DOI:10.25365/yis-2020-5-3

Creating Needs-Based Metadata and Research Data Management Services

Exploring the Requirements of Scientists

Tereza Kalová*

Objective — Detailed needs-assessments of the target groups are of particular relevance when developing new services for eScience. This study aims to identify the demands of researchers in natural sciences regarding metadata for research data.
Methods — This paper applies the key points of Grounded Theory to analyze eight semi-structured interviews with scientists from the University of Vienna and the Medical University of Vienna. The interviews were conducted from April to May 2019 as part of the project e-Infrastructures Austria Plus funded by the Austrian Federal Ministry of Education, Science and Research.
Results — The results illustrate how scientists understand the term "metadata". The study identifies several examples of describing data as well as six areas of researchers' requirements for data management services. These include the need to design efficient solutions in German and English, the demand for technical infrastructure and dedicated data stewards who bring expertise from a relevant discipline. Service providers further need to take interdisciplinary differences into account when developing new forms of training and support.
Conclusion — While the researchers' statements highlight the various ways they described data, the interviews also indicated a lack of understanding of the term "metadata". Together with the six areas of needs and requirements, these demonstrate the need for action by libraries and other service providers in order to further contribute to the development of sustainable research data management.

Keywords — Austria, Research Data, Research Data Management, Metadata, Needs Assessment, Library and Information Science

Entwicklung bedarfsgerechter Services für Metadaten und Forschungsdatenmanagement: Bedarfsanalyse in den Naturwissenschaften

Zielsetzung — Bei der Entwicklung von bedarfsgerechten Services im Bereich eScience sind Untersuchungen der Bedürfnisse und Anforderungen der Zielgruppen von besonderer Relevanz. Ziel dieser Studie war, den Bedarf von Forschenden in den Naturwissenschaften an Metadaten für Forschungsdaten zu ermitteln.
Methoden — Angelehnt an die Grounded Theory wurden acht Leitfadeninterviews mit Naturwissenschaftlern von der Universität Wien und der Medizinischen Universität Wien durchgeführt. Die Interviews fanden von April bis Mai 2019 im Rahmen des vom österreichischen Bundesministerium für Bildung, Wissenschaft und Forschung geförderten Projekts e-Infrastrukturen Österreich Plus statt.
Ergebnisse — Die Ergebnisse demonstrierten ein mangelndes Verständnis des Begriffes »Metadaten« unter Forschenden. Es konnten aber mehrere Beispiele der Beschreibung von Daten in der Praxis identifiziert, sowie sechs Bereiche an Bedürfnissen der Forschenden herausgearbeitet werden. Diese zeigten die Notwendigkeit, effiziente Lösungen für Forschungsdatenmanagement auf

[*] Tereza Kalová, MA (Res), M.A. LIS | Universität Wien / University of Vienna | tereza.kalova@univie.ac.at | ORCID: 0000-0002-1764-72284

Young Information Scientist (YIS) wird vom Verein zur Förderung der Informationswissenschaft (VFI), Wien, herausgegeben. Alle Beiträge unterliegen einem Peer Review. ISSN: 2518-6892

Deutsch und Englisch zu konzipieren, den Bedarf an technischer Infrastruktur und den Wunsch nach dem Einsatz von Data Stewards, die Expertise aus einer relevanten Disziplin mitbringen. Ferner sollten Dienstleister bei der Entwicklung neuer Formen der Trainings- und Beratungsangebote interdisziplinäre Unterschiede berücksichtigen.
Schlussfolgerungen — Während die Aussagen der Befragten die verschiedenen Arten der Datenbeschreibung illustrierten, wiesen die Interviews auch auf ein mangelndes Verständnis des Begriffs »Metadaten« hin. Zusammen mit den sechs Bereichen an Bedürfnissen und Anforderungen, ergibt sich aus den Interviews ein weiterer Handlungsbedarf für Bibliotheken und andere Serviceeinrichtungen, um die Entwicklung im Bereich Forschungsdatenmanagement voranzutreiben.

Schlagwörter — Österreich, Forschungsdaten, Forschungsdatenmanagement, Metadaten, Bedarfserhebung, Bibliotheks- und Informationswissenschaft

Diesem Beitrag liegt folgende Masterarbeit zugrunde / This article is based upon the following dissertation/thesis: Kalová, Tereza: Metadaten für Forschungsdaten: Bedürfnisse und Anforderungen in den Naturwissenschaften. Masterarbeit (M.A. (LIS)), Humboldt-Universität zu Berlin, 2019. DOI: 10.18452/21536.

> *[W]hen you have a data file, you need to somehow understand months and years later, what this data file is about and the metadata is the context that allows you to understand the file.*
>
> *(Scientist E, 2019, para. 12)*

1. Introduction

Metadata represent one of the most common terms in library and information science. They play a key part in describing publications and increasingly also in research data. Scholars and academic librarians have repeatedly demonstrated the importance of creating metadata and documentation for data (Wilkinson et al. 2016; Blumesberger 2018). Rich metadata offer numerous advantages, as they open the way to interdisciplinary research, as well as the reuse of research data (Farnel and Shiri 2014; Park and Park 2019). However, awareness among researchers is often still lacking (Wiley and Burnette 2019).

The main objective of the present study is to identify specific needs and demands of scientists regarding metadata for research data as a basis for development and optimization of library services. In order to do so, I also discuss the topics of the understanding of and familiarity with the term “metadata” as well as common approaches to the description and documentation of research data in the sciences. The study seeks to explore the following research questions in the context of the population studied:

1. Do scientists understand the term “metadata”?
2. How do researchers describe data?
3. What are the needs and requirements regarding metadata for research data in the sciences? What kinds of services do scientists need?

The study attempts to answer these questions using semi-structured qualitative interviews with researchers.

This article is based on a Master’s thesis in Library and Information Science that I completed at the Humboldt University of Berlin and published in June 2020[1]. The work was carried out as part of the Austrian project e-Infrastructures Austria Plus[2]. Preliminary findings were published in a project report in September 2019[3].

[1] For the complete thesis see – https://doi.org/10.18452/21536

[2] The project e-Infrastructures Austria Plus was financed through a grant from the Austrian Research Ministry and ran from January 2017 to December 2019. For more information on the project as well as its many deliverables on various aspects of eScience see https://www.e-infrastructures.at/en/

[3] For the project report see – https://doi.org/10.25651/1.2019.0020

2. Literature Review

The amount of data generated in research is growing exponentially as a result of the ever-increasing digitization (Pryor 2012, p. 2). This development is especially pronounced in natural science disciplines (Hey and Trefethen 2003, p. 2; Büttner et al. 2011, p. 13). Moreover, interdisciplinary and international collaboration necessitate long-term strategies to manage data effectively. Data-driven or data-intensive research, also called the fourth paradigm (Hey, Tansley, et al. 2009), is becoming increasingly established in many disciplines (Zook et al. 2017, p. 1; Rapport and Braithwaite 2018, pp. 3–6). Digital data represent invaluable results of research independent of formal publications well beyond the scope of the original project (Wilkinson et al. 2016, p. 1). As such, there is a clear demand for sustainable research data management (RDM).

New roles such as a data librarian, data manager or data steward have been created as part of RDM infrastructures worldwide (Büttner et al. 2011, p. 206). As is clear from the international study by Matusiak and Sposito, RDM services are predominantly located at academic libraries (Matusiak and Sposito 2017, p. 755). On the one hand, research data management offers libraries the possibility of a new field of activity as the next logical step to their traditional tasks ("Forschungsdatenmanagement und Bibliotheken" 2016, p. 614). On the other, it also presents a challenge that makes libraries rethink their rather passive role in storing research results in order to become active partners of scientists during the entire research process (Martin 2013, p. 20; Morgan et al. 2017, p. 304). Librarians can offer their expertise in many areas including their knowledge of indexing, documentation and metadata. Metadata play a crucial role in making research FAIR. The FAIR data principles demand that "all research objects should be Findable, Accessible, Interoperable and Reusable (FAIR) both for machines and for people" (Wilkinson et al. 2016, p. 3) and are increasingly used by funding agencies and publishers as guidelines for sustainable RDM. Metadata are highly relevant to each of the four attributes of FAIR Data.

Despite the argument that RDM represents further development of traditional library services, "[t]here can be barriers for researchers to accept librarians in this role" (Surkis and Read 2015, p. 156). Therefore, it has been suggested that librarians use "the language of research" instead of librarianship when talking to scientists (ibid., p. 156).

2.1. Researchers and Metadata

Metadata can be defined as structured information that describes analog or digital objects (Jensen et al. 2011, p. 83). There is a vast amount of scholarly literature highlighting the numerous reasons for research data management, including the description of research data in the form of metadata and documentation (Büttner et al. 2011; Pomerantz 2015; Wilkinson et al. 2016). However, when developing library services for RDM concrete requirements of the faculty must be taken into account, usually through a formalized needs assessment using qualitative interviews or online surveys. These types of studies usually address metadata as an aspect of RDM, typically in the form of a single question. It is also important to note that most studies are general rather than subject-specific in nature and intend to represent the needs of all disciplines.

In 2015, the project e-infrastructures Austria[4] conducted an Austrian-wide survey on RDM. The study analyzed responses from more than 3,000 scientists from various research institutions (Bauer et al. 2015, p. 18). The study has shown that 89% of researchers want more institutional support in managing their data properly (ibid., p. 67). This emphasizes the clear need to develop such services. The report mentions that most scientists describe their data, albeit very inconsistently (ibid., p. 31). The University of Warwick conducted eight interviews with scientists working in the field of physics and social research. One of the results was that, compared with the social scientists that wanted to make use of the UK Data Archive[5], metadata "was not a priority" for some of the physicists (Delasalle 2013, pp. 101–102).

[4] The project e-Infrastructures Austria ran from 2014-2016. For more see https://e-infrastructures.univie.ac.at

[5] https://www.data-archive.ac.uk/

Both groups, however, agreed on their reservations about the bureaucratization of science, which they associate with RDM (Delasalle 2013, p. 101). In this regard, Delasalle suggests that scientists should be encouraged to participate actively in the development of new services (ibid., p. 103).

In a survey conducted at the University of Trier, in which almost 200 scientists took part (medicine/science n=51), researchers considered metadata as one of the most important factors enabling the reuse of data (Lemaire et al. 2016, pp.6, 20). Nonetheless, the study also showed a lack of awareness of the importance of metadata among researchers. Lemaire et al. argue that in order to promote a culture of data sharing, "work processes and tools [must] be designed in such a way that, for example, the description of the research data (metadata) does not entail excessive additional effort" (ibid., p. 28). The survey at the Technical University of Hamburg (n=96) also showed that researchers want advice on technical questions including metadata (Feldsien-Sudhaus and Rajski 2016, p. 35). In the survey by the Philipps University of Marburg (n=427) (Krähwinkel 2015, p. 9), the demand for consulting on the topics of metadata and long-term archiving is the most requested service at 66% (ibid., p. 39). Interestingly, this service ranks first among scholars from the fields of medicine and natural sciences (ibid., p. 39). In their study of RDM practices in bioengineering and biomedical research, Wiley and Burnette from the University of Illinois observe notable gaps in the knowledge and awareness of metadata and documentation among researchers (Wiley and Burnette 2019, p. 16). This calls for "instruction in the areas of file organization […], documentation, and metadata standards" by academic libraries (ibid., p. 16).

The results of the studies demonstrate a lack of or insufficient awareness and knowledge in the area of metadata. There is a need for information transfer and consulting on adequate data description. Scientists further require simple and efficient RDM solutions. Given the comparatively low number of studies that deal with metadata as part of their needs assessment, the topic does not seem to be of particular interest to the service centers that conducted the studies.

3. Methods

The terms "research data" and "metadata" are neither immediately clear nor readily and uniformly understood among researchers. For these reasons, I chose to explore the research questions using a qualitative approach in the form of semi-structured interviews with scientists. The study employed the essential points of Grounded Theory (Glaser and Strauss 1967; Glaser 1978; Charmaz 2014; Bryant 2017).

3.1. Sampling

The interviewees were chosen based on the following criteria:

1. They were involved in current research activity in a science field (compulsory).
2. They demonstrated interest in the topic of research data management, such as through previous inquiries to the data management department at the Vienna University Library or the participation in training courses (optional).

The first contact was made either via e-mail, by phone or in person. An attempt was made to use convenience and snowball sampling to generate as heterogeneous a group as possible in terms of seniority, experience and scientific discipline. As it was only possible to recruit five participants from the University of Vienna, the scope was expanded to include three researchers from the Medical University of Vienna. It is crucial to emphasize that the study deviates from Grounded Theory in this point, as due to the difficult field access and limited timeframe it applies no theoretical sampling. The following results should therefore be interpreted with caution.

Scientist	Discipline	Position
A	Medicine	Junior Researcher
B	Chemistry	Senior Researcher
C	Biochemistry	Senior Researcher
D	Biology	Senior Researcher
E	Physics	Senior Researcher
F	Biology	Senior Researcher
G	Medicine	Senior Researcher
H	Medicine	Junior Researcher

Table 1: Sample

All scientists who could be recruited for the study are male. Despite considerable efforts to approach female researchers, none could be recruited for the study. The participants all work full-time and are divided into two groups (Junior Researcher, Senior Researcher) depending on their experience. Although the sample represents different age groups, a larger proportion of junior researchers would have enriched this study. Various research disciplines are distributed unevenly with three participants working in clinical medical research. However, as this is a very broad area, this does not seem particularly problematic. Half of the respondents are originally from Austria, the other half from other European countries. All respondents live in the Greater Vienna area and work at the University of Vienna, or the Medical University of Vienna. Five of the eight interviews (A, B, D, G, H) were conducted in German and translated for this article by the author (for the original quotations see appendix C).

Because of the nature of the study — as part of the project e-infrastructures Austria plus — it was unfortunately impossible to anonymize the data completely. These were instead pseudonymized and all participants formally agreed to the handling of the data through an informed consent form.

3.2. Data Collection and Analysis

The interviews took place after a pre-test in person from April 24th to May 11th, 2019. They lasted between 25 and 40 minutes and were all recorded digitally. All the interviews were conducted using the interview guide (appendix A). The interviews were partially transcribed using the software MAXQDA with the focus on responses relevant to the research questions.

I examined the data using several coding methods. In order not to lose the overview of the data and theoretical considerations, analytical memos were created during the entire research project. The initial coding phase consisted of a combination of in vivo as well as sociologically constructed codes. The in vivo codes, which arose from the three English-language interviews, were merged through comparisons with the corresponding codes in German in order to carry out the further analysis of all interviews in one language. I then further analyzed these in order to systematically draw up another level of descriptive codes. During the next phase of analysis, all codes were grouped into suitable categories as part of selective/focused coding (for the coding system see appendix B).

4. Results

This section highlights the most important findings pertaining to the understanding of the term “metadata”, documentation of data, as well as six areas of needs and requirements of scientists. The data are presented according to the coding system (appendix B).

4.1. Metadata and Data Documentation

The interviews show that half of the respondents could at least partially classify the term metadata correctly (scientists C, E, F, H):

> *“Metadata are basically the way to describe the common properties of the data”* (Scientist C, 2019, para. 9).

The other respondents either did not know the term at all, or unconsciously misused it, referring to other concepts such as “meta-analysis”:

> *“[Metadata] are the second level of data processing. So if, for example, I write a review, like I published it a few weeks ago, where I then summarize all the work on a specific topic”* (Scientist D, 2019, para. 7).

Despite the misunderstandings concerning the term “metadata”, the study identified multiple examples of documentation of research data. As the interviewees did not make a sharp distinction between these two concepts, the article summarizes both as the description of data. All eight scientists described their research data in some form.

4.1.1. *Motivation for Describing Research Data*

The interviews suggest four main reasons for describing data. Findability and visibility (1) played a major role for five of the eight scientists:

> *“[I]t was partly impossible to find where data were available at all. In what form they are available [...] And of course one has to know [...] that they exist”* (Scientist G, 2019, para. 56).

For half of the respondents, metadata and documentation made data usable for themselves, as well as reusable for others (2):

> *"If you just collect the primary data and they are not annotated properly, this information is lost, basically"* (Scientist C, 2019, para. 17).

Furthermore, for three of the eight scientists, data description enabled the reproducibility of the results (3). Two mentioned the requirements of scholarly journals (4) as an essential motivating factor for the description of research data. Three reported on the publication of research data as a requirement of scholarly publishers. However, this did not necessarily mean using sufficient metadata.

4.1.2. Automatically Generated Metadata

Three of the respondents mentioned that automatically generated metadata were relevant in their discipline:

> *"What is always important to me is data on or metadata about the creation and time and processing ...of data [...] so that I can always understand when I did something"* (Scientist H, 2019, para. 11).

4.1.3. Discipline-Specific Approaches

Each science discipline negotiated the concept of describing data differently. In clinical medical research, detailed applications had to be submitted to the ethics committee of the institution before any new research project could begin. These included a study protocol, a document that, among other things, described the handling of data. According to the statements, the logs were usually not kept with the research data, nor were they used during the research process.

Respondents from the fields of physics and biochemistry mentioned the use of laboratory books and electronic lab notebooks to describe data:

> *"The simplest metadata is what the person writes in his or her lab book"* (Scientist C, 2019, para. 26).

The entries figured both as data in their own right and as descriptive metadata. Scientists C and F further reflected on the use of dedicated descriptive files to write metadata.

4.1.4. Publications

Scientists B and D spoke of publications, specifically the section on methodology, as the description of data that allowed for reusability. They considered further separate metadata files or documentation superfluous:

> *"If I want to know how I did it back then, I look at the publication and the methods. Because that would be doubling of descriptive methods for me, why?"* (Scientist D, 2019, para. 14).

4.1.5. Metadata and Collaboration

According to all respondents with one exception (scientist C, where metadata were partially automatically linked to data), metadata did not play a role in active collaboration. When analyzing data in a team, researchers transferred the relevant information about the data directly, such as in conversation or via e-mail:

> *"One makes interpretations with the data. And of course you have to talk to each other very intensively, because a physicist makes different interpretations than a chemist"* (Scientist B, 2019, para. 36).

4.1.6. Metadata When Using Data from Others

Except for scientist D, who mentioned that the use of data generated by others did not play any part in his discipline, all the other researchers had had experience with this type of data. The interviews underscored the key role of this practice in clinical medical research (A, G, H) that often utilized data from large patient registries. Scientist A even stated that he is "dependent" on data from others (2019, para. 35). In this context, scientist H described issues with the understanding of data lacking sufficient description:

> *"There was no explanation for the classification of certain patient characteristics, where a numerical system was used [...] The data were coded in the data set as one, two, three, four, five and I did not know which patient had received chemotherapy"* (Scientist H, 2019, para. 21).

The answers underscored the critical importance of documentation in terms of data reuse.

4.2. Needs and Requirements

As suggested by the variety of research the interviewees engaged in, their needs and requirements were equally diverse. Whenever possible, this section mentions those that relate to the creation of metadata and documentation. Overall, it was difficult to extract this aspect from more general research data management and answers relating to various RDM services are therefore included in this section.

4.2.1. *Differences Among the Disciplines*

The various scientific disciplines represented in this study showed great heterogeneity when it came to the way they used terms like "research data" or "metadata". This was also explicitly discussed by three of the respondents (scientists B, E, H):

> *"We all speak a different language. So, biologists speak differently from chemists, and we in turn speak differently from physicists [...]. Everyone tries explaining things to others [...] in terms of their own data"* (Scientist B, 2019, para. 24).

It was desirable for all scientists from various disciplines to use a more "uniform language" as opposed to a variety of languages currently in use. This demand came into play not only in interdisciplinary research.

4.2.2. *Effort*

Three of the scientists (C, E, F) who, according to their statements, created a lot of metadata, broached the subject of the great effort involved in properly documenting data. In particular, they perceived this task as being too "difficult" and "time-consuming" (Scientist C, 2019, para. 25). This was also evident in the comments on the use of the university's institutional repository:

> *"Just filling in the minimal metadata that's relevant that still takes [...], I mean it's not a long process maybe five minutes [...], but for doing a lot more open data, it's still a bit too much"* (Scientist E, 2019, para. 28).

Although the other respondents did not explicitly mention the time required to create metadata, the issue of efficiency arose in other parts of their interviews. This indicates that it would also matter in terms of describing data.

4.2.3. *Language*

Seven of the eight interviewees claimed that they do research in two languages. In six cases, this was the combination of English-German, in one it was Swedish-English. One respondent mentioned using only one language (English).

The scientists participating in the research preferred using German as a working language due to geographical reasons. However, English was used as a lingua franca in everyday communication because their research teams were often international. The interviews indicated that scholarly activities beyond PhD were carried out mainly in English. All the respondents agreed that they prefer English not only because of the ease of communication in it, but also because it is the prevailing language of scholarly literature:

> *"For someone working in science of nature [...], all the literature's in English"* (Scientist F, 2019, para. 32).

4.2.4. *Information Transfer*

Half of the respondents were already familiar with university consulting and support services in the field of RDM and they perceived them as important. Most common inquiry topics related to the writing of data management plans (scientists E, H), as well as archiving data (scientist D).

The scientists' attitudes towards training and courses on RDM differed greatly. Such services seemed to be neither particularly well known, nor sparked much interest (six of the eight researchers). In contrast, scientists C and E considered courses on RDM including the topic of metadata, to be helpful:

> *"I think an introductory course would be useful. In particular things... how you manage the data, how you annotate them in a reasonable way, what exactly are metadata, which ones are useful for others"* (Scientist C, 2019, para. 50).

C and E had both attended at least one training course. Also, in this context, scientist E emphasized the importance of efficiency.

Researcher H mentioned an unsuitable range of courses offered by the university:

> *"[I] now and then see that there are offers [...] But then these are more like series of lectures or something. Concrete training somehow for maybe, I haven't seen something for junior scientists or something like that"* (2019, para. 43).

This statement suggested a wish to take part in a training course tailored to specific needs. There was also a growing interest in the topic of describing data (scientists B and G):

> *"That was an interesting question. So, I didn't really think about it, but now in retrospect [...], that is actually a relatively good idea to get it right from the start"* (Scientist G, 2019, para. 73).

Two of the respondents also stated that they preferred to acquire the necessary knowledge on their own:

> *"I learn everything by myself, or I draw on the recommendations and experiences of other colleagues"* (Scientist A, 2019, para. 45).

4.2.5. *Staff*

Due to the differences among scientific disciplines, as well as the multitude of data types and methods, three of the researchers (C, E, G) expressed the need for professional and dedicated staff for RDM:

> *"If this has to be done properly, you actually need a person to do just that"* (Scientist C, 2019, para. 40).

RDM requires advanced discipline-specific knowledge, researcher E emphasized:

> *"It's probably true for [...] most natural science research, is that the data is generated in so many different ways that are very specific and no generalist data management support person can be familiar with all the technical nuances. So, I think it kind of has to be in-house"* (Scientist E, 2019, para. 40).

In this regard, scientist C also suggested the establishment of working groups of researchers in order to develop specific RDM solutions and best practices. Two other interviewees (A, B) would have preferred having more staff to process data. This also highlighted the relationship between working with scientific (meta)data and specialist knowledge.

4.2.6. *Technical Infrastructure*

All eight interviews underlined the need for improved technical infrastructure. It should be user-friendly, automated if possible, and integrate local solutions.

The topic of user-friendliness came up in three interviews and pertained mainly to reducing the effort necessary to fill out the metadata fields in the institutional repository. However, this also applied in general to technical solutions for metadata, as can be seen from the quotation from scientist C:

> *"Another thing which is very important, develop some tools, which make it easy (laughs) because this is very hard"* (2019, para. 50).

Scientists C and F proposed the use of existing metadata and the integration of institutional technical solutions:

> *"We have [..] the parameters that are important to replicate the science. I think what would be useful, [...] ask us for our metadata files"* (Scientist F, 2019, para. 65).

Two other respondents (E, H) expressed the need for the automatic transfer of metadata as they often had to "write the metadata by hand" (Scientist E, 2019, para. 30). The specific information that had to be documented in order to make results understandable, reproducible and (re)usable differed in various disciplines. This is why certain flexibility is required when it comes to technical infrastructure.

5. Discussion

Although there are numerous publications which demonstrate the advantages of sustainable RDM including metadata, the findings emerging from the interviews conducted in this research indicate the lack of understanding and awareness of metadata among scientists. Funding agencies such as the European Commission or the FWF Austrian Science Fund, as well as scientific libraries tend to use the words "metadata" and "metadata standards" in their information materials aimed at researchers. The interviews suggest that only half of the researchers could properly understand the term "metadata". This raises the question about the clarity of the term. Can scholars fill out the field "metadata" in the data management plan without support? The interviews confirmed the recommendation of Read and Surkis, "to speak the language of research, not the language of libraries, for example, talk about describing data, not about […] metadata" (Surkis and Read 2015, p. 156). Interestingly, four of the five scientists who placed the term correctly had either studied or worked in research outside of Austria. This could be due to increased awareness of the topic or a highly developed RDM infrastructure in other countries.

However, the interviews also showed that the often used paraphrasing of the concept as "describing data", as recommended by Surkis and Read (ibid., p. 156) and used, for example, in the e-Infrastructures study, did not seem clear enough for the researchers. This outcome of this paper is contrary to the results of the studies conducted at various universities, as well as the e-Infrastructures project (Bauer et al. 2015). The use of adequate standards and the description of data relates to the understanding of the term. Building on Read and Surkis, this study suggests to apply the word "speak" (Surkis and Read 2015, p. 156) – to engage in further discussions with researchers in order to convey the advantages of describing data. The interviews thus reveal a further need for action for libraries in their role as conveyors of information literacy.

Despite the unfamiliarity of the term, the paper identified many different approaches to the description of data in the interviews. This coincides with the findings of the study by Bauer et al. (2015). They could be found in the form of automatically generated metadata, laboratory notebook entries, study protocols in medicine, and as separate descriptive files. Some scientists in the study also considered publications as descriptions of data, which librarians could utilize as "methods metadata" (Chao 2015, p. 83). In her paper, Chao examines articles in the field of soil ecology in order to work out potentially relevant repository metadata fields from the methods section (ibid., p. 86). Furthermore, this finding also holds potential for the communication strategies of libraries. When communicating with scholars, libraries could establish a link to practical examples of data description in the above-mentioned formats, thereby making them more tangible while conveying the additional advantages of metadata and documentation. An in-depth scientific investigation of how these approaches to data description should be translated into more formalized metadata and documentation would therefore be useful.

The interviews indicated that metadata did not play a significant enough role during active collaboration among researchers. However, the advantages of clear documentation became evident when discussing the (re)use of data from others. According to the majority of the interviewed scientists, they would have benefitted from structured and comprehensive information about data. The interviews revealed four motivating factors for the creation of metadata – visibility, findability, (re)usability of data and the reproducibility of results. Not least, they allowed scholars to meet the requirements of scientific journals. There is a growing awareness of the advantages of properly describing data, which libraries can apply to further the creation of metadata and documentation in science.

Throughout the analysis I defined six areas of needs and requirements for metadata and RDM services. Despite the division into six parts, these demands of researchers are interrelated und should therefore not be considered individually. Unsurprisingly, research in science is conducted predominantly in the English language. The interviews thus provide a clear statement also for German-speaking research institutions offering RDM services - they should be designed in both languages. Furthermore, the interviewees observed the need to reflect the differences between various disciplines in RDM services.

Three of the researchers emphasized the issue of discipline-specific expertise when delivering RDM support. The University should provide specialist staff to take on RDM on the faculty level. This corresponds to the latest developments in institutional data management practices, as they increasingly provide data stewards with RDM skills, as well as knowledge from a relevant discipline[6]. This confirms the concerns expressed by Surkis and Read (2015) and Martin (2013) that researchers may have about the role of libraries in RDM.

A further obstacle to the increased creation of metadata according to the respondents was that it was time-consuming. This result correlates with the findings of other universities (Delasalle 2013; Lemaire et al. 2016). The interviews revealed a need

[6] See for example the Technical University of Delft - https://www.tudelft.nl/en/library/current-topics/research-data-management/research-data-management/data-stewardship/
or the Technical University of Graz https://www.tugraz.at/institute/isds/research/groups/orrg/

for solutions that are as simple and as easy to use as possible. The question of efficiency played a part in both technical infrastructure and training.

Other aspects that were important when it came to technical solutions were usability, automation and the integration of local solutions. Thus, approaches such as usability testing of metadata schemata in repositories seem reasonable (Petritsch 2017, p. 202). Two scientists who produced a large amount of metadata suggested for these to be transferred to centralized infrastructures such as repositories automatically.

Further comments by the researchers could be summarized in the category Information Transfer. Half of the researchers considered consulting and support to be an important measure. This reflects the findings of other studies by Feldsien-Sudhaus and Rajski (2016) and Krähwinkel (2015). Two of the respondents had already taken part in RDM training and expressed their positive experiences. In addition, two of the study participants said that they preferred independent learning. These results underline the need for different ways of conveying information based on various learning strategies. These largely coincide with library services already on offer (Tenopir et al. 2017). However, there is also potential to design targeted, practice-oriented, subject-specific training in order to improve the way libraries support researchers in their concrete projects. In combination with the demand for efficiency, the expansion of information transfer to independent learning also seems desirable.

6. Conclusion

Despite the omnipresence of the term “metadata” in the communication between libraries and funding agencies on the one hand, and researchers on the other, the interviews manifested the lack of understanding and awareness of the term among the researchers these institutions try to address. Nevertheless, the study identified various ways researchers described their data. These included automatically generated metadata, entries in laboratory notebooks or the depiction of methods in publications. Findability and (re-)usability of data, as well as reproducibility of results and requirements of academic publishers presented the motivating factors for the creation of metadata and documentation.

These findings suggest that when developing new RDM solutions, service providers should:

- Design services as well as technical solutions in both English and German
- Focus on efficiency
- Take interdisciplinary differences into account
- Provide qualified staff with discipline-specific expertise as data stewards
- Develop user-friendly technical infrastructure that (automatically) integrates local solutions
- Disseminate relevant information in the form of support and training, allowing for independent learning

It is important to note that due to the difficult access to the field and limited timeframe, no theoretical sampling was used and the results should therefore be considered accordingly. Nonetheless, the findings opened up many opportunities for libraries and other providers of research data management services to further increase awareness of metadata and documentation of research data in the natural sciences. According to the Grounded Theory, “theory as process can be presented in publications as a momentary product, but it is written with the assumption that it is still developing” (Glaser and Strauss 1967, p. 32). Although the paper provides relevant results, these are only the first insights into the needs and requirements of scientists in the field of metadata for research data. These demands change continuously and therefore require further research, for example in the form of usability testing of concrete solutions. Furthermore, qualitative observations of focus groups, where researchers can speak with their peers, could complement the information this paper presents.

References

Bauer, B.; Ferus, A.; Gorraiz, J., et al. (2015). *Researchers and their data: results of an Austrian survey : report 2015*. Version 1.0. [Wien]. DOI: 10.5281/zenodo.34005.

Blumesberger, S. (2018). Metadaten als Mehrwerte. In *Praxishandbuch Informationsmarketing: Konvergente Strategien, Methoden und Konzepte*. Ed. by Georgy, U.; Schade, F. De Gruyter, pp. 279–291. 🔒 DOI: 10.1515/9783110539011-018.

Bryant, A. (2017). *Grounded theory and grounded theorizing: pragmatism in research practice*. New York: Oxford University Press.

Büttner, S.; Hobohm, H. C.; Müller, L., eds. (2011). *Handbuch Forschungsdatenmanagement*. Bad Honnef: Bock + Herchen. DOI: 10.34678/opus4-208.

Chao, T. (2015). Mapping methods metadata for research data. In *International Journal of Digital Curation* 10(1), pp. 82–94. DOI: 10.2218/ijdc.v10i1.347.

Charmaz, K. (2014). *Constructing grounded theory*. 2nd edition. Introducing Qualitative Methods. Los Angeles ; London ; New Delhi: SAGE.

Delasalle, J. (2013). Research data management at the University of Warwick: recent steps towards a joined-up approach at a UK university. In *LIBREAS. Library Ideas* 9(2), pp. 97–105. https://libreas.eu/ausgabe23/10delasalle/ visited on January 14, 2021.

Farnel, S.; Shiri, A. (2014). Metadata for research data: current practices and trends. In *Proceedings international Conference on Dublin Core and Metadata Applications*. International Conference on Dublin Core and Metadata Applications, pp. 74–82. https://dcpapers.dublincore.org/pubs/article/view/3714/1937 visited on January 14, 2021.

Feldsien-Sudhaus, I.; Rajski, B. (2016). *Digitale Forschungsdaten für die Zukunft sichern: Umfrage zum Umgang mit Forschungsdaten an der TU Hamburg: Auswertung*. DOI: 10.15480/882.1326.

Forschungsdatenmanagement und Bibliotheken (2016). In *Bibliotheksdienst* 50(7), pp. 614–615. DOI: 10.1515/bd-2016-0075.

Glaser, B. G. (1978). *Theoretical sensitivity: advances in the methodology of grounded theory*. 4. print. Mill Valley, Calif.: Sociology Press.

Glaser, B. G.; Strauss, A. L. (1967). *The discovery of grounded theory: strategies for qualitative research*. New York: Aldine.

Hey, T.; Tansley, S.; Tolle, K. (2009). *The fourth paradigm: data-intensive scientific discovery*. Redmond, Washington: Microsoft Research. https://digital.library.unt.edu/ark:/67531/metadc31516/ visited on January 14, 2021.

Hey, T.; Trefethen, A. (2003). The data deluge: an e-science perspective. In *Grid computing: making the global infrastructure a reality*. Ed. by Berman, F.; Fox, G.; Hey, T. John Wiley & Sons, Ltd, pp. 809–824. DOI: 10.1002/0470867167.

Jensen, U.; Katsanidou, A.; Zenk-Möltgen, W. (2011). Metadaten und Standards. In *Handbuch Forschungsdatenmanagement*. Ed. by Büttner, S.; Hobohm, H. C.; Müller, L. Bock + Herchen Verlag, pp. 83–100.

Krähwinkel, E. (2015). *Forschungsdatenmanagement an der Philipps-Universität Marburg. Die Ergebnisse der Umfrage zum Forschungsdatenmanagement im November 2014*. DOI: 10.17192/es2015.0019.

Lemaire, M.; Rommelfanger, Y.; Ludwig, J., et al. (2016). *Umgang mit Forschungsdaten und deren Archivierung. Bericht zur Online-Bedarfserhebung an der Universität Trier*. https://ubt.opus.hbz-nrw.de/frontdoor/index/index/docId/737 visited on January 14, 2021.

Martin, C. L. (2013). Wissenschaftliche Bibliotheken als Akteure im Forschungsdatenmanagement. In *LIBREAS. Library Ideas* 23. https://libreas.eu/ausgabe23/03martin/ visited on January 14, 2021.

Matusiak, K. K.; Sposito, F. A. (2017). Types of research data management services: an international perspective. In *Proceedings of the Association for Information Science and Technology* 54(1), pp. 754–756. 🔒 DOI: 10.1002/pra2.2017.14505401144.

Morgan, A.; Duffield, N.; Walkley Hall, L. (2017). Research data management support: sharing our experiences. In *Journal of the Australian Library and Information Association* 66(3), pp. 299–305. DOI: 10.1080/24750158.2017.1371911.

Park, M. S.; Park, H. (2019). An examination of metadata practices for research data reuse: characteristics and predictive probability of metadata elements. In *Malaysian journal of library & information science* 24(3), pp. 61–75. DOI: 10.22452/mjlis.vol24no3.4.

Petritsch, B. (2017). Metadaten für Forschungsdaten in der Praxis. In *VOEB-Mitteilungen* 70(2), pp. 200–207. DOI: 10.31263/voebm.v70i2.1678.

Pomerantz, J. (2015). *Metadata*. The MIT Press Essential Knowledge Series. Cambridge, Massachusetts: The MIT Press. 🔒 https://mitpress.mit.edu/books/metadata visited on January 14, 2021.

Pryor, G. (2012). Why manage research data? In *Managing research data*. Ed. by Pryor, G. Facet, pp. 1–16. 🔒 DOI: 10.29085/9781856048910.

Rapport, F.; Braithwaite, J. (2018). Are we on the cusp of a fourth research paradigm? predicting the future for a new approach to methods-use in medical and health services research. In *BMC medical research methodology* 18, 131. DOI: 10.1186/s12874-018-0597-4.

Surkis, A.; Read, K. (2015). Research data management. In *Journal of the Medical Library Association* 103(3), pp. 154–156. DOI: 10.3163/1536-5050.103.3.011.

Tenopir, C.; Talja, S.; Horstmann, W., et al. (2017). Research data services in european academic research libraries. In *LIBER quarterly* 27(1), pp. 23–44. DOI: 10.18352/lq.10180.

Wiley, C. A.; Burnette, M. H. (2019). Assessing data management support needs of bioengineering and biomedical research faculty. In *Journal of eScience Librarianship* 8(1), e1132. DOI: 10.7191/jeslib.2019.1132.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In *Scientific data* 3, 160018. DOI: 10.1038/sdata.2016.18.

Zook, M.; Barocas, S.; boyd, d., et al. (2017). Ten simple rules for responsible big data research. In *PLoS computational biology* 13(3), e1005399. DOI: 10.1371/journal.pcbi.1005399.

Interviews

Scientist A (2019) “Interview with the author”. DOI: 10.5281/zenodo.3897321.

Scientist B (2019) “Interview with the author”. DOI: 10.5281/zenodo.3897321.

Scientist C (2019) “Interview with the author”. DOI: 10.5281/zenodo.3897321.

Scientist D (2019) “Interview with the author”. DOI: 10.5281/zenodo.3897321.

Scientist E (2019) “Interview with the author”. DOI: 10.5281/zenodo.3897321.

Scientist F (2019) “Interview with the author”. DOI: 10.5281/zenodo.3897321.

Scientist G (2019) “Interview with the author”. DOI: 10.5281/zenodo.3897321.

Scientist H (2019) “Interview with the author”. DOI: 10.5281/zenodo.3897321.

A. Interview Guide

- original in German, translated by the author

[Introduction, consent form, installation of recording equipment]

Opening question

- Please introduce yourself and briefly describe your area of research.

Open topic lead-in

- Please define what you consider research data in your discipline.
- Does the term metadata mean anything to you? What do you imagine metadata to be? [present definition if necessary]

Metadata in practice

- When you do research, what kind of metadata do you create for your data? How do you describe your data?
- Please describe what the metadata are used for.
- Please describe the role of metadata when working in a team (based on your last research project).
- If you use research data from others, what has been your experience with metadata from these data sets?

In-depth questions (if necessary)

- In what language do you describe your research data?
- Do you apply metadata standards? Which ones do you use?
- What problems have you encountered with the description of research data?
- What kinds of metadata are important to you when using research data from others?
- Imagine that someone wanted to use your data after the completion of your current research project. How would this be possible?
- Are you aware of any training or consulting services regarding research data management?

Conclusion

- Imagine that anything is possible. What services would you want the university to provide for the creation of metadata?
- Would you like to add anything else?

[Thanks for participation; fact questionnaire including e.g. age]

B. Coding System

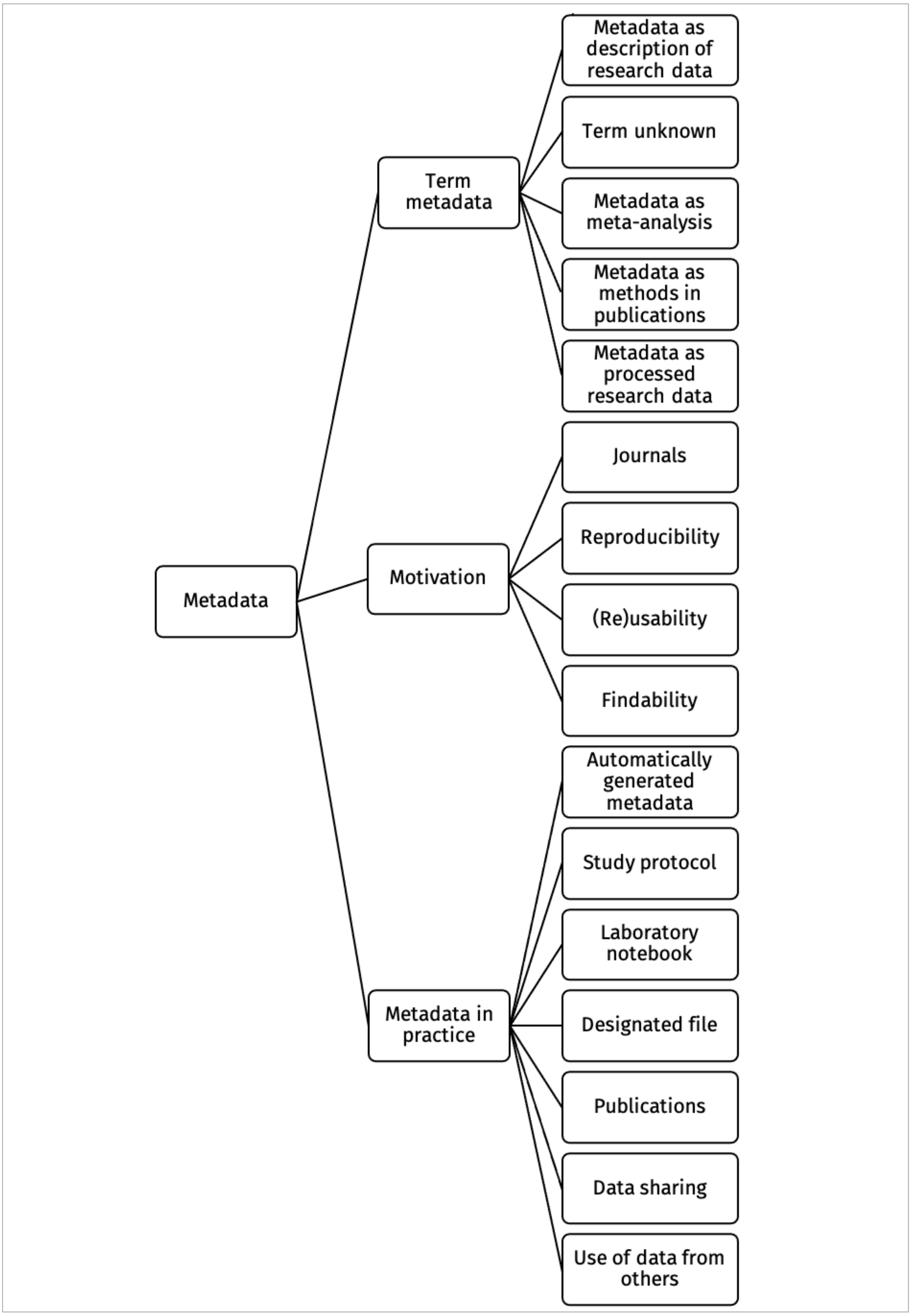

Figure 1: Coding system: category "Metadata" (original in German, translated by the author)

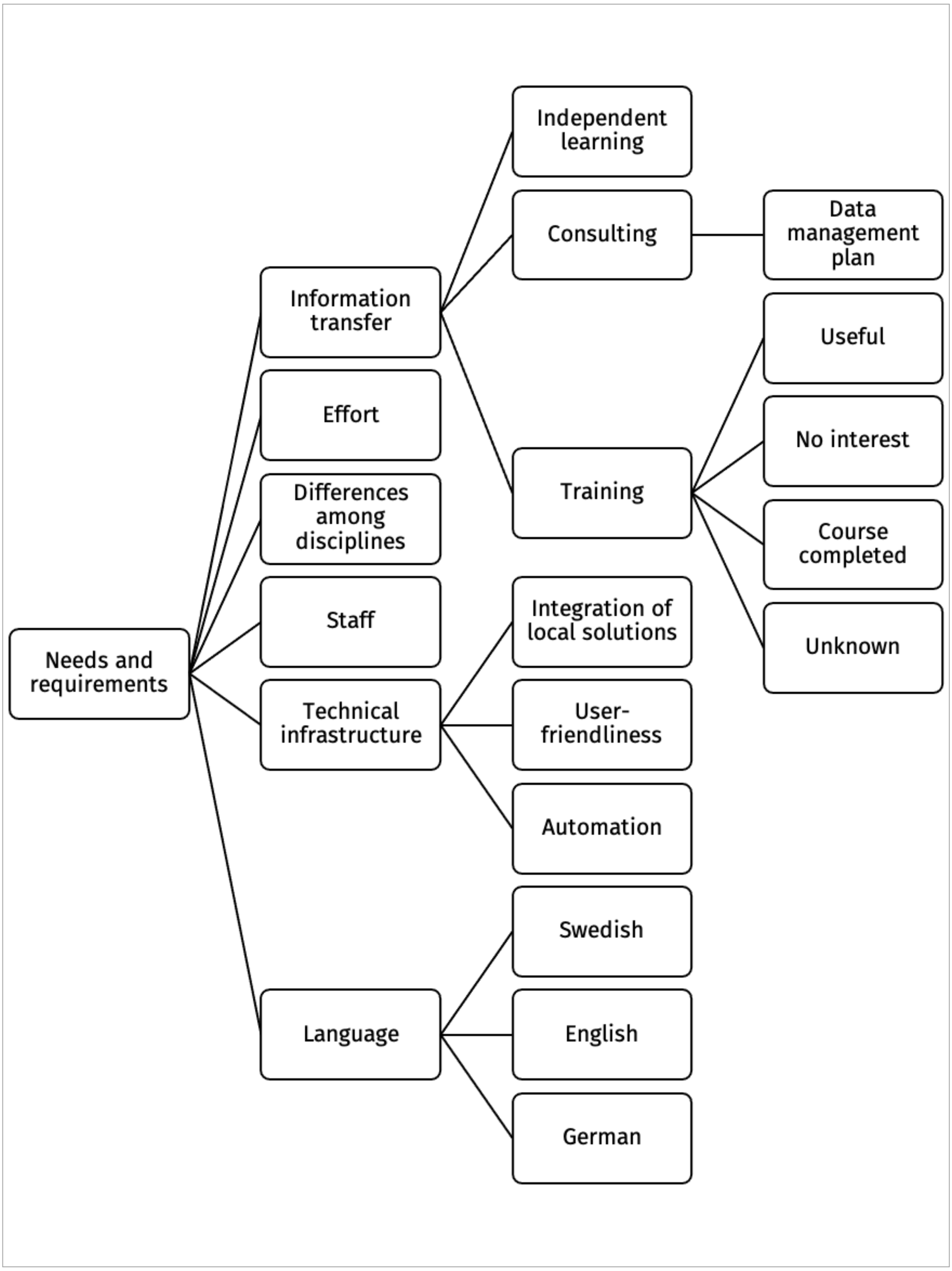

Figure 2: Coding system: category “Needs and requirements” (original in German, translated by the author)

- Categories “Research data” and “Academic discipline” are not part of these figures

C. Original Quotations and Translations

Translation	Original
“[Metadata] are the second level of data processing. So if, for example, I write a review, like I published it a few weeks ago, where I then summarize all the work on a specific topic” (Scientist D, 2019, para. 7)	»[Metadaten], das ist die zweite Ebene der Verarbeitung von Daten. Also, wenn ich zum Beispiel einen Review schreibe, wie ich’s vor paar Wochen veröffentlicht hab’, wo ich dann alle Arbeiten zu einem bestimmten Thema zusammenfasse« (Wissenschaftler D, 2019, A. 7)
“[I]t was partly impossible to find where data were available at all. In what form they are available [...] And of course one has to know [...] that they exists” (Scientist G, 2019, para. 56)	»es war teilweise nicht auffindbar, wo überhaupt Daten vorhanden sind. In welcher Form sie vorhanden sind [...] Und da muss man natürlich auch [...] wissen, dass es das überhaupt gibt« (Wissenschaftler G, 2019, A. 56)
“What is always important to me is data on or metadata about the creation and time and processing ...of data [...] so that I can always understand when I did something” (Scientist H, 2019, para. 11)	»Was für mich immer wichtig ist, sind Daten zur oder Metadaten bezüglich der Erstellung und des Zeitpunkts und der Bearbeitung ... von Daten. [...] dass ich auch immer nachvollziehen kann, wann ich was gemacht habe« (Wissenschaftler H, 2019, A. 11)
“If I want to know how I did it back then, I look at the publication and the methods. Because that would be doubling of descriptive methods for me, why?” (Scientist D, 2019, para. 14)	»Wenn ich wissen möchte, wie ich’s damals gemacht hab, dann schau’ ich in die Publikation nach, in den Methoden nach. Weil das wäre für mich doppelte beschreibender Methoden, wozu?« (Wissenschaftler D, 2019, A. 14)
“One makes interpretations with the data. And of course you have to talk to each other very intensively, because a physicist makes different interpretations than a chemist” (Scientist B, 2019, para. 36)	»Mit den Daten macht man ja entsprechende Interpretationen. Und ähm da muss man natürlich dann sehr intensiv miteinander reden, weil ein Physiker macht andere Interpretationen als ein Chemiker« (Wissenschaftler B, 2019, A. 36)
“There was no explanation for the classification of certain patient characteristics, where a numerical system was used [...] The data were coded in the data set as one, two, three, four, five and I did not know which patient had received chemotherapy” (Scientist H, 2019, para. 21)	»Bei der Klassifizierung von bestimmten Patientenmerkmalen, wo man sich dann halt eines numerischen Systems bedient hat [...] Die Daten waren dann im Datensatz nur codiert als eins, zwei, drei, vier, fünf und ich wusste jetzt nicht welcher Patient hat eine Chemotherapie bekommen« (Wissenschaftler H, 2019, A. 21)
“We all speak a different language. So, biologists speak differently from chemists, and we in turn speak differently from physicists [...]. Everyone tries explaining things to others [...] in terms of their own data” (Scientist B, 2019, para. 24)	»Wir sprechen jeweils eine unterschiedliche Sprache. Also, Biologen reden anders als Chemiker, und wir reden wieder anders als Physiker [...] jeder versucht sich irgendwo darauf einzulassen, dem anderen das [...] mit einem Rahmen zu erklären, was derjenige dann da [...] an eigenen Daten vorliegen hat« (Interview Wissenschaftler B 2019, A. 24)
“[I] now and then see that there are offers [...] But then these are more like series of lectures or something. Concrete training somehow for maybe, I haven’t seen something for junior scientists or something like that” (Scientist H, 2019, para. 43)	»[ich] sehe ab und an, dass es Angebote gibt [...] Aber das sind dann eher Vortragsreihen oder sowas. Konkrete Schulungen irgendwie für vielleicht Junior Scientists oder so ähnlich habe ich [...] jetzt noch nicht gesehen« (Wissenschaftler H, 2019, A. 43)
“[T]hat was an interesting question. So, I didn’t really think about it, but now in retrospect [...], that is actually a relatively good idea to get it right from the start” (Scientist G, 2019, para. 73)	»Das war ’ne interessante Fragestellung. Also, ich hab’ nicht wirklich darüber nachgedacht aber jetzt im Nachhinein [...], also dass das eigentlich eine relativ ganz gute Idee ist, das von Anfang an richtig zu machen« (Wissenschaftler G, 2019, A. 73)
“I learn everything by myself, or I draw on the recommendations and experiences of other colleagues” (Scientist A, 2019, para. 45)	»Ich lerne alles selber, beziehungsweise tu’ ich auf Empfehlungen und Erfahrungen der anderen Kollegen« (Wissenschaftler A, 2019, A. 45)

Table 2: Original quotations and translations (translated by the author)

Eingereicht: 18.09.2020; revidiert: 08.01.2021; akzeptiert: 13.01.2021; publiziert: 01.02.2021