SFB 882
From Heterogeneities to Inequalities

Social Research Data

Documentation, Management, and Technical Implementation at SFB 882

Stefan Friedhoff
Christian Meier zu Verl
Christian Pietsch
Christian Meyer
Johanna Vompras
Stefan Liebig

SFB 882 Working Paper Series ○ No. 16 ○ February 2013
DFG Research Center (SFB) 882 From Heterogeneities to Inequalities
http://www.sfb882.uni-bielefeld.de/

SFB 882
From Heterogeneities to Inequalities

Stefan Friedhoff, Christian Meier zu Verl, Christian Pietsch, Christian Meyer, Johanna Vompras, and Stefan Liebig

Social Research Data
Documentation, Management, and Technical Implementation within the SFB 882

SFB 882 Working Paper Series, No. 16
DFG Research Center (SFB) 882 From Heterogeneities to Inequalities
Project “Information and Data Infrastructure” (INF)
Bielefeld, February 2013

SFB 882 Working Paper Series
General Editors: Martin Diewald and Thomas Faist
ISSN 2193-9624

This publication has been funded by the German Research Foundation (DFG).

SFB 882 “From Heterogeneities to Inequalities”
University of Bielefeld
Faculty of Sociology
PO Box 100131
D-33501 Bielefeld
Germany
Phone: +49-(0)521-106-4942 or +49-(0)521-106-4613
Email: office.sfb882@uni-bielefeld.de
Web: http://www.sfb882.uni-bielefeld.de/

DFG Research Center (SFB) “From Heterogeneities to Inequalities”

Whether fat or thin, male or female, young or old – people are different. Alongside their physical features, they also differ in terms of nationality and ethnicity; in their cultural preferences, lifestyles, attitudes, orientations, and philosophies; in their competencies, qualifications, and traits; and in their professions. But how do such heterogeneities lead to social inequalities? What are the social mechanisms that underlie this process? These are the questions pursued by the DFG Research Center (Sonderforschungsbereich (SFB)) “From Heterogeneities to Inequalities” at Bielefeld University, which was approved by the German Research Foundation (DFG) as “SFB 882” on May 25, 2011.

In the social sciences, research on inequality is dispersed across different research fields such as education, the labor market, equality, migration, health, or gender. One goal of the SFB is to integrate these fields, searching for common mechanisms in the emergence of inequality that can be compiled into a typology. More than fifty senior and junior researchers and the Bielefeld University Library are involved in the SFB. Along with sociologists, it brings together scholars from the Bielefeld University faculties of Business Administration and Economics, Educational Science, Health Science, and Law, as well as from the German Institute for Economic Research (DIW) in Berlin and the University of Erlangen-Nuremberg. In addition to carrying out research, the SFB is concerned to nurture new academic talent, and therefore provides doctoral training in its own integrated Research Training Group. A data infrastructure project has also been launched to archive, prepare, and disseminate the data gathered.

Project “Information and Data Infrastructure” (INF)

By setting up and administering a virtual research environment, this project takes on key service functions in the domain of data and information infrastructure. It will devise and implement standards of documentation for the data generated, develop anonymization concepts for data archiving and provide advice on methodology. The project aims at constructing a virtual research environment composed of three elements:

– A conventional working platform, which provides necessary IT resources for the individual projects and for the Collaborative Research Center as a whole. This working platform offers effective support for research in the social sciences by bringing together various tools for daily work (e.g., storage and administration of documents and publications) and administration (e.g., project management, collaborative documentation and exchange of knowledge, appointment calendars, forum, blogs, etc.) in a single information infrastructure adapted to the specific needs of the projects and the working methods of the researchers.
– A research data platform, which constitutes an innovative aspect of the project. It combines elements of research data management with the further development of social science methodology. The platform provides services for the archiving and subsequent use of datasets and is responsible for the infrastructural and methodological coordination of the data documentation. The entire data lifecycle – from the conceptualization of the project to the analysis and archiving of data – is documented by implementing and applying the metadata standard DDI (Data Documentation Initiative) 3.x in all the projects participating.
– An interface module, which manages the external links to existing information resources (e.g., SOEP, the Federal Employment Agency, the Data Service Center for Business and Organizational Data at the University of Bielefeld) and other tools for exchanging data. For example, a smoothly working data exchange between heterogeneous information resources requires a variety of transformation functions to standardize terminologies (e.g., DDI elements) and data formats (e.g., output of diverse statistic programs). The interface module will also set up and administer the homepage of the SFB and ensure long-term external access to the data.

By extending the classic functions of compiling and servicing the information and data infrastructure, this project assumes developmental and advisory functions in the domain of IT infrastructure and empirical research methods. This implies that it makes its own contribution to the further development of research methods in the social sciences and to the construction of the infrastructure for data gathered using different methods (both qualitative and quantitative) and from different units (situations, persons, companies).

The Authors

Stefan Friedhoff is a member of the SFB 882 project “Information and Data Infrastructure” (INF) and PhD candidate at the Bielefeld Graduate School in History and Sociology (BGHS). In the INF project he assists researchers with problems of everyday documentation and improves documentation practices. His research interests focus on data management, survey methodology, social inequalities and factorial survey research.
Contact: stefan.friedhoff@uni-bielefeld.de

Christian Meier zu Verl is a member of the SFB 882 project “Information and Data Infrastructure” (INF) and PhD candidate at the Bielefeld Graduate School in History and Sociology (BGHS). His research interests focus on scientific practices (within the social sciences), methodology, and qualitative social research.
Contact: christian.meier_zu_verl@uni-bielefeld.de

Christian Pietsch is a computational linguist with a degree from Saarland University who previously worked at the Open University, UK, and at the Center of Excellence in Cognitive Interaction Technology (CITEC) in Bielefeld. At SFB 882, he helped create an IT infrastructure for research data management before moving on to work as a digital library developer at Bielefeld University Library, where he continues to facilitate open access to data and other publications.
Contact: cpietsch+sfb882wp@uni-bielefeld.de

Johanna Vompras is a member of the SFB 882 project “Information and Data Infrastructure” (INF) and is a researcher at the Bielefeld University Library where she works on research data management services. She is involved in the introduction and development of information infrastructures for documentation, archiving, and reuse of research data and their embedding into institutional services within the university.
Contact: johanna.vompras@uni-bielefeld.de

Christian Meyer is currently Senior Fellow at the “Centre for Global Cooperation Research” at the University of Duisburg-Essen. He is a sociologist and anthropologist and one of the directors of the INF project, focusing on qualitative research methods. His research interests include qualitative methodology, philosophy of science, science and technology studies, religion, politics and interaction.
Contact: christian.meyer5@uni-bielefeld.de

Stefan Liebig is Professor of Sociology with a special focus on Social Inequality and Social Stratification at the Faculty of Sociology, Bielefeld University, and one of the directors of the SFB 882 project “Information and Data Infrastructure” (INF). His research interests are methods of empirical research, empirical justice research, and organizations and social inequality.
Contact: stefan.liebig@uni-bielefeld.de

Social Research Data

Documentation, Management, and Technical Implementation at SFB 882

Stefan Friedhoff, Christian Meier zu Verl, Christian Pietsch, Christian Meyer, Johanna Vompras, and Stefan Liebig

Abstract

This paper is a contribution to the methodological and technical discussion of social research infrastructure. The main question is how to store and manage data in a way that meets the increasing demand for secondary data analysis in both quantitative and qualitative social science research. The first two sections focus mainly on aspects of data documentation, in particular on the unification of various documentation requirements that have arisen across ongoing projects of the SFB 882. While the aim of documenting quantitative research processes is to ensure replicability, the aim of documenting qualitative projects is to maintain the understandability and informative value of research data.

In the third section a virtual research environment (VRE) is presented that provides both a generic work platform and a project-specific research platform. The work platform bundles IT resources by bringing together various tools for administration, project management, and time- and location-independent collaboration in a single environment adapted to researchers' specific work processes. The research component combines data management with further developments in social science methodologies. It provides services for the archiving and reuse of data and enables the infrastructural and methodological coordination of data documentation. We also introduce a documentation scheme for qualitative and quantitative social research within the SFB 882. This scheme considers the specific requirements of research projects within the SFB, such as different methods (e.g. panel analysis, experimental approaches, ethnography, and interview research), project work, and requirements of long-term research.

Keywords: reflexivity, replication, data documentation, informative value, documentation practices, quantitative research, qualitative research, data management, reproducible research

Introduction

Two important criteria for conducting social research are replicability and reflexivity. Qualitative social research usually centers on reflexivity while quantitative research focuses on replicability (see Hammersley 2007). A distinction must be made between demands in relation to the research *process* and those in relation to the *publication* of understandable results. Whereas the research process (e.g. developing questionnaires, conducting interviews, etc.) should be made transparent for third parties, the published results must be replicable and understandable. These quality criteria for good scientific practice are achieved, inter alia, through the good management and documentation of data collected and computed. This paper addresses several problems of data management and documentation in qualitative and quantitative social research as well as looking into technical solutions designed to assist social researchers.

During the past few years an increasing sensitivity has emerged to the importance of data documentation and availability, especially in economics and the social sciences. Most recently, Huschka and Wagner (2012) have shown that data documentation and its availability are requirements not only for good scientific practice but also for good methodological work.

Many researchers, however, still lack experience with data management and documentation. Since funding organizations and professional associations are increasingly expecting researchers to share all scientific outcomes (including data) from publicly funded projects, data management and documentation have become important cross-disciplinary issues. Even though several institutions have called for the professionalization of research data management, precise knowledge of effective data management is still missing. However, in contrast to other disciplines, social sciences and humanities are confronted with specific challenges that currently prevent effective data management and documentation. These challenges include, e.g., privacy issues and the particular properties of qualitative data (non-standardized, context-sensitive, etc.). The aim of this paper is, first, to describe the main challenges of data management and documentation in a large-scale social science research project in which different types of data are generated, and second, to introduce possible solutions.

In qualitative social research, standardized methods of research and data documentation are especially critical. The concept of "datum" itself remains in dispute, so that the relevance of contextual knowledge for the understanding of research documents is still unspecified. Both await further elaboration. In addition, varying notions of context are used by different researchers depending upon their respective research paradigms, questions, methods, etc. For the needs of the SFB projects, the information and data infrastructure project (INF) has developed a documentation scheme based, on the one hand, on methodological considerations and, on the other, on observations of *in situ* documentation by researchers. This documentation scheme and its development will be presented in the first section.

Regarding quantitative social research, the lack of a user perspective in current data management strategies impedes the resolution of the problems of granularity and acceptance. By taking up the user perspective, we highlight these problems and show several steps that may help to relieve the (currently external) pressure on researchers by allowing them to document their work as part of their own workflow. The documentation scheme for quantitative projects will be presented in the second section, along with the reasons it is needed and the knowledge that can be gathered by talking to researchers about the way they work and manage their data.

Even though computers are as common as pen and paper today, most researchers are ill-equipped for conducting research in a way that makes repetition easy. Most of them use office software that was never meant to be used in research. Moreover, the user interface of most current operating systems, with its underlying desktop metaphor, do not support research data management, sharing, or archiving at all well. This is where the technical part of the INF project comes into play: our main tasks are to provide advisory and developmental services in the domain of information infrastructure. We aim to analyze the broad diversity of working processes across the projects, collect researchers' requirements, and use them as the basis for setting up a "Virtual Research Environment" (VRE) as part of the information infrastructure. The general aim of the VRE is to collect the projects' output, shorten communication lines, optimize organizational workflows within research groups, and upgrade research work by facilitating data documentation and data reuse. In the third part of this paper, we will discuss various key technical research collaboration issues arising from a user requirements analysis, and introduce the main aspects and components of the research platform.

1. Qualitative Archiving: Some Methodological and Practical Insights

In the field of qualitative social research, researchers are confronted with the challenge of documenting as text the data they have produced. In line with their interpretative orientation, this form of documentation should not only provide the necessary context for a better understanding of the data, but also offer information that allows third parties (especially other researchers) to assess their meaningfulness. In order to ensure an effective archiving system that fulfills this need, it must be decided what information is needed to allow further use of the data and what information can possibly be omitted. This process of selection requires, firstly, anticipated knowledge of possible use cases and, consequently, also of possible scientific trends still unknown during the actual research. Therefore, it will be expedient to document every research project as broadly as possible so that as many different secondary analyses as possible can be conducted. Secondly, the expectations of different researchers with regard to data documentation are not congruent, since their theoretical approaches, research questions, interests, etc., differ. In the field of qualitative social research it is almost impossible to predict all the potential secondary analytic uses, and they are mostly untested (see, e.g., Heaton 2008, Gläser & Laudel 2008). From this perspective, data documentation is still highly problematic and remains explorative. Based on the criteria of "understandability" and conserved "information value," we now discuss **(1)** theoretical requirements for appropriate documentation of the entire research process, and **(2)** empirical practices of documentation within the research process. Subsequently we outline **(3)** the scheme we have developed for the documentation of qualitative data, and **(4)** ongoing and future challenges for improving the documentation practice in the SFB 882.

(1) Before we discuss possible problems in conserving the understandability and informative value of stored qualitative data, we will first address some basic features of the qualitative social research paradigm, as these fundamental theoretical orientations strongly influence the qualitative research process. That research process is distinguished from quantitative social research by a high degree of reflexivity, low standardization, circularity, and parallelism (see Bergmann 2006, Flick 2011, Kalthoff et al. 2008, Silverman 2007, Strübing 2007).

Bergmann (2006) has outlined six common characteristics, which possibly apply for all the approaches that count as qualitative social research. His list can be used as a basic orientation to develop a documentation scheme for qualitative social research.

(a) *Data thickness.* An important motive of qualitative research is the non-reductive description of social phenomena in a way that is adequate and preserves their meaningfulness in the social world. Qualitative researchers try to preserve ambivalences and dynamics of the phenomena under study by keeping the documents, and data representing them, "rich" and "thick" through contextual referencing and exhaustive description. This stands in contrast to quantitative social research, which isolates and reduces social phenomena to countable units (Bergmann 2006: 17). (b) Accordingly, a fundamental *context orientation* of qualitative research unfolds its analytic value by referring to data thickness. For example, a particular utterance within an interview, such as the answer "no" to a question about the social integration of a person, might be relativized by further utterances or through the triangulation of the interview with observations in field research (Suchman & Jordan 1990). While the scope of context is fundamentally variable, in qualitative research meaning is generally captured by the contextual embedding of individual social phenomena. (c) As a consequence, qualitative research gains its concepts of description out of the *exploratory*, thorough, and in-depth investigation of social phenomena, which includes an oscillation between diving into the phenomena under study and its distanced analysis (Bergmann 2006: 19). The standard linear process of data collection, analysis, and theory construction appears inapplicable to the qualitative research process. (d) It is, furthermore, appropriate to conduct qualitative research by circular movement or parallel performance because of the *theoretical ambition* of qualitative research. (e) That said, qualitative social research focuses on *individual cases*, which are explored *exhaustively* in regard to structures of meaning and generative mechanisms of social phenomena. Thus exhaustivity means that analytic descriptions of research topics always include descriptions of ambiguities, contraries, and, importantly, deviant cases. The researcher mobilizes his participant experiences in the field in order to deal with these conflicting descriptions. (f) Finally, Bergmann (2006: 22–24) emphasizes three levels of *reflexivity* in qualitative social research (see also Lynch 2000): (i) The data are shaped by the researcher, so the researcher always encounters him- or herself within his/her data to a certain degree. (ii) In Schütz's (1953) terms, social science constructions are second-order constructions, which build on the mundane first-order constructions of the actors under study. (iii) In the sense of Garfinkel (1967), the actions of social actors are themselves

reflexive, since their performance is designed right from the start in such a way that they can be recognized and interpreted by third parties (such as mundane actors and researchers) as *doings* of a specific kind.

Against this methodological background, the documentation of the entire qualitative research process appears as an important challenge. Social researchers are confronted with the difficulty of preserving their data in a way that conserves their complexity, interconnectedness, and reflexivity and thus, equally, conserves their understandability and informative value for third parties (Walters 2009: 317). Decisions must be documented with respect to the selection of research place, time, people, and material as well as selections of interview questions, focused observations, theoretical considerations, contexts of data collection and analysis, or relations between different kinds of data (e.g. field notes, audiovisual recordings, interviews, and collected documents).

An international standard, DDI, has recently been developed for data documentation. This standard is becoming increasingly common in data archives and among researchers in Germany (e.g. GESIS and DIW). DDI is an XML-based standard that targets the recording of the entire research process. Although it will be extended in the near future, the documentation using DDI currently focuses on quantitative research processes, and is therefore based on the model of a linear research process, isolable data, and the use of homogeneous sorts of data. As a result, DDI does not fulfill the specific requirements of qualitative research. The features of qualitative social research mentioned above (temporal parallelism and circularity, reflexivity, theoretical motivation, orientation towards individual cases) and the consequences of data collection (heterogeneity and complexity of data designed for the analytic and exhaustive understanding of individual cases and the fundamental importance of deviant cases) still await integration into the DDI model. At present, a group within the DDI Alliance, including the INF project, is working on this task.

(2) As part of the demand analysis, the INF project observes projects of the SFB 882 that apply qualitative methods, so that, hopefully, a wide spectrum of explicitly methodologically driven, but also tacit and contingent, decisions and selections will be identified in regard to the research process and its data collection.[1] Taking research practice into consideration, it is

[1] We do this by drawing on the established tradition of ethnography in scientific practice, developed on the

obvious that researchers do in fact already document their data. Through this process, the difference between data (in the strict sense of the term) and researchers' documentation of them becomes blurred (e.g. in field notes in which observational and contextual descriptions are mixed with theoretical ideas and methodological reflections). Empirically, two forms of documentation can be distinguished: forms of *in situ* and *ex situ* documentation.

By *in situ* documentation, we mean the situative utilization and application of data at a certain point of time, place, and complex of meaning (similar to the social science term contextualization). Thus, the *in situ* documentation emerges within specific situations. This kind of documentation practice is part of an even wider reenactment or (re)formulation within the research process which preserves, deletes, and transforms data sections as required (Heritage & Watson 1979: 129). In general, the meaning of an utterance is situatively produced within an interaction in order to serve the aim of its immediate intersubjective application. The same is true for data within everyday research life. *In situ* documentation follows a specific purpose that is bound to the situation and its context. It is also part of negotiation processes and thus contingent upon and subject to the restrictions of social interaction among co-present users. From a phenomenological perspective, knowledge of the research subject accumulates with the researcher. Only in specific situations with particular intentions will it be reconstructed and uttered and thereby made transparent for other researchers. Examples of this kind of situation include analysis and interpretation sessions or lectures, and also mundane day-to-day conversations between researchers. During these interactions, data are used for direct application within a concrete situation. This method of documentation is shaped by particular situations and is therefore contingent upon them: the documentation of data varies depending on the researchers present, their research traditions and methodological and theoretical orientations, or the current research questions of each social researcher. Different purposes are pursued according to the situative configuration, and the understandability and informative value of the data discussed is thus produced interactively.

Ex situ documentation, in contrast, separates the collection from the utilization of the data spatially, temporally, or personally. This implies that researchers always document *ex situ* for third parties. This separation leads to documentation practices that take place exclusively via

example of the natural sciences (Latour & Woolgar 1979, Knorr Cetina 1981).

written communication. Here, all kinds of applications and possible recipients must be anticipated and integrated into the textual documentation so that all imaginable third parties can use the data for their purposes of research. A kind of understandability must be produced without the interactional feedback loops that are present in *in situ* documentation practices.

An example of each documentation method should help to clarify the distinction: (A) A researcher contextualizes his or her thick data by discussing the context of their origin before starting the interpretation session. He relates this context of origin to the current research question so that the data become documented (contextualized) *in situ* and are processed appropriately by the audience. Hence the data are rendered utilizable so that everyone present is able to pursue the situative purpose. In this sense *in situ* documentation is always a reasonable documentation practice within its particular situation. However, it hardly makes sense beyond that concrete situation. (B) *Ex situ* documentation is always necessary when some third party wants to use the data for secondary analysis.[2] In long-term research projects such as the SFB 882 (with a maximum funding period of 12 years), there is a constant need for *ex situ* documentation for those researchers who did not participate in the initial data collection. They should be able to understand these data on the basis of their documentation without direct interaction throughout the whole funding period.

This distinction between *in situ* and *ex situ* documentation, heuristic and provisional as it is, enables us to locate different practices of documentation within the research process. Our assumption is that the ongoing ethnographic research on the qualitative research practices of the SFB 882 will allow us to identify different methods of *in situ* documentation. Furthermore, we assume that these insights will be useful for the development of *ex situ* documentation schemes (such as DDI). Firstly, the observation of *in situ* documentation could tell us something about the relation between possible consecutive usages of data and required documentation methods, and about possible useful methods of formal contextualization. Secondly, an integral view of documentation well founded in the qualitative research paradigm must necessarily be based on day-to-day practices of research. In this way, the documentation scheme remains close to daily research routines yet, through the identification of formal procedures, is abstract enough to represent generalizable mechanisms. Thirdly,

[2] This kind of division of labor between data collection and analysis (quantitative social research) is still uncommon for qualitative social research, for the reasons mentioned (data thickness, context orientation, and reflexivity).

integration in the sense of a bundling of practices of *in situ* documentation would be insufficient. Data thickness, context orientation, and reflexivity also refer to the tacit knowledge of the research phenomenon that the researcher necessarily acquires. This kind of knowledge is indispensable for a detailed understanding of qualitative data and their documentation, so it must be explicated in order to allow further understanding and utilization of the data. An explication cannot be produced merely by bundling *in situ* documentation; it requires the thorough study of those research practices that make something that is socially defined as "scientific data" out of mere mundane observations or question-answer play.

(3) Thus, the documentation of qualitative data must offer plenty of space for parallel and circular processes, for the layering of meaning, and for steps and levels of contextualization so that data can be archived without losing their understandability and informative value.

Currently we are developing a scheme for the documentation of qualitative data within the SFB 882. For the textual documentation of qualitative social research we propose four parts: (I) generic information about the study, (II) stages of fieldwork (II.1 before, II.2 during, and II.3 after fieldwork), (III) information and reflections on the data (memos), on the potentially complementary (or contradictory) relationship between the individual pieces of data, and on the data itself, (IV) information on the analysis and further manipulation of the data (e.g. transcripts, coding).

Part I gives an overview of the study and contains barely more information than the project proposal (including central research question, purpose of the study, short description of the project, literature to be published after conducting the study). Part II contains all information on the stage(s) of fieldwork. Information on sampling, establishment of field contact, and designated methodical approach can be found in subsection II.1 (see Table 1).

Table 1 - Before Fieldwork (Quali. Data)

Before Fieldwork	Content of Inquiry*
	Method(s) & Approach(es)*
	Research Question(s) & Theoretical Assumptions*
	Scheduled Method(s) of Inquiry*
	Plan of Data Management*
	Criteria of Field Sampling

	Changes of Field Sampling
	Communication w.r.t. Field Access
	Evolution of Inquiry Instruments
	Memos of Field Access

* = boxes must be filled out

The procedure for carrying out fieldwork is documented in subsection II.2 (see Table 2),

Table 2 - During Fieldwork (Quali. Data)

During Fieldwork (per Field)	Target Audience*
	Place of Field Contact(s)*
	Time & Place of Field Contact(s)*
	Unit of Study*
	Kind of Sampling*
	Object of Study*
	Study Area (geographical)*
	Preparative(s)
	Method(s) Used
	Type(s) of Inquiry Method*
	Technologies Used

* = boxes must be filled out

and information about steps taken after the fieldwork will be documented in subsection II.3 (including central outcomes of the fieldwork; see Table 3).

Table 3 - After Fieldwork (Quali. Data)

After Fieldwork	Internet Address (project)*
	Record of Field Contact(s)
	Central Finding(s)*

* = boxes must be filled out

It is important to cover all parallel and circular steps of the research process by documenting each research field individually. Our model therefore makes it possible to document more than one data set for a single study by multiplying the subsections of Part II. Part III (see Table 4) documents the complexity and interdependence of the individual data, memos about data, and the relations between them. In other words, every piece of data is stored in its raw

form and documented with regard to how it is embedded within the context of the research as a whole and how it is related to other data. Part IV (also in Table 4) collects all kinds of analyses and manipulations of enriched data. For example, one field note will be represented in different stages of its emergence from a handwritten entry in a diary or a digital text file to an anonymized text, etc.

Table 4 - Thickness, Manipulation & Analysis (Quali. Data)

Thick Data	Primary Data 1 till n*
	Place / Time of Origin, Object of Study of Data*
	Memos of All Primary Data*
	Further Memos (e.g. methodological) & their Relation to Primary Data
	Report(s)*
Manipulation & Analysis of Data	Codes / Categories / Development of Codes / Paraphrases*
	Memos of Development & Explanation of Code Choice / Paraphrases
	Transcript(s) of Primary Data (if provided)*
	Anonymized Transcript(s) of Primary Data (if provided)*
	Arrangements for Anonymization*
	Memos of Interpretation of Every Analysis
	Theoretical / Analytical Memos
	Analyses of Primary Data*
	Method(s) of Interpretation*
	Convention of Transcribing*
	Software Used *

* = boxes must be filled out

(4) Our scheme of data documentation for qualitative data is still under testing. It is being used and modified by the qualitative projects in the SFB 882. The form of documentation is designed for different groups of recipients: (a) for current internal use, (b) for future internal use (by subsequent project staff), and (c) for external use (by third parties and for secondary analysis). The projects will provide specified collections of information for all three of these groups at the end of the funding period.

Mixed-method research documentation is a further issue within the SFB 882. Several research projects are currently working with qualitative and quantitative data. These projects make different demands on documentation than do “pure” qualitative or quantitative approaches.

Bryman (2006) identified five types of mixed-method approach, which combine qualitative and quantitative data in different ways. He distinguished approaches (1) by temporal order of collecting qualitative and quantitative data (simultaneously or sequentially), (2) by priority (what counts more?), (3) by function of integration (e.g. triangulation, explanation, or exploration), (4) by stage of research process, and (5) by data strand (how many research methods were applied and how many sources of data were used?) (Bryman 2006: 98). The documenting of mixed-method research is an unresolved issue, and must additionally incorporate these specific characteristics.

2. Quantitative Archiving: Changing the Perspective for Improving Documentation

Regarding the documentation of quantitative social research, many of the questions mentioned in the previous section have already been answered. There are already publications (e.g. Long 2009) and "best practices" for data management (e.g. Büttner et al. 2011) in quantitative research processes. There is no doubt that it is crucial for any researcher to document their work no matter what method is used, but one general problem for researchers with regard to data management and data documentation is the perspective from which most research in this field is conducted. As most research on aspects of data documentation is conducted by data librarians or people working in data archives, there is almost no research that includes the user perspective. Yet inclusion of that would be helpful, as the researchers are the experts in adapting everyday data documentation practices. To achieve high acceptance and solutions that are close to the way researchers work, it is essential to use this kind of knowledge to optimize documentation. This is why we strive to assume the user's perspective when addressing the problems with the current state of data documentation.

The problem of granularity

The main question around the problem of granularity is: How detailed must documentation be to be sufficient for someone to replicate the conducted research? This problem must be divided into two aspects: the problem of granularity itself stems from the problem of what to document in general. Despite the high degree of standardization in quantitative research, there is no easy answer to this question. That documentation must fit the requirements of journals (e.g. *Social Research & Methods*) for publication is not in question, but it leaves unresolved

the need for further information on the research process that is not required by journals but that enhances data quality. One possible solution is to seek a tool that helps to document work and look at the possibilities it offers for assisting data documentation. One of these tools could be DDI as a metadata standard. As we decided that the method of documentation must comply with DDI3 standards, this initially appears promising. But looking more deeply into the possibilities that DDI offers for documentation of the research process, we see that it is almost impossible to gather all this information during the research process without spending a large amount of time that could otherwise be used to conduct the research itself. It also leaves unanswered the question of what information is important and how it can be easily obtained and documented in a standardized way. DDI is a modular approach; it is not reasonable to use all of DDI's features. If the documentation aims to be comparable to other people's work, it will need to find a way to limit the information gathered. This limit must be set by discussing with researchers what they believe is important information regarding their research process and what information data archives or any secondary user will require. Working out what is relevant information that requires documentation means asking the researchers. Their answers will differ from researcher to researcher, which is why it is essential to discuss the question with as many researchers as possible and gather enough information to find common ground from which to start.

Several steps have been taken by the INF project to solve the granularity problem, with varying results. Since one of the interests of researchers is to keep their workloads as light as possible, they are keen to have a much weaker standard than any other party managing or creating the documentation (e.g. data archives), so speaking with researchers without elaboration has offered little or no advance regarding what should be documented. Since DDI is to be used as the metadata standard, it also seemed logical and worthwhile to look at what can be documented using this standard; however, because DDI offers almost infinite possibilities for documenting research processes, this does not solve the problem either. In the end we focused on both the minimum and optimal data documentation requirements for data archives. The selection process started with a list of requirements from the Data Service Center for Business and Organizational Data (DSZ-BO) at Bielefeld University, as five of the SFB 882 projects will submit their data to this data center. We added some items to their list by looking into the requirements of GESIS (http://www.gesis.org/) and the UKDA (http://www.data-archive.ac.uk/), deleted some items that were not necessary in our context (e.g. organizational criteria that would only apply to those projects that conduct research in

organizations, information that is constant between projects), and by doing so developed one combined list of what should be documented by all projects within our collaborative research center. This sheet is divided into several topics that accompany the research process. For every topic, there are fields with information for the researchers to fill in and instructions on the format of the given field. As not all fields can be completed by every project, we marked the mandatory fields and left the other fields as optional.

The first main topic is "General information on the study," which covers all relevant information on the study that does not change over time (except potential fluctuation in the scientific staff). This section of the documentation is almost identical to the documentation sheet for qualitative working projects (see Table 5). The next section covers general information on the data collection such as the theoretical background, methods of data collection, and sampling methods. The first two topics are to be answered by the researchers even before the first piece of information is gathered in the field. The next two topics concern the fieldwork, which in most of our projects takes the form of a survey with a pretest. The fields are similar on both topics, with slight changes according to the evaluation of the pretest. Table 5 shows the part of our documentation sheet relating to the pretest:

Table 5 - Quanti. Data

Topic	*Field*	*Guide*
Pretest	Realizing instance (Survey)*	Name (e.g. institute)
	Population / Target group / Researched entity*	Name / Description
	Area of research (geographic)*	Address/Geographic region
	Duration / Timespan*	Date (From - To)
	Sample (Planned size)*	Number
	Sample (Realized size)*	Number
	Method of sampling*	Name of method or description
	Measuring instance*	Name/s
	Survey preparation	Description
	Measures to increase response rate	Description
	Research methods used *	Description
	Type of survey method*	Name/s
	Technologies used	Name/s
	Time measurement data available*	Yes/No
	Other meta data (IP-Addresses, Addresses or	Kind of data

	similar)	
	Method used for evaluation of pretest (e.g. Think Aloud)*	Description
	Final questionnaire*	Reference to documents

All fields marked with an asterisk are mandatory while the other fields are optional, as some projects, for example, do not take measures to increase the response rate. The next two topics are post data-collection and preparation. While the post data-collection field is intended to gather all the information on results, reports, and data usage, the data preparation topic covers all manipulation of the raw data collected. Our sheet closes with three topics on different syntax files for the preparation, generation, and analysis of the raw data.

Table 6 - Quanti. Data

Topic	*Field*	*Guide*	*File 1*	*File 2*	*File 3*
Syntax Data III: Analysis (Once per Syntax File)	Author/s*	Name/s			
	Date of last change*	Date			
	Task*	Description			
	Software Version	Program and version			

As can be seen in Table 6, for every syntax file the researchers are required to name the author, the date of the last change, and the purpose of the file. These fields are different from the other topics as they are required for every syntax file. While, for example, the data on the study only has one possible value, it is likely that one project has several files for different analysis. All of these files should be documented by including this information.

However, this data documentation sheet led to a legitimization problem when we tried to simply tell researchers, without further explanation, that this was the list of the items to be documented. The best solution so far seems to be a combination of several practices: Gather the researchers' proposed documentation topics and aspects, combine these with what is required by the archives, and define a mapping to DDI. Any additional piece of information that should be documented but was not mentioned by the researchers has to be legitimized with them by showing them the added value of documenting the additional item. Also helpful

is further information on the worst-case scenarios of documentation gaps and positive examples of the time that can be saved in the long run by using proper documentation. By combining theories, we were able to optimize the list of what needs be documented.

The problem of acceptance

Another major problem that we face is that of the researchers' acceptance or understanding of the need for complete documentation of the research process. Even with the best data management plans, thorough documentation will never be achieved if the researchers do not see why it supports and eases their work. To deal with this issue, it is necessary to talk to researchers about what they think is worth documenting and what is not. It is also important to explain and clarify the ways in which they can profit from thorough documentation. Showing the advantages (e.g. easier access to all relevant information when writing a paper, easier integration of potential new co-workers in the future), on the one hand, and describing the problems that can arise from insufficient documentation (e.g. potential loss of knowledge, time-consuming searches caused by unstructured file management), on the other, might help to solve the acceptance problem. As part of addressing this, we conduct group discussions with all projects in our collaborative research center on how the researchers work, how they document their data, and what they believe is necessary in order to replicate their research process. For this purpose, we created guidelines to ensure that the talks cover all relevant topics for optimizing documentation practice and the projects' coordination of their work. During the group discussions there is a chance to talk about specific problems relating to the projects and their ways of managing data. One of the topics discussed is the list with the proposed data documentation guideline. As this list is meant to be a first draft of our individual solution to the granularity problem, we need to gather information on problems with the list that emerged during the adaptation of our proposed way of documentation. All these sessions are recorded and made into short abstracts on how the projects work and where their specific strengths and weaknesses lie in terms of their current methods of data management. These short papers are intended as a kind of individual data management plan for the projects and a starting point for an ongoing discussion and optimization process. Another positive side effect of the talks is getting in touch with all researchers in the research center and showing them how we can be of assistance for their everyday work.

The essence of these discussions so far has been that although all projects have their own methods of managing data, there are some common problems regarding the documentation

itself. The main problem is the lack of a central file or something similar to combine the diversely stored data in one place, ensuring that another person will understand the logic of the data management. Our requirements file—despite its intended usage as a central location to organize dispersed research documents—has often been used as a guide to what research information is worth documenting. Another finding is that researchers are well aware of the importance of data management, especially in projects where they are based in different locations. These projects benefit from sound data management because it eases communication between project members: everyone knows where to look for specific information. Since some projects mentioned that they had a way of managing their data but not enough time to ensure it was adhered to, it would be useful to have some on-the-fly mechanisms to document and manage data during the research process without much effort. However, we have not yet systematically analyzed what the next step should be with regard to these discussions.

As a way of increasing attention to data documentation, there have been several presentations for all researchers participating in the collaborative research center on the usefulness of data documentation and the way it is carried out. As one of the central projects within our research center, we are constantly reminding researchers of the importance of data documentation—not only by our existence but also by talking to them about current problems and their ways of documenting their work.

3. Data infrastructure

> We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them. Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated coincidence, but with events which, on account of their regularity and reproducibility, are in principle intersubjectively testable. (Popper 1959: 45)

The wave of digitization that came with computers has improved the situation compared to Popper's days, but digital availability does not guarantee sustained access to what is in the data files. We have already discussed the issue of documentation, but what about the data themselves? Will they be readable 5, 10 or 50 years on? Even if the files are stored on a reliable medium, some formats are known to age much faster than others. For instance, there is the well-known problem

that current versions of Microsoft Office are unable to read certain file formats used by older versions of Microsoft Office. Thanks to the re-engineering efforts of the free software community, those legacy file formats can be read by free and open-source software such as LibreOffice. Users of less widespread proprietary software such as specialist research software might not be so lucky, losing their data forever when one company goes out of business or decides to change a file format in a backward-incompatible way (vendor lock-in). The only solution is to use open formats that are supported by open (and often free) implementations.

Experience has shown that documenting as an afterthought does not work. It is vital that researchers document as they go (*in situ*): Donald Knuth pioneered a method and created appropriate tools for what he termed Literate Programming (Knuth 1984), allowing a programmer to write software and corresponding documentation at the same time. This idea was later suggested for other disciplines as well, but did not catch on to a great extent. Nevertheless, the idea lives on in a small but determined group of scientists who push the vision of "reproducible research" (Fomel & Claerbout 2009). One relevant proposal for the social sciences is called Literate Statistical Practice (Rossini & Leisch 2003). It involves using tools such as Sweave, which allows researchers to interleave code written in the statistical programming languages S or R with documentation written in LaTeX, a professional and free typesetting markup language popular among scientists. Unfortunately, these tools have a steep learning curve. Creating intuitive user interfaces for them remains a challenge.

However, as long as they use computers to conduct their research there is hope even for busy researchers who do not find the time to learn how to use new software tools, and for those who cannot set aside any time for documenting. For instance, using a VRE or any other Web application usually generates traces of users' activities in the form of log files. Such data could be used for creating rough outlines of lab diaries (with exact timing information) and other forms of documentation. The most difficult part in implementing this would be to ensure researchers' privacy. The minimal requirement for such data collection activities would be informed consent, i.e. an opt-in process. Even then, researchers should be regularly notified of the data gathered about them, and be given the chance to delete certain data points without compromising the validity and accuracy of the data. Currently, we do not plan to implement this.

INF: Data Infrastructure within the SFB 882

In tune with many funding bodies in leading research countries, the German Research Foundation (DFG) has recently formulated stricter requirements regarding data management. In particular, it is now mandatory to make primary research data available to the research community for at least 10 years after the end of project funding. As a consequence, the INF project is an integral part of the SFB 882, providing advisory and developmental services in the domain of information infrastructure. The aim is to analyze the broad diversity of working processes across the projects, collect researchers' requirements, and use them as the basis for setting up a Virtual Research Environment (VRE) as a part of the information infrastructure. Additional aims are to collect the projects' output, shorten communication lines, optimize organizational processes within research groups, and upgrade research work by facilitating data documentation and reuse. Other core aspects in the INF project are data archiving and long-term preservation of the data generated.

In the case of the SFB 882, the VRE itself will combine both general work and project-specific research tools on one platform:

(1) The **work platform** will bundle IT resources by bringing together various tools for administration, project management, and time- and location-independent collaboration in a single environment adapted to researchers' specific working processes.
(2) The **research component** combines data management with further developments of social science methodologies. It will provide services for archiving and reuse of data sets and is responsible for the infrastructural and methodological coordination of the data documentation.

In the following section, we will discuss various technical key issues of research collaboration as an outcome of the user requirement analysis, and introduce the general components of the research platform.

Virtual Research Environment (VRE)

There is no one-size-fits-all solution for a Virtual Research Environment (VRE) for social sciences that would support all possible working and research processes and all data types. This is confirmed when we look at the UK research foundation JISC, which recently introduced the following general definition of a VRE: "A VRE helps researchers from all disciplines to work

collaboratively by managing the increasingly complex range of tasks involved in carrying out research."[3] According to this definition, a VRE is not a standard piece of software but rather a collective term for context-dependent and discipline-specific tools and technologies needed by researchers to do their research, to collaborate, and to make use of other resources and technical infrastructures in (preferably) one working environment. In the case of the SFB 882, the general tasks to be fulfilled with the VRE are summarized in Figure 1.

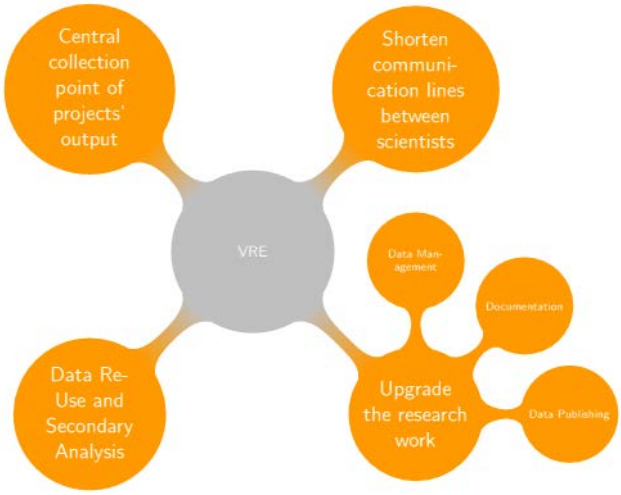

Figure 1: Aims of a Virtual Research Environment in the scope of the SFB 882

Requirements Analysis and Results

To discover the requirements of the interdisciplinary subprojects, at the very outset a requirements evaluation was performed by the INF project to collect and evaluate researchers' working procedures in terms of communication, data management, and requirements for a subsequent data archiving. For this purpose a survey was answered by each of the subprojects, revealing how the researchers involved have worked until now, how they plan to work in future, and what tools and technical frameworks they need. Specific topics around data management or types of data processed in the projects were examined afterwards, systematically and in detail, during personal meetings with the subproject members.

The results of the requirements analysis can be divided into general social science-related and project-specific requirements. They included the following areas:

[3] http://www.jisc.ac.uk/whatwedo/programmes/vre

1. **Collaboration:** Allowing multiple researchers in separate locations to share a common view of the project-specific workspace and to work together on documents and texts stored in the VRE. Another requirement is to support group-based video conferencing.

2. **Data archiving:** As already mentioned, it is a requirement of our funding institution DFG to archive research data for at least 10 years. For a variety of reasons, including Popper's plea, all empirical researchers—independently of such formal requirements—should have a strong self-interest in data archiving as a prerequisite for repeatable research. Repeatable or replicable research is known under many names. In our context, it is instructive to take a closer look at re-analysis and reuse:

   Re-analysis: According to a recent review (Gómez et al. 2010), there are two distinct usages of the term "re-analysis": 1) As mistakes can always happen, a repeated data analysis by a person other than the original author(s) using the same data and same methods can be useful for verification. 2) Data analysis methods continue to evolve, and there are often several options for analyzing a given data set. The original authors may not have had the time or expertise to take into account all relevant methods of data analysis. Therefore, it is very important to preserve a study or research project in such a way that interested researchers can re-analyze the data at a later date.

   Reuse and reproducibility: For many disciplines, empirical data can be reused to shed light on new research hypotheses. The central requirement is that research results should be reproducible at any time.

   Fortunately, all these methods of doing repeatable research place similar demands on a Virtual Research Environment. In particular, we have implemented the following features:

   - ***Data safety:*** File system storage is provided by enterprise-grade redundant storage systems connected to the VMware cluster that hosts our Linux servers. The database (MySQL) underlying our Drupal-based research environment and the local files are backed up nightly on tapes stored in two physical locations at a safe distance apart (Bielefeld and Aachen).
   - ***Data security:*** Before backups are copied to tape, they are encrypted with GnuPG in the OpenPGP format. Sensitive user home directories on the Linux servers are

encrypted using eCryptfs. Laptops used in the field employ full disk encryption provided by TrueCrypt.

- ***Data conversions:*** We automatically convert proprietary data formats to open formats[4] and archive both the pristine original and future-proofed converted versions.
- ***Assigning persistent identifiers:*** We have created a web application that allows researchers to register their data sets with DataCite. DataCite then issues a DOI (digital object identifier). This makes a data set citable as an independent publication and is conducive to long-term availability.
- ***Data sharing within research groups:*** Collaborative writing within a WYSIWYG editor and forum functionalities are supported. Furthermore, various data types can be uploaded into the VRE and shared within the research group, for example to be annotated or commented on.
- ***Versioning to synchronize documentation and data:*** Versioning functionality is now supported for the wiki entries. Thus, the users can visually compare two selected versions and roll back to former versions of their contents. All data we collect is also time-stamped. This makes it possible to synchronize documentation and data.

3. **Data Documentation:** Data documentation should explain how data is created, by whom it is created, the structure of its content, and its meaning. In order to make further research efficient and the subsequent intermediate steps comprehensible, it is also important to agree and follow "best practices" for data organization—before data are created.

Data documentation is supported by the INF project at several levels. First, we give advice to researchers on best practices for data-level documentation. This covers descriptions that may be included within the actual data (e.g. in statistical files, predominantly Stata, SPSS, and R). In many existing statistical software packages, variables, data types, or missing values can be documented in "Variable View" or via syntax files. Furthermore, we raise awareness among researchers of the need to consider the documentation standard—in order to make research data machine-readable and machine-processable—and give guidance on how to generate and

[4] We prefer to archive data in formats for which a free implementation exists. See the GNU project's website at http://www.gnu.org/licenses/ for a list of free and open-source licenses. In practical terms, the existence of free, open-source, and portable software is more important for long-term preservation than the existence of a written standard documenting the format.

use it. The current technical support for generating DDI is as follows: We provide DDI-based templates (see Table 1 to Table 6) adjusted for the specific needs of the projects. The next steps in our developmental work on the VRE in this respect will be to enable the automated generation of DDI files from the data documenting performed within the VRE at particular data life-cycle steps. For example, the "data collection" event is well suited for documentation using a structured metadata format. The result of this effort is that elements such as variables, their descriptions, codes, question text, and question routing instructions can be easily searched, shared, or (re)used within the SFB and afterwards made available to the research community in an interoperable and semantically enriched format.

Another documentation type is related to the organization of data required in order to describe semantic relations between different types and data elements. In this case we recommend researchers to create one or more additional worksheets (e.g. MS Excel or OpenOffice/ LibreOffice spreadsheets) within their shared folders to contain information about the relations between files and data in directories. By giving rules and conventions for file naming, meaningful abbreviations, and versioning (e.g. unified time-stamps for each new version of a file, information on creator, etc.), files and processes are made much more traceable and research more efficient.

A Virtual Research Environment as a part of institutional services

The comprehensibility and replicability of research data is only possible when its visibility is ensured. Therefore, although a VRE should support researchers' discipline-specific working processes, it should also be an integrated component of already established institutional solutions for general research data management. Such examples of institutional services are: a) Publication Management (PUB) providing the infrastructure for managing and visualizing the university's publication output, b) the University Computing Center (HRZ) running global authentication services and access rights management for academics, and c) PEVZ (Staff and Organization Database) as a university-wide directory of staff and departments or affiliated organizations.

The INF project is closely linked with Bielefeld University Library (UB Bielefeld), whose practices are rooted in the management and delivery of publications and in leading expertise in metadata generation, data storage and cataloguing, and data retrieval. Through this cooperation, the UB expands its remit beyond that of traditional academic libraries by taking

on more responsibilities for providing information and data services at the earlier stages of the data and research life cycle—tasks that are increasingly expected from modern libraries (Neuroth et al. 2008). The existing, classic publication services are thus supplemented by post-production and post-publication services such as the provision of digital environments for accessing research data as "scientific records" that could be less dependent on papers and articles and expressed instead in terms of networks of links and associations among diverse research artifacts. The linkage of semantically enriched data is then used to support research comprehensibility, data reuse, and further secondary analysis. A key feature within the publication service PUB is the extension of the existing DOI registration interface for publications to research data—and thus for data generated within the SFB.

Figure 2 summarizes the involvement of the actors within the integrated data infrastructure for the SFB 882. It uses existing interfaces to institutional services such as PEVZ and HRZ, and other "External Resources" such as PUB. By using these global systems, data maintenance is simplified and persistence is ensured (e.g. by using existing user rights for authentication). Specifically, direct usage of PEVZ data facilitates the automatic linkage of data, publication, projects, and people. Another feature implemented within the infrastructure is the automatic mounting of shared drives (as "External Resources") into the internal VRE depending on the current user's Active Directory group memberships.

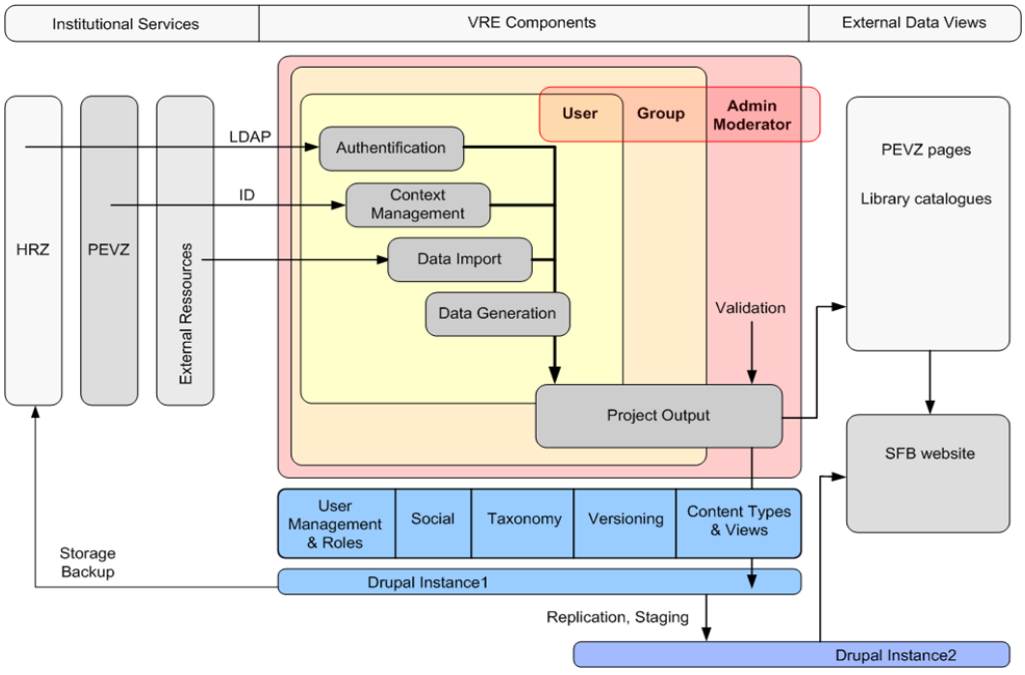

Figure 2: Components of the VRE and its interfaces to institutional services

As our internal "core" platform (VRE Components in Figure 2), which we further developed and modified for the VRE, we use Open Atrium (http://openatrium.com/)—a free and open-source software package whose code is licensed under GPL 2, inherited from Drupal, its backbone (http://drupal.org/). The standard package allows an out-of-the-box creation of group/project spaces in which users can have conversations, preserve knowledge, track progress, and share files.

The "external data view" presents the presentation layer of project output and editorial work, which are transferred to different interfaces. For example, a publication captured by an SFB 882 researcher would automatically appear on the institutional publication site of the SFB or his personal scholarly web page. Analogously, specific "non-sensitive" data nodes (all data objects created in Drupal are called "nodes"), e.g. "news," "calendar events," or "uploaded files" created within the VRE can be made available and visible on the external SFB website.

Summary

We have seen that the development of a documentation scheme which records the entire research process depends primarily on epistemological notions and suitable technical solutions. While quantitative social research tends to follow science-oriented paradigms, qualitative social research tends to follow humanities paradigms. These different choices bias not only the research process, but also data documentation methods.

Qualitative research, in particular, is complex, context-sensitive, and reflexive. If they are not documented well enough, its data are therefore mostly not understandable for third parties. We have also seen that the effort required to store understandable qualitative data is large and so far not all necessary documentation steps have been identified. In the section on documentation practices we distinguished between *in situ* and *ex situ* documentation practices, which refer to data for different purposes. We chose this approach because the characteristics of qualitative research (especially data thickness, context orientation, and reflexivity) refer to tacit knowledge of the research phenomenon that the researcher necessarily acquires. This kind of knowledge is indispensable for a detailed understanding of qualitative data and their documentation. Accordingly, it must be explicated in order to allow further understanding and utilization of the data by studying the research process itself. In

doing so, we can transfer knowledge regarding qualitative data into the documentation schema by observing and adapting day-to-day practices of the researchers.

Thus, a documentation scheme of qualitative data that accounts for parallel and circular processes, layering of meaning, and the steps and levels of contextualization is still in the making. Future improvements will be a more recipient-oriented variation of the documentation outputs and an adaptation to mixed-method requirements. Moreover, it must be taken into account that the requirements of data management could engender negative consequences for the research process itself, such as high consumption of manpower and time or the implicit need for standardization.

In contrast to qualitative research, the problem of quantitative documentation is not so much a lack of a standard or methods, but the challenge of finding a balance between the contents that must be documented and the workload for the individual researcher. One of the reasons for this is the perspective from which research in this field is conducted. The INF project, with its unique constellation of technicians and both qualitative and quantitative social scientists, takes a range of different approaches towards changing that perspective in everyday practice. We have shown in this paper that the problems of acceptance and granularity can only be solved by including researchers in the development of a comprehensive data management strategy. As everyday advisors for people who work with various data, we concluded that collaboration with the researchers is necessary: a patriarchal style, telling researchers what they should do, simply does not work. Since a lack of user perspective has been identified as a problem, the further development and implementation of user-driven documentation solutions will be one of the main tasks for the future.

Data infrastructure projects that deliver solutions for sustainable research data management and provide support for data-intensive research projects are indispensable today. In our heterogeneous case—the collaborative research center SFB 882, where the various work and research processes additionally face us with a large number of data types, evolving routines, and subject-specific documentation standards—it quickly becomes clear that there will never be a one-size-fits-all solution for a general data and information infrastructure and that VREs cannot be expected to fulfill every sort of task. This is why we conducted our requirements analysis to discover the challenges for developing and sustaining a research environment for the SFB, with the greatest emphasis on providing digital technologies to support the research

process, ensure the long-term accessibility, data comprehensibility and visibility of the data, and thus facilitate its use and reuse in the long term.

In the first stage of the project, we implemented several components for collaborating, data sharing, and documenting. The next stage will be to refine particular solutions and react more flexibly to the specific workflows within the different research groups. A considerable upcoming challenge will be to provide a general technical solution for data documentation using the DDI3 metadata standard, which must be enriched with a certain flexibility to make it usable across all projects.

We believe that after a settling-in period researchers will benefit greatly from using a collaborative environment, which indirectly (through its technical framework for discussions and feedback) promotes reciprocal monitoring of data validity and accuracy. The result of this will be well-organized, well-documented, and accessible high-quality data that can form a basis for reliable and trustworthy research results. Whenever the opportunity arises, whether in consultations or when giving technical support, the INF project additionally raises awareness of the data life cycle and long-term preservation issues, its mantras being: (1) avoid vendor lock-in by using free and open-source software, (2) avoid binary formats, favor formats based on plain text (US-ASCII or Unicode) such as XML, (3) keep all versions of your data, and try to take notes of changes.

One of the main findings in our work and this paper is that it is crucial to exchange information between different fields of research in order to achieve a coherent framework for data documentation. It is necessary not only to gather information from researchers, but also to understand the technical possibilities for ensuring “real-life” solutions as opposed to theoretically good ideas that do not apply to all researchers or are impossible to realize from a technical point of view. There is no doubt that good documentation takes time, but if researchers understand the necessity of documentation and if, ideally, they themselves draw profit from this extra work, the need for comprehensive and relevant data documentation will gain more and more acceptance.

References

Bergmann, J. R. (2006), Qualitative Methoden der Medienforschung - Einleitung und Rahmung, in Ruth Ayaß & Jörg R. Bergmann, eds., *Qualitative Methoden der Medienforschung*, Rowohlt, Reinbek bei Hamburg, pp. 13-41.

Bryman, A. (2006), 'Integrating quantitative and qualitative research: How is it done?', *Qualitative Research* **6**(1), 97-113.

Büttner, S., Hobohm, H.-C. & Müller, L., eds. (2011), *Handbuch Forschungsdatenmanagement*, Bock + Herchen, Bad Honnef.

Flick, U. (2011), *An introduction to qualitative research*, Sage, Los Angeles.

Fomel, S. & Claerbout, J. (2009), 'Guest Editors' Introduction: Reproducible Research', *Computing in Science Engineering* **11**(1), 5-7.

Garfinkel, H. (1967), *Studies in Ethnomethodology*, Prentice Hall, Englewood Cliffs, NJ.

Gläser, J. & Laudel, G. (2008), 'Creating Competing Constructions by Reanalysing Qualitative Data', *Historical Social Research* **33**(3), 115-147.

Gómez, O. S., Juristo, N. & Vegas, S. (2010), 'Replication, Reproduction and Re-Analysis: Three Ways for Verifying Experimental Findings', *Proceedings of the 1st international workshop on replication in empirical software engineering research (RESER 2010), Cape Town, South Africa*.

Hammersley, M. (2007), 'The issue of quality in qualitative research', *International Journal of Research & Method in Education* **30**(3), 287-305.

Heaton, J. (2008), 'Secondary Analysis of Qualitative Data: An Overview', *Historical Social Research* **33**(3), 33-45.

Heritage, J. & Watson, R. (1979), Formulations as Conversational Objects, in George Psathas, ed., *Everyday Language. Studies in Ethnomethodology*, Irvington Publishers, New York, pp. 123-162.

Huschka, D. & Wagner, G. G. (2012), 'Data Accessibility is Not Sufficient for Making Replication Studies a Matter of Course' (195), RatSWD Working Paper No. 195, Berlin.

Kalthoff, H., Hirschauer, S. & Lindemann, G., eds. (2008), *Theoretische Empirie. Zur Relevanz qualitativer Forschung*, Suhrkamp, Frankfurt/Main.

Knorr Cetina, K. (1981), *The manufacture of knowledge*, Pergamon Press, Oxford.

Knuth, D. E. (1984), 'Literate programming', *Comput. J.* **27**(2), 97–111.

Latour, B. & Woolgar, S. (1979), *Laboratory life. The social construction of scientific facts*, Sage, London.

Long, J. S. (2009), *The workflow of data analysis using Stata*, Stata Press, College Station, Tex.

Lynch, M. (2000), 'Against Reflexivity as an Academic Virtue and Source of Privileged Knowledge', *Theory, Culture & Society* **17**(3), 26-54.

Neuroth, H., Oßwald, A., Scheffel, R., Strathmann, S. & Huth, K. (2008), *Nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, Version 1.2*, Nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland.

Popper, K. (1959), *The Logic of Scientific Discovery*, Hutchinson & Co, New York. Published in German as *Logik der Forschung*, 1934.

Rossini, A. & Leisch, F. (2003), 'Literate statistical practice' (194), UW Biostatistics Working Paper 194, University of Washington, Seattle, USA.

Schütz, A. (1953), 'Common-Sense and Scientific Interpretation of Human Action', *Philosophy and Phenomenological Research* **14**(1), 1-38.

Silverman, D. (2007), *Interpreting qualitative data*, Sage, Los Angeles.

Strübing, J. (2007), Research as pragmatic problem-solving, in Antony Bryant & Kathy Charmaz, eds., *The Sage Handbook of Grounded Theory*, Sage, London, pp. 580-601.

Suchman, L. & Jordan, B. (1990), 'Interactional troubles in face-to-face survey interviews', *Journal of the American Statistical Association* **85**(409), 232-241.

Walters, P. (2009), 'Qualitative archiving: engaging with epistemological misgivings', *Australian Journal of Social Issues* **44**(3), 309-320.

Previously published SFB 882 Working Papers:

Diewald, Martin / Faist, Thomas (2011): From Heterogeneities to Inequalities: Looking at Social Mechanisms as an Explanatory Approach to the Generation of Social Inequalities, SFB 882 Working Paper Series, No. 1, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Bielefeld.

Busch, Anne (2011): Determinants of Occupational Gender Segregation: Work Values and Gender (A)Typical Occupational Preferences of Adolescents, SFB 882 Working Paper Series, No. 2, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project A3, Bielefeld.

Faist, Thomas (2011): Multiculturalism: From Heterogeneities to Social (In)Equalities, SFB 882 Working Paper Series, No. 3, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project C3, Bielefeld.

Amelina, Anna (2012): Jenseits des Homogenitätsmodells der Kultur: Zur Analyse von Transnationalität und kulturellen Interferenzen auf der Grundlage der hermeneutischen Wissenssoziologie, SFB 882 Working Paper Series, No. 4, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project C3, Bielefeld.

Osmanowski, Magdalena / Cardona, Andrés (2012): Resource Dilution or Resource Augmentation? Number of Siblings, Birth Order, Sex of the Child and Frequency of Mother's Activities with Preschool Children, SFB 882 Working Paper Series, No. 5, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project A1, Bielefeld.

Amelina, Anna / Bilecen, Başak / Barglowski, Karolina / Faist, Thomas (2012): Ties That Protect? The Significance of Transnationality for the Distribution of Informal Social Protection in Migrant Networks, SFB 882 Working Paper Series, No. 6, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project C3, Bielefeld.

Alemann, Annette von / Beaufaÿs, Sandra / Reimer, Thordis (2012): Gaining Access to the Field of Work Organizations with the Issue of "Work-Family-Life Balance" for Fathers, SFB 882 Working Paper Series, No. 7, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project B5, Bielefeld.

Kaiser, Till (2012): Haben gebildetere Mütter gewissenhaftere Kinder? Soziale Herkunft und Persönlichkeitsentwicklung im frühkindlichen Alter, SFB 882 Working Paper Series, No. 8, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project A1, Bielefeld.

Gusy, Christoph / Müller, Sebastian (2012): Social Construction of Heterogeneity Indicators and their Relationship to Law. The Example of Guiding Principles in Immigration Law, SFB 882 Working Paper Series, No. 9, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project C4, Bielefeld.

Liebig, Stefan / May, Meike / Sauer, Carsten / Schneider, Simone / Valet, Peter (2012): Inequality Preferences in Interviewer- and Self-Administered Interviews, SFB 882 Working Paper Series, No. 10, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project A6, Bielefeld.

Fauser, Margit / Voigtländer, Sven / Tuncer, Hidayet / Liebau, Elisabeth / Faist, Thomas / Razum, Oliver (2012): Transnationality and Social Inequalities of Migrants in Germany, SFB 882 Working Paper Series, No. 11, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project C1, Bielefeld.

Freistein, Katja / Koch, Martin (2012): Global Inequality and Development. Textual Representations of the World Bank and UNDP, SFB 882 Working Paper Series, No. 12, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project C5, Bielefeld.

Golsch, Katrin (2013): Shall I Help You My Dear? Examining Variations in Social Support for Career Advancement within Partnerships, SFB 882 Working Paper Series, No. 13, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project A3, Bielefeld.

Bröckel, Miriam / Busch, Anne / Golsch, Katrin (2013): Headwind or Tailwind — Do Partner's Resources Support or Restrict a Promotion to a Leadership Position in Germany?, SFB 882 Working Paper Series, No. 14, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project A3, Bielefeld.

Cardona, Andrés (2013): Closing the Group or the Market? The Two Sides of Weber's Concept of Closure and Their Relevance for the Study of Intergroup Inequality, SFB 882 Working Paper Series, No. 15, DFG Research Center (SFB) 882 From Heterogeneities to Inequalities, Research Project A1, Bielefeld.