-described so that they can be replicated and/or combined in different settings. To achieve that (meta)data need to be richly described with a plurality of accurate and relevant attributes, released with a clear and accessible data usage license, and finally associated with detailed provenance and meet domain-relevant community standards[6].

These principles are now leading the development of research data infrastructures at various level (national, European and worldwide). Open Science and Open Research Data ambitions aim to address not only traditional sciences but also Humanities and Social Sciences (HSS). In fact, all disciplines and fields are now expected to submit Data Management Plans (DMP) as a prerequisite to receive research funding from all major funding agencies and where researchers are asked to specify their concrete intentions for long-term archiving, data sharing and re-use. The European Commission emanated two important reports concerning the establishment of Open Research Data. The first one was published in 2012 where European member states were explicitly requested to ensure that research data when funded with public budget "become publicly accessible, usable and re-usable through digital e-infrastructures" (EC - European Commission and Kroes 2012). In another report published in 2018, the EC further elaborate on the cultural change expected to be pushed forward in the academic contexts and guided by the FAIR principles: "A holistic approach is required, with due attention paid to creating a culture of FAIR, to the needs and priorities of particular research communities and to the technical ecosystem that enables FAIR data and services. [...] The wider FAIR ecosystem must support disciplinary standards while also ensuring to the greatest degree practical that data will be FAIR across traditional disciplines and also in emerging interdisciplinary research areas" (Collins et al. 2018). The ambition is to maximize access and interdisciplinary reuse of research data generated by the publicly funded projects where eventually, in the long-term, data will be opened by default following the principle "as open as possible, as closed as necessary" (Landi et al. 2020) which means balancing openness and protection of scientific information,

[6] Source: https://www.go-fair.org/fair-principles/ searched on date 21.01.2023

commercialization and Intellectual Property Rights, privacy concerns and security (Bfurgelman et al 2019). The EC recognizes that "one size does not fit all" (Open Science Policy platform 2018), but nevertheless, as Leonelli has pointed out "national and international policies tend to implement OS guidelines, tools and principles in a top-down manner and across domains, with some attention paid to disciplinary cultures but no fine-grained consideration of the diverse capacities, motivations and methods characterising different epistemic communities" (Leonelli, 2022, p.4). The same top-down approach with normative and prescriptive characteristics can be found in the institutional models of Research Data Management which are being promoted by data centres, universities libraries and research institutions worldwide as a way to encourage 'good scientific practices' in all fields and disciplines and which should lead to the successful management of research data needed to fulfil the political ambitions for Open Science and Open Research Data.

2.2 Research Data Management and institutional models: academic "best practices"

Over the last twenty years, libraries, data centres and other research institutions have increasingly started to collaborate, build partnerships, define policies and build up information infrastructures in pursuit of OS goals (Oßwald and Strathmann 2012; Reilly 2012; Pampel and Dallmeier-Tiessen 2014). In this context, Research Data Management (RDM) has been established as a normative and prescriptive field which aims at translating the OS and Open Research Data political ambitions into reality by defining requirements and encouraging academic best practices. Research Data Management is commonly defined as "the organization of data, from its entry to the research cycle through to the dissemination and archiving of valuable results" (Whyte and Tedds, 2011, p.1). RDM is characterised by several core practices, such as data curation, metadata documentation, long-term archiving, and data sharing altogether leading to the publishing and successful reuse of research data. They are all different set of practices but strictly intertwined.

One of the most recognized institutional models of RDM is characterized by the 'data lifecycle' which was first developed by a Digital Curation Centre (DCC) in the UK. The model describes an idealized research process that already incorporate the shareability and reuse of data in the process itself which is considered the goal of RDM. The DCC provided a high-level overview of the RDM stages that was later simplified and adapted by other Data Centres and institutions across the globe.[7] The UK Data Archive suggested the data lifecycle could be modelled in six

[7] For the original see: http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf

different stages, in which certain practices and tasks arise and vary in size depending on the field of application (see Figure 1).

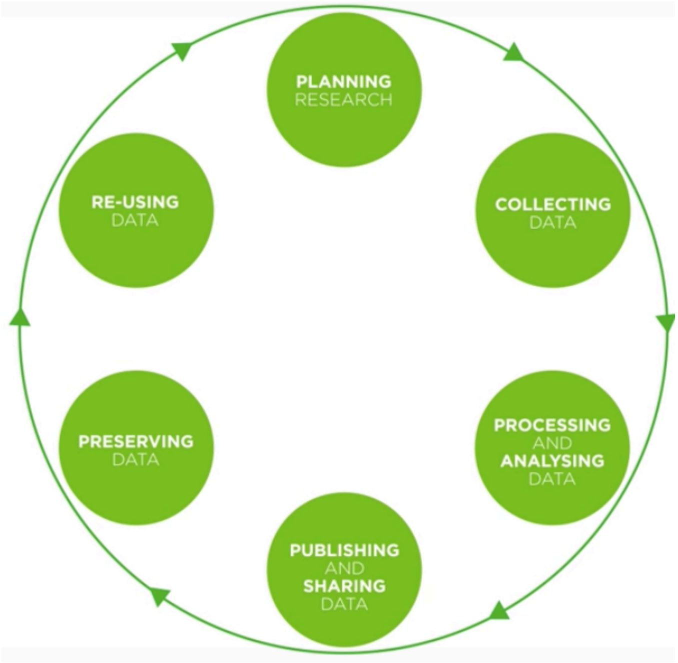


Figure 1: The Data Lifecycle

(1) **planning research**: design research, plan data management, plan consent for sharing, plan data collection, process protocols and templates; explore existing data sources;

(2) **collecting data**: collect data; capture data with metadata; acquire existing third-party data;

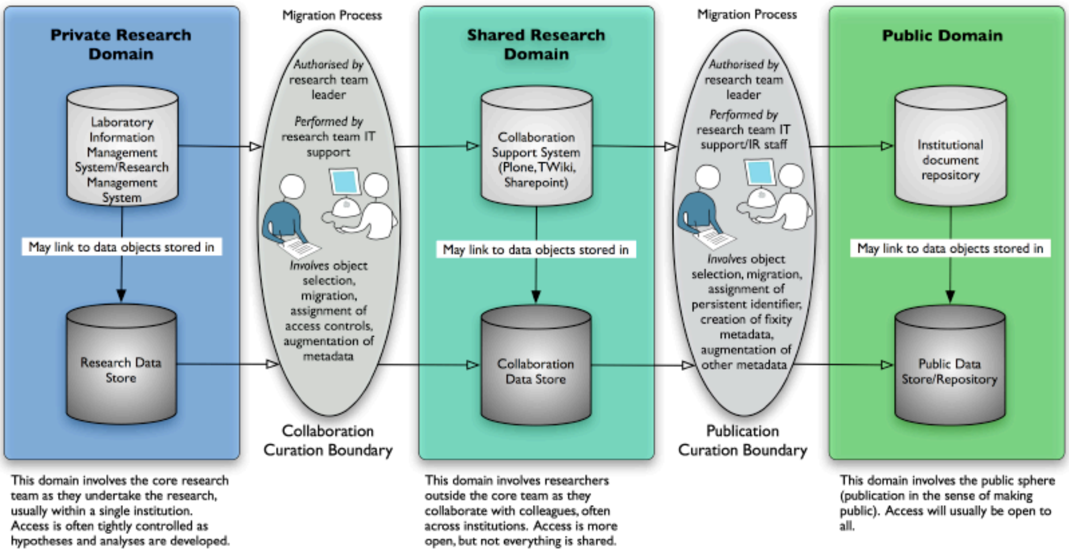
(3) **processing data and analysing data**: enter, digitize, transcribe and translate data; check, validate, clean, anonymize; derive data; describe and document data; manage and store data; analyse and interpret data; produce research outputs; cite data sources;

(4) **publishing and sharing data**: establish copyright; create user documentation; create discovery metadata; select appropriate access to data; publish/share data; promote data

(5) **preserving data**: migrate data to best format/media; store and back up data; create preservation documentation; preserve and curate data;

(6) **re-using data**: conduct secondary analysis; undertake follow-up research; conduct research review; scrutinize findings; use data for teaching and learning.

This abstract model clearly highlights what constitutes the RDM 'best practices' which should take place at different stages of the data life cycle. However, it fails, I would argue, to the provide a good representation of the socio-technical and collaborative infrastructure in which researchers actually engage in the business of RDM. In this sense, 'the Data Curation Continuum' (Treloar and Harboe-Ree 2008) constitutes a more elaborated 'institutional' model which was developed between several Australian universities (led by Monash University[8]). It describes the various domains in which research data are expected to migrate during their life cycle, the actors involved in each domain and the curation boundaries data needs to cross in order to be made publicly available. This model distinguishes between the private domain, where researchers store and organize data for their own purposes, a shared research domain, where researchers might use a variety of tools to exchange and share data with collaborators or partners, and the public domain, where a repository stores data in a relatively permanent and standard form. The aim is the successful publication of digital objects in the public domain at the end of the data life cycle and the public dissemination of valuable results which will be accessed for data re-use.


Private Research Domain
Laboratory Information Management System/Research Management System
May link to data objects stored in
Research Data Store
Migration Process
Authorised by research team leader
Performed by research team IT support
Involves object selection, migration, assignment of access controls, augmentation of metadata
Collaboration Curation Boundary
Shared Research Domain
Collaboration Support System (Plone, TWiki, Sharepoint)
May link to data objects stored in
Collaboration Data Store
Migration Process
Authorised by research team leader
Performed by research team IT support/IR staff
Involves object selection, migration, assignment of persistent identifier, creation of fixity metadata, augmentation of other metadata
Publication Curation Boundary
Public Domain
Institutional document repository
May link to data objects stored in
Public Data Store/Repository

This domain involves the core research team as they undertake the research, usually within a single institution. Access is often tightly controlled as hypotheses and analyses are developed.

This domain involves researchers outside the core team as they collaborate with colleagues, often across institutions. Access is more open, but not everything is shared.

This domain involves the public sphere (publication in the sense of making public). Access will usually be open to all.

Figure 2: Data Curation Continua. In Treloar et al., 2008, pg.6

Figure 2 shows how the migration process involves a combination of human and computer supported actions. As Treloar & Harbor-Ree (2008) put it: "Humans will need to make selection decisions and then use automated assistance to modify and augment the objects as they cross the curation boundary". However, "the process of ongoing curation in the public

[8] Monash has led other projects in this area, such as the institutional repository project (ARROW) and two projects on researcher workflow and data management (DART and ARCHER).

domain relies on provenance metadata that should have been captured during the research process" (Treloar & Harboe-Ree, 2008, p.7). Thus, the Data Curation Continuum model resonates with Jacobs and Humphrey's (2004) position where: "Data archiving is a process, not an end state where data is simply turned over to a repository at the conclusion of a study", and which should include "the creation and preservation of accurate metadata" and where "such practices would incorporate archiving as part of the research method."

The Data Curation Continuum was updated in 2019 into what Treloar & Klump (2019) called the 'Object Curation Continuum'. In the second version, the authors added a set of activitiy layers taking place at at each curation boundary in order to clarify which activities should take place, where and what consequences they imply for the process. These include:

- The **object layer** which shows a variety of research objects (data, models, workflows, software, publications, documentation), and how they decrease in number as the result of a process of intentional selection (from left to right).
- The **storage layer** which distinguishes between discrete storage (different for each domain) and cloud storage (contiguous across each domain) where different combinations of storage are also possible, i.e.: local storage and cloud storage, or even three different cloud storage solutions (one for each domain).
- The **context layer** shows the way in which object context is added as the object(s) transition across the boundaries. This reflects the way in which tacit knowledge needs to be made explicit for audiences broader than the setting in which the objects are being created/used.
- The **provenance layer** can be viewed as just another kind of context, encoded in provenance metadata. The provenance information is added within the domain (by whatever systems for data management/generation are being used) and then simply migrated across the boundary transition.
- The **archival layer** shows the ways in which archival elements can be included in the object life cycle from the point of creation, rather than being added as an afterthought later on (Treloar and Klump, 2019, p.97)

In the institutional approaches, the data management challenge is conceptualized as a series of steps, each of which must be satisfactorily completed for the data to advance to the next step, with the ultimate goal of reusing it and maximizing its value. If the infrastructure for a given step does not exist or if the actors involved do not understand what is required, the data's full potential value cannot be realized (Wilson et al. 2011).

The goal of these models is to promote academic best practices of RDM - which involve the curation, sharing, reuse and long-term archiving of research data - to be achieved by expecting researchers themselves to provide contextual information, in the form of documentation and metadata, and select the appropriate data to be made available across different domains. In this sense, RDM implies that researchers must take on new responsibilities, tasks, and practices to be performed from the beginning of the research process. This ideal scenario, however, does not fully acknowledge how RDM really works in practice, the challenges that researchers encounter in managing and sharing research data and what that really means for them. As Wilms et al. (2018) put it: "most institutions are solely eager to openly share research data without answering researchers' interests, and thereby forget that the future of research data management lies within researchers' hands" (p. 4418). In fact, we are from achieving the results expected by funding bodies and other institutions in many research contexts and disciplines. Researchers still need to be supported with more tools and infrastructures that can connect the different activity layers and allow data to cross from one domain to another. It is this lacuna- the missing detail of practice and its contextual nature, which informs the work described in the thesis.

In the next section, I will highlight previous studies which have identified major challenges researchers face when confronted with RDM. The challenges illustrate the complexity of the issue at hand and call for a better understanding of research data management practices from the bottom-up and for new tools and infrastructural support that can be tailored around specific research communities and their data practices.

2.3 Unresolved challenges for RDM

Research on data management and data sharing are inextricably linked in the literature with an apparent emphasis on documentation, data sharing (Chin and Lansing 2004; Kervin, Cook, and Michener 2014; Tenopir et al. 2011) and reuse (Rolland and Lee 2013; Wallis, Rolando, and Borgman 2013; Faniel and Jacobsen 2010) which have been extensively addressed in the recent years in the CSCW literature and elsewhere.

RDM is clearly problematised by a series of challenges related first of all to contextualization and data documentation. As Koltay (2016) noted, documenting datasets is time-consuming, and researchers frequently disregard standards, conventions, and metadata when formatting their data. Contrary to what RDM models and Open Data policies recommend and expect, the reality is that many researchers do not budget adequate time for metadata generation and consider it a low priority task. This, as mentioned above, is because 'articulation work' is not seen as a primary activity. In fact, researchers are not yet compensated or rewarded for producing data products. They are evaluated for advancing the research field through scientific publications. As a result, many data collection activities are not targeted at sharing and archiving and the resulting products are not well documented or formatted for others to use (Kervin, Cook, and Michener 2014). However, even when documentation is provided, it is frequently the case that a significant portion of the knowledge needed to make sense of datasets is tacit (Birnholtz and Bietz 2003) and therefore has not been recorded in written form for others to understand it. It is not always the case that scientists can easily explain all contextual information necessary to allow someone else to comprehend their work. For example, Rolland and Lee (2013) investigated the data reuse practices of cancer-epidemiology postdocs and found out that even when researchers have direct access to all the documentation relating to original data, they still struggle to understand it and require additional information about the data at different stages of the research lifecycle. The postdocs employed several information seeking strategies, including conversations with their mentors and data managers.

To get access to contextual information and acquire a proper understanding of the data, Birnholtz and Bietz (2003) argue it is imperative to understand 1) the nature of the data, 2) the scientific purpose of its collection, and 3) its social function within the community that created it. Context also determines if something can be considered as data or metadata and the "degree to which those contexts and meanings can be represented influences the transferability of data" (Borgman 2015, p. 18). However, transferring and sharing data is not always a simple process. It's possible that a wide variety of tools and software programs are being utilized, which has implications for interoperability. Even in situations where the software being used is shared, there is still a risk that data could quickly become unreadable due to upgrades in software and hardware (Borgman 2012). Moreover, Borgman (2015) argues that the heterogeneity of the data being produced by a variety of research methodologies and fields results in the data being organized and displayed in a wide variety of unique and idiosyncratic ways. Cultural norms and values also play a role. Vertesi and Dourish (2011) found that the procedures used to generate and acquire data in scientific collaboration influence how the data is shared. The authors discovered a different sense of data ownership when they compared the cultures of two robotic space research teams. One team's communal and interdependent research strategy fostered the idea that data is owned by the group rather than an individual. On the other hand, the more independent research of the second group, in which researchers needed to compete for equipment, time, and resources, gave the impression that data is personally earned and so owned by individuals. Thus, Vertesi and Dourish (2011) argue that the circumstances of data sharing are connected to a broader sense of 'data economy', through which scientific data get produced, used and circulate, and these economies influence how researchers handle data sharing.

Issues of control and trust have also been highlighted. For example, researchers may hesitate to share their data due to the possibility of being publicly criticized for mistakes that others might find in the shared data collection (Birkbeck, Nagle, and Sammon 2022) and by lack of trust concerning what others might do with the shared data (Gupta and Müller-Birn 2018). In general, researchers seem to struggle to even imagine what others might do with their data and this influences the documentation of datasets to facilitate data reuse (Mayernik 2011).

Carlson and Anderson (2007) have noted that it is false to assume that "knowledge can easily and straightforwardly be disembedded from its producers and original contexts to become explicit data for temporally and geographically distributed re-users" (Carlson an Anderson, 2007, p.647). Drawing on an original observation by Bowker (2005), Gitelman (2013) points out that this is bound up with the fact that, 'raw data is an oxymoron'. Instead, "data produce and are produced by the operations of knowledge production more broadly. Every discipline and disciplinary institution has its own norms and standards for the imagination of data, just as every field has its accepted methodologies and its evolved structures of practice" (Gitelman, op.cit, p.3).

All the issues mentioned above exist regardless of the particular research area under consideration. In the case of HSS, however, where qualitative and ethnographic methods prevail, the problem is even more complex.

2.4 The additional challenges for qualitative and ethnographic data

The studies mentioned above have mainly focused on computation and/or data intensive research endeavours in scientific domains like natural science and other fields that rely on highly structured (or structure-able) data and the routinized processes of analysis (Korn et al. 2017). The management of qualitative and ethnographic data with the purpose of sharing and reuse, however, is an emerging focus as yet not fully understood, and present additional challenges that can be characterized as epistemological, methodological, and ethical in nature (Feldman and Shaw 2019; Ryen 2011).

In fact, researchers applying qualitative and ethnographic methods gather less structured data, follow less routinized processes, and engage in more fluid, flexible, and open-ended research practices. Corti (2007) includes as qualitative data, "interviews … fieldwork diaries and observation notes, structured and unstructured diaries, personal documents, annotations, or photographs" (Corti 2007). Most of these types of data may be created in a variety of formats: digital, paper (typed and hand-written), audio, video and photographic. However, data is increasingly "born digital", e.g., texts are word-processed, and audio recordings are often collected and stored as MP3 files.

Ethnographic research requires more than 'just data'. Researchers gather over a long period of time a unique 'insider view' of the phenomena they study due to the nature of qualitative research methods and the circumstances in which the data are generated (Creswell and Poth 2016). The researcher focuses on capturing and interpreting human phenomena that are particular to the individuals' lived experience that they study in a particular context, and at a specific time. The context is situationally constrained by historical, cultural, social, and political factors (Coltart et al. 2013) relating to the individuals which cannot be replicated easily. Moreover, ethnographic approaches are generally based on a relationship of trust between researchers and research participants, often in sensitive domains. Data often include critical personal information (e.g., political, or religious views, diseases, corruption, even genocide) that requires high sensitivity in its handling (Eberhard and Kraus 2018). As researchers often spend long periods of time interacting with others in the field, they gather personal reflections and experiences in the form of field diaries which are not meant to be shared with third parties (Caton 1990; Eberhard and Kraus 2018). As Tsai et al. (2016) put it, it is "one thing to make available several hundred pages of interview transcripts […]. It is another thing to make available thousands of pages of field notes and journal entries – some of which may be intensely personal in content" (Tsai et al., 2016, p. 195). In an era where the researcher's positionality is seen as a central aspect of ethnographic enquiry, it is entirely possible that researchers may select or otherwise alter the data by removing material they do not want to be published and creating private 'shadow files' beyond the official material (Tsai et al. 2016).

The human aspects of data collected via interviews and through observations also lead to legal and ethical concerns. Here one of the most significant challenges confronting qualitative and ethnographic data sharing is the preservation of participant anonymity. Sharing a qualitative study and ensuring it conforms with prevailing legal and ethical guidelines (especially in the light of recent EU GDPR legislation) is quite problematic. Anonymization strategies are often mentioned as a solution but the greater the amount of anonymization the greater the risk of losing relevant information needed to make use of the data and interpret them adequately. Another pressing issue is that the majority of RDM support and training providers are often only universities, libraries, and librarians. These institutions frequently lack the personnel or expertise to provide guidance on a vast array of disciplines and heterogeneous research data practices (Hamad, Al-Fadel, and Al-Soub 2021; Kervin, Cook, and Michener 2014; Pinfield, Cox, and Smith 2014). Therefore, they may not be able to satisfy the rising demand for RDM skills applicable in various research contexts. For example, in my Collaborative Research Centre, the researchers themselves are called to engage with RDM practices. Only limited support is offered by our INF project but there are no data managers or data specialists on site. To sum up, the expectation of funding agencies and governments for Open Data and RDM practices to be developed to ensure appropriate curation, sharing and reuse of data is still very far from being realised in practice. Some issues and barriers in achieving these demands exist independently of disciplinary specificities whilst others are clearly dependent on the specific of methodological and epistemic characteristics. Other barriers, however, can be found in the lack of available infrastructures and tools that can support the development of the daily activities and workflows which constitute practice. Of course, data management in some form has always been part of the practice of researchers. It can, nevertheless, potentially be transformed through a better understanding of what the existing practices look like and how they can be both supported and transformed. In advocating a contribution for qualitative research, I suggest it provides an alternative to the over-generalising tendency to be observed in top-down RDM approaches and calls for a better and more nuanced view of what data management is, and can be.

2.5 Available tools and infrastructures for RDM

Some obstacles to the appropriation of RDM practices have their origins in the interaction with sociotechnical infrastructures or in the absence of adequate ones (Christine L Borgman 2010; P N Edwards et al. 2013; Sebastian S. Feger et al. 2020b). To date, many of the currently available solutions are research storage facilities that take the form of a repository. These can either be generic, like Globus[9], Zenodo[10], Dryad[11] or DataverseNO[12] which support many

[9] https://www.globus.org/data-sharing
[10] https://zenodo.org/
[11] https://datadryad.org/stash
[12] https://dataverse.no

different types of research data and are therefore appropriate for a wide variety of scientific fields; or they can be discipline-specific and community-driven, like QualiService[13], GESIS[14], and SowiDataNet[15] which are examples of solutions suitable for social science research (Linne and Zenk-Möltgen 2017). University repositories are also progressively being built by all major institutions and they frequently cover various fields and disciplines at once, presenting similar characteristics to those of generic repositories.

However, these infrastructures and services focus only on two specific aspects of the RDM data life cycle: long-term archiving and sharing. They do not necessarily solve the upstream problem of how to effectively support researchers in curating, documenting, and managing their data during the research process. Archiving data in a repository, for the purpose of sharing or simply for long-term preservation, is then seen by researchers as the ultimate step, the archiving process not being directly connected to the daily practices and environments in which data are generated, processed, and analysed. It is perceived simply as an additional burden, with no direct benefits, especially in the absence of a strong mechanism of rewards (Chawinga and Zinn 2020; Curdt and Hoffmeister 2015; Donner 2022). As evident by the work of Feger et al. (2020) and by my own work (Mosconi et al. 2019a), researchers have been driven to use often haphazard, ad hoc techniques because of a lack of sufficient infrastructure, knowledge, and skills which ultimately lead to unstructured archives or refusal to archive data completely. In addition, open data portals or data repositories often focus on the organization of data and the policies that surround it, such as the number of datasets, the number of formats, the open licenses, and so on. Even though formats, standards, and licenses are necessary for the long-term preservation of data and their retrieval, there are still very few design solutions that specifically support the practices and workflows that are necessary for the ongoing curation and sharing that will lead to interdisciplinary collaboration around data (Feger et al. 2020). In fact, only a very small amount of work has been focused on developing innovative digital solutions to these problems (Feger et al. 2019; Garza et al. 2015; Mackay et al. 2007). One notable example is Touchstone (Mackay et al. 2007) a platform which aims to facilitate reuse and replication of research on interaction design and allows exporting and importing of experimental designs and log data. The platform enables researchers to specify their experiments and provides support throughout the evaluation process. While, for the Humanities, a good example is PECE (worldpece.org), an open-source, Drupal-based platform

[13] https://www.qualiservice.org/de/
[14] https://www.gesis.org/en/research/research-data-management
[15] https://www.re3data.org/repository/r3d100011062

designed to support a wide range of collaborative humanities projects. It pays considerable attention to the way data artefacts get collaboratively shared, archived, and potentially reused (Fortun et al. 2021; Poirier 2017). Nonetheless, as previous research has shown (Feger et al. 2020) more tools and infrastructures in support of RDM practices are needed and the role of CSCW and HCI can be crucial "in supporting the transition to effective digital RDM through a design-focused understanding of the roles and uses of technology" (Feger et al. 2020). Now I turn to literature on information infrastructure and infrastructuring emerged mainly from CSCW to ground my work into a processual and relational perspectives which is needed to design new tools and infrastructure in service of new RDM practices.

2.6 Information Infrastructures and Infrastructuring

CSCW extensively contributed to the study, design, development of information infrastructures. While some researchers took a techno-centric perspective which mainly focused on studying and analysing the technical components of an infrastructure (Tanenbaum 2002; Dourish 1999), other scholars proposed a relational and a socio-technical perspective. For example, Star and Bowker (2002) went beyond the mere analysis of the physical and technical components of an infrastructure and looked into the role of actors involved in their use and their relationships. In a study of distributed information system within a scientific community, they defined eight salient characteristics of an information infrastructure (Star and Bowker 2002; Star and Ruhleder 1996):

- embeddedness in other social and technological structures;
- transparency in invisibly supporting tasks;
- spatial and temporal reach or scope;
- the taken-for-grantedness of artifacts and organizational arrangements, learned as part of membership in a community;
- infrastructures shape and are shaped by conventions of practice;
- infrastructures are plugged into other infrastructures and tools in a standardized fashion, though they are also modified by scope and conflicting (local) conventions;
- infrastructures do not grow de novo but wrestle with the inertia of the installed base and inherit strengths and limitations from that base;
- normally invisible infrastructures become visible upon breakdown.

In this view, an information infrastructure is always to be considered as a relationship between human situated practices and the technologies that enable and support those practices. As Star and Ruhleder put it, an infrastructure happens “in practice, for someone, and when connected to some particular activity” (Star and Ruhleder 1996, p. 112). The concept, then, reflects the interdependencies between technical and social contexts. This relational quality clearly differs from the views on information infrastructures as technical artefacts/objects (i.e., discrete, standalone entities) typically applied in engineering and design fields. At the same time, its situated nature stresses the local and micro-level, which differs from those large-scale information infrastructures considered as macro-level systems: a focus that can be extensively found in the Large Technical Systems field in the Science and Technology Studies (STS) tradition (Simonsen, Karasti, and Hertzum 2020).

For a long time, researchers have tended to see infrastructures as a static entity fixed at the point of design. Subsequent research has addressed infrastructures in terms of on-going processes and purposeful activities and therefore the concept of *infrastructuring* has been proposed (Pipek and Wulf 2009; Star and Bowker 2002). Björgvinsson, Ehn, and Hillgren (2010) stated: “Infrastructuring can be seen as an ongoing process and should not be seen as being delimited to a design project phase in the development of a free-standing system. Infrastructuring entangles and intertwines potentially controversial ‘a priori infrastructure activities’ (like selection, design, development, deployment, and enactment), with ‘everyday design activities in actual use’ (like mediation, interpretation and articulation), as well as ‘design in use’ (like adaptation, appropriation, tailoring, re-design and maintenance)”. Pipek and Wulf (2009) understand infrastructuring as the practice of “re-conceptualizing one’s own work in the context of existing, potential, or envisioned IT tools”. In this view, ‘design’ is not a task exclusively in the hand of the designated designer but designing emerges from the various interactions among developers, designers, users, and the technology. It is a negotiated and decentred ongoing activity. Therefore, the concept of infrastructuring highlights a processual, in-the-making perspective (Karasti and Baker 2004; Karasti 2014; Karasti and Syrjänen 2004; Star and Bowker 2002; Pipek and Wulf 2009) whose temporalities and scales need to be considered and carefully analyzed. Many infrastructuring processes and phenomena emerge from an ‘installed base’ (from what is already there) and are strongly influenced by the network of existing dependencies. However, with some very innovative or radically new technological concepts, the infrastructures that emerge are not only influenced by existing relations and dependencies, but also by imagined or envisaged relations. This is the case in my work, which is driven by imagining new socio-technical relationships to be built in support of new practices not yet in place and shaped by the new requirements and expectations that are introduced by the Open Science agenda. By following an infrastructuring approach synergistically with these previous works, my research aims at building socio-technical processes able to relate different research contexts and actors to one another and create new (social) relationships 'from within'.

Of relevance here also the concepts of 'points of infrastructure'. A point of infrastructure can be thought of as the point where routine and invisible technical and organizational matters become visible, usually when problems arise, or innovative possibilities are introduced. As Pipek and Wulf point out, "While our framework describes a single point of infrastructure, it should be obvious that in any concrete work environment, points of infrastructure may show up repeatedly" (op cit., p. 459; see also Björgvinsson et al. 2010).

PoIs do not happen arbitrarily. Instead, there are specific factors that are likely to trigger a (socio-technical) reconsideration especially where there is a dependency between a (work) practice and its supporting (work) infrastructure that has developed previously and that hence becomes largely invisible to the actors who engage the practice in question (Ludwig, Pipek, and Tolmie 2018, p.4). The fracturing of the dependency between (work) practices and (work) infrastructure is what causes its reconsideration, and this can happen based on four motivational forces (Pipek and Wulf 2009):

- *Actual infrastructure breakdown:* The infrastructure is not able to deliver the service it is expected to provide, often because parts of the technologies have become inoperable (e.g. power failure when trying to stream a video).
- *Perceived infrastructure breakdown:* The infrastructure does provide its service technologically, but not to the level of expectations of its user (e.g. the low quality of a streamed video in a mobile network when there is limited bandwidth available).
- *Extrinsically motivated practice innovation:* The framing conditions, the task, and goals associated with a practice, have changed in such a way that it is impossible to maintain the old practice (e.g. a video streaming platform develops a new pricing/subscription scheme and the customer requires a new device, accompanied by new process documentation).
- *Intrinsically motivated practice innovation:* The framing conditions, tasks and goals associated with a practice remain unchanged, but practitioners discover the potential for performing the practice in a new way, possibly because it is more cost efficient, simpler, quicker, or simply more fun (e.g. equipping the home with new sensors and

technology to be able to start streaming a video two minutes after arrival in the living room).

Points of infrastructure (PoI) in turn provide for 'resonance activities', which include observing and communicating aspects of what has become visible. Ludwig et al. (2018) describe resonance activities as," … currently underexplored aspects of infrastructuring [which] can be understood to be all those kinds of activities that may become visible to other users engaged in related practices, or to technology developers who laid the technological foundation of an ongoing practice innovation (ibid; 113). This is clear from my own experience in INF, and indeed in the development of research hub.

In my view, the Open Science agenda and the expectation of funding agencies for Research Data Management practices are causing Point of Infrastructures. What this means is that, due to these new requirements which cannot be met by current tools and infrastructure, the successful sharing and reuse of research data is problematised. Infrastructuring efforts are needed in which researchers, designers and IT developers can work together, and reflect on how to meet these new demands in a meaningful way.

2.7 Research Gap

As highlighted in the literature review, the management of qualitative and ethnographic data with the purpose of curation, sharing and reuse is an emerging focus but as yet understudied. It presents unique challenges which have not yet been addressed. Moreover, new tools and infrastructures still need to be developed in order to facilitate the appropriation of RDM practices in ethnographically driven research contexts.

My work then aims to fill these gaps. First, by closely investigating data practices of researchers working in an interdisciplinary collaborative context – mainly applying qualitative and ethnographic methods – recently affected by the funding agency demands of embracing more openness and transparency aligned with the Open Science agenda and to develop RDM practices. Second, by conceptualizing a new design solution and a research data infrastructure developed together with the researchers affected by these new demands. The approach I followed is to integrate bottom-up practices and top-down policies in the development of a research infrastructure in order to facilitate the realization of the Open Science agenda in a participatory and sustainable way especially for the researchers impacted by it. I now move on with outlining the context of my research and the research approach I followed.

3

Context and Research Approach

In this chapter, I briefly describe the INF project and the research context at the centre of my dissertation: the Collaborative Research Centre (CRC). More details can be found in the publications following chapter 4. After a brief description of the context, I outline the research approach I followed which combined elements of ethnography and design work, and it was inspired by 'embedded research' approaches (Lewis and Russell 2011a; Jenness 2008). Finally, I outline two conceptual influences – articulation work and infrastructuring – relevant to the work I carried on in the INF project concerning the implementation of Research Data Management solutions developed to support the appropriation of new data practices.

3.1. The Collaborative Research Centre and the INF project

My research took place in an information infrastructure project called INF connected to a Collaborative Research Centre (CRC) based in a middle size town located in the region of North Rhein Westphalia (Germany). The CRC is an interdisciplinary research network consisting of 14 projects with more than 60 scientists coming from a variety of disciplines and research fields (i.e.: media studies, ethnology, sociology, anthropology, philosophy, German studies, and computer science as well as history, education, law, and engineering). The CRC is an exemplary academic context characterized by the interdisciplinarity of every project (with one or more disciplines working together) and by the predominance of ethnographic and qualitative research methods applied by most of its members. Each project is composed by two Principal Investigators (PIs), coming from different fields, leading a team composed by one or two postdocs and/or one or two PhD students depending on the size and scope of the project.

The discipline of media studies is leading the CRC's research program which focuses on the exploration of digitally networked data-intensive media, no longer conceptualized as 'standing alone media', but as being cooperatively produced by infrastructures and publics. The CRC investigates cooperative media with a praxeological approach (Burkhardt et al. 2022; Schmidt 2016) that mediates between history and the present and focuses on the cooperative practices that are created by media and from which media emerge. The interdisciplinarity and collaborative nature of the centre is promoted through several formats such as annual retreats, seminars, lecture series, summer schools, and workshops organized every year by the CRC's members themselves.

The CRC started its first funding phase in January 2016, completed it in December 2019, and begun its second phase in January 2020 (funded until December 2023). The funding body is the DFG (in German Deutsche Forschungsgemeinschaft; in English: German Research Foundation), one of the most prestigious research institutions in Germany who funds a consistent number of CRCs every year and several other research programs.

Within the organization of a CRC, an INF project usually has the goal of supporting the development and implementation of a data management strategy, as well as providing an appropriate information infrastructure to all its projects. My own INF project was tasked from the beginning with supporting the appropriation of RDM practices, providing infrastructural solutions, and developing new design concepts for RDM in support of the whole CRC. It being the case that the CRC is mainly composed by researchers applying qualitative and ethnographic methods, the focus of the INF project and of my own research focused on understanding the challenges of managing this type of data and at the same time developing design solutions that could address the identified challenges. My INF project consisted of two sides collaborating with each other: 1) the IT service provider of the University, composed by three members, in charge of providing infrastructural support; and 2) the CSCW chair - represented by myself, two student assistants, and my academic advisor Prof. Volkmar Pipek - in charge of leading the empirical research and develop new design concepts grounded on the empirical data. The INF project started in January 2016, but I personally joined in September 2016. Since then, I have worked in the CRC as an affiliated member and I therefore carried multiple roles in the field: member, researcher, and designer.

3.2 The DFG agenda for RDM

Since 2010, the DFG defined and adopted "Principles for the Handling of Research Data" which highlighted the importance of long-term archiving and accessibility of research data that should be applied to all fields and disciplines while observing subject-specific requirements[16] (DFG, 2010). With those Principles, aligned with the global trend of Open Science highlighted in chapter 2, the DFG wishes to promote future cooperative research activities at a national and international level, thus providing useful insights for the support of innovative research in other disciplinary contexts as well. The principles are expected to be followed by all DFG funded projects and in fact long-term preservation and the sharing of materials with a wider public form part of both CRC's proposals (phase 1 and phase 2). This institutional mandate, however,

[16] https://www.mpg.de/230783/principles_research_data_2010.pdf

is particularly challenging for my own CRC due to the nature of the data collected and the methods applied by most of the projects, mainly qualitative and ethnographic. As presented in the related work, the curation, long-term archive and sharing of these kinds of materials is still an unsolved challenge and data management practices are not yet established or consolidated. My research and related activities were then constrained. meaning that its aims were restricted by the institutional framework and expectations formulated by the funding agency. Therefore, my work aimed at investigating how to support CRC's researchers in appropriating RDM practices aligned with the expectations of the DFG while respecting the methodological, epistemological, and ethical concerns specific to qualitative and ethnographic data management.

3.3 Ethnography and design

Ethnographic approaches have, by now, become commonplace in areas like HCI and CSCW. They draw on a history of anthropological and sociological research dating back to the 1920s, when anthropologists engaged in "strange tales in faraway places". Examples include Boas (1914), Tylor (1882), and Malinowski (1922; 1929). Subsequently ethnographic techniques were deployed in more familiar environments such as urban life (Park and Burgess 1925; W. I. Thomas and Florian 1927) by sociologists in the Chicago school. Even at that time, some work focused on the working lives of ordinary people (Hughes 1958). It was also clear even then that so-called ethnography could involve many different methods, including participant observation, interviewing, document analysis, diary materials and so on. Rather, ethnography could be understood as being a methodological commitment, involving epistemological and ontological commitments as well as certain methods. It was the work of, for instance, Blumer (1954; 1940), that established what this entailed. Blumer was a critic of the idea of 'universal laws' in the social sciences and argued further that interpretation was fundamental to sociological research methods. Later work by, for instance, Goffman; the ethnomethodologists, and not least Anselm Strauss, was influenced by this insistence on developing 'sensitizing' or 'illuminating' concepts rather than precise and invariant ones. They constitute an important link to the use of ethnographic approaches in HCI and CSCW, not least with the development by Glaser and Strauss of grounded theory, and equally concepts such as articulation work which are now often deployed in CSCW (see e.g., Schmidt and Bannon 1992).

Ethnography, then, is not in any simple sense, a method, but is a kind of analytic commitment, one which I now describe in the context of my own work. This approach to research is made more complicated by the fact that, in CSCW, ethnography is deployed to a purpose. It is associated with design. A number of authors have commented on the nature of ethnographic work in the design context (David Randall, Harper, and Rouncefield 2007; Crabtree, Rouncefield, and Tolmie 2012; Blomberg and Karasti 2013). In Siegen, careful attention has been paid to the how, when, where and when of ethnographic research (Wulf et al. 2015a; Rohde et al. 2017), sometimes influenced by participatory design (Small and Uttal 2005; McIntyre 2007; Hayes 2011; Hearn et al. 2008). PD, PAR, along with ethnographic commitments, have influenced my own work, particularly when I began to look at 'embedded research' (see below) as a way of thinking about my long-term involvement. I say more about the conceptual influences on my work below.

3.4 Embedded Research

Over the course of my long-term engagement, which started in November 2016, I followed an ethnographic approach comprise of participatory observations, qualitative interviews, and facilitation of design activities, meetings, and events. The first two years of my work were dedicated to the exploration of the research context and the research data management practices of CRC's members. Between 2017 and 2019, over thirty qualitative interviews were performed where I investigated researchers' data lifecycle with specific attention to documentation and data sharing practices.

ID	Pseudonym	Background	Academic Role	Relation to qualitative and ethnographic methods[17]
#1	Sophie	Media Science	Principle Investigator	QM + others
#2	Joe	Media Science	PhD Student	QM + others
#3	Alvin	Sociology	Post-Doc, Project Leader	Trained in QM + E
#4	Lucy	Sociology	PhD Student	Trained in QM + E
#5	Mary	Law	PhD Student	IP applying QM +E
#6	Rupert	History	Principle Investigator	Oral history interviews
#7	Lukas	Sociology	Post-Doc, Project Leader	Trained in QM + E
#8	Mark	Political Science	Project Leader	Trained in QM + E
#9	Paul	Sociology	Principle Investigator	Trained in QM + E
#10	Carl	Sociology	PhD Student	Trained in QM + E
#11	Rob	Media Science	Principle Investigator	Oral history interviews
#12	Colin	History	Post-Doc, Project Leader	Oral history interviews
#13	Julian	Anthropology	PhD Student	Trained in QM + E

[17] Relation to qualitative and ethnographic methods, key:
QM + others = Qualitative Methods complementary to other methods
Trained in QM + E = It means strongly trained in Qualitative methods and Ethnography
IP applying QM +E = It refers to an individual working in an Interdisciplinary Project applying qualitative methods and Ethnography. The subject could apply those methods or a collaborator.

#14	Aaron	Business Information System	PhD Student	IP applying QM +E
#15	Philip	Computer science	Principle investigator	IP applying QM +E
#16	Cliff	Business Information System	Post-Doc	IP applying QM +E
#17	Nolan	Business Information System	PhD Student	IP applying QM +E
#18	Trey	Business Information System	PhD Student	IP applying QM +E
#19	Victor	Business Information System	PhD Student	IP applying QM +E
#20	Will	Anthropology	Principal Scientist	Trained in QM + E
#21	Beth	Political science	PhD Student	Trained in QM + E
#22	Tom	Sociology	PhD student	Trained in QM + E
#23	Robert	Physiology	Project Leader	IP applying QM +E
#24	Erik	Human Computer Interaction	Post-Doc	IP applying QM +E
#25	Susanne	Social Science	Principle Investigator	Trained in QM + E
#26	Alan	Computer Science	PhD Student	IP applying QM +E
#27	Carolyn	Human Computer Interaction	Project Leader and PhD student	IP applying QM +E
#28	Kevin	Economy	PhD student	IP applying QM +E
#29	Julie	Sociology	Project Leader and PhD student	QM + E
#30	Danny	Business Information System	Project Leader and PhD student	IP applying QM +E

Table 2. List of the interviewees with their disciplinary background, academic position and their relation to qualitative methods (see the key, footnote 9).

Since 2017, I led a forum called ‘Research Tech Lab’ where CRC’s members were invited to participate in open discussions about methods, tools, and research data practices including specific sessions which targeted RDM issues. In some of these sessions, I invited external speakers with expertise in RDM who gave input concerning data archiving, anonymization practices, data sharing and reuse.

Based on the analysis of the interviews, interactions during Tech Lab sessions and informal meetings, I also collaborated with the IT service provider of the University, where I specifically helped developers over the years to customise several open-source tools (i.e.: *RDMO*: for creating Research Data Management plans; *DSpace:* a long-term repository; and *Humhub,* a platform for team collaboration and sharing). In particular Humhub, later renamed ‘Research-hub’, was established in 2020 to customise, test, and study new RDM concepts and collaborative workflows expected to be implemented by INF in the long-term (see chapter 4).

Due to my role in the CRC context where I was an affiliated member actively working in the INF project, my methodology is best described as ‘embedded research’ (Lewis and Russell 2011a). The arrangements of an embedded researcher have roots within both anthropological and sociological traditions and are not tied to either a specific methodological approach or to a singular discipline. According to McGinity and Salokangas (2014) embedded researchers are “those who work inside host organisations as members of staff, while also maintaining an affiliation with an academic institution. Their task is seen as collaborating with teams within the organisation to identify, design and conduct research studies and share findings which respond to the needs of the organisation and accord with the organisation's unique context and culture" (op.cit. 2014, p.3). A salient aspect of this research is a sustained didactic element in the engagement (Jenness 2008) where research findings are shared early on with the research participants to stimulate discussions relevant for the institutions to improve reflexivity and practices. In the case of my research context, the DFG agenda, RDM concepts and technicalities needed to be explained, discussed, and negotiated according to the interests, needs and practices of the CRC's researchers, and my research was used as a vehicle to do so.

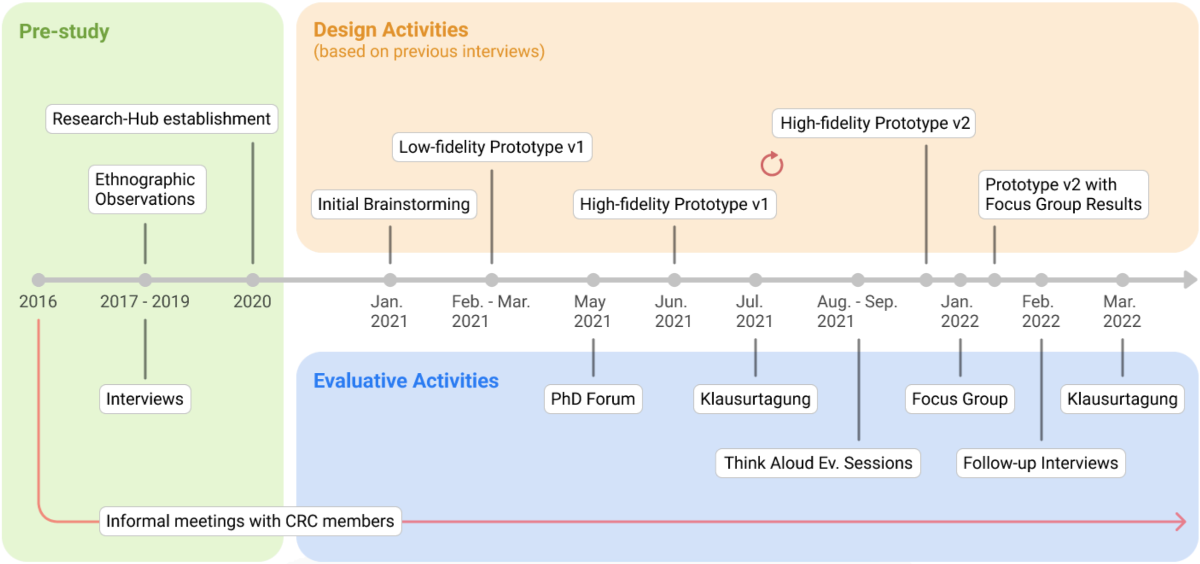


Figure 3: timeline of research and design activities

As shown in the timeline above, the research started in 2016 when I joined as affiliated member. Ethnographic observations and interviews took place between 2017 and 2019 while in 2020 the platform Research-hub was established. The time frame between 2016 and 2020 can be considered the pre-study of the research where I gathered information regarding researchers' RDM practices, related data life cycle and laid the foundations for the design efforts. I also helped researchers in compiling their research data management plans and used those meetings to discuss further researchers' challenges, expectations and personal wishes towards new tools and infrastructural support. Specifically, some concerns were expressed especially towards the establishment of an infrastructure for RDM and long-term archiving:

> “My problem in that discussion that we had was more like ‘wow ok’ they want to store for ten years and neither me nor my interviews have a control on who in the future will look at this data, who will use it and for what purposes, it sounded a bit threatening but on the other side I see the intention no?! To make research more transparent, more comparable, so this is ok, it’s a legitimate intention, yes, but for example if I do an interview with somebody and I tell him, “it’s only me looking at your data, I will just use it anonymised version of it and afterwards we will delete the data we have from you”, so we have an informed consent with my interviewees and I don’t know how they would react if I tell them hem so I don’t know what’s happening with the data in five or ten years” [#3: Alvin, Sociology].

In fact, many CRC researchers were not aware of the funding agency’s goals, and they were confronted for the first time with the idea of archiving, curating, and sharing their data. Most researchers did not engage with documentation or curation practices of any kind, and in most cases, they did not even know what metadata are and how to make use of them in their research processes.

> “I don't know if I create metadata. Maybe I do in doing those Citavi things and keywords, it’s kind of information about the information that I collected, right? […] I will create lots of reflection on how I gathered my material. But it's more reflection and not exactly metadata. Maybe you could say it's kind of metadata because its, you look at the way you gather the data and the way you work. So, if that is the thing you meant with metadata then I would say it is definitely a big part in an anthropological dissertation. But I don't know, I think myself, I am not a metadata person” [#13: Julian, anthropology].

Moreover, no tools used by researchers allowed them to engage with data curation practices and the demand for archival and sharing were perceived simply as an extra burden. However, in the ongoing interactions I had with the researchers some of them reported a strong interest and need for a new technological aid that would allow them to engage with collaborative data sense making practices and a better organization of their research materials. In fact, researchers were not completely against to share their data, in fact they were keen to learn from others how to do data interpretation and analysis and to show their own materials, but they reported a lack of technological support that would allow them to engage in this exchange.

> “you can also suggest (…) to talk to other projects who have similar research data in order to maybe, yeah, think about standardization. Do we need that, do we not need it because we’re so small, are there even standards for archiving these types of research data? (…) also, for presenting this invisible work, because making interviews is very time consuming, but it doesn’t really show a lot, so to have something like a representation of that would be great” [#Colin, Media History, Research Data Management plan meeting on January 2020].

This apparent need led me to conceptualize a new design concept called ‘Data Story’ where the partial sharing of ‘data nuggets’ or ‘data snippets’ would allow researchers to develop curation and sharing practices initially for their own sake. Initial brainstorming and design sketches were made in January 2021 and shortly after that a first low fidelity prototype was designed. From there informal and formal evaluations of the design concept and related prototype took place following what can be called “embedded evaluation” meaning that there were no obvious demarcations between investigative, design, and evaluative work. All can be seen as being mutually constitutive (details can be found in chapter 8).

3.5 Positionality

My position within the CRC was not always clear-cut; instead, it was mainly left for me to interpret and navigate. During my research in the field, I faced opposition from some researchers who saw our funding agency’s goals as a threat to their established workflows and practices. Despite this pushback, I made efforts to establish myself as both an embedded researcher and ongoing participant in conversations with members of the CRC. To gain deeper insight into data sharing methods utilized by those involved with the project, I opted for qualitative research approaches that allowed me more nuanced understandings of existing protocols used at CRC. Additionally - leaning on design principles - I worked collaboratively together with fellow members crafting tailored solutions better aligned towards specific needs of the CRC.

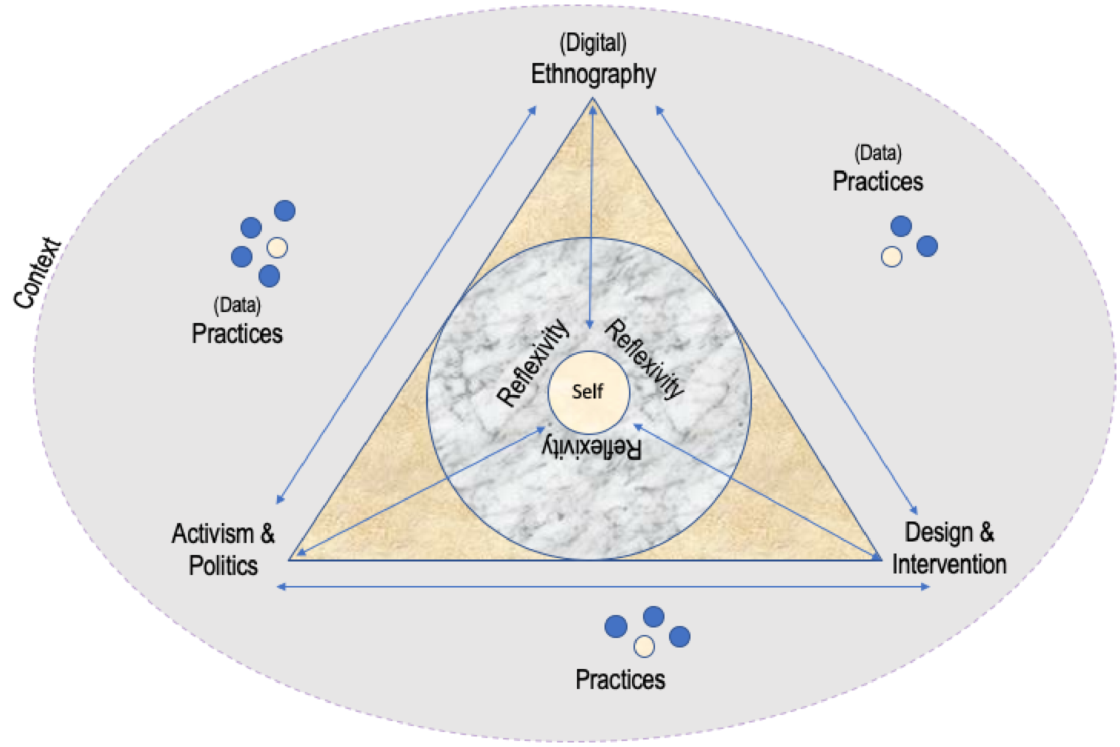


Figure 4: Relations between ethnography, design, and politics

Furthermore, I took on the role of a translator, communicating policies and best practices regarding research data management to the members of the CRC, facilitating communication and understanding between myself and the members of the CRC, and ensuring that the new data sharing and management practices are in compliance with institutional and national policies. I acknowledge that my multiple roles have brought challenges and limitations, such as the potential influence of my designer role on the research approach and the possible limitations of my translator role in fully understanding and representing the members' perspectives. Nonetheless, my role as an affiliated member of the CRC allowed me to gain a deeper understanding of their data sharing and management practices and promote new practices tailored to their needs.

In addition to the challenges posed by conflicting perspectives and goals, the development of research data management practices within the CRC was further complicated by the lack of consensus among the researchers. As an affiliated member of the CRC, I observed that each researcher had their own individual research methods, data types, and data management practices. This made it difficult to identify and implement a one-size-fits-all solution. Instead, it was necessary to work with each researcher to identify their specific needs and develop tailored solutions that met their unique requirements. Through ongoing dialogues with the researchers, I aimed to promote a shared understanding of the importance of data sharing and management, and to facilitate the development of collaborative practices that were sensitive to the varying needs of the CRC's members.

3.6 Conceptual influences

Now I will introduce two important concepts relevant to my work in the INF project: Articulation Work and Infrastructuring. Firstly, the concept of articulation work, introduced by Anselm Strauss in the 1980s, refers to the work that is necessary to coordinate and integrate the activities of individuals or groups who are working towards a common goal (Strauss 1985). In the context of the INF project, understanding and supporting articulation work is necessary to ensure that the different disciplines and research fields involved in the CRC can work together effectively towards the common goal of exploring digitally networked data-intensive media. Secondly, the concept of infrastructuring, as developed by Susan Leigh Star and Karen Ruhleder, refers to the process of designing and developing infrastructures that support collaborative work and social practices (Star and Ruhleder 1996). In the context of the INF project, I followed an infrastructuring approach in providing an appropriate information infrastructure to all the projects within the CRC, as well as supporting the development and implementation of a data management strategy. Overall, the concepts of articulation work and infrastructuring are important in understanding the role of the INF project in facilitating the collaboration and coordination necessary for the successful exploration of digitally networked data-intensive media within the CRC.

3.6.1 Articulation Work

Articulation work is a concept that refers to the coordination activities that are performed by workers in order to align their work and achieve a common understanding of the work at hand. This coordination work can be carried out explicitly or implicitly, and it can involve communication, negotiation, and problem-solving. One of the earliest works on articulation work was proposed by Gerson and Star (1986). They introduced the concept of boundary objects, which are artifacts or concepts that serve as a means of communication and coordination between different groups or individuals. Boundary objects can be physical, such as a blueprint, or abstract, such as a technical term. They allow for the different groups to understand each other's perspectives and work towards a common goal. Other studies have expanded on the concept of boundary objects to include more types of artifacts that facilitate articulation work. Bowker and Star (1999) identified three types of boundary objects: standardized forms, standardized procedures, and classification systems. Standardized forms and procedures are used to ensure that information is exchanged in a consistent and predictable way, while classification systems allow for different perspectives to be organized and compared. In addition to boundary objects, other types of artifacts have been identified as important for articulation work. Such artifacts include job aids, checklists, and protocols, which help workers to coordinate their actions and ensure that tasks are completed accurately and efficiently (Koschmann 1996).

Articulation work is not only important for the coordination of work between different groups or individuals but also within a single team. Such coordination activities can be seen in various settings, including emergency response teams, healthcare teams, and software development teams. For example, in healthcare, nurses and doctors work together to provide coordinated care for patients. This coordination requires articulation work to ensure that each member of the team understands their role and the responsibilities of others (Hendy et al. 2009). Another example can be found in software development teams. These teams often work on complex projects that require the coordination of multiple tasks and the integration of various components. In such settings, articulation work is important for ensuring that everyone has a shared understanding of the project goals and the work required to achieve them (Carstensen, Schmidt, and Spanner 2010). Articulation work has also been studied in the context of distributed teams, where workers are physically separated and communicate through technology. In such settings, articulation work becomes even more important as workers cannot rely on informal communication and must rely on explicit coordination mechanisms (Dourish and Bellotti 1992). Articulation work is especially important in complex and distributed settings, where it becomes necessary to rely on explicit coordination mechanisms to ensure successful completion of tasks.

Articulation work can be applied in the context of RDM by identifying and making explicit the different types of work involved in managing research data. For example, this can include activities such as creating metadata, ensuring data quality, and ensuring compliance with legal and ethical requirements. By making these activities visible, researchers and other stakeholders can gain a better understanding of the work involved in RDM and can more effectively coordinate their efforts to ensure that research data is managed in a systematic and effective way. Moreover, managing research data involves a complex set of activities that require coordination and communication among different stakeholders, such as researchers, data managers, and IT staff. Articulation work can help to facilitate this coordination and communication by making visible the often-invisible work of managing research data.

3.6.2 Infrastructuring

Infrastructuring describe the processes and practices involved in the development and maintenance of information infrastructures. Information infrastructures are socio-technical systems that provide a shared and evolving foundation for coordinating and performing work in organizations, communities, and other social settings. Infrastructuring involves a range of activities, such as designing, configuring, deploying, maintaining, and evolving information technologies and associated social practices, norms, and values. Infrastructuring is not just about designing and building technical infrastructures. It also involves the ongoing processes of negotiating, interpreting, and enacting shared meanings, norms, and values around the use and development of technologies. Such processes are particularly relevant in situations where different stakeholders have diverse and sometimes conflicting goals, perspectives, and interests. In this context, infrastructuring can be seen as a way of generating and maintaining coherence and alignment across different levels and scales of action, from individual practices to organizational routines and beyond.

As we have seen in chapter 2, the concept of infrastructuring has its roots in the work of scholars such as Star and Bowker (2002; 2000) who studied the development of scientific data infrastructures. They argued that such infrastructures are not simply technical artifacts but are socially constructed and maintained through ongoing work practices, negotiations, and institutional arrangements. Other scholars have since extended this notion to a wide range of contexts, including healthcare (Kuziemsky et al., 2010), community informatics (Gurstein 2007), and civic engagement (Mosconi et al. 2017).

Infrastructuring has also been applied to the design and development of information infrastructures for research data management (RDM). For example, Tenopir et al. (2011) explored the factors that influence the adoption and use of RDM infrastructures in academic libraries. They found that the success of such infrastructures depends not only on technical factors such as functionality and usability but also on social factors such as institutional policies, cultural norms, and individual motivations. Similarly, Borgman (2015) argued that RDM infrastructures need to be designed as part of larger socio-technical systems that take into account the diverse needs and goals of different stakeholders, including researchers, librarians, funders, and publishers.

Infrastructuring is also relevant to the development of information infrastructures for collaborative work and knowledge sharing. For example, Hara et al. (2003) studied the development of an online community of practice in a large, distributed organization. They found that the success of the community depended on a range of factors, including the design of the online platform, the development of shared norms and values around participation and contribution, and the establishment of formal and informal governance structures.

In the context of my project INF, infrastructuring is a key concept for understanding the development and management of research data infrastructures. RDM involves a range of practices and technologies for managing and sharing research data across the data lifecycle, from planning and collection to preservation and reuse. These practices and technologies are situated within larger socio-technical systems that involve diverse stakeholders with different roles, goals, and values. Infrastructuring can help to understand how these stakeholders interact and negotiate around the use and development of RDM infrastructures and associated practices. It can also provide guidance for designing and evolving RDM infrastructures that are responsive to the changing needs and goals of these stakeholders.

To conclude, infrastructuring is a concept that is central to the development and management of information infrastructures in a wide range of contexts, including RDM. Infrastructuring involves the ongoing processes of designing, configuring, deploying, maintaining, and evolving socio-technical systems that support and coordinate work practices and knowledge.

Now I will move on with illustrating the infrastructuring work I led which is centred around the platform, ‘Research-hub’.

4

Platform development: Research-hub modules and concepts

Since late 2016, as explained above, I investigated researchers' individual and collaborative RDM practices and their use of existing infrastructures and tools. In parallel (the work has been ongoing since March 2019), a platform for research collaboration, and sharing has been, and continues to be, developed by a team composed by myself, two other PhD students and a small group of student assistants. The platform chosen for this work is called Humhub (https://www.humhub.com/en), open-source software built for team communication and collaboration. Our own Humhub installation was later renamed 'Research-hub' and with it the goals were: 1) to explore design concepts to be integrated in the collaborative platform in the long-term; 2) to facilitate the development of new practices in the direction of data curation and sharing. The platform was chosen for its highly customizable features. In fact, Humhub is a free (community edition), flexible, and open-source social network kit based on the Yii2 PHP framework, it has a big open-source marketplace containing a significant and growing selection of modules and a quite active OS community. Thanks to the module system, it is possible to extend Humhub by using third party tools, writing completely new and independent modules, and connecting existing software. Overall, the platform afforded for a person and process-centric approach rather than a data-centric approach which is more typical of databases.

Once we selected the OS software deemed suitable to our needs, we downloaded a version and installed it on our university server. We created a project space and invited all members of the development team to join the space. Based on the interviews and observations conducted between 2017 and 2019, and by our own use of the platform, we begin to conceptualize and develop three major modules and concepts for RDM:

(1) Online Drives;

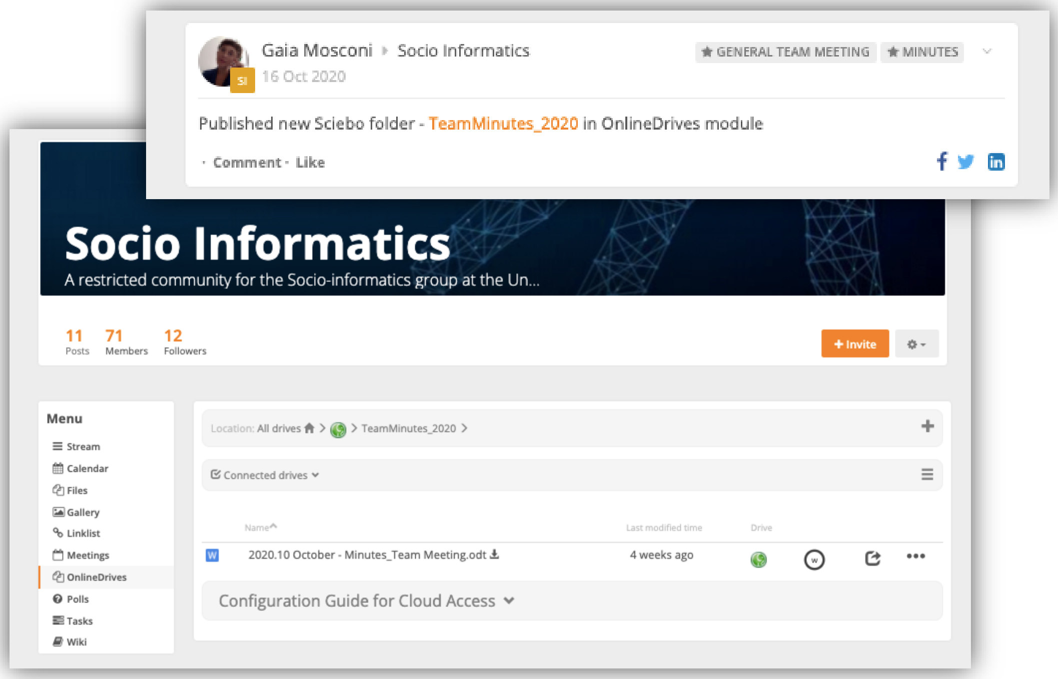
(2) Metadata Interface Processing;

(3) Data Story Module (main contribution of this thesis).

4.1 Supporting data sharing: Online Drive module

The aim of Research-hub is that it should be a small-scale research infrastructure for research collaboration, data management and sharing. However, it is not, and was never, intended to be an alternative to existing data storage and sharing solutions. My first publication (Mosconi et al. 2019) indicated that there was a diversity of views concerning preferences for individual data management solutions as against the need for a collaborative infrastructure. It seemed to us that transplanting currently used applications to our platform entailed a significant overhead and might well mitigate against people's willingness to use it. We took the view that our solution should work 'on top' of current file sharing solutions in order to increase levels of collaboration.

The online Drive module is the first module fully designed and implemented by the development team. It aims at interconnecting Research-hub to the most used file sharing system used at our university: Sciebo (https://www.sciebo.de/en/). Sciebo is a non-commercial cloud storage service for research developed and customized from Own Cloud Open-Source software. It allows automatic data synchronization from various devices and file sharing with collaborators for joint work on documents. The data is only saved and processed at university locations in NRW (Münster, Bonn, Duisburg-Essen). As a result, researchers' data are protected by the strict German data protection law.



Gaia Mosconi ▸ Socio Informatics
16 Oct 2020

Published new Sciebo folder - TeamMinutes_2020 in OnlineDrives module

Socio Informatics
A restricted community for the Socio-informatics group at the Un...

Location: All drives > > TeamMinutes_2020 >

Name	Last modified time	Drive
2020.10 October - Minutes_Team Meeting.odt	4 weeks ago	

Configuration Guide for Cloud Access

Figure 5: Online Drive screenshot from Research-hub. Connection with Sciebo displayed (below) and automatic post in the Research-hub group stream (above)

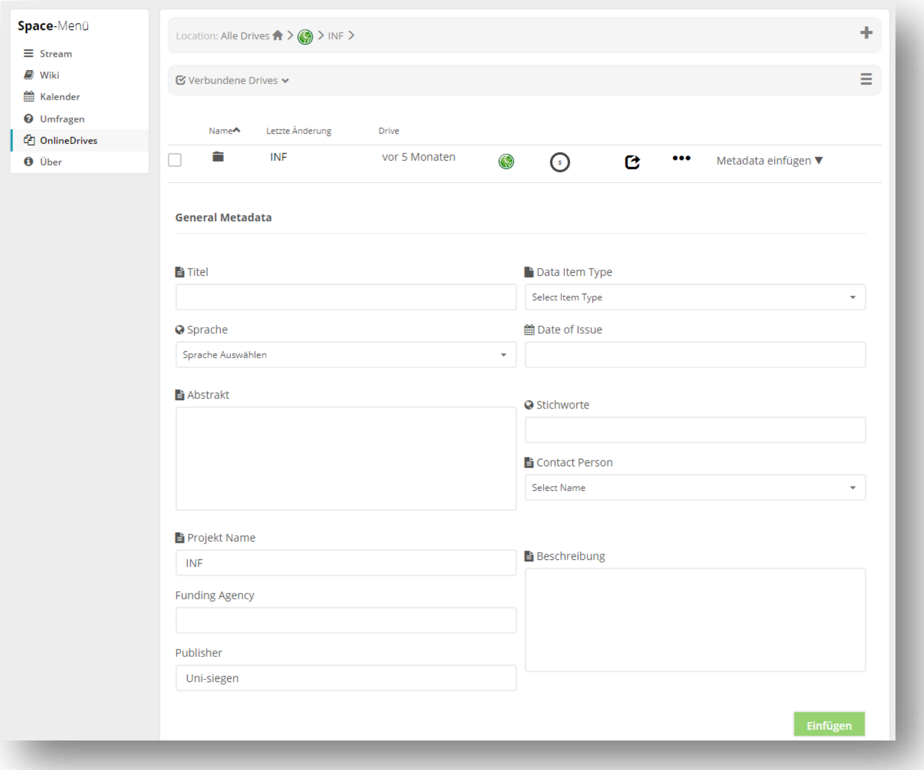
One vexed problem here is the fact that collaboration between multiple heterogenous entities like university, industry partners and public services is commonplace. Data-repositories typically have strict rules concerning access rights, which can be troublesome given the relatively fluid nature of research collaboration. To access a repository, one normally needs an account. In bigger organizations this then requires that one goes through a process where the user has a centrally managed account created in a user-repository, for example, Open LDAP or Microsoft AD. Maintaining these accounts (keeping them up to date, deleting them when the cooperation is no longer needed, and so on) is a burden for the responsible person (such as a project leader) and as yet are not supported. The users then have also to remember yet another password to get access.

In our Online Drive module, we implemented an approach where the leader of a project sets up a repository and then gives access to every member of the project through his or her (the leader's) credentials. The Online Drive module allows one to select only specific files and folders and synchronize them to user profiles, project or community spaces in the platform. In this way, only the selected files or folders are made accessible while protecting other folders that might be restricted.

The integration of the Sciebo module with Research-hub links chosen files and folders to an activity stream, thus enhancing collaborative possibilities. In the activity stream, users can visualize who is sharing files/folder, comment on it, keep track of the most important files and of major activities. Effective collaboration entails knowing who has access to data, what rights they have over it (e.g., in respect of editing), and being aware of the document history (for instance, knowing who has worked on a document and when). The development of research hub on top of Sciebo allows for an interface layer which visualizes the above practices.

4.2 Metadata interface processing and annotation manager

In respect to RDM, our immediate plans are to examine how Research-hub could support data annotation and metadata editing (currently not generally supported by file sharing systems) by developing a user-friendly interface in which performing this kind of curation work. In fact, when it comes to file sharing systems, solutions like Sciebo, Sharepoint, Google Drive and Dropbox do not support any metadata creation or tagging during the research process. As already expressed elsewhere in the literature (Bietz et al. 2012) metadata, if at all, can be collected idiosyncratically in a variety of ways and the databases used by researchers do not adequately support metadata creation. In our context specifically, researchers reported that they were not engaging with any type of documentation practices and noticed how they had no opportunity at all to curate data with appropriated metadata catalogue as expected by the RDM best practices promoted by the funding agency and data centres. Therefore, metadata or tags are required for effective research collaboration, and which can be quickly edited by researchers during the course of a study, elaborated according to need, then eventually exported, shared with colleagues or uploaded in institutional repositories. Currently, once researchers upload documents in a file sharing system as the principal repository of empirical data, they cannot attach any type of metadata to files or visualize summaries/overviews of their interviews or fieldnotes. We believe that this gap constitutes an opportunity for further innovation, and the development team is currently working on integration between Sciebo and Research-hub. Starting from the connection between Sciebo and the activity stream which now announces when a file or a folder is created or uploaded, we are implementing a standard template to annotate those files with descriptive metadata (insert picture). The implementation of tabs in the stream could allow users to retrieve and visualize all at once those files/folders synchronized in Sciebo and to work on metadata editing via the interface.



Name	Letzte Änderung	Drive
INF	vor 5 Monaten	

General Metadata

Titel:
Data Item Type: Select Item Type
Sprache: Sprache Auswählen
Date of Issue:
Abstrakt:
Stichworte:
Contact Person: Select Name
Projekt Name: INF
Beschreibung:
Funding Agency:
Publisher: Uni-siegen

Figure 6: Metadata interface processing interface

Moreover, a link between metadata interface processing and the long-term archive FoDaSi was implemented to support the migration of research data from the private/common research domain to the public domain (Treloar, 2008, see figure 8). In this approach, researchers have the ability to curate their data through the platform and send their data, initially stored in Sciebo (see above), directly to the long-term archive, where data collection is completed and metadata is made accessible for searching. The goal here is to allow the researchers to prepare their own data for curation with as little additional effort as possible. The metadata of the individual research data and folders will be stored in a database. During the collaborative phase of the data, researchers can modify, add, and remove their research data and/or the corresponding metadata. At the end of the "hot phase" of research data, the platform provides the option to select only research data worthy of archiving. When archiving, the metadata is packaged with the corresponding research item and imported into FoDaSi. Finally, the system automatically sends an email notification to the users who archived the research data, and a unique DOI is directly assigned to the data collection.

4.3 Data Story Module

The ‘Data Story’ is a module for describing heterogenous data collected during a study. It represents an alternative approach to the business of characterizing qualitative data in such a way that it is made useful to others.

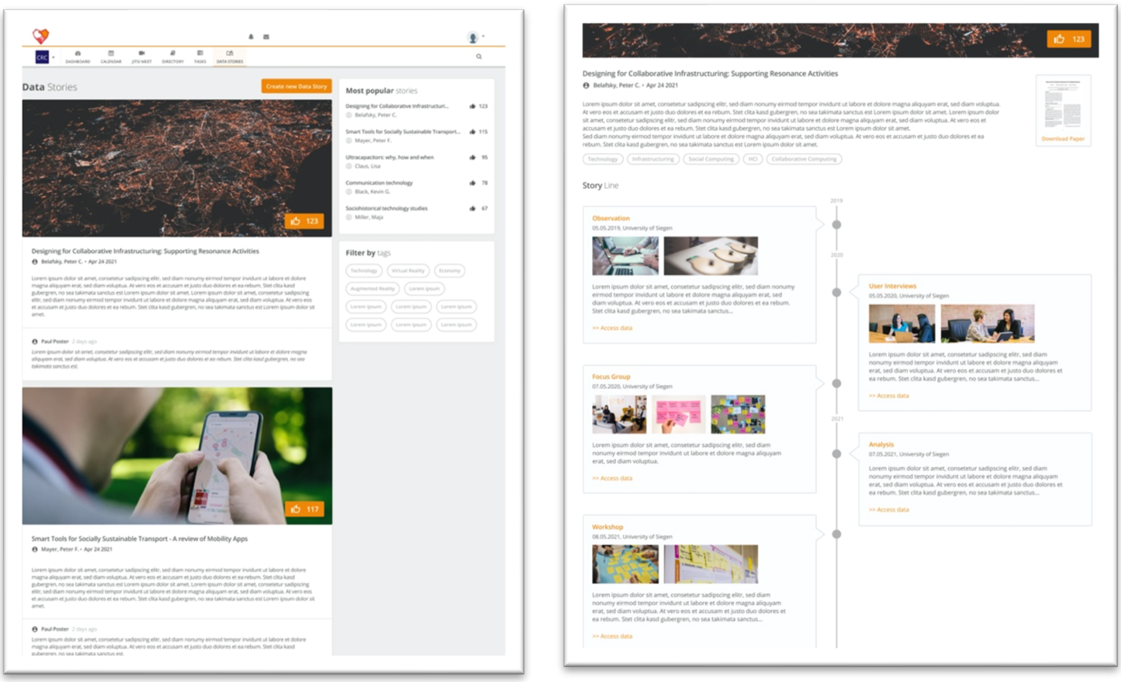

CRC
DASHBOARD
CALENDAR
JITSI MEET
DIRECTORY
TASKS
DATA STORIES
Data Stories
Create new Data Story
123
Designing for Collaborative Infrastructuring: Supporting Resonance Activities
Belafsky, Peter C. • Apr 24 2021
Paul Poster 2 days ago
117
Smart Tools for Socially Sustainable Transport - A review of Mobility Apps
Mayer, Peter F. • Apr 24 2021
Paul Poster 2 days ago
Most popular stories
Designing for Collaborative Infrastructuri... 123
Belafsky, Peter C.
Smart Tools for Socially Sustainable Transport... 115
Mayer, Peter F.
Ultracapacitors: why, how and when 95
Claus, Lisa
Communication technology 78
Black, Kevin G.
Sociohistorical technology studies 67
Miller, Maja
Filter by tags
Technology
Virtual Reality
Economy
Augmented Reality
Lorem ipsum
123
Designing for Collaborative Infrastructuring: Supporting Resonance Activities
Belafsky, Peter C. • Apr 24 2021
Download Paper
Technology
Infrastructuring
Social Computing
HCI
Collaborative Computing
Story Line
2019
2020
2021
Observation
05.05.2019, University of Siegen
>> Access data
User Interviews
05.05.2020, University of Siegen
>> Access data
Focus Group
07.05.2020, University of Siegen
>> Access data
Analysis
07.05.2021, University of Siegen
>> Access data
Workshop
08.05.2021, University of Siegen
>> Access data

Figure 7: Data Story module landing page (left), Data Story module overview (right)

In the data story module, researchers will be able to provide a "data driven" narrative of the data collected for a specific project or for a specific paper. The data story module aims at 'show casing' a portion of data collected (for a specific purpose) by supporting data 'sense making' intended as a creative and active endeavour that can be made explicit by the researchers who conduct the study.

The data story aims at displaying only a minor portion of selected and curated data accompanied by annotation and metadata as chosen by the researcher(s) who collected the data and performed the analysis. Major issues regarding data sharing related to qualitative data, as identified by Mosconi et al (2019), are: 1) data cannot be shared in full because they often describe personal and sensitive material; 2) qualitative data are difficult to understand and metadata are not often collected by researchers; 3) making sense of the data is difficult and you need to provide a lot more additional information (which takes time). The data story tries to overcome these issues and it starts from the assumption that data can be shared only partially and that a certain narrative should be provided in order to make sense of the data.

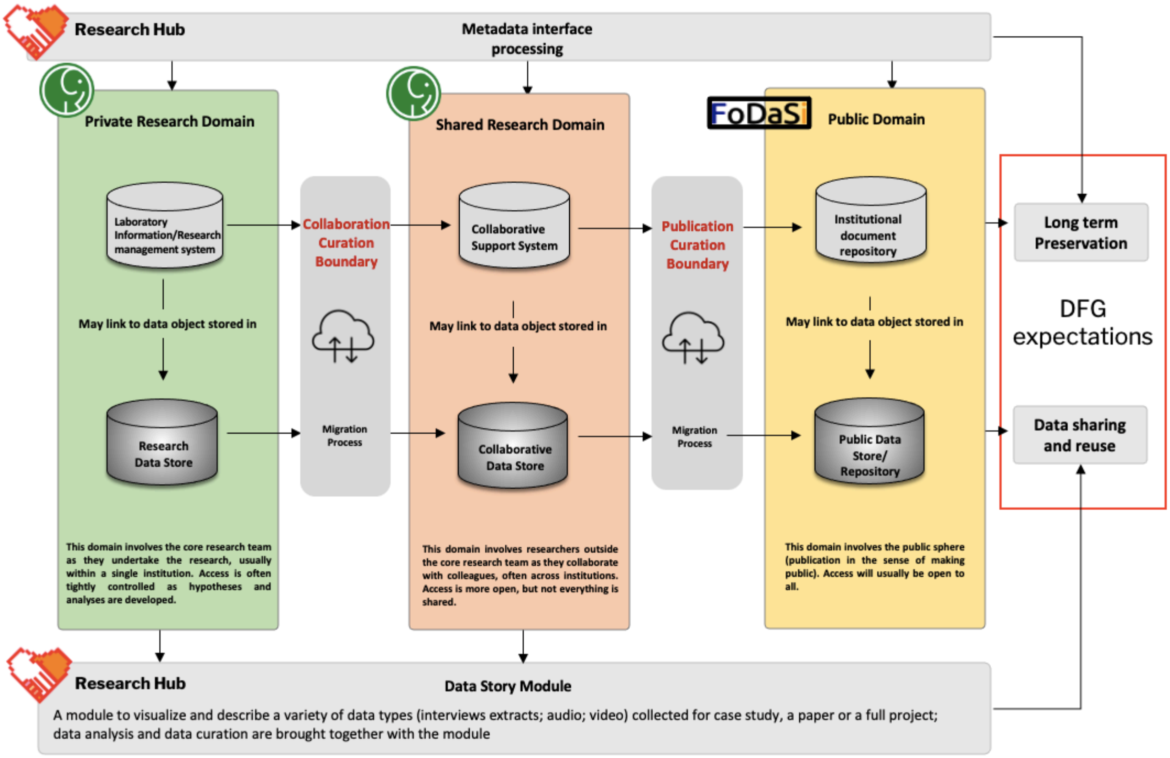

Research Hub
Metadata interface processing
Private Research Domain
Shared Research Domain
FoDaSi
Public Domain
Laboratory Information/Research management system
Collaboration Curation Boundary
Collaborative Support System
Publication Curation Boundary
Institutional document repository
Long term Preservation
May link to data object stored in
May link to data object stored in
May link to data object stored in
DFG expectations
Research Data Store
Migration Process
Collaborative Data Store
Migration Process
Public Data Store/ Repository
Data sharing and reuse
This domain involves the core research team as they undertake the research, usually within a single institution. Access is often tightly controlled as hypotheses and analyses are developed.
This domain involves researchers outside the core research team as they collaborate with colleagues, often across institutions. Access is more open, but not everything is shared.
This domain involves the public sphere (publication in the sense of making public). Access will usually be open to all.
Research Hub
Data Story Module
A module to visualize and describe a variety of data types (interviews extracts; audio; video) collected for case study, a paper or a full project; data analysis and data curation are brought together with the module

Figure 8: RDM vision represented along the Data Curation Continua (Treloar et al. 2008)

The picture above highlights the overall vision for the research data infrastructure where research data management and data curation activities are considered as daily practices which should be developed and supported along a processual and interconnected workflow (as a continuum).

In the chapters that follow, I will present the major findings of my thesis represented by four major publications:

Chapter 5. Three Gaps in Opening Science (JCSCW)
Chapter 6. Designing a Data Story: A Storytelling Approach to Curation and Sharing in Support of Ethnographically-driven Research (American CSCW)
Chapter 7. Designing a Data Story: An Innovative Approach for the Selective Care of Qualitative and Ethnographic Data (Book Chapter)
Chapter 8. Fostering Research Data Management in Collaborative Research Contexts: Lessons learnt from an 'Embedded' Evaluation on 'Data Story' (JCSCW)

Part II – Collected Findings

The second part of this thesis presents the collected findings mainly in the field of CSCW with regard to the overall objective of designing novel concepts for the appropriation of RDM practices in collaborative contexts where researchers applying mainly qualitative and ethnographic data. Chapters 5, 6 and 8 have already been published in peer-reviewed journals and have been adapted to the format of this thesis without modifications. While Chapter 7 was published as book chapter in an edited collection called "Interrogating Datafication - Towards a Praxeology of Data" reviewed and edited by CRC members.

Chapter 5 presents the early ethnographic study of the CRC's researchers data practices and highlights the challenges encountered by researchers applying ethnographic methods when confronted for the first time with the Open Science agenda and RDM expectations. We identified three major gaps in the development of the Open Science agenda hindering the appropriation of RDM practices on the part of the researchers. These are 1) Policy and practices gap, 2) knowledge gap, 3) tools gap.

Chapter 6 reacts to the identified challenges and gaps and presents the first ideation of a design concept called 'Data Story' grounded on empirical findings and on my ongoing engagement in the CRC. As a possible design solution for Research Data Management, Data Story offers: 1) a collaborative workflow for data curation; 2) a story-like format that can serve as an organizing principle; 3) a means of enhancing and naturalizing curation practices through storytelling.

Chapter 7 discusses the 'act of selective care' afforded by the Data Story concept and speculates on how the concept could become a recognized publication format to be promoted in different collaborative data infrastructures or databases.

Chapter 8 reports on how the design concept and related prototype was iteratively designed based on evaluation results. The prototype was evaluated trough my 'embedded engagement' meaning that evaluation opportunities spontaneously emerged from my ongoing presence in the field and interaction with researchers. An important element of this is that there are no obvious demarcations between investigative, design, and evaluative work. All can be seen as being mutually constitutive.

Three gaps in Opening Science

This chapter was published in JCSCW: Mosconi, Gaia, Qinyu Li, Dave Randall, Helena Karasti, Peter Tolmie, Jana Barutzky, Matthias Korn, and Volkmar Pipek. "Three gaps in opening science." *Computer Supported Cooperative Work (CSCW)* 28 (2019): 749-789.

Abstract. The Open Science (OS) agenda has potentially massive cultural, organizational and infrastructural consequences. Ambitions for OS-driven policies have proliferated, within which researchers are expected to publish their scientific data. Significant research has been devoted to studying the issues associated with managing Open Research Data. Digital curation, as it is typically known, seeks to assess data management issues to ensure its long-term value and encourage secondary use. Hitherto, relatively little interest has been shown in examining the immense gap that exists between the OS *grand vision* and researchers' actual data practices. Our specific contribution is to examine research data practices *before* systematic attempts at curation are made. We suggest that interdisciplinary ethnographically-driven contexts offer a perspicuous opportunity to understand the Data Curation and Research Data Management issues that can problematize uptake. These relate to obvious discrepancies between Open Research Data policies and subject-specific research practices and needs. Not least, it opens up questions about how data is constituted in different disciplinary and interdisciplinary contexts. We present a detailed empirical account of interdisciplinary ethnographically-driven research contexts in order to clarify critical aspects of the OS agenda and how to realize its benefits, highlighting three gaps: between policy and practice, in knowledge, and in tool use and development.

5.1 Introduction

The digitalization of information at scale has had profound consequences for the conduct of scientific activity. Some even claim we are experiencing the emergence of a 4th paradigm in science (Hey, Tansley, and Tolle 2009). Various terms have been deployed to convey the shifts that have taken place in relation to the collection, organization, management and sharing of scientific data. These include things like cyberinfrastructure, eScience, eResearch, Science 2.0, Digital Humanities, Open Science or Open Research, emphasizing various aspects of the 'data revolution' (Kitchin 2014; Fecher and Friesike 2014)After public consultation by the European Commission, 'Open Science' has become the preferred term to address this putative transformation of scientific practices.[18] Principles of openness, sharing and collaboration across the whole research process are foundational to its precepts. The aim is "… making scientific research and *data* accessible to all" by removing barriers to sharing, regardless of the type of output, resources, methods or tools used and independently of the actual research

[18] European Commission, Public Consultation: 'SCIENCE 2.0': SCIENCE IN TRANSITION. Available at: http://ec.europa.eu/research/consultations/science-2.0/background.pdf (searched at 02.09.2018)

process. The Open Science movement has successively elaborated principles[19] that have aimed to influence the political debate around these issues. One aspect of this apparent revolution has particularly drawn attention: Open Research Data. Open Research Data is considered especially critical in order to facilitate data reuse, ensure verifiability and good scientific practice, provide greater returns on public investment in research (Wallis, Rolando, and Borgman 2013; Arzberger et al. 2004; OECD 2007), and promote computational data-intensive research across all disciplines.

Significant research has gone into investigating the issues associated with managing Open Research Data (Bechhofer et al. 2010; Erickson et al. 2014; Murray-Rust 2008; Wallis, Rolando, and Borgman 2013; Choi and Tausczik 2017; Pasquetto, Sands, and Borgman 2015). Data curation, as it is typically known, focuses on the movement of data and its management (Research Data Management) to ensure its long-term value (so-called digital preservation) and to encourage secondary use. Over the last twenty years, libraries, data centres and other institutions have increasingly attempted to collaborate, build partnerships, define policies and build up information infrastructures in pursuit of those goals (Oßwald and Strathmann 2012; Pampel and Dallmeier-Tiessen 2014; Reilly 2012). Alongside of this, many funding bodies have mandated the creation of research data management plans (RDMP) and institutional Open Research Data policies. Knowing how to create a data management plan and how to efficiently structure and manage data has become a *sine qua non* condition for receiving research funding from all the major funding agencies. One obvious response to these demands has been the creation of numerous general-purpose data repositories, at scales ranging from the institutional (e.g., a single university) to the globally-scoped.[20] In 2016, stakeholders from academia, industry, publishers and funding agencies published a concise and measurable set of principles called the FAIR Data Principles (Findable, Accessible, Interoperable and Re-usable). These were adopted by the European Commission, who released new Guidelines on FAIR Data Management in Horizon 2020 (Commission 2016).

Of course, policy and practice do not always align. The Open Science agenda is clearly geared to promoting a cultural, organizational and infrastructural change in academia that is pervasive and massive in scope. However, despite all the political effort geared towards developing and

[19] Budapest Open Access Initiative, 2001; Panton Principles, 2009; Amsterdam Call for Action on Open Science presented to Dutch Presidency of the Council of the European Union, May 2016. (Search date 22.09.2018)

[20] Dataverse, FigShare, Dryad, Mendeley Data, Zenodo, DataHub, DANS, and EUDat. These digital repository systems are used by social science data archives and may be implemented locally, though they are not open source and may involve payment. They offer a range of data management and online data analysis features.

facilitating polices, standards, infrastructures and sustaining the required cultural shifts, realization of the possibilities inherent in Open Science is still some way off across all disciplines, especially for humanities and social sciences (HSS) and for those researchers applying qualitative and ethnographic methods. This should not surprise us. In respect of data collection methods, conceptual formulations, theory use and, more generally, epistemological and ontological issues, there are clear discrepancies between the requirements and wishes of the funding bodies, subject-specific research practices and needs (Eberhard and Kraus 2018) and, ultimately, how those specificities influence data management and data sharing.

CSCW, we suggest, has much to contribute to our understanding of the potential of so-called 'Open Science' ambitions. This paper presents two years of ongoing research (with findings based on preliminary analysis of 30 interviews and observations) performed in two research contexts in which scholars are working in interdisciplinary project teams and typically applying qualitative and ethnographic approaches for data collection. Through a careful examination of the practices of researchers engaged in collaborative and interdisciplinary research, we aim to show that their understanding of what data is, how it is to be organized and shared, on what occasions, for what purposes, when, and using what resources, has consequences for these ambitions. We argue that an examination of an environment where researchers come from a variety of different disciplinary origins, have heterogeneous knowledges, skills, and have different mundane practices in respect of choices about how to organize, store and represent data, ought to be fruitful.

Our reasons for taking an interest in this work lie in two broad research questions:

1. Whether interdisciplinary work entailing substantial ethnographic input problematizes Open Science assumptions.
2. Whether the Open Science agenda adds layers of complexity to questions concerning the collection, storage, analysis, sharing of data and requires new assemblages of tools.

5.2 Foundations

In this section, we start by examining the field of digital curation through a historical lens. We present two intuitional models, the data life cycle and the data curation continua, which address Research Data Management and Open Science concerns (data sharing, long-term preservation, data reuse) with prescriptive intentions. In contrast to this, we further present pragmatic models, developed in the field of digital curation in recent years, which ground data curation in actual research practices. We move on by illustrating how CSCW previously addressed collaborative research practices and especially focus on literature with similarities to the pragmatic models. We identify a connection between CSCW and digital curation literature but also a research gap, and therefore motivate the need to develop CSCW's interest in the scientific collaboration exercise under the auspices of the Open Science agenda. Finally, we outline the major tensions identified in previous work related to Open Research data in interdisciplinary contexts and in particular for qualitative and ethnographic data.

5.2.1 Institutional and pragmatic models of digital curation

The term 'digital curation' was coined by John Taylor, Director General of the UK's joint Research Councils, in an e-science policy meeting in London in 2001. He wanted ''to distinguish the actions involved in caring for digital data beyond its original use, from digital preservation''. Taylor wanted the ''[a]cquisition and curation of very large valuable collections of primary data'' to be a key function of the e-Science information infrastructure (Dallas 2016; Taylor 2001). In a report published in 2003 it was claimed:

> We are entering an era in which digital data resources are becoming a central pillar of scientific research. […] The data generated in this deluge requires active management to meet basic needs of access and re-use (Lord and Macdonald 2003).

In the UK, e-Science programs received significant amounts of funding to study grid application pilots in all areas of science, to strengthen cooperation between academia and industry, create a skilled pool of expertise in digital curation and to develop services for networking and other infrastructure.[21]

This included the establishment of the Digital Curation Centres (DCC), and demanded of different stakeholders that they develop policies and guidelines for long-term preservation and secondary use.[22] The DCC considered data in this context to be "any information in binary digital form", comprising: "(1) Simple Digital Objects: such as textual files, images or sound files, along with their related identifiers and metadata; (2) Complex Digital Objects: made by combining a number of other digital objects, such as websites; (3) Structured collections of records or data stored in a computer system" (Abbott 2008).

The DCC was one of the first centres to develop and officially accept the "data life cycle" as a model for describing a research process with the idea of shareability of data embedded in the process itself. It was even promoted as an academic "best practice". The DCC provided a high-

[21] Wikipedia re. "e-Science" (search date 04.10.2018)

[22] DDC website: http://www.dcc.ac.uk/about-us/history-dcc/history-dcc (search date 10.10.2018)

level overview of the curation stages of research data that was later simplified and adapted by other Data Centres and institutions across the globe, implicating a six-stage life cycle model (see Figure 1). The term Research Data Management (RDM) refers to all activities involved in handling research data during the data life cycle:

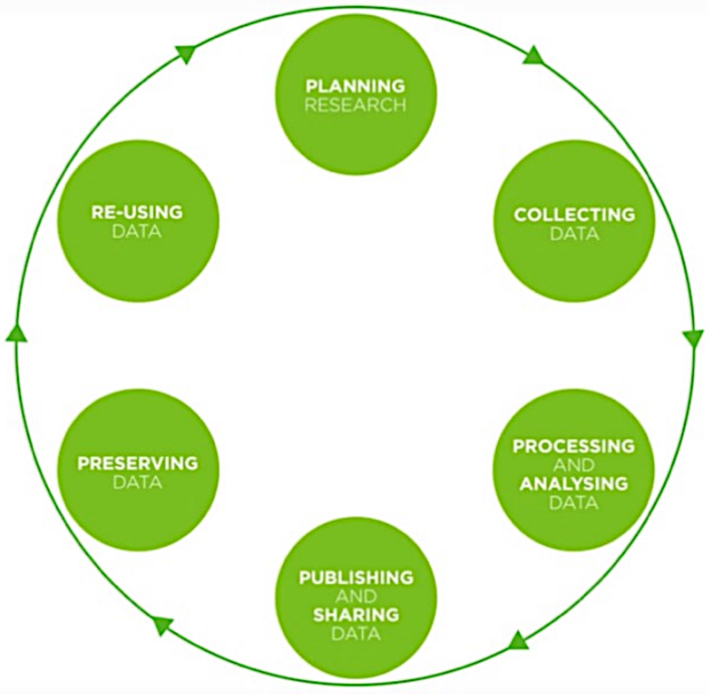


Figure 1: Data life cycle model (UK Data Archive).

While this abstract model helps us understand what constitutes "good research data management" and the related "best practices" requested by funding bodies, it does not, we argue, provide a good representation of the collaborative infrastructure in which researchers actually engage in the business of storing, managing and archiving data. In this sense, "the data curation continua" (Treloar and Harboe-Ree 2008), developed between several Australian universities, constitutes a more elaborated "institutional" model. It describes the various domains in which research data migrate during their life cycle, the actors involved in each domain and the curation boundaries.

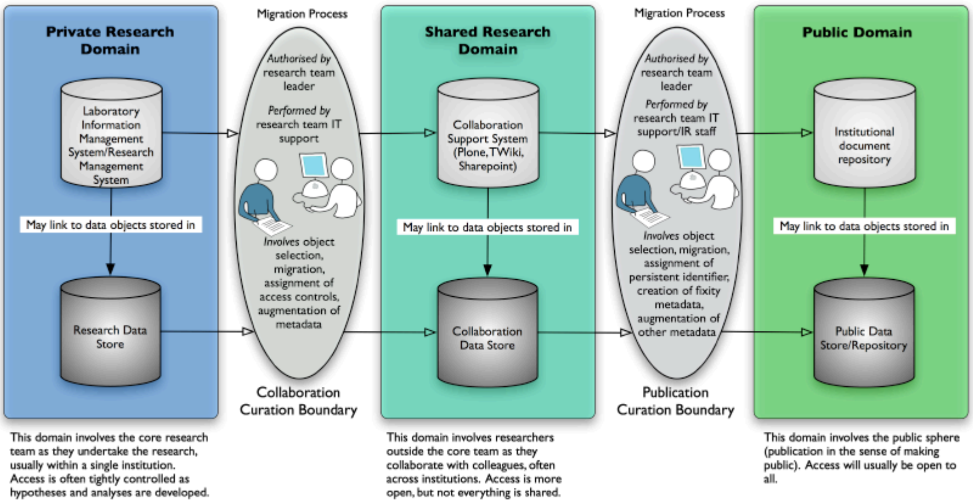

Private Research Domain
Laboratory Information Management System/Research Management System
May link to data objects stored in
Research Data Store
This domain involves the core research team as they undertake the research, usually within a single institution. Access is often tightly controlled as hypotheses and analyses are developed.
Migration Process
Authorised by research team leader
Performed by research team IT support
Involves object selection, migration, assignment of access controls, augmentation of metadata
Collaboration Curation Boundary
Shared Research Domain
Collaboration Support System (Plone, TWiki, Sharepoint)
May link to data objects stored in
Collaboration Data Store
This domain involves researchers outside the core team as they collaborate with colleagues, often across institutions. Access is more open, but not everything is shared.
Migration Process
Authorised by research team leader
Performed by research team IT support/IR staff
Involves object selection, migration, assignment of persistent identifier, creation of fixity metadata, augmentation of other metadata
Publication Curation Boundary
Public Domain
Institutional document repository
May link to data objects stored in
Public Data Store/Repository
This domain involves the public sphere (publication in the sense of making public). Access will usually be open to all.

Figure 2. Data Curation Continua. In Treloar et al., 2008, pg.6

Figure 2 shows how the migration process involves a combination of human and computer actions. Treloar et al. (2008) acknowledge how "researchers are not, in general, focused on curating their data. This is a task more suited to the professionals who will take responsibility for the data in the publication domain". However, "the process of ongoing curation in the public domain relies on provenance metadata that should have been captured during the research process" (Treloar et al. 2008, p. 7). That said, what the set of skills and knowledges that researchers need to acquire in order to perform "good" Research Data Management is as yet unclear. Equally, what the appropriate tool set for such activities might be is equally opaque.

Note, here, that both the data life cycle and the data curation continua embed "sharing" in the process they aim to describe but, first of all, promote. In this sense, digital curation appears to be prescriptive rather than descriptive of digital curation practices that happen on the "wild frontier" (Dallas, 2016).

In recent years, the field of digital curation has developed more pragmatic views on digital curation. The Sheer curation approach is a good example of this. Sheer curation is a term first used by Alistair Miles in the ImageStore project[23] and the UK DCC's SCARP project. A key feature of this approach is the recognition that digital curation activities have to be integrated into the workflow of the researchers as they create or capture data (Hedges et al. 2012)The

[23] https://alimanfoo.wordpress.com/category/the-imagestore-project/ (search date 10.09.2018)

word “sheer” is used in the sense of “lightweight and virtually transparent”. The idea is that curation should be integrated into normal working practices with minimal disruption (ibid). The approach depends both on curators ‘immersing’ themselves in data creators’ working practices and on the data capture process being so embedded within researchers' working practices that data capture is effectively invisible to them. Similarly, Dallas (2016) advocates an approach to digital curation inspired by McDonald (1995) and Hedstrom (1997) that calls for attention to practice across the ‘wild frontier’, but also calls for prioritization of human agency, pragmatics, historicity, and the sociotechnical (Dallas 2007).

An increasing number of scholars suggest a less prescriptive approach and advocate a more practice-based view, as indicated in some CSCW studies. CSCW has also emphasized a ‘pragmatic’ approach to data curation and has influenced our work. This pragmatic approach, we argue, highlights some of the tensions inherent in data curation management and emphasizes the possible consequences for scientific data which is expected to be transparent, traceable and accessible.

5.2.2 CSCW and collaborative research practices

CSCW and HCI have for some time been interested in collaborative research practices and infrastructure. Here, relevant studies focused on research practices within large, long-term, and distributed research projects and investigated the sociotechnical infrastructure needed to support shared common resources, access to dataset and special tools for data storage and processing (Jirotka, Lee, and Olson 2013; Ribes and Lee 2010; Karasti and Baker 2004; Karasti, Baker, and Halkola 2006; Karasti et al 2010; Ribes and Finholt 2009; Lee, Dourish, and Mark 2006; Bietz, Baumer, and Lee 2010; Edwards et al. 2011; Jackson et al. 2007).

Karasti et al. (2006) undertook an ethnographic study of the practices involved in a pioneering exercise in research data management and sharing associated with a long-term program in the field of ecology. Observing and giving voice to both scientists and data managers working collaboratively at long-term ecological research (LTER) sites, they provide insights into, and understandings of, the complexities involved in actual local data stewardship. They also describe how data managers, in an ongoing manner, have collaboratively worked to develop their ways of doing data management since the establishment of the US LTER Network in 1980 (see also Karasti and Baker 2004). Similar to the Sheer curation argument, they suggest looking “carefully at concrete ways of conducting science, curating data and the complicated relations of data in their environments of scientific (re)use and curation/management” because, in doing that, “more consistent understandings of existing and emerging data curation and stewardship practices” will potentially manifest themselves (Karasti et al. 2006, p. 351). The authors warn that, “while the idea of open access to publicly funded research data is an admirable one, it is also an unresolved concept in practice and poses unprecedented challenges to the actual conduct of science, curation of good quality data, and understanding of long-term stewardship” (Karasti et al. 2006, p. 350).

Bietz and Lee (2009) and Bietz et al. (2010), in a study of metagenomics, show how the design of databases for scientists to use in this context is ‘an immense challenge’ because of divergent needs, metadata assumptions and tools used. They point to the way that, even in a community of users who might otherwise be thought to be fairly homogeneous, it turns out that there are several different stakeholder communities. Moreover, the emergence of such cyberinfrastructures depends on, as they put it, purposeful activities with a ‘synergizing’ effect. CSCW research can, then, be largely associated with ‘pragmatic’ approach to digital curation issues, one which emphasizes the practices of researchers. What is clear from these and other studies is that, both in communities which, from the outside, appear to be homogenous and in those which are more self-evidently interdisciplinary, careful attention needs to be paid to the subtleties of practice and that a cultural change will evolve over a long period of time. Such an insight, we suggest, is even more pressing given the Open Science agenda where there is a stronger demand for the institutionalization and standardization of the research in all disciplines. We would suggest that the need to develop CSCW’s interest in the scientific collaboration exercise, particularly in opening research data, is predicated on a number of developments:

(1) Open Science implies an audience for data which encompasses not only primary users but also the wider scientific community and, ultimately, members of the general public, corporations and other interested parties;

(2) Open Science is characterized by a ‘top down’ policy push which may impact on the otherwise collegial desire to share data;

(3) The agenda does not recognize the very heterogeneous nature of what ‘science’ might be, and specifically does not encompass the difficulties inherent in sharing *qualitative* data in interdisciplinary contexts. This, we will argue, has to do with an impoverished and decontextualized view of what data is.

In the next two sections, we will present the issue of Open Research Data in interdisciplinary contexts and then dive into the particular case of qualitative and ethnographic data. We will argue that the emphasis on storing and archiving data has not concerned itself substantially with the practices that go into the curation process.

5.2.3 Open Research Data in interdisciplinary contexts

Neelie Kroes (2012), vice president of the European Commission responsible of the Digital Agenda, claimed: "To make progress in science, we need to be open and share". Open Research Data is considered especially critical to realize the Open Science agenda and with 'open' is often indicated free data access, re-use and sharing[24].

Data sharing and consequently data reuse have been extensively addressed by in the last decades by CSCW literature and beyond, where the force of the critique has run counter to seeing data as a final 'packaged' item. In and across almost every discipline, one of the most critical issues has been proposed as the sharing of context information to enable proper reuse (Faniel and Jacobsen 2010). To get access to contextual information and acquire a proper understanding of the data, Birnholtz and Bietz (2003) argue it is imperative to understand 1) the nature of the data, 2) the scientific purpose of its collection, and 3) its social function within the community that created it. Context also determines if something is data or metadata and the "degree to which those contexts and meanings can be represented influences the transferability of data" (Borgman 2015, p. 18). However, data is not necessarily easy to transfer. A range of tools and software applications might be in use, with ramifications for interoperability. The degree to which assumptions about data structures are held in common, whether the conceptual bases underpinning decisions about data structures are shared and the nature of motivations governing local policy on sharing, all turn out to be relevant. Even where the software in use is shared, data can rapidly become unreadable because of software and hardware updates (Christine L. Borgman 2012). Borgman (2015) also argues that the diversity of the data arising across different research approaches and fields leads to it being structured and represented in many individual and specific ways. This makes it hard to transfer and understand the context and meaning of the data for sharing and reuse.

Rolland and Lee (2013) have found that even researchers with direct access to all the original material and data from a study may struggle to understand it. As Carlson and Anderson (2007) have noted, it is false to assume that "knowledge can easily and straightforwardly be disembedded from its producers and original contexts to become explicit data for temporally and geographically distributed re-users" (Carlson and Anderson 2007, p. 647). This leads to

[24] Open Knowledge definition. Source: http://opendefinition.org/. (search date 4.02.2019)

what Edwards et al. (2011) call “metadata friction”. Drawing on an original observation by Bowker (2005), Gitelman (2013) points out that this is bound up with the fact that, ‘raw data is an oxymoron’. Instead, “data produce and are produced by the operations of knowledge production more broadly. Every discipline and disciplinary institution have their own norms and standards for the imagination of data, just as every field has its accepted methodologies and its evolved structures of practice” (Gitelman op cit., p.3). To continue the analogy, if data is always ‘cooked’, then careful examination of how the data dish is prepared and, later, conserved ought to be a valuable exercise.

It can also be argued that researchers’ data practices are frequently guided by individual benefit and equally by idiosyncratic ways of working (Fecher, Friesike, and Hebing 2015; Fecher et al. 2017a). The reality is that many researchers do not budget adequate time for metadata generation and consider this a low priority task. Nor are researchers compensated for producing data products, for they are typically evaluated for advancing science through research publication. Many data collection activities are not targeted at archiving and the resulting products are not well documented or formatted for others to use (Kervin, Cook, and Michener 2014). As a consequence, collaborative research will remain limited until there is an understanding of how to efficiently prepare and reuse data (Rolland and Lee 2013). Another critical factor is uncertainty about who has access (Gupta and Müller-Birn 2018). Researchers sometimes avoid sharing data because they are unsure who might use it. Thus, there is a need to inform researchers about the potential users and uses of their data (Borgman 2012) and provide better control of use and access (Eschenfelder and Johnson 2011).

The issues mentioned above exist regardless of the particular research area under consideration. In the case of HSS, however, where qualitative and ethnographic methods prevail, the problem is even more complex.

5.2.4 Open Research Data in ethnographic contexts

The CSCW contributions to data sharing mentioned above have mainly focused on computation and/or data intensive research endeavours in scientific domains and other fields that rely on highly structured (or structure-able) data and the routinized processes of analysis (Korn et al. 2018). Sharing of qualitative and ethnographic data, however, is as yet less studied. Corti (2007) includes as qualitative data, “interviews … fieldwork diaries and observation notes, structured and unstructured diaries, personal documents, annotations, or photographs” (Corti 2007). Most of these types of data may be created in a variety of formats: digital, paper (typed and hand-written), audio, video and photographic. However, some data is increasingly "born digital", e.g. the text is word-processed and audio recordings are collected and stored as MP3 files (Corti 2007). Beyond this, ethnographic research requires more than "just data". If 'contextual' information is significant for data reuse, we need a good sense of what the 'context' in question might be from the point of view of the researcher. Ethnographic approaches are generally based on a relationship of trust between researchers and participants, often in sensitive domains. Data can include critical personal information (e.g. political or religious views, diseases, corruption, even genocide) that requires particular sensitivity in its handling (Eberhard and Kraus 2018). As researchers often spend long periods of time interacting with others in the field, it is also necessary to reflect on the relationship between proximity and distance - which is also reflected in parts of the data such as field diaries. Field research is and has always been a borderline personal experience (Caton 1990; Eberhard and Kraus 2018).

The human aspects of data collected via interviews and through observations, lead to legal and ethical concerns. It is commonly argued that one of the most significant challenges confronting qualitative data sharing is the preservation of participant anonymity and the need to specify exactly what 'informed consent' might look like once data is more widely shared (and after it has been available for an extended period of time). Sharing a qualitative study and ensuring it conforms with prevailing legal and ethical guidelines is a problematic exercise. What guarantees need to be made to subjects in the light of widespread data sharing (and especially in the light of recent EU GDPR legislation) is likely to prove contentious. A further challenge relates to the kind of data. It is "one thing to make available several hundred pages of interview transcripts […]. It is another thing to make available thousands of pages of field notes and journal entries – some of which may be intensely personal in content" (Tsai et al. 2016, p. 195). It is entirely possible that researchers may select or otherwise alter the data by removing material they do not want to be published and creating private "shadow files" beyond the official material (Tsai et al. 2016, p. 195).

Our point here is that data sharing brings with it a number of complex problems, some of which exist largely independently of disciplinary specificities whilst others are clearly dependent on the specific methodological features of things like qualitative and ethnographic work. Thus, digital curation and the contextual information on which it depends can only be derived from a close understanding of research practices and concerns. As we will show in the following sections, our research focuses on interdisciplinary contexts with an eclectic but typically qualitative and ethnographic approach to methodology, with research taking place over a range of projects and where researchers come from different disciplinary origins. Our specific contribution is to examine these practices *before* systematic attempts at curation are made. The heterogeneity of this environment gives us an opportunity to take Digital Curation and Research Data Management issues seriously by examining the obvious discrepancies between the Open Research Data policies, distinct subject-specific research practices and the delicate business of managing data across disciplines.

5.3 Research settings and methodological approach

5.3.1 Research settings

To date, we have been engaged in an investigation of interdisciplinary research practices for 2 years, starting from November 2016 (the research is ongoing). We report here findings based on analysis of 30 interviews and observations. Our objective has been to examine data management and research processes 'on the ground', with an eye on how individuals describe their tool-use, their practices, and their data use. We especially focus on practices concerning the organization of research materials, documentation and metadata creation, data sharing, data archive, and finally data reuse.

We investigated two contexts within the same university: (1) 15 semi-structured interviews and observations were conducted within an interdisciplinary university department where most of the researchers we engaged with specialized in either human-computer interaction, business information systems or in sociology and anthropology. These researchers have received some training in qualitative and ethnographic methods (at different levels of depths) that they often apply in their research-projects; (2) At the same time and subsequently (the work is ongoing), we have conducted 15 semi-structured interviews and observations with members of an interdisciplinary Collaborative Research Center (CRC) funded by the German Research Foundation (DFG). The Collaborative Research Centres[25] are long-term university-based research institutions, funded generally for a period of up to 12 years. In particular, the research centre we engaged with is composed by 14 sub-projects funded for 4 years (2016-2019) under the name "Media of cooperation". Across 14 individual research projects at the Centre, its aims are to investigate the cooperative practices that arise in media and from which, vice versa, media arise. Almost every project of the Centre is characterized by interdisciplinary

[25] Collaborative Research Centre (CRC), source: http://www.dfg.de/en/research_funding/programmes/coordinated_programmes/collaborative_research_centres/. (search date 4.02.2019)

cooperation across fields of specialization and faculties with more than sixty researchers coming from media and cultural studies, sociology, anthropology, history, political science, law, socio-informatics, and computer science.

Out of thirty researchers we engaged with, three are both research associates of the interdisciplinary department and members of the CRC. Moreover, three authors of the paper (including the first one) are affiliated to the CRC, doing research in a project called “INF” (Infrastructural Concepts for Research on Cooperative Media) which is one of the fourteen projects. In the CRC context, the project “INF” is officially called to investigate research practices established within this centre, cooperate with the IT service provider of the university and provide infrastructural support to all CRC members. In this sense, our research might reasonably be termed an example of what Wulf et al. (2018) call ‘meta research’, or ‘research on research’ (Dachtera, Randall, and Wulf 2014).

Both contexts, the single department and the CRC, are characterized by the interdisciplinary aspect of their projects and by a specific focus on practices: many of the projects (and researchers themselves) ascribe to methodological approaches which include, among others, qualitative and ethnographic methods, ethnomethodology, participatory design, appropriation studies, and various digital (online) methods. We sought to understand sharing activities in both contexts, looking at what might need to be shared both ‘individual to individual’ and ‘project to project’, work in progress, and project histories. Comparison of work within the department (with a relatively consistent methodological philosophy), and across different departments with different philosophies was useful in that we were able to compare data sharing and data organization practices in that light. As we will show in section 4.2.1 of the findings we did not note any particular differences in sharing behaviours and data organization.

The study involved observations and interviews with the following persons (anonymised). In order to protect the anonymity of our interviewees, information about their affiliated projects and related institutions is not given. However, in table 1 we address the ways in which each interviewee stated their relation to qualitative and ethnographic methods.

ID	Pseudonym	Background	Academic Role	Relation to qualitative and ethnographic methods[26]
#1	Sophie	Media Science	Principle Investigator	QM + others
#2	Joe	Media Science	PhD Student	QM + others

[26] Relation to qualitative and ethnographic methods, key:
QM + others = Qualitative Methods complementary to other methods
Trained in QM + E = It means strongly trained in Qualitative methods and Ethnography
IP applying QM +E = It refers to an individual working in an Interdisciplinary Project applying qualitative methods and Ethnography. The subject could apply those methods or a collaborator.

#3	Alvin	Sociology	Post-Doc, Project Leader	Trained in QM + E
#4	Lucy	Sociology	PhD Student	Trained in QM + E
#5	Mary	Law	PhD Student	IP applying QM +E
#6	Rupert	History	Principle Investigator	Oral history interviews
#7	Lukas	Sociology	Post-Doc, Project Leader	Trained in QM + E
#8	Mark	Political Science	Project Leader	Trained in QM + E
#9	Paul	Sociology	Principle Investigator	Trained in QM + E
#10	Carl	Sociology	PhD Student	Trained in QM + E
#11	Rob	Media Science	Principle Investigator	Oral history interviews
#12	Colin	History	Post-Doc, Project Leader	Oral history interviews
#13	Julian	Anthropology	PhD Student	Trained in QM + E
#14	Aaron	Business Information System	PhD Student	IP applying QM +E
#15	Philip	Computer science	Principle investigator	IP applying QM +E
#16	Cliff	Business Information System	Post-Doc	IP applying QM +E
#17	Nolan	Business Information System	PhD Student	IP applying QM +E
#18	Trey	Business Information System	PhD Student	IP applying QM +E
#19	Victor	Business Information System	PhD Student	IP applying QM +E
#20	Will	Anthropology	Principal Scientist	Trained in QM + E
#21	Beth	Political science	PhD Student	Trained in QM + E
#22	Tom	Sociology	PhD student	Trained in QM + E
#23	Robert	Physiology	Project Leader	IP applying QM +E
#24	Erik	Human Computer Interaction	Post-Doc	IP applying QM +E
#25	Susanne	Social Science	Principle Investigator	Trained in QM + E
#26	Alan	Computer Science	PhD Student	IP applying QM +E
#27	Carolyn	Human Computer Interaction	Project Leader and PhD student	IP applying QM +E
#28	Kevin	Economy	PhD student	IP applying QM +E
#29	Julie	Sociology	Project Leader and PhD student	QM + E
#30	Danny	Business Information System	Project Leader and PhD student	IP applying QM +E

Table 1. List of the interviewees with their disciplinary background, academic position and their relation to qualitative methods (see the key, footnote 9).

The DFG funding carries an expectation that results of the INF project will provide a basis for systematic data management "best practices". In fact, principles such as long-term preservation and the sharing of materials with a wider public formed part of the original CRC proposal for the research being undertaken. The DFG wishes to promote future cooperative research activities at a national and international level, thus providing useful insights for the support of innovative research in other disciplinary contexts as well. This requirement, new to HSS, and in general to researchers applying qualitative and ethnographic methods, allowed us to investigate the gaps between the Open Science vision embedded in the DFG expectations and the scientific research practices we observed in the field.

5.3.2 Ethnographic approach

We followed an ethnographic approach consisting of participatory observations and semi-structured interviews. The fieldwork was conducted by two researchers (first two authors) and is still ongoing.

The interviewees were recruited via personal contact based on their position, field of specialization and experience in dealing with qualitative and ethnographic methods. The first two authors constructed a sample representing all disciplines and also sought representativeness in relation to institutional position, including PIs, post-docs and PhD students. Having explained to prospective participants our interest in research data management practices, they were given detailed consent forms that explicitly stated the purpose of our research and our interest in examining their research materials and infrastructure. The consent forms turned out to be extremely helpful in "preparing the setting" by sensitizing respondents to what physical and digital materials might be of interest. They also facilitated a discussion on the role of such "formal consent" in ethnographic field research.

The interviews always started with a nondirective open question: "What is research data for you?" in order to capture the meaning ascribed to data by researchers and its perceived value. After that, the interviews continued with four more open questions: "How do you store and organize your digital research materials?"; "What are your experiences and considerations for sharing research materials with different audiences?"; "How do you document and prepare data for long-term preservation?"; "What are your experiences and considerations of reusing data gathered by anybody else?". With these last questions, we were primarily concerned with understanding and identifying researchers' practices, in comparison to the data life cycle model, unpacking the various existing practices and relating them to the Open Science perspectives.

To better ground the interviews in actual research materials and data practices, we asked respondents to walk us through the materials stored on their personal computers and any shared folders. When the interviewee granted consent, we took screenshots and video-recorded data folder organization and software application use. This enabled us to understand and record research data management practices from the bottom up, including what kinds of socio-technical boundaries researchers encountered in dealing with qualitative and ethnographic data and how data was transformed to meet different research purposes. All interviews were conducted in English, recorded and subsequently fully transcribed. The average length of the interviews was 75 minutes (range from 45 min to 126 min).

The interview data was open coded (Strauss and Corbin 1998), after repeated readings of the data, into approximate categories that reflected the issues raised by the respondents and organizing those issues into similar statements. Iterative data analysis sessions took place from April 2017 to January 2018. The first two authors, as data collectors, were leading the sessions. Emerging themes from the analysis were captured using Annotations, a qualitative analysis software package. In the very first analysis sessions, the two first authors and more experienced researchers met to discuss, adapt, and sometimes align the emerging themes, following a broadly inductive analytic procedure (see: Thomas 2006). The two first authors expanded those themes to the full material and checked for inconsistencies. The video material was used to dive into specifics when the transcript was not sufficient to understanding certain issues like folder structure and organization of research material, or was otherwise difficult to grasp solely from the interview transcriptions.

It should be noted that the collection and analysis process was itself also a (self)reflective process. As researchers, we were ourselves involved in many of the same considerations and many of the issues reflected challenges that we faced ourselves. The close work with the IT service provider, the deep study of Open Science literature and policies made us realize the relevance of this agenda, its impact on academic work and the limitations that still exist for qualitative and ethnographic data. We soon realized that we became the medium through which meanings emerged and negotiations between institutional points of view and actual practices took place. We were 'the translator'. We became aware that our work aimed at 'making visible the invisible work' of data, tool and infrastructure use without imposing or defending a specific position. In the next section we illustrate the major findings or our study.

5.4 Findings

In what follows, we present our findings, aiming to highlight discrepancies between the researchers' data management practices and the institutional approaches mandated in the data life cycle model, explained in section 2.1 of the literature. The findings show how researchers from a variety of disciplines organize their collaborative daily work (without any help from data managers), starting with setting up a data infrastructure and outlining the socio-technical issues they face when doing so. They also reveal researcher attitudes to the fundamental concerns present in the Research Data Management and Open Science discourse (data sharing, preserving data, data reuse) and highlight how the envisaged socio-technical transition is impacting upon their work in practice.

5.4.1 Research Data Management practices bottom-up

5.4.1.1 Setting up a data infrastructure

The UK Data Archive considers the data lifecycle to start with planning the research. Major activities like planning data management, getting consent for sharing, data collection, processing protocols and templates, and exploring existing data sources are all held to be core processes at this stage. While none of the interviewees mentioned any specific data management plan or templates to guide their work, most of them described, as a first step, the choice of a file hosting system, either for themselves or for collaboration. They also selected a digital location to store and actively work upon scientific data (interviews, pictures, videos, literature etc.) for the duration of a project. All of the interviewees were involved in projects that required some sort of sharing (information, data, resources) with project partners, superiors or collaborators. In this context, t of INFRA[27], the IT service provider for the University, maintains the IT infrastructure such as file hosting sharing systems, collaboration solutions for workgroups, mail and network services. Some flavour of the frustrations experienced, however, is provided here:

> "They just say, "here we have Sciebo. Here we have SharePoint", but you have to figure out how to use it. I mean they give you a manual which says "This is how you log in and this is how you create a folder". But they don't suggest any use cases or any structures or any ways of showing how you can actually use this for something useful. So, it's of course important that they provide new options, or that they provide proper options for new stuff. But, you know, we have to figure out how we are using it and we are endlessly trying things […] It's a mess. SharePoint, we have Sciebo, we have the old BCSW thing. And we have other stuff. We have Dropbox and we have stuff that's not going through INFRA [the university's IT service provider]" [#16: Cliff, Business Information Systems]

Tools, software choices (storage system, groupware solution, etc.), the appropriate data infrastructure and how to make best use of it is all, according to the researchers, left for them to discover by themselves. Cliff continues:

> "I mean these things just come up and I just try to make the best out of it. I just use, you know, what I am familiar with. What I find useful, what is easy to learn […] whatever it is, it just has to blend in very nicely

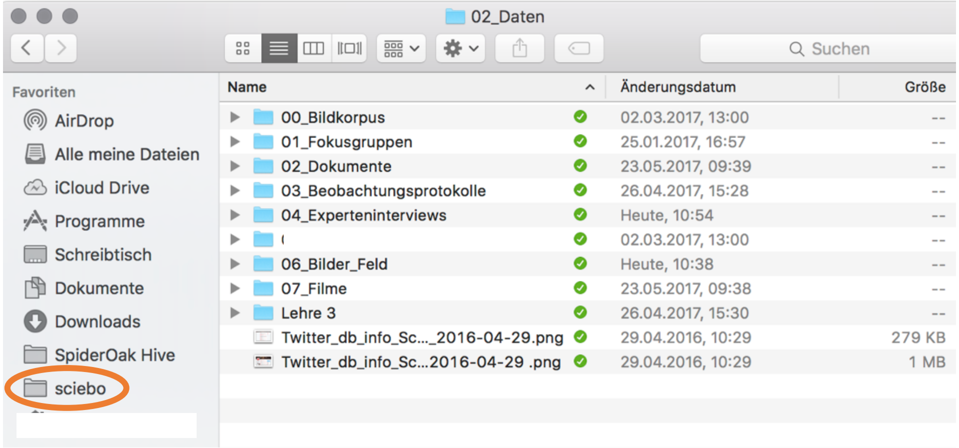
[27] Anonymized

> with my current web practices, be quick to adopt and learn because it's like I don't have the time, you have to adapt your processes and the way you do things! [#16: Cliff, Business Information Systems]"

Over the years, INFRA has offered different solutions and new ones are always in development. Sharepoint was currently the most popular file sharing system for group collaborations, despite a variety of functionality problems, including a lack of drag and drop and incompatibilities with certain operating systems. Erik works on a project (BMBF) with five partners (eighteen people overall). At the beginning of the project in 2016 they agreed to use Sharepoint but, in the end, Erik says: "It didn't work out, we kept losing things too easily, it is not the most intuitive tool to use. Today, everything we need for the project is there but when you need some things you just can't find it!".

Mark, a Post-doc in political science working in the CRC, argues: "a chain is just as strong as the weakest part of it", meaning that an "online collaboration only works well if even the not internet savvy people are trained to use it and are willing to use it and motivated to use it, so you need to have some sessions with everybody to try to accommodate the workflow, I actually wrote or re-wrote together some pieces of document in which we describe typical workflows". Mark spent a considerable amount of time learning how Sharepoint actually works, reading blogs and exchanging emails with the university's IT service providers to understand how it might best service a team distributed across Germany: "distance is the major problem, and coming with distance also scheduling appointments, so, cloud-based online collaboration is obviously a very good solution, so when I talked about struggling, it's not really fighting people, it's more about them fighting with infrastructure". After two years, he is now moving everything into another file sharing system offered by the university called "Sciebo[28]", whose interface and functionalities are similar to Dropbox (see Figure 3). Susanne, similarly, points out how "so much time, so much energy is invested in this journey, it is really a journey through all these collaborative tools".

[28] Further information on https://www.sciebo.de/.



Name		Änderungsdatum	Größe
00_Bildkorpus	✓	02.03.2017, 13:00	--
01_Fokusgruppen	✓	25.01.2017, 16:57	--
02_Dokumente	✓	23.05.2017, 09:39	--
03_Beobachtungsprotokolle	✓	26.04.2017, 15:28	--
04_Experteninterviews	✓	Heute, 10:54	--
(	✓	02.03.2017, 13:00	--
06_Bilder_Feld	✓	Heute, 10:38	--
07_Filme	✓	23.05.2017, 09:38	--
Lehre 3	✓	26.04.2017, 15:30	--
Twitter_db_info_Sc..._2016-04-29.png	✓	29.04.2016, 10:29	279 KB
Twitter_db_info_Sc...2016-04-29 .png	✓	29.04.2016, 10:29	1 MB

Figure 3: Screenshot from Luka's Sciebo project folder

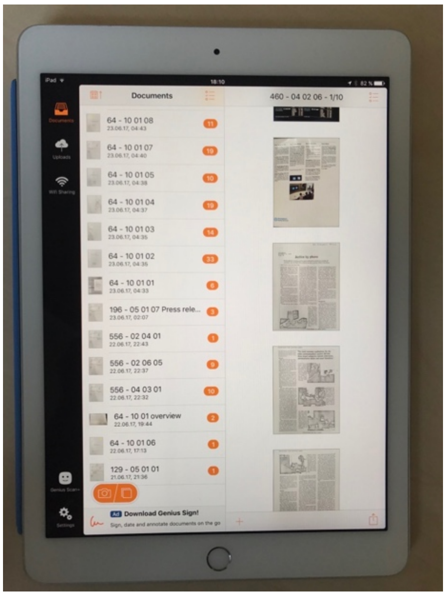
After setting up a collaborative infrastructure and tools, another preparatory step is the working up of statements regarding data protection and data handling. For large projects this is effected through a *consortium agreement*: "in the consortium agreement it is specified that nothing will be published without agreement, data will be handled with care, and it will not be disclosed." Informed consent is, of course, another hurdle. Informed consent typically identifies explicit conditions such as: 1) the scope of the research; 2) the anonymization of data; 3) how long data will be kept and where; and 4) intentions to publish the data. Most interviewees were following an orally-based consent protocol:

> "I am not as thorough as you are with your form which I really liked and it's really the proper way of doing this I guess, I didn't have a form in which all of that was stated explicitly but of course I talked to the people I asked them if it's ok to record the interview for example and I also told that this is going to be transcribed and of course every name will be removed and so on and try my best to preserve their anonymity and talked about the purpose of the project" [#7: Lukas, Sociology]

Informed consent (oral or written) can be seen as the first step in Research Data Management, whereby researchers make the conditions of storing and accessing data explicit. While researchers always mentioned confidentiality, not everyone was aware that the DFG intended to make data available or that there was an expectation of long-term preservation. Indeed, it is quite obvious from our data that little or nothing has been done at the private level (Figure 1) to facilitate or otherwise progress this requirement.

Messy folders and software support

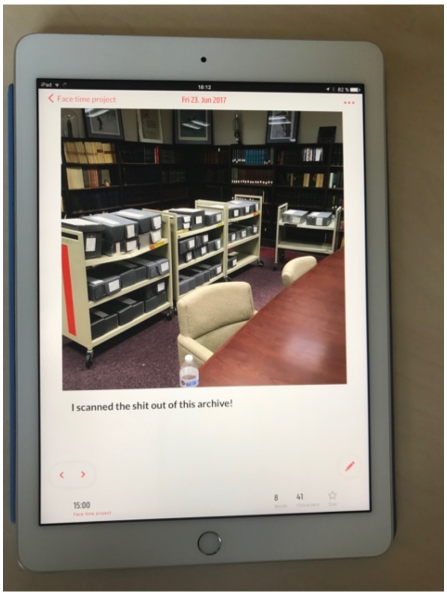
Colin is a post-doc in the History department. He is working “in a media historical project” and his “research has more to do with archival material then with ethnographical data”. However, he has wondered: “and this is experimental […] if I could use some of the approaches from grounded theory for instance for bring all this together”. In one of his first visits to an historical archive, he took 3000 photos in just a few days. To do so, he used a “user-friendly” document scanning app that can speed-up the process of scanning: “that was very efficient but it doesn’t do an automatic text recognition so what I need to do is I need to do the text recognition later. With Acrobat, it’s is not so bad but it's another step”.



460 - 04 02 06 - 1/10

64 - 10 01 08 23.06.17, 04:43 11
64 - 10 01 07 23.06.17, 04:40 19
64 - 10 01 05 23.06.17, 04:38 10
64 - 10 01 04 23.06.17, 04:37 19
64 - 10 01 03 23.06.17, 04:35 14
64 - 10 01 02 23.06.17, 04:35 33
64 - 10 01 01 23.06.17, 04:33 6
196 - 05 01 07 Press rele... 23.06.17, 02:07 3
556 - 02 04 01 22.06.17, 22:43 1
556 - 02 06 05 22.06.17, 22:37 9
556 - 04 03 01 22.06.17, 22:32 10
64 - 10 01 overview 22.06.17, 19:44 2
64 - 10 01 06 22.06.17, 17:13 1
129 - 05 01 01 21.06.17, 21:36 1

Ad Download Genius Sign!
Sign, date and annotate documents on the go



I scanned the shit out of this archive!

Figure 4: (left): Colin’s Document Scanning App

Figure 5: (right): Using the Scanning App to capture reflections on the fieldwork

The application was connected to Google Drive, where he stored the scans as PDFs together with videos and pictures captured in the field. Apart from the Cloud, he also has a big local folder in which “I basically have all my articles and research papers and presentations that I’m working on, so this is more like my actual work, no matter what it is”. Due to space constraints he also uses Dropbox for uploading yet more material:

> “and then of course restrictions like Dropbox and Google drive is only so many gigabytes and maybe the research is much more so I need to put them in the different systems just to get what I want, which is a good backup. Of course we could use a University solution which may have unlimited or I don't know 50 GB or 15 and of course I could probably put more stuff together”.

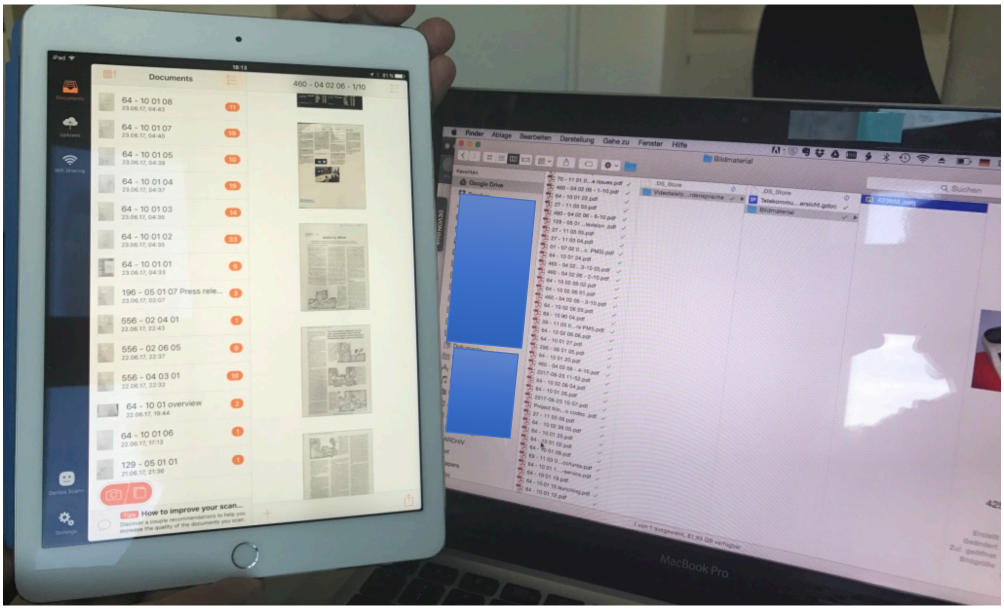

Documents
460 - 04 02 06 - 1/10
How to improve your scan...
Finder Ablage Bearbeiten Darstellung Gehe zu Fenster Hilfe
Bildmaterial
Google Drive
MacBook Pro

Figure 6: Screenshot of Colin's document scanning App and the related Google Drive folder where he saved and stored the files

He did express a willingness to move to an institutional solution, but only "if it works in the same way as Dropbox or GDrive!"

While Colin prefers commercial, user-friendly cloud solutions connected to applications, Lucy, a PhD student in sociology, has a local folder in the centre of her desktop. She has all the important materials she is currently working on under her direct view. In the folder she mainly has the interviews, pictures and videos she captured in the field, but also a back-up of Maxqda (a qualitative data analysis software tool): "in Maxqda I don't have all the interviews I have at the moment but I will have, we have protocols from the fieldwork and observations too but these are in my notebook, I haven't transferred it yet into digital form".

Lucy writes up her ethnographic data in a notebook and she mainly focuses on interviews. Many ethnographers work with notebooks in this way and, once again, this underscores the way in which what counts as data is constituted in a set of discipline-specific and situated practices. Notebooks are typically indexical of the larger body of fieldwork in ways that are highly particular to the individual researcher. Yet this is usually lumped into the basket of 'ethnographic data' with little hesitation. This is further elaborated in the following observation: Julian, an ethnographer and anthropologist by training, collects ethnographic data as a core part of his work. He started his PhD in 2016 and spent the first six months in the field. From the outset he was concerned with how to organize his data collection:

> "the only real thing that I did before I went to do my field research was to think about how I wanted to organize my data collection…. I decided to use Citavi for most of it because I worked with Citavi before to manage my literature, I decided it might be also a good tool to write my notes. Because I knew how to work with it already and the most interesting thing for me was that I can just search globally everything that I put down in Citavi. Because if I thought about making like Word documents for each day like a diary but the problem that I came up with was, if after this year I remember that I once wrote something about this and that situation, how am I able to find it? Do I remember the date? I thought … It's highly not sense to do your project that way.. So I thought it's best to put everything into Citavi because then you can just like search it" [#13: Julian, anthropology]

His whole data collection is organized and structured in a project folder saved in the cloud with Citavi. He found this convenient because he could comment, tag, search and organize data according to his needs. He was also already familiar with the application. Using Citavi as an ethnographic diary allowed him to create a project in which to manage every note written. The "fieldnote project" created in Citavi contained several single files divided by month of observation and every single note was tagged with annotations about its content. The drawback of this is having his data collection bound to Citavi itself. Thus, he will only be able to access his data collection as long as Citavi remains in business.

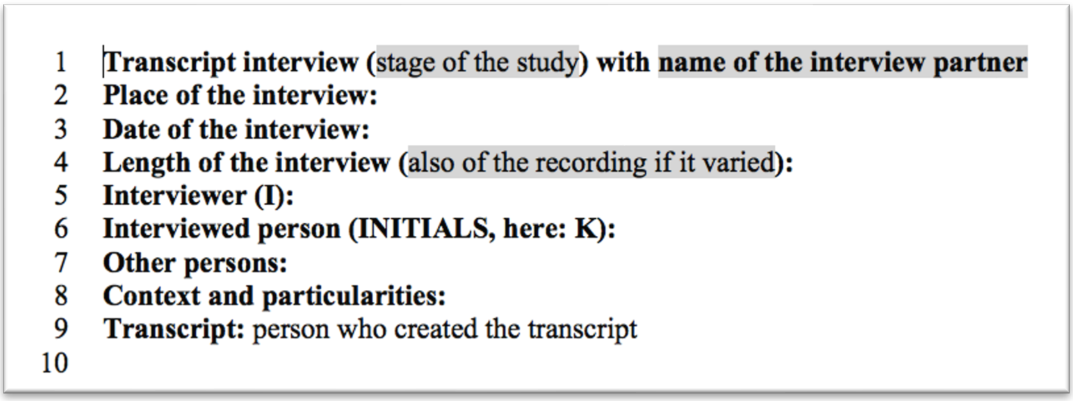
5.4.1.2 Metadata: what is metadata?

Institutional approaches in RDM presume metadata creation to be a fundamental activity of the research process, closely connected to the collection and organization of data but also critical for documentation and secondary use. However, when asked about metadata creation, most of the interviewees said they had little or no understanding of what metadata actually is, what its definition might entail, or what it might be used for. Thus, it is hardly surprising that they typically chose to either ignore it completely or, in rare cases, tag data in local and informal ways. Metadata is often described as "data about data" or "information about data". Edwards et al. (2011) define it as the information needed to share with others in a meaningful way, a sort of "everything you need to know about my data". If so, then - *prima facie* - systematic data sharing is not currently taking place. When asking researchers what is needed to share their data with someone else in a meaningful way, a list of contextual information is usually provided:

> "so these are some protocols of the interviews with some information, like the name, the age, what the people are doing, how the interview came about, what the communication was before the interview, what was

> the interview like, where it took place, how was the atmosphere, were there breaks or pauses for something for some reasons, what the people look like, what are some aspects there were in the minds of the people who did the interview that could be interesting for further research and so on … if we would give or share data it would be useful to have also these protocols and also the questions we actually asked to understand what we did" [#3: Alvin, Sociology]

This suggests that metadata in qualitative research is provided by describing the context in which protocols are made use of. Field protocols are data but also metadata. The protocols are often text files, most often Word documents, where detailed information is displayed. Researchers normally provide information in these documents about how they approached the field, what was memorable or relevant, the physical layout of the setting, the 'atmosphere', and so on. What is striking is that, although this information is often present, it is seldom structured in any consistent way, although people using software packages such as Maxqda or f4 transcription say they find the headers extremely useful:



Transcript interview (stage of the study**) with name of the interview partner**
Place of the interview:
Date of the interview:
Length of the interview (also of the recording if it varied**):**
Interviewer (I):
Interviewed person (INITIALS, here: K):
Other persons:
Context and particularities:
Transcript: person who created the transcript

Figure 7: Screenshot of the header of an interview file highlighting possible "metadata"

Given that researchers usually provide information like this somewhere in their documentation, it is reasonable to assume they find it useful. The length of interviews, for instance, is used to calculate how much data in total has been recorded during a study. This information often features in the methodology sections of published papers. However, it can be difficult to distinguish between metadata and data *per se*:

> "I don't know if I create metadata. Maybe I do in doing those Citavi things and keywords, it's kind of information about the information that I collected, right? […] I will create lots of reflection on how I gathered my material. But it's more reflection and not exactly metadata. Maybe you could say it's kind of metadata because its, you look at the way you gather the data and the way you work. So if that is the thing you meant with metadata then I would say it is definitely a big part in an anthropological dissertation. But I don't know, I think myself, I am not a metadata person" [#13: Julian, Anthropology]

What Julian recognizes is the fundamental role of reflection and contextual information about his own material, which he classifies with keywords and tags using Citavi. Given its unstructured nature, however, it is not clear it can be construed as metadata in the sense that Edwards et al. (2011) use it, e.g. meaningfully shareable. It also suggests that the point made earlier about 'raw data' extends also to metadata. It is the reasoned situated product that cannot be divorced from the specific research practices and preoccupations associated with its production. Alvin expresses further concerns about the shareability of ethnographic data when it constitutes "private documents for the people who wrote them, their personal emotions, experiences in the field so it would need a lot of trust to trust in other colleagues to share that, at least to share that with unknown people". Again, this resonates with what we already know about reluctance to share data with a broad public (Gupta and Müller-Birn 2018; Kervin, Cook, and Michener 2014; Eschenfelder and Johnson 2011)

5.4.2 Open Science perspectives

5.4.2.1 Publishing and sharing data

While there is scepticism about sharing data with unknown audiences (both in the public and scientific domain), there are cases of informal sharing across the research contexts we investigated. We encountered two such examples, respectively in the CRC and in the interdisciplinary department.

In the CRC, interview data was shared with researchers from other projects in order to have collaborative analysis sessions. The researchers found this useful because they considered getting an outside perspective on the data to be important by potentially improving the quality of the analysis and giving them an opportunity to learn from more experienced peers. Excerpts of anonymized data were sent to participants via email a few days before. The overall interview data was described in an introduction where the data collector explained any relevant background that might prove useful. This included:

(1) The research question(s): *"Our interests in these interviews centre on…"* (where the research object and field of study were specified). *"We are interested in…"* (where the research questions were made explicit);

(2) The reason for choosing one specific segment: *"The present material is a 20-minute excerpt from a 2-hour interview. The material is really hard to anonymize when we share transcripts in full – which led us to this unconventional selection that we are comfortable with*

sharing only in this restricted group. As customary with the data sessions, please do not share the material any else".

(3) A summary of the rest: *"The whole interview proceeds through several phases. It starts with a biographical section about the profile, disciplinary background, and experience of the interviewee".*

(4) Biographical information about the interviewee: *"The interviewee is male, has 4+ years of research experience and some (limited) computer literacy. The interviewee uses qualitative empirical methods in his work".*

The structure of the data provided, and its content, reflected specific local needs. Data was added, truncated, withheld and otherwise managed with a view to the work to be undertaken. In the interdisciplinary department, a PhD student decided to share his own project folder on Sciebo and asked via a group telegram channel if others wanted access to it. He also created another folder in which he asked people to upload books and shared knowledge across projects. Immediately, ten out of twenty PhD students in the department accepted the invitation and got access to the folder. Cliff commented on this, saying:

> "I am happy that he does it. I wouldn't share my whole working project folder with all of the group and I don't see so much direct use of him sharing it with us […] But maybe it is more interesting to have like folders collecting all the proposals across projects. Or collecting all the milestone presentations across projects […] I don't want to go into each project and figure out like, where is the budget in this messy project, I want a folder with all the budgets. For now, it's nice that he shares it, but I don't know if he should share it because there is also empirical material there, there is personal information in there". [#16: Cliff, Business Information System]

This objective, here, was to increase the degree of awareness across different projects. Such actions are unusual. Our data shows very little evidence of data sharing between groups. Indeed, there is little overall awareness of what others are doing outside of one's own group:

> "That's a mess. Like we use some of the stuff of INFRA [the university IT service provider], we use some of the stuff from our own IT support, and then some projects do their own stuff and no one knows, there is no overview, there is no shared resources, there is no awareness of what other projects are doing "Oh you did it like this and that and we could have done it like that as well". But, you know, no one knows" […] I would like to have a shared data storage again … Like having a better infrastructure for getting a better awareness of what's going on… I just would like to know more about what other colleagues do". [#16: Cliff, Business Information System]

In this department, several projects were being conducted in the same domain, but there was little or no evidence that data was shared between them:

> "I would love to have time check the qualitative data, we did like sixty or seventy interviews […] Susanne (a colleague working in the same domain) doesn't have any access at all, because it stored on the BSCW […] and she would, she needs to know that this exists […] I don't know how if the others have also folder like this but we have a lot of work but no one except people that belong to this project know about this data" [#14: Aaron, Business Information System]

5.4.2.2 Preserving data: archive and documentation

After data sharing, long-term preservation is the most fundamental concern of data curation. The data lifecycle suggests this stage involves activities like: migrating data to the best format/media; storing and backing up data; creating preservation documentation; and actually preserving and curating data. Philipp is a computer scientist. Using machine learning as an example, he explains the difficulty of storing large volumes of data for long periods of time, something that is compounded by machine/hardware updates:

> "This paper for example has 5 tables and 23 figures. So, you can imagine how much effort it would be for a single paper to have this process for each of the graphs stored? I don't know how to do that, I have no idea. Without hiring five people doing that […] Sometimes we have that problem when we try to compare our results to other results then we get software from somewhere else which is older we have the same problem, to make the machine to run that software […] So, I don't know how to take care of it. So, I ignore it, even if I know I shouldn't. But I have no solution to that" [#15: Philipp, Computer Science]

Long-term preservation is also associated with the documentation that forms the basis of data sharing. Without documentation it is impossible for others to understand the context in which data was created, collected and analysed. However, as we have already noted in our examination of bottom-up practices above, both social scientists and computer scientists engage in practices that are highly idiosyncratic, writing notes, codes or ethnographic reports mainly for themselves in their own style. As Carl put it "protocols are written by me for me … a memory tool in order that I do not forget what I experienced in the field".

Apart from being potentially idiosyncratic and intended for personal use (or only limited sharing), research and research data is also experimental, with "very chaotic", "messy" procedures. This impacts the possibility of documenting something that might not be finished or useful:

> "We have to set priorities and we just don't have time for this documentation. I just try to insert some comments for me and maybe for another person but it's not always possible because something I implement some functions as a test function let's say, then I implement it and it's already changing and doesn't make sense to describe it if I still don't know what this function exactly does […] That's a little bit chaotic and its maybe a lack of time" [#26: Alan, Computer Science]

The "main work" is not preservation of the information. Curation, rather, from the viewpoint of the researcher, can be thought of as another kind of articulation work (Strauss 1985). The pressure for a publication outcome influences how research data management is performed and the quality of the archive, documentation and preservation. A researcher's priority is typically to get as many publications as possible, get a PhD, or provide project results as soon as possible:

> "It's not only my personal problem, I have seen different programs done by another researcher and its normal if you are a developer and code for a problem, you just do it in support for your publication […] It's not done to be read by another person. But in some cases, it will be done and, in that case, it will be very difficult to understand the code". [#26: Alan, Computer Science]

To add to the point of how data may be shaped according to specific concerns and practices, data is collected and structured "around publication outcomes", around the need to find novelty in the field of research.

5.4.2.3 Re-using data

Data reuse closes and at the same time reopens the lifecycle. This step allows "data objects" to gain, in principle, a new life and purpose through secondary use. It allows the cycle to start again, iteratively. Once again, the problem is the type of data and the documentation needed for it to be understood by others. Paul's data collection is created with "a very specific purpose", making it hard and time consuming to prepare for others, such that "the problems heavily outweigh the benefits".

> "We have a very specific research question, that we will follow and the data would only be useful for somebody who has the same […] you need so much extra information from the observations, from being there, from talking to the people in order to correctly frame what they say in the interviews. It's not only extremely

> time consuming to process it in a way for others to be able to use it and then if you would, it would be useless to them […] So I would be happy if the university would store it and would say "I give you a lifelong access to our service. You keep the University … Email address and with that, and you log in, you can always log into and get back to your data. But then again we can just keep it personally" [#9: Paul, Sociology]

Note that he supports the idea of having an infrastructure for secondary use, but only so that researchers can revisit their own data in the future. Indeed, researchers find it difficult to imagine what the characteristics of a secondary use infrastructure might be. They give little thought to what kind of data should be published, for what reasons and for whom. They are also mistrustful of the intention of the funding bodies regarding Research Data Management and tend, when discussing such matters, to do so at a relatively abstract level.

> "I think if you are planning or the DFG are planning storing all these data or information one should carefully looking at the type of data which is intended to be stored […] I don't know what these infrastructures would look like and who has access now, later maybe you and your colleagues can establish an infrastructure which will give me the trust that everything will be work out for the good in the end, I don't know how I could judge it even if I could see it" [#3: Alvin, Sociology]

However, Lukas was less sceptical about secondary use of interview data, at least for internal use or learning/teaching purposes: "interviews are not as personal as ethnographic data I think, you have the transcripts which are kind of an objective translation of what people said on the audio tapes […] I wouldn't have a problem with the sharing these interview data if some other maybe a younger researcher comes to me and say "why you did these interviews, can I use them this project with another research questions you had in your own project so if they formulate their own research questions because you can always answer several research questions with audio data I guess yeah why not?!" Lukas mentioned a seminar in which students collected interviews and he, as tutor, and the professors, asked the students to give them the interviews to prepare a publication:

> "we asked the students if they can give us the interviews for this publication and this was kind of considered ok back then, but why?! maybe because they were "just" students doing interviews, I am not sure if I would ask another qualitative researcher for their interview data, maybe if it's old data like the students, or the younger researcher I have just imagined, maybe if it's really old data and I would rephrase the initial research question, "ah! didn't you do interviews on topic X, and asked question Y?! I want to do, I want to take these interviews and show something completely or answer completely different question with that" […] I would frame it very specifically very, because is a kind of a sensitive topic again" [#7: Lukas, Sociology]

Note also the assumption here that interview transcripts will somehow constitute ‘objective’ data. Clearly, however, the conduct of interviews and their transcription is embedded in a body of associated research practices that remain unexplicated within the transcripts themselves, posing questions again about the extent to which data might be considered ‘raw’ or ‘objective’.

5.5 Discussion

Open Science is held to be crucial for the future of academia but, as we have argued, it remains currently little more than an ambition for the kinds of cases we have described. Understanding why this might be so necessitates a careful consideration of the practices of researchers themselves, taking into account the overall research process and its complex ecosystem with its tasks, tools and workflows. Each and every socio-technical element we have analysed relates to data creation, transformation and eventually migration from the private to the public domain. Above, we have shown how the negotiated order manifests itself through a series of tensions that implicate: researcher biographies and their history of tool use, including things like relative status and individual motivations; individual and heterogeneous practices and awareness of the overhead contained in metadata work, along with a lack of awareness as to how it might be produced; naivety about the nature of metadata and how it is to be construed; the difficulty of making metadata ‘fit’ the realities of local practices and in particular the contingent nature of sharing practices at a local level; and various disciplinary and methodological specificities. Below, we tackle these issues under three main headings that capture what we see as the three main ‘gaps’: (1) the policies and practices gap; (2) the knowledge gap; and (3) the tools gap. We suggest it is critical to understand these to address the Open Science vision and allow policies and practices to be aligned in the future.

5.5.1 Policies and Practices Gap: standardization and idiosyncratic heterogeneity

We characterized our work in relation to a ‘gap’ between Open Science policy and the ordinary practices of researchers which may affect and constrain the potential for realization. Here, then, we decompose that general question into two elements. The first one we highlight has to do with the general organizational mandate devolving from the Open Science policy initiative; the second one refers to the nature of data itself.

5.5.1.1 Organizational mandate

The CRC context is especially useful to explain this first element. The CRC is funded by the DFG who demands that researchers release data in institutional repositories at the end of a project and mandate that data be documented and delivered with metadata according to specific standards. Moreover, the DFG claims that, while observing subject-specific requirements, "standards, metadata catalogues and registries are to be developed in such a way that interdisciplinary use is also possible" (DFG 2010). This request sounds extremely ambitious and burdensome considering that, in the interdisciplinary contexts we examined, researchers themselves are called upon to organize data for long-term preservation and secondary use. Currently this is without any help from data managers or curation specialists. This is an important difference between our case and the US LTER network studied by Karasti et al. (2006), where data managers have developed expertise in RDM over decades. Their approach to data stewardship initially aimed to support ongoing long-term ecological research at local research sites. Only later on – with the funder's mandate – did they integrate long-term preservation of data for public reuse. The LTER case is emblematic of the gap between the real-world laborious, ongoing processual endeavour (Karasti et al. 2006) and the demands at a policy level where it is simply assumed that the Open Science initiative will bring about change (European Commission 2010).

In our institution this process is still at a very early stage. The IT service provider of the university struggles to develop solutions that could support data sharing and reuse for the CRC context. Very few "best practices" can be shared so far among other INF projects funded by the DFG. From how to construct a Research Data Management Plan to how to develop solutions for long-term preservation and data reuse is left to each INF project to discover independently (no suggestions are provided from the funders). On the one hand, funders and IT service providers are at the very beginning of this process and they have yet to develop the requisite know-how concerning OS strategy. On the other hand, the researchers have just started to realize and reflect upon the potential impact of OS over their work.

5.5.1.2 Ethics and epistemology

The interdisciplinary research environments we studied present other challenges as well because of the specific characteristics of the data gathered and the particular ethical and legal restrictions associated with this kind of work. Eberhard and Kraus (2018) call the "obvious inconsistencies" between Open Science expectations and the epistemological peculiarities of ethnographic field research the "elephant in the room". The principles of findability, accessibility, interoperability and reusability in these contexts, as demanded by the FAIR Data

Principles, will be implementable only to a limited extent because the "ethical code" intrinsic to ethnographic approaches imposes on researchers the obligation to ensure the confidentiality and anonymity of their informants (ibid). Furthermore, whilst anonymization of data (e.g. to comply with EU GDPR legislation) is typically offered as a solution to confidentiality concerns, this also presents challenges because, the greater the amount of anonymization, the greater the risk of losing contextual information necessary to making sense of ethnographic data.

There is also a question of how to distinguish what counts as metadata and how the contextuality of qualitative research metadata is to be established. The epistemological consequences of this are significant. We have pointed above to Gitelman's observation that 'raw data is an oxymoron', whereby she alludes to the fact that the apparent objectivity of data disguises a variety of factors that go into its selection, its description and its narrative form. In ethnographic approaches the data itself, for instance, often includes reflections by researchers on their own positioning in the field. This can take many forms and be extensive – especially in its unanalysed state. Beyond this, it is hard to see what possible value large amounts of unanalysed data could have to external readers, especially in the absence of detailed contextual information (that may only be in a researcher's head). Furthermore, ethnographic approaches are not commensurate with staged process models of research and data curation. Instead they adhere to a model that is more complexly interleaved. For instance, initial analysis and interpretation of 'data' already starts in the field and continues up until publication. Interpretation, reflection and documentation also continue throughout the research process, incrementally adding descriptions to the materials collected.

A further tension lies in the fact that the drive to harmonization and standardization ignores the idiosyncratic heterogeneities we have identified. Our findings show a huge variety of practices developed by researchers over the course of their careers, influenced by their biographical situation, by their IT skills, their research interests and methodological choices, and their academic backgrounds. Standardization can be imposed from above, but this requires unproblematic 'translation' processes and a tightly disciplined research environment. This will not be arrived at in the short term. Given the significant overheads implied and the possible epistemic limitations inferred by top-down standardization, one wonders whether this can ever be achieved. If, as the motto of the Digital Curation Centre (DCC) attests, *"good research needs good data",* then some serious attention needs to be paid to how those who collect and analyse the data construe the idea of 'good' and, indeed, the idea of 'data' itself. Our findings show that what is "good data" in current ethnographic research is still an unresolved question for practitioners themselves, let alone imagining what it might connote in the context of Open Data and Open Science. How to deal with potential incommensurabilities probably lies in reaching agreements about the kinds of metadata that best represent the nature of the work done and the epistemological assumptions embedded in the data. This is, to say the least, no easy task.

5.5.2 The Knowledge Gap: data awareness

The second gap we identified relates to knowledge in the digital curation domain. The level of knowledge about Research Data Management (RDM) and digital curation amongst the kinds of researchers we studied is generally poor. Our subjects were knowledgeable, aware and concerned about some of the ethical issues and possible legal consequences implied by data sharing in relation to ethnographic research, but the more technical aspects of data curation were not fully understood by many. Thus, for some researchers, the term 'metadata' is not something they can explicitly relate to their own research practices. Research Data Management and digital curation demands the acquisition of specific skill sets together with a certain kind of 'data awareness'. Clearly, training around these topics will help but there is little value in this being purely generic. As an example, in November 2016, the American Anthropological Association organized a panel about the specific work of anthropologists regarding data organization, preservation, metadata cores, access and retrieval, archiving and policies at individual, institutional and federal levels. Freeman and Crowder (2016) in their contribution, recognized as an imperative that anthropologists understand both the technical side of RDM (organizing, sharing and storing their data) and its ethical implications (e.g. who will have access to this data and what they will – or can – do with it). How this is to be done is entirely non-trivial. There is, so to speak, an issue to do with the social distribution of expertise. While there is considerable expertise 'out there' in relation to the character of data and its subject-specific management, and there is considerable expertise 'out there' in relation to the general principles of data curation, these expertises are not always co-located. It would follow that institutionally knowledgeable parties need to work closely with researchers from specific disciplines to align institutional knowledge and expectations with the epistemological and methodological understandings of particular groups of researchers. One area where the organizational structures, as thus far constituted, seem inadequate lies in the fact that no provision has as yet been made for ongoing data curation. The literature discussed above, and notably Karasti et al. (2006), strongly suggests that 'success' results from taking curation seriously and from the ongoing development of the necessary skills. Identifying where those skills are located would be a necessary first step.

We have also identified a knowledge gap regarding studies of the actual practices of researchers applying qualitative ethnographic approaches from the point of view of data management and digital curation. The majority of the studies here (Van den Eynden et al. 2016; Scaramozzino et al. 2012; Tenopir et al. 2011; Gooch 2014) report data from surveys that only partly cover HSS research (but see Broom et al. 2009; Asher and Jahnke 2013). Furthermore, discussion of the major ethical, legal, and technical concerns is not tackled from a practice perspective. Some other texts provide normative instructions (UKDA 2014) and application cases regarding how to use secondary qualitative data for teaching purposes (Bishop 2012). However, when it comes to discussing in detail how to provide metadata for the wealth of different kinds of ethnographic data and materials so that it may meet the needs of long-term preservation and reuse, little to nothing is available. This study is the first attempt to highlight this gap. Through our findings we have been able to show something of how researchers practically deal with metadata. However, it is clear there is confusion and some serious imponderables here so, whilst metadata creation is an activity already performed by the researchers we have studied and central to the conduct of ethnographic and qualitative research, there is an urgent need for more investigation to understand how to better support it, reduce the overheads and link it to the requirements of long-term preservation and reuse. More than this, though, a key gap is that many interdisciplinary researchers do not currently see themselves as re-users of ethnographic data.

The notion of an ‘Open Ethnography’, where ethnographers use as a matter of course ethnographic data collected and curated by someone else is thus far entirely unrealistic. There are very few studies that make use of curated and archived ethnographic data (exceptions include: Kelder 2005; Gillies and Edwards 2005) or that engage with the challenges it might present. Curating data and reusing data are two sides of the same coin – one can learn from re-using archived data about how to improve data management and curation practices – but at present this is a near vacuum and we need studies of ethnographic data reuse. Our own work here has surfaced several possible issues, such as what to describe about the ethnographic research process and what kinds of information would be relevant for reuse. Clearly, the only solution here is further research.

5.5.3 Tools Gap: new tools for digital curation and data reuse

As we shown in our findings, empirical data from interviews, fieldnotes, audio, video files and literature are processed through specific tools created to perform specific tasks (e.g.: data analysis or literature management). Keeping track of what is happening to data within these individual tools is challenging if not impossible. All information eventually gets "packaged" into the tools themselves. While coding and tagging are critical features of some of the tools mentioned in our findings, it is difficult to export processual information in a way that would enable researchers themselves or others to make sense of the processed data or of the analytic process itself.

When it comes to file sharing systems, solutions like Sciebo, Sharepoint, Google Drive and Dropbox do not support any structured metadata creation or tagging during the research process. As already expressed elsewhere (Bietz and Lee 2010), metadata are collected idiosyncratically in a variety of ways and the databases used by researchers do not adequately support metadata creation. Metadata or tags are required that can be quickly edited by researchers during the course of a study, elaborated according to need, then eventually exported, shared with colleagues or uploaded in institutional repositories. Currently, once researchers upload documents in a file sharing system as the principal repository of empirical data, they cannot attach any type of metadata to files or visualize summaries/overviews of their interviews or fieldnotes. No data curation tasks can be performed within the private or shared project domain (see Figure 2).

The example of the anthropologist using Citavi to manage most of his ethnographic data highlights an urgency for new tools that can support the everyday "data work", which in the case of ethnography consists of data collection, analysis and interpretation steps that iteratively influence one another. When appropriate tools do not exist yet, some researchers try to adapt existing tools to meet unsolved needs. Data and tools are naturally intertwined, so new tools need to be developed that can specifically register and monitor data flows, data activities and analysis. New tools also need to be designed to support digital curation, including functionalities for iterative and ongoing documentation, the creation of metadata, process descriptions, (partial) anonymization, etc., to be used as close as possible to the data source and allowing for editing by the data creator. Of the many tools for qualitative research that are currently used by researchers, none are specifically designed with data curation, long-term data preservation and reuse in mind.

While we believe metadata and more structured procedures are needed, they will require better technological support to reduce the overhead. As noted by Birnholtz and Bietz (2003) and others (Zimmerman 2007; Edwards et al. 2013), metadata alone will not be sufficient for meaningful data reuse. Thus, tools will need to support "data negotiation" between data producers and data consumers. Researchers who create the data need to be able to choose who to share it with and whether to offer extra information that might not have been recorded in the original metadata.

Based on our current findings and analysis, these new kinds of tools would need to: (1) Support ongoing research whilst also enabling curation in situ and being long-term preservation oriented; (2) Reduce the overhead of describing data, processes etc. by supporting automatic extraction of metadata/contextual information that can then be edited by the researcher, while the final say regarding what to extract, include and display for sharing will thus reside with the researcher; (3) Raise awareness of research data management and prompt researchers to undertake data management and curation activities; (4) Make use of a data management plan (this is already required by research funders and would encourage researchers to refine it and make it relevant to their own research process); (5) Support communication between data producers, data consumers and, potentially, data re-users, to facilitate "data negotiation". To properly design such tools, however, requires more research regarding actual research practices in diverse settings. Our own research raised many questions that are still unsolved: To what extent should awareness development, knowledge and skills enhancement be provided? Should workflows be tailorable? Should there be completely new tools for research data management, curation, and preservation or should new functionality be built into existing software tools for qualitative research?

Literature in CSCW has previously investigated file sharing activities (Lindley et al. 2018, Voida et al. 2006) and collaborative information management (Rader 2009; Marshall and Tang 2012; Marshall et al. 2012; Voida and Mynatt 2006) in the contexts of academic practices but also beyond. Several prototypes have been explored and developed that tried to solve the issues here addressed (Yoon et al. 2016; Chang et al. 2017; Cadiz et al. 2000; Voida et al 2006). Although the challenges that Open Science pose have been, to a degree, recognized, they entail a new level of complexity. The institutionalization of data curation practices and its challenges is likely to change the way research is performed. This requires a better understanding of the use of data in practice but also the development of reliable infrastructure and tools built in a way to help negotiate OS objectives, stimulate self-reflective and learning processes and support discipline-specific data practices.

5.6 Conclusion

This paper has concerned itself with the relationship between generic policy and heterogeneous practice. It is unique insofar as it constitutes a study of existing interdisciplinary and largely qualitative data practices which take place before policies are implemented and which will undoubtedly affect the success or failure of possible futures. Our aim has been to bring out certain specificities that have been understudied in the literature but that are of fundamental interest to Open Science. We suggest that careful analysis of this work setting demonstrates both the presence of gaps and reflect on how they might be closed. We have shown empirically that there are obvious discrepancies between the Open Research Data mandate and the subject-specific research practices and needs identified above. "Openness" should ultimately, in principle, help to increase the quality of research, improve research methods and enhance reflexivity in our own work. However, at the same time, "good data quality", how it is to be construed and what development processes and implementation procedures are to be followed remains underexamined. CSCW has consistently demonstrated the gap between policies, mandates, rules and procedures and the pragmatic ways in which they are oriented to and negotiated. We pointed out above that, in the context of scientific collaboration, CSCW research has developed this argument through a focus on socio-technical infrastructures, cyberinfrastructures and the infrastructuring process. As we have shown, Open Science agendas evidence the same issues but, given the features we describe in section 2.4, with additional levels of complexity. Our data suggests certain features of possible salience that we summarize below.

Local data sharing routinely takes place in heterogeneous ways. For obvious reasons, much of it takes place within projects or across projects. These familiar occasions of sharing data offer opportunities for researchers to reflexively address data management and sharing issues regarding, for instance, recording of project histories, methodological decisions, the various kinds of data collected and used within projects, and bibliographic material. Insight into local collaborative and individual practice, we have shown, provides a basis for development of relevant and useful data management and curation practices.

The description of data storage practices and a concomitant understanding of the practices of *data sharing*, we suggest, are the first steps in the managing and curating of data over the long-term. Data sharing for a wider audience is likely to be a more complex issue. This cannot be left only to researchers. As we have seen, they are not motivated, lack the necessary knowledge and/or tools, are often not granted the necessary resources, and do not see data sharing to be an important feature of their day-to-day work. At the same time, curation cannot be left to professionals who have the technical skills but lack knowledge of the disciplinary and interdisciplinary specificities of the work. Instead, researchers and data managers and curators need to learn from each other to evolve a mutual understanding that can facilitate the development of new practices, methods and tools.

Furthermore, as with the proliferation of new data specialist job descriptions in 'big data' environments, our research suggests a need to consider what kinds of new roles for data managers or curators are needed for qualitative/ethnographic research. These roles should provide support and knowledge about the standards and regulations policymakers constantly update. However, they should also be able to encompass negotiation and a deeper understanding of research practices, as evinced in the sheer curation and US LTER examples we've described.

We call for a negotiation of standards between researchers, data curators and policy makers that recognizes the practicalities of data work. Just as participatory design principles are founded on mutual learning (Halskov and Hansen 2015; Simonsen and Robertson 2013). We see the development of the necessary skills in the same light. The evolution of research data management and its sociotechnical solutions will be an ongoing, long-term, process that entails learning. This has to be predicated on a consideration of the division of labour and how that is negotiated, on an awareness of the kinds of contingency that arise and that might problematize development, and on a recognition of the different understandings of organizational members.

Lastly, we have identified a technological gap that needs to be filled and that could be supported by CSCW research. Open Science objectives will not be met without the development of new technological solutions that can support digital curation, long-term preservation and data reuse. While we can anticipate some of the tools that might be needed (e.g. for metadata recording and editing, data negotiation, etc.) this also calls for further investigation. In this sense, this paper also calls upon the CSCW community to join the Open Science discussion in order to get a better sense of the various contexts in which digital curation activities will evolve over time and the tools which will prove relevant and useful.

Implementation involves complex socio-technical elements and has to be regarded as a long-term, evolving, objective. It is likely that many different kinds of attempts will emerge to address data management, curation and preservation challenges in ethnographic research. The necessary expertise for dealing with the kinds of sociotechnical issues we have raised in this paper lies within the CSCW community, for it is in this community more than any that socio-technicality is recognized as being to do with practice. This paper has therefore sought to give researchers, scientists, decision-makers, politicians, IT service providers and other stakeholders an overview of the *grand vision* behind the current changes in the fields of data management, preservation and curation and to surface how this ramifies for, and is influenced by, current practices.

6

Designing a Data Story: A Storytelling Approach to Curation, Sharing and Data Reuse in Support of Ethnographically-driven Research

This chapter was published in: Mosconi, Gaia, Dave Randall, Helena Karasti, Saja Aljuneidi, Tong Yu, Peter Tolmie, and Volkmar Pipek. "Designing a Data Story: A Storytelling Approach to Curation, Sharing and Data Reuse in Support of Ethnographically-driven Research." *Proceedings of the ACM on Human-Computer Interaction* 6, no. CSCW2 (2022): 1-23.

Abstract. In this paper, we introduce an innovative design concept for the curation of data, which we call ‘Data Story’. We view this as an additional resource for data curation, aimed specifically at supporting the sharing of qualitative and ethnographic data. The Data Story concept is motivated by three elements: 1. the increased attention of funding agencies and academic institutions on Research Data Management and Open Science; 2. our own work with colleagues applying ethnographic research methods; and 3. existing literature that has identified specific challenges in this context. Ongoing issues entailed in dealing with certain contextual factors that are inherent to qualitative research reveal the extent to which we still lack technical design solutions that can support meaningful curation and sharing. Data Story provides a singular way of addressing these issues by integrating traditional data curation approaches, where research data are treated as ‘objects’ to be curated and preserved according to specific standards, with a more contextual, culturally-nuanced and collaborative organizing layer that can be thought of as a “Story”. The concept draws on existing literature on data curation, digital storytelling and Critical Data Studies (CDS). As a possible design solution for Research Data Management and data curation, Data Story offers: 1) a collaborative workflow for data curation; 2) a story-like format that can serve as an organizing principle; 3) a means of enhancing and naturalizing curation practices through storytelling. Data Story is currently being developed for deployment and evaluation.

6.1 Introduction

For at least two decades, academic institutions have had to deal with the major changes implicated by a move towards the so-called Open Science agenda. This has the potential to reshape the cultural, organizational and infrastructural academic landscape (Bartling and Friesike 2014). In fact, most Western governments and all their major funding institutions fully embrace this agenda, with the clear intent of ensuring the verifiability of findings, promoting good scientific practice, and providing greater returns on public investment by encouraging data reuse (Wallis, Rolando, and Borgman 2013). To satisfy these objectives, data repositories and data centres are proliferating and many funding bodies now mandate the creation of research data management plans (RDMP) and the implementation of Open Data policies that embrace the "FAIR Data Principles" (Wilkinson et al. 2016), i.e., research data deposited into archives should be Findable, Accessible, Interoperable and Reusable. Knowing how to efficiently structure, manage and curate data in order to fulfill expectations regarding long-term preservation, sharing and data reuse is becoming a sine qua non condition for receiving research funding. However, despite political and infrastructural efforts, the Open Science agenda remains some way from being realized and its ambitions have proven to be especially challenging for Humanities (Fenlon 2019; Rawson and Muñoz 2016) and Social Sciences (HSS) scholars (Mozersky et al. 2020) for whom these requirements are relatively new. Indeed, not all data are created equally and for some disciplines is much harder to adjust to these demands due to the nature of the data collected and the methods applied. Within the Social Sciences, researchers working with qualitative and ethnographic data are confronted with particular legal and ethical issues (Mosconi et al. 2019; Eberhard and Kraus 2018), the personal character of the data can make researchers unwilling to share it in its totality, it can be hard to see what counts as metadata or how to curate qualitative data for sharing, and the sheer heterogeneity of data and data management practices can make standardization massively problematic (Ryen 2011). As a result, the sharing and meaningful reuse of qualitative data remains rare, outside of teaching contexts (Bishop 2014; 2012), nor are concrete solutions – beyond data archives or data repositories – being successfully implemented and regularly used as yet (Mannheimer et al. 2018).

In our view, critiques of openness should be taken seriously. There is growing consensus that the mere release of data is not enough to realize the full potential of openness (Zuiderwijk et al. 2012; Mosconi et al. 2019). In particular, open data portals or data archives are prone to becoming 'data dumps', where the number of published datasets is more significant than their quality or utility (Nelson and Simek 2011). Open data portals or data repositories are typically all about the structuring of data and the policies that surround it: how many datasets, how many formats, which open licenses and so on. While formats, standards and licenses are necessary for the long-term preservation of 'data objects' and their retrieval, there are still few design solutions that specifically support the practices and workflows necessary for interdisciplinary collaboration around those objects (Mosconi et al. 2019; Feger et al. 2020b). In response to these challenges and critiques, we present an exploratory and conceptual design solution, called 'Data Story', that offers a particular way of curating and sharing heterogeneous data sources collected by ethnographically-driven research projects that can be seen to better resonate with the interests and expectations of qualitative researchers. The solution aims to support the partial curation of data by encouraging a pre-selection of relevant data that researchers might wish to share that can then be contextualized by making use of storytelling practices. The concept grew out of a long-term engagement within an interdisciplinary Collaborative Research Centre, where we observed researchers working in interdisciplinary ethnographically-driven contexts and engaged in conversations with them about data their practices. The Data Story design can be seen as a way of building upon the current informal sharing practices we observed and of addressing the unsolved Research Data Management issues we surfaced.

The Data Story, as an exploratory and conceptual design solution, has its roots in literature relating to *data curation and sharing* (Bishop 2012; 2014; Dalton and Thatcher 2014; Treloar and Harboe-Ree 2008; Tsai et al. 2016). However, it also takes inspiration from works relating to *data storytelling* (Duarte 2019; Knaflic 2015; Ojo and Heravi 2017), and *Critical Data Studies* (Dalton and Thatcher 2014; Dalton, Taylor, and (alphabetical) 2016; Kitchin 2021). Ethnographic and other qualitative data, historically associated with the social sciences but increasingly deployed in HCI and CSCW contexts, are inherently narrative in character. It follows that something akin to 'storytelling' might be an appropriate focus for the data sharing agenda. As we will be elaborating below, Data Story, as a concept, seeks to supplement traditional data curation approaches by adding a more contextual, cultural and collaborative *organizing layer:* "the Story".

6.2 Related work

There are three principal bodies of literature that delineate the research space this paper is addressed to. One of these reconstitutes data management as a sociotechnical issue and stands as a critique of approaches that assign a certain fixity to what counts as data. Another focuses more specifically upon the sharing of qualitative data and the unique challenges this can pose. The third is concerned with data narratives and data storytelling and the extent to which this has already featured in approaches to data management. We look at each of these in turn below.

6.2.1 Critical Data Studies and the myth of 'raw data'

As Dourish and Cruz (Dourish and Cruz 2018) have pointed out: "Data makes sense only to the extent that we have frames for making sense of it, and the difference between a productive data analysis and a random-number generator is a narrative account of the meaningfulness of their outputs" (Dourish and Cruz 2018). We see this, above all, as an issue of rationale. *Why* is data collected, organized and represented in the way that it is? The desire to embed rationale into data can be traced back to the literature on 'design rationales' in the context of software design (Moran and Carroll 1996; Burge et al. 2008; Demian and Fruchter 2009; Lee 1997). As Lee (Lee 1997) sums up the concern as follows: "Reuse/redesign/extension support... can serve as indices to past knowledge (similar designs, parts, problems encountered)." Big data, however, has prompted an epistemological shift away from relatively mechanical, model-based approaches to problems of storage and retrieval and towards a more practice-oriented view, at least for some. This has been a motivating force behind 'Critical Data Studies' (Dalton and Thatcher 2014; Iliadis and Russo 2016) and various practice-oriented studies in CSCW, HCI and STS. Critical data studies are largely concerned with "questions about the nature of data, how they are being produced, organized, analyzed and employed, and how best to make sense of them and the work they do" (Kitchin and Lauriault 2014). As noted, this was occasioned by a 'step change' in the production and employment of data.

At heart, critical approaches recognize that political, social, ethical, organizational, and economic elements shape data management as much as technical problems. If so, data can no longer be treated as having some kind of 'objective' status. Data, as Gitelman (Gitelman 2013) has suggested, is always "cooked" and "raw data is an oxymoron". The construction and reconstruction of data formats depends on an array of factors, among others the cultural norms of the groups that created them (op. cit.). By way of example, Vertesi and Dourish (Vertesi and Dourish 2011) have shown how data management, including sharing practices, is mediated by the nature of research cultures. Thomer and Wickett (Thomer and Wickett 2020) underscore this in their analysis of the various material forms that the 'database' can take, arguing that "'best practices' for data management are in tension with the realities and priorities of scientific data production", and "understanding pluralism in data practices is crucial to supporting the needs of those traditionally marginalized by information technologies—whether in their personal or disciplinary identity" (Thomer and Wickett 2020). As we shall see, curating for data work as a pluralistic and contextual endeavor has, as yet, not been fully realized.

6.2.2 Challenges for qualitative data sharing

Data sharing have been a topic of intense interest across a number of disciplines in recent years (Heaton 2008; Faniel and Jacobsen 2010b; van den Berg 2008) motivated by the requests for Open Data increasingly mandated by all major funding institutions. Most of the literature points to the many unresolved challenges inherent in preparing data for sharing purposes. Documenting and providing sufficient context for others to understand how data has been gathered, analyzed and processed, and the lack of incentives and motivation on the part of the researchers are seen as the most critical issue (Birnholtz and Bietz 2003; Zimmerman 2007). These apply equally to all types of data and disciplines. However, some disciplines such as the natural sciences have managed to better adjust to these new demands and with time have developed internal policies to ensure the sharing and eventually the reuse of research data (Zuiderwijk and Spiers 2019). For other disciplines these requirements are relatively new and researchers and institutions are still struggling to understand how to meet these expectations. In the Social Sciences specifically, it is recognized that the sharing of qualitative and ethnographic data presents particular challenges because of the epistemological, methodological and ethical complexities associated with this type of data that do not directly apply to quantitative data and/or other disciplines (Tsai et al. 2016; Mozersky et al. 2020).

The epistemological difficulty with qualitative data lies in the fact that it is challenging to grasp *what* the 'context' may be in any precise way and *how to* describe it (Moore 2006). Context determines whether something can be viewed as data or metadata and the "degree to which contexts and meanings can be represented influences its transferability" (Borgman, Scharnhorst, and Golshan 2019). Others have questioned the legitimacy of data when removed from the original contexts, packaged in repositories, and disentangled from the knowledge and expertise of the researchers who performed the study (Walters 2009).

With regard to methodological challenges, it is important to recognize the reflexive character of this type of research (Davies 2008; Marcus 1994). The collection of qualitative data is inherently intersubjective, its analysis is iterative, and interpretation is always a key aspect of the work. Data are often rich with personal content and are neither collected nor analyzed in a linear manner (Tsai et al. 2016). Nor are many data collection activities targeted at sharing and archiving, so the resulting products are not well documented or formatted for others to use (Kervin, Cook, and Michener 2014).

In relation to the ethical challenges, preserving the anonymity of study participants is of key concern. Informed consent stands to become significantly more complex if the sharing of the associated data for public consumption becomes commonplace (Neale 2013; Ruggiano and Perry 2019; Bishop 2009). As ethnographic approaches are generally based on a trust relationship between researchers and participants and can often focus on sensitive domains, there is a risk of this being undermined by the prospect of sharing data with unknown and potentially unaccountable parties. Anonymization of data (e.g. to comply with EU GDPR legislation) is typically offered as a solution, but the greater the amount of anonymization, the greater the risk of losing the contextual information needed to make sense of ethnographic data.

There is also a lack clear standards regarding how to describe and prepare qualitative data for sharing (Antes et al. 2018; Tsai et al. 2016). Data formats are difficult to identify due to the heterogeneous nature and idiosyncrasy of researchers' data practices. Beyond this, time issues may also arise because "the burden of organizing qualitative data for inspection or reuse could easily exceed the work of writing the manuscript itself" (Tsai et al. 2016).

Evidently, the sharing of qualitative data is anything but trivial. Effective sharing and potential reuse will remain problematic until there is an understandable and efficient way of preparing and curating data in a way that is aligned with researchers' practices and data work. In this way, we second the work of Rawson and Muñoz (Rawson and Muñoz 2016) who advocate for articulating new paradigm and practices in the field of Research Data Management that should support the humanistic way of dealing with data and its specific way of producing knowledge. As we shall see, Data Story offers an innovative and lightweight way of addressing some of these complex issues.

6.2.3 Data Storytelling: Guiding principles and insights

The social sciences and humanities have long stressed the role that narrative plays in human life, education and research. As Game and Metcalfe (Game and Metcalfe 1996) argue: "Research is always an interpretative process that involves conversations and storytelling, though the research framework traditionally applies other names such as aims, methods and conclusions. Research conventions are a particular form of storytelling that allows sociologists and historians 'to tell stories as if they weren't' storytellers'" (Game and Metcalfe 1996). Social scientists tell these stories for a range of purposes. In doing so, they attempt to contextualize the 'data' that they work with. However, there is a difference between context as an analytic construct – something that researchers, curators, etc. define – and something that emerges in and is enacted by the work of the participants. Thus, 'context' has no existence outside of the way in which it is ongoingly constructed by participants to an activity. Data, in other words, is a process of enactment. Digital storytelling, we want to argue, is a useful mechanism for reconstructing this process.

Digital storytelling simply refers to the digitally-mediated practices adopted by everyday professionals and organizations to tell a story. They can seek to stimulate emotional responses in recipients and can offer interactive elements. Digital storytelling can be found across numerous fields, including: therapy, education, arts and culture, library science, and management and business (Barrett 2006; Denning 2006; Restrepo and Davis 2003; Kervin, Cook, and Michener 2014; Vecchi et al. 2016; Sturm and Nelson 2016; McDowell 2018). Over the last decade, the advent of big data and the data revolution (Kitchin 2014) has led to western economies and governments becoming increasingly data-driven, leading to a growing focus specifically on '*Data Storytelling'* (Ojo and Heravi 2017). The main argument is that, to understand and use 'data' effectively, it needs to communicate a clear message (a narrative) in intelligibly human terms that enable us to make sense of it ('data sense-making') and understand why it looks (is reconstructed) the way it is.

At heart, data storytelling consists of three main elements: 1) explaining the context; 2) identifying a coherent narrative; and 3) providing effective visualization. Note, again, the emphasis upon *context* here, with it being the producer of the narrative who identifies the relevant context. At the same time, and as with all human communication, the narrative that is produced involves assumptions about its potential audience, which, in turn, recognizes the second active principle in data storytelling, i.e., *narration*. A narrative can stimulate learning, emotions, and drive action through discursive constructions. A story has a beginning and an end, it has a goal, sometimes a moral, and, as already mentioned, an audience that it is designed to engage. The power of narrative can help to share norms and values, develop trust and commitment, share tacit knowledge, facilitate unlearning, and generate emotional connections (Sole and Wilson 2002). The third principle is related to *effective visuals*. Once data is analyzed, and the message and the story are developed, the data needs to be visualized accordingly.

Other literature, largely associated with critical data studies and STS, has focused on scientific storytelling and the creation of stories in a more qualitative fashion (Karasti, Baker, and Bowker 2002; Kitchin 2021; Linde 2001; Vertesi et al. 2016). Based on their fieldwork in the LTER network, Karasti et al. (Karasti, Baker, and Bowker 2002) showed how storytelling is integral to the practice of doing science, but also highlighted the challenges inherent in recalling, identifying and articulating stories while members are immersed in everyday work activities. Vertesi et al. (Vertesi et al. 2016) have also demonstrated that a narrative account of data management practices can help to uncover tensions in personal data management and allows the emergence of what they call "moral economy of data management", which express the "complexity, ongoing tradeoffs, emotional and reflexive components" of individuals' decisions and actions in respect to their own ecosystem of tools and data. Despite the recognition of narratives and storytelling as a useful mean to describe and talk about data and data practices little attention has been paid to how to translate such insights into design solutions. As Karasti et al. (Karasti, Baker, and Bowker 2002) point out, "Stories of everyday technical aspects of data work may be lacking due to data-work being considered something so mundane, even boring that it would be 'oddly inappropriate for an experienced worker to tell another experienced worker a story about daily routine (Linde 2001)'. Yet it is just this tacit knowledge and contextual understanding that is essential for the analysis and design of 'narrative knowledge management systems' (Karasti, Baker, and Bowker 2002)".

To conclude, relatively little attention has been paid to data storytelling for design purposes and even less to its use in the construction and reconstruction of qualitative data for the purposes of data curation and sharing. We will be arguing here that the concept of a Data Story is particularly appropriate when making sense of qualitative research data. Storytelling can thus be used as an organizing principle when curating and sharing excerpts (snippets) of data from heterogenous data collections to facilitate its contextualization.

6.3 Background and Approach

As background to the Data Story, this section details the fieldwork and practical experience of working within a research infrastructure project (INF) that together guided its conceptualization and design. The project INF is connected to a Collaborative Research Center (CRC) that started in January 2016 and is still ongoing. The fieldwork was characterized by an ethnographic approach and comprised observations and semi-structured interviews, as well as long-term engagement and member participation. Interdisciplinary discussions concerning Research Data Management and data practices within CRC's projects took place regularly in the CRC and the principal author's involvement in these provided an opportunity for numerous formal and informal conversations with researchers. These conversations highlighted relevant RDM issues that make it difficult to meet the expectations of funding agencies for data sharing and reuse and therefore motivates the development of a new approach.

6.3.1 Empirical setting

Our research took place in the CRC, an interdisciplinary research center consisting of 14 projects and more than 60 scientists from a variety of disciplines (i.e.: cultural studies, media studies, social sciences, digital humanities, engineering and computer science). Research projects in the center characteristically involve interdisciplinary cooperation, with most researchers using qualitative and ethnographic methods. This interdisciplinarity is further promoted by seminars, lecture series, workshops, PhD forums, and annual retreats, the latter being focused on discussing project updates and aligning research interests and findings. The CRC started in 2016 and completed its first funding period in December 2019. A second phase began in January 2020 (funded until December 2023[29]). The funding agency, DFG (in English: German Research Foundation), first defined and adopted its "Principles for the Handling of Research Data" in 2010. These highlighted the importance of long-term archiving and the accessibility of research data, across all fields and disciplines. The principles are expected to be followed by all DFG-funded projects. A key element of our project, INF, has been investigating how to achieve this goal. INF's overall objective is to support the sustainable handling of research data, to develop and implement Research Data Management concepts, and to maintain the necessary infrastructure for the whole CRC. The INF project, which forms the principal background for this paper, therefore has a double focus: a) the provision of infrastructural services, led and represented by the IT service provider of the university; and b) design-oriented empirical research, conducted by the first author.

6.3.2 Fieldwork activities in the INF project and the research approach

Since 2016, the first author has engaged in monthly meetings with the IT service provider and its developers. These have included technical meetings to discuss unsolved RDM challenges and the brainstorming of design possibilities for a new research data infrastructure that could meet the expectations of the funding agency. One of the major challenges discussed was the lack of standard solutions for curating and sharing qualitative-ethnographic data. There was also a concern about how to resolve the top-down nature of the RDM mandate with the specific needs and interests of individual researchers. This prompted several rounds of empirical research to gauge how to proceed. Ethnographic observations and qualitative interviews were undertaken, largely between 2017 and 2019, that involved nineteen researchers representing all the major disciplines, roles and positions.

ID	Pseudonym	Background	Academic Role
#1	Sophie	Media Science	Principle Investigator
#2	Joe	Media Science	PhD Student
#3	Alvin	Sociology	Post-Doc, Project Leader
#4	Lucy	Sociology	PhD Student
#5	Mary	Law	PhD Student
#6	Rupert	History	Principle Investigator
#7	Lukas	Sociology	Post-Doc, Project Leader
#8	Mark	Political Science	Project Leader

[29] CRCs can be funded for up to twelve years across three separate evaluation stages (Phase 1; Phase 2 and Phase 3).

#9	Paul	Sociology	Principle Investigator
#10	Carl	Sociology	PhD Student
#11	Rob	Media Science	Principle Investigator
#12	Colin	History	Post-Doc, Project Leader
#13	Julian	Anthropology	PhD Student
#14	Aaron	Business Information System	PhD Student
#15	Philip	Computer science	Principle investigator
#16	Cliff	Business Information System	Post-Doc
#17	Susanne	Social Science	Principle Investigator
#18	Beth	Political science	PhD Student
#19	Will	Anthropology	Principal Scientist

Table 1. List of the interviewees with their disciplinary background and academic position

The fieldwork focused on understanding research data management practices from the bottom up, with a specific focus on documentation and sharing practices. More detail about these practices and the challenges we uncovered can be found in our previous work (reference omitted). The interviews revealed frictions between the expectations of the funding agency and researchers' actual practices. The funding agency's vision of RDM and the data life-cycle implied that research practices should be targeted at the long-term preservation of research data and ideally support both data sharing and reuse. In fact, while curation, sharing and consequent data reuse are central to the OS agenda, these practices are currently not much of a feature of qualitative research and are not well-supported by any of the tools qualitative researchers typically use. However, over the course of our long-term engagement, during which we undertook plenary discussions, group meetings, and supported researchers in drafting their Research Data Management plans, the researchers reported an interest in innovative solutions that might help them to represent and share their highly heterogenous research data in ways that would help them to organize it and underpin the work of collaborative interpretation. The Data Story concept was grounded in this apparent need. From it we came to see that the showcasing of data 'snippets' and the integration of storytelling practices could potentially support the organization, curation and eventually the sharing of research data, in greater synergies with the researchers' practices.

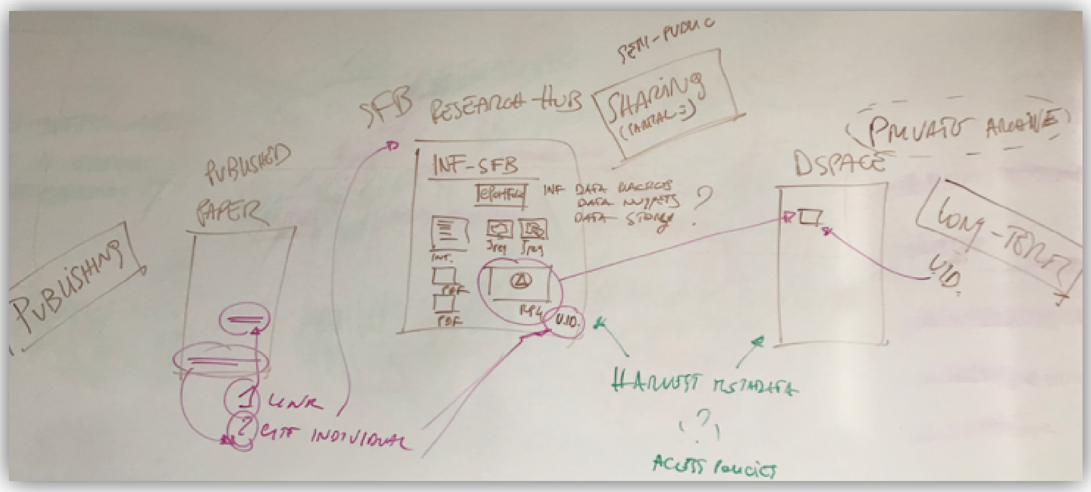


Figure 1: The first sketch of the Data Story idea, generated during a meeting in July 2019.

The first outline of the Data Story idea arose as a sketch[30] during a group meeting in July 2019 (Figure 1). It was inspired by the way that researchers were seen to share data snippets and engage with them on an ad hoc basis during internal meetings. This partial and purposeful sharing was the point of departure. We returned to it and developed the idea further by designing a low-fidelity prototype between January and March 2021. This integrated storytelling components as a way of providing contextual information and complementing it with basic metadata that could support data retrieval. Although the prototype has not been formally evaluated, its design is grounded in informal sharing practices and RDM issues reported by our CRC members during interviews and meetings. We will report on own our observations and related issues in the next section.

6.4 Empirical insights

6.4.1 Data sharing: Informal practices and workarounds

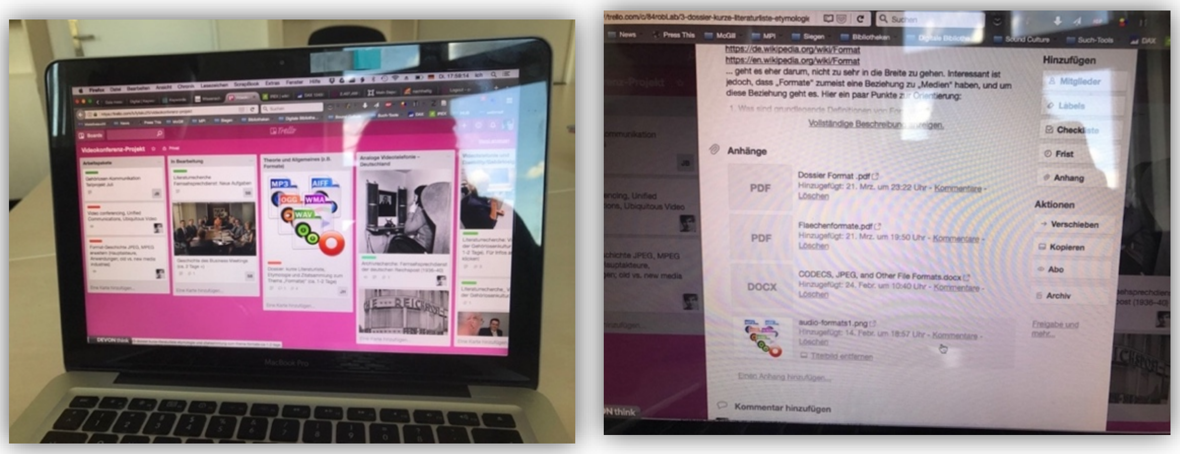
Most CRC projects involve two or more disciplines working together, typically social scientists, anthropologists or media scholars working with computer scientists, designers and/or software developers. Methods for data collection are heterogeneous, often local to the disciplines involved. For ethnographically-oriented projects, the full data collection usually comprises interview files, ethnographic fieldnotes (often on paper), archival documents and

[30] The middle side ‘INF-SFB’ represents the interactive interface of the collaborative platform ‘Research-hub’ through which share heterogenous data. ‘Data nuggets’ or ‘data stories’ are also imagined to be linked to published papers (left side) in order to make other researchers aware of these additional materials. The right side represents a long-term repository that could also be linked to a Data Story.

other types of media, such as audio, pictures and videos. Data is often stored on personal hard drives and/or in Cloud Systems (like Dropbox, OneDrive, Sciebo or Sharepoint). Generally, only the researcher(s) who actually engage firsthand in the fieldwork activities have full access to it. Different data types are distributed across several repositories and almost no one in the project has an overview, not least because few people anticipate a need for full access. Instead, during research meetings, data that exemplifies putatively 'important' themes is usually presented. This data is pre-extracted from larger datasets with a view to meeting presumed analytic agendas, engaging in collaborative interpretation, discussing major findings, developing design ideas, structuring publication outcomes, and so on. Thus, it is common practice to share 'data snippets' in collaborative analysis sessions with members of the same project (but with different disciplinary backgrounds) and/or with researchers from other projects. These snippets of anonymized data are often enriched with contextual information (e.g.: time and place of collection, atmosphere, informant background, etc.) and sent to participants via email a few days before the analysis session.

At the very beginning of the actual session, a *narration* or, if you will, a *story* that contextualizes the data is often provided by the data collector in written form (i.e. as text), and/or in oral form. The data itself is then often displayed to guide the conversation and promote interpretative work. Through these oral and written narratives, qualitative data is constantly evolving and being co-constructed in a collaborative effort that can engage team members, research advisors, student assistants, fellow researchers and even study participants. Claude, a PhD student told us: "*I like storytelling and I even catch myself sharing data that way, I share snippets of my fieldwork and I add some sort of storytelling to it to give others an idea of what I did or what's the background to a short piece of data I might want to talk about"(Claude, PhD student from HCI, forum discussion on May 2020).* Full ethnographic datasets and, in particular, fieldnotes, are not fully shared, due to concerns about both its potentially sensitive nature and time constraints. As another PhD student put it: "*It doesn't necessarily help if I make my whole notes accessible to all the team members, it will take too much time to read it all, and also, I wouldn't want that either, because it's a rather personal thing" (Julian, PhD student from anthropology, interview on April 2019).* So, even among colleagues from the same research team, qualitative data is often shared only partially. There is pre-work involved in selecting the most relevant data. This may then circulate via email, but also often ends up on different commercial software platforms, like Dropbox or Google Drive, where researchers organize it to foreground what they consider to be most important. Many researchers feel this informal data sharing practice is 'not ideal', mainly because it implies the use of commercial platforms, but also because it results in different chunks of data being spread across multiple platforms or file sharing systems without any consistent structure. It should be emphasized that researchers resort to using these commercial platforms because they are the only obvious solutions that can efficiently support simultaneous collaborative interpretative work around written data narratives.

As platforms change and evolve, so researchers have to constantly come up with new techniques and tools to assist them in communicating ideas and interpreting data. As an example, an historian made use of a Trello board (see Figure 2) to collect and structure heterogeneous data sources. The most important pieces of files, pictures and historical documents were organized into thematic sections, annotated and collaboratively discussed with student assistants.



https://de.wikipedia.org/wiki/Format
https://en.wikipedia.org/wiki/Format
... geht es eher darum, nicht zu sehr in die Breite zu gehen. Interessant ist jedoch, dass „Formate" zumeist eine Beziehung zu „Medien" haben, und um diese Beziehung geht es. Hier ein paar Punkte zur Orientierung:

Anhänge

Dossier Format .pdf
Hinzugefügt: 21. Mrz. um 23:22 Uhr - Kommentare - Löschen

Flaechenformate.pdf
Hinzugefügt: 21. Mrz. um 19:50 Uhr - Kommentare - Löschen

CODECS, JPEG, and Other File Formats.docx
Hinzugefügt: 24. Febr. um 10:40 Uhr - Kommentare - Löschen

audio-formats1.png
Hinzugefügt: 14. Febr. um 18:57 Uhr - Kommentare - Löschen

Figure 2: Picture taken by the first author during an interview. It represents the Trello board of a historian who used it to organize heterogeneous data collaboratively collected together with his team.

Here, Trello provided a great workaround for structuring heterogenous data sources. However, each data snippet or document is not made searchable, single data entries cannot be easily exported from the tool in order to continue analytic work, nor can the data be officially shared, presented or cited by another scholar. This specific researcher was in fact interested in experimental publication formats that would allow him to share and publish oral history interviews as video material together with transcripts and other supporting data.

Something important we found was that qualitative researchers are not opposed to the notion of data sharing in principal. Rather, they actively want to learn from one another and seek to understand what type of data other projects collect and how to organize, share and represent research data in innovative ways: *"you can also suggest (...) to talk to other projects who have similar research data in order to maybe, yeah, think about standardization. Do we need that, do we not need it because we're so small, are there even standards for archiving these types of research data? (...) also, for presenting this invisible work, because making interviews is very time consuming, but it doesn't really show a lot, so to have something like a representation of that would be great" (Colin, Postdoc from Media History, Research Data Management plan meeting on January 2020).*

Based on these observations, Data Story started to emerge as a solution that researchers could engage with at any stage of a project in order to represent snippets of heterogenous research data, engage in discussion concerning data interpretations and develop bottom-up curation standards. What we particularly took from the above is the fact that there are already mechanisms that qualitative researchers have in place for sharing data. However, what does not happen is the sharing of all of the data all at once, if ever. Rather facets of their data are shared, having been pre-treated in certain ways, and those facets are embedded in narrative structures that *premise* the data in certain ways, according to its expected recipients, just as stories are shaped for their anticipated audience in certain ways. Another key observation was that qualitative researchers currently struggle to find consistent ways of doing this, but rather adopt formats, structures and platforms in a piecemeal fashion according to whatever currently seems to be at-hand. Data Story, as a concept, focuses on this nodal point, between the data and those with whom it might be shared. It is not an end-to-end solution, but it seeks to draw upon what already happens naturally and to imbue it with more structure and to make it more conducive to meeting some of the more formal demands associated with the Research Data Management agenda. Thus, it might be seen as a way of facilitating the readier sharing of qualitative research data than is currently the case. In the next section, we explore how Research Data Management, as it stands, is seen primarily as a source of tension.

6.4.2 Research Data Management issues

When the CRC started its "first phase" in January 2016, not all projects were fully aware of the DFG agenda relating to the long-term preservation and accessibility of data that would imply commitment on their part to share data. CRC members showed skepticism regarding this agenda and questioned the expectations. As Carl, a sociologist, told us:

"now we did the interviews and we didn't even know if there was going to be one repository or what that would look like (…) yes, we are a bit skeptical because again when we filed the application for the research project nobody came up with the idea yeah somebody would eventually need to anonymize all that data. Eventually you need to have somebody who does that and that work power was not sort of calculated within the original calculation right?!" (Carl, PhD student from Sociology, Interview on July 2017).

At the time, the university had no Open Access policy guidelines[31] and no long-term repository to offer as an archive service to CRC projects. The INF projects wrote the policy in late 2017 and the repository infrastructure was finalized in June 2021. Meanwhile researchers were assisted with the creation of RDM plans for phase II, but no research projects formally agreed to allow their data to be publicly accessible. They only committed to engage with the long-term archive and even the archival process generated a number of concerns, especially with regard to metadata and the documentation to be deposited with the data. Some researchers even said they would not provide any documentation because *"that is a practice currently not in place"*. Some researchers were curious to know more and wanted to learn how to create the right metadata and documentation, but they were disappointed by the replies they were given. The IT service provider, for instance, suggested they use the Dublin Core metadata and simply shared a link with them. After consulting the link, they came back to us and said: *"we literally have no idea how and when we should be using this?! The standard names and definition are expressed in a very technical language that makes it difficult to understand what is asked exactly, what is the coverage?"* (Lukas, Postdoc from Sociology, RDM plan meeting on December 2019). The proposed metadata had no clear link with the data the researchers collected in folders via file sharing repositories or in their personal hard drives. This made it impossible to support a workflow where documentation practices and metadata entries could be embedded in the everyday business of their data work. The solutions offered at the time largely focused on long-term preservation and sharing, not on the more immediate problem of how to choose what data to share and how to share it as a part of one's everyday work.

A particularly evident problem was the heterogeneity of people's data and the metadata used to organize it. As one sociologist put it, *"these are some protocols of the interviews with some information, like the name, the age, what the people are doing, how the interview came about, what the communication was before the interview, what the interview was like, where it took place, how the atmosphere felt, were there breaks or pauses for various reasons, what the people looked like, how I felt, how they seemed to feel, and so on ... if we share data it needs*

[31] These were finally implemented in March 2017

to have this information, along with things like the questions we asked." (Alvin, Postdoc from Sociology, Interview on April 2017). Other data often found in qualitative work includes descriptions (written and pictorial) of physical layouts, the positionality of the researcher, and difficulties encountered. Some researchers even incorporated information about what they had failed to find out. We would argue that contextual information of this kind is not easily represented in existing metadata structures. Some approaches, such as ethnography, are not at all commensurate with the step-by-step process idealized in the data life cycle. For example, initial analysis and interpretation of 'data' often begins as the fieldwork itself starts and continues until publication. Interpretation, reflection, and documentation also continue throughout the research process, incrementally adding descriptions to the materials collected, which are often enriched with personal reflections and emotions. Sharing, then, is difficult without some level of 'curation' and pre-selection of data. Currently, no processes exist to afford ongoing curation and partial sharing in this way, even though this is a routine feature of qualitative research. The expressed need for flexible contextual metadata, pre-selection and partial sharing all resonates strongly with the existing practices and requirements identified in Section 4.1.

Some researchers wanted to explore solutions that would allow them to record aspects of the analytic process in support of methodological reflexivity: *"that's something I am super interested in. How do you kind of make sense of the different data sources that you are working with? (...) How do you make sense with it in a research process, what kind of decisions are being made and where? And so being kind of reflexive and accountable of your methodological steps is something that I am interested like both like intellectually and also then that motivates to open up not only the data but also the decision process that comes with it" (Sophie, PI in Media Science, Interview on April 2017).* Some researchers were clearly interested in having a tool that would allow them to organize different data sources and support analytic reflection. They argued that this would pay dividends by making sharing and courting feedback more straightforward. Data Story was therefore also focused upon providing such tools to support both ongoing and completed research. In particular, we envisaged an interactive interface that could present their data for comparative purposes, allow for intermediate feedback, promote the ongoing evolution of research data, and potentially provide an alternative publication format.

6.5 The ‘Data Story’ concept and its design

The preceding materials provide a backdrop to the development of the Data Story concept and its design. This concept takes the notion of a ‘story’ as a design metaphor and uses it as a source of inspiration for the representation, organization and description of *partial and situated* research data to be shared with colleagues, and/or with external audiences. As mentioned above, it is not an end-to-end solution. Instead, it takes existing practices, concerns and requirements as a point of departure and seeks to facilitate the establishment of curation and sharing practices. The core idea is to showcase anonymized ‘data snippets’ (interview excerpts, pictures, videos, sketches or any other relevant material) that are organized in such a way as to elicit storytelling practices (in oral and written form) to contextualize the data. Above we noted that we recurrently observed researchers telling stories about their data, but in a relatively unstructured way. Our design seeks to give more structure to that practice, while affording other aspects of the data curation process that meet researchers recorded wishes.

The concept draws on all three affordances of data storytelling identified in the literature by providing: a) a way to contextualize collected data; b) a narrative structure to demonstrate its analytical potential, c) a vehicle for the integration of additional representational elements. We discuss below how ‘Data Story’ was envisaged in accordance with these principles. It should be emphasized that Data Story has not yet been deployed and evaluated. To date, we have only developed a low-fidelity prototype[32]. Therefore, at present, it only has the status of being a conceptual design, albeit grounded in our empirical work. We plan to implement this design in the up-and-coming months as an *independent module* in an existing and established platform called ‘Research-hub’, which is built for team collaboration and sharing and that is already used by multiple research groups in our university.

6.5.1 Research-hub

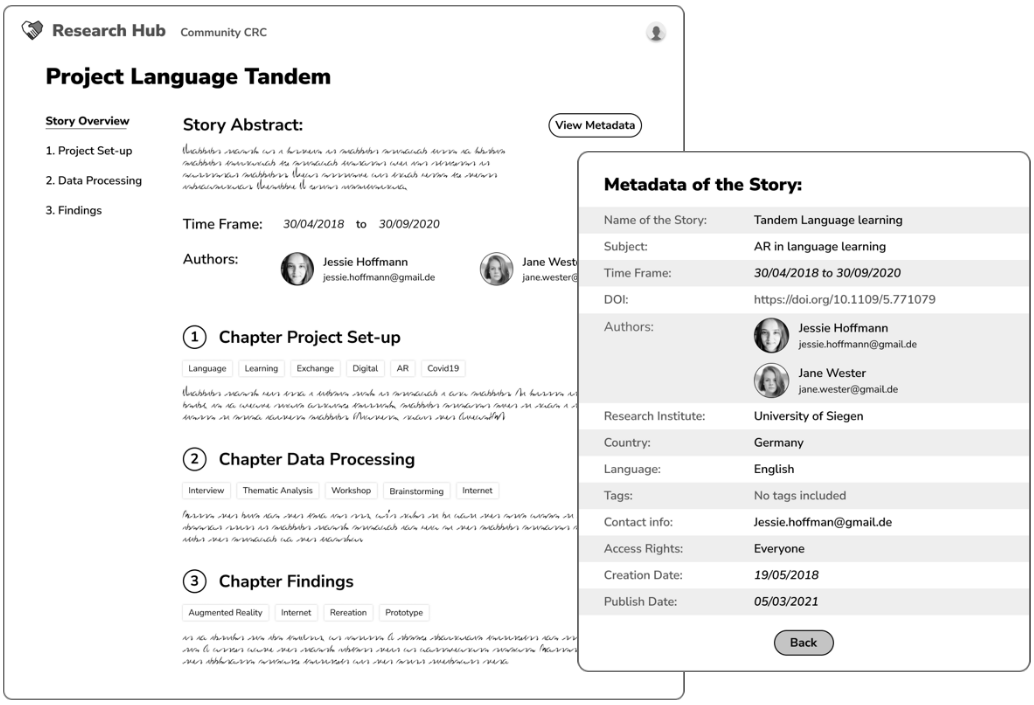
‘Research-hub’ is a customized platform based on Humhub open source software for team communication and collaboration (see https://www.humhub.com/en). Research-hub is already in use in our university as a resource for research project management, academic collaboration (collaborative paper writing, reading groups, etc.) and teaching. In the future, the goal is to also support curation and data sharing practices as well. The platform has a three-level hierarchy: 1) *User profile*; 2) *Spaces*: smaller collaborative units (e.g., research projects); and 3)

[32] **The full prototype can be accessed at this link**: https://www.figma.com/proto/TtFgWU2Oau7njVk9klyZgI/Data-Story-Module?node-id=209%3A12&scaling=scale-down&page-id=0%3A1&hide-ui=1

Communities: larger organizational and institutional units (e.g., Departments, or Research Centres). Spaces are linked to a specific community if they belong to the same institution. We intend to develop Data Story as a module that will be connected to the User and Space levels (i.e., smaller collaborative units). Certain outputs - once published - will also be displayed at a Community level for broader sharing within the institution. Hanging Data Story off of Research-hub facilitates easy cross-discipline, cross-project and cross-department sharing. This reflects the existing character of many meetings within which data-sharing takes place. It also stands as an example of what the Open Data agenda might be seen to be about, but in miniature.

6.5.2 Data Story Design in a nutshell

Data Story provides a preliminary structure or template to help researchers organize and describe the context of a specific study by making use of written narratives (stories). The interface is organized into chapters, so that shared data can be sorted into sections, aiding navigation through the story. The sequential organization of the chapters creates a timeline of actions, events, and decisions regarding the study being shared.



Project Language Tandem

Story Abstract:

Time Frame: 30/04/2018 to 30/09/2020

Authors: Jessie Hoffmann jessie.hoffmann@gmail.de Jane West jane.wester@

1 Chapter Project Set-up

Language Learning Exchange Digital AR Covid19

2 Chapter Data Processing

Interview Thematic Analysis Workshop Brainstorming Internet

3 Chapter Findings

Augmented Reality Internet Rereation Prototype

Metadata of the Story:

Name of the Story:	Tandem Language learning
Subject:	AR in language learning
Time Frame:	30/04/2018 to 30/09/2020
DOI:	https://doi.org/10.1109/5.771079
Authors:	Jessie Hoffmann jessie.hoffmann@gmail.de Jane Wester jane.wester@gmail.de
Research Institute:	University of Siegen
Country:	Germany
Language:	English
Tags:	No tags included
Contact info:	Jessie.hoffman@gmail.de
Access Rights:	Everyone
Creation Date:	19/05/2018
Publish Date:	05/03/2021

Figure 3: Example of a Data Story structure: chapters' overview and related metadata.

Each chapter might contain multiple documents and 'data snippets' that help to clarify the overall story. Questions and tips are highlighted in the interface of each chapter to support researchers in crafting their own narratives, to encourage reflexive thinking and elicit discussions. Story authors can add few selected metadata to the ‘data snippets’ to enrich the explanation (and support future retrieval), but in a way that allows for learning about and questioning the role of metadata as well. Authors can also introduce themselves and their research institution and give their contact information, etc. This is needed to connect a Data Story with a specific researcher or research team (so as to be publicly acknowledged and possibly contacted).

To exemplify the possibilities, we provide a possible structure for three different chapters within a Data Story (see Figure 3): (1) project set-up; (2) data processing (with snippets of anonymized data); (3) main findings. Each chapter provides a focused insight into the study conducted and suggests a narrative structure threaded through the chapters. There is some general information regarding the story that is provided in the overview screen. This can give information about the time frame and the project to which it belongs (a single publication, a complete research project, a PhD dissertation, etc.). Across the chapters, authors are encouraged to enrich the Data Story with various kinds of contextual information that echoes many of the practices we observed during our own fieldwork, where not just pre-selected data was made available to prospective meeting attendees, but also information about it, such as when and where it was collected, the informants, etc. This can serve to overcome a number of the issues we identified, such as the fact that data snippets are not currently provided in ways that make them searchable, easy to export to continue analytic work, or open to broader sharing or re-use. It also oversteps many of the current concerns being expressed about metadata and documentation by providing a natural way for this to be embedded in the preparation of data for sharing that respects its potential heterogeneity.

The project set-up chapter. The project set-up chapter introduces the overall story outline, thus providing an understandable context for the study. Information related to the study's domain, topic, research questions, methods, author contact information, motivation and aims can all be included. Tips and questions are highlighted in the interface. In this chapter, researchers who write the story are encouraged to consider the following questions (and include their answers in their story narrative):

- How does your story start and where is it situated?
- What is the topic of the story?
- What is/are the research question(s)?

- Why are you sharing the story? What is the goal?
- For whom would this story/study be interesting?
- Who would be interested in your data?

The data processing chapter. The ‘data processing’ chapter encapsulates the actual ‘data snippets’. It also provides a more detailed contextual narrative that explains important milestones in the data collection and analysis process. As with the project set-up chapter, the processing narrative is aimed at resolving common queries to support analytical reflections of the shared data nuggets. This feature of Data Story took inspiration from, and is actively designed to reflect existing practices.

We saw how data was pre-selected from larger datasets to illustrate putatively important themes, in a way that could support collaborative interpretation, discussion, and the development of further ideas. It also avoids a need for researchers to proceed in radically different ways because they are already investing effort in the pre-selection of the most relevant data to support their interactions with other researchers. One of the key advantages to proceeding in this way is that it may serve to eliminate the current tendency for data snippets to be spread across disparate platforms and file-sharing systems without any consistency of structure.

In this chapter, sub-sections can be created to categorize and group data, based on the data type and methodology, thereby easing navigation. Authors are advised to create and fill the sub-sections with relevant data in a way that supports the storyline and its sequence, with sub-sections being ordered sequentially (see Figure 4). Authors can position and relocate sub-sections by simply dragging them to their desired location on the storyline. Example of data types in this sub-section are: informed consent; interview guidelines; observations; interview data; focus groups results; workshop protocols; evaluation outcomes; etc. Customized sub-sections can be created where desired data categories are missing.

Data Story supports the sharing of different data formats. Some snippets might be extracted from a text file and have a text format, e.g. interview questions, transcripts, notes, etc. Other data snippets might take the shape of audio or video files, presentations, posters, pictures, sketches and design materials, etc. All of this can be seen to reflect a need to support existing heterogeneous data collection practices with a variety of data formats. As before, the authors are provided with a list of questions to help them structure the story, support the ‘sense-making’ of the shared data, and enrich the contextual layer. For example:

- What methods were used to collect and analyze the material?
- When was the data collected (timeframe)?
- What data types have you considered during the analysis?
- Keep in mind: what/why/with whom are you sharing?

Only selected and anonymized data will be displayed. There are three principal reasons for this, each clearly articulated in the empirical insights presented above: 1) to protect study participants and avoid the disclosure of any private and sensitive information; 2) to decrease 'data overload' by encouraging researchers to display only the most relevant data; 3) time constraints - it is not possible to provide a complete and carefully crafted narrative in a relatively short period of time that will adequately contextualize all of the data collected during a study. It also in no way breaches existing practices where the original data is stored on personal hard drives or in the cloud and is only accessible to the original data gatherers.

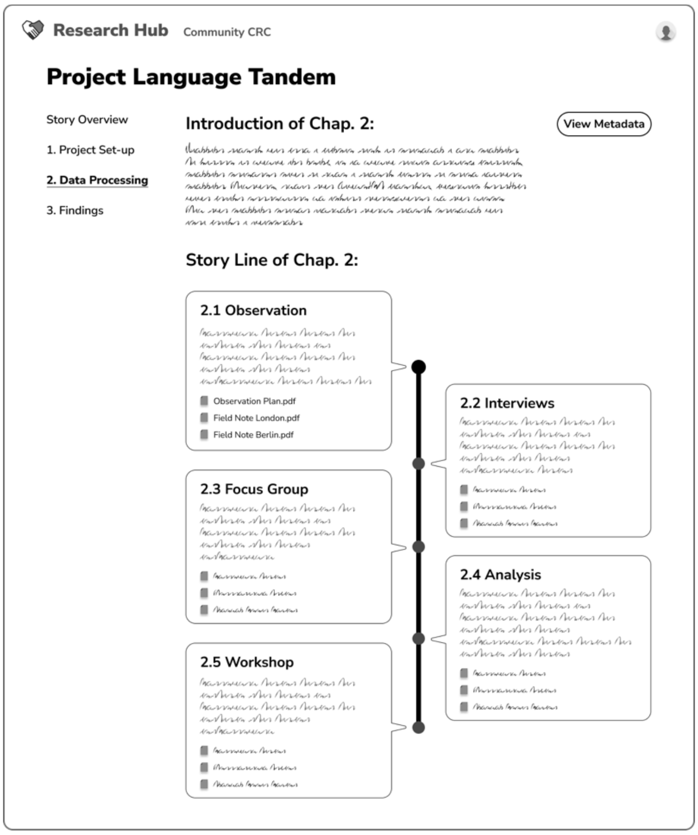


Figure 4: Data processing chapter and its structure once completed. Each sub-chapter contains relevant data and contextual information that helps to make sense of the data itself and of the methods applied.

The findings chapter. Last but not least is the findings chapter, where the narrative ends and future plans are explained. Published materials and citation and review data can be included in this chapter. Again, guiding questions and tips are visible upfront to help researchers structure the information and narrative, e.g.:

- What findings came out of your data?
- To whom/in which fields is this data story specifically useful?
- Bring the story to an end.

6.5.3 Integrating metadata standards

In general, Data Story gives the option to annotate, tag and add metadata to every chapter. Keywords and relevant tags can be assigned to both the story in general and to individual chapters. This provides a quick overview of the general context and topic of the story. Data Story suggests few basic metadata (i.e.: the Dublin Core or DDI) as a standard source for elements. They can, however, be edited quickly and/or a new folksonomy can be created to explain the data. As mentioned earlier, Data Story invests effort in bringing the data and its metadata together by integrating many of the important metadata fields in its interface. This makes metadata an important pillar of the narrative and a driver of discussions. It promotes 'data literacy' and 'awareness' by providing an opportunity for researchers to learn about and reflect on the role of metadata and finally adapt it to their needs. We also envision that the metadata elements will change depending on the data type: i.e., some will be suggested for interview snippets, but different ones for ethnographic notes, focus groups, design sketches, etc. Further research is needed to identify just which elements best match different data types.

6.5.4 Supporting processual workflows: plugin solution

Data Stories can be posted with key data and story milestones at any time throughout the course of a study. In fact, Data Story aims to promote ongoing curation activities as a feature of everyday workflows. To achieve this, Data Story will be connected to routinely-used tools for collecting, analyzing and processing data. We therefore envision a plugin solution. The plugin can be connected to text-editing software like Microsoft Word, data analysis tools like MaxQDA, literature management tools like Citavi, cloud storage tools like Sciebo[33], etc. The idea is to provide researchers with an opportunity to feed their Data Stories with new input at all times by creating direct connections between Research-hub and their own data stores. In

[33] Info on Sciebo: https://hochschulcloud.nrw/en/index.html

this way, researchers can select key segments (text, files, etc.) while organizing and analyzing their data and send them to a Data Story as '*data snippets'*. They will also be able to add annotations, descriptions, comments, and metadata that clarify the context of the chosen data. The transferred data snippets can be previewed and further annotated in Research-hub. They will be located in the data processing chapter unless directed otherwise.

In this ways, Data Stories can be assembled piece by piece as a natural extension of researcher's own data processing workflows, instead of trying to organize the data at the end of a study. This has a better fit with existing approaches where it is important to allow for interpretation, reflection and documentation to continually be enriched throughout the research process.

6.5.5 Publishing: DOI and accessibility rights

Once researchers have completed a Data Story and feel secure with the provided data and narrative, they will be able to publish it. Once published, the story will be visualized in the respective community. A DOI (Digital Object Identifier) can also be automatically assigned to the Data Story. A Data Story's DOI can also be promoted in papers, so that potential collaborators or interested parties can see additional data. Individual sharable links will also be automatically generated for single data entries so that researchers can give others direct access to a specific data snippet. It is up to authors to decide the amount of data to include in a Data Story. Some authors may decide to share very small snippets, others more substantial chunks. Critically, they can also decide what to share with whom. They can share certain parts with some recipients and other parts with some other audience, all within the same Data Story. This is facilitated by having different accessibility rights provided in the Data Story for each data snippet added to the storyline.

Data Story will be accessible within the existing Research-hub platform via a web browser. By integrating it into this platform, it will be possible to engage in discussion (if desired) with people interested in the data. All in all, Data Story hopes to trigger collaborative discussion, negotiation, awareness, sense-making and reflexivity around shared data by whichever parties have an interest, thus neatly tying it into the Open Science agenda.

6.6 Discussion

Above, we have described an approach, inspired by storytelling insights and designed to support a collaborative workflow for the curation and sharing of data which can be used in conjunction with more standard approaches and data descriptions. We have shown how we might, in this way, address some of the more problematic aspects of Research Data

Management and how to meet the expectations of the Open Science agenda. We discuss below some of the key points that can be seen to arise out of adopting this approach. First, we consider how Data Story can serve to articulate a RDM-related collaborative workflow. We then look at the specific advantages of promoting the notion of a 'story' as an organizing principle. We conclude by reflecting upon how Data Story, while not necessarily reducing the overhead of data management, can play an important role in naturalizing the RDM process.

6.6.1 Data Curation as collaborative workflow

In our fieldwork, we have noted how the curation and sharing practices implicit in the Open Science agenda are currently not yet visible or are only being performed in very haphazard ways. At the same time, however, we have seen that there is actually a willingness to share data amongst qualitative researchers we engaged with and, in particular, an interest in how to undertake collaborative work around their data. Data Story actively evolved out of our observations as a potential solution to manage evident RDM issues, to do that in ways that resonate with existing concerns, interests and practices, but, at the same time, to bring qualitative data curation closer to what is increasingly being demanded by funding bodies. While Data Story is not intended to offer a comprehensive RDM solution, it does view RDM and data curation as a process to be embedded in daily practices within a collaborative workflow. In fact, current tools are not yet interconnected in a workflow and miss to offer the opportunity to engage with curation elements such as adding metadata and annotations 'on the go'. Data Story integrates those missing elements but also provides a 'mechanism for narration' with an interactive interface that helps researchers to contextualize and organize their own data with written narratives, a practice more aligned with their way of doing research and dealing with data. Composing Data Stories will still be time consuming for researchers and curation practices will remain an overhead, however, we hope researchers will have the opportunity to gain personal benefits such as organizing heterogeneous data but also structuring relevant findings and analytical reflections to be used in collaborative discussions, publications and future work.

As (Birnholtz and Bietz 2003) have already suggested, to get more effective data sharing systems, designers and IT developers need to go beyond current metadata models and take into account social interactions around data abstractions. Data Story takes this recommendation seriously and stresses the collaborative and social dimensions inherent in data and data practices. At this stage, we cannot yet anticipate how the narrative work will actually play out in practices, which difficulties researchers will encounter in engaging with its workflow and with whom Data Stories will actually be shared. However, we believe that a system like this would be a first attempt to bring the invisible data work to the forefront, to embrace the difficulty of making curation activities 'fit' the realities of local practices and the contingent nature of sharing practices. In this sense, Data Story promotes data awareness and reflexivity, and involves making curation activities and their concerns, technicalities, and specificities visible, while articulating workflows and processes that encourage social interaction and collaboration around data. An important part of how this will be accomplished is the embedding of Data Story within an existing platform, 'Research-hub', that is already purposed for communication, collaboration and sharing across a diverse set of research groups and disciplinary interests. Data curation and sharing practices are, in a sense, a new concern for some academic groups. As such, the concern has yet to be consolidated and supported by collaborative tools. As Mosconi at al. (Mosconi et al. 2019) have already noted "the institutionalization of data curation practices […] requires a better understanding of the use of data in practice but also the development of reliable infrastructure and tools built in a way to help negotiate OS objectives, stimulate self-reflective and learning processes and support discipline-specific data practices" (Mosconi et al. 2019). Data Story is an attempt to address all these issues within the CSCW research tradition and its core themes (Blomberg and Karasti 2013).

6.6.2 The story as organizing principle

The practice of storytelling invites data handlers to think about their data in a way that encourages data reflexivity. Reflexivity has a special status in HSS disciplines, where there is a particular focus upon the relationship between researchers and their data. Through the organization of data snippets as data stories, researchers are specifically invited to reflect upon: 1) what they are sharing (i.e.: what are data, what are metadata etc., what methods were applied, etc.); 2) who they are sharing with; 3) why they are sharing in the first place; and 4) how the data's recipients might understand it. As our fieldwork highlighted, reflexivity is not typically prompted by standard approaches to data curation due the technical and generic language in which metadata standards are normally expressed. Therefore, with Data Story we wish to support data sharing practices while at the same time encouraging greater reflexivity during the process of curation and sharing.

We envision how Data Stories can be seen as a potential solution to the challenges outlined in section 4.2 by accompanying the self-archiving process. Writing a Data Story could be seen as a first step prior to depositing the data into an official archive. The storytelling approach is not intended to replace data curation activities or data curators. Instead, Data Story can enhance the curation process by opening up the 'black box' of research and providing the cultural and contextual circumstances in which data are generated, enabling this to happen during the research process, before formally archiving the data. It can fit organically between the moment of data production and its formal archiving, acting as an interface that can meet the practical issues researchers are currently confronting. As indicated in the empirical insights (section 4.1), researchers already tell stories about their data in their everyday work and undertake some of the activities encompassed within Data Story. Data Story takes those existing practices and concretizes their formulation and pursuit within a visible narrative. It also facilitates the sharing of those everyday stories about data across larger cohorts of researchers. In research about situated storytelling practices (Sacks 1992), much is made of how storytelling provides a mechanism for the sharing of experience. There is a sense in which Data Story builds upon that sentiment by recognizing the power stories have to promote sharing and engagement. The same research also emphasizes how storytelling practices require mundane competences that most people engage in willingly. It doesn't prevent data management and curation from being work, but it does make that work more familiar and routine.

Longer term, our approach will be suitable for data reuse. The central question in this respect is, how does Data Story provide a narrative that not only contextualizes the production of data but also renders it relevant for those who might use it. There is no simple answer to this question, for the value of data in reuse depends as much on the reasons for reuse as it does on the reasons for its production. Nevertheless, Data Story can do a number of useful things as its chapters' structure affords certain data relevancies: the project set-up can tell re-users why the data exists in the first place, its potential value in relation to existing knowledge, and information about the disciplinary origins of researchers; the data processing and the 'snippets' can answer some of the queries re-users may have about the methods adopted, the amount of data and its format and give examples of the data, etc.; the findings can provide a link between snippets and results, enable judgements about accuracy, reliability and validity to be made, reveal literature deemed to be relevant, point to reviews of the work, suggest options for future progress. Finally, the overall narrative positions both data creators and potential re-users as active agents in the construction of meaningful data. Stories invite both data creators and the data re-users to reflect on what messages can be found in the data, what questions can be evoked and answered, and what uses the data can be put to.

6.7 Conclusion and future work

To conclude, organizing, communicating and understanding data are crucial issues in a 'datafied society' (Van Es and Schäfer 2017). Yet, in our digital world it is not always clear what counts as data, how best to make sense of it, and what is at stake when it is put to use (Kitchin 2021).

Data Story aims to foster exchange around data storytelling that can serve as medium to explore data sense making, support data awareness and reflexivity. Although we have focused here upon qualitative data, the concept is, in principle, agnostic as to what is deemed to count as data. Instead, it is able to embrace a plurality of data practices and approaches. Data Story drew upon insights from CSCW, Critical Data Studies and related disciplines. These emphasize the sociality of work practices and the co-construction of meaning. Through Data Story, we want to promote more inclusive data practices that embrace a broader audience and provide diverse and faceted entry points for personal explorations. Our wish is to promote a smooth transition toward open science principles while remaining "as open as possible and as closed as necessary" (EC - European Commission 2016). At present, Data Story remains a conceptual contribution - it has not yet been deployed or evaluated. However, we plan to develop the work further by implementing and evaluating it across a range of projects. In this way, we hope to refine the concept and to gain deeper insights into how it might best support researchers and the recipients of their data. Clearly, we cannot wholly predict what the outcomes of this process might be, though we can speculate. For instance, innovations of this kind might form the basis of new publication formats in the longer term and help to incentivize the work of data curation, which is currently largely seen as unrewarding.

As Rob Kitchin has pointed out, the cooking of data does not take place in a vacuum. Data-driven endeavors are socio-technical in nature. They are as much a result of human values, desires and social relations as they are of scientific principles and technologies (Kitchin 2021). Such a view, we would argue, is fundamental to the CSCW tradition and we would encourage researchers in the field to use the Data Story concept we have presented here as a starting point for examining how alternative approaches to data sharing and reuse can be developed. How data is socially constructed, and how the stories researchers naturally tell about their data feed into its subsequent sharing, reuse and appropriation, should be a fertile field for CSCW research and may have much to offer in turn about effective data design.

7

Designing a Data Story: an innovative approach for the selective care of qualitative and ethnographic data

This chapter was published as book chapter in the edited volume “Praxeology of Data”. Mosconi, Gaia, Helena Karasti, Dave Randall, and Volkmar Pipek. "Designing a Data Story: An Innovative Approach for the Selective Care of Qualitative and Ethnographic Data." *Media in Action*| *Volume 3* (2022): 207.

7.1 Introduction

In this chapter, we present an explorative design concept for the sharing and reuse of qualitative-ethnographic data, that we call Data Story, which is inspired by data storytelling principles. Recent critics of data science have pointed to the need for a contextual approach to data, one which reflects the view that, “data doesn’t speak for itself, it needs a storyteller” (Duarte 2019, 5). However, approaches to data storytelling have hitherto mainly been contingent on the deployment and use of quantitative and statistical data. Our contribution suggests that considerable benefit might result from the use of new tools and methodologies inspired by data storytelling principles for qualitative data as well. We believe this approach has the potential to advance the Open Science agenda at large, which remains some way from realization, especially so for Humanities and Social Sciences (HSS) and for those researchers applying qualitative and ethnographic methods (Mosconi et al. 2019).

Policies that demand or encourage the release of data are predicated on the assumption that others will find the data useful and that data will thus be reused (Christine L. Borgman 2012), but there is evidence indicating that secondary use of data is not yet an established practice (Christine L. Borgman 2012; Bishop 2012; 2014; Mannheimer et al. 2018; Corti 2013). In our view, to make qualitative research data reusable means that, in addition to formats, (metadata) standards and licenses, we must pay attention to the practices of creating, structuring, analyzing and interpreting data (Mosconi et al. 2019; Feger et al. 2020). In order to foreground this largely invisible work as a form of data care, we developed the concept of a Data Story and argue, along with (Maria Puig de la Bellacasa 2010), that care is a useful conceptual anchor for this work specifically because it concerns itself with the “politics of knowledge”. Caring is conceived of as entailing concern for the three dimensions of “labor/work, affect/affections, ethics/politics”. Moreover, caring is interpreted as an act of doing and as a relational act of thinking-with data (Bellacasa 2012). Our concept aligns with this insight, and in fact the Data

Story supports collaborative mechanisms for narration around data snippets that are situated at the center of its design. With it we propose the idea of data curation as an act of selective care that is foregrounded in the interface design.

The purpose of creating a Data Story is to provide a solution for the curation and sharing of data as it is expected by major funding agencies and institutions. In fact, this demand is seldom met in practice, and there aren't any tools available yet that clearly support this additional work of caring for the reusability of data (Mosconi et al. 2019). Therefore, with the Data Story concept, we wish to fill this gap. With our design, we aim to support researchers who do empirical work in organizing the data they care about and make explicit the context. In doing so, we hope to make easier the curation and sharing of qualitative and ethnographic data on the one hand, and the potential reuse by other researchers on the other hand. Software implementations of the Data Story concept will provide researchers with guides and templates supporting them to build stories around the most relevant data they have collected while at the same time envisioning a potential audience. We speculate on how this concept could potentially become a recognized publication format to be promoted in different collaborative data infrastructures or digital databases. In this way, researchers will have the opportunity to get recognition for this unrewarded and invisible work.

Our research concerns itself with the question: How can we best describe qualitative-ethnographic research data practices while respecting epistemological, methodological and ethical challenges, in order to facilitate data sharing? Data Story, as an exploratory conceptual design solution, is an attempt to give an answer to this question. With it we wish to contribute to the international debate around Open Science, and encourage further engagement in such matters by scholars from various disciplines interested in the issues of openness and data care. This chapter brings together various streams of literature on *Critical Data Studies* (Dalton and Thatcher 2014; Dalton et al. 2016; Kitchin 2021), *data curation and sharing of qualitative-ethnographic work* (Bishop 2012; 2014; Corti 2013; Tsai et al. 2016; Treloar and Harboe-Ree 2008; Irwin 2013)and finally *data storytelling* (Duarte 2019; Cole N. Knaflic 2015; Ojo and Heravi 2017). Against the background of the outlined literature, we conducted empirical work and gained practical experiences within a research infrastructure project (INF) in which we engaged in formal and informal conversations with researchers working with qualitative-ethnographic data. Finally, we outline the exploratory design concept, Data Story, and discuss the act of selective care it affords.

7.2 Related work

7.2.1 Data as matter of care

As Dourish and Gómez have pointed out: “Data makes sense only to the extent that we have frames for making sense of it, and the difference between a productive data analysis and a random-number generator is a narrative account of the meaningfulness of their outputs” (2018, 8). The arrival of big data has been a motivating force for what is termed Critical Data Studies (Dalton and Thatcher 2014; Iliadis and Russo 2016). As Kitchin and Lauriault (2014) point out, critical data studies are largely concerned with questions about the nature of data, how they are being produced, organized, analyzed and employed, and how best to make sense of them and the work they do, occasioned by a step change in the production and employment of data. The principal force of a critical approach, then, lies in the recognition that political, social, ethical, organizational, and economic elements shape data management as much as technical problems in much the way Bellacasa (2011) suggests in her critique of technoscience. As Bowker (2005) suggested:

> "We need to open a discourse – where there is no effective discourse now – about the varying temporalities, spatialities and materialities that we might represent in our databases, with a view to designing for maximum flexibility and allowing as much as possible for an emergent polyphony and polychrony. Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care" (Bowker 2005, 184).

Thomer and Wickett (2020) further demonstrate the point through an analysis of the various material forms that the database can take, arguing that “‘best practices’ for data management are in tension with the realities and priorities of scientific data production”, and “understanding pluralism in data practices is crucial to supporting the needs of those traditionally marginalized by information technologies—whether in their personal or disciplinary identity” (Thomer and Wickett 2020, 3). Curating for data work as a pluralistic and contextual endeavor has, as yet, not been fully realized.

7.2.2 Challenges for qualitative data sharing

Data sharing and consequently data reuse have been extensively addressed (Heaton 2008; van den Berg 2008; Faniel and Jacobsen 2010). The vast part of the literature, however, deals with practices embedded in the natural and applied sciences. Our matter of care, however, is the additional complexity entailed in the management of qualitative data, where most of the challenges can be characterized as epistemological, methodological and ethical in nature. For qualitative data, paying attention to the context of their collection and possible re-use becomes an overarching concern. However, what context is, and how to describe it, is non-trivial (Moore 2006). Context determines whether something can be viewed as data or metadata and the “degree to which those contexts and meanings can be represented influences its transferability” (Borgman et al. 2018). Data loses meaning when removed from the original contexts, packaged in repositories, and disengaged from the knowledge and expertise of the researchers who performed the study (Walters 2009). When dealing with qualitative data we need to recognize the essentially reflexive character of data and that it is often rich with personal content (Tsai et al. 2016). Ethnographic approaches are generally based on a relationship of trust between researchers and participants, often in sensitive domains. This leads us to a consideration of the ethical challenges, where protecting the privacy of participants typically is one of the central aims (for more details see contribution by Kraus and Eberhard in this volume, and Eberhard and Kraus 2018).

Other challenges related to describing and preparing these types of data for sharing are: the lack of clear standards (Tsai et al. 2016; Antes et al. 2018) which are difficult to identify due to the heterogeneous nature and idiosyncrasy of researchers’ data practices; not knowing how one might access and use the data in the future and for which purposes (Broom, Cheshire, and Emmison 2009); and finally time-constraints where “the burden of organizing qualitative data for inspection or reuse could easily exceed the work of writing the manuscript itself” (Tsai et al. 2016, 5). As we shall see below, data storytelling provides us with inspiration as to how to best design for the curation and sharing of these types of data while addressing some of these complex issues.

7.2.3 Data Storytelling: guiding principles

The social sciences and humanities have long stressed the role that narrative plays in human life, in education and in research. As Game and Metcalfe argue:

“Research is always an interpretative process that involves conversations and storytelling, though the research framework traditionally applies other names such as aims, methods and conclusions. Research conventions are a particular form of storytelling that allows sociologists and historians ‘to tell stories as if they weren’t’ storytellers’” (1996, 65).

Social scientists tell stories for a range of different purposes. In doing so, they attempt to contextualize the 'data' that they work with. They do so largely for analytic purposes. In relation to this, and to return to the question of what context is and how to describe it, there is a difference between context as an analytic construct – something that researchers, curators, etc. define – and something that emerges in and is enacted by the work of the participants. Put simply, context in this view has no existence outside of the way in which it is ongoingly constructed by participants to an activity. Data, in other words, is a process of enactment. Digital storytelling, we want to argue, is a useful means to reconstruct what has previously been constructed or enacted.

Digital storytelling describes the practice of everyday professionals and organizations who make use of digital tools in order to tell a story. Digital stories can stimulate emotional responses in recipients and potentially offer interactive elements. Storytelling approaches have been applied to several fields: therapy, education, arts and culture, management and business, among others (Barrett 2006; Vecchi et al. 2016; Yuksel, Robin, and McNeil 2011; Restrepo and Davis 2003; Denning 2006). In the last decade, however, due to the advent of big data and the "data revolution" (Kitchin 2014) western economies and governments are becoming progressively more data-driven, and therefore we have seen growing contributions and approaches focusing specifically on *Data Storytelling* (Duarte 2019, Knaflic 2015; Ojo and Heravi, 2018). The main argument being made is that to understand and use data effectively, data needs to communicate a clear message (a narrative) and speak a human language to allow us to make sense of data (data sense making) and the reasons why it is presented (reconstructed) the way it is.

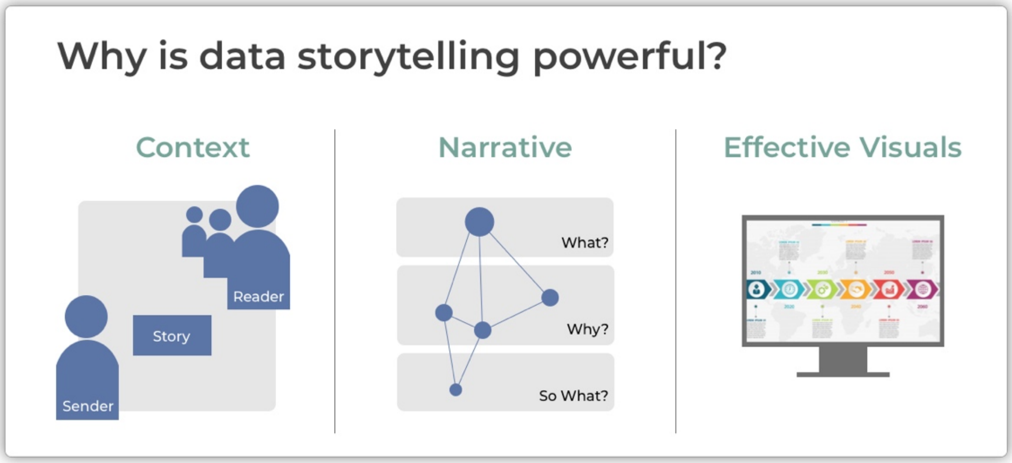


Figure 1: Main principles of Data Storytelling. Source from: https://www.nugit.co/what-is-data-storytelling/. Accessed in April 2021.

As shown in the picture, three main principles summarize what data storytelling is about and how to achieve it: 1) explaining the context; 2) identifying a coherent narrative; 3) working on effective visualization. In data storytelling, the second principle, *narration,* is a crucial element. A narrative can, additionally, have emotional elements. A story has a beginning and an end, it has a goal, sometimes a moral, and, obviously, a story has an audience. Narrative helps to "share norms and values, develop trust and commitment, share tacit knowledge, facilitate unlearning, and generate emotional connections" (Sole and Wilson 2002). The third principle is related to *effective visuals*. As Lee et al. (2015) suggest, relatively little attention has been paid in the visualization literature to the ways in which the stories in question are actually crafted.

To conclude, the concept of a Data Story for qualitative research data, as proposed here, combines all three affordances of data storytelling identified in the literature: a) it offers researchers an opportunity to provide contextual information to their collected data, b) it employs a narrative structure to demonstrate its analytical potential, c) and it allows for the integration of visual elements.

7.3 Background and approach

Our research takes place in a research infrastructure project (INF), connected to the Collaborative Research Centre (CRC) "Media of Cooperation" funded by the DFG (in English: German Research Foundation) since January 2016 and it's currently ongoing. Our CRC is characterized by interdisciplinary cooperation across disciplines and faculties, and most of researchers apply qualitative and ethnographic methods. Being tasked with providing suitable solutions for both ongoing research and long-term preservation as well as the sharing of materials with a wider public, the focus of our project is on developing new RDM practices and infrastructures for qualitative-interpretative research contexts. Collaboration with the IT service provider of the University – partner of the project – has been going on since the beginning of the funding period and this entailed interdisciplinary work with developers where we worked on metadata structures, restructured database hierarchies and classification schemes. Drawing on insights from CSCW and socio-informatics (Wulf et al. 2018), our project roots conceptual design and technology development itself in qualitative and long-term situated research. Therefore, we engaged in participatory observations, semi-structured interviews and informal conversations with CRC's projects, where we particularly investigated data practices, salient Research Data Management and data sharing issues that could inform our design.

The fieldwork we conducted as part of our infrastructural research was not straightforward and unproblematic. Some researchers felt annoyed and irritated by the work of our project. Its objectives were often met with indifference, questioned or overtly criticized on multiple occasions. In particular, *metadata critiques* emerged repeatedly during fieldwork. Researchers we talked to struggled to understand the meaning and the applicability of metadata standards such as the Dublin Core[34] which was often mentioned by the IT service provider as the existing metadata standard that researchers should use in describing data for long-term preservation (and potentially for data reuse). However, in practice, qualitative researchers in particular lack familiarity with such standards and struggle to understand, or fail to see the point of, its technical language.

The agenda of the funding agency and the institutional top-down narratives around Research Data Management were not always matched by the immediate and practical objectives of research teams. Nonetheless, our approach was dialogic. Through interviews, observations and informal conversations we oriented reflexively to the often conflicting viewpoints expressed. We questioned design solutions, discussed current or new practices and the connection between the two in relation to design possibilities. As Schön (1983) pointed out, “design, in practice, is not a linear process.” This pragmatic-reflexive approach led us to consider the need to embrace narrative as a focus for our deliberations in relation to data. The idea developed into what we call Data Story here which came about gradually after reflecting over a long period of time with local research groups. Their own narratives regarding data sharing and related challenges inspired the approach we describe. This led us to envision a system in which the showcasing of data snippets (or data nuggets) could potentially support the organization, curation and eventually sharing and reuse of research data, and therefore allow to meet the expectations of the funding body.

In the next section, we explain the major insights which led to the Data Story concept. We do so by grounding the concept in researchers’ practices where storytelling emerges as an integral part of (collaborative) analytical work with qualitative data and therefore synergetic with these types of research approaches.

7.3.1 Grounding the concept in practices

The conceptualization of a Data Story gradually emerged during fieldwork, especially in our interaction via observations and interviews with researchers. Over three years, we paid

[34] https://www.dublincore.org/specifications/dublin-core/dces/

particular attention to situations in which (informal) data sharing practices took place, and we observed how qualitative-ethnographic data was analyzed, collaboratively discussed and represented with the support of (digital) media.

We began to notice, for instance, the common practice in qualitative research of sharing *data snippets* in collaborative analysis sessions with members of the same project (but with different disciplinary backgrounds) and/or with researchers from other projects. In these situations, snippets of anonymized data are often selected, enriched with context and sent to participants via email a few days before the analysis session. A *narration* or, if you will, a *story* which contextualizes the data is often provided by the data collector in written form (i.e. as text), and/or in oral form at the very beginning of the session. The piece of data in question then is often projected in the room in order to guide the conversation and to promote interpretative work. Through this collaborative practice, as Dourish and Cruz (2018) expressed it, data is "put to work in particular contexts, sunk into narratives that give them shape and meaning, and mobilized as part of broader processes of interpretation and meaning-making" (Dourish and Cruz 2018, 1). Data are not collected and analyzed in a vacuum, but are always shaped, co-created, (partially) shared and narrated based on the specific circumstances in which data are needed and "put to work". Another example is Rose, who said: *"in our team we couldn't really do very close readings of the data together, due to lack of time and the overload of data we collected, so we just selected a few data and sketches that we could talk about in order to collaboratively develop our thinking."* Her team developed 'ad hoc' visualization techniques around data snippets, as we might call it, in order to elicit a collaborative narrative and which partly inspired our conceptual design.

Another researcher, Sophie, told us that direct access to data (even if partial) could foster interdisciplinary collaboration and new research approaches: "*sometimes you see a paper, but you do not realize all the kinds of data and fieldwork that has been done, and if you look at the data then it makes you think of other collaboration that you could have with this person."* In fact*,* Sophie had collaborated with a social scientist in the past, but only after looking at some examples of ethnographic data was she capable of understanding what kinds of collaboration might be possible and what research questions could be answered. But she also added that *"there aren't really good solutions to represent and share ethnographic data just yet"* and *"we had to share the data via email which obviously wasn't ideal!"* Another important element connected to data sharing and reuse is the messiness of ethnographic work. The majority of researchers we talked to expressed discomfort in sharing their qualitative data due to the "messiness" which often comes with it. We noticed their need to have better tools and techniques that could support the organization of the heterogenous data and the non-linear way of conducting research typical of ethnographic work. The Data Story started to emerge then as a form of digital data storyboarding to support collection, organization, collaboration, and data sense-making.

The above vignettes point to the way in which a *storytelling approach* to data curation can be called into action, one which is more aligned with researchers' practices, and as possible inspiration to organize the heterogeneous data and to support collaborative data sense-making. In the following, we demonstrate how the Data Story is envisaged to work by showing the design sketches of the low-fidelity prototype we have developed so far. We will then discuss more extensively the idea of selective care that it affords.

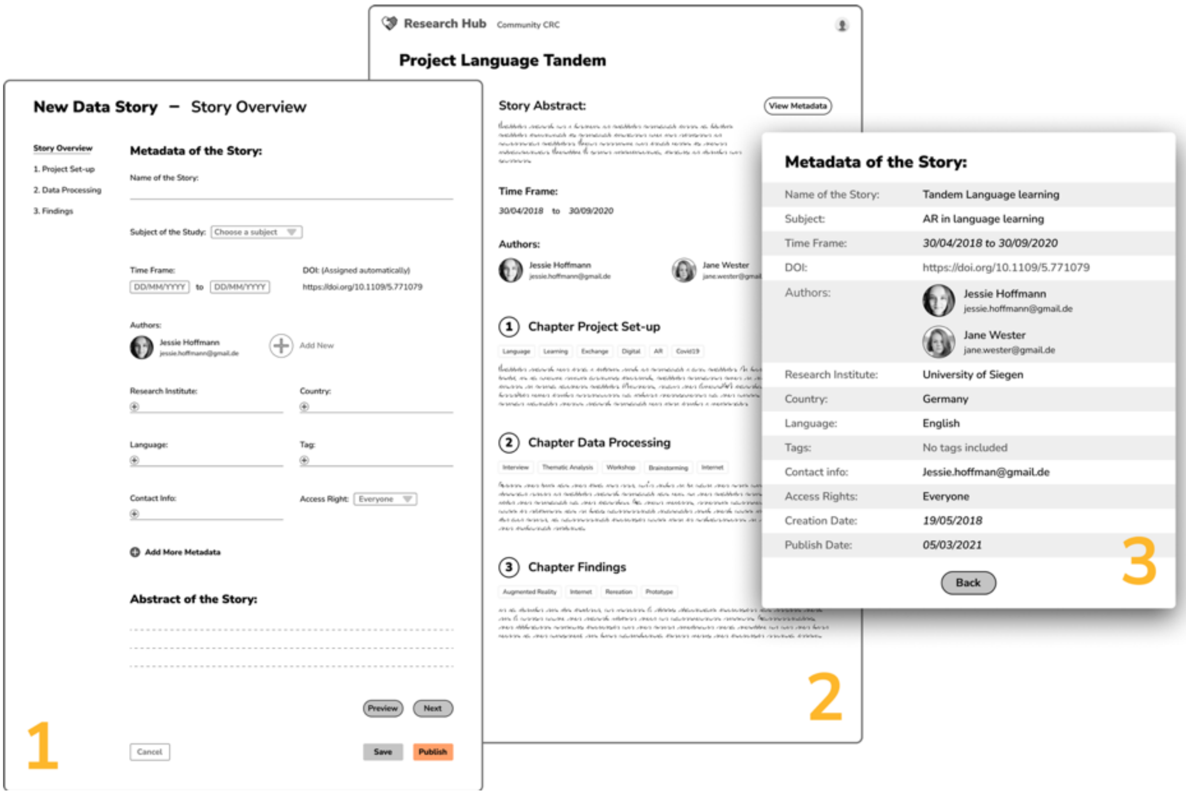
7.4 The Data Story process and its components

This design concept is meant to be an organizing device in support of (collaborative) storytelling practices as a major component of data analysis and sense making. By engaging with its process and its interactive interface researchers will have the opportunity to perform data curation practices resulting in selected data snippets. In this way, we wish to make easier the sharing of these types of data on the one hand, and the potential reuse by external researchers on the other hand.

The interface is organized into chapters to sort the shared data into sections and better help in navigating through the story. The chapters sequence creates a timeline of the actions, events, and decisions regarding the study being shared. Each chapter might have multiple data snippets that help clarify the overall story. Questions and tips are highlighted in the interface of each chapter to support reflexivity, elicit discussions and help researchers to construct their narrative. To exemplify the possibilities, we provide a possible structure with an initial overview screen (0) followed by three main chapters for the story: (1) project set-up; (2) data processing (with snippets of anonymized data), and (3) main findings. As mentioned before, each chapter provides a focused insight into the study conducted but also it invites to make explicit the context and to define a coherent narrative.

(0) Overview screen

In the overview screen, general information regarding the study will be given, like the time frame and to which project it belongs (a single publication, a complete research project, a PhD dissertation, and so on). Moreover, the authors can introduce themselves, their research institution, their contact information, etc. This is needed to connect a Data Story with a specific researcher or research team (in order to be publicly acknowledged, and possibly contacted).



New Data Story – Story Overview

Metadata of the Story:

Name of the Story:

Subject of the Study: Choose a subject

Time Frame: DD/MM/YYYY to DD/MM/YYYY

DOI: (Assigned automatically) https://doi.org/10.1109/5.771079

Authors: Jessie Hoffmann jessie.hoffmann@gmail.de

Access Right: Everyone

Abstract of the Story:

Research Hub Community CRC

Project Language Tandem

Story Abstract:

Time Frame: 30/04/2018 to 30/09/2020

Authors: Jessie Hoffmann jessie.hoffmann@gmail.de; Jane Wester jane.wester@gmail

1 Chapter Project Set-up

2 Chapter Data Processing

3 Chapter Findings

Metadata of the Story:

Name of the Story:	Tandem Language learning
Subject:	AR in language learning
Time Frame:	30/04/2018 to 30/09/2020
DOI:	https://doi.org/10.1109/5.771079
Authors:	Jessie Hoffmann jessie.hoffmann@gmail.de Jane Wester jane.wester@gmail.de
Research Institute:	University of Siegen
Country:	Germany
Language:	English
Tags:	No tags included
Contact info:	Jessie.hoffman@gmail.de
Access Rights:	Everyone
Creation Date:	19/05/2018
Publish Date:	05/03/2021

Figure 2: Data Story module overview: Figure 2.1 is the view of the author, 2.2 is the view of the reader, and 2.3 is an overview of some of the included metadata

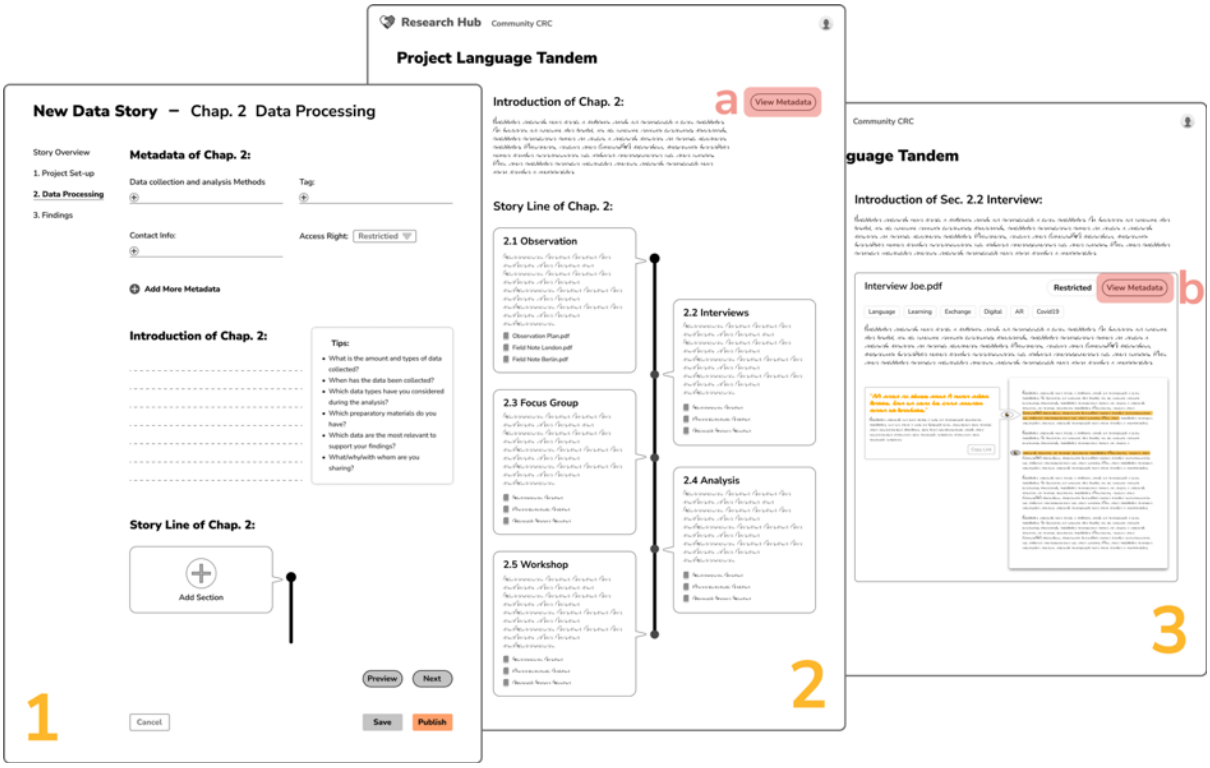
(1) The project set-up chapter

The project set-up chapter introduces the overall story outline, in order to provide an understandable context for the study. Information related to the research field, topic, and research questions of the study, as well as methods used, a short summary about the motivation and aim of the study can be included. Tips and questions are highlighted in the interface in order to elicit reflexive thinking while support data sense-making.

(2) The data processing chapter

The data processing chapter encapsulates the actual *data snippets*. It also provides a more detailed contextual narrative that explains important milestones in the data collection and the analysis process. As with the project set-up chapter, the process narrative is aimed at resolving common queries to support the sense making of the shared data nuggets.

The chapter provides the possibility to create sub-sections which categorize and group data, based on the data type, to ease navigation through it. It is advised to create and fill the sub-sections with relevant data in a way that supports the storyline and sequence of the story. Moreover, this chapter creates a storyline by ordering the created sub-sections sequentially. Authors of the stories will have the ability to relocate the created sub-section if necessary by dragging it to the desired location on the storyline



Tips:

- What is the amount and types of data collected?
- When has the data been collected?
- Which data types have you considered during the analysis?
- Which preparatory materials do you have?
- Which data are the most relevant to support your findings?
- What/why/with whom are you sharing?

Figure 3: Data processing chapter: Figure 3.1 shows the view of the story writer, 3.2 shows the story from the reader's view after publishing, 3.3 shows the interview sub-section.

The Data Story supports the sharing of different data formats. Some snippets might be extracted from a text file and thus have a text format, e.g. interview questions, transcripts, notes, etc. Other data snippets might take the shape of audio or video files, presentations, posters, pictures, sketches and design material, etc. As in the chapter before, the author will be provided with a list of questions that might add a better structure to the story and support the sense making of the shared data as well as enrich the contextual layer.

As already mentioned, only selected and anonymized data will be displayed. This is for three reasons: (1) facilitate the protection of the study participants and avoid the disclosure of any private and sensitive information; (2) decrease data overload by encouraging researchers to display only the most relevant pieces of data; (3) time constraints: as it is not possible to provide a deliberate narrative, in a relatively short time, that is rich of context to all the collected data of the study.

(3) The Findings Chapter

Last but not least is the Findings chapter, where the narrative is brought to an end and future visions can be explained. Any publications or material, citation and review data can be included in this chapter. Again, guiding questions and tips for contextualizing the chapter will be visible upfront and will help researchers in structuring the information and narrative.

7.4.1 Supporting processual workflows: plugin solution

The Data Story aims to promote curation activities to be carried out as soon as possible, as close as possible to the data source, and in support of workflow. It is a proposal for embeddedness. In order to achieve this, the Data Story will be connected to tools used routinely while collecting, analyzing and processing data. Therefore, a plugin solution is envisioned. The plugin is to be connected to text editing software like Microsoft Word, data analysis tools like MaxQDA, literature management tools like Zotero, cloud storage tools like Sciebo[35] or other tools that researchers routinely use. As mentioned earlier, the idea is to provide the researchers the opportunity to feed their Data Story with data at all times by creating such direct connections between a collaborative research infrastructure already in use and the researcher's data storage. In other words, researchers can select key data pieces (text, file, etc.) while organizing and analyzing their data, and send them to the Data story as *data snippets*. Moreover, researchers will be given the chance to add annotations, descriptions, comments, and metadata that clarify the context of the chosen data. The transferred data snippets can be previewed and further annotated via the interface.

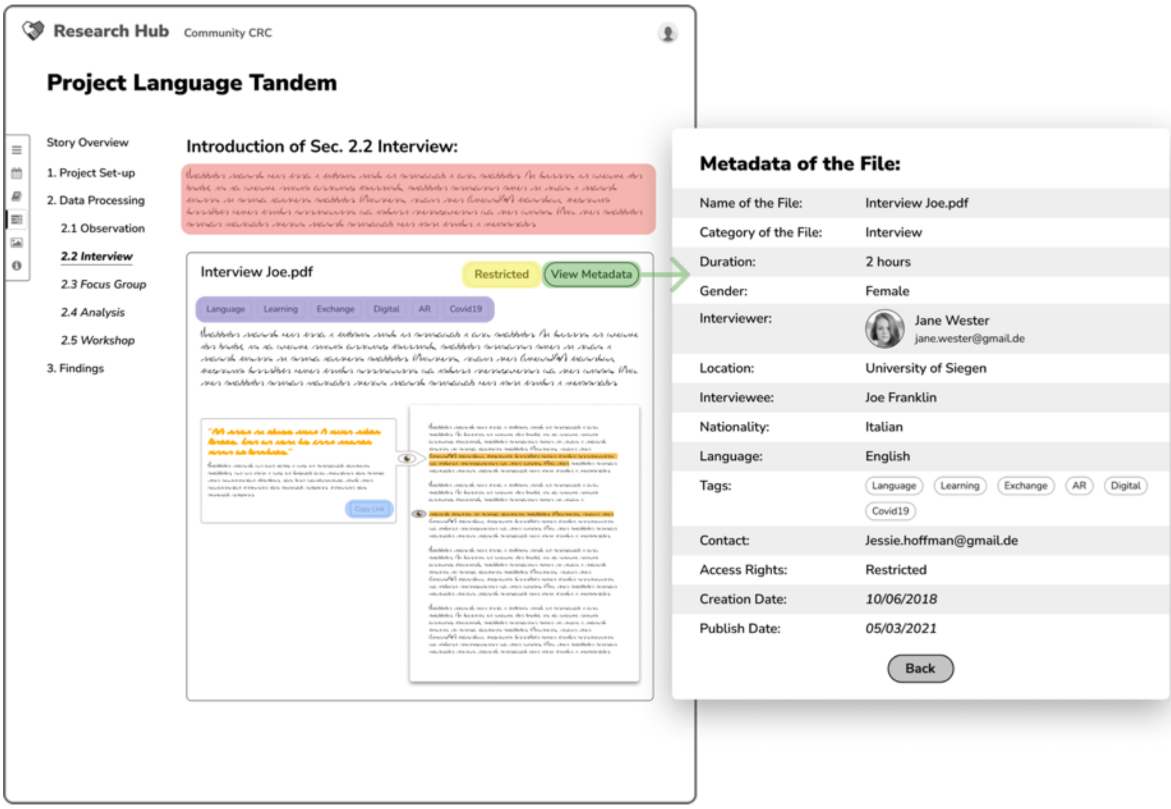
7.4.2 Publishing: DOI and accessibility rights

Once researchers have completed their Data Story, and feel secure with the provided data and narrative, they will be able to publish it. A DOI (Digital Object Identifier) can also be (automatically) assigned to the Data Story [see Figure 4, blue highlight]. We envision a new practice that could emerge from this: the DOI link of the Data Story web-interface might be promoted in papers where potential collaborators or interested parties could see additional data.

[35] Info on Sciebo: https://hochschulcloud.nrw/en/index.html

Moreover, share links will be (automatically) generated for single data entries to indicate a clear reference to a specific data snippet.

Researchers can share parts of the data with some recipients and some other parts with some other audience using the same Data Story. This is facilitated by different accessibility rights provided in the Data Story for each data snippet added in the storyline. Taking inspiration from Jones et al. (2018) we considered the following accessibility rights: open, restricted, controlled, and closed (these categories can be assigned to the whole Data Story, or to specific data snippets). The accessibility right *Open* means that data is available to be accessed by anyone; *Restricted* means to be accessed by some specific audience; *Controlled*, means that the author has to grant permission to access it after assessing the request. Lastly, *Closed* means "data deposit and citation exist for archival purposes but no data are currently available (could be embargoed until publication of results, change in sensitive situation, death of a participant, or certain duration of time from collection)" (Jones et al. 2018, 21). Figure 4 highlights how accessibility rights will be shown in the design (highlighted in yellow).



Metadata of the File:

Name of the File:	Interview Joe.pdf
Category of the File:	Interview
Duration:	2 hours
Gender:	Female
Interviewer:	Jane Wester jane.wester@gmail.de
Location:	University of Siegen
Interviewee:	Joe Franklin
Nationality:	Italian
Language:	English
Tags:	Language Learning Exchange AR Digital Covid19
Contact:	Jessie.hoffman@gmail.de
Access Rights:	Restricted
Creation Date:	10/06/2018
Publish Date:	05/03/2021

Figure 4: Visual of metadata, tags, DOI, data snippet and the story [Purple: tags, Red: Story. Orange: Data Snippets, Blue: DOI, Green: Metadata, Yellow: access rights]

In our view, the Data Story should be promoted as a new publication format that is centered around relevant data points. Data Stories can act as intermediate format between a larger dataset to be stored and secured in long-term archives and the official publications (paper, books etc.). Data Story could offer insights into the content of a dataset but also offering some reflections on the data that might not be part of the final publications. By promoting a Data Story as a new publication format that can be cited, researchers will have the incentives to actually engage with this type of work and get rewarded for this additional effort. Being a Data Story an additional step is important so that researchers will get compensated for this work. By envisioning an accessible open link, Data Stories can circulate freely through the web, and can act as entry points for engagement with the data that have been collected.

We are planning to implement this design in a collaborative research infrastructure, called Research-hub, that is already in use in our research center. However, we believe this design with its modular and customizable characteristics has the potential to be integrated as interface layer of any other (collaborative) data infrastructure or digital database.

7.5 Discussion

7.5.1 Data Story as an act of selective care

Above, we have described an approach, inspired by storytelling insights and designed to support a workflow for the organization, curation and sharing of data which can be used in conjunction with more standard approaches and data descriptions (i.e.: metadata). The purpose of creating the Data Story is to provide all those with an interest in the possible uses of data with an easy way to access and understand how a data collection was assembled and the reasons for it. This, we do by supporting researchers who collected the data in the first place to envision a possible audience and to make the context of their work explicit, using both metadata and a narrative. So, this design concept is meant to be an organizing device in support of (collaborative) storytelling practices as a major component of data analysis and sense making. As we have seen, however, complex issues intervene. They include the nature of the work, ethical concerns and the reflexive nature of the engagement with data, all of which have methodological and epistemological consequences.

We take on board the injunction of Van Es and Schäfer (2017) that, "[r]ather than import questions and methods from the hard sciences, we must develop our own approaches and sensitivities in working with data that will reflect the humanities' traditions" (2017, pg. 16). The authors here include a call for action, inviting humanities scholars to develop their own research questions and methods to stay consistent with their epistemological positions. We have shown how we might translate these ideas to the field of Research Data Management and curation. If solutions to data sharing and curation need to be found, as expected and demanded by funding agencies, then we argue, those technical solutions, tools or infrastructures will need to embrace and embed in the design cultural values, methodological practices and epistemological understandings of the communities they are designed for. In doing so, we again connect to the concept of care as pushed forward by Bellacasa (2011): "… representing matters of fact and sociotechnical assemblages as matters of care is to intervene in the articulation of ethically and politically demanding issues. The point is not only to expose or reveal invisible labors of care, but also to generate care" (Bellacasa 2011, 94). We discuss below two lines of argument in which we explicit how the Data Story reveal the invisible labor of data care while at the same time generate care for both, the data producer and the data re-user.

7.5.2 Complementing metadata standards with a Story

As we have seen, it is now accepted that context is critical to our understanding of data (Christine L. Borgman 2015; Carlson and Anderson 2007) as a representational mechanism bridging data producers and data re-users. Within the Research Data Management domain this contextual role is typically assigned to metadata standards and data descriptions. Formal and standardized metadata such as the Dublin Core or the Data Documentation Initiative (DDI) assume not only a contextual role but also, it is claimed, are essential for the discovery, comprehension, and reuse of data. Metadata are often interpreted as the "bridges" because they can, in principle, convey the information essential for discovery and secondary analysis: "secondary users must rely on the amount of formal metadata that travels along with the data in order to exploit their full potential." (Ryssevik, DDI). However, and as is evidenced both in our own practical experiences with researchers and in Feger et al. (2020), cleaning the data, and filling metadata requirements is a quite tedious and rather technical practice. The inherent difficulties, along with the fact that researchers do not see this as their primary purpose, means it is frequently poorly done or not done at all. Moreover, analysts of qualitative data often do not have enough time to fully explore their data given the richness and the amount of the data in question (Fielding and Fielding 2000; Yoon 2014). Therefore, the Data Story provides the opportunity to display only selected data snippets and narrate them coherently. This we argue could potentially make it easier for a researcher interested in certain data sets to understand how the data collection and analysis came about. At the same time, the researcher(s) who collected the data is supported in explaining the whole data process, displaying what, for them, is the most important aspect in the data and envisioning a potential re-user.

The Data Story interface makes visible the act of care by articulating the tasks of data care needed in order to organize the data, retrieve them, present them, share them, and possibly reuse them. In fact, it provides every chapter with the option to annotate, tag and add metadata. The Data Story suggests metadata (i.e.: the Dublin Core or DDI) as the standards source for elements set. They can, however, be adapted quickly and added as new folksonomy. In this way metadata are treated as "living things" that can grow and develop based on a bottom-up understanding. As mentioned earlier, the Data Story invests noticeable effort in bringing the data and its metadata together by integrating many of the important metadata fields in its interface in a way that makes metadata an important pillar of the story narrative and driver of discussions. It promotes data literacy and awareness, as it is an opportunity for researchers to learn about the role of metadata but also put it into question and adapt it to their needs.

With our contribution, we complement the role that formal data descriptions (metadata) bring to the table when they are provided, and suggest an alternative when they are not, depending on the institutional investment in data curation. By focusing on narrative as an organizational layer and as a useful method to make explicit the context, we aim to make the interpretative work – essential to make use of data – less onerous for both parties: data producer(s) and data re-user(s). Stories, then, can serve a further purpose, that of inviting re-users to reflect on what messages can be found in the data, what questions can be evoked and answered, and what uses the data can be put to. The Data Story is then a complementary organizing layer – flexible, culturally, collaborative and context sensitive – that can be added to the formal and structured way of organizing and preserving data. Finally, by promoting the Data Story as a possible intermediate publication format, we allow researchers to get rewarded for this additional step and we show care for their additional curation work.

7.5.3 Designing for situated data

That knowledge is situated is hardly a discovery by now and, indeed, has been a central tenet of the sociology of knowledge at least since Mannheim (1936). It can be traced through the work of, for instance, Vygotsky (1980), Garfinkel (1967) and many others, but has been reinvigorated in practice-oriented thinking (see Randall et al. 2018) and in feminist standpoint theory (Haraway 1991; D'Ignazio and Klein 2020). Critical Data Studies (Dalton and Thatcher, 2014; Dalton, Taylor, and Thatcher 2016; Kitchin 2021) draws on these insights to address "the situated, partial, and constitutive character of knowledge production" (Drucker 2011, 2), in order to show how the meaning of data is derived from its context of production and use. This is particularly true for qualitative data because qualitative research is characterized as an "insider activity" (Mauthner, Parry, and Backett-Milburn 1998), its knowledge "is highly contextual and experience dependent" (Niu and Hedstrom 2008), its practice uses "the self … as the primary instrument of knowing" (Ortner 2006), and it involves interpretation and subjectivities not concrete (or transportable) enough for information to be documented and reused in its entirety (Broom, Cheshire, and Emmison 2009).

Kitchin (2021) suggests that, for *all* datasets, "we tell stories *about* data, and stories *with* data, in which there are inherent politics at play in how they are discursively figured" (Kitchin 2021, 5). D'Ignazio and Klein in their book "Data Feminism" (2020) also pose interesting questions such as, "How can we use data to remake the world? […] or, more precisely, whose information needs to become data before it can be considered as fact and acted upon?" (D'Ignazio and Klein 2020, 36). Embracing the partiality and situatedness of data means designing with these questions in mind, to question what is data, what is metadata, how do we construct facts and information, how are they disseminated, how they get curated and shared. In this way, the Data Story concept engages in "politics of knowledge" (Bellacasa 2011). Our design helps to address the questions raised above and tries to give some answers applied to the context of curation and data sharing. With our design, we wish to support pluralism in research data (management) practices, embrace situated knowledge, without excluding data collection efforts which might not fit neatly into current standards and categories.

Concerning the issue of reuse, the question is how does the Data Story provide a narrative which can not only contextualize the production of the data but also render it relevant for the re-user. Of course, there is not, and cannot be, any simple answer to such a question, for the value of data in reuse will depend as much on the reasons for reuse as it does on the reasons for its production. Nevertheless, the Data Story can do a number of useful things (bearing in mind that it is a complement to, and not a replacement for, established metadata schemes). Firstly, and most obviously, it renders certain features of the data more visible which otherwise would not be (at least immediately) the case. The proposed three-chapters structure affords a number of data relevancies and highlight specific data points. Thus, the project set-up might tell the re-user why the data exists in the first place, what value it is believed to add to existing knowledge, information about the disciplinary origins of researchers (and possibly the backgrounds of participants). The data processing section affords snippets which go some way to answering the queries that re-users may have about methods adopted, the amount of data and its formats, examples of the data in question, and so on. The findings section provides a link from the snippets to results, enabling judgements about accuracy, reliability and validity to be made, literature deemed to be relevant to the researchers, reviews of the work, and so on. Overall, it offers the possibility of comparison with the aims that re-users might have, the options they may have with regard to methods and forms of analysis, insights into the kinds of questions and answers embedded in the data, insights into the number and type of people they may wish to engage with, and even suggest options for future progress.

7.6 Conclusion

Organizing, communicating and understanding data are crucial issues of our modern "datafied society" (Van Es and Schäfer 2017). Yet, in our digital world it is not always clear what data are, how best to make sense of them, and what is at stake (Kitchin 2021, 1). With our design concept of the Data Story, we aim at fostering exchange around data storytelling which should not be limited to quantitative data, data visualizations, infographics, statistics and standard approaches, but should embrace a plurality of data practices and approaches.

Bellacasa (2012) argued: "We cannot possibly care for everything, not everything can count in a world, not everything is relevant in a world…" (Bellacasa 2012, 204). For this reason, the Data Story aims at showcasing only anonymized data snippets (such as interview excerpts, pictures, videos, sketches or any other relevant material) that researchers are encouraged to select based on the relevance for their own research findings and for an envisioned audience i.e. what they care about. This act of selective care is organized along a timeline and enhanced with storytelling practices (in oral and written form). STS scholars have already demonstrated how formal data descriptions wrapped in informal descriptions might increase the usefulness of the data (Bowker and Star 1999). The Data Story concept embraces this insight. In fact, it integrates traditional metadata standards but also allows the creation of bottom-up folksonomies. Metadata elements, folksonomy and data snippets are then visualized and glued together, enriched and situated with the addition of a storyline. In this way, Data Story brings the invisible work of data care to the forefront, it promotes data awareness and reflexivity, and calls for making visible (and supporting) curation activities, its concerns, technicalities, and specificities while articulating workflows and processes for collaborative activities. In all, the notion of care and more specifically how selective caring (or caring about caring) provide a conceptual anchor for a range of issues that have hitherto been only addressed in very limited ways. The Data Story, we suggest, is an explorable avenue for more sophisticated approaches to data management and reuse.

Fostering Research Data Management in Collaborative Research Contexts: Lessons learnt from an 'Embedded' Evaluation of 'Data Story'

This chapter was published in JCSCW: Mosconi, Gaia, Aparecido Fabiano Pinatti de Carvalho, Hussain Abid Syed, Dave Randall, Helena Karasti, Volkmar Pipek. "Fostering Research Data Management in Collaborative Research Contexts: Lessons learnt from an 'Embedded' Evaluation of 'Data Story'." *Computer Supported Cooperative Work (CSCW), 1-39.*

Abstract. Recent studies suggest that RDM practices are not yet properly integrated into daily research workflows, nor supported by any tools researchers typically use. To help close this gap, we have elaborated a design concept called 'Data Story' drawing on ideas from (digital) data storytelling and aiming at facilitating the appropriation of RDM practices, in particular data curation, sharing and reuse. Our focus was on researchers working mainly with qualitative data in their daily workflows. Data Story integrates traditional data curation approaches with a more narrative, contextual, and collaborative organizational layer that can be thought of as a 'story'. Our findings come from a long-term 'embedded' evaluation of the concept and show that: (1) engaging with Data Story has many potential benefits, as for example peer learning opportunities, better data overview, and organization of analytical insights; (2) Data Story can effectively address data curation issues such as standardization and unconformity; and (3) it addresses a broader set of issues and concerns that are less dealt with in the current state of play such as lack of motivation and stylistic choices. Our contribution, based on lessons learnt, is to provide a new design approach for RDM and for new collaborative research data practices, one grounded in narrative structures, capable of negotiating between top-down policies and bottom-up practices, and which supports 'reflective' learning opportunities – with and about data – of many kinds.

8.1. Introduction

Problems related to collaborative practices are frequently related to 'infrastructural' work that may well benefit some practitioners, or a community as a whole, but not the practitioners who need to do the work (Grudin 1988). In those cases, these 'beneficial' rules and procedures are often well-known and acknowledged, but their appropriation into actual practices often proves difficult. This challenge also applies to research contexts, where, in principle, the Open Science (OS) agenda can provide a beneficial framework for successful collaborations. In fact, the OS mandate – strongly supported by funding and research agencies who aim to facilitate research verifiability, 'good' scientific practices, and data reuse – is simultaneously changing the dynamics of research (Wallis et al., 2013) and promoting massive infrastructural investments. The mandate implies, (or explicitly states) that future research funding will depend on data sharing. Therefore, governments and research institutions worldwide are imposing from above a specific rhetoric of 'good' RDM practices which often implies the use of institutional infrastructures, standards, and guidelines (EU, 2020). The top-down policy-driven adoption of OS initiatives is often constrained due to funding agencies' insistence on a generic view of research data practices, and a strong emphasis on data storage and recovery as the primary issue. In fact, the OS movement has been conceptualised, within the FAIR (Findability, Accessibility, Interoperability, and Re-use) data principles, as entailing guidelines to improve Research Data Management (RDM) which has been realised in an ever-increasing proliferation of data hubs and repositories acting as storage and recovery media in research (Borgman et al., 2019; Wilkinson et al., 2016).

However, more recently, concerns for how data is to be understood across disciplinary boundaries, and how re-use is to be facilitated, have come to the fore (Feger et al. 2020), implying that discipline and methodological-specific norms and data practices need to be investigated and understood (Borgman 2012, 2015; Mayernik 2016; Pasquetto et al., 2016; Tenopir et al., 2011; Velden 2013). For example, in Humanities and Social Sciences (HSS), and more specifically for those researchers applying qualitative and ethnographic methods, collaborative and data-intensive research endeavours, the plurality of research methods, standards and traditions, ethical and legal implications, and heterogeneous practices in storing, processing, sharing and analysing data indicate higher barriers to the implementation of OS initiatives (Eberhard and Kraus, 2018; Korn et al., 2017; Mosconi et al., 2019).

To close this gap, since 2016, we have explored socio-technical contexts in which qualitative-ethnographic data are produced, curated, and eventually shared. Our initial insights allowed us to delineate the gaps that still exists between the OS and related RDM top-down agenda and the bottom-up practices of researchers affected by it (Mosconi et al., 2019). Indeed, not all data are created equally and for some disciplines it is much harder to adjust to the new expectations due to the nature of the data collected and the methods applied. This issue calls for the development a new approach for RDM specifically in support of qualitative and ethnographic data but that could potentially serve other disciplines struggling with the OS mandate and RDM expectations.

RDM, in itself, is a complex and long-term endeavour spanning the entire research lifecycle and beyond, requiring attention to the specifics of data creation, curation, storage, sharing and reusability (Treloar and Harboe-Ree, 2008; Whyte and Tedds, 2011). They are different practices but at the same time intertwined. 'Good' RDM asserts the notion of reusability through openness, sharing and collaboration throughout the whole research process

(Reichmann et al., 2021) but the implications for RDM when confronted with disparate data practices applied by different disciplines, methodologies, and research communities are still not fully understood. Another layer of complexity in RDM is added by the overhead (additional work, time, and costs) implied in the appropriation of data curation and the sharing practices which require researchers to engage in systematic organization of data (i.e., metadata creation, contextualization and structuring the storage of data) in on-going research projects and in anticipation of reuse.

To tackle some of these complex problems, new tools, and research data infrastructures for RDM are emerging (Borgman et al., 2019; Kaltenbrunner 2017; Khan et al., 2021; Lee et al., 2009). In our view, these solutions typically address the guidelines of findability and accessibility, but they do not necessarily solve the issue upstream of how to curate and manage data effectively during the research process. It is clear that tools for the meaningful appropriation of RDM as a long-term processual phenomenon are as yet lacking. Here, we argue, data storytelling approaches can come in handy.

Over the past few years, data storytelling – i.e., the use of narrative and visual elements to effectively communicate data insights (Dykes 2015) – has been emerging as "a promising approach for supporting more accessible and appealing human-data-interactions" (Concannon et al. 2020, p. 2). However, as we will argue in section 2, very little work (Riche et al., 2018; Showkat and Baumer 2021) has been done to support researchers working in an interdisciplinary context to use data storytelling insights to curate, share and potentially reuse data – a notable exception is the work of Showkat and Baumer (2021), who have addressed the relationship between journalism and data scientist work practices, by investigating the exploration process in investigative data journalism. Our current contribution specifically addresses this gap and seeks to provide conceptual and socio-technical answers to some of the issues above.

Since 2016 we have explored the challenges that qualitative and ethnographic researchers encountered when confronted with OS and RDM mandates for the first time (Mosconi et al. 2019). These investigations have been carried out within an information management (INF) project, connected to a Collaborative Research Centre (CRC), and funded by the German Research Foundation (German acronym: DFG from the original German Deutsche Forschungsgemeinschaft), where the DFG expects INF to provide support and develop RDM solutions for the qualitative and ethnographic-oriented research projects (representing the majority in our CRC).

Driven by these institutional constraints and drawing on empirical findings, we developed a conceptual solution for RDM called ‘Data Story’ (Mosconi et al., 2022) which offers a means of enhancing and naturalizing curation practices through storytelling. The name itself *Data Story* is not new. We credit the term to Nancy Duarte (2019) who has been applying data storytelling principles to support decision-making processes within the business sector. The novelty here, however, lies in the application of data storytelling insights to the field of RDM and in the use of the ‘Story’ as a metaphor and design principle used to implement a socio-technical system in support of data curation and sharing practices not yet established and that in the long-term might lead to a re-use of research data.

Our work is, therefore, driven by the following wider question: How can a Data Story approach support with the establishment and appropriation of RDM practices of researchers – mainly working with qualitative and ethnographic data – in collaborative research contexts? And more specifically:

SQR1. How can we best support researchers in curating, sharing and potentially re-using data through a Data Story?

SQR2. What features should a Data Story have in order to allow for the appropriation of new practices – data curation, sharing, and re-use – not yet established for qualitative and ethnographic research contexts?

Our previous publication (Mosconi et al. 2022) presented in detail the Data Story concept and the first low-fidelity prototype and showed how its design was grounded in researchers' practices and wishes concerning new tools for RDM. It speculates on the benefits of applying data storytelling principles to the field of RDM mainly by drawing on a literature review without including any direct feedback from the researchers. On the other hand, this paper reports on the Data Story design as it was iterated, based on users' evaluation gathered through formal and informal interactions. We define our engagement and evaluation as ‘embedded’ – (see e.g., Barry et al., 2018; Lewis and Russell, 2011 on embedded research) meaning that researchers and research participants are ongoingly immersed in the research context in which the technology is to be used. In fact, since September 2016 the first author has been an affiliated member of the CRC. In this way, ‘Data Story’ became both the topic and the medium through which we were able to understand how RDM practices can be introduced into researchers' daily workflows, how they are adjusted to elaborated on by researchers – therefore appropriated – and how collaborative research contexts can profit from them. Our contribution, based on lessons learnt, is to provide a new design approach for RDM and for new collaborative research data practices, one grounded in narrative structures, capable of negotiating between top-down policies and bottom-up practices, and which supports ‘reflective’ learning opportunities - with and about data – of many kinds.

8.2 Related Work

Adding some form of narrative to data forms and structures has been advocated and implemented in a variety of contexts. This can be seen, for instance, in both the literature on ‘digital’ and ‘data’ storytelling.

Previous research has investigated the use of storytelling in non-profit organisations (Erete et al., 2016) and in educational contexts – e.g., (Martinez-Maldonado et al., 2020; Xu et al., 2022). The InfoVis community, as it is sometimes termed, has invested considerable effort in providing tools for generating effective visualisations to aid narrative - see e.g., (Fekete 2004; Fekete et al., 2008; Liu et al., 2014; Méndez et al., 2017; Pantazos and Lauesen, 2012). Recently, some attention has been paid to the differences in meaning that users in different contexts might experience (Lallé and Conati, 2019). This latter issue is of central importance to our own work. Elsewhere, ‘digital storytelling’, as it is sometimes called, has explored the use of visuals in different domains, as for example, education (Wu and Chen, 2020), health (Moreau et al. 2018; West et al. 2022), and business (Duarte 2019; Cole Nussbaumer Knaflic 2015). However, to the best of our knowledge, no previous work has used such data storytelling insights to develop socio-technical solutions for addressing RDM issues and support the appropriation of related practices, in particular data curation, sharing and reuse.

Below we uncover three major streams of literature relevant to our work: first, we concentrate on work discussing the challenges of RDM, paying special attention to specific issues concerning qualitative and ethnographic research methods – the focus of our research; second, we go on to introduce existing solutions and infrastructures for RDM, especially in regard to data curation and sharing; last but not list, we address contributions concerning recommendations for the design of RDM tools and infrastructure. These three strands speak directly to our wider research question, which concentrates on RDM practices, and the more specific research questions (SQR) addressed in this work, which respectively focus on the concept that we are proposing – i.e., the Data Story – (SRQ1) and the features that such support should include to allow for appropriation of new practices more effectively (SRQ2).

8.2.1 Challenges for RDM: the issues with Data Curation and Sharing for qualitative and ethnographic data

Research Data Management (RDM) is commonly defined as "the organization of data, from its entry to the research cycle through to the dissemination and archiving of valuable results" (Whyte and Tedds, 2011, p.1). RDM is characterised by several core practices, such as data curation, metadata documentation, long-term preservation, and data sharing altogether leading to the publishing and successful reuse of research data.

Ethical issues, privacy concerns, technical limitations, lack of skills, restricted access, and lack of a rewards systems are among the most discussed barriers to effective RDM in all major disciplines and fields (Feger et al., 2020; Tenopir et al., 2011). In fact, curating, preserving, and sharing research data require appreciable overhead and technical skills but the current scientific culture and rewards system do not directly incentivise or yet, recognise these endeavours (Fecher et al., 2017). Moreover, issues in sharing data are intrinsic to the complex and contextual nature of data itself. Data are not 'natural kinds' but are constructs, existing in contexts of production, use and reuse (Christine L. Borgman 2015).

Nonetheless, some disciplines, such as the natural sciences, have managed to adjust better to OS and RDM expectations, and progressively, have developed internal policies to ensure the curation, sharing and eventually reuse of research data (Zuiderwijk and Spiers, 2019; Witt et al., 2009). For other disciplines these requirements are relatively new, and researchers and institutions are still struggling to understand how to meet these new demands.

For Humanities and Social Sciences (HSS), and specifically for those researchers working with qualitative data, the expectations for data curation and sharing pose some additional challenges characterised as epistemological, methodological, and ethical in nature (Feldman and Shaw, 2019; Ryen 2011). For instance, with these data, legal and ethical issues can abound, the personal character of the data can make researchers unwilling to share it in its totality; it can be hard to see what counts as data and/or metadata, and the sheer heterogeneity of RDM practices can make standardisation massively problematic. Therefore, data sharing concepts and infrastructures for quantitative data cannot be translated directly to qualitative data. As Tsai et al. (2016) puts it:

> "… the iterative nature of qualitative data analysis, and the unique importance of interpretation as part of the core contribution of qualitative work, [makes data] verification likely to be impossible" (p. 192).

Other critical factors are the protection of study participants expected by ethics bodies, or self-imposed through researchers' lack of familiarity with ethical data sharing practices. Trust-related issues are also relevant: researchers lack the knowledge on who might have access to their data once shared and what they will do with it, fearing a loss of control over the data and subsequent risk to study participants (Eberhard and Kraus 2018). Another pressing problem is connected to the messiness of qualitative data which are often overwhelming to work with (Jiang et al., 2021). A final issue is that for the most part, only major universities, libraries, and librarians are the service providers for RDM support and training. These institutions are often understaffed and/or unqualified to advise on a huge variety of disciplines and heterogenous research data practices (Hamad et al., 2021; Kervin et al., 2014; Johnston et al., 2018; Pinfield et al., 2014). Therefore, they might fail to satisfy the increasing demand for skills in RDM applied in different research contexts.

It is evident that data curation and sharing still has unresolved and nuanced challenges. In our contribution, we seek to address some of the abovementioned issues by examining a solution that is innovative, flexible, epistemologically nuanced, and which has been designed by closely looking into situated, collaborative research data practices.

8.2.2 Existing Solutions and Infrastructures for RDM, Data Curation and Sharing

Researchers have devoted considerable attention to promoting large-scale, distributed scientific collaboration that can facilitate new scientific discovery. Cyberinfrastructure, eScience and OS initiatives have been at the forefront of these efforts (Jirotka et al., 2013; Mosconi et al., 2019). Initial attempts to support these collaborations had a technology-centric focus, with a particular emphasis on providing advanced computing capabilities such as high-speed processing, data repositories, and specialized analytical tools (Finholt 2002; Neang et al., 2020; Olson et al., 2008). However, developers and researchers alike quickly realized that scientific collaboration presented sociotechnical challenges, with technology, social practices, and social structures all being closely intertwined (Downey et al. 2019; Neang et al. 2020).

For RDM in particular, some major barriers to the appropriation of data curation, sharing and re-use practices can be rooted in the interaction with socio-technical infrastructures or in the lack of suitable ones (Borgman 2010; Edwards et al., 2013; Feger et al., 2020). Most existing solutions are repository-styled research storage facilities: they can be generic, such as

Zenodo[36], Dryad[37] or DataverseNO[38], supporting many types of research data and therefore suitable for a wide variety of scientific fields; or they can be discipline specific and community-driven, e.g., for social science research, examples being QualiService[39], GESIS[40], and SowiDataNet[41] (Linne and Zenk-Möltgen, 2017). Universities' repositories are also being increasingly developed by all major institutions, and they often address multiple disciplines similar to existing generic repositories.

Research repositories, however, largely target two specific aspects of the RDM data life cycle: long-term preservation and sharing. They do not necessarily solve the issue upstream on how to curate and manage date effectively during the research process (Mosconi et al. 2019a). Archiving data in a repository is then seen by researchers as the ultimate step, not always directly connected to daily practices in which data get generated, processed, and analysed, causing the archiving process to be perceived simply as an extra burden, with no direct benefits, especially in the absence of a strong mechanism of rewards (Chawinga and Zinn, 2020; Curdt and Hoffmeister, 2015; Donner 2022).

Moreover, open data portals or data repositories are typically all about the structuring of data and the policies that surround it: how many datasets, how many formats, which open licenses and so on. While these are necessary for the long-term preservation of 'data objects' and their retrieval, there are still few design solutions that specifically support the practices and workflows necessary for interdisciplinary collaboration around data objects (Feger et al. 2020; Mosconi et al., 2019; Tuna et al., 2022). These previous studies shown that lack of suitable infrastructure, knowledge and skills has forced researchers to adopt haphazard, ad hoc, practices that lead to unstructured archives. In response to these challenges, Johnston et al. (2018) elaborated the Data Curation Network (DCN), a curation-as-service model designed to support network partners to foster local curation expertise, aiming at a resilient and distributed expertise network capable of sustaining central services and supporting its expansion. The network has established itself and as of today presents itself as a network of "professional data curators, data management experts, data repository administrators, disciplinary scientists and scholars" representing "academic institutions and non-profit data repositories that steward research data for the future use" (DCN, 2023).

[36] https://zenodo.org/
[37] https://datadryad.org/stash
[38] https://dataverse.no
[39] https://www.qualiservice.org/de/
[40] https://www.gesis.org/en/research/research-data-management
[41] https://www.re3data.org/repository/r3d100011062

A thorough understanding of RDM in practice is clearly indicated (Cragin et al., 2010) especially if, as Feger et al. (2020) suggest, HCI research is to have a role "in supporting the transition to effective digital RDM through a design-focused understanding of the roles and uses of technology". Our prior work on the use of data storytelling in the context of RDM (Mosconi et al., 2022) has demonstrated at a conceptual level the potential role of narrative structures in providing relevance for data curation and sharing. Our current contribution resonates with the *data journeys* approach (see e.g. Leonelli and Tempini, 2020; Bates et al., 2016) which aims at highlighting "the socio-material conditions that frame activities of data production, processing and distribution, and resultantly influence the form and use of data and their movement across infrastructures" (Bates et al., 2016, p.2).

However, only a very limited amount of work has been aimed at innovative digital solutions which address these problems (Feger et al., 2019; Garza et al., 2015; Mackay et al., 2007). One notable example for the Humanities is PECE (worldpece.org), an open-source, Drupal-based platform designed to support a wide range of collaborative humanities projects, which pays a considerable attention to the way data artefacts get collaboratively shared, archived, and potentially reused (Fortun et al., 2021; Poirier 2017). Another example is "Making a Tea" a design elicitation approach used to implement a digital lab notebook with the aim at making available to the general public experimental records from the chemistry field (Dix, 2009). Lastly, worth mentioning is Data Curation Profile (Witt et al., 2009), a tool that has been developed for academic librarians which provides a template of different metadata representing relevant information concerning a variety of data collections to be used in institutional repositories.

8.2.3 Existing recommendations for the design of RDM tools and infrastructures

Recent literature has identified design recommendations for new tools and infrastructure in support of RDM (Feger et al., 2020; Koesten et al., 2019; Witt et al., 2009), and more specifically for data curation and sharing (Birnholtz and Bietz, 2003; Feger et al., 2019; Jahnke and Asher, 2012; Rowhani-Farid et al., 2017; Zimmerman 2007). Because these two practices (data curation and sharing) directly imply the additional work needed to make data understandable for a potential audience, they are often described in relation to reuse.

For instance, Koesten and Simperl (Koesten and Simperl, 2021) argue that in order to better facilitate reuse, the creation of structured textual data documentation (or descriptions such as Readme files) are of importance, as they often constitute the first points of interaction between a user and a dataset. Therefore, their creation should be supported during the act of curation and sharing. As they put it:

> we cannot see datasets as usable end products without telling the story of how they were made. Because the story is complex, the user experience of data relies on tools and environments that try to do exactly that: embedding datasets in the rich context of their creation and use (Koesten and Simperl, 2021, p.99)

Other studies (Birnholtz and Bietz, 2003) underline how research infrastructures also need to improve communication channels around research artefacts because anything that is shared should in principle be of interest for somebody else and data creator and recipient need to be allowed to exchange information. More recent studies (Allen et al., 2019) however found out that that researchers are often unmotivated because there are no incentive structures, or retain a degree of uncertainty about their results (e.g. Van Der Bles et al., 2019). Rowhani-Farid et al. (2017) and Feger et al. (2021), on the other hand, concentrated on tools for sharing and reproducibility and stressed the importance of mechanism of reward, to increase motivation and benefit, which could be promoted through OS badges and gamification elements.

Technical standards, legal frameworks, and guidelines are also crucial and need to be considered while designing new tools and infrastructure but most of the literature in this direction has focused on operational problems such as interoperability and machine readability and not so much on readable metadata for human interpretation. Only a few solutions have been proposed so far to document data context beyond what is typically considered and stored as metadata (Chao 2015; Gebru et al., 2021; Preuss et al., 2018). One example comes from recent information-research scholars (Sköld et al., 2022) who suggested a term called *paradata* which signifies information about the means (procedures, tools, activities) by which a certain body of information came into being.

Feger et al. (2020) suggest investigating how RDM tools could compensate for the lack of formal training in RDM and state that new tools should be developed to remove current barriers and more specifically to integrate RDM practices into the research workflow. In our view, RDM, metadata, and curation work have focused too much on interoperability and machine readability. The issue here for us is how do we produce a meaningful (possibly asynchronous and distant) interaction between users in and through the data they use. In what follows, then, we describe the iterative process by which we designed and evaluated a new technological aid, called 'Data Story', devised to provide for meaningful organisation, curation and sharing of heterogenous data which in the long-term could include all the above suggestions and recommendations made by previous studies.

8.3 Methodology and approach

In this section, we describe the ethnographic, long-term (and ongoing) engagement taking place within the aforementioned INF project. Our involvement, which started back in September 2016, has *inter alia* produced the Data Story design concept. This concept, as introduced above, was meant to support researchers to engage in data storytelling as a way to support the appropriation of RDM practices and in particular with the curation, sharing, and potential reuse of qualitative ethnographic data. In what follows, we introduce our research design and then provide some more contextual information on the data collection and analysis activities of the study.

8.3.1 A Design Case Study

In order to enhance the likelihood of designing a useful and usable concept, which can be integrated in current research data practices and appropriated accordingly, we drew on a practice-centred approach predicated on constant engagement with the user and their contexts (Wulf et al. 2015b). Therefore, the interests and concerns of all parties guided our interaction in the field, and continuously shaped our design and evaluation activities from within.

More specifically, our initiative has been predicated on the Design Case Study (DCS) framework, a research design for design research. The framework is mainly composed of (1) contextual investigations, usually predicated on qualitative research approaches, very often of the ethnographic kind; (2) (participatory) design activities, engaging different stakeholders in decision processes concerning the technology under elaboration, and using different design methods, as for example, sketching and prototyping; (3) appropriation studies, also predicated on qualitative research approaches, focusing on how users adopt, integrate and tweak the design for their own purposes in their practices, and how these practices evolve (Wulf et al. 2015b).

8.3.1.1 Pre-study: Uncovering the Context and Existing RDM Practices in Place

The Collaborative Research Centre (CRC)[42] is composed of 14 projects with over sixty researchers, representing several major disciplines and faculties, and where the majority of them apply qualitative and ethnographic methods. As expected by our funding agency, the DFG, (acronym from the original German: Deutsche Forschungsgemeinschaft) and defined by the project proposal, the goal of the INF project is to develop (and establish) infrastructural solutions and design concepts which should lead to the curation, sharing, and potential reuse of research data in our CRC.

Since September 2016, the first author has joined the CRC as affiliated member working in the INF project. Being an affiliated member means that the first author joined the CRC and regularly participated to seminars, lecture series and events which took place over the years. In November 2016 she began her fieldwork where she started to investigate the difficulties of qualitative data sharing and the practical challenges that the OS agenda is presenting specifically in qualitative-ethnographic driven research contexts (Mosconi et al., 2019). She has also been collaborating with the IT service provider of the University, helping developers to customise several open-source tools (i.e.: *RDMO*: for creating Research Data Management plans; *DSpace:* a long-term repository; and *Humhub,* a platform for team collaboration and sharing). In particular Humhub, which is now named 'Research-hub' (https://research-hub.social/dashboard), was established to customise, test, and study new RDM concepts and workflows. These are expected to be implemented by INF in the long-term. In parallel, she has conducted nineteen semi-structured qualitative interviews (see table 1) and ethnographic observations, run meetings to discuss RDM issues with CRC's projects, and supported them in creating their RDM plans.

Pseudonym	**Pseudonym**	**Background**	**Academic Role**	**Date**
Interviews	Sophie *	Media Science	Principle Investigator	4.4.2017
	Joe	Media Science	PhD	16.4.2017
	Alvin	Sociology	Post-Doc, Project Leader	20.4.2017
	Lucy	Sociology	PhD	4.5.2017
	Mary	Law	PhD	19.05.2017
	Rupert	History	Principle Investigator	31.5.2017

[42] CRCs can be funded for up to twelve years across three separate evaluation stages (Phase I; Phase II and Phase III). Our CRC started in January 2016 and completed its first funding period in December 2019. A second phase began in January 2020 (funded until December 2023). All CRC's projects are interdisciplinary in nature.

	Lukas	Sociology	Post-Doc, Project Leader	31.5.2017
	Mark	Political Science	Project Leader	6.6.2017
	Paul	Sociology	Principle Investigator	7.6.2017
	Carl	Sociology	PhD	14.7.2017
	Rob	Media Science	Principle Investigator	10.7.2017
	Colin *	History	Post-Doc, Project Leader	25.7.2017
	Julian	Anthropology	PhD	12.2.2018
	Aaron *	Business Informatics	PhD	3.3.2018
	Philip	Computer science	Principle investigator	7.5.2018
	Cliff	Business Informatics	Post-Doc	6.7.2018
	Susanne	Social Science	Principle Investigator	15.1.2019
	Beth	Political science	PhD	23.3.2019
	Will	Anthropology	Principal Scientist	5.5.2019
RDM plan meetings	Presence of two members per project: total 26 researchers			October 2019
Total participants involved: 45 Researchers (between 2017 and 2019)				

Table 1: Pre-study participants' overview: type of interaction, background, role and date. All participants have an interdisciplinary background and apply qualitative and ethnographic methods in their research with various degree of expertise. Three participants marked with * were involved in evaluation activities as well (see table 2).

These first interviews, observations and meetings took place between 2017 and 2019 and were useful for investigating the CRC researchers' data life cycle and bottom up RDM practices from the outset. Moreover, interdisciplinary discussions concerning Research Data Management and data practices within CRC's projects took place regularly in the CRC - during seminars and other events - and the first author's involvement in these provided an opportunity for numerous formal and informal conversations with researchers. These conversations highlighted relevant RDM issues that make it difficult to meet the expectations of funding agencies for data sharing and reuse and therefore motivates the development of a new approach. Due to the nature of our interaction is difficult to provide an exact number of participants, however, we estimate that at this stage we involved forty-five participants.

Our initial insights allowed us to discover major gaps that still exist between the top-down OS policies, standards, and infrastructure (see i.e., FAIR data principles promoted by funding agencies and research institutions worldwide) and the bottom-up research data practices observed in the field (Mosconi et al., 2019). It was evident that, while sharing and curation practices are expected by all major funding agencies, these practices are not yet supported by any tool that researchers use daily, nor they are integrated in researchers' workflows. If at all, they are performed informally or in a haphazard way. Consequently, as already highlighted by previous literature (Begley and Ellis, 2012; Collaboration 2012; Fecher et al., 2017), data curation and sharing practices, needed to meet the OS goals, are perceived by many as an unrewarding chore, especially when targeted at preserving and sharing data for other researchers to benefit from. Put another way, their primary work tasks are typically separate from any additional work they might need to perform for others to benefit.

In the context of data curation and sharing, the beneficiaries are, or appeared to be, mainly future (unknown) data re-users. Indeed, much of the scepticism about the funding agency's agenda that we encountered early on in our work was a function of these factors. Others, however, showed an interest in innovative solutions that might help them to represent and share their highly heterogenous research data, initially for their own purposes. They were specifically interested in how to organise different data sources and underpin the work of collaborative interpretation and sense-making, and potentially organize their data for their own future use.

These early investigations led us to envision a digital system called Data Story (Mosconi et al., 2022) to be embedded as a module in the platform, Research-hub, in which researchers could organise portions of pre-selected data to be curated with written narratives, storytelling, tags and metadata elements, ultimately to share them with colleagues and/or with an external audience. We organized the system in three main chapters distributed over a timeline (more details in section 3.2). Ultimately, Data Story integrates traditional data curation approaches, where research data are treated as 'objects' to be curated and preserved according to specific standards, with a more contextual, culturally nuanced, and collaborative organizing layer that can be thought of as a "Story". We anticipated that, in the long-term, the Data Story would help to introduce and support the new RDM practices expected by the DFG, first and foremost curation and sharing and potentially data re-use. In the next section, we highlight the initial design and low-fidelity prototype.

8.3.1.2 Initial design: Data Story design rationale, sketches and low-fidelity prototype

The Data Story concept was inspired by the way researchers were seen to share 'data snippets' and engage with them on an ad hoc basis during meetings, collaborative analyses sessions or paper discussions (for more details, see Mosconi et al., 2019, 2022). In those meetings, portions of selected data are contextualised to others with the support of written or oral narratives and collaboratively interpreted and analysed. Through collaborative research data practices, as Dourish and Cruz (2018) expressed it, data is "put to work in particular contexts, sunk into narratives that give them shape and meaning, and mobilised as part of broader processes of interpretation and meaning-making" (p.1). Therefore, the main rationale behind the concept was to allow the sharing of heterogenous qualitative data accompanied with 1) written narratives or storytelling practices for data contextualization, analysis, and sense-making; and 2) technical element and standards, such as metadata, tags and DOI for data curation and retrieval.

Initial prototype sketches were made between January and February 2021. Figure 1 shows the Data Story as an independent module already integrated and accessed through the Research-hub platform menu (already established in 2019).

We took the story as a design metaphor and organizing principle and as such, we translated this into 'design features' that would reflect a Story-like structure. Therefore, we organised its interface with chapters and a panel that would allow movement across them. The sketches developed further into a low-fidelity prototype designed between February and March 2021.

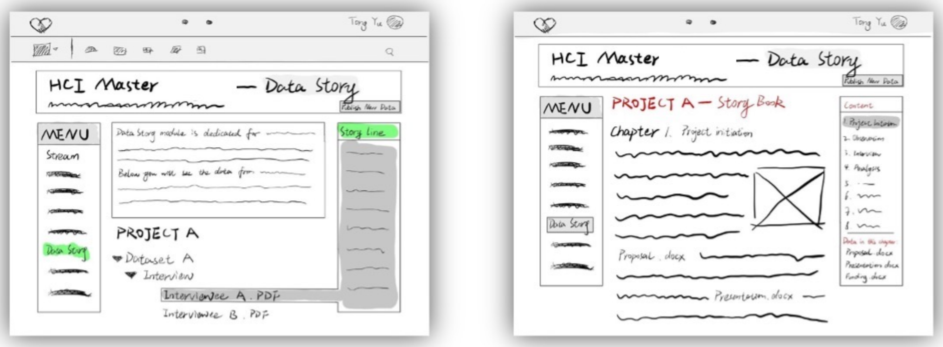
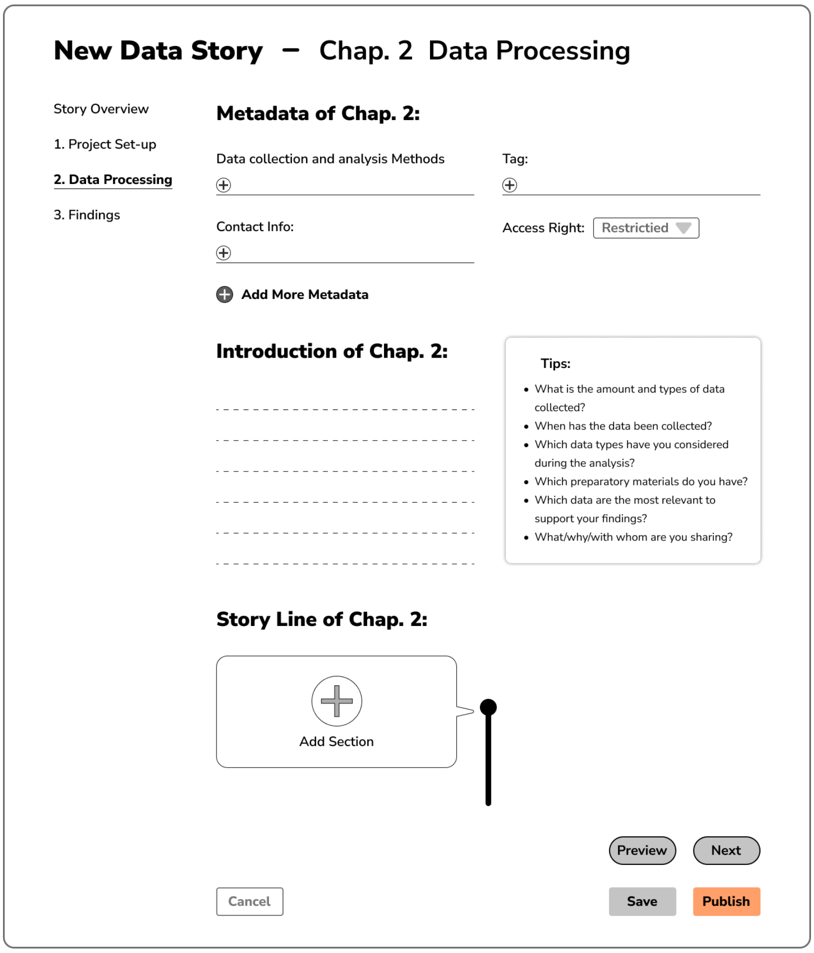


Figure 1: First sketches of the Data Story concept (February 2021).

To simplify the possibilities, we created three main chapters: 1) project set-up; 2) data processing; 3) findings (see Figure 2 below: Data Story overview). Open text fields for writing narratives, tags, relevant metadata and a DOI were organised all along the three interface chapters. Especially in the data processing chapter, researchers would showcase pre-selected data, organised them in sub-sections, and visualised them along a timeline. To better support the data creators in engaging with narrative and storytelling practices, we highlighted relevant guiding questions called 'tips' next to each open text field, that researchers would use to structure their stories and contextualise their data. Finally, we envisioned a plugin for different tools (i.e., Word, Sciebo, Maxqda etc.) that would allow researchers to easily add new data to their stories 'on the go' while still actively working on their research projects.



New Data Story – Chap. 2 Data Processing

Story Overview

1. Project Set-up

2. Data Processing

3. Findings

Metadata of Chap. 2:

Data collection and analysis Methods

Tag:

Contact Info:

Access Right: Restrictied

Add More Metadata

Introduction of Chap. 2:

Tips:

- What is the amount and types of data collected?
- When has the data been collected?
- Which data types have you considered during the analysis?
- Which preparatory materials do you have?
- Which data are the most relevant to support your findings?
- What/why/with whom are you sharing?

Story Line of Chap. 2:

Figure 2: Data Story processing chapter. Link to low fidelity the prototype: https://bit.ly/3ry9mH2

Although providing an overview of the design features and activities is important to understand the core of our contribution, it is not our focus here to provide detailed on those, as they have been already explored in the previous publications mentioned above. In this contribution we want to focus the on the evaluation-based appropriation studies of the designed concept and related prototypes. In what follows, we provide details concerning the evaluative work we conducted and illustrate how the prototype changed accordingly and how progress was made on the wider question of supporting the appropriation of RDM practices. The next section includes few quotes from our evaluation activities to directly show participants' point of view and reactions and how those changed the design of the Data Story.

8.3.2 'Embedded' evaluation: Shaping design through users' feedback

Evaluation, of course, can take many forms. It can be conceived of, for brief mention, as 'summative', 'formative', 'diagnostic', 'situated', and so on (Chambers 1994; Irani 2010; Kaye 2007; Ledo et al., 2018; MacDonald and Atwood, 2013; Remy et al., 2018; Twidale et al., 1994). The character of each is shaped by epistemological assumptions, pragmatic considerations, and overall purpose.

As mentioned above, our overall ethnographic approach is characterised by a long-term engagement and by member participation, while the type of evaluation conducted can be described as 'embedded' (Lewis and Russell, 2011), due to the nature of our participation in the CRC, which is long term, involves ongoing interaction with participants during formal and informal meetings, and is participative but at the same time constrained: the aims of all participants are restricted by the institutional framework and expectations of our own funding agency – DFG.

An important element of this work is hence the double role that the first author assumed in the field, being both a researcher and affiliated member of the CRC in question. As members of the CRC ourselves, we are part of the context we were called to design for (and with). We always positioned ourselves in a constant dialog with the researchers involved whom we met regularly during informal encounters, official plenary meetings, and seminars organised by ourselves or others in the CRC. Therefore, all interactions (formal and informal) were part of our evaluative work.

Another salient aspect of being an embedded researcher is a sustained didactic element in the engagement (Jenness 2006) where research findings are shared early on with the research participants to stimulate discussions relevant for the institutions to improve reflexivity and practices. In this research context, the DFG agenda, RDM concepts and technicalities needed to be explained, discussed, and negotiated according to the interests, needs and practices of the CRC's researchers, and our research was used as a vehicle to do so. We quickly became the medium through which meanings emerged and negotiations between institutional points of view and actual practices took place. We were 'the translator'. Our work aimed at 'making visible the invisible work' of data work and, for this reason, our research was perceived by some researchers as threatening and frustrating while for others was seen as an opportunity to discuss how to better improve current data practices.

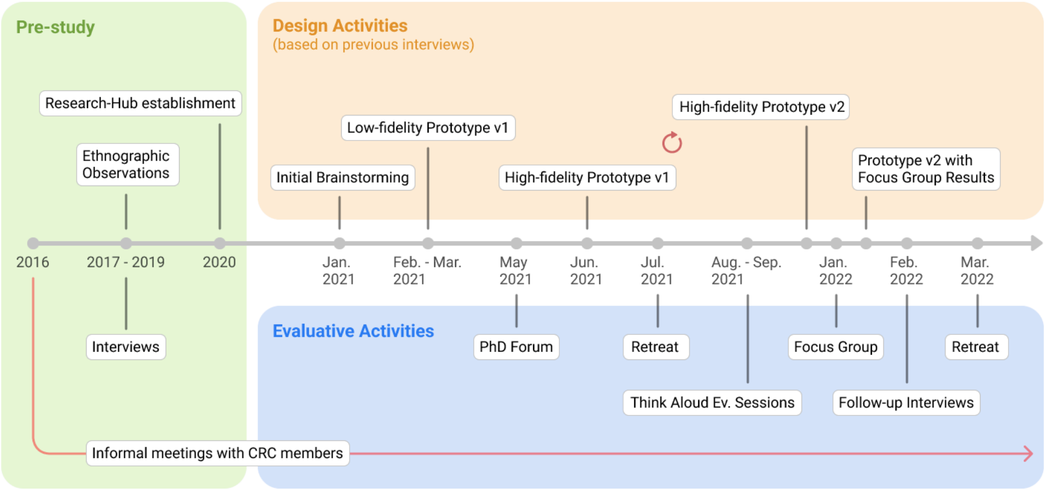


Figure 3: Embedded evaluation: overview of fieldwork, design, and evaluation activities

As shown in Figure 3, initial brainstorming and the low-fidelity prototype were grounded on previous interviews and observations, while evaluation of feedback on our conceptual design was done initially in a PhD forum (May 2021, with twelve participants see table 2 for participants overview), and in a strategic planning meeting locally known as 'Retreat' (July 2021) where all CRC's projects (including our own) were invited to discuss their latest updates concerning publications and research findings (thirty-four participants attended). On both occasions, the first author shared with the participants the low-fidelity prototype and the draft of a conceptual paper which described it. Researchers were enthusiastic with our initial concept, with our interpretation of their RDM issues, and with the new opportunities that a Data Story could offer. As one PhD student told us:

> I really like the idea of combining (just a) few metadata and organised the data and information across the research process that you divided in chapters. I like the fact that you could use a Data Story over time and add more data to it. In this way, you could use the interface to discuss relevant data with your colleagues and even with others who do not directly work with you. (PhD forum, May 2021; PhD Student in HCI)

Another Postdoc said during the Retreat:

> Data Story could be used to collaborative craft publications outcomes based on specific relevant data but also as a possibility to present to a wider audience how data practices actually unfold. I find this approach very exciting. I really want to use it at some point to see how it works. (Retreat, July 2021; Postdoc in Media History)

Activity	**Pseudonym**	**Background**	**Academic Role**	**Date**
PhD forum	Alfred	HCI	PhD	May 2021
	Franka	Media Science	PhD	

	Elvis	Media Science	PhD	
	Sophie*	Media Science	Principle Investigator	
	Bob	Media Science	Postdoc	
	Jack	Business Informatics	PhD	
	Julie	Computer science	PhD	
	Aaron*	Business Informatics	Post-Doc	
	Sarah	Social Science	PhD	
	Carol	Political science	PhD	
	Will	Anthropology	Principal Investigator	
	Elijah	HCI	PhD	
Retreat	Number participants: thirty-four researchers			July 2021
Think aloud evaluation	Claudia	HCI	Ph.D.	13.09.2021
	Oliver*	Media history	Postdoc	05.08.2021
	Karl	Computer Science	Postdoc	06.08.2021
	Paul	STS and Media Studies	Ph.D.	06.08.2021
	Rose	Economics	Ph.D.	16.08.2021
	Marie	Educational Science	Postdoc	24.08.2021
Focus group + interview	Alex	Software Engineering	Ph.D.	20.01 + 10.02.2022
	Franziska	Media Science	Ph.D.	20.01 + 25.02.2022
	Dave	Computer Science	Master	20.01.2022
	Max	Sociology	Postdoc	20.01+15.02.2022
Total participants: 56				

Table 2: Evaluation activities with participants' overview: background, role, type of evaluation performed with them and date. All participants have an interdisciplinary background and apply qualitative and ethnographic methods in their research with various degree of expertise.

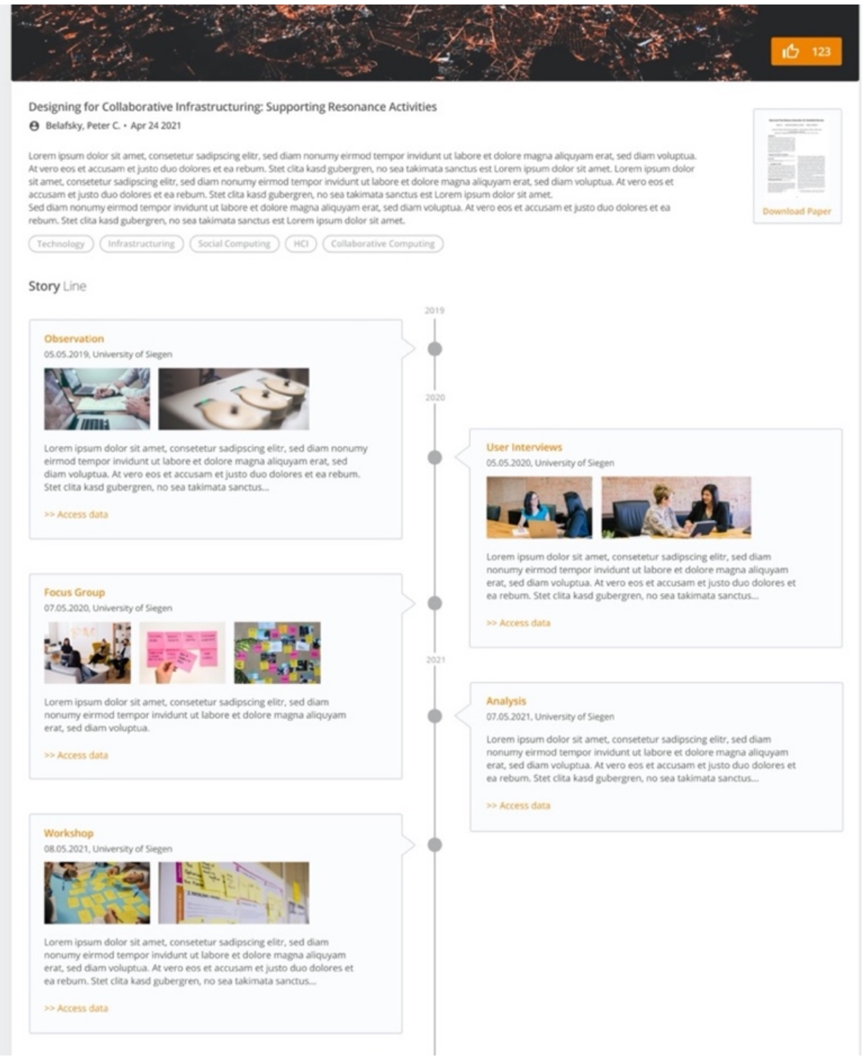
8.3.2.1 Thinking aloud evaluation sessions

After this initial positive feedback, we decided to evaluate the prototype workflow in the actual interface of the Research-hub platform where the Data Story is planned to be fully implemented. We especially wanted to find out what researchers liked or disliked about our design, how they would engage with its workflow, what was missing or unclear, and what further ideas or expectations researchers might have. We then designed a high-fidelity prototype that mimicked the Research-hub platform interface but with the same features and structure of the low-fidelity described in section 3.2. With it, we ran six individual thinking aloud evaluation sessions between July and August 2021.

Thinking aloud is a technique traditionally associated with usability testing, where users verbalise what they are thinking while they interact with the system or technological artefact – both in terms of positive and negative thoughts about the interface and difficulties to interact with it (Nielsen 1993). In our evaluation activities, we were not concerned with usability per se. We were more interested in the participants views on our concept, their ideas for improvements, and predisposition to engage with the Data Story approach in their everyday work practices.

Overall, three graduate students and three Postdocs representing all major disciplines were invited to join the sessions via Zoom (see Table 2 for participants' overview). Each participant received the clickable prototype link at the beginning (access here: https://broad-smoke-1273.animaapp.io/data-story-02), then the first author instructed them to share their screens, engage with the Data Story workflow and provide feedback by thinking aloud (Van Den Haak et al., 2003).

The initial feedback, collected in the PhD forum and Retreat, were enthusiastic and positive. However, when confronted with the first high fidelity prototype, researchers were more critical, and some scepticism was again expressed. Researchers were especially discouraged by the amount of metadata and number of input fields distributed across all sections. They spotted some redundancies concerning metadata and tags, and they found some metadata confusing and difficult to fill in. In general, they were confused with the purpose of a Data Story in the first place and wondered why one would put to so much effort into it.



Designing for Collaborative Infrastructuring: Supporting Resonance Activities

Belafsky, Peter C. • Apr 24 2021

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.
Sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Story Line

Observation
05.05.2019, University of Siegen

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus...

>> Access data

User Interviews
05.05.2020, University of Siegen

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus...

>> Access data

Focus Group
07.05.2020, University of Siegen

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

>> Access data

Analysis
07.05.2021, University of Siegen

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus...

>> Access data

Workshop
08.05.2021, University of Siegen

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus...

>> Access data

Figure 4: Second version of the high-fidelity prototype redesigned according to the feedback of Thinking Aloud Evaluation Sessions. Link to the prototype: https://bit.ly/3ehmFEN.

Based on this feedback, we modified the prototype and created a second version with fewer sections and less metadata. We removed the option to provide metadata for single files and focused the design on open narratives and open input text fields. As shown in the Figure 4, the prototype lost the rigid chapter structure but maintained the timeline of data and related methods. More emphasis is given to the narrative itself, data, and methods, to be described with open text fields.

8.3.2.2 Focus groups and follow-up interviews

During all evaluative activities, participants mentioned repeatedly how they missed the opportunity to engage with the actual writing flow, they were concerned with how long the writing would take, and how a Data Story would look in the end. Therefore, we organised a focus group to discuss specifically the writing process and with the goal of creating the first sample of users' Data Stories. The focus group was organised around two solo-writing timeslots (40 min each) and two plenary discussions timeslots (45 min each). Four different participants were invited this time (see Table 2 for overview). Researchers were invited to selected beforehand a few sample data (pictures, interviews, surveys etc.) that they had collected during their research project and that they imagined sharing with an external audience via the Research-hub platform.

At the beginning of the workshop, we briefly introduced the Data Story concept and showed the high-fidelity prototype. We created an online form with the tool Tripetto (https://tripetto.com/product/) that allowed us to collect and save all written stories and sample data in a WordPress database uploaded by the participants (link to the Tripetto online form used for the research: https://tripetto.app/run/LPJKU480IY). After the focus group, we copied and pasted all stories and data researchers uploaded (via Tripetto) into the new interface design. We also included social media features, such as likes and comments, as suggested by one of the participants during the discussion to provide with a stronger feeling of the potential interactions. Finally, we had one-hour follow-up interview with the focus group participants to discuss the Data Story visualization and interface navigation. One week after the follow-up interview, one of the participants came back to us a with the following feedback:

> This has been fun. I made some reference to the tool at today's meeting on the annual conference because we were talking about the need for new forms of presentation (actually, also briefly discussed the upcoming Retreat). I guess there's plenty of interest, at least on the doc/postdoc level (email sent by Max, a Postdoc, to the first author).

8.3.3 Data collection and analysis

All interactions mentioned up to this point – the PhD forum, Retreat, thinking aloud evaluation sessions, focus group (plenary sessions), follow-up interviews – took place via zoom due to the pandemic restrictions. They were all video recorded and transcribed 'ad litteram'. For all the other informal interactions, meetings, or seminars we wrote fieldwork reports. The thinking aloud evaluation sessions and the follow-up interviews lasted in average 1 hour.

After repeated reading, all data were open coded (Strauss and Corbin, 1998), structured into approximate categories, and thematically analysed (Gibson and Brown 2009), Iterative data analysis sessions took place between September and October 2021 (for the thinking aloud evaluation sessions) and between February 2022 and April 2022 (for all data combined). The first author, as data collector, was leading the sessions. In the very first analysis sessions, the first author and more experienced researchers met to discuss, adapt, and sometimes align the emerging codes, following a broadly inductive analytic procedure (cf. Thomas 2006). Then, the first author conducted an initial round of thematic analysis using the software MaxQDA. Subsequently, the second author reviewed the transcripts and discussed preliminary code groups, such as “data overview” or “data organization” with the first author. Two more rounds of iterative coding were performed to consolidate similar code groups into higher order categories, such as “collaborative benefits” and “RDM issues”. All authors regularly reviewed and revised the codes and categories to uncover the connections between the categories and eventually defined the broader themes, leading to the three main findings that are discussed below: 1) personal and collaboration benefits connected to sharing data, 2) RDM issues and expectations, 3) open issues and fears.

The focus of the evaluation and analysis was not on the tool or the interface itself, but rather on what we had learned through this evaluation process concerning how to foster new research (management) practices. The focus was on how to analyse the way in which researchers reasoned about how to think, select, describe, and write about data when engaging with the Data Story, and what issues emerged in doing so. The ongoing evaluation, then, was critical to our emerging understanding of how to foster RDM practices in collaborative research contexts. It enabled us, simply, deeper into researchers’ expectations, hopes, and fears.

8.4 Findings

In this section we report on the findings concerning the above-mentioned research question. The first section highlights the benefits that researchers hoped to get from a tool like the Data Story, and stresses those benefits connected to sharing and collaboration research (data) practices. The second section explores issues concerning metadata and curation work while pointing to how researchers could increase their awareness and learn to do this type of work through Data Story. The last section digs deeper into general issues or open questions and explores some anticipated issues that researchers talked about when imagining a Data Story becoming commonplace in academia. Each of those sections is an important building block of the overall answer to our research question. The implications of this are discussed in section 5.

8.4.1 Identified benefits for research collaboration and sharing

In the focus group, participants engaged in an animated discussion and spelled out several benefits and concrete use cases in which a Data Story could be helpful. For example, Max mentioned how he sees a lot of value in the concept, in the data contextualization and visualization suggested in the prototype. He hoped, for instance, that it might replace the sharing of long papers in CRC's meetings, such as the Retreat and research forums, because in the end "nobody reads papers in detail for lack of time". Data Stories, thus, provide a quick entry point into ongoing collaborative research projects where authors can explain essential information and even display relevant data like interviews or observations. As Max put it:

> … it can open opportunities for different discussions and different type of questions to be asked in plenary meetings. […] it forces you to write the essential and test if others understand what you want to say and what your aims are (Focus group, January 2022; Postdoc in Sociology).

In general, participants saw benefit in the time they spent in "sitting with their data" which was useful to them for structuring the major insights of their research process while also having a format specifically targeted to show these insights to others.

Peer learning opportunities were also highlighted. In fact, Alex had graduated in software engineering and when he joined an HCI department few years ago, he struggled in adapting to the new research environment. He joined an already existing project that had started two years earlier. Data collected from other colleagues were not accessible, so it was even harder to understand what had been done up until that point, to learn from others and/or to start analysing materials already collected from other graduate students. If Data Stories had been available when he joined, he said, he might have had the chance to learn faster how the HCI and CSCW communities deal with data, which methodologies are applied and how. Franziska had a similar experience. She started her project one year later and she needed the overview of what they did before her time, so she decided to visualise her own data in order to get an overview and prompt discussion with other colleagues:

> We created a lot of data, and it was also difficult for myself to have an overview. I also visualised it. I discussed the visualization with my colleagues from the other faculty because they didn't know everything that was happening, so it was very good to discuss it together and we used it also as a basis for writing papers just to know what kind of data do we have, what kind of insights did we get (Follow-up interview, February 2022, Ph.D. student in Media science)

In general, participants highlighted the need for an overview and data organization which many of researchers struggle to have. It seems they are in a constant search for tools or new methods to visualise what has been done collaboratively. Franziska added that she has been searching for quite some time for a tool where she can present their results to the funding agency, as a way to provide them with a quick overview of their data collection and research achievements.

The Data Story fits this specific need, where links to stored data folders can be established to prove that data exists somewhere, and they are stored safely. Others envisioned Data Stories to be used as prop to collect data in the field, inviting participants for example to create their Data Stories and collaboratively gather data. This is a need that was expressed by one CRC's project where researchers interested in 'decolonizing ethnography' (Bejarano et al. 2019) have been searching for tools where participants could be engaged from the beginning in the data collection to support researchers' claims.

Finally, others stressed the impact that Data Stories could have in the long-term, specifically for re-use or for guiding new line of research and research questions. As Oliver put it:

> Funders want research data to be collected and archived and the question is 'where would it be?' Should I put them on an anonymous archival environment and then it's there for eternity? Or wouldn't we have to invent new formats of decentralised devices connected through the DOI, so that the published texts are somehow connected to their materials?' (Thinking aloud session, August 2021, Post Doc in Media History).

In fact, 'anonymous', remote archives, which are removed from where data are actually created, are often perceived as an additional burden and researchers do not see a benefit in archiving data there. Data Stories instead emphasise the organization, the overview, and analytical insights that researchers want to get from 'their' data, initially for themselves, and later, potentially, provide it to others.

8.4.2 Data curation and metadata issues

The first high-fidelity prototype integrated technical elements such as the tags, metadata and DOIs to support data curation and retrieval along with open text fields for open contextualization and narratives. However, during the thinking aloud evaluation sessions, researchers found it surprising but also confusing to see these technical elements. For example, Rose was confused, because she wasn't sure what metadata really are and what purpose they might have in the process. As mentioned in section 3, after this feedback from the focus group, the prototype was redesigned. The majority of metadata were removed and we left the categories as more open-ended, and we almost lost the 'traditional' curation aspect. However, in one of the plenary sessions we discussed the issue of standardization which can be connected to the role of metadata. Researchers agreed that standardization would make the process faster and could help in mapping the major methodologies used within a specific research group but also it could generate internal discussions concerning the development of methods by showing in-depth descriptions and sample data that could be compared and might trigger new research collaboration. Max suggested having a workshop in the CRC where together researchers could come up collectively with their own metadata and categories starting with their methodologies and research interests. A couple of researchers also mentioned some metadata elements that could be added as a way of organizing and detailing the data. For example, for interview data they mentioned "place of interview, date of interview, length of interview etc." as metadata that could be helpful to describe single data items and organizing them along the timeline. Interesting to note is that on a different occasion, after a seminar organised on the topic of RDM and curation, another CRC researcher approached the first author to say:

> After the session, I started to think about metadata, and I started doing it, but I am not sure if I am doing right and how to do it, where the metadata should be stored or how to better organise my data" (Informal meeting with a PhD student, Sociology).

As highlighted in our previous work (Mosconi et al. 2019) the tools that researchers use daily do not offer the possibility to enter metadata and link them to each data item. Metadata writing is a task currently being done, if at all, in the end of the research process shortly before the archival submission. What Data Story suggests is an interface through which uploading a specific data item and engaging with metadata work while still working on the research process is possible and desirable. It is also available for sharing information with colleagues and/or an external audience in a timely fashion.

Lastly, researchers suggested to provide info boxes that could explain in detail the technical features, such as the DOI, the metadata and the tags, so that users could learn about them and understand why they are there and how to make use of them. Other info boxes might be included in the data upload section to explain anonymization, ethical and legal policies. These are important aspects that are often not explained anywhere. They influence how to curate the data and what one can share, but researchers often lack knowledge. In fact, in multiple occasions, researchers asked the project INF to organise seminar sessions on this specific issue which proves the need for more information, training, and support in the field of RDM and the technicalities involved.

8.4.3 General issues, concerns, and fears

Early on, we decided to provide researchers with a vague definition of what a Data Story actually is in order to allow participants to come up with their own scenarios. However, especially during the thinking aloud evaluation sessions, basic questions came up from the beginning: What is a Data Story? What does it do? Why and how should I write one? For most researchers the three-chapter structure (project set-up, data processing and findings) resonated too much with the structure of academic papers and they wondered in what way a Data Story differs.

Besides stylistic choices, some researchers struggled with the documentation and with the selection of data to show in their Data Stories. For example, Paul asked: *"How would I document that so that people actually understand the interesting insights I had with this story?"*. Paul and a colleague participated at a summer school where they had to illustrate a case study on users' interactions with apps and present the methodology. They wrote a presentation but, they said, it was hard to convey some of the most interesting questions they had from the dataset, conceptually but also methodologically. During the focus group, the guiding tips were proven helpful in supporting researchers in crafting their narrative and the structuring of the data processing chapter. However, researchers suggested to have a clear separation between the data uploaded and the insights derived from it so that potential reader could better distinguish between a piece of data, personal interpretation, and reconstruction of the analytical process.

To better accommodate Data Stories that are connected to ongoing research, Oliver encouraged us to offer the possibility of starting writing data stories from the data and method section, because:

> To what extent do I have to know my story in advance? Am I able to create my story by feeding new bits and pieces and kind of bringing them into an order and swapping them around this storyline until I find that it has somehow become a narrative? That would be something I'd love to know from a design perspective. If it would somehow help to find the narrative, that's something that could be really interesting as a tool (Thinking aloud session, August 2021, Post Doc in Media History).

In his view, this would potentially allow for bottom-up categories to emerge and to use the Data Story also as an analytical tool. Again, this refers to personal benefits that researchers might see while engaging in data work and their interests in having tools that could support ongoing research.

Our participants also voiced some opinions about Data Story becoming commonplace in academia. Max stressed how some features, similar to those found on social media, could hinder user engagement because some academics might not want to be exposed. Finally, in the focus group, the fear of losing control of the data and data protection came up as an important topic. Concerning this, Max suggested a feature called "visible for a day" because some people might feel uncomfortable "with having data openly accessible in perpetuity".

8.5 Discussion

The findings illustrated above demonstrate the evolving nature of user reaction to the design as it iterated. As we have stressed, because of our participation as members in the institution, our ongoing interactions with CRC members, and our active research into the issues over a long period of time, we conceive of our efforts as being ‘embedded’ (Lewis and Russell, 2011). This means that separating evaluation from other investigative processes was neither possible nor desirable. Data Story became both the topic and the medium through which we were able to understand how data curation and sharing practices can be introduced in researchers’ daily workflows and how researchers can profit from them. Our contribution highlights lessons learnt through our embedded engagement and provides a new design approach for RDM and for new research data practices. This implies: 1) establishing a consensual and gradual process for data curation practices to unfold over time; 2) negotiating metadata readability, flexibility, and standardization through interface design; 3) prompting conversations and learning opportunities with and about data.

8.5.1 Introducing RDM into collaborative research practices: Lessons Learned

Our initial aim with Data Story was, then, to investigate the priorities that researchers had in respect of data curation, sharing and reuse. These RDM endeavours require the acquisition of data management skills, but the current scientific culture and rewards system do not directly incentivise or yet, recognise these endeavours (Fecher et al., 2017; Feger et al., 2020; Kervin et al., 2014). We had no preconceptions about researchers’ priorities but had, in previous work, identified many of the issues they faced when confronting a top-down mandate (Mosconi et al., 2019). We saw the initial scepticism of some researchers but also a recognition of potential benefits connected to sharing and collaboration research practices that are otherwise not traditionally considered in the RDM discourse. In fact, researchers showed interest in learning from others how to do research, how to meaningfully show their own work to others, how best to collaborate together asynchronously, and how to provide an overview of what has been done. The emphasis is also on the user-orientation with transparency in roles and profiles of data workers and collaborators (RfII 2016). The ‘data overview’ is something that both researchers who collect the data and others interested in the data struggle with. At times, researchers come up with informal practices to visualise their own fieldwork activities and their most important data (as shown by Franziska). As our research participants confirmed, the effort of curation, facilitated and supported through Data Stories, can positively impact how researchers work, and can repay them in providing a structure, assisting them in keeping their data organised or deepen their analysis. In turn, it could make the process of writing publications faster because people can organise and reflect on their findings in and through their curatorial activities elaborated with written narratives.

8.5.2 Curation as a consensual and gradual process

Our findings suggest that a solution like Data Story will need firstly to provide features that researchers benefit directly from (i.e.: having the overview, drafting papers, collaborate etc.) and then gradually also introduce curation elements. It also requires a long-term processual perspective for RDM activities which allows researchers to learn new practices as part of their membership of the research infrastructure (Feger et al., 2020; Mosconi et al., 2019). Thus, a gradually emerging consensus around mutual benefit, we anticipate, will consolidate RDM practices and provide learning opportunities (Cox and Verbaan, 2018). The first thinking aloud evaluative sessions focused on a very advanced version of the Data Story concept and the related prototype. It had plenty of metadata. It had a lot of different sections. It had metadata for the story and metadata for files, leading to non-uniformity in practices for metadata curation. Researchers found this type of non-uniformity in data descriptions and the amount of it quite overwhelming. They were confused about the purpose of a Data Story in the first place and wondered why one would put to so much effort into it. Indeed, our earliest prototype proved somewhat paralysing and counter-productive because it attempted to provide an all-encompassing solution. We subsequently adopted what one might term a ‘gradualist’ solution, one which emphasised the immediate benefits of sharing by focusing on the Data Story as an iterative process, focused on what researchers were interested in but which also, through flexible design, would allow for the addition of other elements. The gradual expansion of metadata is an example of this. With time, from within, we anticipate that we will be able to build a workflow process, based on new standards and folksonomies that will emerge directly from users’ interaction and needs; and support the appropriation of new practices of curation, sharing and reuse that can be data-driven, negotiate between top-down policies and bottom-up practices, and that can grow and evolve so as to service more distant needs (Pryor 2014; Pryor et al., 2013).

8.5.3 Negotiating metadata readability, flexibility, and standardization

The work of Koesten and Simperl (2021) has previously stressed the importance of narratives and textual documentation needed in order to facilitate data sharing and reuse. Data Story embraces this finding and supports the elaboration of narratives, conceived as "readable metadata for human interpretation", which can highlight the "social function of data" (Birnholtz and Bietz, 2003). Especially with qualitative data, narratives are the vehicle through which researchers perform interpretations, engage reflexively and elaborate data through sense-making (Pepper and Wildy, 2009). The guiding questions (called 'tips') included in the interface design (see section 3.2) aim specifically at supporting such a narrative structure by helping researchers to explicate and organise the implicit knowledge gathered through interactions and observations in the field. It resonates to a degree with the Data Curation Profiles project (Witt et al., 2009) but instead of gathering only metadata and sample data our design aims at making explicit a broader context with open-ended narratives combined with the addition of metadata, data files and other relevant materials (i.e.: interview guidelines, informed consent etc.).

There were evident issues in the emergent logic of the Data Story in relation to, on the one hand, the need for some kind of structure but, on the other, the need for a flexibility in representation which allowed researchers to order matters in ways which were relevant to their work. That flexibility, allowing for their rationales to become visible in their ordering practices, was a useful adjunct in respect of acting as a medium for their own reflections, providing an ongoing, visibly historical document, and providing a medium for engagement with others at various points in project endeavours (Whyte 2014). Specific benefits brought out included the idea that the Data story provided a quick overview, obviating the need for tedious reading; provided a prop for future data collection and analysis; and could replace other forms of sharing which are typically more difficult to find and access. These added degrees of flexibility, however, will need to be negotiated and balanced with some requirements of standardization, for example represented by the metadata elements, which are needed specifically for data retrieval. As suggested by Max, we plan in our future work to identify (through participatory workshops) relevant categories and metadata standards useful to describe methods and data that will be used in conjunction with flexible narratives.

8.5.4 Prompting conversations and opportunities for learning with and about data

As mentioned in section 2, research infrastructures should channel improvements in communication around research artefacts because anything that is shared can in principle be of interest for somebody else so both data creator and recipient need to be allowed to exchange information (Birnholtz and Bietz, 2003; Neang et al., 2021; Thomer et al, 2022). Data Story, even at an early stage, seemed to prompt reflections and conversations about data and its uses. Participants argued that it both stimulated and facilitated conversations with colleagues (and others), encouraged them to be more reflective about their data (the act of building the Story was itself part of an ongoing analytic process), prompting precisely the kinds of thinking about data that methodologies such as grounded theory (see e.g., Muller and Kogan, 2010) seem to recommend. As researchers like Max said, *"it encourages you to think of data, what is the most interesting insights in your data"*. Highlighting what are the most interesting insights from the data at hand is otherwise difficult, especially when drawing the attention of others to it. Data Story encourages researchers to record thinking through practices such as dropping notes into it. It makes data-work visible and present and, as such, facilitates the building of analytic insights while being in conversation (with yourself or) with someone else. We conceptualise these various opportunities as 'reflective' learning opportunities (Boyd and Fales, 1983). Reflective learning is the internal examination and exploration of a concern prompted by an experience, which produces and clarifies meaning in terms of self and leads to a shift in conceptual viewpoint (Boyd and Fales, 1983). In fact, not everyone is equally familiar with the ways in which data is collected, organised, and used in research. In the interdisciplinary contexts we have been involved with, dealing with qualitative data is a new experience for many new researchers and the existence of prior examples which provide rationale for methods adopted or for analytic choices made has proven valuable. Therefore, Data Story can be thought of as an interface which affords learning opportunities (with and about data) of many kinds, above all in relation to research methodology and RDM. It encourages researchers to sit together with their data, curate them and share them, while at the same time supporting them with the organization of their materials and reflection on what they are sharing, who are their sharing with and why. As we move on in this RDM era, data skills are crucial but to learn them, we will need more than just standard routines or pre-defined guidelines, fixed metadata, and categories. As data (and data skills) are the results of ongoing, even serendipitous, learning opportunities and personal (internal) explorations - in relation with a vast ecology of tools, methods, practices - in constant evolution.

8.6 Conclusion

Solutions to support RDM collaborative workflows are clearly needed. First and foremost, these solutions need to provide benefits to data creators in order to motivate them in using them (Feger et al., 2020). As already highlighted by Rolland and Lee (2013) “investigators need ways to engage in data curation in support of tomorrow’s research without delaying today’s.” (p. 443). In the above, we have demonstrated the opportunities and challenges associated with an alternative approach to RDM which might support these activities in a meaningful way. We have done so through a focus on ‘narrative’ and the construction of useful and reusable narrative structures. The need for this comes out of the complex and interwoven strands we have examined, and which are not easily reduced to single constitutive elements.

Our work is predicated on an investigative policy we have called ‘embedded evaluation’, involving ongoing work by ourselves and others as joint participants to a number of research projects where data curation, sharing and potential reuse has become an issue. Our design was guided by an attempt to negotiate between various interests, and it was in a sense constrained by the funding agency agenda, the INF goals connected to it, and researchers’ concerns and wishes. Our motivation for the work emanated from the realization that the people we worked with in a largely interdisciplinary context are often not trained in, nor used to, data curation and sharing. For the most part they have few resources with which to develop an understanding of the way qualitative data can be organised, what it might be used for, or who it might be used by, nor there are solutions yet that really support the development of a (data) sharing culture within and beyond research groups. What we describe are some steps thus far taken towards meeting that objective. In fact, Data Story offers a simple, and structured way to gain, so to speak, a flavour of the work in question, its epistemic assumptions, its methodologies and specific methods, and its positioning with respect to other work. Naturally, future potential re-users should be kept in mind. We foresee that Data Story can potentially be used for what we would term ‘anticipatory’ articulation work, meaning supporting not only articulation work in respect of current cooperation, but also the work for future cooperation not yet known. The point there is that, in normal organizational life, the kinds of articulation work that are necessary are more predictable. Roles and responsibilities, at least to a degree, are known. That is not the case here. There is no clear agreement about what the responsibilities of active researchers might be, and it is very difficult to anticipate what uses shared data might be put to, and who by. In this sense 'anticipatory' articulation work would refer to the work to make future cooperative work possible, in a situation where data work will be fluid, dynamic and mediated by heterogeneous purposes. The Data Story, we argue, provides an entry point into the sensemaking work that will be needed. The focus, then, is on a development from 'anticipation work', i.e., "the practices that cultivate and channel expectations of the future, design pathways into those imaginations, and maintain those visions in the face of a dynamic world" (Steinhardt and Jackson 2015, p. 443). We plan in our future work to examine practical implications for research collaboration and RDM in more detail by looking at the kinds of sensemaking that go into narrative structures and the way they are received by others in real contexts.

To conclude, the Data Story, as we call it, is predicated on an amalgam of some orthodox data science constructions and a more flexible, narrative approach. The latter aims to embed the history and the emergent rationale behind the organization of the data and that can highlight "the social function of data in the community that created it" (Birnholtz and Bietz, 2003).We do not imagine that the Data Story will, in and of itself, produce radical and systemic changes to data curation, sharing and reuse practices. Data curation and sharing practices are very much contingent on when and for what reason, and with whom data is to be shared (there will, for instance, be a significant difference between sharing data with other team members, re-using data oneself, and curating it for unknown future users). We do, however, see, in embryo and along with our colleagues, how we can address the need to start developing sharing and RDM strategies step by step, building bottom-up communities of (data) sharing practices in and through the progressive adoption of the solution we describe. We take on board the injunction of Feger et al. (2020) regarding the transition to effective digital RDM and the role of HCI in it: we, as HCI and CSCW researchers, can facilitate the design of interfaces that can support collaborative data work, learning opportunities, encourage reflective thinking, and making data work visible, so that it can be better organised, meaningful, and worthy of our time.

III

Part III - Discussion and Conclusion

In this chapter, I return to the research questions mentioned at the beginning of the thesis and try to provide some answers connecting relevant previous studies and the findings presented in part II. My work demonstrates the sheer complexity of the issues. The problems of how to curate and share data, what data, who to share it with, and when, have been examined from a variety of points of view including so-called data science, research data management, data curation, data sharing, 'storytelling' and other narrative approaches, and so on. All share one common feature, which is that dealing with data is much more than just a technical problem (Tenopir et al. 2011; K. Kervin, Cook, and Michener 2014a; Christine L. Borgman 2012; Curty et al. 2016; Tsai et al. 2016; Birnholtz and Bietz 2003a; Velden 2013; Feldman and Shaw 2019; Plantin, Lagoze, and Edwards 2018) implicating problems of 'overhead', organizational structure, self-interest, timeliness, and audience. None of these have as yet been fully resolved. The major gap in the literature examined in this thesis, however, has been the problem how to deal with qualitative data when drawn upon by an interdisciplinary audience. Qualitative data brings with it issues that are even more problematic than those encountered with data of a more structured kind. These problems include (Mosconi et al. 2019a) the very special status of data privacy when conducting ethnographic research and storing data; the heterogeneous nature of assumptions in interdisciplinary working and the varied terminologies used; the lack of training and skill and the absence of data managers. In Mosconi et al. 2019, we argued for 'sheer curation' as a general way in which one might address some of these problems. Sheer curation is an approach to information management that emphasizes the importance of designing digital infrastructures that enable users to effortlessly organize, structure, and curate data as part of their everyday work practices, without the need for separate and distinct data management activities. The approach involves designing digital infrastructures that are flexible, malleable, and adaptable, allowing users to structure and curate data as part of their everyday work practices. In other words, the approach aims to make data management an integral and seamless part of work practices, rather than a separate and distinct activity. This is in fact what the Data Story and the Research-hub platform strive for.

Discussion

Below, I structure my findings under headings which reflect the research questions I originally raised. I then go on to discuss the challenges that remain and speculate to some degree on what the future may hold.

RQ1: What are the socio-technical challenges for the appropriation of RDM practices in qualitative ethnographically driven research contexts?

9.1 Standardization and idiosyncratic heterogeneity: what is good data?

The CRC is funded by the DFG who demands that researchers release data in institutional repositories at the end of the project. This means data will have to be documented and delivered with metadata according to specific standards. Moreover, the DFG claimed that while observing subject-specific requirements, "standards, metadata catalogues and registries are to be developed in such a way that interdisciplinary use is also possible" [43]. This request sounds extremely ambitious and burdensome considering that in the interdisciplinary contexts we engaged with, researchers themselves are called to organize data for sharing, long-term preservation and ideally secondary use without any direct help from data managers. This creates a significant overhead in terms of time, effort, and resources for researchers who are already stretched thin by their other research commitments. Furthermore, interdisciplinary research often involves idiosyncratic heterogeneity, meaning that data sets may be very different from one another and may not easily fit into standardized categories. Researchers must engage in extensive articulation work to bridge these gaps and make their data sets interoperable and understandable by external audiences. This includes developing data sharing agreements, data management plans, and metadata standards that meet the specific needs of their research contexts. This work can be time-consuming and challenging, particularly for researchers who are not trained in data management. In this way, our case differs from the US LTER network studied by Karasti et al. (2006) in which data managers were involved in understanding and supporting data stewardship and where data managers developed expertise in RDM domain for more than forty years. The LTER case is emblematic of the laborious and ongoing processual endeavor which Open Science initiated and requires all disciplines to undertake. Even in the LTER, after more than forty years, Open Research Data is still an

[43] https://www.mpg.de/230783/principles_research_data_2010.pdf

unresolved issue in practice and posed unprecedented challenges to the actual conduct of science (Helena Karasti, Baker, and Halkola 2006a).

The case of interdisciplinary ethnographically driven research environments I engaged with poses, as my papers show (Mosconi et al. 2022; 2019a), even more challenges in respect to research data management and long-term stewardship, due to the specific characteristics of the data gathered - which imply ethical and legal restrictions not present in other disciplines - but also due to the absence of data managers in dealing with this process who could otherwise relieve researchers from at least some of the overheads caused by it. Igor Eberhard & Wolfgang Kraus (2018) used the metaphor of ‘the elephant in the room’ to describe what they call ‘obvious inconsistencies’ between Open Science expectations and the epistemological peculiarity that distinguish ethnographic field research approaches from many others. In fact, the principles of findability, accessibility, interoperability and reusability, in these contexts, as demanded by the FAIR Data Principles and adopted by all major funding agencies, will be possible to implement only to a limited extent. This is due to the ‘ethical code’ intrinsic to ethnographic methods that impose on researchers the responsibility to ensure the confidentiality and anonymity of their informants (Eberhard & Kraus, 2018). Another issue is related to the so called ‘good scientific practice’ of metadata creation that is expected to facilitate secondary use and interdisciplinary collaboration. As we have seen, in ethnographic approaches ‘metadata’, more appropriately called ‘reflections’, are a crucial element of ethnographic research which is seldom highly structured. Moreover, the metadata, unless redacted, cannot be released without revealing critical information. Anonymization has to be done carefully and it is likely this will cause loss of data density or even uselessness. Ethnographic data can only be interpreted from the social and cultural context – i.e. supplementary information on the framework conditions – therefore it is questionable how fruitful secondary use could be achieved without betraying the ethical norms that ethnographic research has to respect.

This brings us back to the tension between standardization and idiosyncratic heterogeneity. My findings show a huge variety of highly idiosyncratic practices developed by researchers over the course of their career. These are influenced by their intellectual history, by their IT skills, their research interests and by their academic backgrounds. Standardization, if imposed from above, without a deep understanding of epistemological and methodological needs, and the specificity of different disciplines, might have a disciplinary (in another sense) consequence through the labeling of ‘good’ or ‘bad’ researchers by imposing from above standard criteria expected to be met in the long term. In the DDC, a UK data center, it is written on the website the motto *"because good research needs good data"*. I strongly believe 'good data' criteria should be developed not by following general conventions but considering the epistemological value of data in each discipline. From my findings, it can be seen that what is 'good data' in ethnographic research is still an unresolved question. With my work, I suggest a serious debate should take place that aims at increasing the quality of research and data collected without the prescriptive attitude of data sharing. I recognize the need for a more transparent discussion aimed at acknowledging the benefit of 'openness' in increasing quality of research, improving research methods and reflexivity on our own work. However, at the same time, I suggest that good data quality should be identified by researchers themselves together with data curators and policy makers. This calls for an ongoing negotiation of standards between researchers, data curators and policy makers - the core of sheer curation, mentioned above.

9.2 Lack of discipline-specific training on tools: supporting 'resonance activities'

In my findings (Mosconi et al. 2019; 2022a; 2022b), I showed the struggle that researchers have with a variety of tools they use, from simply managing the data storage space with Commercial solutions like Dropbox or Google Drive; organizing, naming and searching files in data sharing system like Sharepoint; to finally analyzing a vast number of interviews with Maxqda following a specific methodological paradigm. Researchers at the beginning of their projects do not often receive specific training on how to (1) to set up a complete data infrastructure or how (2) to use and how to choose from the variety of tools available by the University service provider or by the market. Infrastructure and tools appropriation/usage is something left to researchers to find out on their own. They themselves are called to discover how to make the best use out of the tools array available.

In order to use a tool (and process data) there is a sense-making activity that needs to be performed that has little relation with learning or implementing specific functionalities. It is rather aligned with with the specific (inter)disciplinary practices of using a tool (or data) for certain purposes. In order to perform 'good' research data management, researchers should be trained in tools' usage but tools should be learnt in relation to specific methodological and discipline-specific data practices. We believe this kind of training should be made available by IT-support University services but what is needed should be defined together with researchers from each discipline in order to align institutional knowledge and the expectations of Research Data Management with the epistemological and methodological understandings which are discipline-specific.

As showed in the findings, Research Data Management comes to be somehow problematic and difficult to perform especially with regard to data sharing, the last step of the data life cycle. Researchers, in my view, should receive training and support in order to perform efficiently any activities of the data life cycle (collecting data, storing, data analyzing). I suggest that one way to go is to increase support with training but also building infrastructure for peer-to-peer data practices' appropriation or what Ludwig et al (2018) called 'resonance activities'. In the context of RDM practices in qualitative ethnographically driven research contexts, resonance activities could involve creating opportunities for researchers to collaborate and learn from each other about the use of specific tools and data practices that are relevant to their discipline. This could be achieved through activities such as workshops, peer-to-peer mentoring, or the creation of discipline-specific communities of practice. This is what I have been trying to promote through the establishment of Research-hub: a community-based platform for academic data sharing. Research-hub can promote resonance activities in RDM by providing a platform for peer-to-peer data practices' appropriation, as discussed by Ludwig et al. (2018). Research-hub facilitates resonance activities by providing a shared space for researchers to collaborate, share data, and exchange information about their RDM practices. For example, the Online Drives module allows researchers to share files and collaborate on data analysis. The Metadata Interface Processing module ensures that metadata is captured and stored in a consistent manner, making it easier for researchers to find and reuse data. Finally, the Data Story Module provides a way for researchers to tell the story of their data, making it more accessible and understandable to others. Through these modules and concepts, Research-hub can help to promote resonance activities by facilitating communication, collaboration, and the exchange of ideas and practices among researchers. By promoting resonance activities, researchers can collectively develop a shared understanding of best practice for RDM in their discipline, which can help to reduce the burden of individual researchers having to learn and navigate a complex array of tools and practices on their own. This, in turn, can facilitate the adoption and effective use of RDM practices and tools, ultimately leading to better quality data and research outcomes.

RQ2: How can tools and infrastructures support of the establishment of RDM practices in qualitative and ethnographically driven contexts?

9.3 Focus on narrative. Going beyond metadata models

My findings showed the need to develop tools that could go beyond standard metadata models and to consider the inclusion of a more fluid and narrative-driven approach. In fact, it is widely accepted that data cannot be understood without context (Borgman, 2015; Carlson and Anderson, 2007). However, within the Research Data Management domain this contextual role is assigned to metadata standards and data descriptions. Formal and standardized metadata such as the Dublin Core or the Data Documentation Initiative (DDI) assume not only a contextual role but also, it is claimed, they are essential for the discovery, comprehension, and reuse of data. As stated, on the website of the DDI alliance: "DDI is designed to make research data independently understandable. DDI provides a standard structure for all of the metadata that accompanies a dataset and helps users of that dataset to interpret its contents.[44]". Metadata are often interpreted as 'the bridges' between the producers of data and their re-users, because they should convey the information essential for discovery and secondary analysts. However, filing metadata in order to meet the standards for long-term preservation for potential reuse is a quite tedious and rather technical practice. It could cause delays in the sharing and presentation of collected data and most importantly, due to its complexity and technicalities, it requires the support of data curators or data managers (see LTER as example). In fact, 'metadata critiques' emerged repeatedly during fieldwork. Researchers we talked to struggled to understand the meaning and the applicability of metadata standards such as the Dublin Core which was often mentioned by the IT service provider as the existing metadata standards that researchers should use. However, in our view this specific standard is elaborated in a very general and technical language not useful to, nor familiar to, qualitative researchers. It is very difficult to understand what many of the categories mean in practice. In our specific CRC context, in fact, we don't have data specialists, curators or data managers. In our case, we have researchers (from different fields and disciplines - mainly working with ethnographic/qualitative data) who are asked to provide basic metadata and description for long-term preservation, and potentially sharing and reuse. If they are called to do this work, they should be able to understand what is asked of them clearly and in practice (with limited possibility for misunderstanding). We argue that a storytelling approach to data curation which could integrate metadata and embed stories could be a fruitful way forward, more aligned with researchers' practices.

With the design of a Data Story the intention was to provide researchers with a way of narrating, curating and eventually sharing the heterogeneous data collected during a study. Through the

[44] Source: https://ddialliance.org/training/why-use-ddi searched on 15.02.2021

Data Story, researchers are able to provide a 'data-driven' narrative of the major findings, highlighting interesting results which emerged from a specific project or study (i.e.: snippets of anonymized interviews, pictures, design sketches, short video, personas etc.). The solution aims at 'show casing' a selected portion of data (collected for a specific purpose) by supporting data sense making and Data Stories intended as a creative and active endeavour that should be made explicit by the researchers who conduct the study. It can be thought of as storyboarding for the digital age. The main vision behind the Data Story is to consider 'the story' as a technique to organize and describe *partial and situated* research data. In our view, it could support curation and data sharing practices, because a story might trigger 'data reflexivity'. As De Carteret (2008) puts it: "Stories and conversations create transitional and exploratory spaces in-between the thinking, doing and reporting of research" (De Carteret, 2008, p. 6).

This concept is motivated by the lack of clarity and purpose of data repositories and archives in the large. In fact, researchers from the CRC stated how they saw no benefit in archiving data in anonymous repositories. If we look at data repositories, as potential 'data producers', and reflect on what these infrastructures need in order to function (standard metadata and data descriptions), who the data consumer will be is unclear and undefined, as is how these 'users' or 're-users' will navigate our data and documentation. Researchers are supposed to diligently clean, organize and document using metadata standards, investing a considerable amount of time in this process. How they do this when the data consumer or re-user is unknown remains unclear. Stories, to the contrary, invite you think about and clarify who are we preparing the data for, who are we telling stories for. They invite you to clarify what messages can be found in the data, what questions can be evoked and answered. We can think of this, quite simply, as metadata being for machines to read while stories are for people to read. The Data Story concept invites us to think about how to develop new interfaces and infrastructures which are able to negotiate between metadata standards (machine readable content) and stories (human readable content). Moreover, it suggests using stories as a method to organize and spell out the tacit/implicit knowledge researchers accumulate during the ethnographic research. As De Carteret (2008) reflecting on 'storytelling as research praxis' puts it: "Narrative processes are an interactive activity that organizes experience and knowledge of the world. It is the potential of life stories to raise conscious awareness of the social and ideological roots of self-understanding that is useful, providing opportunities to change (Dhunpath, 2009, 544). Narration is the displacement of an inner reality to an outer reality (Dhunpath 2000, 547). "(….) Writing practices that transgress the traditional academic genre resists objective indifference, as well as sentimental authenticity and empathy (Lather 2000, 16)". A system like the Data

Story, my findings suggest, could help organize in a synthetic fashion the essential information needed to understand at a glance the reasoning behind a specific research setting, behind a research project and related data collections. Data Story is a way to organize the tacit knowledge and the first-hand experience researchers have in the field.

Data don't get collected and analyzed in a vacuum, nor they are shared in a vacuum, they are always shaped, co-created, (partially) shared and narrated based on the specific circumstances in which data are needed and 'put to work'. Storytelling then can be seen an integral part of (collaborative) analysis with qualitative data and a mode of inquire in itself. De Carteret (2008) for example reflects on storytelling as research praxis and illustrates how the storytelling emerged "as a method of inquire and a mode of representing the research […] *where* research stories and conversations create transitional and exploratory spaces in-between the thinking, doing and reporting of research" (de Carteret, 2008, pg.xx). The additional values of stories, moreover, is that they represent the "bridge between the tacit and the explicit, allowing tacit social knowledge to be demonstrated and learned" (Linde, 2001, pg.5) and our work question is how to make use of this knowledge, how to represent it by experimenting in the development of new tools. As Karasti et al. (2021) pointed out "there is a need for method devices that are both agile and flexible enough to be able to deal with configuring, multiplicities, open-endedness, unpredictabilities, and emergence as well as with relations across multiple boundaries" (p.22).

So, the vision of a Data Story is to provide a visualizing (Karasti et al. 2021) and organizing device in support of already existing storytelling practices as a major component of data analysis and sense making. It aims at facilitating (semi)-structured data illustrations useful to organize and/or elicit storytelling practices with and about 'snippets of data' (whether they be interviews excerpts, pictures, video, sketches or any other relevant material). The Data Story vision could then be seen as a digital data storyboard to support collection, organization and data sense-making. As Linde (2001) puts it: "recognizing that narrative is fundamentally social, relying on interactions between people, suggests different ways to capture and transmit it effectively. Rather than focusing on archival storage, it is important to understand and create social mechanisms for narration" (Linde 2001, p.12).

9.4 Tools as boundary object to support interdisciplinary collaboration

Tools can function as boundary objects that facilitate interdisciplinary collaboration by providing a shared language and understanding of complex concepts, enabling individuals from different disciplines to collaborate and share information effectively (Star & Griesemer, 1989).

In interdisciplinary collaborations, team members may come from diverse backgrounds with varying levels of expertise and different perspectives, which can create communication barriers and difficulties in sharing knowledge (Borgman, 2007). Boundary objects, such as tools, can help overcome these barriers by providing a common ground for collaboration (Carlile, 2002), bridging the gap between different disciplines and facilitating collaboration in various ways. For example, data visualization tools can help researchers from different fields to analyze and interpret complex data sets by providing a common way to represent data visually (Friedman et al., 2008). Similarly, project management software can help interdisciplinary teams organize and manage their work, providing a shared space for collaboration and communication (Wenger, 1998).

I believe a Data Story can function as a boundary object in several ways. Firstly, it can serve as a shared reference point for researchers from different disciplines who may have different understandings and interpretations of the data being analyzed. The narrative structure of a Data Story can provide a common language and understanding of the data, allowing for collaboration and communication between researchers from different backgrounds. Secondly, a Data Story can function as a boundary object between researchers and stakeholders outside of academia. The narrative structure of a Data Story can make complex research findings more accessible and understandable to a broader audience, including policymakers and the general public. This can help to bridge the gap between research and practice and facilitate the translation of research findings into actionable insights. Thirdly, a Data Story can function as a boundary object between different stages of the research process. By incorporating both data curation and narrative elements, a Data Story can provide a bridge between the data collection and analysis stage and the dissemination and sharing stage. It can also help to ensure that data is properly documented and contextualized, which can facilitate the reuse of data by other researchers. Finally, a Data Story can function as a boundary object between top-down policies and bottom-up practices in research data management. Data stories can provide a framework for researchers to organize and contextualize their data, while still allowing for flexibility and creativity in how the data is presented. This can help to bridge the gap between the often-rigid policies and guidelines surrounding research data management and the diverse and complex practices of researchers.

9.5 Supporting 'Anticipatory' Articulation work and other kinds

Articulation work and the need to minimize the effort involved has been a known issue in CSCW and elsewhere since Anselm Strauss introduced the notion (e.g. Strauss 1988). Strauss in that paper argued that: "Projects characteristically have narrative histories: they evolve over time. [...] Although project participants may be relatively unreflective about how they get their work done, we must develop a theoretical framework to understand analytically this organizational process." (p. 163).

With the Data Story, I am providing a technical rather than a theoretical, framework but this quote has a particular resonance here, in part because where people are doing is working around their own research with a prospective collaboration in view. They have in mind that there will be some degree of collaboration in the future. Data story then supports a new kind of articulation work, one which we can call, 'anticipatory' articulation work, (see Steinhardt and Jackson 2015). In fact, by engaging with the Data Story, researchers prepare the groundwork, the mechanics for the collaboration that will take place, not knowing in advance if, how and when a collaboration will happen in the future. Data Story offers an interface that is already designed to provide an essential structure and relevant information for interested parties who might want to access a specific data collection. In doing so, the Data Story provides an entry point into the sensemaking work that will be needed. The focus, then, is on a development from 'anticipation work', i.e., "the practices that cultivate and channel expectations of the future, design pathways into those imaginations, and maintain those visions in the face of a dynamic world" (Steinhardt and Jackson 2015, p. 443). However, if researchers do not engage in curation and sharing practices that allow them or others to understand data after some time, potential collaborations, or reuses of data, will most likely not happen or will be hindered. In this sense, data stories can support the articulation work of researchers by providing them with a tool to curate and share their data in a way that enables others to understand and potentially reuse it in the future.

Moreover, Data Stories can be used to support different kinds of articulation work. For example, they can be used to document the data collection and curation process, allowing others to understand the data and how it was collected. They can also be used to visualize the data, enabling others to make sense of it and identify patterns and trends. In addition, Data Stories can support the articulation work of researchers by providing them with a platform to collaborate with other researchers. By sharing their data and visualizations, researchers can work together to identify patterns and trends, make sense of the data, and generate new insights. This collaboration can be facilitated through tools that enable researchers to comment on and annotate data, share visualizations, and communicate with each other in real-time.

Data stories can also support the articulation work of researchers by providing them with a platform to communicate their findings to different audiences. By creating data stories that are tailored to specific audiences, researchers can communicate their findings in a way that is accessible and understandable to non-experts. This can be done through the use of visualizations, infographics, and other interactive elements that engage the audience and enable them to explore the data in more detail. Finally, Data Stories can support the articulation work of researchers by providing them with a platform to explore and experiment with their data. By creating different visualizations and exploring different data sets, researchers can gain new insights and identify new patterns and trends. This experimentation can be facilitated through tools that enable researchers to manipulate and explore data in real-time, allowing them to quickly test and refine their themes, codes or other analytic constructs.

9.6 Rethinking the overhead: making it worthwhile

As discussed by previous studies (Begley and Ellis 2012; Collaboration 2012; Fecher et al. 2017), RDM inherently involves overhead, which comes with the additional practices of curation and sharing. Inn our research context, however, they are currently not performed at all, or only in haphazard way. Clearly, the absence of specific tools and infrastructure in support of this additional data work - perceived not to be primary in researchers' daily activities and workflows, and clearly not rewarded - motivate researchers' struggles in appropriating these new practices. However, from our observations and exchanges with researchers in the field, we argue that the overhead involved in RDM is unavoidable and should be embraced by those who decide to engage in it. Nevertheless, we also argue that when it comes to designing for RDM the overhead involved needs to be worthwhile and not necessarily effort-less. In fact, when researchers evaluated the Data Story prototype, they saw value in the time invested in sitting together with their data to select the most relevant data, organize it and describe it. Therefore, Data Story will most likely not reduce the time invested in curation, sharing and related overhead. Engaging with it, however, we anticipate will be motivating. With it, we aim at supporting data work processes by producing a structure which will afford an ongoing meaningful workflow in support of several stages of the research process (analyzing, organizing, reflecting, curating, publishing etc.) that researchers, as data creators (and potential data re-users), will and need to benefit from. We believe that more tools and interfaces for RDM should be designed in such a way that allow the data creators to profit from this exercise: better organization, more comprehensive and relevant notes about data, tools for thinking about data not only for future potential reader but for the primary researchers themselves, to deepen their analysis, insights, and interpretations (Pryor 2014; Rolland and Lee 2013). This resonates with Neylon and Wu's (2009) position: "whether they be social networking sites, electronic laboratory notebooks, or controlled vocabularies, (tools) must be built to help scientists do what they are already doing, not what the tool designer feels they should be doing" (ibid., p. 543).

We take on board the injunction of Feger et al. (2020) regarding the transition to effective digital RDM and the role of HCI in it: we, as HCI and CSCW researchers, should facilitate the design of interfaces that can support collaborative data work, learning opportunities, encourage such reflective thinking, and making data work visible, so that it can be better organized, meaningful, and worthy of our time.

The data story, as yet, is only a concept. It cannot be fully implemented without input from researchers, designers and data managers. Nevertheless, in my view, it points the way towards a more relevant and fruitful approach to data use and reuse. As has been pointed out by others, the evolution of successful data management practices is long term and involves the articulation of, and negotiation with, a wide range of heterogeneous interests (Borgman 2010; 2012)

RQ3: In what ways can infrastructuring support the development of new data practices (first and foremost curation and sharing) and eventually lead to data re-use across different disciplines?

9.7 Infrastructuring Research-hub

Infrastructuring, as a socio-technical process, can support the development of new data practices by establishing frameworks, tools, and platforms that enable effective curation, sharing, and re-use of data across different disciplines. By fostering collaboration, improving data management, and addressing the unique challenges associated with diverse data types, infrastructuring can facilitate the development of innovative and interdisciplinary research practices. In this thesis, I presented Research-hub as an example of a socio-technical infrastructure that embodies these principles and demonstrates the potential benefits of infrastructuring in and for the research community.

One of the primary ways in which infrastructuring can support the development of new data practices is by facilitating collaboration among researchers. Platforms like Research-hub provide an accessible and customizable space for communication and cooperation, connecting researchers from different disciplines and backgrounds. By enabling seamless interaction and exchange of ideas, infrastructuring helps create an environment conducive to interdisciplinary research and potentially data re-use. Research-hub, for instance, is built on the open-source platform Humhub, which is designed for team communication and collaboration. Its highly customizable features allow for the integration of various tools and modules, enabling researchers to tailor the platform to their specific needs. This adaptability promotes the development of new practices in data curation and sharing, fostering interdisciplinary research. However, the platform alone will not produce systemic changes. What will be beneficial is to create community rules and internal decisions concerning what to share and with whom while also promoting peer format (like Research Tech Lab, or PhD forum) that support and encourage research interaction and exchange offline and via the platform.

Infrastructuring also supports the development of new data practices by improving data management and accessibility. By integrating existing data storage and sharing solutions, infrastructuring enables researchers to work with familiar tools while leveraging the added benefits of collaboration and metadata management. In the case of Research-hub, the Online Drive module connects the platform to the widely used file-sharing system Sciebo, allowing researchers to synchronize and share files and folders within the Research-hub environment. This integration enhances collaboration by linking shared files to an activity stream where users can visualize, comment, and track important files and activities.

Addressing the challenges associated with diverse data types is another way infrastructuring supports new data practices. For example, qualitative data often require additional context and metadata to be useful and understandable to other researchers. Infrastructuring can help by providing user-friendly interfaces for data annotation and metadata editing, enabling researchers to curate their data more effectively.

Research-hub's Metadata Interface Processing module addresses this challenge by allowing researchers to create and edit metadata for their files and folders within the platform. This capability is essential for effective research collaboration and is not generally supported by conventional file-sharing systems like Sciebo, Sharepoint, Google Drive, or Dropbox. By enabling metadata creation and editing, Research-hub supports better data curation and sharing practices among researchers.

Moreover, Research-hub's Data Story Module offers an innovative approach to presenting qualitative data in a manner that is both useful and accessible to other researchers. This module allows researchers to create a "data-driven" narrative that showcases a selected portion of their collected data, complete with annotations and metadata. By providing a narrative, the Data Story Module enables researchers to engage in data sense-making and to effectively communicate their insights to others, thus addressing some of the main challenges associated with sharing qualitative data.

Lastly, infrastructuring can support data re-use across different disciplines by streamlining the process of archiving and making data accessible for searching. Research-hub's integration with the long-term archive FoDaSi, for example, allows researchers to migrate their curated data from Sciebo to the public domain, where it becomes searchable and accessible to others.

9.8 Infrastructuring Open Science

As mentioned above, the concept of ‘infrastructuring has been valuable in articulating my thoughts about data management and its evolution. I have argued in my papers (Mosconi et al 2019; Mosconi et al. 2023) that, as collaborative work on data increases, so does the need for clarity about the responsibility for its curation. This constitutes an enormous opportunity but also a very significant challenge. Success or otherwise is contingent on a range of factors including the nature of the data to be shared, when it is to be shared, who has rights over it, and the socio-technical infrastructure upon which sharing is to be built. I have made the case for ‘sheer’ curation, an approach which sees curatorial activities going hand in hand with the normal, natural, working lives of those who collect, share and use data. My approach, is strongly influenced, as argued above in a section called ‘conceptual influences’, on the concept of infrastructuring. In my view, the concept aids considerably in understanding how the integration of bottom-up and top-down interests and expectations in the development of e-infrastructures might take place.

An information infrastructure is a relationship between situated practices and the technologies that enable them. An infrastructure “... occurs when local practices are afforded by a larger-scale technology which can then be used in a natural, ready-to hand fashion” (Star and Ruhleder 1996). The concept of ‘infrastructuring’ was developed by Pipek and Wulf (2009). It constitutes a “ framework for designing organizational information systems that focuses on the role of IT as a work infrastructure.” By following an *infrastructuring* approach, I have demonstrated the way in which designing socio-technological affordances as ongoing infrastructure is necessary. Pipek and Wulf developed the notion of ‘points of infrastructure’ and ‘resonance activities’. A point of infrastructure can be thought of as the point where routine and invisible technical and organizational matters become visible, usually when problems arise or innovative possibilities are introduced (Pipek and Wulf, 2009). This may happen, according to them, in different ways:

- *Actual infrastructure breakdown:* The infrastructure is not able to deliver the service it is expected to provide
- *Perceived infrastructure breakdown:* The infrastructure does provide its service technologically, but not to the level expected .
- *Extrinsically motivated practice innovation:* The framing conditions, the task, and goals associated with a practice, have changed in such a way that it is impossible to maintain the old practices.
- *Intrinsically motivated practice innovation:* The framing conditions, tasks and goals associated with a practice remain unchanged, but practitioners discover the potential for performing the practice in a new way.

These four elements describe much of what happens as data infrastructures emerge. The 'breakdown', in this instance, lies in the fact that existing practices do not provide for reuse, perceived as such by the authorities which mandate changes. Goals are imposed which mean that old practice swill have, in time, to change. The fourth condition, however, in this case, is the critical one. Practitioners have not yet, on their own, found new means to perform their data curation tasks and need support. That is what the work in this thesis seeks to provide.

Observation, participation, and design were crucial features of the infrastructuring process I followed in supporting the development of new data practices. Observation involved studying existing practices to understand how they currently work and identify areas for improvement. Participation involved engaging with stakeholders, including researchers, IT service providers, developers, to understand their needs and perspectives and involve them in the design process. Design involved creating new infrastructures that align with the needs and goals of the stakeholders and support the development of new data practices.

By combining these features, the infrastructuring approach helped me to build a socio-technical system - Research-hub - that integrates bottom-up and top-down interests and expectations. Research-hub is an example of infrastructuring as it aims to provide a digital platform to support open science practices and facilitate collaboration among researchers across different disciplines. The platform is designed to be a boundary object, a common ground that can be used by researchers, funders, institutions, and the wider public to share, discover and access research outputs. By providing a common ground for researchers, funders, and institutions to share and access research outputs, Research-hub supports the development of new data practices and can lead to greater data re-use and knowledge production.

10

Conclusion

In conclusion, this work has aimed to contribute to the understanding of data curation and sharing practices and the role of infrastructuring in supporting their development. The concept of infrastructuring has been instrumental in shaping the approach taken, which has emphasized the importance of designing socio-technical processes as ongoing infrastructure. Through observation, participation, and design processes, this work has identified a range of challenges and opportunities in the development of RDM practices for qualitative and ethnographic research context. While progress has been made in advancing the sheer curation approach and supporting bottom-up data practices through Data Stories, significant obstacles remain. These include issues around the ownership and control of data, the need for increased capacity-building, sustain the long-term appropriation, and the importance of developing sustainable funding models to support new tools and infrastructures.

Looking ahead, there are exciting possibilities for the further development of data practices and infrastructures. The ongoing growth of digital technologies and the increasing recognition of the value of data for research and innovation are creating new opportunities for collaboration and knowledge sharing. As the sheer curation and infrastructuring approach continue to evolve, they offer promising avenues for addressing the challenges and unlocking the potential of data for the benefit of society as a whole.

One area for future work is the development of more robust and sustainable funding models for research data management infrastructure. As data becomes increasingly valuable for research and innovation, there is a growing need for institutions and funding agencies to support the development of data infrastructure and tools. However, many researchers and institutions struggle to secure long-term funding for these initiatives, which can limit their ability to develop and maintain effective data management practices. To address this challenge, there is a need for more coordinated efforts between funders, institutions, and researchers to develop sustainable funding models that prioritize the development of data infrastructure and tools.

Another area for future research is the development of more effective approaches to data sharing and collaboration across disciplines. While the use of tools as boundary objects can facilitate interdisciplinary collaboration, there are still challenges around sharing and accessing data across different disciplines. These challenges can be exacerbated by differences in data formats, ethical considerations, and institutional policies. To address these challenges, there is a need for more research on effective approaches to data sharing and collaboration across disciplines, including the development of common standards and practices for data sharing and the development of tools and infrastructure to support interdisciplinary collaboration. Additionally, there is a need for continued research on the social and ethical implications of data curation and sharing practices. As data becomes increasingly valuable, there is a growing need to consider the potential social and ethical implications of data curation and sharing practices, particularly around issues of privacy, confidentiality, and data ownership. There is a need for more research on the ethical and social implications of data curation and sharing practices, including the development of guidelines and best practices for ethical data management.

Finally, there is a need for more research on the role of infrastructuring in supporting the development of new data practices and facilitating the reuse of data across different disciplines. Infrastructuring has emerged as a key concept in the development of effective data management practices, and there is a need for more research on its role in facilitating collaboration and knowledge sharing across different disciplines. This research could include case studies of successful infrastructuring initiatives, as well as more theoretical work on the role of infrastructure in shaping social and technical systems.

In summary, while significant progress has been made in the development of data curation and sharing practices, there are still significant challenges and opportunities for future research. Through continued research and development of effective infrastructures and tools, there is potential to unlock the full potential of data for the benefit of society as a whole.

References

Abbott, Daisy. 2008. "What Is Digital Curation? DCC Briefing Papers: Introduction to Curation." Edinburgh.

Antes, Alison L., Heidi A. Walsh, Michelle Strait, Cynthia R. Hudson-Vitale, and James M. DuBois. 2018. "Examining Data Repository Guidelines for Qualitative Data Sharing." *Journal of Empirical Research on Human Research Ethics* 13 (1): 61–73. https://doi.org/10.1177/1556264617744121.

Arzberger, P, P Schroeder, A Beaulieu, G Bowker, K Casey, L Laaksonen, D Moorman, P Uhlir, and P Wouters. 2004. "Promoting Access to Public Research Data for Scientific, Economic, and Social Development." *Data Science Journal* 3: 135–52. https://doi.org/10.2481/dsj.3.135.

Barrett, Helen. 2006. "Researching and Evaluating Digital Storytelling as a Deep Learning Tool." Edited by C. Crawford, R. Carlsen, K. McFerrin, J. Price, R. Weber, and D. Willis. *Society for Information Technology & Teacher Education International Conference* 2006 (1): 647–54. http://www.editlib.org/p/22293/.

Barry, Danika, Leighann E Kimble, Bejoy Nambiar, Gareth Parry, Ashish Jha, Vijay Kumar Chattu, M Rashad Massoud, and Don Goldmann. 2018. "A Framework for Learning about Improvement: Embedded Implementation and Evaluation Design to Optimize Learning." *International Journal for Quality in Health Care* 30 (suppl_1): 10–14.

Bartling, Sönke, and Sascha Friesike. 2014. *Sönke Bartling & Sascha Friesike*. Edited by Sönke Bartling and Sascha Friesike. *Opening Science: The Evolving Guide on How the Internet Is Changing Research, Collaboration and Scholarly Publishing*. London: Springer Nature.

Bechhofer, Sean, David De Roure, Matthew Gamble, Carole Goble, and Iain Buchan. 2010. "Research Objects: Towards Exchange and Reuse of Digital Knowledge." *Nature Precedings*. https://doi.org/10.1038/npre.2010.4626.1.

Begley, C Glenn, and Lee M Ellis. 2012. "Raise Standards for Preclinical Cancer Research." *Nature* 483 (7391): 531–33.

Berg, Harry van den. 2008. "Reanalyzing Qualitative Interviews from Different Angles: The Risk of Decontextualization and Other Problems of Sharing Qualitative Data." *Historical Social Research* 6 (1): 179–92. https://doi.org/10.12759/HSR.33.2008.3.179-192.

Bietz, Matthew J, Eric P S Baumer, and Charlotte P Lee. 2010. "Synergizing in Cyberinfrastructure Development." *Computer Supported Cooperative Work (CSCW)* 19 (3): 245–81. https://doi.org/10.1007/s10606-010-9114-y.

Bietz, Matthew J, and Charlotte P Lee. 2009. "Collaboration in Metagenomics: Sequence Databases and the Organization of Scientific Work." In *ECSCW 2009*, edited by Ina Wagner, Hilda Tellioğlu, Ellen Balka, Carla Simone, and Luigina Ciolfi, 243–62. London: Springer London.

Bietz, Matthew J., Andrea Wiggins, Mark Handel, and Cecilia Aragon. 2012. "Data-Intensive Collaboration in Science and Engineering." Edited by Steven Poltrock, Carla Simone, Jonathan Grudin, Gloria Mark, and John Riedl. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion - CSCW '12*. New York, New York, USA: ACM Press. https://doi.org/10.1145/2141512.2141515.

Birkbeck, Gail, Tadhg Nagle, and David Sammon. 2022. "Challenges in Research Data Management Practices: A Literature Analysis." *Journal of Decision Systems* 31 (sup1): 153–67. https://doi.org/10.1080/12460125.2022.2074653.

Birnholtz, Jeremy P., and Matthew J. Bietz. 2003a. “Data at Work: Supporting Sharing in Science and Engineering.” *In Proceedings of the SIGGROUP Conference on Supporting Group Work (GROUP’03)*, 339–348. https://doi.org/10.1145/958160.958215.

———. 2003b. “Data at Work: Supporting Sharing in Science and Engineering.” *In Proceedings of the SIGGROUP Conference on Supporting Group Work (GROUP’03)*, 339–348. https://doi.org/10.1145/958160.958215.

Bishop, Libby. 2009. “Ethical Sharing and Reuse of Qualitative Data.” *Australian Journal of Social Issues* 44 (3): 255–72. https://doi.org/10.1002/J.1839-4655.2009.TB00145.X.

———. 2012. “Using Archived Qualitative Data for Teaching: Practical and Ethical Considerations.” *International Journal of Social Research Methodology* 15 (4): 341–50. https://doi.org/10.1080/13645579.2012.688335.

———. 2014. “Re-Using Qualitative Data: A Little Evidence, On-Going Issues and Modest Reflections.” *Studia Socjologiczne* 3 (214): 167–76. http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-657a6a2d-6222-4613-a6f3-85e17b08f124.

Björgvinsson, Erling;, Pelle; Ehn, and Per-Anders; Hillgren. 2010. “Participatory Design and ‘Democratizing Innovation.’” In *Proceedings of the 11th Biennial Participatory Design Conference (PDC ’10)*, edited by ACM, 41–50. New York.

Blomberg, Jeanette, and Helena Karasti. 2013. “Reflections on 25 Years of Ethnography in CSCW.” *Computer Supported Cooperative Work* 22 (4–6): 373–423. https://doi.org/10.1007/S10606-012-9183-1.

Blumer, Herbert. 1940. “The Problem of the Concept in Social Psychology.” *American Journal of Sociology* 45 (5): 707-719.

———. 1954. “What Is Wrong with Social Theory?” *American Sociological Review* 19 (1): 3–10. https://doi.org/10.2307/2088165.

Boas, Franz. 1914. “Mythology and Folk-Tales of the North American Indians.” *The Journal of American Folklore* 27 (106): 374–410. https://doi.org/10.2307/534740.

Borgman, C. L. 2012. “The Conundrum of Sharing Research Data.” *Journal of the American Society for Information Science and Technology* 63 (6): 1059–1078.

Borgman, Christine L. 2010. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. MIT press.

Borgman, Christine L. 2012. “The Conundrum of Sharing Research Data.” *Journal of the American Society for Information Science and Technology* 63 (6): 1059–78. https://doi.org/10.1002/ASI.22634.

———. 2015. *Big Data, Little Data, No Data. Scholarship in the Networked World*. Cambridge, Massachusetts: The MIT Press.

Borgman, Christine L., Andrea Scharnhorst, and Milena S. Golshan. 2019a. “Digital Data Archives as Knowledge Infrastructures: Mediating Data Sharing and Reuse.” *Journal of the Association for Information Science and Technology* 70 (8). https://doi.org/10.1002/asi.24172.

———. 2019b. “Digital Data Archives as Knowledge Infrastructures: Mediating Data Sharing and Reuse.” *Journal of the Association for Information Science and Technology* 70 (8): 888–904. https://doi.org/10.1002/ASI.24172.

Bowker, Geoffrey C. 2005. *Memory Practices in the Sciences*. Cambridge, MA : MIT Press .

Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things out : Classification and Its Consequences*. London: MIT Press.

———. 2000. *Sorting Things out: Classification and Its Consequences*. MIT press.

Boyd, Evelyn M, and Ann W Fales. 1983. “Reflective Learning: Key to Learning from Experience.” *Journal of Humanistic Psychology* 23 (2): 99–117.

Broom, Alex, Lynda Cheshire, and Michael Emmison. 2009. "Qualitative Researchers' Understandings of Their Practice and the Implications for Data Archiving and Sharing." *Sociology* 43 (6): 1163–80. https://doi.org/10.1177/0038038509345704.

Burge, Janet E., John M. Carroll, Raymond McCall, and Ivan Mistrik. 2008. *Rationale-Based Software Engineering. Rationale-Based Software Engineering*. Springer Science & Business Media. https://doi.org/10.1007/978-3-540-77583-6.

Burgelman, Jean-Claude, Corina Pascu, Katarzyna Szkuta, Rene von Schomberg, Athanasios Karalopoulos, Konstantinos Repanas, and Michel Schouppe. 2019. "Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century." *Frontiers in Big Data* 2. https://www.frontiersin.org/articles/10.3389/fdata.2019.00043.

Burkhardt, Marcus, Daniela van Geenen, Carolin Gerlitz, Sam Hind, Timo Kaerlein, Danny Lämmerhirt, and Axel Volmar, eds. 2022. *Interrogating Datafication: Towards a Praxeology of Data*. transcript Verlag.

Carlson, Samuelle, and Ben Anderson. 2007. "What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use." *Journal of Computer-Mediated Communication* 12 (2): 635–51. https://doi.org/10.1111/J.1083-6101.2007.00342.X.

Carstensen, Peter H., Kari Schmidt, and Stefan Spanner. 2010. "Challenges in Articulation Work: A Study of Coordinative Practices in Distributed Software Projects." *Computer Supported Cooperative Work* 19 (3–4): 285–315.

Caton, Hiram. 1990. *The Samoa Reader. Anthropologists Take Stock.* . Lanham, Maryland: University Press of America.

Chambers, Fred. 1994. "Removing Confusion about Formative and Summative Evaluation: Purpose versus Time." *Evaluation and Program Planning* 17 (1): 9–12.

Chawinga, Winner Dominic, and Sandy Zinn. 2020. "Research Data Management at a Public University in Malawi: The Role of 'Three Hands.'" *Library Management*.

Chin, George, and Carina S Lansing. 2004. "Capturing and Supporting Contexts for Scientific Data Sharing via the Biological Sciences Collaboratory." In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, 409–18. CSCW '04. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1031607.1031677.

Choi, Joohee, and Yla Tausczik. 2017. "Characteristics of Collaboration in the Emerging Practice of Open Data Analysis." In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 835–46. CSCW '17. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2998181.2998265.

Collaboration, Open Science. 2012. "An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science." *Perspectives on Psychological Science* 7 (6): 657–60.

Collins, Sandra, Francoise Genova, Natalie Harrower, Simon Hodson, Sarah Jones, Leif Laaksonen, Daniel Mietchen, Rūta Petrauskaitė, and Peter Wittenburg. 2018. "Turning FAIR into Reality: Final Report and Action Plan from the European Commission Expert Group on FAIR Data."

Coltart, Carrie, Karen Henwood, and Fiona Shirani. 2013. "Qualitative Secondary Analysis in Austere Times: Ethical, Professional and Methodological Considerations." *Historical Social Research / Historische Sozialforschung* 38 (4 (146)): 271–92. http://www.jstor.org/stable/24142699.

Commission, European. 2016. *Open Innovation, Open Science, Open to the World : A Vision for Europe*. Publications Office. https://doi.org/doi/10.2777/061652.

Concannon, Shauna, Natasha Rajan, Parthiv Shah, Davy Smith, Marian Ursu, and Jonathan Hook. 2020. “Brooke Leave Home: Designing a Personalized Film to Support Public Engagement with Open Data.” *Conference on Human Factors in Computing Systems - Proceedings*, 1–14. https://doi.org/10.1145/3313831.3376462.

Corti, Louise. 2007. “Re-Using Archived Qualitative Data – Where, How, Why?” *Archival Science* 7 (1): 37–54. https://doi.org/10.1007/s10502-006-9038-y.

———. 2013. “Infrastructures for Qualitative Data Archiving.” In *Forschungsinfrastrukturen Für Die Qualitative Sozialforschung*, edited by Denis Huschka, Hubert Knoblauch, Claudia Oellers, and Heike Solga, 35–62. Scivero.

Crabtree, Andrew, Mark Rouncefield, and Peter Tolmie. 2012. *Doing Design Ethnography* . Springer Science & Business Media .

Creswell, John W., and Cheryl N. Poth. 2016. *Qualitative Inquiry and Research Design: Choosing among Five Approaches*. Sage publications.

Curdt, Constanze, and Dirk Hoffmeister. 2015. “Research Data Management Services for a Multidisciplinary, Collaborative Research Project: Design and Implementation of the TR32DB Project Database.” *Program: Electronic Library and Information Systems*.

Curty, Renata, Ayoung Yoon, Wei Jeng, and Jian Qin. 2016. “Untangling Data Sharing and Reuse in Social Sciences; Untangling Data Sharing and Reuse in Social Sciences.” *Proceedings of the Association for Information Science and Technology*. https://doi.org/10.1002/pra2.2016.14505301025.

Dachtera, Juri, Dave Randall, and Volker Wulf. 2014. “Research on Research: Design Research at the Margins: Academia, Industry and End-Users.” *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 713–22. https://doi.org/10.1145/2556288.2557261.

Dallas, Costis. 2007. “An Agency-Oriented Approach to Digital Curation Theory and Practice .” In *ICHIM’07. Proceedings of the International Cultural Heritage Informatics Meeting*, edited by J Trant and D Bearman. Toronto: Archives & Museum Informatics .

———. 2016. “Digital Curation beyond the ‘Wild Frontier’: A Pragmatic Approach.” *Archival Science* 16 (4): 421–57. https://doi.org/10.1007/s10502-015-9252-6.

Dalton, Craig M, Linnet Taylor, and Jim Thatcher (alphabetical). 2016. “Critical Data Studies: A Dialog on Data and Space:” *Big Data & Society* 3 (1). https://doi.org/10.1177/2053951716648346.

Dalton, Craig M., and Jim Thatcher. 2014. “What Does a Critical Data Studies Look like, and Why Do We Care? Seven Points for a Critical Approach to ‘Big Data.’” *Society and Space* 29. https://www.societyandspace.org/articles/what-does-a-critical-data-studies-look-like-and-why-do-we-care.

Davies, Charlotte Aull. 2008. *Reflexive Ethnography : A Guide to Researching Selves and Others*. Routledge.

Demian, Peter, and Renate Fruchter. 2009. “Effective Visualisation of Design Versions: Visual Storytelling for Design Reuse.” *Research in Engineering Design* 19 (4): 193–204. https://doi.org/10.1007/S00163-008-0051-4.

Denning, Stephen. 2006. “Effective Storytelling: Strategic Business Narrative Techniques.” *Strategy and Leadership* 34 (1): 42–48. https://doi.org/10.1108/10878570610637885.

D’Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. Mit Press.

Donner, Eva Katharina. 2022. “Research Data Management Systems and the Organization of Universities and Research Institutes: A Systematic Literature Review.” *Journal of Librarianship and Information Science*, 09610006211070282.

Dourish, Paul. 1999. “Software Infrastructures.” *Computer Supported Cooperative Work, John Wiley & Sons*, 195–219.

Dourish, Paul, and Victoria Bellotti. 1992. “Awareness and Coordination in Shared Workspaces.” *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work*, 107-114.

Dourish, Paul, and Edgar Gómez Cruz. 2018. “Datafication and Data Fiction: Narrating Data and Narrating with Data:” *Big Data & Society* 5 (2). https://doi.org/10.1177/2053951718784083.

Drucker, Johanna. 2011. “Humanities Approaches to Graphical Display.” *Digital Humanities Quarterly* 5 (1): 1–21.

Duarte, Nancy. 2019. *Data Story : Explain Data and Inspire Action through Story*. Ideapress Publishing.

Dykes, Brent. 2015. “Data Storytelling: What It Is and How It Can Be Used to Effectively Communicate Analysis Results.” *Applied Marketing Analytics* 1 (4): 299–313.

Eberhard, Igor, and Wolfgang Kraus. 2018. “Der Elefant Im Raum. Ethnographisches Forschungsdatenmanagement Als Herausforderung Für Repositorien.” *Mitteilungen Der Vereinigung Österreichischer Bibliothekarinnen Und Bibliothekare* 71 (1): 41–52.

EC - European Commission. 2016. “H2020 Programme Guidelines on FAIR Data Management in Horizon 2020.”

EC - European Commission, and Neelie Kroes. 2012. “Recommendation on Access to and Preservation of Scientific Information in Europa.” *Official Journal of the European Union*, 1–125. https://doi.org/http://dx.doi.org/10.4403/jlis.it-8649.

Edwards, P N, S J Jackson, M K Chalmers, G C Bowker, C L Borgman, D Ribes, and S Calvert. 2013. “Knowledge Infrastructures: Intellectual Frameworks and Research Challenges. University of Michigan.” *Ann Arbor. Retrieved from Http://Deepblue. Lib. Umich. Edu/Handle/2027.42/9755* 2.

Edwards, Paul N, Matthew S Mayernik, Archer L Batcheller, Geoffrey C Bowker, and Christine L Borgman. 2011. “Science Friction: Data, Metadata, and Collaboration.” *Social Studies of Science* 41 (5): 667–90. https://doi.org/10.1177/0306312711413314.

Erickson, I, K Eschenfelder, S Goggins, L Hemphill, S Sawyer, K Shankar, and K Shilton. 2014. “The Ethos and Pragmatics of Data Sharing.” *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 109–12. https://doi.org/10.1145/2556420.2556852.

Es, Karin Van, and Mirko T Schäfer. 2017. *The Datafied Society. Studying Culture through Data.* Amsterdam University Press.

Eschenfelder, Kristin, and Andrew Johnson. 2011. “The Limits of Sharing: Controlled Data Collections.” *Proceedings of the ASIST Annual Meeting* 48: 1–10. https://doi.org/10.1002/meet.2011.14504801062.

Faniel, Ixchel M., and Trond E. Jacobsen. 2010a. “Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues’ Data.” *Computer Supported Cooperative Work* 19 (3–4): 355–75. https://doi.org/10.1007/s10606-010-9117-8.

———. 2010b. “Reusing Scientific Data.” In *Computer Supported Cooperative Work*, 19:355–75. Kluwer Academic Publishers PUB879 Norwell, MA, USA. https://doi.org/10.1007/S10606-010-9117-8.

Fecher, Benedikt, and Sascha Friesike. 2014. “Open Science: One Term, Five Schools of Thought.” In *Opening Science: The Evolving Guide on How the Internet Is Changing Research, Collaboration and Scholarly Publishing*, edited by Sönke Bartling and Sascha Friesike, 17–47. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_2.

Fecher, Benedikt, Sascha Friesike, and Marcel Hebing. 2015. “What Drives Academic Data Sharing? .” *PloS One* 10 (2).

Fecher, Benedikt, Sascha Friesike, Marcel Hebing, and Stephanie Linek. 2017a. “A Reputation Economy: How Individual Reward Considerations Trump Systemic Arguments for Open Access to Data.” *Palgrave Communications* 3 (1): 1–10.

———. 2017b. “A Reputation Economy: How Individual Reward Considerations Trump Systemic Arguments for Open Access to Data.” *Palgrave Communications* 3 (1): 1–10.

Feger, Sebastian S, Sünje Dallmeier-Tiessen, Albrecht Schmidt, and Paweł W Woźniak. 2019. “Designing for Reproducibility: A Qualitative Study of Challenges and Opportunities in High Energy Physics.” In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Feger, Sebastian S., Paweł W. Wozniak, Lars Lischke, and Albrecht Schmidt. 2020a. “‘Yes, I Comply!’ Motivations and Practices around Research Data Management and Reuse across Scientific Fields.” In *Proceedings of the ACM on Human-Computer Interaction*, 4:1–26. New York, NY, USA. https://doi.org/10.1145/3415212.

———. 2020b. “‘Yes, I Comply!’ Motivations and Practices around Research Data Management and Reuse across Scientific Fields.” In *Proceedings of the ACM on Human-Computer Interaction*, 4:1–26. New York, NY, USA. https://doi.org/10.1145/3415212.

Feger, Sebastian Stefan, Paweł W Woźniak, Jasmin Niess, and Albrecht Schmidt. 2021. “Tailored Science Badges: Enabling New Forms of Research Interaction.” In *Designing Interactive Systems Conference 2021*, 576–88.

Fekete, Jean Daniel. 2004. “The InfoVis Toolkit.” *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*, 167–74. https://doi.org/10.1109/INFVIS.2004.64.

Fekete, Jean Daniel, Jarke J. Van Wijk, John T. Stasko, and Chris North. 2008. “The Value of Information Visualization.” *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4950 LNCS: 1–18. https://doi.org/10.1007/978-3-540-70956-5_1.

Feldman, Shelley, and Linda Shaw. 2019. “The Epistemological and Ethical Challenges of Archiving and Sharing Qualitative Data.” *American Behavioral Scientist* 63 (6): 699–721.

Fenlon, Katrina. 2019. “Modeling Digital Humanities Collections as Research Objects.” In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 138–47.

Fielding, Nigel G., and Jane L. Fielding. 2000. “Resistance and Adaptation to Criminal Identity: Using Secondary Analysis to Evaluate Classic Studies of Crime and Deviance.” *Sociology* 34 (4): 671–89. https://doi.org/10.1177/S0038038500000419.

Fortun, Mike, Lindsay Poirier, Alli Morgan, Brian Callahan, and Kim Fortun. 2021. “What’s so Funny ‘bout PECE, TAF, and Data Sharing.” Collaborative Anthropology Today: A Collection of Exceptions. Ithaca, NY ….

Game, Ann, and Andrew Metcalfe. 1996. *Passionate Sociology*. Sage.

Garfinkel, Harold. 1967. *Studies in Ethnomethodology*. Englewood Cliffs: Prentice-Hall.

Garza, Kristian, Carole Goble, John Brooke, and Caroline Jay. 2015. “Framing the Community Data System Interface.” In *Proceedings of the 2015 British HCI Conference*, 269–70.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92.

Gerson, Elihu M, and Susan Leigh Star. 1986. “Analyzing Due Process in the Workplace.” *ACM Trans. Inf. Syst.* 4 (3): 257–70. https://doi.org/10.1145/214427.214431.

Gitelman, Lisa, ed. 2013a. *Raw Data Is an Oxymoron*. MIT press.

———. 2013b. *“Raw Data” Is an Oxymoron. Infrastructures Series.* Cambridge, MA: MIT Press.

Grudin, Jonathan. 1988. “Why CSCW Applications Fail: Problems in the Design and Evaluationof Organizational Interfaces.” In *Proceedings of the 1988 ACM Conference on Computer-Supported Cooperative Work*, 85–93.

Gupta, Shivam, and Claudia Müller-Birn. 2018. “A Study of E-Research and Its Relation with Research Data Life Cycle: A Literature Perspective.” *Benchmarking* 25 (6): 1656–80. https://doi.org/10.1108/BIJ-02-2017-0030.

Gurstein, Michael. 2007. *What Is Community Informatics (and Why Does It Matter)?* . Vol. Vol. 2. Polimetrica sas.

Haak, Maaike Van Den, Menno De Jong, and Peter Jan Schellens. 2003. “Retrospective vs. Concurrent Think-Aloud Protocols: Testing the Usability of an Online Library Catalogue.” *Behaviour & Information Technology* 22 (5): 339–51.

Hamad, Faten, Maha Al-Fadel, and Aman Al-Soub. 2021. “Awareness of Research Data Management Services at Academic Libraries in Jordan: Roles, Responsibilities and Challenges.” *New Review of Academic Librarianship* 27 (1): 76–96.

Hara, Noriko, Paul Solomon, Seung-Lye Kim, and Diane H Sonnenwald. 2003. “An Emerging View of Scientific Collaboration: Scientists’ Perspectives on Collaboration and Factors That Impact Collaboration.” *Journal of the American Society for Information Science and Technology* 54 (10): 952–65. https://doi.org/https://doi.org/10.1002/asi.10291.

Haraway, Donna. 1991. *Simians, Cyborgs and Women: The Reinvention of Nature*. New York: Routledge.

Hayes, Gillian R. 2011. “The Relationship of Action Research to Human-Computer Interaction.” *ACM Transactions on Computer-Human Interaction (ToCHI)* 18 (3): 1–20. https://doi.org/10.1145/1993060.1993065.

Hearn, Gregory N., Jo A. Tacchi, Marcus Foth, and June Lennie. 2008. *Action Research and New Media: Concepts, Methods and Cases*. Hampton Press.

Heaton, Janet. 2008. “Secondary Analysis of Qualitative Data: An Overview.” *Historical Social Research* 33 (3): 33–45. https://doi.org/10.12759/HSR.33.2008.3.33-45.

Hedges, Mark, Tobias Blanke, Stella Fabiane, Gareth Knight, and Eric Liao. 2012. “Sheer Curation of Experiments: Data, Process, Provenance.” *Journal of Digital Information* 13 (1).

Hedstrom, Margaret. 1997. “Building Record-Keeping Systems: Archivists Are Not Alone on the Wild Frontier.” *Archivaria* 44: 44–71.

Hendy, Jane, Naomi Fulop, Scott Reeves, Ann Hutchings, and Sally Collin. 2009. “Implementing Change: What Can We Learn from the NHS’s Implementation of Electronic Booking Systems?” *Health Services Management Research* 22 (3): 108–16.

Hey, Tony, Stewart Tansley, and Kristin Tolle. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research. https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

Hughes, Everett C. 1958. *Men and Their Work*. Glencoe, Ill: Free Press.

Iliadis, Andrew, and Federica Russo. 2016. “Critical Data Studies: An Introduction:” *Big Data & Society* 3 (2). https://doi.org/10.1177/2053951716674238.

Irani, Lilly. 2010. “HCI on the Move: Methods, Culture, Values.” In *CHI’10 Extended Abstracts on Human Factors in Computing Systems*, 2939–42.

Irwin, Sarah. 2013. “Qualitative Secondary Data Analysis: Ethics, Epistemology and Context:” *Progress in Development Studies* 13 (4): 295–306. https://doi.org/10.1177/1464993413490479.

Jacobs, James A., and Charles Humphrey. 2004. “Preserving Research Data.” *Communications of the ACM* 47 (9): 27–29.

Jahnke, Lori M., and Andrew Asher. 2012. “The Problem of Data: Data Management and Curation Practices among University Researchers.”

Jenness, Valerie. 2008. "Pluto, Prisons, and Plaintiffs: Notes on Systematic Back-Translation From an Embedded Researcher." *Social Problems* 55 (1): 1–22. https://doi.org/10.1525/sp.2008.55.1.1.

Jirotka, Marina, Charlotte P Lee, and Gary M Olson. 2013. "Supporting Scientific Collaboration: Methods, Tools and Concepts." *Computer Supported Cooperative Work (CSCW)* 22 (4): 667–715. https://doi.org/10.1007/s10606-012-9184-0.

Kaltenbrunner, Wolfgang. 2017. "Digital Infrastructure for the Humanities in Europe and the US: Governing Scholarship through Coordinated Tool Development." *Computer Supported Cooperative Work (CSCW)* 26 (3): 275–308.

Karasti, H, and K S Baker. 2004. "Infrastructuring for the Long-Term: Ecological Information Management." In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of The*, 10 pp.-. https://doi.org/10.1109/HICSS.2004.1265077.

Karasti, Helena. 2014. "Infrastructuring in Participatory Design." In *13th Participatory Design Conference (PDC '14)*, edited by ACM, 141–150. New York: ACM.

Karasti, Helena, Karen S. Baker, and Geoffrey C. Bowker. 2002. "Ecological Storytelling and Collaborative Scientific Activities." *ACM SIGGROUP Bulletin* 23 (2): 29–30. https://doi.org/10.1145/962185.962197.

Karasti, Helena, Karen S. Baker, and Eija Halkola. 2006a. "Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network." *Computer Supported Cooperative Work* 15 (4): 321–58. https://doi.org/10.1007/s10606-006-9023-2.

———. 2006b. "Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network." *Computer Supported Cooperative Work (CSCW) 2006 15:4* 15 (4): 321–58. https://doi.org/10.1007/S10606-006-9023-2.

Karasti, Helena, and Anna-Liisa Syrjänen. 2004. "Artful Infrastructuring in Two Cases of Community PD." In *Proceedings of the Eighth Conference on Participatory Design: Artful Integration: Interweaving Media, Materials and Practices - Volume 1*, 20–30. PDC 04. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1011870.1011874.

Kaye, Joseph'Jofish'. 2007. "Evaluating Experience-Focused HCI." In *CHI'07 Extended Abstracts on Human Factors in Computing Systems*, 1661–64.

Kervin, Karina, Robert B Cook, and William K Michener. 2014a. "The Backstage Work of Data Sharing." In *Proceedings of the 18th International Conference on Supporting Group Work*, 152–56. GROUP '14. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2660398.2660406.

Kervin, Karina, Robert B. Cook, and William K. Michener. 2014b. "The Backstage Work of Data Sharing." Edited by Sean Goggins, Isa Jahnke, David W McDonald, and Pernille Bjørn. *Proceedings of the 18th International Conference on Supporting Group Work - GROUP '14*. New York, New York, USA: ACM Press. https://doi.org/10.1145/2660398.2660406.

———. 2014c. "The Backstage Work of Data Sharing." Edited by Sean Goggins, Isa Jahnke, David W McDonald, and Pernille Bjørn. *Proceedings of the 18th International Conference on Supporting Group Work - GROUP '14*. New York, New York, USA: ACM Press. https://doi.org/10.1145/2660398.2660406.

Kervin, Karina E., Robert B. Cook, and William K. Michener. 2014. "The Backstage Work of Data Sharing." In *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work*, 152–56. Association for Computing Machinery. https://doi.org/10.1145/2660398.2660406.

Khan, Nushrat, Mike Thelwall, and Kayvan Kousha. 2021. “Are Data Repositories Fettered? A Survey of Current Practices, Challenges and Future Technologies.” *Online Information Review*.

Kitchin, Rob. 2014. *The Data Revolution. Big Data, Open Data, Data Infrastructures & Their Consequences*. London: SAGE.

———. 2021. *Data Lives: How Data Are Made and Shape Our World*. Policy Press.

Kitchin, Rob, and Tracey P Lauriault. 2014. “Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work.” In *Thinking Big Data in Geography: New Regimes, New Research*, edited by Jim Thatcher, Andrew Shears, and Josef Eckert. University of Nebraska Press. http://ssrn.com/abstract=2474112http://www.nuim.ie/progcity/.

Knaflic, Cole N. 2015. *Storytelling with Data: A Data Visualization Guide for Business Professionals*. John Wiley & Sons.

Knaflic, Cole Nussbaumer. 2015. *Storytelling with Data: A Data Visualization Guide for Business Professionals*. John Wiley & Sons.

Koesten, Laura, Emilia Kacprzak, Jeni Tennison, and Elena Simperl. 2019. “Collaborative Practices with Structured Data: Do Tools Support What Users Need?” In *CHI Conference on Human Factors in Computing Systems Proceedings*. Glasgow, Scotland Uk.

Koesten, Laura, and Elena Simperl. 2021a. “UX of Data: Making Data Available Doesn’t Make It Usable.” In *Interactions*, 97–99.

———. 2021b. “UX of Data: Making Data Available Doesn’t Make It Usable.” In *Interactions*, 97–99.

Koltay, Tibor. 2016. “Data Governance, Data Literacy and the Management of Data Quality.” *IFLA Journal* 42 (4): 303–12. https://doi.org/10.1177/0340035216672238.

Korn, Matthias, Marén Schorch, Volkmar Pipek, Matthew Bietz, Carsten Østerlund, Rob Procter, David Ribes, and Robin Williams. 2017. “E-Infrastructures for Research Collaboration: The Case of the Social Sciences and Humanities.” In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 415–20.

Koschmann, Timothy. 1996. “Paradigm Shifts and Instructional Technology: An Introduction.” In *CSCL: Theory and Practice of an Emerging Paradigm*, edited by Timothy Koschmann, 116:1–23.

la Bellacasa, Maria Puig de. 2010. “Matters of Care in Technoscience: Assembling Neglected Things.” *Social Studies of Science* 41 (1): 85–106. https://doi.org/10.1177/0306312710380301.

la Bellacasa, María Puig de. 2012. “‘Nothing Comes Without Its World’: Thinking with Care.” *The Sociological Review* 60 (2): 197–216. https://doi.org/10.1111/j.1467-954X.2012.02070.x.

Lallé, Sébastien, and Cristina Conati. 2019. “The Role of User Differences in Customization: A Case Study in Personalization for Infovis-Based Content.” *International Conference on Intelligent User Interfaces, Proceedings IUI* Part F1476: 329–39. https://doi.org/10.1145/3301275.3302283.

Landi, Annalisa, Mark Thompson, Viviana Giannuzzi, Fedele Bonifazi, Ignasi Labastida, Luiz Olavo Bonino da Silva Santos, and Marco Roos. 2020. “The ‘A’ of FAIR – As Open as Possible, as Closed as Necessary.” *Data Intelligence* 2 (1–2): 47–55. https://doi.org/10.1162/dint_a_00027.

Ledo, David, Steven Houben, Jo Vermeulen, Nicolai Marquardt, Lora Oehlberg, and Saul Greenberg. 2018. “Evaluation Strategies for HCI Toolkit Research.” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–17.

Lee, Bongshin, Nathalie Henry Riche, Petra Isenberg, and Sheelagh Carpendale. 2015. “More than Telling a Story: Transforming Data into Visually Shared Stories.” *IEEE Computer Graphics and Applications* 35 (5): 84–90.

Lee, Charlotte P, Paul Dourish, and Gloria Mark. 2006. “The Human Infrastructure of Cyberinfrastructure.” In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, 483–92. CSCW ’06. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1180875.1180950.

Lee, Jae W, Jianting Zhang, Ann S Zimmerman, and Angelo Lucia. 2009. “DataNet: An Emerging Cyberinfrastructure for Sharing, Reusing and Preserving Digital Data for Scientific Discovery and Learning.”

Lee, Jintae. 1997. “Design Rationale Systems: Understanding the Issues.” *IEEE Expert-Intelligent Systems and Their Applications* 12 (3): 78–85. https://doi.org/10.1109/64.592267.

Leonelli, Sabina. 2022. “Open Science and Epistemic Diversity: Friends or Foes?” *Philosophy of Science*, 1–21. https://doi.org/DOI: 10.1017/psa.2022.45.

Lewis, Susan J, and Andrew J Russell. 2011a. “Being Embedded: A Way Forward for Ethnographic Research.” *Ethnography* 12 (3): 398–416.

———. 2011b. “Being Embedded: A Way Forward for Ethnographic Research.” *Ethnography* 12 (3): 398–416.

Linde, Charlotte. 2001. “Narrative and Social Tacit Knowledge.” *Journal of Knowledge Management* 5 (2): 160–71.

Linne, Monika, and Wolfgang Zenk-Möltgen. 2017. “Strengthening Institutional Data Management and Promoting Data Sharing in the Social and Economic Sciences.” *Liber Quarterly* 27 (1).

Liu, Shixia, Weiwei Cui, Yingcai Wu, and Mengchen Liu. 2014. “A Survey on Information Visualization: Recent Advances and Challenges.” *Visual Computer* 30 (12): 1373–93. https://doi.org/10.1007/s00371-013-0892-3.

Ludwig, Thomas, Volkmar Pipek, and Peter Tolmie. 2018. “Designing for Collaborative Infrastructuring: Supporting Resonance Activities.” *Proc. ACM Hum.-Comput. Interact.* 2 (CSCW). https://doi.org/10.1145/3274382.

MacDonald, Craig M, and Michael E Atwood. 2013. “Changing Perspectives on Evaluation in HCI: Past, Present, and Future.” In *CHI’13 Extended Abstracts on Human Factors in Computing Systems*, 1969–78.

Mackay, Wendy E., Caroline Appert, Michel Beaudouin-Lafon, Olivier Chapuis, Du Yangzhou, Jean-Daniel Fekete, and Guiard. Yves. 2007. “Touchstone: Exploratory Design of Experiments.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1425–34.

Malinowski, Bronislaw. 1922. “Ethnology and the Study of Society.” *Economica*, no. 6: 208–19. https://doi.org/10.2307/2548314.

———. 1929. “Practical Anthropology.” *Africa* 2 (1): 22–38. https://doi.org/DOI: 10.2307/1155162.

Mannheim, Karl. 1936. *Ideology and Utopia*. London: Routledge.

Mannheimer, Sara, Amy Pienta, Dessislava Kirilova, Colin Elman, and Amber Wutich. 2018. “Qualitative Data Sharing: Data Repositories and Academic Libraries as Key Partners in Addressing Challenges.” *American Behavioral Scientist* 63 (5): 643–64. https://doi.org/10.1177/0002764218784991.

Marcus, George E. 1994. “On Ideologies of Reflexivity in Contemporary Efforts to Remake the Human Sciences.” *Poetics Today* 15 (3): 383–404.

Martinez-Maldonado, Roberto, Vanessa Echeverria, Gloria Fernandez Nieto, and Simon Buckingham Shum. 2020. “From Data to Insights: A Layered Storytelling Approach for

Multimodal Learning Analytics." *Conference on Human Factors in Computing Systems - Proceedings*, 1–15. https://doi.org/10.1145/3313831.3376148.

Mauthner, Natasha S., Odette Parry, and Kathryn Backett-Milburn. 1998. "The Data Are Out There, or Are They? Implications for Archiving and Revisiting Qualitative Data:" *Sociology* 32 (4): 733–45. https://doi.org/10.1177/0038038598032004006.

Mayernik, Matthew S. 2011. "Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators." In .

McDonald, John. 1995. "Managing Records in the Modern Office: Taming the Wild Frontier." *Archivaria* 39: 70–79.

McDowell, Kate. 2018. "Storytelling: Practice and Process as Non-Textual Pedagogy." *Education for Information* 34: 15–19. https://doi.org/10.3233/EFI-189003.

McGinity, Ruth, and Maija Salokangas. 2014. "Introduction: 'Embedded Research' as an Approach into Academia for Emerging Researchers." *Management in Education* 28 (1): 3–5.

McIntyre, Alice. 2007. *Participatory Action Research*. Sage Publications.

Méndez, Gonzalo Gabriel, Uta Hinrichs, and Miguel A. Nacenta. 2017. "Bottom-up vs. Top-down: Trade-Offs in Efficiency, Understanding, Freedom and Creativity with Infovis Tools." *Conference on Human Factors in Computing Systems - Proceedings* 2017-May: 841–52. https://doi.org/10.1145/3025453.3025942.

Moore, Niamh. 2006. "The Contexts of Context: Broadening Perspectives in the (Re)Use of Qualitative Data:" *Methodological Innovations Online* 1 (2): 21–32. https://doi.org/10.4256/MIO.2006.0009.

Moran, Thomas P., and John M. Carroll. 1996. "Overview of Design Rationale." In *Design Rationale: Concepts, Techniques, and Use*, edited by Thomas P. Moran and John M. Carroll, 1–19. CRC Press. https://doi.org/10.1201/9781003064053-1/OVERVIEW-DESIGN-RATIONALE-THOMAS-MORAN-JOHN-CARROLL.

Moreau, Katherine A., Kaylee Eady, Lindsey Sikora, and Tanya Horsley. 2018. "Digital Storytelling in Health Professions Education: A Systematic Review." *BMC Medical Education* 18 (1): 208. https://doi.org/10.1186/s12909-018-1320-1.

Mosconi, Gaia, Matthias Korn, Christian Reuter, Peter Tolmie, Maurizio Teli, and Volkmar Pipek. 2017. "From Facebook to the Neighbourhood: Infrastructuring of Hybrid Community Engagement." *Computer Supported Cooperative Work (CSCW)* 26 (4): 959–1003. https://doi.org/10.1007/s10606-017-9291-z.

Mosconi, Gaia, Qinyu Li, Dave Randall, Helena Karasti, Peter Tolmie, Jana Barutzky, Matthias Korn, and Volkmar Pipek. 2019a. "Three Gaps in Opening Science." *Computer Supported Cooperative Work* 28 (3–4): 749–89. https://doi.org/10.1007/S10606-019-09354-Z.

———. 2019b. "Three Gaps in Opening Science." *Computer Supported Cooperative Work (CSCW)* 28 (3–4): 749–89. https://doi.org/10.1007/s10606-019-09354-z.

Mosconi, Gaia, Dave Randall, Helena Karasti, Saja Aljuneidi, Tong Yu, Peter Tolmie, and Volkmar Pipek. 2022. "Designing a Data Story: A Storytelling Approach to Curation, Sharing and Data Reuse in Support of Ethnographically-Driven Research." *Proc. ACM Hum.-Comput. Interact.* 6 (CSCW2). https://doi.org/10.1145/3555180.

Mozersky, Jessica, Heidi Walsh, Meredith Parsons, Tristan McIntosh, Kari Baldwin, and James M. DuBois. 2020. "Are We Ready to Share Qualitative Research Data? Knowledge and Preparedness among Qualitative Researchers, IRB Members, and Data Repository Curators." *IASSIST* 43 (4).

Murray-Rust, Peter. 2008. "Open Data in Science." *Serials Review* 34 (1): 52–64.

Neale, Bren. 2013. "Adding Time into the Mix: Stakeholder Ethics in Qualitative Longitudinal Research." *Methodological Innovations Online* 8 (2): 6–20. https://doi.org/10.4256/MIO.2013.010.

Nelson, Sharon D, and John W Simek. 2011. “Data Dumps: The Bane of E-Discovery.” *OR. ST. B. BULL* 71: 36–36.

Niu, Jinfang, and Margaret Hedstrom. 2008. “Documentation Evaluation Model for Social Science Data.” In *Proceedings of the American Society for Information Science and Technology*, 45:11–11. John Wiley & Sons, Ltd. https://doi.org/10.1002/MEET.2008.1450450223.

OECD. 2007. “Annual Report.”

Ojo, Adegboyega, and Bahareh Heravi. 2017. “Patterns in Award Winning Data Storytelling: Story Types, Enabling Tools and Competences.” *Digital Journalism* 6 (6): 693–718. https://doi.org/10.1080/21670811.2017.1403291.

Organizations, Non-profit, Sheena Erete, Emily Ryou, Geoff Smith, Khristina Fassett, and Sarah Duda. 2016. “Storytelling with Data : Examining the Use of Data By.” *In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1273–83.

Ortner, Sherry B. 2006. *Anthropology and Social Theory: Culture, Power, and the Acting Subject*. Duke University Press.

Oßwald, Achim, and Stefan Strathmann. 2012. “The Role of Libraries in Curation and Preservation of Research Data in Germany: Findings of a Survey.” In *IFLA World Library and Information Congress 78th IFLA General Conference and Assembly*. Helsinki, Finland.

Pampel, Heinz, and Sünje Dallmeier-Tiessen. 2014. “Open Research Data: From Vision to Practice.” In *Opening Science: The Evolving Guide on How the Internet Is Changing Research, Collaboration and Scholarly Publishing*, edited by Sönke Bartling and Sascha Friesike, 213–24. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_14.

Pantazos, Kostas, and Soren Lauesen. 2012. “Constructing Visualizations with InfoVis Tools: An Evaluation from a User Perspective.” *GRAPP 2012 IVAPP 2012 - Proceedings of the International Conference on Computer Graphics Theory and Applications and International Conference on Information Visualization Theory and Applications*, 731–36. https://doi.org/10.5220/0003860507310736.

Park, Robert E., and Ernest Burgess. 1925. “Chicago.” *Condor* 102: 848-854.

Pasquetto, Irene V, Ashley E Sands, and Christine L Borgman. 2015. “Exploring Openness in Data and Science: What Is ‘Open,’ to Whom, When, and Why?” *Proceedings of the Association for Information Science and Technology* 52 (1): 1–2. https://doi.org/https://doi.org/10.1002/pra2.2015.1450520100141.

Pasquetto, Irene V., Ashley E. Sands, Peter T. Darch, and Christine L. Borgman. 2016. “Open Data in Scientific Settings.” *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI ’16*, 1585–96. https://doi.org/10.1145/2858036.2858543.

Pepper, Coral, and Helen Wildy. 2009. “Using Narratives as a Research Strategy.” *Qualitative Research Journal*.

Pinfield, Stephen, Andrew M Cox, and Jen Smith. 2014. “Research Data Management and Libraries: Relationships, Activities, Drivers and Influences.” *PLoS One* 9 (12): e114734.

Pipek, Volkmar, and Volker Wulf. 2009. “Infrastructuring: Toward an Integrated Perspective on the Design and Use of Information Technology.” *Journal of the Association for Information Systems (JAIS)* 10 (5): 447–73.

Plantin, Jean-Christophe, Carl Lagoze, and Paul N Edwards. 2018. “Re-Integrating Scholarly Infrastructure: The Ambiguous Role of Data Sharing Platforms.” *Big Data & Society* 5 (1): 205395171875668. https://doi.org/10.1177/2053951718756683.

Poirier, Lindsay. 2017. “Devious Design: Digital Infrastructure Challenges for Experimental Ethnography.” *Design Issues* 33 (2): 70–83.

Preuss, Nils, Georg Staudter, Moritz Weber, Reiner Anderl, and Peter F Pelz. 2018. "Methods and Technologies for Research-and Metadata Management in Collaborative Experimental Research." In *Applied Mechanics and Materials*, 885:170–83. Trans Tech Publ.

Pryor, Graham. 2014. "Who's Doing Data? A Spectrum of Roles, Responsibilities, and Competencies." *Delivering Research Data Management Services: Fundamentals of Good Practice*, 41–58.

Pryor, Graham, S Jones, and A Whyte. 2013. "Options and Approaches to RDM Service Provision." *Delivering Research Data Management Services: Fundamentals of Good Practice*, 21.

Randall, Dave, Markus Rohde, Kjeld Schmidt, and Volker Wulf. 2018. "Introduction: Socio-Informatics—Practice Makes Perfect?" In *Socio-Informatics*, edited by Volker Wulf, Volkmar Pipek, David Randall, Markus Rohde, Kjeld Schmidt, and Gunnar Stevens. Vol. 1. Oxford University Press. https://doi.org/10.1093/OSO/9780198733249.003.0001.

Randall, David, Richard Harper, and Mark Rouncefield. 2007. *Fieldwork for Design: Theory and Practice.* Springer Science & Business Media.

Rawson, Katie, and Trevor Muñoz. 2016. "Against Cleaning." *Curating Menus* 6: 1–14.

Reilly, Susan. 2012. "The Role of Libraries in Supporting Data Exchange." In *IFLA World Library and Information Congress 78th IFLA General Conference and Assembly*. Helsinki, Finland.

Remy, Christian, Oliver Bates, Alan Dix, Vanessa Thomas, Mike Hazas, Adrian Friday, and Elaine M Huang. 2018. "Evaluation beyond Usability: Validating Sustainable HCI Research." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.

Restrepo, Elizabeth, and Lisa Davis. 2003. "Storytelling: Both Art and Therapeutic Practice." *International Journal of Human Caring* 7 (1): 43–48. https://doi.org/10.20467/1091-5710.7.1.43.

RfII. 2016. "German Council for Scientific Information Infrastructures: Enhancing Research Data Management: Performance through Diversity. Recommendations Regarding Structures, Processes, and Financing for Research Data Management in Germany."

Ribes, David, and Charlotte P Lee. 2010. "Sociotechnical Studies of Cyberinfrastructure and E-Research: Current Themes and Future Trajectories." *Computer Supported Cooperative Work (CSCW)* 19 (3): 231–44. https://doi.org/10.1007/s10606-010-9120-0.

Rohde, Markus, Peter Brödner, Gunnar Stevens, Matthias Betz, and Volker Wulf. 2017. "Grounded Design-a Praxeological IS Research Perspective." *Journal of Information Technology* 32 (2): 163–79.

Rolland, Betsy, and Charlotte P. Lee. 2013a. "Beyond Trust and Reliability: Reusing Data in Collaborative Cancer Epidemiology Research." In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 435–44. San Antonio, Texas, USA: ACM Press. https://doi.org/10.1145/2441776.2441826.

———. 2013b. "Beyond Trust and Reliability: Reusing Data in Collaborative Cancer Epidemiology Research." *In Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'13)*, 435–44. https://doi.org/10.1145/2441776.2441826.

Rowhani-Farid, Anisa, Michelle Allen, and Adrian G. Barnett. 2017. "What Incentives Increase Data Sharing in Health and Medical Research? A Systematic Review." *Research Integrity and Peer Review* 2 (1): 1–10.

Ruggiano, Nicole, and Tam E Perry. 2019. "Conducting Secondary Analysis of Qualitative Data: Should We, Can We, and How?:" *Qualitative Social Work* 18 (1): 81–97. https://doi.org/10.1177/1473325017700701.

Ryen, Anne. 2011. "Ethics and Qualitative Research." *Qualitative Research* 3: 416–238.

Sacks, Harvey. 1992. *Lectures on Conversation: Volume I*. Massachusetts: Blackwell.

Schmidt, Kjeld, and Liam Bannon. 1992. "Taking CSCW Seriously: Supporting Articulation Work." In *Cooperative Work and Coordinative Practices*, 45–71. Springer.

Schmidt, Robert. 2016. "The Methodological Challenges of Practising Praxeology." In *Practice Theory and Research*, 59–75. Routledge.

Schön, Donald. 1983. *The Reflective Practitioner*. London: Temple Smith.

Simonsen, Jesper, Helena Karasti, and Morten Hertzum. 2020. "Infrastructuring and Participatory Design: Exploring Infrastructural Inversion as Analytic, Empirical and Generative." *Computer Supported Cooperative Work (CSCW)* 29 (1): 115–51.

Small, Stephen A, and Lynet Uttal. 2005. "Action-oriented Research: Strategies for Engaged Scholarship." *Journal of Marriage and Family* 67 (4): 936–48.

Sole, Deborah, and Daniel Gray Wilson. 2002. "Storytelling in Organizations: The Power and Traps of Using Stories to Share Knowledge in Organizations." *LILA, Harvard, Graduate School of Education*, 1–12.

Star, Susan Leigh, and Geoffrey C. Bowker. 2002. "How to Infrastructure." Incollection. In *Handbook of New Media*, edited by L. A. Lievrouw and S. Livingstone, 151–62. London, United Kingdom, UK: SAGE Pub.

Star, Susan Leigh, and Karen Ruhleder. 1996. "Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces." *Information Systems Research* 7 (1): 111–34. https://doi.org/10.1287/isre.7.1.111.

Strauss, Anselm. 1985. "Work and the Division of Labour." *The Sociological Quarterly* 26 (1): 1–19. https://doi.org/https://doi.org/10.1111/j.1533-8525.1985.tb00212.x.

Strauss, Anselm, and Juliet Corbin. 1998. "Basics of Qualitative Research Techniques."

Sturm, Brian W, and Sarah Beth Nelson. 2016. "With Our Own Words: Librarians' Perceptions of the Values of Storytelling in Libraries." *Storytelling, Self, Society* 12 (1): 4–23. https://doi.org/10.13110/storselfsoci.12.1.0004.

Tanenbaum, Andrew. 2002. *Computer Networks*. 4th ed. Prentice Hall.

Taylor, John M. 2001. "The UK E-Science Programme [Powerpoint Presentation], e-Science London Meeting."

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev U. Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. "Data Sharing by Scientists: Practices and Perceptions." *PloS One* 6 (6). https://doi.org/10.5061/dryad.6t94p.

Thomas, David R. 2006a. "A General Inductive Approach for Analyzing Qualitative Evaluation Data." *American Journal of Evaluation* 27 (2): 237–46.

———. 2006b. "A General Inductive Approach for Analyzing Qualitative Evaluation Data." *American Journal of Evaluation* 27 (2): 237–46.

Thomas, William I., and Znaniecki Florian. 1927. *The Polish Peasant in Europe* . New York, NY.

Thomer, Andrea K, and Karen M Wickett. 2020. "Relational Data Paradigms: What Do We Learn by Taking the Materiality of Databases Seriously?" *Big Data & Society* 7 (1). https://doi.org/10.1177/2053951720934838.

Treloar, Andrew, and Cathrine Harboe-Ree. 2008. "Data Management and the Curation Continuum: How the Monash Experience Is Informing Repository Relationships." *VALA2008 Proceedings*, 13. http://www.vala.org.au/vala2008-proceedings/vala2008-session-6-treloar.

Treloar, Andrew, and Jens Klump. 2019. "Updating the Data Curation Continuum: Not Just Data, Still Focussed on Curation, More about Domains." *International Journal of Digital Curation* 14 (1): 87–101.

Tsai, Alexander C., Brandon A. Kohrt, Lynn T. Matthews, Theresa S Betancourt, Jooyoung K. Lee, Andrew V. Papachristos, Sheri D. Weiser, and Shari L. Dworkin. 2016. "Promises and Pitfalls of Data Sharing in Qualitative Research." *Social Science & Medicine* 169 (November): 191–98. https://doi.org/10.1016/J.SOCSCIMED.2016.08.004.

Twidale, Michael, David Randall, and Richard Bentley. 1994. "Situated Evaluation for Cooperative Systems." In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, 441–52.

Tylor, Edward B. 1882. "Notes on the Asiatic Relations of Polynesian Culture." *The Journal of the Anthropological Institute of Great Britain and Ireland* 11: 401–5. https://doi.org/10.2307/2841767.

Vecchi, Nadia De, Amanda Kenny, Virginia Dickson-Swift, and Susan Kidd. 2016. "How Digital Storytelling Is Used in Mental Health: A Scoping Review." *International Journal of Mental Health Nursing* 25 (3): 183–93. https://doi.org/10.1111/INM.12206.

Velden, Theresa. 2013. "Explaining Field Differences in Openness and Sharing in Scientific Communities." Edited by Amy Bruckman, Scott Counts, Cliff Lampe, and Loren Terveen. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work - CSCW '13*. New York, New York, USA: ACM Press. https://doi.org/10.1145/2441776.2441827.

Vertesi, Janet, and Paul Dourish. 2011a. "The Value of Data: Considering the Context of Production in Data Economies." In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 533–42. Association for Computing Machinery. https://doi.org/10.1145/1958824.1958906.

———. 2011b. "The Value of Data : Considering the Context of Production in Data Economies." *Cscw2011*, 533–42. https://doi.org/10.1145/1958824.1958906.

Vertesi, Janet, Jofish Kaye, Samantha N. Jarosewski, Vera D. Khovanskaya, and Jenna Song. 2016. "Data Narratives: Uncovering Tensions in Personal Data Management." *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, 477–89. https://doi.org/10.1145/2818048.2820017.

Vygotsky, Lev S. 1980. *Mind in Society: The Development of Higher Psychological Processes*. Harvard university press.

Wallis, Jillian C., Elizabeth Rolando, and Christine L. Borgman. 2013. "If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology." *PLOS ONE* 8 (7): e67332. https://doi.org/10.1371/JOURNAL.PONE.0067332.

Walters, Peter. 2009. "Qualitative Archiving: Engaging with Epistemological Misgivings." *Australian Journal of Social Issues* 44 (3): 309–20. https://doi.org/10.1002/J.1839-4655.2009.TB00148.X.

West, Christina H., Kendra L. Rieger, Amanda Kenny, Rishma Chooniedass, Kim M. Mitchell, Andrea Winther Klippenstein, Amie-Rae Zaborniak, Lisa Demczuk, and Shannon D. Scott. 2022. "Digital Storytelling as a Method in Health Research: A Systematic Review." *Systematic Reviews* 21: 1–25. https://doi.org/10.1177/16094069221111118.

Whyte, Angus. 2014. "A Pathway to Sustainable Research Data Services from Scoping to Sustainability." *Delivering Research Data Management Services: Fundamentals of Good Practice*, 59–88.

Whyte, Angus, and Jonathan Tedds. 2011. "Making the Case for Research Data Management." *A Digital Curation Centre Briefing Paper*, no. September: 1–8. https://doi.org/10.5281/ZENODO.817936.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 1–9. https://doi.org/10.1038/sdata.2016.18.

Wilms, Konstantin, Stefan Stieglitz, Alina Buchholz, Raimund Vogl, and Dominik Rudolph. 2018. "Do Researchers Dream of Research Data Management?" In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Wilson, James AJ, Luis Martinez-Uribe, Michael A. Fraser, and Paul Jeffreys. 2011. “An Institutional Approach to Developing Research Data Management Infrastructure.”

Wu, Jing, and Der Thanq Victor Chen. 2020. “A Systematic Review of Educational Digital Storytelling.” *Computers and Education* 147 (July 2018): 103786. https://doi.org/10.1016/j.compedu.2019.103786.

Wulf, Volker, Claudia Müller, Volkmar Pipek, David Randall, Markus Rohde, and Gunnar Stevens. 2015a. “Practice-Based Computing: Empirically Grounded Conceptualizations Derived from Design Case Studies.” In *Designing Socially Embedded Technologies in the Real-World*, edited by Volker Wulf, Kjeld Schmidt, and David Randall, 111–50. London, UK: Springer London. https://doi.org/10.1007/978-1-4471-6720-4_7.

———. 2015b. “Practice-Based Computing: Empirically Grounded Conceptualizations Derived from Design Case Studies.” In *Designing Socially Embedded Technologies in the Real-World*, edited by Volker Wulf, Kjeld Schmidt, and David Randall, 111–50. London, UK: Springer London. https://doi.org/10.1007/978-1-4471-6720-4_7.

Wulf, Volker, Volkmar Pipek, David A. Randall, Markus Rohde, Kjeld Schmidt, and Stevens Gunnar. 2018. *Socio-Informatics. A Practice-Based Perspective on the Design and Use of IT Artifacts*. Oxford: Oxford University Press.

Xu, Xian, Leni Yang, David Yip, Mingming Fan, Zheng Wei, and Huamin Qu. 2022. “From ‘Wow’ to ‘Why’: Guidelines for Creating the Opening of a Data Video with Cinematic Styles.” *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3491102.3501896.

Yi, Ji Soo, Youn Ah Kang, John T. Stasko, and Julie A. Jacko. 2007. “Toward a Deeper Understanding of the Role of Interaction in Information Visualization.” *IEEE Transactions on Visualization and Computer Graphics* 13 (6): 1224–31. https://doi.org/10.1109/TVCG.2007.70515.

Yoon, Ayoung. 2014. “‘Making a Square Fit into a Circle’: Researchers’ Experiences Reusing Qualitative Data.” In *Proceedings of the American Society for Information Science and Technology*, 51:1–4. John Wiley & Sons, Ltd. https://doi.org/10.1002/MEET.2014.14505101140.

Yuksel, Pelin, Bernard R. Robin, and Sara McNeil. 2011. “Educational Uses of Digital Storytelling All around the World.” In *Society for Information Technology & Teacher Education International Conference. Association for the Advancement of Computing in Education (AACE)*, 1264–71.

Zimmerman, Ann. 2007. “Not by Metadata Alone: The Use of Diverse Forms of Knowledge to Locate Data for Reuse.” *Journal on Digital Libraries 7, No. 1-2 (20* 7 (1–2): 5-16.

Zuiderwijk, Anneke, Marijn Janssen, Sunil Choenni, Ronald Meijer, and Roexsana Sheikh Alibaks. 2012. “Socio-technical Impediments of Open Data.” *Electronic Journal of E-Government* 10 (2): 156–72.

Zuiderwijk, Anneke, and Helen Spiers. 2019. “Sharing and Re-Using Open Data: A Case Study of Motivations in Astrophysics, International Journal of Information Management.” *International Journal of Information Management* 49.