Three gaps in Opening Science

Gaia Mosconi, Qinyu Li, Dave Randall, Helena Karasti, Peter Tolmie, Jana Barutzky, Matthias Korn, Volkmar Pipek
University of Siegen
{gaia.mosconi;qinyu.li;dave.randall;helena.karasti;peter.tolmie;jana.barutzy; matthias.korn;volkmar.pipek}@uni-siegen.de

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.

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 et. al. 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.[1] 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 process. The Open Science movement has successively elaborated principles[2] 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 (Arzberger et al. 2006; Wallis et al. 2013; 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; Pasquetto et al. 2015; Wallis et al. 2013; Choi and Tausczik 2017). 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 (Pampel and Dallmeier-Tiessen 2014; Osswald and Strathmann 2012; 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.[3] 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 (European Union 2016).

[1] 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)

[2] 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)

[3] 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.

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 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 Wolfgang 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.

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.

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 (Taylor 2001 in Dallas 2016, p.4). 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.[4]

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.[5] 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-level overview of the curation stages of research data that was later simplified and adapted by other Data Centres and

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

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

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:

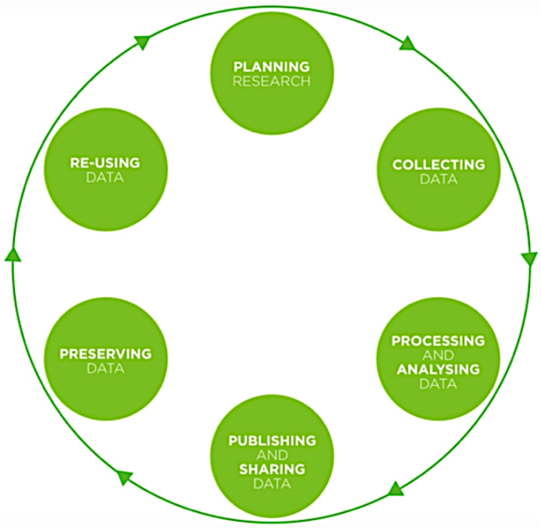
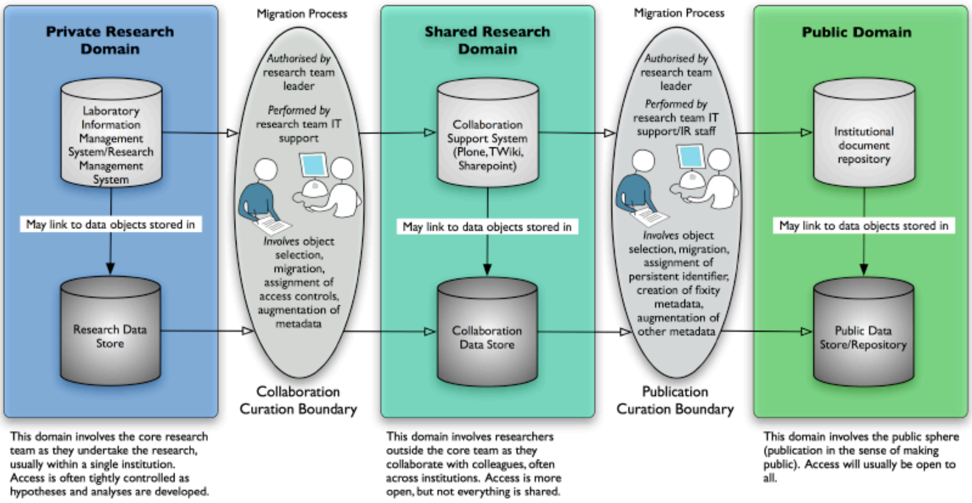


Figure 1: Data life cycle model (UK Data Archive).

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 et al. 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
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 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[6] 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 at al. 2012). The 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.

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

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

sociotechnical infrastructure needed to support shared common resources, access to dataset and special tools for data storage and processing (Jirotka et al. 2013; Ribes and Lee 2010; Bietz et al. 2010; Lee et al. 2006; Jackson et al. 2007; Karasti et al. 2006; Edwards et al. 2013; Karasti and Baker 2004; Karasti et al 2010; Ribes and Finholt 2009).

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.

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[7].

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 (Borgman 2012). Borgman (2015) also argues that the diversity of the data arising across different research approaches and fields leads to it being structured

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

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 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 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 et al. 2015a, 2015b). 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 et al. 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.

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.

3 Research settings and methodological approach

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[8] 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

[8] 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 et al. 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[9]
#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

[9] 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.

#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	Anthropologist	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.

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 studentsHaving 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.

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.

4.1 Research Data Management practices bottom-up

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[10], 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]

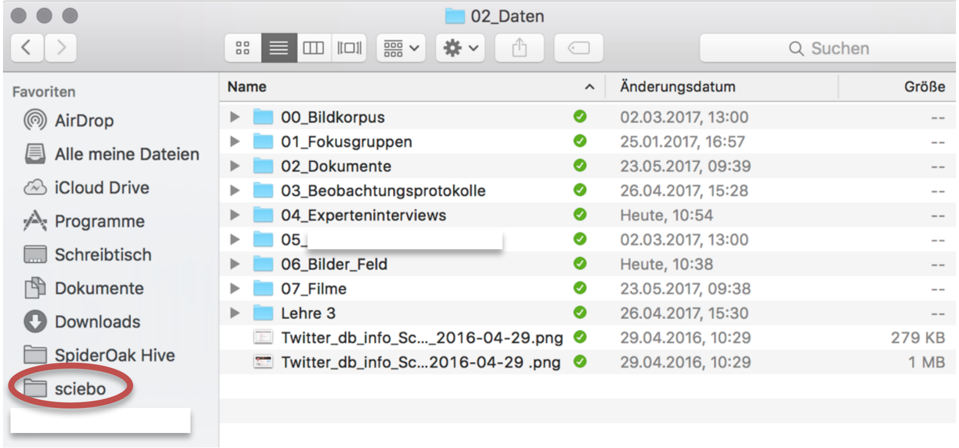
Tools, software choices (storage system, groupware solution, etc.), the appropriate data infrastructure and how to make best use of it is all, according to the researchers, left for them to discover by themselves. Cliff continues:

> "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 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!".

[10] Anonymized

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[11]”, 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”.



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	--
05_[redacted]	✓	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

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

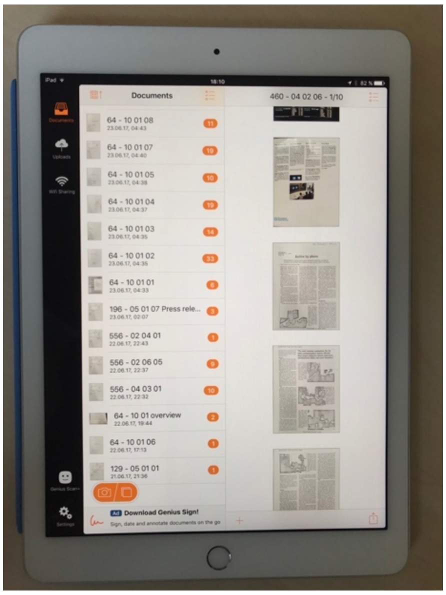
data will be kept and where; and 4) intentions to publish the data. Most interviewees were following an orally-based consent protocol:

> "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.

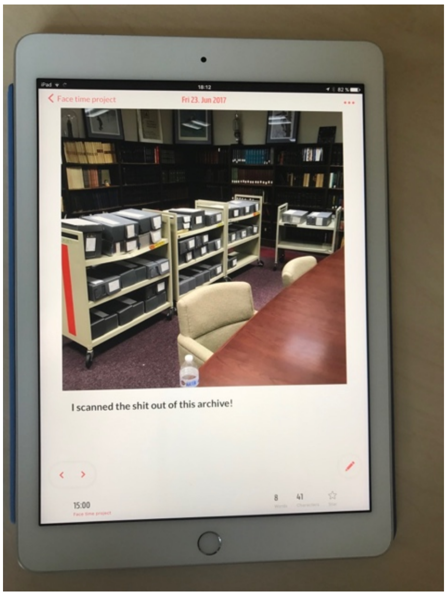
4.1.2 Messy folders and software support

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".



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

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



I scanned the shit out of this archive!

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".

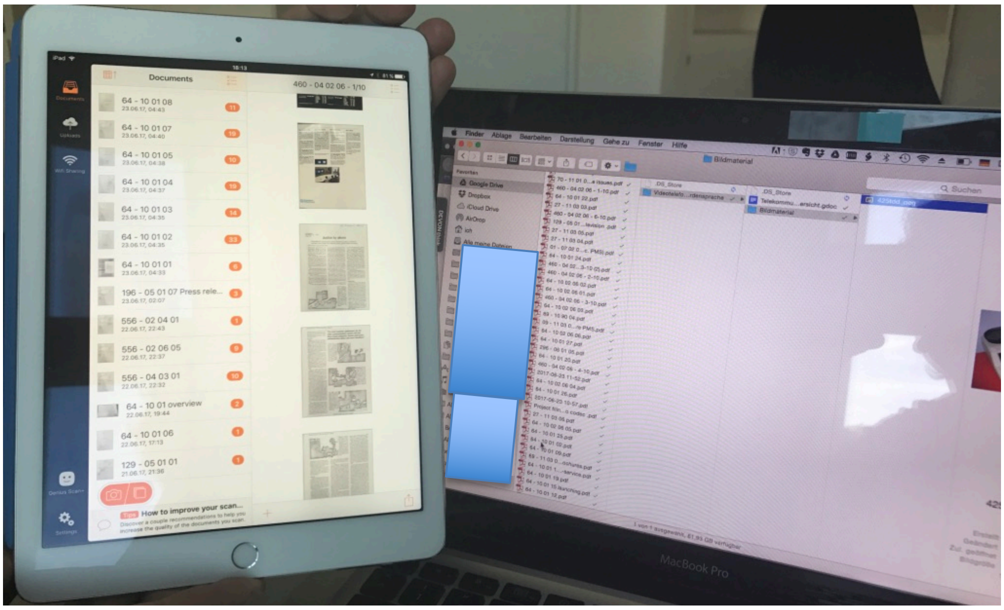


Figure 6: Screenshot of Colin's document scanning App and the related Google Drive folder where he saved and stored the files

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.

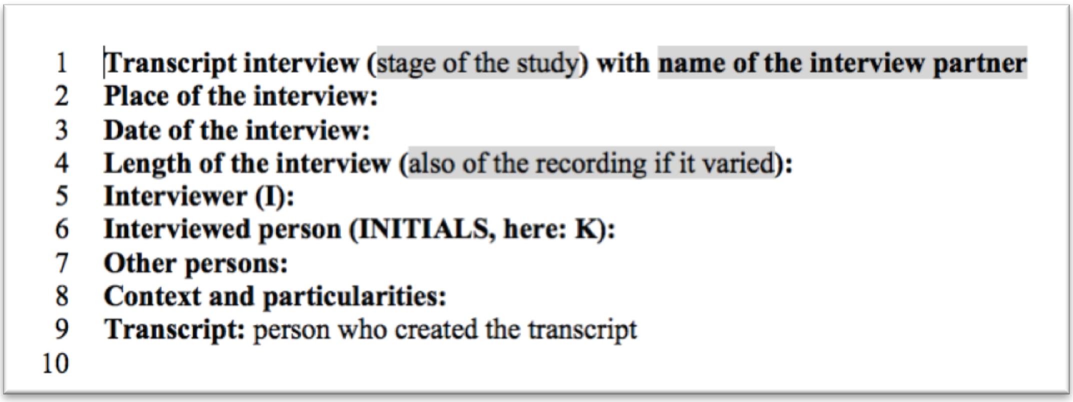
4.1.3 Metadata: what is metadata?

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 et al. 2014; Eschenfelder and Johnson 2011).

4.2 Open Science perspectives

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]

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.

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 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.

5.1 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.

5.1.1 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.

5.1.2 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.

5.2 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.

5.3 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.

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 tolls 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.

Acknowledgements

This research has been possible thanks to the engagement of many scholars, the CRC “Media of Cooperation” organization board and the IT service provider with whom we have worked with and learned from. The findings in this paper originate from the project INF funded by a grant of the DFG (SFB 1187).

References

Abbott, Daisy (2008). *What is Digital Curation?* DCC Briefing Papers: Introduction to Curation. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation. Accessed 13 February 2019.

Arzberger, Peter; Peter Schroeder; Anne Beaulieu; Geof Bowker; Kathleen Casey; Leif Laaksonen; David Moorman; Paul Uhlir; and Paul Wouters (2006). Promoting Access to Public Research Data for Scientific, Economic, and Social Development. *Data Science Journal,* vol. 3, pp. 135–152.

Asher, Andrew; and Lori M. Jahnke (2013). Curating the Ethnographic Moment. *Archive Journal,* no. 3. Available online http://www.archivejournal.net/essays/curating-the-ethnographic-moment/. Accessed 13 February 2019.

Bechhofer, Sean; David De Roure; Matthew Gamble; Carole Goble; and Buchan Iain (2010). Research Objects: Towards Exchange and Reuse of Digital Knowledge. In *FWCS 2010. Proceedings of The Future of the Web for Collaborative Science, Raleigh, USA, April 26, 2010.* Nature Proceedings. 6 pages.

Bietz, Matthew J.; Eric P. Baumer; and Charlotte P. Lee (2010). Synergizing in Cyberinfrastructure Development. *Computer Supported Cooperative Work (CSCW),* vol. 19, no. 3-4, pp. 245–281.

Bietz, Matthew J.; and Charlotte P. Lee (2009). Collaboration in Metagenomics: Sequence Databases and the Organization of Scientific Work. In I. Wagner, H. Tellioğlu, E. Balka, C. Simone and L. Ciolfi (eds): *ECSCW 2009. Proceedings of the 11th European Conference on Computer Supported Cooperative Work, Vienna, Austria, 7-11 September 2009.* London: Springer London, pp. 243–262.

Birnholtz, Jeremy P.; and Matthew J. Bietz (2003). Data at work: Supporting sharing in science and engineering. In M. Pendergast, K. Schmidt, C. Simone and M. Tremaine (eds): *GROUP'03: Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, Sanibel Island, FL, United States, 9 November – 12 November 2003.* New York: ACM Press. pp. 339–348.

Bishop, Libby (2012). Using archived qualitative data for teaching: practical and ethical considerations. *International Journal of Social Research Methodology,* vol. 15, no. 4, pp. 341–350.

Borgman, Christine L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology,* vol. 63, no. 6, pp. 1059–1078.

Bowker, Geoffrey C. (2005). *Memory practices in the sciences.* Cambridge, MA: MIT Press.

Broom, Alex; Lynda Cheshire; and Michael Emmison (2009). Qualitative Researchers’ Understandings of Their Practice and the Implications for Data Archiving and Sharing. *Sociology,* vol. 43, no. 6, pp. 1163–1180.

Cadiz, J. J.; Anop Gupta; and Grudin Jonathan (2000). Using Web annotations for asynchronous collaboration around documents. In W. Kellogg and S. Whittaker (eds): *CSCW’00: Proceedings of the 2000 ACM conference on Computer supported cooperative work, Philadelphia, PA, USA, December 2-6, 2000.* New York: ACM Press, pp. 309–318.

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,* vol. 12, no. 2, pp. 635–651.

Caton, Hiram (1990). *The Samoa reader. Anthropologists take stock.* Lanham, Md: University Press of America.

Chang, Yuan-Chia; Hao-Chuan Wang; Hung-kuo Chu; Shung-Ying Lin; and Wang Shuo-Ping (2017). AlphaRead: Support Unambiguous Referencing in Remote Collaboration with Readable Object

Annotation. In C. P. Lee, S. Poltrock, L. Barkhuus, M. Borges and W. Kellogg (eds): *CSCW'17. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, Portland, Oregon, USA, 25 February – 01 March 2017.* New York: ACM Press, pp. 2246–2259.

Choi, Joohee; and Yla Tausczik (2017). Characteristics of Collaboration in the Emerging Practice of Open Data Analysis. In C.P. Lee, S. Poltrock, L. Barkhuus, M. Borges and W. Kellogg (eds): *CSCW'17. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, Portland, Oregon, USA, 25 February – 01 March 2017.* New York: ACM Press, pp. 835–846.

Corti, Louise (2007). Re-using archived qualitative data – where, how, why? *Archival Science,* vol. 7, no. 1, pp. 37–54.

Dachtera, Juri; Dave Randall; and Volker Wulf (2014). Research on research. In M. Jones, P. Palanque, A. Schmidt and T. Grossman (eds): *CHI'14. Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, Toronto, Canada, 26 April – 1 May 2014.* New York: ACM Press, pp. 713–722.

Dallas, Costis (2007) An agency-oriented approach to digital curation theory and practice. In J. Trant and D. Bearman (eds): *ICHIM'07. Proceedings of the International Cultural Heritage Informatics Meeting.* Toronto: Archives & Museum Informatics. Available online: http://www.archimuse.com/ichim07/papers/dallas/dallas.html. Accessed 13 February 2019.

Dallas, Costis (2016). Digital curation beyond the "wild frontier": a pragmatic approach. *Archival Science,* vol. 16, no. 4, pp. 421–457.

DFG (2010). *Principles for the Handling of Research Data*. Available: https://www.wissenschaftsrat.de/download/archiv/Allianz-Principles_Research_Data_2010.pdf. Accessed 19 February 2019.

Eberhard, Igor; and Wolfgang Kraus (2018). Der Elefant im Raum. Ethnographisches Forschungsdatenmanagement als Herausforderung für Repositorien. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare,* vol. 71, no. 1, pp. 41-52.

Edwards, Paul N.; Steven J. Jackson; Melissa K. Chalmers; Geoffrey C. Bowker; Christine .L Borgman.; David Ribes; Matt Burton; and Calvert Scout (2013). *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. Ann Arbor: Deep Blue.

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,* vol. 41, no. 5, pp. 667–690.

Erickson, Ingrid; Kristin Eschenfelder; Sean Goggins; Libby Hemphill; Steve Sawyer; Kalpana Shankar; and Katie Shilton (2014). The ethos and pragmatics of data sharing. In *CSCW'14. Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing, Baltimore, Maryland, USA, 15 February – 19 February 2014.* New York: ACM Press, pp. 109–112.

Eschenfelder, Kristin; and Andrew Johnson (2011). The Limits of sharing: Controlled data collections. *Proceedings of the American Society for Information Science and Technology,* vol. 48, no. 1, pp. 1–10.

European Commission (2016). *H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020.* Available online: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf. Accessed 13 February 2019.

European Union (2010). *Riding the wave. How Europe can gain from the rising tide of scientific data*. Final report of the High Level Expert Group on Scientific Data. Available online: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=707. Accessed 13 February 2019.

European Union (2015). *Access to and preservation of scientific information in Europe. Report on the implementation of Commission Recommendation C(2012) 4890 final*, Luxembourg: Publications Office of the European Union. Available online: https://ec.europa.eu/research/openscience/pdf/openaccess/npr_report.pdf. Accessed 13 February 2019.

Faniel, Ixchel M.; and Trond E. Jacobsen (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Computer Supported Cooperative Work (CSCW),* vol. 19, no. 3-4, pp. 355–375.

Fecher, Benedikt; and Sascha Friesike (2014). Open Science: One Term, Five Schools of Thought. In S. Bartling and S. Friesike (eds): *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing*. London: Springer, pp. 17–47.

Fecher, Benedikt; Sascha Friesike; and Marcel Hebing (2015a). What drives academic data sharing? *PloS one,* vol. 10, no. 2, e0118053.

Fecher, Benedikt; Sascha Friesike; Marcel Hebing; Stephanie Linek; and Armin Sauermann (2015b). A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing. *DIW Berlin Discussion Paper,* no. 1454.

Freeman, Richard; and Jerome Crowder (2016) Abstract: Digital Files and the future of anthropological data: ethics and organization. In *ORGANIZE THIS!: Data management for anthropology in the digital age, preserving our evidence for future discovery*. Minneapolis, MN. 2016 American Anthropological Association, pp. 1–2.

Gillies, Val; and Rosalind Edwards (2005). Secondary Analysis in Exploring Family and Social Change: Addressing the Issue of Context. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research,* vol. 6, no. 1, Art. 44.

Gitelman, Lisa (2013). *"Raw data" is an oxymoron. Infrastructures series*. Cambridge, MA: MIT Press.

Gooch, Amanda J. (2014). *Data Storage and Sharing: A Needs Assessment Survey of Social Science Researchers and Information Professionals for Developing a Data Management Curriculum. A Master's Paper for the M.S. in L.S degree.*

Gupta, Shivam; and Claudia Müller-Birn (2018). A study of e-Research and its relation with research data life cycle: a literature perspective. *Benchmarking: An International Journal,* vol. 25, no. 6, pp. 1656–1680.

Halskov, Kim; Nicolai Brodersen Hansen (2015). The diversity of participatory design research practice at PDC 2002–2012. *International Journal of Human-Computer Studies,* vol. 74, pp. 81–92.

Hedges, Mark; Tobias Blanke; Stella Fabiane; Gareth Knight; and Eric Liao (2012). Sheer Curation of Experiments: Data, Process, Provenance. *Journal of Digital Information,* vol. 13, no. 1. https://journals.tdl.org/jodi/index.php/jodi/article/view/5883. Accessed 06 April 2019.

Hedstrom, Margaret (1997) Building record-keeping systems: archivists are not alone on the wild frontier. *Archivaria*, vol. 44, pp. 44–71. https://archivaria.ca/index.php/archivaria/article/viewFile/12196/13210. Accessed 07 April 2019.

Hey, Anthony J. G.; Stewart Tansley; and Kristin M. Tolle (eds) (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.

Jackson, Steven J.; Paul N. Edwards; Geoffrey C. Bowker; and Cory P. Knobel (2007). Understanding infrastructure: History, heuristics and cyberinfrastructure policy. *First Monday,* vol. 12, no. 6. https://www.firstmonday.org/ojs/index.php/fm/article/view/1904/1786. Accessed 06 April 2019.

Jirotka, Marina; Charlotte P. Lee; and Gary M. Olson (2013). Supporting Scientific Collaboration: Methods, Tools and Concepts. *Computer Supported Cooperative Work (CSCW),* vol. 22, no. 4-6, pp. 667–715.

Karasti, Helena; and Karen S. Baker (2004). Infrastructuring for the long-term: ecological information management. In *HICSS'3. Proceedings of the Hawaii International Conference on System Sciences 2004,* Hawaii, USA, January 5 - 8, 2004. IEEE. 10 pages.

Karasti, Helena; Karen S. Baker; and Eija Halkola (2006). 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),* vol. 15, no. 4, pp. 321–358.

Karasti, Helena; Baker, Karen S. Baker; and Florence Millerand (2010). Infrastructure Time: Long-term Matters in Collaborative Development. *Computer Supported Cooperative Work (CSCW),* vol. 19, no. 3-4, pp. 377–415.

Kelder, Jo-Anne (2005). Using Someone Else's Data: Problems, Pragmatics and Provisions. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research,* vol. 6, no. 1. http://www.qualitative-research.net/index.php/fqs/article/view/501. Accessed 06 April 2019.

Kervin, Karina; Robert B. Cook; and William K. Michener (2014). The Backstage Work of Data Sharing. In S. Goggins, I. Jahnke, D. W. McDonald and P. Bjørn (eds): *Group'14. Proceedings of the 18th ACM*

International Conference on Supporting Group Work, Sanibel Island, Florida, USA, 09 November – 12 November 2014. New York: ACM Press, pp. 152–156.

Kitchin, Rob (2014). *The data revolution. Big data, open data, data infrastructures & their consequences.* London: SAGE.

Korn, Matthias; Marén Schorch; Volkmar Pipek; Matthew Bietz; Carsten Østerlund; Rob Procter; David Ribes; and Robin Williams (2017). E-Infrastructures for Research Collaboration. In C.P. Lee, S. Poltrock, L. Barkhuus, M. Borges and W. Kellogg (eds): *CSCW'17 Companion. Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, Portland, Oregon, USA, 25 February – 01 March 2017.* New York: ACM Press, pp. 415–420.

Kroes, Neelie (2012). *Opening science through e-infrastructures.* (Speech-12-258) Available at: europa.eu/rapid/press-release_SPEECH-12-258_en.pdf. Accessed 07.01.2019.

Lee, Charlotte P.; Paul Dourish; and Gloria Mark (2006). The human infrastructure of cyberinfrastructure. In P. Hinds and D. Martin (eds): *CSCW'06. Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, Banff, Alberta, Canada, 04 November - 08 November 2006.* New York: ACM Press, pp. 483-492.

Lindley, Siân E.; Gavin Smyth; Robert Corish; Anastasia Loukianov; Michael Golembewski; Ewa A. Luger; and Sellen Abigali (2018). Exploring New Metaphors for a Networked World through the File Biography. In R. Mandryk, M. Hancock, M. Perry and A. Cox (eds): *CHI'18. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21 April – 26 April 2018.* New York: ACM Press, pp. 1–12.

Lord, Philip.; and Alison Macdonald (2003). *e-Science Curation Report: Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision.* The JISC Committee for the Support of Research (JCSR).

Marshall, Catherine C.; Ted Wobber; Venugopalan Ramasubramanian; and Terry Douglas B. (2012). Supporting research collaboration through bi-level file synchronization. In T.A. Finholt, H. Tellioğlu, K. Inkpen and T. Gross (eds): *GROUP'12. Proceedings of the 17th ACM international conference on Supporting Group Work, Sanibel Island, Florida, 27 October – 31 October 2012.* New York: ACM Press, pp. 165–174.

Marshall, Cathy; and John C. Tang (2012). That syncing feeling: early user experience with the cloud. In *DIS'12. Proceedings of the Designing Interactive Systems Conference, Newcastle Upon Tyne, United Kingdom, 11 June – 15 June 2012.* New York: ACM Press, pp. 544–553.

McDonald, John (1995). Managing records in the modern office: taming the wild frontier. *Archivaria,* vol. 39, pp. 70–79. https://archivaria.ca/archivar/index.php/archivaria/article/view/12069/13047. Accessed 07 April 2019.

Murray-Rust, Peter (2008). Open Data in Science. *Serials Review,* vol. 34, no. 1, pp. 52–64.

OECD (ed). *Annual Report 2007.*

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, 11.-17. August 2012.* 10 pages.

Pampel, Heinz; and Sünje Dallmeier-Tiessen (2014). Open Research Data: From Vision to Practice. In S. Bartling and S. Friesike (eds): *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing.* London: Springer, vol. 40, pp. 213–224.

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? In *Proceedings of the Association for Information Science and Technology,* vol. 52, no. 1, pp. 1-2

Rader, Emilee (2009). Yours, mine and (not) ours: social influences on group information repositories. In D.R. Olsen, R.B. Arthur, K. Hinckley, M.R Morris, S. Hudson and S. Greenberg (eds): *CHI'09. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 04 April – 09 April 2009.* New York: ACM Press, pp. 2095–2098.

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, 11.-17. August 2012.* 7 pages.

Ribes, David; and Thomas A. Finholt (2009). The Long Now of Technology Infrastructure: Articulating Tensions in Development. *Journal of the Association for Information Systems,* vol. 10, no. 5, pp. 375–398.

Ribes, David; and Charlotte P. Lee (2010). Sociotechnical Studies of Cyberinfrastructure and e-Research: Current Themes and Future Trajectories. *Computer Supported Cooperative Work (CSCW),* vol. 19, no. 3-4, pp. 231–244.

Rolland, Betsy; and Charlotte P. Lee (2013). Beyond trust and reliability: reusing data in collaborative cancer epidemiology research. In A. Bruckman, S. Counts, C. Lampe and L. Terveen (eds): *CSCW'13. Proceedings of the 2013 conference on Computer supported cooperative work, San Antonio, Texas, 23 February – 27 February 2013.* New York: ACM Press, pp. 435–444.

Scaramozzino, Jeanine M.; Marisa L. Ramírez; and Karen J. McGaughey (2012). A Study of Faculty Data Curation Behaviors and Attitudes at a Teaching-Centered University. *College & Research Libraries,* vol. 73, no. 4, pp. 349–365.

Simonsen, Jesper; and Toni Robertson (eds) (2013). *Routledge international handbook of participatory design. Routledge international handbooks.* London: Routledge.

Strauss, Anselm (1985). Work and the Division of Labor. *The Sociological Quarterly,* vol. 26, no. 1, pp. 1–19.

Strauss, Anselm L.; and Juliet M. Corbin (1998). *Basics of qualitative research. Techniques and procedures for developing grounded theory.* Thousand Oaks: Sage Publications.

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,* vol. 6, no. 6.

Treloar, Andrew; and Cathrine Harboe-Ree (2008). Data management and the curation continuum: how the Monash experience is informing repository relationships. In *Proceedings of VALA 2008, Melbourne,* February, vol. 13.

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,* vol. 169, pp. 191–198.

UK Data Archive (2014). *Qualitative data collection ingest processing procedures* (8th ed.).

van den Eynden, Veerle; Gareth Knight; and Vlad Anca. (2016). *Open Research: practices, experiences, barriers and opportunities.* Colchester, Essex: UK Data Archive.

Voida, Amy; and Elizabeth D. Mynatt (2006). Challenges in the analysis of multimodal messaging. In P. Hinds, and D. Martin (eds): *CSCW'06. Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, Banff, Alberta, Canada, 04 November - 08 November 2006.* New York: ACM Press, pp. 427–430.

Voida, Stephen; W. Keith Edwards; Mark W. Newman; Rebecca E. Grinter; and Nicolas Ducheneaut (2006). Share and share alike: exploring the user interface affordances of file sharing. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries and G. Olson (eds): *CHI'06. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montréal, QC, Canada, 22 April – 27 April 2006.* New York: ACM Press, pp. 221-230

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,* vol. 8, no. 7, e67332.

Wulf, Volker; Volkmar Pipek; David A. Randall; Markus Rohde; Kjeld Schmidt; and Gunnar Stevens (eds) (2018). *Socio-informatics. A practice-based perspective on the design and use of IT artifacts.* Oxford: Oxford University Press.

Yoon, Dongwook; Nicholas Chen; Bernie Randles; Amy Cheatle; Corinna E. Löckenhoff; Steven J. Jackson; Abigail Sellen; and François Guimbretiére (2016). RichReview++: Deployment of a Collaborative Multi-modal Annotation System for Instructor Feedback and Peer Discussion. In D. Gergle, M.R. Morris, P. Bjørn and J. Konstan (eds): *CSCW'16. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, San Francisco, California, USA, 27 February – 02 March 2016.* New York: ACM Press, pp. 194–204.

Zimmerman, Ann (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries,* vol. 7, no. 1-2, pp. 5–16.