FernUniversität in Hagen

– Fakultät für Mathematik und Informatik –

Lehrgebiet Datenbanken und Informationssysteme

A Requirements Review Based on a User Survey for a University-Wide User-Centered Research Data Management System

Abschlussarbeit zur Erlangung des akademischen Grades

Bachelor of Science (B.Sc.)

vorgelegt von

Sabine Diemt

Matrikelnummer: 8589453

Referentin : Prof. Dr. Uta Störl

Betreuerin : Dr.-Ing. Tanja Auge

DECLARATION

Ich erkläre, dass ich die Bachelorarbeit selbstständig und ohne unzulässige Inanspruchnahme Dritter verfasst habe.

Ich habe dabei nur die angegebenen Quellen und Hilfsmittel verwendet und die aus diesen wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht.

Die Versicherung selbstständiger Arbeit gilt auch für enthaltene Zeichnungen, Skizzen oder graphische Darstellungen.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form weder derselben noch einer anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Mit der Abgabe der elektronischen Fassung der endgültigen Version der Arbeit nehme ich zur Kenntnis, dass diese mit Hilfe eines Plagiatserkennungsdienstes auf enthaltene Plagiate geprüft werden kann und ausschließlich für Prüfungszwecke gespeichert wird.

Manama (Kingdom of Bahrain), 22-February-2024

Signiert von:	Sabine Diemt
Datum:	22.02.2024 21:26:56

Dieses Dokument ist digital signiert!
Dieses mit einer qualifizierten elektronischen Signatur versehene Dokument hat gemäß Art. 25 Abs. 2 der Verordnung (EU) Nr 910/2014 vom 23. Juli 2014 ("eIDAS-VO") die gleiche Rechtswirkung wie ein handschriftlich unterschriebenes Dokument.

Prüfinformation: Informationen zur Prüfung der elektronischen Signatur finden Sie unter: www.a-trust.at/pdf

Sabine Diemt

ABSTRACT

This thesis evaluates the needs of potential users for a university-wide user-centered Research Data Management System (RDMS). Through a literature review and online survey (n = 269), possible RDMS functionalities in accordance with the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) were explored.

Key Findings:

- Respondents highlighted a lack of familiarity with Research Data Management (RDM) and RDMS.
- The survey identified critical attributes users would prefer in a new RDMS, including support across all research data cycle phases, integration with familiar software, and user-friendly interfaces.
- While digital natives showed no greater willingness to use an RDMS, familiarity with RDM principles significantly increased user openness to adopting an RDMS across the entire research data cycle.
- User needs were analyzed by scientific field, position, and preferred usage methods (web application, mobile application, local installation).
- Integration with commonly used research software and additional features requested by users were also assessed.

Recommendations: The most important aspect to a user-centered system that can be used across all disciplines is a Graphical User Interface (GUI) with adaptations, depending on the scientific field of the user. The thesis proposes best practices and a comprehensive list of requirements for developing a user-centered university-wide RDMS, aligned with the FAIR principles, addressing identified user needs, and based on the devised criteria for a GUI. Additionally, the thesis evaluates existing RDM concepts at two German universities for their user-centeredness.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

API	Application Programming Interface
DCP	Data Curation Profile
DFG	German Research Foundation
DMP	Data Management Plan
DOI	Digital Object Identifier
FOS	Field of Science and Technology
GDPR	General Data Protection Regulation
GUI	Graphical User Interface
IDE	Integrated Development Environment
NDA	Non-Disclosure Agreement
NFDI	German National Research Data Infrastructure
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
ÖFOS	Austrian Fields of Science and Technology Classification
RDM	Research Data Management
RDMS	Research Data Management System
STEM	Science, Technology, Engineering and Math
UR	University of Regensburg
URL	Uniform Resource Locator
UX	User Experience
W3C	World Wide Web Consortium
WWTF	Vienna Science and Technology Fund
ZB	Zettabyte

1

INTRODUCTION

The introduction clarifies the problem at hand and provides a brief overview of the most important terms. It continues to describe the objectives and closes with the general structure of the thesis.

1.1 PROBLEM

Nowadays, the majority of information produced is digital. Each person generates 1.7 MB of data per second on average. [BM21] Whilst the amount of data and information created, captured, copied, and consumed worldwide was only 2 Zettabyte (ZB) in 2010, it had already reached 33 ZB in 2018, and was estimated at 175 ZB for 2025 (1 ZB = 1 billion terabytes).[1] Equally, the amount of research data is growing.

Research data refers to all data that was collected for scientific purposes, including working with measurements, surveys, and documentary material, but excluding interpretative publications such as journal articles or dissertations. This data represents a base for research and its results.[2]

Research Data Management (RDM) covers the entire research cycle, from data collection to digital preservation. It takes into account methods and concepts as well as organizational and technical procedures. The research data cycle as per Biernacka and Schulz is shown in Figure 1.1.

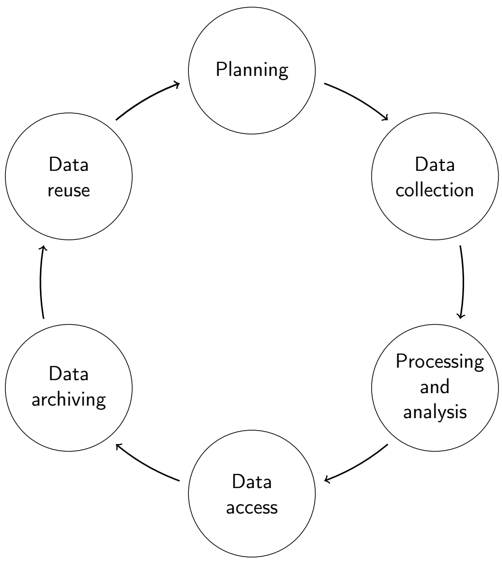


Figure 1.1: Research Data Cycle. Adapted from [BS22]

1 `https://www.statista.com/chart/17727/global-data-creation-forecasts/`, accessed 29-December-2023

2 `https://forschungsdaten.info/praxis-kompakt/glossar/`, accessed 30-July-2023

It starts with the planning phase during which a *Data Management Plan (DMP)* is drafted and existing research data located. Following the data collection (measurements, experiments, simulations, surveys, etc.) and the description of the data with metadata comes the processing and analysis phase. The data is being cleaned, processed, analyzed, documented, and stored. Data access refers to the access for the public but also for project members. In order to allow future use of the data, it shall be archived accordingly. For potential data reuse, the collected research data has to be made available in a proper manner, and there has to be an option to cite the data. Therewith, existing research data can be collected and used for future projects. Ideally, the entire research data cycle can be managed in one system.

A *Research Data Management System (RDMS)* is such a system and is defined as "the technical framework to collect, describe, and provide research data [..], also called data repository, archive or project database". [CH15]

Research data has to fulfill certain requirements. In particular, it shall be FAIR: *Findable, Accessible, Interoperable, and Reusable*. [Kon23] Those *FAIR principles* according to FORCE11[FOR16] are:

Metadata is data about data and can be divided into descriptive, technical, preservation, rights, and structural metadata and markup languages. [Ril17]

> "**To be Findable:**
>
> F1. (meta)data are assigned a globally unique and eternally persistent identifier.
>
> F2. data are described with rich metadata.
>
> F3. (meta)data are registered or indexed in a searchable resource.
>
> F4. metadata specify the data identifier.
>
> **To be Accessible:**
>
> A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
>
> A1.1 the protocol is open, free, and universally implementable.
>
> A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
>
> A2 metadata are accessible, even when the data are no longer available.
>
> **To be Interoperable:**
>
> I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
>
> I2. (meta)data use vocabularies that follow FAIR principles.
>
> I3. (meta)data include qualified references to other (meta)data.
>
> **To be Re-usable:**
>
> R1. (meta)data have a plurality of accurate and relevant attributes.
>
> R1.1. (meta)data are released with a clear and accessible data usage license.
>
> R1.2. (meta)data are associated with their provenance.
>
> R1.3. (meta)data meet domain-relevant community standards."

Whilst the FAIR principles already define requirements for an RDMS, the individual scientific fields have different methods for data collection and different types of data to be processed and therefore unique requirements for their RDMS.

Many organizations do not utilize a proper RDMS yet, thus they struggle with (parts of) their RDM. However, this is an important aspect of research. On one hand, it is good scientific practice to have FAIR data. On the other hand, several funders have already defined requirements for RDM in order to issue and release grants. Some examples are:

- The German Research Foundation (DFG) is committed to open access for research data. Since 2006, grantees should make their results available online and free of charge.[3] Applicants have to consider data reuse and storage of research data in appropriate repositories.[4] The DFG expects to receive a description of the handling of research data.[5]
- All projects receiving a grant from the Vienna Science and Technology Fund (WWTF) have to adhere to an open access and open science policy since 2022.[6] In addition to offering online and free access to the results, grantees have to follow the principle of openness during the entire research process whenever possible, and apply the FAIR principles.

1.2 OBJECTIVE

The University of Regensburg would like to develop its own RDMS to be used by researchers of all scientific fields. This bachelor's thesis evaluates the needs of potential future users through a survey and studies RDM concepts of universities in terms of user-centeredness. Based on the results of the survey, as well as best practices, it provides a requirements review with the user needs in mind for the upcoming development process.

The results of the survey shall also validate three hypotheses about RDMS:

1. People working in technically oriented fields (e.g. Engineering and Technology, Natural Sciences) rather prefer rich features, whilst people in less technical fields (e.g. Humanities, Social Sciences) prefer an RDMS which is easy to use.
2. People who are not familiar with RDM and/or an RDMS are less likely to be willing to use an RDMS for the entire research data cycle.
3. "Digital natives" [Pre01] are more likely to use an RDMS than previous generations.

3 https://www.dfg.de/de/foerderung/antrag-foerderprozess/faq/open-access-faq, accessed 15-December-2023

4 https://www.dfg.de/de/foerderung/antrag-foerderprozess/faq/forschungsdaten-faq, accessed 15-December-2023

5 https://www.dfg.de/de/grundlagen-rahmenbedingungen/grundlagen-und-prinzipien-der-foerderung/forschungsdaten, accessed 15-December-2023

6 https://www.wwtf.at/funding/our-principles/downloads/open-access-and-open-science-policy, accessed 15-December-2023

A complementary bachelor's and master's thesis cover existing RDMS in Germany and their advantages, and expert interviews with representatives of different domains of the association German National Research Data Infrastructure (NFDI). In her bachelor's thesis, Franz [Fra24] provides an interdisciplinary overview of best practices and their challenges, based on the standards and methods identified through expert interviews with the NFDIs. Together, these three works shall consider different aspects and therefore provide a holistic view of the requirements, standards, and best practices for a university-wide user-centered RDMS.

1.3 STRUCTURE

This thesis outlines the current state of research and a possible lack thereof, and examines the relevance of possible criteria for such a system in Chapter 2. Chapter 3 describes the methodology and scientific approach used. It explains the cleaning and preparation of the survey data before analyzing the collected data and concludes with an evaluation of RDM concepts of universities. In Chapter 4, this thesis is compared with existing research, the role of ontologies is explored, and critical comments are addressed. Chapter 5 summarizes the thesis and its results and provides an outlook for future works. Chapter 6 contains supplemental data on the user survey as well as more detailed data of the analysis.

2 BACKGROUND

Chapter 2 contains a literature review to assess which data was already collected on the topic. It evaluates if there are criteria and best practices for a user-centered RDMS in those papers. The relevance of existing literature for a university-wide user-centered system is assessed, and criteria for an RDMS are derived.

2.1 CURRENT STATE OF RESEARCH

Most research about RDMS was done with an emphasis on data sharing and the technical aspects of such a system, especially for libraries. When user needs were evaluated, this was done mostly as an incentive for researchers to share their data rather than how they could make the best use of an RDMS. We concentrated our literature review on papers covering user or system design needs for RDMS, regardless of the scientific field. On the one hand, these were the results of surveys and interviews conducted with current or future users of such a system. On the other hand, we looked into papers referring to any requirements relevant for usability. The subsequently mentioned papers in chronological order addressed at least some user needs.

Anderson et al. [And+07] looked at the needs of academic biomedical researchers in the Pacific Northwest (USA), especially the University of Washington, and what barriers exist to address those needs. Their main focus was data management and analysis needs. They combined an online survey with in-depth one-on-one semi-structured interviews. The most pressing need identified is support for managing and analyzing large datasets, such as searchability and export functionality, alongside organizational factors. Given the age of that study as well as the rapid development of technology and the increasing familiarity of the users thereof, the needs of the researchers may have shifted in the meantime.

Mückschel et al. [Müc+07] wrote about their experiences with collaborative research centers. Those are characterized by long-term (up to twelve years) research projects with scientists of several fields, faculties, and universities. The data management is mostly dictated by the funder's instead of the user's needs, thus the user acceptance is low. Another challenge is insufficient communication between users and software developers, resulting in systems which are not designed in line with the user needs and subsequently not utilized. They indicate that control of the data release date may increase the use of such a system and emphasize the importance of metadata for future use. The use of open source software not only shows monetary advantages but also offers the opportunity to tailor it to the (changing) user needs. Mückschel et al. identified graphical, standardized, web-based user-interfaces as one criteria for a user-friendly system. Based on user surveys, they defined mandatory and desirable metadata components. The mandatory metadata should cover the source of the data and five key words at the most. Useful metadata shall cover (1) **who** created a data file, (2) **what** is the content of the data file, (3) **where** was the data collected, and (4) **how** was the data collected.

Carlson [Car12] concentrated his studies on librarians who want to support researchers with their needs for managing, sharing, or curating their data. The project was conducted at the Purdue University Libraries and the Graduate School of Library and Information Science at the University of Illinois. As a result, a workshop for reference librarians[1] on the data interview process was designed to use the previously developed *Data Curation Profile (DCP)* Toolkit[2]. Librarians can use this DCP Toolkit for data interviews with researchers in order to build appropriate Data Curation Profiles. One of the main requirements of the researchers that can translate to an RDMS was control over the release of their data. Other than that, the needs varied greatly, depending on the discipline, sub-discipline, and person.

Curdt and Hoffmeister [CH15] reported about the design and implementation of a project database for TR32DB, which is a multidisciplinary, collaborative research project in Germany. They mainly look at the technical aspects in this paper. However, the use of a wizard for metadata input was welcomed by researchers utilizing the RDMS and seems to meet their needs to deal with this task. Curdt and Hoffmeister identified the early and continuous participation of future users throughout the development process as one of the success factors for their project.

Knight [Kni15] focused on the needs of health researchers at the London School of Hygiene and Tropical Medicine, gathered through a survey. He identified that the main challenges for researchers when creating, managing, and sharing data are uncertainties on practices related to data archiving, followed by issues when preparing a data sharing agreement and documentation standards to apply.

Krähwinkel [Krä15] evaluated several aspects of RDM at the University of Marburg through an online survey. Her goal was to achieve an overview of available data and experiences, current practices, existing data cultures, and possible service requirements. She received detailed information on the sources of research data (e.g. lab experiments, gene data, simulations, surveys, etc.) as well as data and file types used. This shows a wide variety of file types, including proprietary ones. Information on the storage requirements were also gathered. The most pressing need of users of an RDMS Krähwinkel found in the area of technical questions (e.g. metadata, standards, archiving), followed by legal questions (e.g. restricted access, handling of sensitive data, use of licenses). Asked for the most important functionality of an RDMS, the majority of the respondents require long-term

1 "Reference librarians recommend, interpret, evaluate, and/or use information resources to help patrons with specific information needs." according to the American Library Association.

2 `https://docs.lib.purdue.edu/dcptoolkit/`, accessed 30-July-2023

safe storage (minimum ten years), quality testing of the data (e.g. readable without errors), a permanent addressing/citability, protection of sensitive data (e.g. de-personalization of data), an agreement with the data users, the possibility to grant differentiated access rights, user-friendly access (e.g. direct download of research data), a web-interface with search functionality, provisioning for reuse online, the possibility to select a license for data reuse (e.g. a Creative Commons License), and the possibility to describe and index the data (e.g. with additional information and metadata).

Wiley and Mischo [WM16] conducted in-depth interviews at the University of Illinois Urbana-Champaign amongst members of the engineering and atmospheric science faculty which all produce quantitative data. They were interested in the faculty data management practices and needs. Their literature review uncovered that the needs in data management differ greatly amongst various disciplines. This applies to metadata tagging, and data format and size storable alike. The main issues of researchers concerned the data volume, long-term preservation, making big data user friendly, data backup and distribution, and the organization of the research workflow.

Eifert, Muckel, and Schmitz [EMS16] wanted to implement RDM as a service suite at RWTH Aachen, looking not only into the IT solution but also training and support. Their goal was to achieve acceptance amongst a broad spectrum of future users by addressing both the top management as well as the scientists for their practical requirements. Based on surveys conducted by other researchers combined with the results of their own interviews, they identified a need for supportive services which are user-oriented, intuitive, and can be seamlessly integrated in current processes. Those processes differ, depending on the size of the environment and whether it is national or international, and so do the functional and non-functional requirements of an RDMS. The user requirements that Eifert, Muckel, and Schmitz identified are long-term stability, an appropriate metadata infrastructure, traditional file system access as well as web-based access, data access through the RDMS and the cloud, fast access to resources of any size, a function for easy attaching and linking of metadata, a user-centered handling and functions, and the integration into existing workflows.

Amorim et al. [Amo+17] have compared RDM platforms in terms of architecture, flexible metadata, and interoperability. In the systems compared, they identified domain-specific metadata production, collaboration, and integration of the RDM platforms with research management tools and realtime collaboration as items that need improvement.

Kaye, Bruce, and Fripp [KBF17] examined the development of a research data shared service for the UK by Jisc. Jisc is the non-profit UK digital, data, and technology agency focused on tertiary education, research, and innovation. They collected the requirements of the higher education sector in 2015 by conducting a survey on research systems, doing desk research on the existing requirements of previous Jisc programs [Kay15], and holding requirements gathering workshops. Focus groups on metadata [Fer16] identified issues and requirements of researchers. One of the crucial points on metadata was an early creation with the option to update the metadata at any time. Suggestions for metadata fields were aggregated.

Bugaje and Chowdhury [BC18] have researched the design requirements of a user-centered RDMS at universities in the United Kingdom, in particular the component research data repository. They used four different methods: (1) online questionnaire survey, (2) face-to-face interviews, (3) market appraisal and review, (4) technical experiment comparing data retrieval with traditional information retrieval. According to them, no standards or best practices have yet been defined for evaluating usability or user-centeredness.

Vock [Voc19] conducted an online user survey at Friedrich-Schiller-Universität Jena amongst users of the BExIS (Biodiversity Exploratories Information System) platform. The greatest potentials for improvement seem to be in the areas of data search as well as statistics and visualization. He has received the following specific inputs, amongst others: use of auto-complete, search for date, sorting function (more recent/old first, matching words, etc.), filters, both simple and advanced search, help function (e.g. explanatory fields), possibility of modification of already uploaded data, and implementation of a data transformation tool.

Brenger et al. [Bre+19] worked on project UNEKE at the University of Duisburg-Essen and the RWTH Aachen with the goal to develop a strategic approach for a suitable research data storage. They were focusing on the needs of the scientists, and taking into account currently available solutions as well as the long-term perspectives. The methodology used for getting an overview of the currently used RDM and identifying overlapping needs was an online survey in which 13 German universities participated. In particular, the survey collected current information on the origin of the data, data formats and data management, data privacy and access to the data, logging of data access, entry of metadata, and application of metadata schemata. The rest of the survey dealt with data sharing and its impact. Shared access for the duration of a project is needed for almost half of the project data, thus access control could be an important feature for an RDMS. Given that most departments have very limited time available for data management and mostly no extra staff for data maintenance, the use of an RDMS should not add additional workload.

Arndt [Arn22] lead a survey among the scientific staff of Helmholtz centers in Germany to investigate RDM practices – specifically metadata handling and data publication – and identify service demands of the different communities. All the respondents did their research in Science, Technology, Engineering and Math (STEM) fields. Especially supporting services for RDMS, best practices, and metadata were sought-after. Furthermore, the users are very interested in technical support. The majority of the scientists questioned stored their data in an unstructured way which does not follow any international standard, although they were intrinsically motivated to document metadata. Almost 20 % of the researchers reused existing data, which highlights the importance of proper data management inside an RDMS.

The FAIR principles were only known to around one third of the respondents, and less than half of them applied them.

Fichtner et al. [Fic+22] led the project BE-KONFORM at the Faculty of Medicine of the University of Freiburg in Germany. They used qualitative and quantitative methods to evaluate the needs of future users of a proper infrastructure for RDM, including consulting and training. Half of the data processed at the faculty is personal data, and thus data privacy acts apply. Faculty staff mostly requires support for data archiving. 21.4 % of lab notebook users work solely, and one third predominantly with analog lab notebooks. Most of those groups would like to utilize electronic lab notebooks. Some concerns voiced against digitalization are a lack of technical infrastructure and support, high costs, and a lack of a proper system, amongst others. This study also contains information on how researchers would like to receive support (preferably through information on websites as well as personal support), although this may very well depend on the age group. However, this data was not collected.

2.2 RELEVANCE

Most of the existing research points to the importance of adequate metadata, whilst this is also one of the major issues scientists have. An easy and time-saving handling of metadata could therefore be one of the most important aspects for achieving a user-centered RDMS. Accordingly, several questions about the researchers' preferences in relation to metadata were added to the online survey.

Nevertheless, the user needs for an RDMS were only evaluated on a few occasions in existing papers. The focus of previous studies was mainly on specific scientific fields instead of catering to the requirements of all areas of research. This thesis shall close the gap.

Through systematic analysis of this relevant literature, a set of evaluation criteria for a RDMS were identified. These criteria were subsequently supplemented with additional considerations to ensure comprehensive evaluation:

- Open source [Amo+17]
- Existence of Application Programming Interface (API) [Fer16] [Amo+17]
- Unique identifiers [Raz11] [Amo+17]
- Support for flexible research workflows [EMS16] [Fer16] [Amo+17]
- Compliance with Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [Amo+17]
- Data location/control [Amo+17]
- Multilingual interface [Amo+17]
- Collaborative environment for teams/groups with an access rights system [Amo+17]
- Domain-specific metadata schemata available [EMS16] [Fer16] [Amo+17]

- Mandatory and optional metadata elements [Raz11] [BC18]
- Possibility to set up multiple metadata schemata [Amo+17]
- Automated metadata tagging [Raz11] [Fer16] [KBF17]
- Data export in different formats [Amo+17]
- Data validation (enforcing correct data and metadata structure, file format) [Amo+17]
- Tracking of content changes [Raz11] [Fer16] [Amo+17]
- Control over data release dates [Raz11] [Car12] [Amo+17]
- Control over data expiry date [Raz11]
- Interoperability [Amo+17] [KBF17]
- Free-text search capability [Fer16] [Amo+17]
- Faceted search (narrowing down the search result by applying parameters) [Tun09] [Fer16] [Amo+17]
- Ease of use vs. rich features [EMS16] [Amo+17]
- Web interface and/or local installation [EMS16] [Amo+17]
- Data identification via Digital Object Identifier (DOI) or Uniform Resource Locator (URL) (if DOI does not exist) [Fer16] [Amo+17]
- Tracking of interactions of external entities with the dataset (e.g. citations) [Fer16] [Amo+17]
- Transformation of data within the RDMS [Amo+17]
- Visualization of datasets [BC18]
- Feature for previewing dataset content [BC18]
- Linking between data objects to enhance discoverability [Raz11] [Fer16] [BC18]
- Data provenance (how was the data created, modified, and used) [Raz11] [Fer16]
- Compliance with data regulations (e.g. General Data Protection Regulation (GDPR), ethical guidelines, Non-Disclosure Agreement (NDA)) [Fer16]
- Maximum file size of typical output [WM16]
- Wizards that guide users through processes [CH15]
- Support options (helpdesk, user forum, workshops/training, etc.) [Arn22]
- Encryption of files
- Easy data citation
- Functionality to anonymize research data
- Specific tools that shall be integrated with the RDMS

3 ANALYSIS

We first cover the methodology applied to collect useful data for this thesis. Then we go into the classification of criteria defined, and how they are relevant for research. We examine the steps involved in preparing and cleaning the data for the subsequent analysis. Following that section, the data analysis is divided into four groups: demographic data of survey respondents, the status quo on RDM and an RDMS for their research, the hypotheses presented, and analysis of user needs for an RDMS. The chapter concludes with a summary about RDM concepts of two universities in Germany.

3.1 METHODOLOGY

The software requirements of the future users can be evaluated through interviews (qualitative analysis), a survey (quantitative analysis), or a combination of both. Although the latter would have been ideal to get more in-depth information, due to the time constraint of three months a survey was chosen. This provided a bigger sample for more feedback whilst still allowing for free text. Anonymity is expected to have led to more honest answers. To reduce the risk of repeated participation, a cookie was set. Financial incentives for the participation (e.g. vouchers) were not offered. Since the survey was distributed internationally, the respondents could choose between a questionnaire in German or English.

The design of the survey was based on Mayer [May13]. Demographic questions were placed at the end of the survey, although this meant that data about the scientific field of respondents who only partially filled the survey is missing, thus most of their answers cannot be used. In the interest of keeping questions nonambiguous and avoiding lengthy explanations, the phase "data access" was left out when asking about the research data cycle. A potential confusion of "data access" and "data reuse" was too likely. Survey pretests were conducted after each modification, taking into account the input of the pretest participants. In total, approximately ten to 15 pretests with people conducting research in or studying educational sciences, philosophy, economics and business, computer and information sciences, and medical sciences were performed.

Valuable input as to the requirements of an RDMS can come from people identifying as (former) scientists, researchers, academics, professors, research assistants, data curators, data scientists, and funders, all of which are potential future users of such a system. [Amo+17] Therefore, this was defined as the desired target group.

The FernUniversität in Hagen[1] and the University of Regensburg[2] have 369 professors in total as per the summer term 2023. Together, they cover six major scientific fields and 25 sub-disciplines.

The calculation of our sample size is based on Mayer [May13]:

$$n = \frac{N}{1 + d^2(N - 1)} \tag{3.1}$$

with n: sample size, N: population, d: sampling error (accuracy of the conclusion of ± 10 % applied -> d = 0.1). The significance level is α = 5 %, therefore the certainty of the conclusion is 95 %.

With this calculation, we were looking for at least n = 79 in total. An equal distribution over the six major scientific fields means 14 samples each. Due to a lack of knowledge about the exact number of professors and research assistants per faculty and subject area, a more accurate sample size per scientific field was not determined.

The sample was selected both randomized (through an open call on social media and other platforms) and non randomized (by directly addressing potential respondents). More specifically, the survey was sent out through distribution lists of the FernUniversität in Hagen, the university commissioning this bachelor's thesis. Further researchers and universities were contacted by the thesis advisor and the author. The survey was promoted through social media and distributed to researchers active in online groups (Discord, Microsoft Teams). A simple random sampling was not possible because there is no access to the population. On the other hand, quota sampling was not necessary since input from any respondent of the desired target group can be helpful. The goal was to reach a wide spectrum of potential RDMS users to obtain the best possible input for a future system. It was however desirable to have equal number of responses for each scientific field throughout all age groups for a more representative sample. [May13] The total number of people contacted (population) and thus the response rate is unknown.

To prevent multiple answers from a single respondent, the option "Set Cookie to Prevent Repeated Participation" of the survey tool was used. Repeated participation cannot be ruled out completely if respondents used several devices or deleted the cookies. However, it reduced the probability.

The survey was conducted with the university-owned local installation of the open source tool LimeSurvey of the FernUniversität in Hagen. The analysis of the data received was done using IBM SPSS for Windows, version 29.0.0.0 (241) and Microsoft Excel (Microsoft 365).

3.2 SCIENTIFIC APPROACH

Based on the answers provided by the subjects, a list of requirements for an RDMS will be compiled. It will further be evaluated if those require-

1 https://www.fernuni-hagen.de/universitaet/zahlen.shtml, accessed 27-December-2023

2 https://www.uni-regensburg.de/universitaet/profil/zahlen-und-fakten/index.html, accessed 27-December-2023

ments lead to FAIR data. To support this evaluation, the functional and non-functional requirements asked about in the survey are assigned to the FAIR principle(s) they contribute to (Tables 3.1 and 3.2). Functional requirements are functions or properties (e.g. exception handling) and external interfaces (e.g. format of input/output data); they describe what the system is expected to do. Non-functional requirements cover technical (e.g. operating system) and quality requirements (e.g. reliability); in other words, how the system should work. [SW21]

	Findable	Accessible	Interoperable	Reusable
Existence of an API		•	•	
Integration of other tools			•	
Collaborative environment for teams/ groups with access rights system		•		
Free-text search capability	•			
Search filters (faceted search)	•			
Support for flexible research workflows			•	
Wizards to guide through processes				
Feature for previewing dataset content	•			
Visualization of datasets	•			
Functionality to anonymize research data		•		•
Control over data location				
Tracking of content changes				•
Control over data release dates				
Control over data expiry dates				
Data export in different formats		•	•	•
Possibility of data citation	•			•
Transformation of data within RDMS		•	•	•
Tracking of interactions of external entities with dataset			•	•
Data provenance				•
Unique identifiers for data	•			
Logical links between related data	•		•	•
Mandatory/optional metadata elements	•			•
Automated metadata tagging	•			•
Possibility to set up multiple metadata schemata	•			•
Domain-specific metadata schemata	•			•
Data validation				•
Data identification via DOI/URL	•	•		

Table 3.1: RDMS requirement criteria vs. FAIR principles (functional requirements)

Some system requirements may even be at odds with the FAIR principles. By giving researchers full control over their own data, they can prevent others from finding and accessing it. The same applies to control over data release and expiry dates as well as data location. Nonetheless, these could especially be important for commercial research, thus making an implementation necessary. Wizards guiding users through processes in an RDMS facilitate the usability but do not directly contribute to the FAIR principles.

	Findable	Accessible	Interoperable	Reusable
Availability of preferred usage method		•		
Multilingual interface	•			
Open source		•	•	•
Support options		•		
File encryption				
Compliance with data regulations				•
Full control over own data				
Compliance with OAI-PMH	•	•	•	

Table 3.2: RDMS requirement criteria vs. FAIR principles (non-functional requirements)

File encryption can be an important feature to protect data, but it is also not relevant to them. Neither of the aforementioned requirements are therefore assigned to any FAIR principles.

3.3 DATA CLEANING AND PREPARATION

The number of questionnaires that were filled in during the survey period of 25-October-2023 to 18-December-2023 is summarized in Table 3.3. Any questionnaire with all mandatory fields filled in counted as fully filled questionnaire. Any questionnaire that was opened and closed again without input is counted by LimeSurvey as a partially filled questionnaire.

Number of questionnaires filled (full):	207
Number of questionnaires filled (partial):	202
Total number of questionnaires returned:	**409**
Number of empty questionnaires:	- 139
Number of invalid questionnaires:	- 1
Total number of valid questionnaires returned:	**269**

Table 3.3: Metrics Questionnaires

To filter out potential responses where respondents only clicked through the questionnaire without actually reading the questions, a column called "duration" was added. This column contains the duration in seconds, calculated based on the timestamps of beginning and ending of the processing time. The average duration of fully filled questionnaires was 13 minutes and 16 seconds. 127 cases with a duration of 0 seconds (respondent opened the questionnaire and closed it again without input) were removed. Twelve cases without data were removed. One case was fully filled with only the mandatory fields within 1 minute and 13 seconds; it was considered as an invalid questionnaire and was removed. The remaining fully filled questionnaires were answered within 3 minutes or more with more than only the mandatory fields and therefore, they were considered valid. After data cleaning, 269 valid questionnaires were left, 63 of which were partially filled and 206 fully filled.[3]

The questionnaires filled in German contained some German text in the free text fields. Those answers were translated into English.

The application of the k-anonymity protection model was considered. K-anonymity means that the information that could identify a person differs from at least $k-1$ other persons. Therefore, the correct mapping of the data with a person is aggravated. The bigger k, the better the anonymity but the smaller the information content. [Swe02][Sam01] Recognizing the potential vulnerabilities of k-anonymity, including background knowledge and homogeneity attacks, and the consequent risk of rendering the data unusable, the application of this model was deemed unsuitable. To avoid any identification of respondents, the data received was de-facto anonymized instead. [Sch+22] The combination of position, institution name, and scientific field was a direct personal identifier. A complete recoding of the institution name would not have provided any benefit to this thesis. Therefore, the institution name was removed from the data (deletion of the variable). Software names can constitute indirect identifiers if this software is used exclusively in certain institutions. Therefore, any software name used by less than ten respondents was removed and replaced with the software category (aggregation of individual values of a variable). The input of all free text fields was searched for the aforementioned data, and such data was removed (aggregation of individual values of a variable). [EM15]

3.4 DATA ANALYSIS

The data analysis starts with the demographics of the survey respondents before assessing the status quo on RDM and an RDMS for their research. Then we validate three hypotheses and analyze the user needs for an RDMS.

Whenever "major scientific fields" are mentioned, the 1-digit reporting categories of the Field of Science and Technology (FOS) classification in the Frascati manual (e.g. Humanities) is meant. "Sub-discipline" refers to the respective revised 2-digit classification (e.g. Humanities – History and Archaeology). [Com02]

3.4.1 *Demographics*

We examine first in which countries the survey respondents conduct their research and which scientific fields are represented. Then we check which positions they hold within their respective institutions and within the major scientific fields. We conclude this section with statistics on age and gender of the respondents.

3 The anonymized answers to the survey are available on request through the eLabFTW of the University of Regensburg (unique eLabID: 20240129-b5d5aac48707eb0df0859089da2cdc9170731287).

COUNTRIES: Respondents from all continents participated in the survey (Figure 3.1) which provides insight into RDM practices and user requirements around the world, where regulations may differ. Research cannot be seen inside a local bubble. Instead, it is an international endeavor. Scientists often move to other countries to conduct their research. They bring their individual points of view and needs. Therefore, it is helpful to consider those for a requirements review of an RDMS.

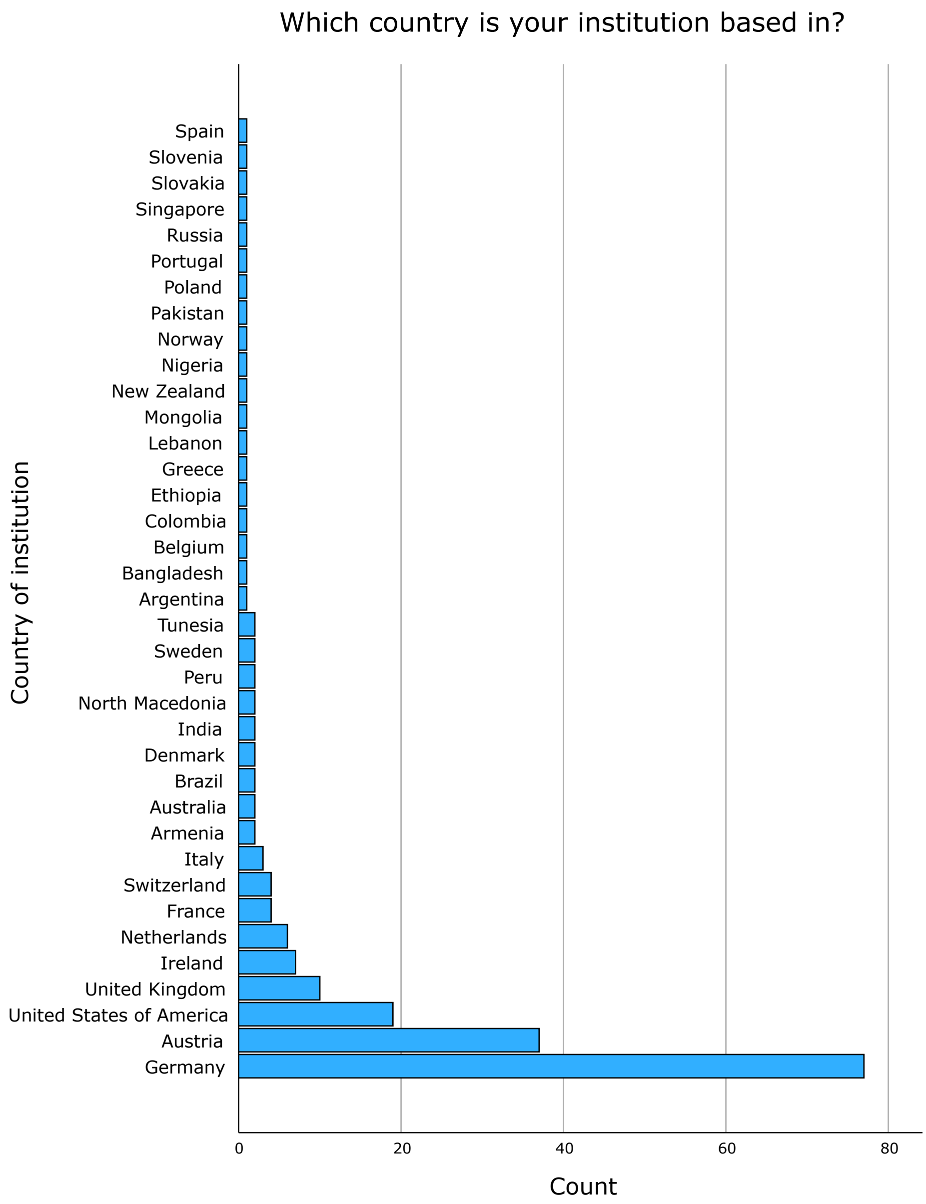


Figure 3.1: Countries of Respondents

The majority of answers are from Germany (77 respondents/37.7 %), followed by Austria (37 respondents/18.1 %), the USA (19 respondents/9.3 %), and the UK (10 respondents/4.9 %). The remaining countries represented are in the single digit range.

SCIENTIFIC FIELDS: The answers of researchers in the fields of natural and social sciences comprise more than half of the data. The rest is shared almost equally between humanities, agricultural sciences, medical and health sciences, and engineering and technology, while the latter claims the biggest portion with 14.5 % (Figure 3.2).

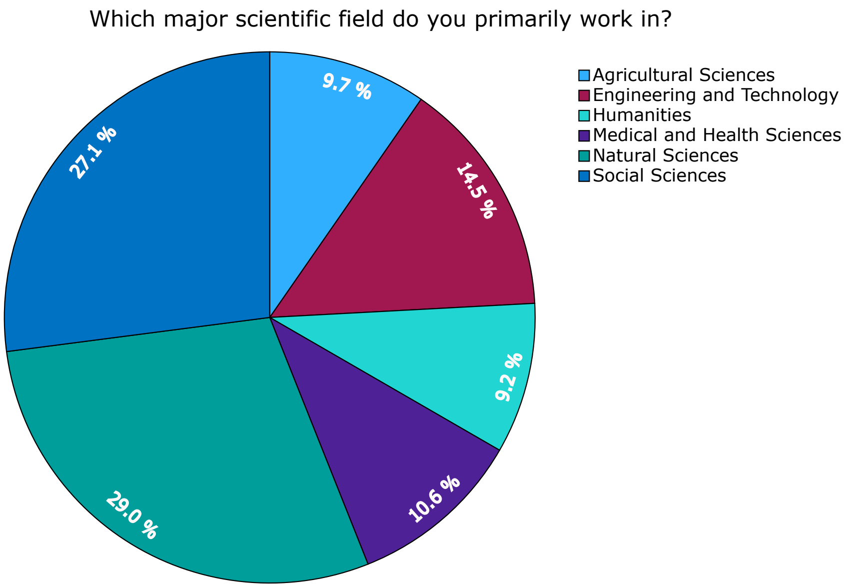


Figure 3.2: Major Scientific Fields of Respondents

POSITION: Respondents could select one or more positions they currently hold (professor, postdoc, research assistant, data curator, data scientist, funder) or enter one which was not listed. During the preparation for data analysis, unlisted positions were either merged into the existing ones or into new groups (PhD student, student, researcher, data steward, manager), according to the information provided. The positions entered with the allocation of groups is detailed in Section 6.5.

Some of the positions can overlap, for example, PhD student and researcher. However, a researcher could also work for a private research institute and not be a PhD student, postdoc, or professor. Therefore, multiple selections were possible. During the survey period, it emerged that the position of a research assistant varies greatly, depending on the institution and country. It was intended for people working in research and assisting, for example, a professor or postdoc. Through conversations it turned out that also students not involved in research hold the title of research assistant. Someone in the position of data steward supports researchers with long-term care and discoverability of their data [Wil], and also provides consulting, training, needs assessment, and requirements engineering[4]. A data curator on the other hand is responsible for the management of the data[5], as well as

4 `https://forschungsdaten.info/praxis-kompakt/glossar/`, accessed 16-January-2024
5 `https://forschungsdaten.info/praxis-kompakt/glossar/`, accessed 16-January-2024

for the quality of the data and metadata, and the compliance with the FAIR principles[6]. Data scientists work with large amounts of data and can support research by preparing, analyzing, and interpreting this data. Funders work for organizations that provide grants to researchers and define the terms under which grants are approved.

The least represented group is the one of funders (2 respondents). Data curators (5 respondents), PhD students (6 respondents), data stewards (7 respondents), and managers (9 respondents) can also be found in the single digit range, closely followed by data scientists (10 respondents), researchers (13 respondents), and students (16 respondents). Most of the answers come from postdocs (37 respondents), professors (41 respondents), and research assistants (82 respondents) (Figure 3.3).

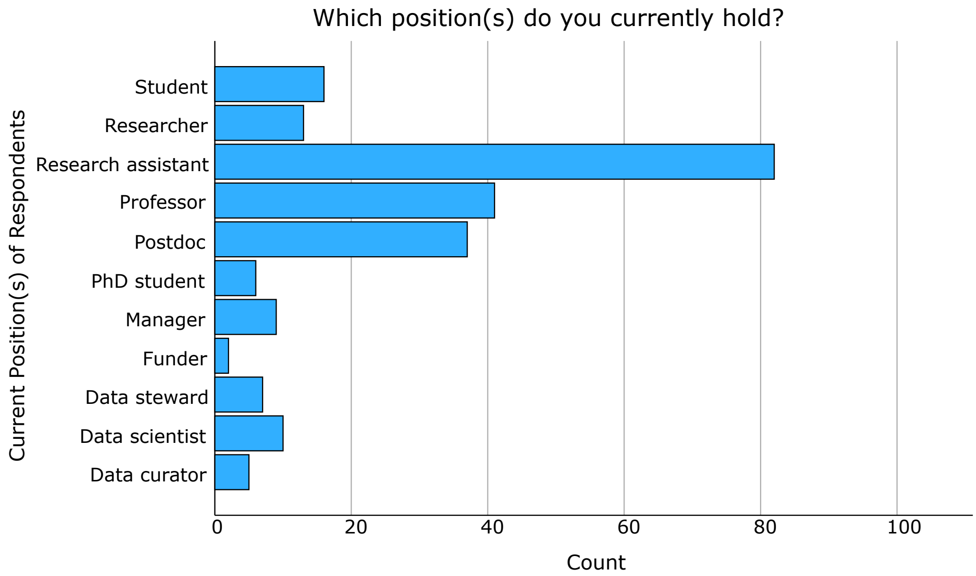


Figure 3.3: Current Position(s) of Respondents

Grouped by the major scientific fields, the following positions are most commonly represented (Figure 3.4):

- Natural sciences:
  Research assistants 42 %, professors and postdocs 17 % each
- Engineering and technology:
  Research assistants 31 %, students 23 %, postdocs and data scientists 11 % each
- Medical and health sciences:
  Professors 36 %, research assistants 16 %, postdocs 12 %
- Agricultural sciences:
  Professors 38 %, research assistants 29 %, postdocs 14 %
- Social sciences:
  Research assistants 40 %, postdocs 21 %, professors 14 %

6 https://www.ugent.be/en/research/openscience/datamanagement/support/data-stewards-curators.htm, accessed 16-January-2024

- Humanities:
  Research assistants 47 %, professors and postdocs 16 % each

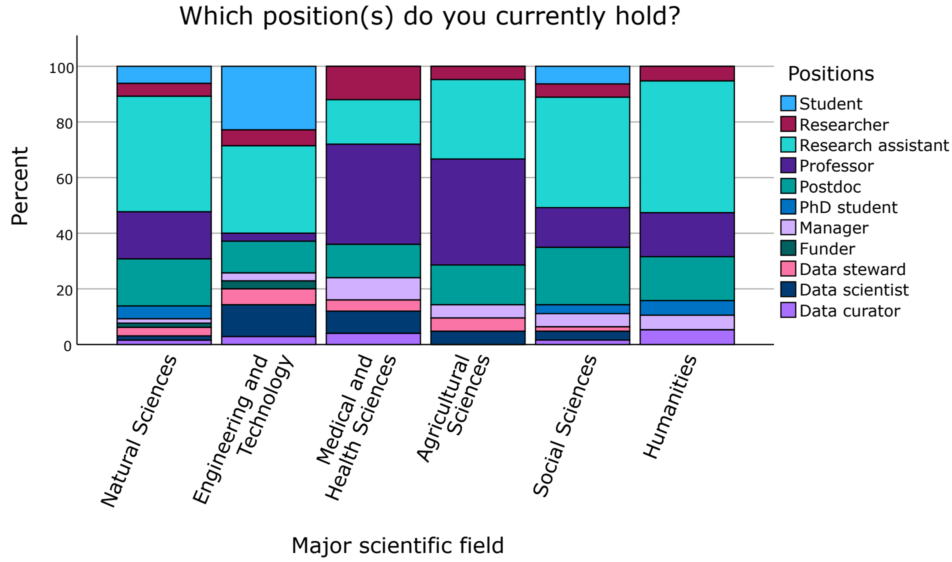


Figure 3.4: Current Position(s) of Respondents within Major Scientific Fields

Considering those positions, it can be argued that students have the least experience with RDM. Therefore, we will take a closer look at the user needs submitted for engineering and technology in Section 3.4.4 to evaluate if the feedback from the students stands out compared to the one of people in other positions.

AGE: A quarter of the respondents, and thus the largest group, are between the ages of 30 and 39. An almost equal share did not provide their age group. The smallest group age-wise are the 60 years and older with only 4.46 % (Figure 3.5).

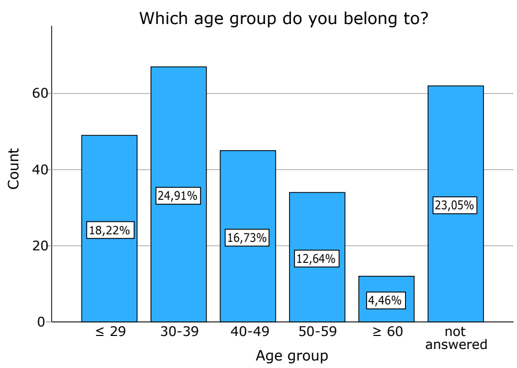


Figure 3.5: Age Groups of Respondents

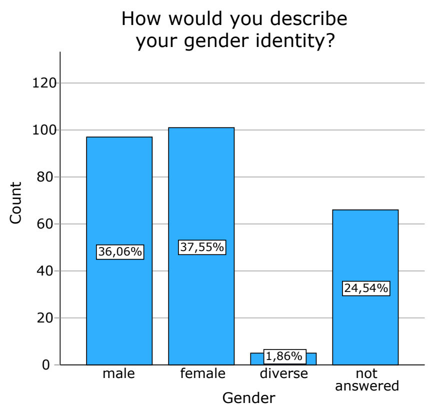


Figure 3.6: Gender of Respondents

GENDER: A quarter of the respondents did not share information about their gender. Male and female almost balance each other, whilst the group of diverse people amounts to 1.86 % (Figure 3.6).

3.4.2 *Status Quo*

The status quo details the familiarity and experience respondents have with RDM and/or an RDMS. It provides information about the current RDMS they use, including the pros and cons. Then it compares the current and desired future use of an RDMS for the phases of the research data cycle. It moves on to important software applications used by respondents. The section closes with insights into types of analog research data, why they are not digitalized, and the storage required for digital data.

FAMILIARITY/EXPERIENCE WITH RDM(S): Respondents were asked if they are familiar with RDM and/or an RDMS. Only five of them decided not to answer. 50.93 % are either not sure or definitely not familiar with RDM and/or an RDMS, whilst 47.21 % have some knowledge of it (Figure 3.7).

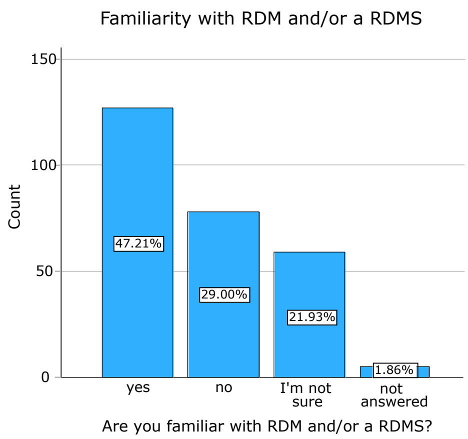


Figure 3.7: Familiarity with RDM and/or an RDMS

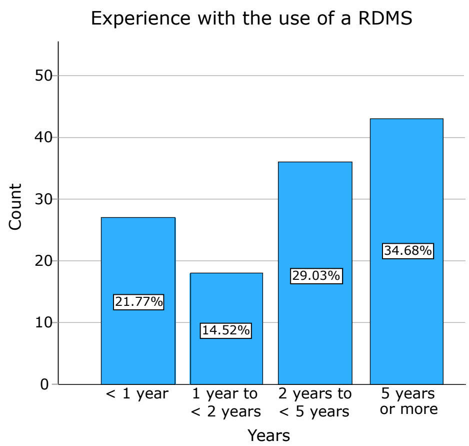


Figure 3.8: Experience with RDMS

Of the 127 participants who are familiar with RDM and/or an RDMS, 124 provided some insight into their experience. Most of them can look back at several years. Slightly more than a third started using such a system five years ago or more, 29.03 % two to less than five years, and 14.52 % one to less than two years ago. 21.77 % of respondents reported having less than one year of experience with an RDMS (Figure 3.8).

Figure 3.9 suggests that RDM and/or an RDMS are equally well established at universities and non-university research institutions, with slightly more than half of the researchers there being familiar with the same. At other institutions, only two out of nine people knew about the concept and technical framework.

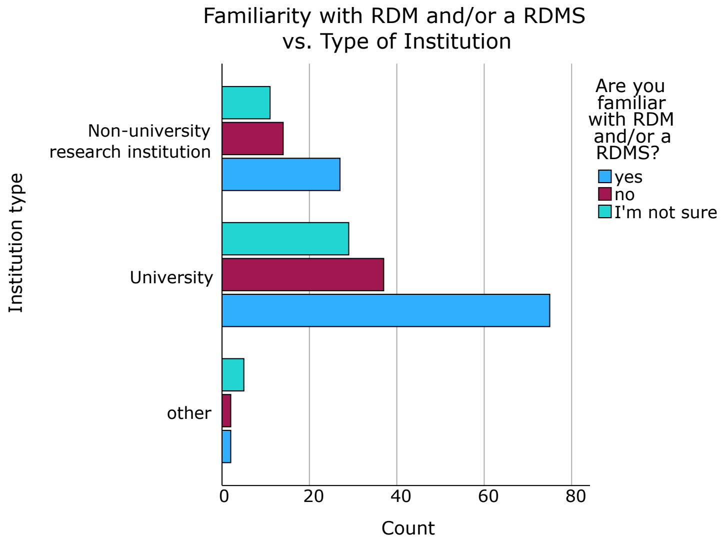


Figure 3.9: Familiarity with RDM and/or an RDMS vs. Type of Institution

Looking at the top three countries where the researchers questioned are based at (Figure 3.10), 49 % of respondents in Germany, 53 % in Austria, and only 39 % in the USA are familiar with RDM and/or an RDMS.

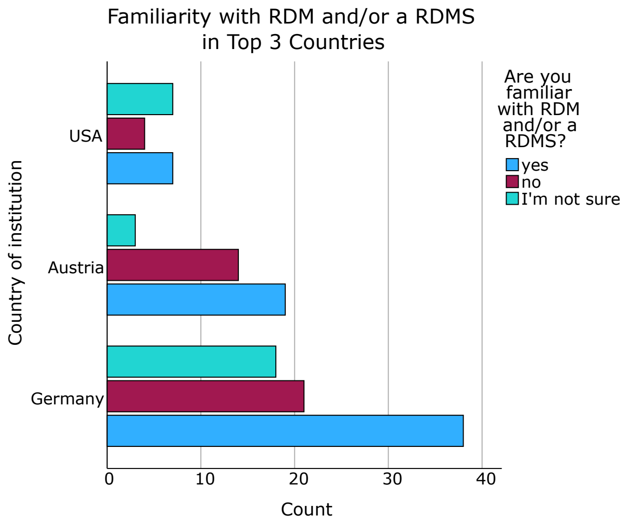


Figure 3.10: Familiarity with RDM and/or an RDMS in Top 3 Countries

Based on the major scientific fields, on average around half of the respondents are familiar with RDM and/or an RDMS, showing the highest familiarity within engineering and technology with 59 % and the lowest within natural sciences with 43 % (Figure 3.11).

The familiarity with RDM and/or an RDMS put into relation to respondent's positions (Figure 3.12), shows the highest familiarity amongst data stewards (100 %), followed by data curators (80 %) and postdocs (68.57 %). PhD students claim almost the same familiarity (66.67 %) as postdocs. Approximately half of the data scientists (55.56 %), professors (53.66 %), and

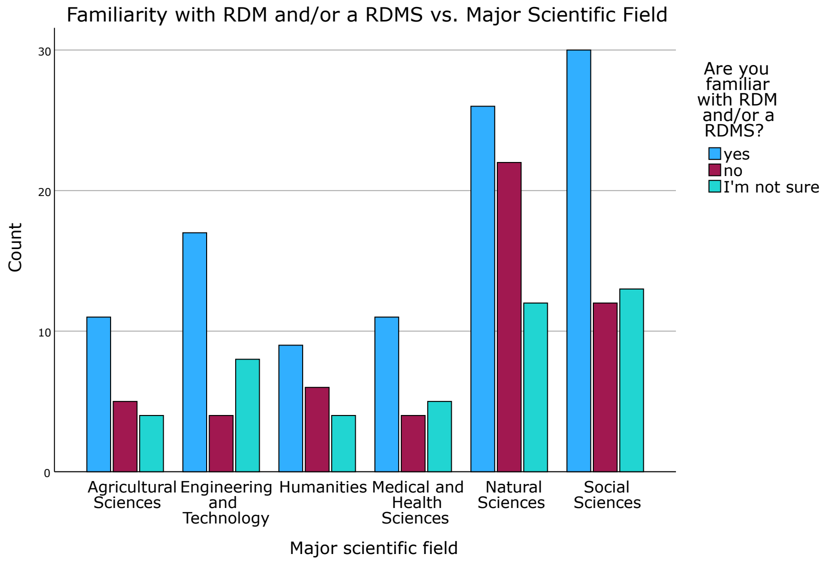


Figure 3.11: Familiarity with RDM and/or an RDMS vs. Major Scientific Field

funders (50 %) are familiar with RDM and/or an RDMS. 46.34 % of the research assistants and 46.15 % of the researchers know about RDM(S), but at the same time, people in these two groups are also among the front-runners of respondents not familiar with RDM and/or an RDMS. The smallest share of individuals who are familiar with RDM and/or an RDMS can be found in the group of managers (33.33 %) and students (31.25 %).

Conclusion: People working in research or with research data know comparatively little about RDM although this is an increasingly important topic. It shows the need for education in this area. Integrating RDM into the curricula of universities when they teach about academic writing could be one approach to tackle this issue early on.

CURRENT RDMS: 60 respondents provided us with the names of their RDMS, eleven of which use more than one system. Since none of the systems were utilized by at least ten respondents, the names were anonymized. Of the 75 systems named, 30 were actually RDMS. The category of 22 applications could not be determined because they were either internally developed or the name was not mentioned. Eight systems were databases. Further categories found were software for data visualization, survey development, statistical analysis, reference management, file hosting, museum management, and data management plans. Besides an actual library, a digital library, data archive, and electronic lab notebook were named as well. The conception of the term RDMS varied throughout all scientific fields.

62 participants rated the degree to which their current RDMS fulfills their needs. The scale was presented with smileys as shown in Figure 3.13, combined with a slider on a scale colored red on the far left side, green on the far right side, and gradually changing to yellow in the middle. In the background, an RDMS not fulfilling the researcher's needs was coded with 1,

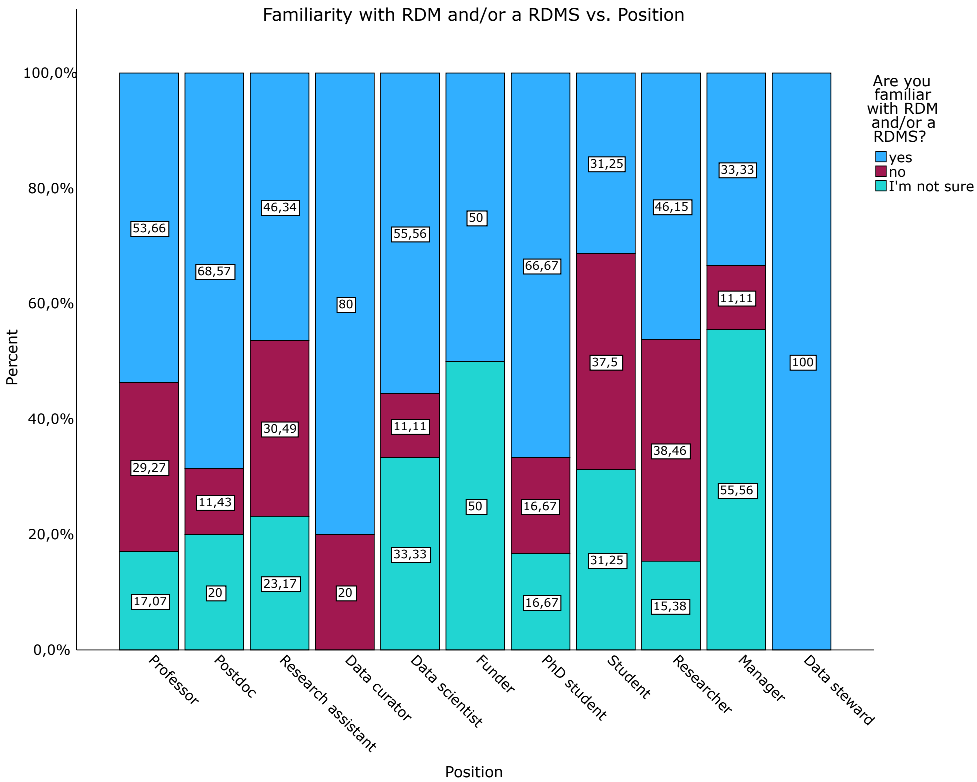


Figure 3.12: Familiarity with RDM and/or an RDMS vs. Position

fulfilling them very good with 5, and values in between. Most ratings can be found in the middle and towards the positive side of the scale, with a mean value of 3.53 and a standard deviation of 0.74.

Conclusion: While most respondents lack an RDMS in their workflow, those who do utilize a diverse range of systems. This variety unfortunately precludes direct comparisons of their experiences. They generally express satisfaction with the tools they use.

PROS AND CONS OF RDMS: To gain insights into the advantages and disadvantages of currently used systems, we asked the participants for their feedback on their experiences with their own systems. 38 of them shared the pros and cons. The answers were grouped into four categories: general, access, (meta-)data, functions (Table 3.4).

Four participants do not like anything about their system, six consider it easy to use, and three like the support available as well as the large storage size. Six participants mention their time-consuming and slow system, whilst three do not like the user interface, and another three have nothing negative to say at all.

When it comes to access, three respondents find their system easy to access, whereas two struggle with that. Having the system accessible online and being able to enter data from everywhere is seen as an advantage. The access rights system seems to be an area for improvement.

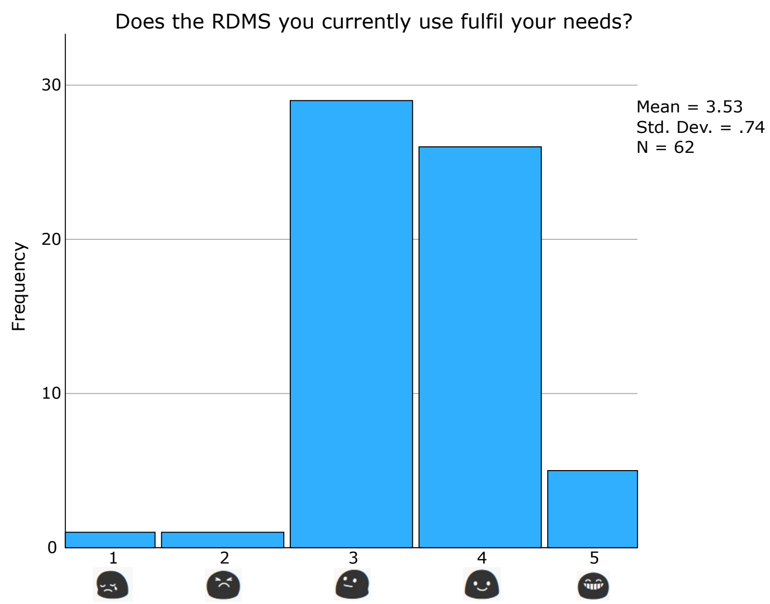


Figure 3.13: Satisfaction with Current RDMS

It seems that existing systems deal with data and metadata in different ways. Some of them have all data together and linked, and enable deposits of any type of digital object. Others require duplicate data entries, store data in a distributed manner, and do not enforce metadata for preserved datasets.

The list of useful functions is long, but visualization is definitely something that users like. Four systems, however, are lacking integration with other tools and databases.

Conclusion: A slow and complicated system with a lack of integration with other tools could pose a problem. Poor data management and an insufficient access rights system may not be well received by users. These and other aspects could be addressed in detail in a future study. On the other hand, potential users value good support and extensive storage space. They consider proper visualization options important.

	Is there anything in particular about your current RDMS...	
Category	...that you like? (#)	...that you don't like? (#)
General	Easy to use/user friendly/intuitive input (6)	Time-consuming/slow (6)
	No (4)	No (3)
	IT support/on-demand customer service (3)	User interface (3)
	Storage space/large repository size (1TB) (3)	It's not broken (1)
	Expandable/upgradable (2)	Price (1)
	Tailor-made (2)	Non-intuitive processes (1)
	Tailored to specific research domain (1)	Help page (1)
	Can be used by everyone (1)	High personnel requirements for management (1)
	Free (1)	
	It's a must (1)	
	It's really good (1)	
	It does what it needs to do (1)	
	Little learning effort (1)	
	Organized (1)	
	Good overview/structure (1)	
	Structured but flexible (managed separately for each research group) (1)	
	Simple set up (1)	
	Yes (1)	
Access	Easy to access (3)	Access/sign-in (2)
	Accessible online/data entry from anywhere with Internet (2)	Access rights system (2)
	Fully integrated into our electronic medical record databases (1)	External access (1)
	Scriptable API (1)	
	Secured cloud upload (1)	
(Meta-)Data	All data together and linked (2)	Distributed/duplicate data (3)
	Enables deposit of any type of digital objects (1)	No mandatory metadata for preserved datasets (2)
	Suitable for worldwide collection of [...] data (1)	No domain-specific metadata schema (1)
	It uses Dublin Core metadata (1)	Messes indicators to collect (1)
	Review of metadata by data experts (1)	Sometimes data seem irrelevant (1)
	Up-to-dateness of data from acquisition to usability (1)	
Functions	Can grab the molecular structure from NIST (1)	Lack of tool/database integration (4)
	Combines tools required for experiment design with data management system and lab notebook (1)	Adjustments needed for functions/elements (2)
	Complex survey design (1)	Manageable range of functions (1)
	Detailed cross tabs and stats functionality (1)	It's not possible to combine multiple datasets and create new ones (1)
	Easiness to visualize results (1)	Electronic Lab Notebook doesn't store data (1)
	Graphical overview options (1)	Long/complicated to change existing data (1)
	Graphical representation of the data (1)	Optimization needed for analysis section and reverse scoring of data (1)
	Has several long-in levels (1)	Lacks easy tracking of content/data changes (1)
	Able to remove noise data when analyzing movements/measuring time reactions (1)	Lack of tackling storage aspect (1)
	In-text search function (1)	Lack of data curation (1)
	Open text analysis (1)	Lack of identifying sub-collections of data (1)
	Provides clear outputs (1)	Doesn't link to PIDs (Personal Identifiable Data) (1)
	Recruitment with piped email names (1)	Doesn't do tables well (1)
	Research panel management (1)	Doesn't do qualitative analysis coding of focus groups (1)
	Some data collection systems provide information for writing projects' justification/repoeting (1)	Only designed to develop plans in format required by specific agencies (1)
	Used for developing Data Management plans for NSF grants (1)	
	Ability to internally share data across different people/teams (1)	
	Confidential data protection (1)	
	Multiple data safety redundancies (1)	

Table 3.4: Pros and Cons of Currently Used RDMS

RESEARCH DATA CYCLE: Table 3.5 combines respondents with same answers; i.e. six participants currently use an RDMS for five phases of the research data cycle and would like to use it for the same number of phases. It can be seen that 204 participants would like to use such a system for the same number or even more phases they currently do (upper part of Table 3.5).

# Participants	6	2	1	3	3	5	4	4	2	5	2	6	3	32	22	48	31	20	5
# Phases Current Use	5	4	4	3	3	3	2	2	2	2	1	1	1	0	0	0	0	0	0
# Phases Future Use	5	5	4	5	4	3	5	4	3	2	4	3	2	5	4	3	2	1	0
# Participants	1	3	1	1	1	1													
# Phases Current Use	5	4	4	3	3	2													
# Phases Future Use	0	3	2	2	1	1													

Table 3.5: Current vs. Future Use of RDMS

There were eight participants who prefer to reduce their use of an RDMS (lower part of Table 3.5). They all work in different scientific fields (computer and information sciences; industrial biotechnology; agriculture, forestry, and fisheries; archaeology; other engineering and technology; economics and business; psychology; other medical sciences). Their experiences with RDM and/or an RDMS ranges from less than one year to five years or more. Surprisingly, they all rated their satisfaction with their current RDMS with three to four (whilst five is the best and one the worst).

One postdoc, for example, who works with a database which was specifically designed to the researcher's needs currently uses it for all phases of the research data cycle except the planning phase. However, they would prefer to use it only for processing and analysis, as well as data archiving. The only criticism mentioned against the system is the need for some tweaking every now and then.

The other participants who currently use their RDMS for four phases work as research assistants, data managers, data scientists, and managers. The phases for which they currently use it and the ones for which they would like to use it are completely different. Their existing systems are a data management system, a digital lab notebook, and an internally developed system. The main reasons for critique can be found in "shortcomings" of those systems (issues with tables and the search function, lack of connection with other systems) and the lack of proper usage by other researchers.

Another research assistant uses a business intelligence and analytics software; they would only like to see a better help page for the tool and currently use it for planning, data processing and analysis, and data reuse. They would prefer to reduce this to data archiving.

A university research assistant would like to use their research data management platform not for planning and data archiving anymore, but rather for data processing and analysis because they consider it lengthy.

Someone working in research support at a university does not use their lab management software themselves. They do not like to see that utilization of the tool as an absolute must. Having no possibility to store data through the software, and thus having the data separate from protocols and other information, is seen as a downside of the software.

Conclusion: There is no common denominator apparent to the reluctance of using an RDMS.

SOFTWARE INTEGRATION: Participants were asked about the three most important software applications they use. Those should be either the most important ones for their research, or the most frequently used ones in this context. Of the 254 people who provided this information, 166 use Excel, 52 SPSS, and 47 Zotero. Other commonly used software can be found in the word cloud in Figure 3.14. Software used by less than ten respondents was anonymized.

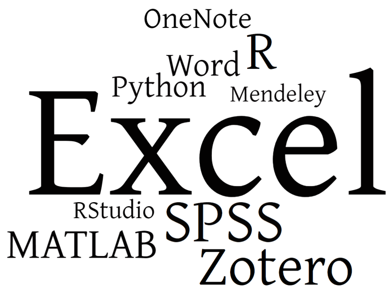


Figure 3.14: Most commonly used software in research[7]

The software applications provided were grouped into 46 categories. The top three categories named were spreadsheets such as Excel (167 mentions), software for statistical analysis such as SPSS (129 mentions), and programming software/languages such as R or Integrated Development Environment (IDE) such as RStudio (99 mentions) (Figure 3.15).

Figure 3.15: Most commonly used software categories in research[8]

7 Created with `http://worditout.com/`
8 Created with `http://worditout.com/`

Conclusion: Some of the software used such as reference management can be replaced by an RDMS to reduce the number of tools required for the daily work of the researchers.

ANALOG RESEARCH DATA: More than half of the respondents (54.28 %) solely work with digital research data. 40.52 % still store at least some research data in analog form, such as biological samples or handwritten notes. The rest did not supply information on that (Figure 3.16).

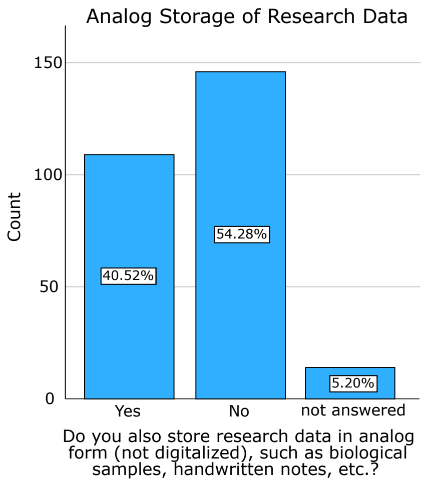


Figure 3.16: Analog Storage of Research Data

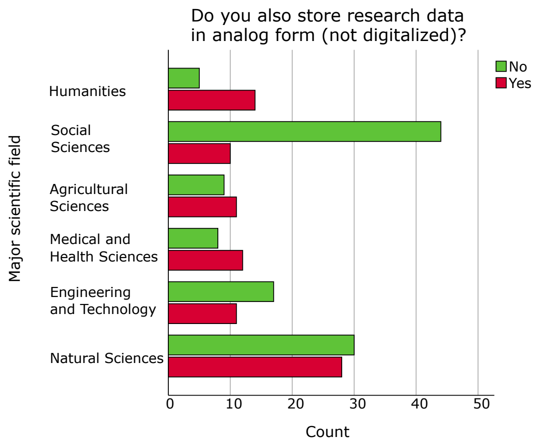


Figure 3.17: Analog Storage vs. Major Scientific Fields

Based on the answers provided, researchers in social sciences are trailblazers when it comes to digitalization – 81 % of them store no analog data anymore. Also in engineering and technology, as well as in the natural sciences, the majority of respondents keep only digital data. On the other hand, humanities is the field with the lowest ratio of digital data (26 %). Furthermore, in the fields of agricultural sciences and medical and health sciences, analog data still plays an important role (Figure 3.17).

Looking into the kind of data which is not digitalized, it is obvious that part of it cannot be fully digitalized, such as biological (21 mentions) or geological samples. However, other data such as handwritten notes (35 mentions), lab journals (9 mentions), or questionnaires (7 mentions) could be moved from analog to digital storage.

A common reason to keep this data analogous is a lack of time (16 mentions). Some institutes are currently in the process of digitizing the data (9 mentions). Part of it is also kept in analog form as a backup or due to legal requirements (19 mentions). Some answers indicate a lack of trust in (3 mentions) or knowledge of the available options (8 mentions), or because of organizational restrictions (5 mentions).

When being asked for the reasons why data is not yet digitalized, a respondent working in the upper management of a biotech company pointed out specific challenges in this field. They suggested to read "about the attempt of the biotech industry to digitalize batch records in GMP[9] manufacturing", since "there are significant obstacles that might be relevant for RDMS". A full list of the types of research data stored in analog form, along with the reason why it is not yet digitalized, can be found in Section 6.6.

Conclusion: Certain data, such as biological samples, will remain analogous in the future, as it is not possible to digitalize everything. However, by offering efficient options for input and addressing security concerns, data such as handwritten notes and questionnaires could be moved into an RDMS.

REQUIRED STORAGE: The file size of a typical digital output varies greatly between the major scientific fields and within (Figure 3.18). At first glance, both engineering and technology and medical and health sciences seem to be hungriest for data – the former has a median of 250 MB and a standard deviation of around 558.2 GB, the latter a median of 225 MB and a standard deviation of around 270 GB. Although natural sciences only has a median of 50 MB, the standard deviation is at around 1.4 TB. Humanities also shows a high standard deviation of 306.3 GB with a median of 30 MB due to the more storage intensive field of history and archaeology. A median of 50 MB and a standard deviation of around 1.2 GB can be found for social sciences, and agricultural sciences has a median of 10 MB and a standard deviation of 204 MB. No scientific field was selected for the last group with a median of 100 MB and a standard deviation of around 1.6 TB.

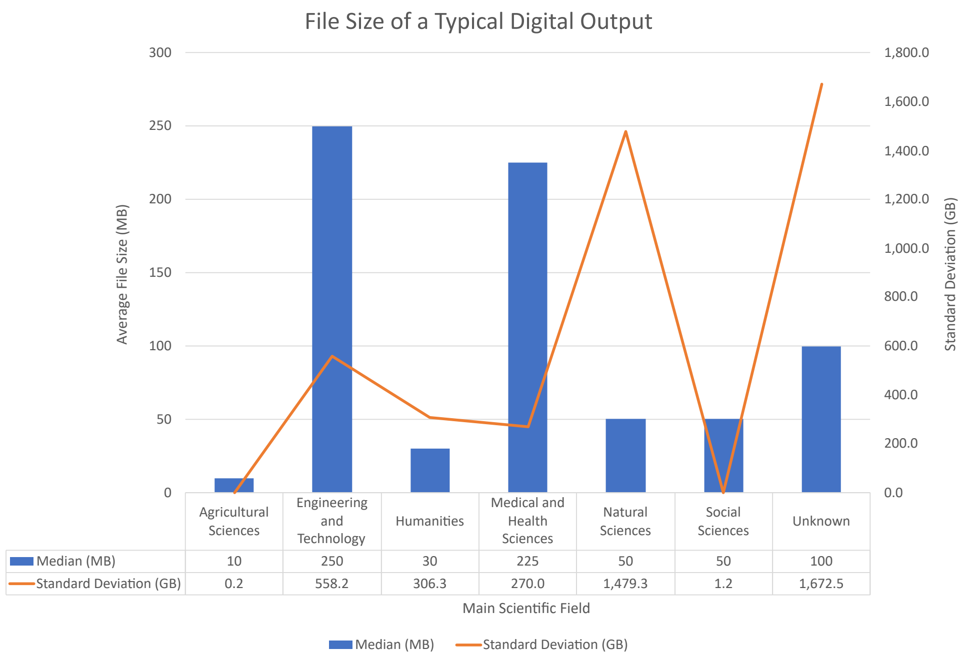


Figure 3.18: File Size of Typical Digital Output

9 GMP = Good Manufacturing Practices

Conclusion: Some high standard deviations show that the required storage can differ greatly from the median. Therefore, providing flexible storage options for an RDMS would be recommended.

Conclusion on the status quo: The current state of knowledge about RDM shows a need for more targeted education on this topic. In regard to RDMS, a subsequent study could look at existing systems and the issues their users have, in order to get more information on what to avoid in a new system. This may also increase the willingness of researchers to use an RDMS. Integrating some of the tools frequently used into an RDMS can ease the workload of users. Offering efficient input options for such a system eliminates the burden on researchers to digitalize their analog data. The storage required is already substantial for some sciences; having flexible storage options available is especially important considering the continuous increase in digital data.

3.4.3 *Hypotheses*

Three hypotheses were defined, which shall be validated with the help of the survey data. Through inductive statistics, it can be examined whether the results of the sample can be transferred to the population or not, thus a hypothesis can be either confirmed or disproven. This can be achieved through univariate, bivariate, or multivariate analysis. [May13]

3.4.3.1 *Hypothesis 1*

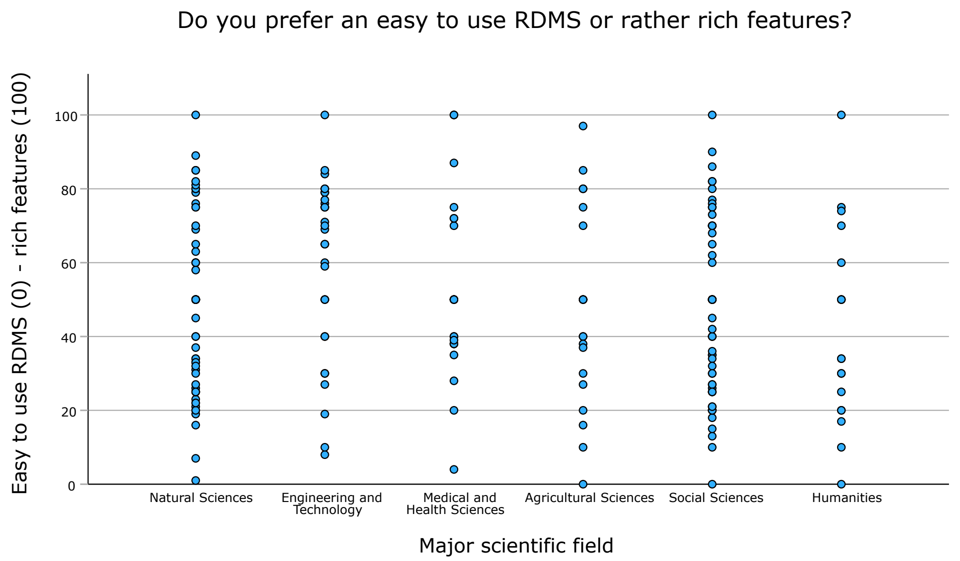
People working in technically oriented fields (e.g. engineering and technology, natural sciences) prefer rich features, whilst people in less technical fields (e.g. humanities, social sciences) prefer an RDMS which is easy to use.

We consider an RDMS easy to use if it can be utilized without extensive training or studying a user manual. Since any software should be easy to use, there should be a good balance between the two extremes, and the ideal value for the purpose of our hypothesis can be seen as 50 (easy to use = 0, rich features = 100).

A bivariate analysis with the help of a scatter plot (Figure 3.19) was used. Data points are distributed throughout the scale. There is no distinct tendency visible. Humanities lean slightly towards easy usage, whilst engineering and technology has more data points closer to the top of the scale. The mean values of those major scientific fields confirm the tendency. The mean values of the other fields can be found within 3 % of the middle.

The technically oriented fields of engineering and technology and natural sciences have a combined average of 53.68 %. The less technical fields of humanities and social sciences have a combined average of 44.54 %. This indicates a slight inclination towards a confirmation of the hypothesis.

A chi-square test examines the association between two or more categorical variables. It determines whether or not the observed data differs signifi-

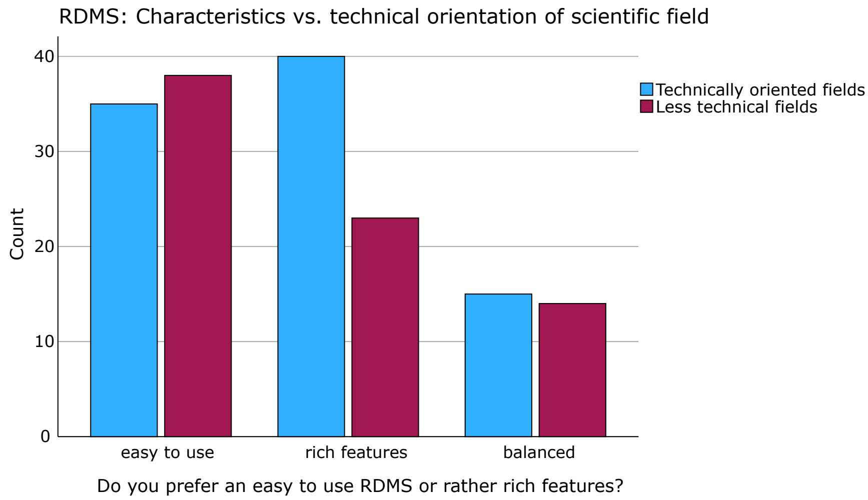


Major scientific field	Mean	N
Natural Sciences	48,08	60
Engineering and Technology	59,27	30
Medical and Health Sciences	52,09	22
Agricultural Sciences	47,25	20
Social Sciences	47,02	56
Humanities	42,05	19
Total	49,21	207

Figure 3.19: Preference for Easy Usage or Rich Features

cantly to the expected data. The null hypothesis is tested of no association. The degrees of freedom refer to the number of independent variables. The p-value stands for the probability. The null hypothesis is not rejected with a p-value greater than the significance level. In that case, the data would not provide statistically significant evidence to conclude that an association between the variables exists. However, a p-value close to the significance level could point to a weak association between the variables. The larger the sample size, the robuster the result.

To facilitate a chi-square test, the values of the questionnaire were divided into three categories: easy to use (0-49), balanced (50), rich features (51-100). The major scientific fields engineering and technology and natural sciences were grouped into technically oriented fields, whilst humanities and social sciences were grouped into less technical fields. We used a significance level of $\alpha = 5$ %. The clustered bar chart (Figure 3.20) indicates an association between the technical orientation of the major scientific field and the preference for rich features in the RDMS. The value of the chi-square test statistic is 3.410. According to the footnote in Figure 3.20, the expected cell counts are all greater than 5, therefore this assumption was met ([May13]). The degrees of freedom (df) for the test statistic is 2. The corresponding p-value of the test statistic is $p = 0.182$. Since the p-value is greater than our chosen significance level ($\alpha = 0.05$), we do not reject the null hypothesis, which states that there is no relationship between the two variables.



Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	3,410[a]	2	,182
Likelihood Ratio	3,437	2	,179
Linear-by-Linear Association	,711	1	,399
N of Valid Cases	165		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 13,18.

Figure 3.20: Chi-Square Test: Preference for Easy Usage or Rich Features

Conclusion: There is actually not enough evidence to suggest an association between the technical orientation of the scientific field and the demand for either an easy to use RDMS or rich features ($\chi^2(2) \geq 3.410, p = 0.182$). Therefore, this hypothesis has been **disproven**.

3.4.3.2 *Hypothesis 2*

People who are not familiar with RDM and/or an RDMS are less likely to be willing to use an RDMS for the entire research data cycle.

Participants could choose from none to all phases of the research data cycle for which they would like to use an RDMS for. 203 participants who provided information about their knowledge on RDM and/or an RDMS did make a selection. 46 of those have selected all phases of the research data cycle (Figure 3.21). Almost two thirds of them are familiar with RDM and/or an RDMS, the rest which is grouped together is either not or not sure.

			For which phases of the research data cycle would you like to use a RDMS?						
			Planning	Data collection	Data processing and analysis	Data archiving	Data reuse	Total	
Are you familiar with RDM and/or a RDMS?	yes	Count	29	29	29	29	29	29	**63 %**
	no	Count	5	5	5	5	5	5	**37 %**
	I'm not sure	Count	12	12	12	12	12	12	
Total		Count	46	46	46	46	46	46	

Figure 3.21: RDMS Desired for all Research Data Cycle Phases

Additionally, Figure 3.22 shows how often each of the research data cycles was selected since, besides the hypothesis, it is equally interesting to look at the individual phases. Also here participants who are familiar chose cycles more frequently: 44 % of them want to use an RDMS for planning, 76 % for data collection, 74 % for data processing and analysis, 77 % for data archiving, and 62 % for data reuse. The numbers for participants who are not sure if they are familiar with RDM and/or RDMS are almost the same. Differences can be seen for those not familiar. Only 21 % of them would like to use an RDMS for planning, 67 % for data archiving, and 38 % for data reuse, whilst the percentages for data collection and data processing and analysis are almost the same as for the other groups.

				For which phases of the research data cycle would you like to use a RDMS?					
			Total	Planning	Data collection	Data processing and analysis	Data archiving	Data reuse	
Are you familiar with RDM and/or a RDMS?	yes	Count	106	47	81	78	82	66	**354 / 55 %**
	no	Count	52	11	40	38	35	20	**144 / 22 %**
	I'm not sure	Count	45	20	35	33	36	26	**150 / 23 %**
Total		Count	203	78	156	149	153	112	**648 / 100 %**

Figure 3.22: RDMS Desired for any Research Data Cycle Phase

Conclusion: 63 % of participants familiar with RDM and/or an RDMS would like to use an RDMS for the entire research data cycle. Furthermore, they also account for more than half of the selections for the individual research data cycle phases made. Therefore, this hypothesis is **confirmed**.

3.4.3.3 *Hypothesis 3*

"Digital natives" are more likely to use an RDMS than previous generations.

According to Prensky [Pre01], a digital native is a person who has grown up in the information age and considers technology as part of their everyday lives. It is widely regarded that those are persons born in or after 1980. Kennedy et al. [Ken+08], amongst others, criticize this definition, given the differences in importance technology plays in the lives of that age group. Those differences may be due to socio-economic, cultural, or ethnic background, geographic location, gender, and discipline specialization. Studies suggest that the real digital natives are younger. Consequently, we consider participants up to the age of 39 (born in or after 1984) as digital natives.

					For which phases of the research data cycle would you like to use a RDMS?					
				Total	Planning	Data collection	Data processing and analysis	Data archiving	Data reuse	
What age group do you belong to?	≤ 29	Count	**56 %**	48	21	38	40	33	22	**max. possibilities: 565 selected: 349 => 61.8 %**
	30-39	Count		65	22	47	40	51	35	
	40-49	Count	**44 %**	43	20	30	33	34	28	**max. possibilities: 440 selected: 297 => 67.5 %**
	50-59	Count		33	12	28	26	26	19	
	≥ 60	Count		12	3	12	9	10	7	
Total		Count		201	78	155	148	154	111	

Figure 3.23: RDMS Desired for any Research Data Cycle Phase based on Age

201 participants who supplied their age group made a selection for one or several phases of the research data cycle they would like to use an RDMS for (Figure 3.23). The sample contains slightly more digital natives than older age groups. Amongst the digital natives, 61.8 % would like to use an RDMS for at least one phase of the research data cycle, whilst amongst participants of 40 years or older this share is 67.5 %.

Conclusion: More than half of the digital natives as well as the older age groups want to use an RDMS for some of the phases of the research data cycle. However, the inclination toward an RDMS is smaller among digital natives. Therefore, this hypothesis has been **disproven**.

3.4.4 *User Needs*

In the user needs section we assess a list of possible criteria for an RDMS by scientific fields and positions of the participants. Then we evaluate which type of application users prefer. We check if and which kind of software they would like to have integrated in an RDMS and conclude with additional features the respondents have requested for such a system.

CRITERIA LIST BY SCIENTIFIC FIELD: A list of 35 criteria for an RDMS was presented in the survey (Section 6.3). Participants were asked to rate each criterion on a Likert-like scale from "1 - not at all important" to "5 - very important". In case a criterion would not be relevant for a specific scientific field or the respondent would not know about it, the option "no answer" was offered as well.

Based on that data, the overall average was calculated, the average for each major scientific field (for example Humanities), and the average of each sub-discipline (for example Humanities – History and Archaeology). An overview is available in Table 3.6, with detailed data in Section 6.7. The color coding in Table 3.6 depends on the rating: values between 3.5 and 3.9 are light gray, between 4.0 and 4.4 medium gray, and all values of 4.5 or more are dark gray.

207 respondents who also provided their scientific field rated the criteria. Overall, the three most important criteria are (the average ratings are set in parentheses):

- Full control over own data (e.g. deletion, editing, export, etc. of data) (4.5)
- Tracking of content changes (4.2)
- Data export in different formats / Search filters (faceted search) / Unique identifiers for data (4.1 each),

whilst the three least important criteria are:

- Compliance with OAI-PMH (2.3)
- Domain-specific metadata schemata available (2.5)
- Multilingual interface (2.6)

Although these average ratings mostly agree with the average ratings of the major scientific fields, the order is different. Full control over own data (e.g. deletion, editing, export, etc. of data) is rated as most important for everyone except humanities, where it is only in second place. Compliance with data regulations (GDPR, NDA, ethics, etc.) clearly has a higher priority for medical and health as well as social sciences, and also file encryption is important in the medical and health field. Tracking of content changes cannot be found amongst the top three of humanities only.

Figure 3.24 shows the overall average ratings, including the minimum and maximum deviations of the major scientific fields from this overall average. The greatest deviations from the average ratings stand out for the support for flexible research workflows, control over data release dates, as well as the tracking of interactions of external entities with the dataset (e.g. citations) and the availability of domain-specific metadata schemata.

The top three criteria of each major scientific field in detail are (the average ratings are set in parentheses):

- Agricultural Sciences:
  - Full control over own data (e.g. deletion, editing, export, etc. of data) (4.9)
  - Tracking of content changes / Data export in different formats (4.6 each)
  - Free-text search capability (4.4)
- Engineering and Technology:
  - Full control over own data (e.g. deletion, editing, export, etc. of data) (4.5)
  - Unique identifiers for data (4.4)
  - Tracking of content changes / Search filters (faceted search) (4.2 each)
- Humanities:
  - Free-text search capability (4.6)
  - Search filters (faceted search) / Full control over own data (e.g. deletion, editing, export, etc. of data) (4.4 each)
  - Availability of preferred usage method (via web application, mobile application, local installation) (4.3)
- Medical and Health Sciences:
  - Compliance with data regulations (GDPR, NDA, ethics, etc.) / Full control over own data (e.g. deletion, editing, export, etc. of data) / Unique identifiers for data (4.5 each)
  - Availability of preferred usage method (via web application, mobile application, local installation) / Search filters (faceted search) (4.4 each)
  - Tracking of content changes / Free-text search capability / File encryption (4.3 each)

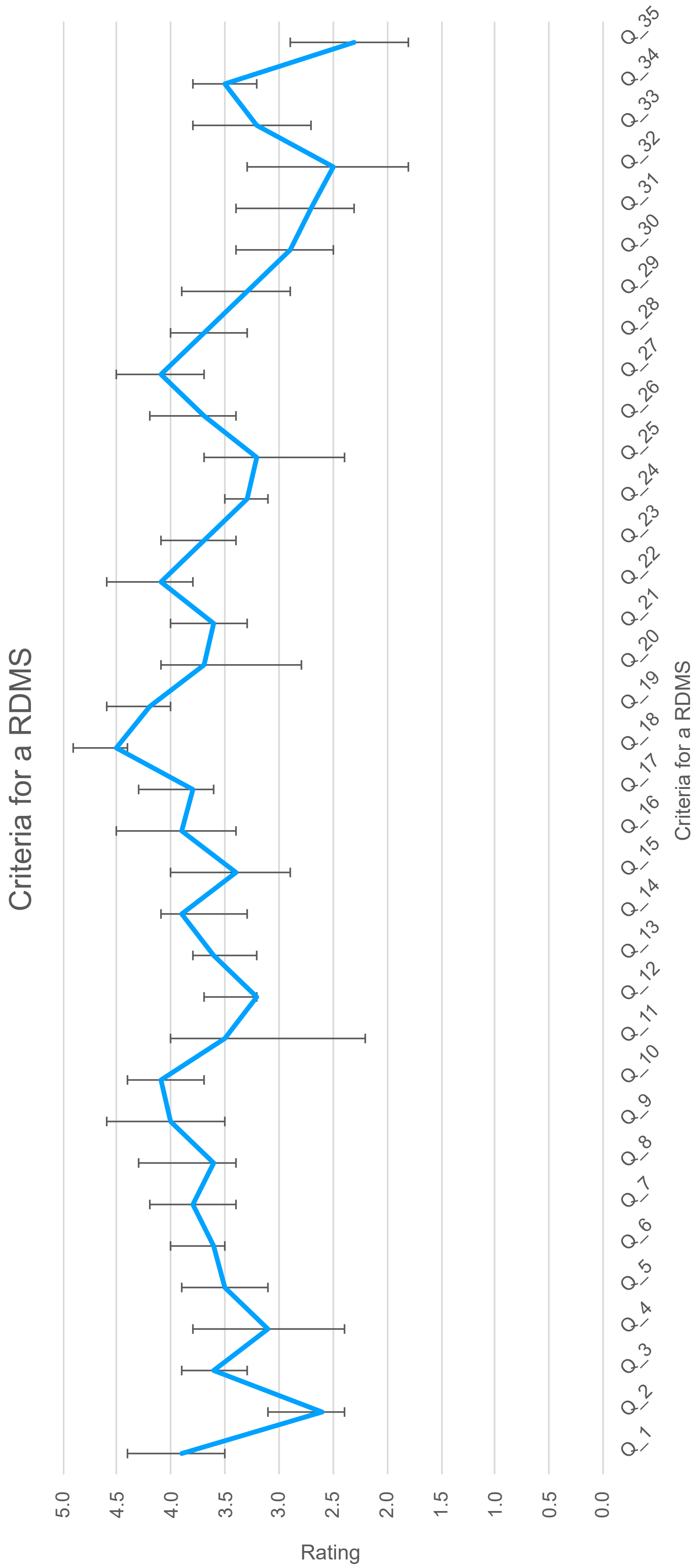


Figure 3.24: Importance of Criteria for an RDMS with Deviations

- Natural Sciences:
  - Full control over own data (e.g. deletion, editing, export, etc. of data) (4.5)
  - Availability of preferred usage method (via web application, mobile application, local installation) / Search filters (faceted search) / Tracking of content changes / Data export in different formats / Unique identifiers for data (4.2 each)
  - Collaborative environment for teams/groups with an access rights system / Free-text search capability / Data provenance (How was the data created/modified/used?) (4.0 each)
- Social Sciences:
  - Full control over own data (e.g. deletion, editing, export, etc. of data) (4.5)
  - Compliance with data regulations (GDPR, NDA, ethics, etc.) (4.2)
  - Tracking of content changes (4.1)

Zooming in to natural sciences where most of the respondents (n = 60) work, the two sub-disciplines with the most input are biological sciences (n = 16) and computer and information sciences (n = 22). For the major (natural sciences) as well as both sub-disciplines (biological sciences, computer and information sciences), full control over own data is the most important feature. Tracking of content changes and unique identifiers for data can also be found in each top three criteria. Overall, the top three of those scientific fields differ, which means we cannot conclude from the average of the major scientific fields to their respective sub-disciplines.

The top three criteria for these sub-disciplines are:

- Natural Sciences – Biological Sciences:
  - Full control over own data (e.g. deletion, editing, export, etc. of data) / Collaborative environment for teams/groups with an access rights system / Tracking of content changes / Data export in different formats / Unique identifiers for data (4.4 each)
  - Availability of preferred usage method (via web application, mobile application, local installation) / Data provenance (How was the data created/modified/used?) (4.3 each)
  - Free-text search capability / Search filters (faceted search) (4.2 each)
- Natural Sciences – Computer and Information Sciences:
  - Full control over own data (e.g. deletion, editing, export, etc. of data) (4.5)
  - Control over data location (locale/cloud/specific country) / Tracking of content changes / Unique identifiers for data (4.2 each)
  - Compliance with data regulations (GDPR, NDA, ethics, etc.) (4.1)

In the field of humanities, we can find almost entirely low ratings for the 35 criteria. Discussions with scientists of those fields revealed that this may be due to the fact that most sub-disciplines of humanities (e.g. philosophy, languages and literature) deal with interpretation of existing literature, thus all their research data is published in full. The only exception to that may be archaeology, dealing with artifacts from excavations.

Conclusion: There are several criteria that are of different importance to users depending on the scientific field in which they work. To have an RDMS which is easy to use and thus provides a higher acceptance rate amongst users, it may be helpful to offer a Graphical User Interface (GUI) tailored to the scientific field of the user. This can reduce cluttering for sciences that do not need a lot of functions (e.g. philosophy), and would be the result of subsequent works.

	#	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}	Q_{12}	Q_{13}	Q_{14}	Q_{15}	Q_{16}	Q_{17}	Q_{18}	Q_{19}	Q_{20}
Total	207	3.9	2.6	3.6	3.1	3.5	3.6	3.8	3.6	4.0	4.1	3.5	3.2	3.6	3.9	3.4	3.9	3.8	4.5	4.2	3.7
Agricultural Sciences	20	3.8	2.5	3.7	3.4	3.9	3.9	4.0	3.6	4.4	4.1	3.9	3.7	3.8	4.1	3.4	3.6	4.3	4.9	4.6	4.1
Engineering and Technology	30	3.8	3.1	3.5	3.8	3.7	3.6	3.9	3.8	3.8	4.2	3.7	3.2	3.7	4.0	3.2	3.8	3.7	4.5	4.2	3.9
Humanities	19	4.3	2.6	3.5	3.3	3.5	3.5	3.9	3.5	4.6	4.4	2.3	3.2	3.2	3.8	3.2	3.4	3.8	4.4	4.2	3.9
Medical and Health Sciences	22	4.4	2.6	3.4	3.6	3.8	4.0	4.2	4.3	4.3	4.4	4.0	3.5	3.6	4.1	4.0	4.5	4.2	4.5	4.3	4.0
Natural Sciences	60	4.2	2.5	3.9	3.2	3.7	3.6	4.0	3.4	4.0	4.2	3.7	3.2	3.6	3.8	2.9	3.6	3.8	4.5	4.2	3.6
Social Science	56	3.5	2.4	3.3	2.4	3.1	3.5	3.4	3.4	3.5	3.7	2.2	3.2	3.5	3.3	3.6	3.9	3.6	4.8	4.0	2.8

	#	Q_{21}	Q_{22}	Q_{23}	Q_{24}	Q_{25}	Q_{26}	Q_{27}	Q_{28}	Q_{29}	Q_{30}	Q_{31}	Q_{32}	Q_{33}	Q_{34}	Q_{35}
Total	207	3.6	4.1	3.7	3.3	3.2	3.7	4.1	3.7	3.3	2.9	2.7	2.5	3.2	3.5	2.3
Agricultural Sciences	20	4.0	4.6	3.7	3.5	3.5	3.8	3.9	4.0	3.3	3.0	2.8	2.9	3.1	3.5	2.1
Engineering and Technology	30	3.6	3.8	3.7	3.4	3.5	3.8	4.4	4.0	3.6	3.4	3.4	3.3	3.8	3.7	2.9
Humanities	19	3.4	4.1	4.1	3.4	2.4	3.4	3.7	3.7	3.3	2.5	2.3	2.2	2.9	3.8	2.5
Medical and Health Sciences	22	4.0	4.2	3.4	3.4	3.7	4.2	4.5	3.8	3.9	3.2	3.2	2.9	3.5	3.3	2.9
Natural Sciences	60	3.6	4.2	3.8	3.4	3.2	4.0	4.2	3.7	3.3	3.1	2.8	2.6	3.4	3.6	2.2
Social Science	56	3.3	4.0	3.4	3.1	2.9	3.4	3.9	3.3	2.9	2.5	2.4	1.8	2.7	3.2	1.8

1 = not at all important / 5 = very important / 0 = no answer

Q_1: Availability of preferred usage method
Q_2: Multilingual interface
Q_3: Open source
Q_4: Existence of an API
Q_5: Integration of other tools
Q_6: Support options
Q_7: Collaborative environment for teams/groups with access rights system
Q_8: File encryption
Q_9: Free-text search capability
Q_{10}: Search filters (faceted search)
Q_{11}: Support for flexible research workflows
Q_{12}: Wizards to guide through processes
Q_{13}: Feature for previewing dataset content
Q_{14}: Visualization of datasets
Q_{15}: Functionality to anonymize research data
Q_{16}: Compliance with data regulations
Q_{17}: Control over data location
Q_{18}: Full control over own data
Q_{19}: Tracking of content changes
Q_{20}: Control over data release dates
Q_{21}: Control over data expiry dates
Q_{22}: Data export in different formats
Q_{23}: Possibility of data citation
Q_{24}: Transformation of data within the RDMS
Q_{25}: Tracking of interactions of external entities with dataset
Q_{26}: Data provenance
Q_{27}: Unique identifiers for data
Q_{28}: Logical links between related data
Q_{29}: Mandatory/optional metadata elements
Q_{30}: Automated metadata tagging
Q_{31}: Possibility to set up multiple metadata schemata
Q_{32}: Domain-specific metadata schemata
Q_{33}: Data validation
Q_{34}: Data identification via DOI/URL
Q_{35}: Compliance with OAI-PMH

Table 3.6: Importance of Criteria for an RDMS

CRITERIA LIST BY POSITION: In Section 3.4.1 we have seen that a relatively large portion of the questionnaires for the area of engineering and technology were filled out by students. Since it can be argued that students have the least experience with RDM, we took a closer look into the user needs submitted for this group. The eight students who provided feedback work in the areas of materials engineering (one student), mechanical engineering (one student), electrical/electronic/information engineering (three students), medical engineering (one student), and other engineering and technologies (two students). Half of them (all the materials and electrical/electronic/information engineering students) stated that they are familiar with RDM and/or an RDMS. However, none has experience of more than two years, and only one uses an internal system for data processing and analysis, which they consider to be an RDMS. To compare the answers of students with people in other positions, we excluded those from respondents who had not provided their position, as well as those of the one student in medical engineering, being the only case in this scientific field at all. The mechanical engineering student solely selected the middle value (3.0); therefore, his answers can be considered invalid and were also left out of this specific comparison.

The user needs submitted by students were set in contrast to the ones of other positions. Any difference in rating greater than 0.5 was highlighted in Table 3.7. This reveals that students and people in other positions barely have any common ground. The only questions where the deviation between their answers was minor across all three scientific fields were about open source, support options, functionality to anonymize research data, full control over own data, possibility of data citation, logical links between related data, automated metadata tagging, and the possibility to set up multiple metadata schemata (eight of 35 criteria).

Looking in more detail, the three students working in electrical/electronic/information engineering stated that they are familiar with RDM and/or an RDMS. Of the five people in other positions of the same scientific field, two are familiar and three are either not familiar or not sure about their knowledge of RDM and/or an RDMS. In this scientific field, students and people in other positions rated 21 of 35 criteria with the same level of importance ($\pm$ 0.5). In the field of materials engineering, we have one student and one person in another position familiar with RDM and/or an RDMS. A second person in another position did not answer the question about their familiarity. Also in this scientific field, students and people in other positions rated 21 criteria out of 35 with the same level of importance ($\pm$ 0.5). The two students working in other engineering and technology fields are either not familiar with RDM and/or an RDMS or not sure about it. Of the ten people in other positions, eight are familiar and two not sure. Here, we can see that students and people in other positions only rated 15 of 35 criteria with the same level of importance ($\pm$ 0.5).

	#	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}	Q_{12}	Q_{13}	Q_{14}	Q_{15}	Q_{16}	Q_{17}	Q_{18}	Q_{19}	Q_{20}
Electrical/Electronic/Information Engineering - Student	3	4.0	4.0	3.3	4.7	4.3	3.7	3.7	4.0	4.0	4.7	4.7	3.7	4.7	3.7	3.3	3.7	4.3	4.7	4.7	3.7
Electrical/Electronic/Information Engineering - Rest	5	4.2	3.6	3.6	3.8	4.0	3.6	4.4	3.0	3.2	4.0	2.8	3.2	4.0	4.2	3.6	4.6	4.2	4.4	4.4	4.6
Materials Engineering - Student	1	4.0	3.0	5.0	5.0	5.0	4.0	4.0	2.0	3.0	5.0	5.0	3.0	4.0	5.0	2.0	2.0	1.0	5.0	5.0	4.0
Materials Engineering - Rest	2	4.0	3.5	5.0	5.0	3.5	3.5	3.0	4.0	4.0	4.0	4.5	2.5	1.5	2.5	1.5	3.5	3.0	4.5	5.0	4.5
Other Engineering and Technology - Student	2	5.0	5.0	3.0	3.0	3.5	4.0	3.0	5.0	4.5	4.5	5.0	4.5	3.5	3.5	4.0	3.5	4.0	4.5	5.0	4.5
Other Engineering and Technology - Rest	10	3.8	2.2	3.4	3.6	3.8	3.7	4.2	4.5	3.7	4.0	3.8	3.3	4.0	4.2	3.5	4.4	3.9	4.8	4.4	3.6

	#	Q_{21}	Q_{22}	Q_{23}	Q_{24}	Q_{25}	Q_{26}	Q_{27}	Q_{28}	Q_{29}	Q_{30}	Q_{31}	Q_{32}	Q_{33}	Q_{34}	Q_{35}
Electrical/Electronic/Information Engineering - Student	3	4.3	4.7	3.3	4.7	3.7	2.0	4.3	4.7	4.0	3.7	3.7	3.3	4.3	2.3	2.3
Electrical/Electronic/Information Engineering - Rest	5	4.0	3.6	3.8	3.8	4.0	4.4	4.6	4.6	4.4	4.2	4.2	4.0	4.2	4.2	4.0
Materials Engineering - Student	1	4.0	3.0	4.0	4.0	5.0	4.0	5.0	4.0	3.0	3.0	3.0	3.0	5.0	5.0	5.0
Materials Engineering - Rest	2	2.5	2.5	4.5	2.0	3.5	4.5	5.0	4.5	4.0	2.5	2.5	2.5	4.5	4.5	2.5
Other Engineering and Technology - Student	2	5.0	3.0	4.0	3.5	3.5	5.0	3.5	3.5	4.0	3.5	3.5	4.0	4.5	4.5	4.0
Other Engineering and Technology - Rest	10	3.2	4.0	3.7	3.3	3.1	3.9	4.5	3.9	3.2	3.1	3.2	3.0	3.6	3.4	2.7

1 = not at all important / 5 = very important / 0 = no answer

Q_1: Availability of preferred usage method
Q_2: Multilingual interface
Q_3: Open source
Q_4: Existence of an API
Q_5: Integration of other tools
Q_6: Support options
Q_7: Collaborative environment for teams/groups with access rights system
Q_8: File encryption
Q_9: Free-text search capability
Q_{10}: Search filters (faceted search)
Q_{11}: Support for flexible research workflows
Q_{12}: Wizards to guide through processes
Q_{13}: Feature for previewing dataset content
Q_{14}: Visualization of datasets
Q_{15}: Functionality to anonymize research data
Q_{16}: Compliance with data regulations
Q_{17}: Control over data location
Q_{18}: Full control over own data
Q_{19}: Tracking of content changes
Q_{20}: Control over data release dates
Q_{21}: Control over data expiry dates
Q_{22}: Data export in different formats
Q_{23}: Possibility of data citation
Q_{24}: Transformation of data within the RDMS
Q_{25}: Tracking of interactions of external entities with dataset
Q_{26}: Data provenance
Q_{27}: Unique identifiers for data
Q_{28}: Logical links between related data
Q_{29}: Mandatory/optional metadata elements
Q_{30}: Automated metadata tagging
Q_{31}: Possibility to set up multiple metadata schemata
Q_{32}: Domain-specific metadata schemata
Q_{33}: Data validation
Q_{34}: Data identification via DOI/URL
Q_{35}: Compliance with OAI-PMH

Table 3.7: Importance of Criteria for an RDMS - Students in E&T

Conclusion: People who are not familiar with RDM and/or an RDMS have different priorities for such a system. Once they start working with one, these priorities may shift.

USAGE METHOD: 210 participants provided their preferred usage method for an RDMS. They could choose between web application, mobile application, local installation, or a combination thereof. The most popular choice was a web application combined with a local installation (60 participants), closely followed by a web application on its own (55 participants). 36 participants wanted only a local installation, and 28 a combination of all three options. The combination of a web and mobile application (15 participants) and mobile application and local installation (10 participants) was less popular. The mobile application on its own trails behind the rest, asked for by six participants.

On average, the availability of all three platforms together, as well as web and mobile application, and web application and local installation, was rated with more than 4 on a scale of 5 being very important and 1 not at all important. The other options are rated between 3 and 4.

This means that – either stand-alone or in combination – 75 % of the participants want to have a web application available, whilst 64 % still like to have a traditional local installation in the mix. Only 28 % of the participants seek at least in addition a mobile application.

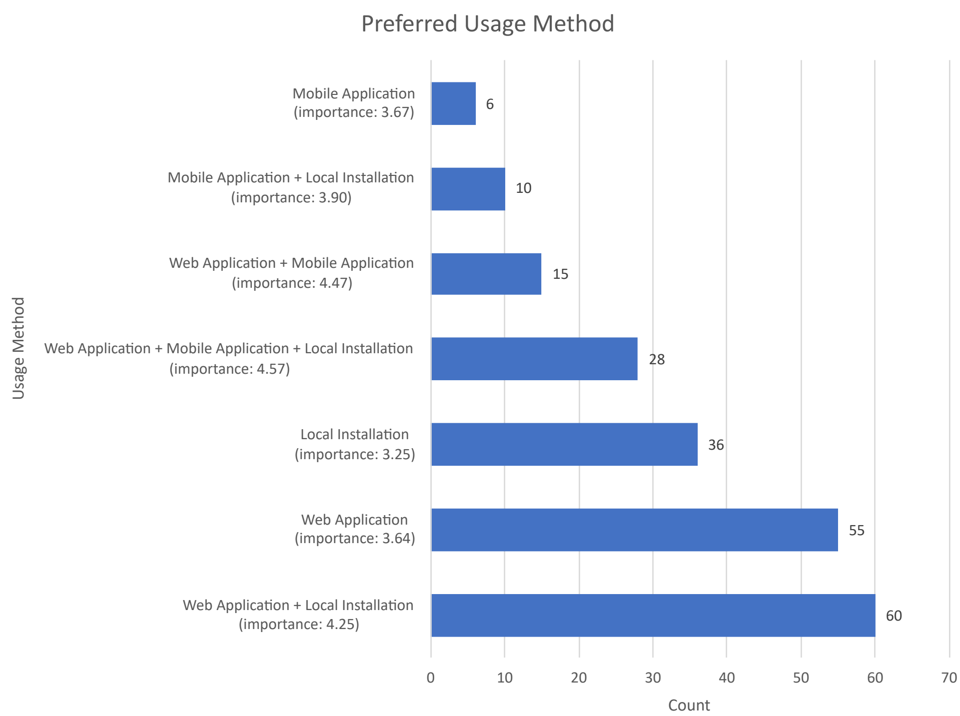


Figure 3.25: Preferred Usage Method for an RDMS

Five of 210 participants did not provide their demographics. Looking into the remaining 205 people, distinct differences between the major scientific fields surface. Sorted from the most to the least popular usage method, 43 % of people working in engineering and technology prefer having a web application and a local installation. In medical and health sciences as well as natural sciences, the share is around one third. For social sciences, humanities, and agricultural sciences, between 20 and 25 % of potential users look for these usage methods. The greatest deviation for a stand-alone web application can be seen in agricultural sciences – only 5 % of respondents prefer it, compared to the most popular choice of a local installation (30 %) for this scientific field. For humanities, a local installation is the second best choice together with a local installation in combination with a web application. For social sciences, a local installation is the third most popular choice, for natural sciences and engineering and technology it ranks forth. For medical and health sciences, it only takes the fifth place of seven. People researching in agricultural sciences are not only advocates for a local installation, but are equally enthusiastic for the combinations web application and local installation, mobile application and local installation, and all three options together (all 20 %). All three options together share the last place for humanities and medical and health sciences, whilst it can be found in the middle ranks for natural sciences, engineering and technology, and social sciences. The medical and health sciences are by far the most mobile scientific fields; 23 % of respondents prefer a combination of web and mobile application, and 14 % would choose only a mobile application. Both options are next to irrelevant for other scientific fields (0-10 %). The reason for that could be that scientists in those fields may tend to collect their research data away from a desk (e.g. on-site with patients), thus having the need to enter them on a tablet or smartphone. The combination of a mobile application with a local installation is out of the question for engineering and technology and social sciences (0 %). For agricultural sciences, it shares the second place with other options, humanities shows it in third place, and for medical and health sciences and natural sciences it is a rather unpopular choice with only 5 %.

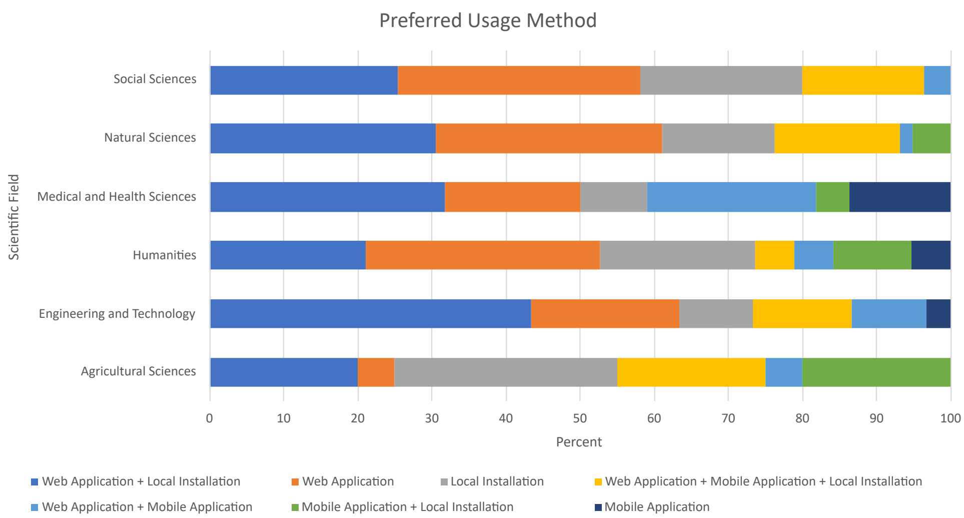


Figure 3.26: Preferred Usage Method by Scientific Fields

Conclusion: When developing an RDMS, it is important to offer both a local installation as well as a web application. The mobile application needs for medical and health sciences have to be investigated further. Other than that, a mobile application can be considered as nice-to-have.

SOFTWARE INTEGRATION: Participants were asked if they would like specific tools integrated into an RDMS, and which ones that would be. In particular, they could choose between either (1) the three software applications most important for their research or the most frequently used ones in this context, (2) other tools, or (3) both options. 159 people shared their requirements. 135 of them wanted the most important software applications they currently use (mentioned in Figures 3.14 and 3.15), eight of those also named additional applications, and another 24 only wanted the applications they specified for this question.

In total, the ranking for specific software resembles very much the one in Figure 3.14, with Excel being the most important one (96 mentions), SPSS in second place (36 mentions), and Zotero as number three (26 mentions). The software applications named were grouped into 40 categories. The top three categories were spreadsheets (97 mentions), software for statistical analysis (76 mentions), and programming software/languages or IDEs (66 mentions) (Figure 3.27). New software categories provided were backup software, software for DMPs, digital libraries, software to display maps with the data, media player, medical records software, project management software, and programs for data synchronization.

Figure 3.27: Software to be integrated into an RDMS.

10

Conclusion: This shows that future users mainly need an integration of their applications they use in their daily work.

ADDITIONAL FEATURES: We gave respondents the opportunity to add any other functions or features they would like an RDMS to offer. 42 people made use of that option. The inputs emphasize the importance of a user-friendly and inclusive system. In terms of features, several interesting functions were on the users' wish list. Interoperability is mentioned a couple of times, as well as several options for visualization. Respondents are worried about the future of their data and want to be sure it is available tamper-proof and in the long-run. The full list of functions and features requested can be viewed in Section 6.8.

Conclusion: When designing a user-centered system, the needs of people with impairments must be taken into account as well. The GUI has to be thoroughly tested using the accessibility options built in for different operating systems. Any accessibility options not available in operating systems should be included in the RDMS. Especially visualization features should be integrated for specific scientific fields. Data has to be secure without an expiration date.

Conclusion on the user needs: It would be beneficial to offer a GUI tailored to the scientific field of the user to cater to the differences in needs. Options can provide an individual look and feel and satisfy shifting priorities of users from when they start working with an RDMS for the first time up to the time they have extensive experience with it. The types of applications most important are a local installation as well as a web application. Future users mainly need an integration of their applications they use in their daily work. Even when offering a lot of features, sight must not be lost of usability and accessibility. Visualization features are demanded by the respondents. Another important topic is data security and persistency.

3.5 RDM CONCEPTS OF UNIVERSITIES

The RDM concepts of two universities in Germany are summarized and evaluated in terms of user-centeredness. Specifically, we focus on the FernUniversität in Hagen and the University of Regensburg.

The FernUniversität in Hagen and the University of Regensburg are two very different universities. Both are public research universities in Germany. The FernUniversität, however, is the only public university in Germany focused entirely on distance learning. Likewise, many staff members work remotely. There are five faculties: Humanities and Social Sciences, Psychology, Mathematics and Computer Science, Economics and Business Administration, and Law. Despite the limited number of faculties, the FernUniversität is the largest in Germany in terms of student numbers (around 70,000 students

10 Created with `http://worditout.com/`

and 1,894 staff members in the summer semester 2023[11]). The University of Regensburg is a traditional university with twelve faculties, almost 21,000 students, and 3,419 staff members in the winter semester 2023.[12]

FERNUNIVERSITÄT IN HAGEN: The FernUniversität in Hagen has regulations in place to ensure good scientific practice. These require researchers to store research data needed for replicating the results of research, archive data for at least ten years, and comply with the FAIR principles, amongst others. An ombudsman and delegates can be consulted for questions about scientific integrity, handling security-related research, or ethics.[13]

The Faculty of Psychology offers more comprehensive research support that covers questions about the entire research data cycle, which is available to all researchers in this faculty. They have a link collection with information on RDM, DMPs, data curation, and archiving. They refer to the Leibniz Institute for Psychology for a repository (PsychArchives) and a platform for study planning and data analysis (PsychNotebook). The faculty supports Open Science through an initiative that aims to strengthen Open Science in the research process, in teaching, and in committees alike.[14]

Conclusion: One faculty has already taken first steps towards RDM. The development of a university-wide RDM concept is an important topic for the university. An introductory event has already taken place in February 2024, but the ongoing process is expected to last several more months.

UNIVERSITY OF REGENSBURG: Through the University of Regensburg (UR) publication server, users have access to the UR Data Hub. The services offered include consulting (e.g. workflows, RDM), digital services (e.g. electronic lab notebooks), as well as training and events (e.g. Q&A sessions).[15]

A policy on RDM has been set up in 2022. At its core are the FAIR principles. The following guidelines have to be complied with: regulation of the UR Data Hub, guidance on the handling of research data, regulations for ensuring good academic practice, Open Access policy. The university already has 241 users for the elabFTW electronic lab notebook, 2101 electronic labbooks, 258 research data on the UR publication server, and 19 DMPs through their Research Data Management Organiser instance.[16]

Conclusion: The University of Regensburg and its library put a lot of effort into supporting researchers with their every RDM needs. They make it as easy as possible for their staff to apply a proper RDM. Only when it comes to tools, they mostly use standardized software instead of a platform tailored to the needs of their users, except for the publication server.

11 `https://www.fernuni-hagen.de/universitaet/zahlen.shtml`, accessed 27-December-2023
12 `https://www.uni-regensburg.de/universitaet/profil/zahlen-und-fakten/index.html`, accessed 27-December-2023
13 `https://www.fernuni-hagen.de/forschung/wisspraxis.shtml`, accessed 10-February-2024
14 `https://www.fernuni-hagen.de/psychologie/forschung/forschungssupport.shtml`, accessed 10-February-2024
15 `https://www.uni-regensburg.de/bibliothek/elektronisches-publizieren/forschungsdaten/index.html`, accessed 09-February-2024
16 `https://epub.uni-regensburg.de/datahub.html`, accessed 09-February-2024

Conclusion on RDM concepts of universities: The FernUniversität in Hagen is making positive progress on implementing first RDM initiatives within the Faculty of Psychology. While a university-wide RDM concept is under development, these initiatives already demonstrate a commitment to RDM best practices. The University of Regensburg also shows experience with RDM, particularly through the UR Data Hub. Their university-wide RDM concept is currently in its early stages. Recognizing the importance of RDM, both universities actively responded to the recent mandatory RDM requirement introduced by the DFG, one of Germany's largest research funders, in March 2022.[17] This thesis aims to further support their efforts by providing recommendations for a university-wide user-centered RDMS.

17 https://www.dfg.de/en/news/news-topics/announcements-proposals/2022/info-wissenschaft-22-25, accessed 15-December-2023

4 DISCUSSION

The importance of a university-wide user-centered RDMS is demonstrated in Section 4.1. Ontologies' roles are explored. This bachelor's thesis is placed in the context of other studies, and it is shown how it can contribute to good scientific practice. Critical comments that were made about the user survey are addressed in Section 4.2.

4.1 DISCUSSION

Due to the vastly growing amount of data produced, the increasing pressure of reproducibility of research results, and the strict requirements of funding agencies, institutions acknowledge the importance of RDM. However, the answers of many respondents show an ongoing need of **education** in this area for people involved in research (Section 3.4.2). **Flexible storage options** for an RDMS are recommended to address researchers' storage needs and secure data without an expiration date.

Especially emerging multi- or interdisciplinary sciences, such as computational linguistics (FOS classification in the Frascati manual: Humanities – Languages and Literature [Com02]) or environmental and geological engineering (FOS classification Engineering and technology – Environmental engineering), make it difficult to use an RDMS specific to a single scientific field like some institutes do. Biotechnology, for example, was recommended to be divided into four components assigned to three different major scientific fields ([Com02]). The same applies to institutes which are not specialized in a single scientific field. They often end up utilizing several RDMS. Although sciences become more intertwined, most studies on RDMS either evaluated requirements for specific scientific fields or did not focus on user needs.

Particularly in institutes housing several scientific fields, a single system is easier to maintain. At the same time, there is an increased risk that it is either too complicated or does not satisfy all user's needs. That makes a system for all scientific fields challenging to implement. This work shows that there is no "one-size-fits-all". Instead, it is recommended to **use a single RDMS with adaptations on the GUI**, depending on the scientific field of the user. When using the FOS classification, this should not only be based on one of the six major scientific fields but at least the 2-digit classification (e.g. Humanities – History and archaeology). An **integration of software** such as reference management into an RDMS can reduce the number of tools required for the daily work of researchers. At the same time, it can reduce the maintenance effort and licenses required by institutions. An RDMS which has **efficient options for input and addresses security concerns** of users and institutions can promote the transition from analog to digital data, reducing additional workload to digitize data after the fact.

Another approach to a single RDMS can be **ontologies**, which are also covered in the literature ([BC18][BS22]). According to the World Wide Web Consortium (W3C), "ontologies are formalized vocabularies of terms, often covering a specific domain and shared by a community of users. They specify the definitions of terms by describing their relationships with other terms in the ontology"[1]. Ontologies improve the interoperability and findability of data. They can be represented by knowledge graphs ([VKP23]). A search function, implemented as free-text search and/or faceted search (search filters), is among the top three criteria for an RDMS overall and for all major scientific fields except social sciences (Section 3.4.4). This demonstrates the importance that respondents see in the findability of data. Therefore, ontologies could contribute to a user-centered system, accessible through another search option and when entering and updating metadata.

With the increasing level of digitalization and availability of mobile devices in workplaces, user needs are changing. This thesis provides a fresh perspective on user needs, some of which were evaluated in earlier studies several years ago and are listed in Section 2.1. The evaluated user needs for an RDMS collectively encompass all four FAIR principles: Findability, Accessibility, Interoperability, and Reusability (Tables 3.1 and 3.2). While no single user need addresses all principles alone, their combined consideration ensures an RDMS aligned with FAIR data management practices. Notably, **the three most important criteria identified by the respondents (Section 3.4.4) also collectively cover all four FAIR principles**, further highlighting their critical role in user-centered RDMS design. In the first place is **full control over own data**, followed by **tracking of content changes**. The third place is shared by **data export in different formats, search filters (faceted search), and unique identifiers for data**. Thus, by taking into account the user needs, the way is paved for repeatable research results and reusable data.

Considering the availability of mobile devices in workplaces, **a local installation as well as a web application** should be developed. Such a combination provides an offline availability, and therefore a spontaneous access without a PC or the Internet, which was seen as an issue in presently available electronic lab notebooks. [Fic+22] During the development process, the **needs of people with impairments** should be clearly considered. Participants voiced their desire for **visualization features**, which in return could be satisfied through integration of other software.

Three hypotheses were tested with the survey data. The first hypothesis (Section 3.4.3.1) showed that more technically oriented scientific fields such as engineering and technology and natural sciences have a slight tendency to favor an RDMS with rich features whilst sacrificing a bit of its simplicity, compared to less technical fields such as humanities and social sciences. However, there was not enough evidence to suggest an association between the technical orientation of the scientific field and the demand for either an

1 `https://www.w3.org/TR/owl2-overview/`, accessed 13-February-2024

easy to use RDMS or rich features at this point, disproving the hypothesis. The second hypothesis (Section 3.4.3.2) supports the demand for education on RDM. It shows that people familiar with RDM and/or an RDMS are more likely to use an RDMS for the entire research data cycle. The link between the age of a person and the desire to use an RDMS was disproven with the third hypothesis (Section 3.4.3.3).

In summary, a university-wide user-centered RDMS must fulfill the following organizational and technical requirements, without implying any specific order of importance:

- Education on RDM and integration of this topic into the curricula of universities
- Availability of both local installation and web application
- Single RDMS with adaptations on the GUI to reduce cluttering, depending on the scientific field of the user
- Accessible GUI for people with impairments
- Implementation of ontologies through knowledge graphs, search options, and integration into metadata
- Flexible storage options
- No expiration date for data (long-term storage)
- Integration of existing software into the RDMS
- Efficient options for input to avoid double entries (first analog, then digital)
- Full control over data for the data owner
- Tracking of content changes
- Data export in different formats
- Search filters (faceted search)
- Unique identifiers for data
- Proper access-rights system and file encryption
- Visualization features

4.2 CRITICAL COMMENTS

During the survey period, some critical comments were submitted. These shall be addressed in this section, along with an evaluation of their impact on the results.

Before starting the actual survey period, a pretest was conducted and amendments made according to the input received. However, a few participants made remarks about some obscurities or shortcomings of the questionnaire, only surfacing when it was already online. They shall be addressed as follows:

- The design of the questionnaire was collating several criteria within a matrix to be graded according to a Likert-like scale (see Figure 4.1). One expert in questionnaire design and User Experience (UX) shared the opinion that it could have been confusing for some respondents not having individual questions to answer. Furthermore, the scale should not have been labeled at the poles only. This may potentially have led to different interpretations of the scale amongst respondents, although this labeling is shown in literature as well (see [May13], Figures 31 and 33).

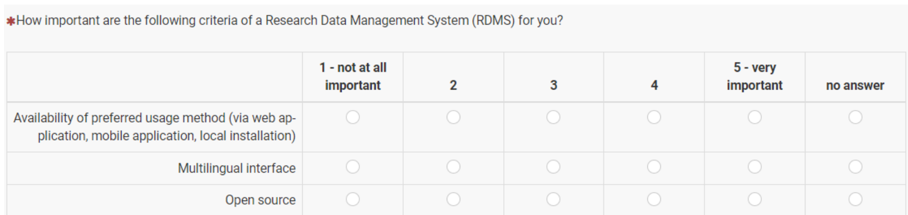


Figure 4.1: User Needs[2]

Evaluation: The presentation of those questions in the form of a matrix was well received by most pretest subjects. The matrix was kept small to be entirely visible inside the browser. A responsive design displayed the scale for every question on small screens. The question format is widely used by pollsters, thus familiar to many people. It is assumed that the incomplete labeling was treated as a linear scale due to its size.

- The presentation of the yes/no button for the question about the analog storage of research data on the survey platform was ambiguous for one respondent. The screenshot in Figure 4.2 shows the appearance when "No" (German: "Nein") is selected. By selecting "Yes", the respondents received additional questions about the analog storage, thus it is less likely that "Yes" was selected by accident. It should be taken into consideration that some respondents may have falsely selected "No". However, given that initially both buttons were blue and only changed color when selecting one, the risk for a false selection is considered minor.

Figure 4.2: No Analog Storage[3]

Evaluation: 146 respondents selected "No" and 109 selected "Yes". Considering how forthcoming many of the participants of the first group were, it cannot be ruled out that some of them misinterpreted the selection. Therefore, the percentage of analog data in Section 3.4.2 may

2 Created with `https://www.limesurvey.org` (LimeSurvey GmbH)
3 Created with `https://www.limesurvey.org` (LimeSurvey GmbH)

not be accurate. However, the answers received in the subsequent filter questions on the type of analog data and the reason for not digitalizing were numerous and provided a good overview.

- A professor of philosophy commented that many of the questions had no relevance for humanities, specifically for this field of science.

  Evaluation: Discussions with researchers in the field of humanities revealed that there are no questions missing.

- One respondent noted that only a small selection of terms was explained in the questionnaire. Additional ones were used and lacked an explanation, although they think that those terms may not generally be known in research (e.g. data citation, flexible research workflows). This could have increased the risk of misinterpretation by respondents and thus wrong prioritization.

  Evaluation: "No answer" was chosen 25 times for "support for flexible research workflows" and 15 times for "possibility of data citation". In comparison, "no answer" was chosen 24 times for "transformation of data within the RDMS" which had an explanation added. Therefore, it can be expected that there was no significant lack of explanations that led to incomplete answers.

- A participant perceived that the definition of research data in the questionnaire was not precise enough, especially in regard to a possible inclusion of documentation.

  Evaluation: Documentary material is explicitly mentioned in the definition. It is based on the definition provided by forschungsdaten.info.[4]

- The question about the three software applications that are most important for the respondent's research or that the respondent uses most frequently "in this context" was not precise enough for one participant, despite the examples given. There was a risk that applications like Word or PowerPoint could be entered.

  Evaluation: Some applications such as Word or PowerPoint were mentioned by respondents, but the majority of software names entered seem to be directly related to research.

- The question "Which kind of research data?", referring to whether research data is stored in analog form, was ambiguous for one person.

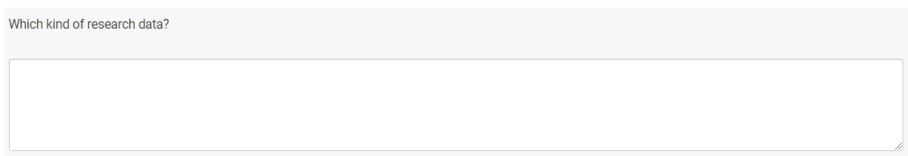


Figure 4.3: Research Data Type[5]

4 https://forschungsdaten.info/
5 Created with https://www.limesurvey.org (LimeSurvey GmbH)

Evaluation: The size of the text box (Figure 4.3) was an indicator that the expected input is not only "quantitative" or "qualitative" but rather if the data is biological samples, handwritten notes, questionnaires on paper, etc. Only one respondent out of 101 who filled in this field answered with "Sociodemographic variables – quantitative and qualitative", all others entered the expected input.

To summarize, the impact of the potential flaws in the questionnaire on the answers provided was minimal. Useful data could be collected to evaluate user needs through a thorough analysis and gain valuable results.

5 SUMMARY AND OUTLOOK

This chapter summarizes the bachelor's thesis. It also provides an outlook for topics of future studies.

5.1 SUMMARY

The objective of this bachelor's thesis was the evaluation of the needs of potential future users of a university-wide user-centered RDMS. It results in best practices and a requirements review for an upcoming development process of such a system.

In Chapter 1, this bachelor's thesis looked into the importance of RDM, and how an RDMS can support the research data cycle. The FAIR principles were explained. A literature review on RDMS was conducted in Chapter 2, which resulted in a list of criteria to be evaluated for such a system. The criteria, along with background information on the (potential) users, were assessed in an online survey. During the eight weeks survey period, a total of 409 questionnaires were filled in. 269 questionnaires were considered valid after the data cleaning.

The data analysis first gathered demographic data on the respondents (Section 3.4.1). It covers the countries the respondents were based at, the scientific field they worked in, the position(s) they hold, and their age group and gender. Section 3.4.2 details the respondents' status quo on familiarity and experience with RDM and an RDMS. It demonstrates a lack of knowledge in this area. Data on existing RDMS, along with the pros and cons, was collected. Users were predominantly happy with their system. However, based on their experiences with their system, they provided useful input on the attributes and functions that could be important or potentially problematic in an RDMS. This section analyzes the discrepancies between respondents' current RDMS usage across research data cycle phases and their desired future use. It continues with the most important software applications the participants used in their research. The evaluation of analog research data shows the amount of data still not digitalized and the reasons for keeping analog data. The file size of the typical digital outputs reveal the storage requirements of the six major scientific fields.

Three hypotheses were validated in Section 3.4.3. Hypothesis 1 does not provide enough evidence to suggest an association between the technical orientation of the scientific field and the demand for either an easy to use RDMS or rich features. Hypothesis 2 confirms that people who are not familiar with RDM and/or an RDMS are less likely to be willing to use an RDMS for the entire research data cycle. Hypothesis 3 is disproven, because the survey data did not show that digital natives are more likely to use an RDMS than previous generations.

The data analysis concludes with a section on the user needs (Section 3.4.4). The criteria defined for an RDMS were compared by scientific field and position. Preferred usage methods (web application, mobile application, local installation) were evaluated. Possible software integrations, based on the answers analyzed in Section 3.4.2, were considered. The respondents' requests for additional features were assessed in the end.

This chapter closes with RDM concepts of universities. The concepts of the FernUniversität in Hagen and the University of Regensburg were summarized and evaluated in terms of user-centeredness.

5.2 OUTLOOK

A subsequent study could investigate further the needs of users working in those scientific fields and their respective sub-disciplines that were represented only by a small group of respondents in this survey. It could clarify the needs for a mobile application for medical and health sciences, which showed a great demand in this survey (Section 3.4.4). People in agricultural sciences had a high preference for the combination of a mobile application with a local installation, which could also be looked at in more detail. Ontologies and their efficient and user-friendly implementation in an RDMS can be another potential topic for future work.

The GUI design criteria resulting from this thesis, as well as the complimentary bachelor's and master's theses, can be implemented into a horizontal prototype. A horizontal prototype represents the complete GUI. However, the menus and functions are predominantly not implemented, therefore it is easy to develop. On the contrary, a vertical prototype has the GUI and a few functions implemented in detail. Therefore, complex functions can be tested with vertical prototypes. They are also used for feasibility studies. Since the user can see the entire GUI and the functions that will be offered a prototype can seem like a finished system. [GKP22] A diverse range of future users in respect to scientific field and experience with RDM and RDMS can give their feedback on the prototype in an interview study.

6 APPENDIX

The appendix contains the lists of scientific fields and countries used in the online survey, followed by the questionnaire in English and German. The responses of the participants about their positions (Section 6.5) as well as the type of analog research data and the reasons for not yet having it digitalized (Section 6.6) are listed. The criteria for an RDMS grouped by the major scientific fields are available in Section 6.7. In the end is a list of functions and features respondents would like to see in an RDMS (Section 6.8).

6.1 SCIENTIFIC FIELDS

The list of scientific fields used for the questionnaire is based on the FOS classification in the Frascati manual [Com02]. Table 6.1 contains the English list as well as the German translations of the *Austrian Fields of Science and Technology Classification (ÖFOS)* [Dir11].

6.2 COUNTRIES

Tables 6.2 and 6.3 list the countries of the questionnaire in English and German.

Major Scientific Field	Sub-Discipline (English)	Sub-Discipline (German)
Natural Sciences / Naturwissenschaften	Mathematics	Mathematik
	Computer and information sciences	Informatik
	Physical sciences	Physik, Astronomie
	Chemical sciences	Chemie
	Earth and related environmental sciences	Geowissenschaften
	Biological sciences	Biologie
	Other natural sciences	Andere Naturwissenschaften
Engineering and technology / Technische Wissenschaften	Civil engineering	Bauwesen
	Electrical engineering, Electronic engineering, Information engineering	Elektrotechnik, Elektronik, Informationstechnik
	Mechanical engineering	Maschinenbau
	Chemical engineering	Chemische Verfahrenstechnik
	Materials engineering	Werkstofftechnik
	Medical engineering	Medizintechnik
	Environmental engineering	Umweltingenieurwesen, Angewandte Geowissenschaften
	Environmental biotechnology	Umweltbiotechnologie
	Industrial biotechnology	Industrielle Biotechnologie
	Nano-technology	Nanotechnologie
	Other engineering and technologies	Andere Technische Wissenschaften
Medical and Health sciences / Humanmedizin, Gesundheitswissenschaften	Basic medicine	Medizinisch-theoretische Wissenschaften, Pharmazie
	Clinical medicine	Klinische Medizin
	Health sciences	Gesundheitswissenschaften
	Medical biotechnology	Medizinische Biotechnologie
	Other medical sciences	Andere Humanmedizin, Gesundheitswissenschaften
Agricultural sciences / Agrarwissenschaften, Veterinärmedizin	Agriculture, Forestry, and Fisheries	Land- und Forstwirtschaft, Fischerei
	Animal and Dairy science	Tierzucht, Tierproduktion
	Veterinary science	Veterinärmedizin
	Agricultural biotechnology	Agrarbiotechnologie, Lebensmittelbiotechnologie
	Other agricultural sciences	Andere Agrarwissenschaften
Social sciences / Sozialwissenschaften	Psychology	Psychologie
	Economics and Business	Wirtschaftswissenschaften
	Educational sciences	Erziehungswissenschaften
	Sociology	Soziologie
	Law	Rechtswissenschaften
	Political science	Politikwissenschaften
	Social and economic geography	Humangeographie, Regionale Geographie, Raumplanung
	Media and communications	Medien- und Kommunikationswissenschaften
	Other social sciences	Andere Sozialwissenschaften
Humanities / Geisteswissenschaften	History and Archaeology	Geschichte, Aräologie
	Languages and Literature	Sprach- und Literaturwissenschaften
	Philosophy, Ethics and Religion	Philosophie, Ethik, Religion
	Arts (arts, history of arts, performing arts, music)	Kunstwissenschaften
	Other humanities	Andere Geisteswissenschaften

Table 6.1: List of Scientific Fields

English	German	English	German
Afghanistan	Afghanistan	Greece	Griechenland
Albania	Albanien	Grenada	Grenada
Algeria	Algerien	Guatemala	Guatemala
Andorra	Andorra	Guinea	Guinea
Angola	Angola	Guinea-Bissau	Guinea-Bissau
Antigua and Barbuda	Antigua und Barbuda	Guyana	Guyana
Argentina	Argentinien	Haiti	Haiti
Armenia	Armenien	Holy See	Vatikan
Australia	Australien	Honduras	Honduras
Austria	Österreich	Hungary	Ungarn
Azerbaijan	Aserbaidschan	Iceland	Island
Bahamas	Bahamas	India	Indien
Bahrain	Bahrain	Indonesia	Indonesien
Bangladesh	Bangladesch	Iran	Iran
Barbados	Barbados	Iraq	Irak
Belarus	Belarus	Ireland	Irland
Belgium	Belgien	Israel	Israel
Belize	Belize	Italy	Italien
Benin	Benin	Jamaica	Jamaika
Bhutan	Bhutan	Japan	Japan
Bolivia	Bolivien	Jordan	Jordanien
Bosnia and Herzegovina	Bosnien und Herzegovina	Kazakhstan	Kasachstan
Botswana	Botswana	Kenya	Kenia
Brazil	Brasilien	Kiribati	Kiribati
Brunei	Brunei	Kuwait	Kuwait
Bulgaria	Bulgarien	Kyrgyzstan	Kirgistan
Burkina Faso	Burkina Faso	Laos	Laos
Burundi	Burundi	Latvia	Lettland
Côte d'Ivoire	Elfenbeinküste	Lebanon	Libanon
Cabo Verde	Kap Verde	Lesotho	Lesotho
Cambodia	Kambodscha	Liberia	Liberia
Cameroon	Kamerun	Libya	Libyen
Canada	Kanada	Liechtenstein	Liechtenstein
Central African Republic	Zentralafrikanische Republik	Lithuania	Litauen
Chad	Tschad	Luxembourg	Luxemburg
Chile	Chile	Madagascar	Madagaskar
China	China	Malawi	Malawi
Colombia	Kolumbien	Malaysia	Malaysia
Comoros	Komoren	Maldives	Malediven
Congo (Congo-Brazzaville)	Kongo (Kongo-Brazzaville)	Mali	Mali
Costa Rica	Costa Rica	Malta	Malta
Croatia	Kroatien	Marshall Islands	Marshallinseln
Cuba	Kuba	Mauritania	Mauretanien
Cyprus	Zypern	Mauritius	Mauritius
Czechia (Czech Republic)	Tschechien	Mexico	Mexiko
Democratic Republic of the Congo	Demokratische Republik Kongo	Micronesia	Mikronesia
Denmark	Dänemark	Moldova	Moldau
Djibouti	Dschibuti	Monaco	Monaco
Dominica	Dominica	Mongolia	Mongolei
Dominican Republic	Dominikanische Republik	Montenegro	Montenegro
Ecuador	Ecuador	Morocco	Marokko
Egypt	Ägypten	Mozambique	Mosambik
El Salvador	El Salvador	Myanmar (formerly Burma)	Myanmar (vorm. Burma)
Equatorial Guinea	Äquatorialguinea	Namibia	Namibia
Eritrea	Eritrea	Nauru	Nauru
Estonia	Estland	Nepal	Nepal
Eswatini (fmr. "Swaziland")	Eswatini (vorm. "Swaziland")	Netherlands	Niederlande
Ethiopia	Äthiopien	New Zealand	Neuseeland
Fiji	Fidschi	Nicaragua	Nicaragua
Finland	Finnland	Niger	Niger
France	Frankreich	Nigeria	Nigeria
Gabon	Gabun	North Korea	Nordkorea
Gambia	Gambia	North Macedonia	Nordmazedonien
Georgia	Georgien	Norway	Norwegen
Germany	Deutschland	Oman	Oman
Ghana	Ghana	Pakistan	Pakistan

Table 6.2: List of Countries (1/2)

English	German	English	German
Palau	Palau	Spain	Spanien
Palestine State	Palästina	Sri Lanka	Sri Lanka
Panama	Panama	Sudan	Sudan
Papua New Guinea	Papua-Neuguinea	Suriname	Suriname
Paraguay	Paraguay	Sweden	Schweden
Peru	Peru	Switzerland	Schweiz
Philippines	Philippinen	Syria	Syrien
Poland	Polen	Tajikistan	Tadschikistan
Portugal	Portugal	Tanzania	Tansania
Qatar	Katar	Thailand	Thailand
Romania	Rumänien	Timor-Leste	Timor-Leste
Russia	Russland	Togo	Togo
Rwanda	Ruanda	Tonga	Tonga
St. Kitts and Nevis	St. Kitts und Nevis	Trinidad and Tobago	Trinidad und Tobago
St. Lucia	St. Lucia	Tunesia	Tunesien
St. Vincent and the Grenadines	St. Vincent und die Grenadinen	Türkiye	Türkei
Samoa	Samoa	Turkmenistan	Turkmenistan
San Marino	San Marino	Tuvalu	Tuvalu
Sao Tome and Principe	Sao Tome und Principe	Uganda	Uganda
Saudi Arabia	Saudi-Arabien	Ukraine	Ukraine
Senegal	Senegal	United Arab Emirates	Vereinigte Arabische Emirate
Serbia	Serbien	United Kingdom	Vereinigtes Königreich
Seychelles	Seychellen	United States of America	Vereinigte Staaten von Amerika
Sierra Leone	Sierra Leone	Uruguay	Uruguay
Singapore	Singapur	Uzbekistan	Usbekistan
Slovakia	Slowakei	Vanuatu	Vanuatu
Slovenia	Slowenien	Venezuela	Venezuela
Solomon Islands	Salomonen	Vietnam	Vietnam
Somalia	Somalia	Yemen	Jemen
South Africa	Südafrika	Zambia	Sambia
South Korea	Südkorea	Zimbabwe	Simbabwe
South Sudan	Südsudan		

Table 6.3: List of Countries (2/2)

6.3 QUESTIONNAIRE (ENGLISH)

The text of the questionnaire is fully reflected here. Typos in the original text remained here as well. All footnotes are written in blue. The questionnaire begins below.

Research Data Management Systems (RDMS)

This survey is being carried out as part of the bachelor thesis of Sabine Diemt (bachelor student of computer science) at the FernUniversität in Hagen (Prof. Dr. Uta Störl, Chair of Database and Information Systems), in cooperation with the University of Regensburg (thesis advisor Dr.-Ing. Tanja Auge, Data Engineering).

The bachelor thesis evaluates the needs of potential users of a Research Data Management System (RDMS) and examines universities' research data management concepts with regard to their user focus. Based on the results of this survey, best practices and a requirements review for such a system are developed.

Please support us by sharing your personal requirements with us. The survey is available in English and German and takes about 10 minutes to complete.

RDMS

We have some questions about your experience with RDMS (Research Data Management Systems). In case you are not familiar with the term, we'd like to provide you with a brief explanation:

*By **research data** we mean all data which was collected for scientific purposes, including working with measurements, surveys, and documentary material, but excluding interpretative publications like journal articles or dissertations. They are the base for research and document the results.*

***Research Data Management (RDM)** covers the whole research cycle from data collection to digital preservation. It takes into account methods and concepts as well as organizational and technical procedures.*

*A **Research Data Management System (RDMS)** is "the technical framework to collect, describe, and provide research data [..], also called data repository, archive or project database."*

Question: Are you familiar with RDM and/or a RDMS?
Choose one of the following answers

- yes *-> Filter*
- no
- I'm not sure

Filter: What's your experience with the use of a RDMS?
Choose one of the following answers

- less than 1 year
- 1 year to less than 2 years
- 2 years to less than 5 years
- 5 years or more

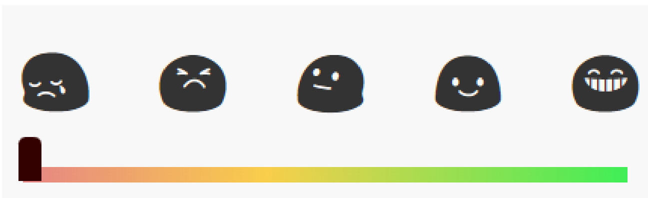
Question: Does your institution have a RDMS (self developed or purchased/external)?
Choose one of the following answers

- yes *-> Filter*
- no
- I don't know

Filter: What's the name of the RDMS?

Answers of less than 10 respondents will be anonymized in the thesis to prevent identification of the individual.

Filter: Does the RDMS you currently use fulfil your needs?



Question type: 5 point choice. Slider layout: Slider with emoticon. Created with `https://www.limesurvey.org` *(LimeSurvey GmbH)*

Filter: Is there anything in particular that you like about your current RDMS?

____

Filter: Is there anything in particular that you don't like about your current RDMS?

____

Filter: For which phases of the research data cycle do you use a RDMS?
Check all that apply

- ☐ Planning
- ☐ Data collection
- ☐ Data processing and analysis
- ☐ Data archiving
- ☐ Data reuse

Question: Which three software applications are most important for your research or do you use most frequently in this context (e.g. Excel, SPSS, Gocad, Zotero, specific databases, MATLAB, Eclipse, ArcGIS, OneNote etc.)?

____

____

____

Answers will be anonymized in the thesis if a software is used by less than 10 respondents.

Question: Do you also store research data in analog form (not digitalized), such as biological samples, handwritten notes, etc.?

Question type: Yes/No. Display type: Button group. Created with `https://www.limesurvey.org` *(LimeSurvey GmbH)*

Filter (only if the question about analog storage was answered with "Yes"): Which kind of research data?

____

Filter (only if the question about analog storage was answered with "Yes"): Why have you not digitalized this data?

____

Question: What is the max. file size of a typical digital output (e.g. databases, Excel spreadsheets, interview recordings, etc.) of your research in MB?

Only numbers may be entered in this field.

____

User Needs 1/4

The following four sections are about your wishes and needs for a research data management system (RDMS). We would like to know what your ideal RDMS looks like.

Question: How important are the following criteria of a Research Data Management System (RDMS) for you?[a] *(selection mandatory)*

	1 - not at all important	2	3	4	5 - very important	no answer
Availability of preferred usage method (via web application, mobile application, local installation)	◯	◯	◯	◯	◯	◯
Multilingual interface	◯	◯	◯	◯	◯	◯
Open Source	◯	◯	◯	◯	◯	◯
Existance of an API (application programming interface)	◯	◯	◯	◯	◯	◯
Integration of other tools	◯	◯	◯	◯	◯	◯
Support options (e.g. integrated manual, help via chat, support hotline, etc.)	◯	◯	◯	◯	◯	◯
Collaborative environment for teams/groups with an access rights system	◯	◯	◯	◯	◯	◯
File encryption	◯	◯	◯	◯	◯	◯
Free-text search capability	◯	◯	◯	◯	◯	◯
Search filters (faceted search)	◯	◯	◯	◯	◯	◯
Support for flexible research workflows	◯	◯	◯	◯	◯	◯
Wizards to guide you through processes (e.g. a dialog box guiding through a data import)	◯	◯	◯	◯	◯	◯

a All answer options had the same size in the original questionnaire.

User Needs 2/4

The following three sections are about your wishes and needs for a research data management system (RDMS). We would like to know what your ideal RDMS looks like.

Question: How important are the following criteria of a Research Data Management System (RDMS) for you?[a] *(selection mandatory)*

Terms marked with a number are briefly explained in the help underneath.

	1 - not at all important	2	3	4	5 - very important	no answer
Feature for previewing dataset content	○	○	○	○	○	○
Visualization of datasets	○	○	○	○	○	○
Functionality to anonymize research data	○	○	○	○	○	○
Compliance with data regulations (GDPR, NDA, ethics, etc.)	○	○	○	○	○	○
Control over data location (locale/cloud/specific country)	○	○	○	○	○	○
Full control over own data (e.g. deletion, editing, export, etc. of data)	○	○	○	○	○	○
Tracking of content changes	○	○	○	○	○	○
Control over data release dates	○	○	○	○	○	○
Control over data expiry dates	○	○	○	○	○	○
Data export in different formats	○	○	○	○	○	○
Possibility of data citation	○	○	○	○	○	○
Transformation (1) of data within the RDMS	○	○	○	○	○	○

(1) **Transformation** of data includes converting, cleaning and structuring the data into a format useful for further processing (e.g. deleting duplicates, converting values, converting to other file formats, etc.).

a All answer options had the same size in the original questionnaire.

User Needs 3/4

The following two sections are about your wishes and needs for a research data management system (RDMS). We would like to know what your ideal RDMS looks like.

Question: How important are the following criteria of a Research Data Management System (RDMS) for you?[a] *(selection mandatory)*

Terms marked with a number are briefly explained in the help underneath.

	1 - not at all important	2	3	4	5 - very important	no answer
Tracking of interactions of external entities with the dataset (e.g. citations)	◯	◯	◯	◯	◯	◯
Data provenance (How was the data created/modified/used?)	◯	◯	◯	◯	◯	◯
Unique identifiers for data	◯	◯	◯	◯	◯	◯
Logical links between related data to enhance discoverability	◯	◯	◯	◯	◯	◯
Mandatory and optional metadata elements	◯	◯	◯	◯	◯	◯
Automated metadata tagging	◯	◯	◯	◯	◯	◯
Possibility to set up multiple metadata schemata	◯	◯	◯	◯	◯	◯
Domain-specific metadata schemata (1) available	◯	◯	◯	◯	◯	◯
Data validation (enforcing correct data/metadata (2) structure, file format, etc.)	◯	◯	◯	◯	◯	◯
Data identification via DOI (3) or URL	◯	◯	◯	◯	◯	◯
Compliance with OAI-PMH (4)	◯	◯	◯	◯	◯	◯

(1) **Domain-specific metadata schemata** are special schemas for a scientific field (e.g. Darwin Core for natural history, Text Encoding Initiative for humanities, etc.)

(2) **Metadata** is data about data (e.g. author, publication year, version, etc.) and can be divided into descriptive, technical, preservation, rights, and structural metadata and markup languages.

(3) **DOI** stands for Digital Object Identifier and is a unique number. It allows objects (e.g. journal articles) to be permanently identified and reliably accessed.

(4) **OAI-PMH** is the Open Archives Initiative Protocol for Metadata Harvesting and is used to harvest and process metadata. This shall improve the findability and usability of electronic publications on the Internet.

a All answer options had the same size in the original questionnaire.

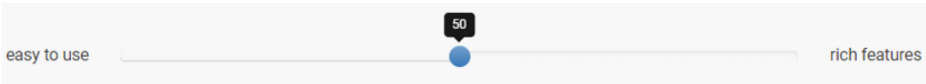
User Needs 4/4

The following section is about your wishes and needs for a research data management system (RDMS). We would like to know what your ideal RDMS looks like.

Question: Is there anything else you would like a RDMS to offer? Please mention any function/feature that comes to mind.

____

Question: Do you prefer an easy to use RDMS or rather rich features?



Question type: Multiple numerical input. Use slider layout: On. Created with `https://www.limesurvey.org` *(LimeSurvey GmbH)*

We consider it easy to use if you can use the RDMS without extensive training or studying a user manual.

Question: Do you prefer to use a RDMS via the Internet browser (web application), via smartphone/tablet (mobile application), via the computer (local installation), or a combination thereof?

Check all that apply

- ☐ Web application
- ☐ Mobile application
- ☐ Local Installation

Question: If you would like any specific tools integrated with the RDMS, which ones would that be?

Check all that apply

- ☐ The three software applications mentioned before
- ☐ Other: ______________________

Question: For which phases of the research data cycle would you like to use a RDMS?
Check all that apply

☐ Planning
☐ Data collection
☐ Data processing and analysis
☐ Data archiving
☐ Data reuse

Demographics

We'd like to finish with some statistical data.

Question: Name of institution

____
Answers will be anonymized in the thesis to prevent identification of the individual.

Question: Country of institution
Choose one of the following answers

List with all country names according to Tables 6.2 through 6.3

Question: Which position(s) do you currently hold? *(selection mandatory)*
Check all that apply

☐ Professor
☐ Postdoc
☐ Research assistant
☐ Data curator
☐ Data scientist
☐ Funder
☐ Other: ______________________

Question: Which kind of institution do you belong to?

Choose one of the following answers

- ○ University
- ○ Non-university research institution
- ○ Other: ____________________

Question: Which scientific field do you primarily work in? *(selection mandatory)*

Choose one of the following answers

List with scientific fields according to Table 6.1

Question: What age group do you belong to?

Choose one of the following answers

- ○ ≤ 29
- ○ 30-39
- ○ 40-49
- ○ 50-59
- ○ ≥ 60

Choose one of the following answers

- ○ male
- ○ female
- ○ diverse

Thank you very much for your help in shaping a new RDMS. We really appreciate your contribution.

If you would like to receive the resulting bachelor thesis, you can leave your email address here. The separate link ensures that it can't be associated with your answers. The collection is subject to the same data protection regulations as this survey.

6.4 QUESTIONNAIRE (GERMAN)

The text of the questionnaire is fully reflected here. A missing comma in the original help text for metadata was left out here as well. All footnotes are written in blue. The questionnaire begins below.

Forschungsdatenmanagementsysteme (FDMS)
Diese Umfrage wird im Rahmen der Bachelorarbeit von Sabine Diemt (Studierende des Lehrgangs Bachelor Informatik) durchgeführt, die in Zusammenarbeit zwischen der FernUniversität in Hagen (Prof. Dr. Uta Störl, Inhaberin des Lehrgebiets Datenbanken und Informationssysteme) und der Universität Regensburg (Betreuerin Dr.-Ing. Tanja Auge, Lehrstuhl Data Engineering) entsteht.

Die Bachelorarbeit evaluiert die Bedürfnisse potenzieller Benutzer eines Forschungsdatenmanagementsystems (FDMS) und untersucht Forschungsdatenmanagementkonzepte von Universitäten in Hinblick auf ihren Nutzerfokus. Basierend auf den Ergebnissen dieser Umfrage werden Best Practices sowie eine Anforderungsanalyse an ein solches System entwickelt.

Bitte unterstützen Sie uns dabei, indem Sie Ihre ganz persönlichen Anforderungen mit uns teilen. Die Umfrage ist auf Deutsch und Englisch verfügbar und dauert in etwa 10 Minuten.

FDMS

Wir haben einige Fragen zu Ihren Erfahrungen mit FDMS (Forschungsdatenmanagementsystemen). Falls Sie mit dem Begriff nicht vertraut sind, möchten wir Ihnen eine kurze Erklärung geben:

*Unter **Forschungsdaten** verstehen wir alle Daten, die für wissenschaftliche Zwecke erhoben wurden, einschließlich der Arbeit mit Messungen, Umfragen und Dokumentationsmaterial, jedoch mit Ausnahme interpretativer Veröffentlichungen wie Zeitschriftenartikel oder Dissertationen. Diese sind die Grundlage für die Recherche und dokumentieren die Ergebnisse.*

***Forschungsdatenmanagement (FDM)** deckt den gesamten Forschungszyklus von der Datenerfassung bis zur digitalen Aufbewahrung ab. Es berücksichtigt Methoden und Konzepte sowie organisatorische und technische Abläufe.*

*Ein **Forschungsdatenmanagementsystem (FDMS)** ist der technische Rahmen zum Sammeln, Beschreiben und Bereitstellen von Forschungsdaten, auch Datenrepositorium, Archiv oder Projektdatenbank genannt.*

Frage: Sind Sie mit FDM und/oder FDMS vertraut?
Bitte wählen Sie eine der folgenden Antworten:

- ○ ja -> *Filter*
- ○ nein
- ○ ich bin nicht sicher

Filter: Wie viel Erfahrung haben Sie mit der Arbeit mit einem FDMS?
Bitte wählen Sie eine der folgenden Antworten:

- ○ weniger als 1 Jahr
- ○ 1 bis weniger als 2 Jahre
- ○ 2 bis weniger als 5 Jahre
- ○ 5 Jahre oder mehr

Frage: Hat Ihre Institution ein FDMS (selbstentwickelt oder zugekauft/extern)?
Bitte wählen Sie eine der folgenden Antworten:

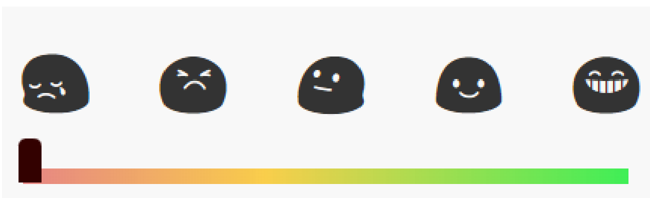
- ○ ja -> *Filter*
- ○ nein
- ○ ich weiß nicht

Filter: Wie lautet der Name des FDMS?

____

Antworten von weniger als 10 Personen werden in der Bachelorarbeit anonymisiert um Rückschlüsse auf einzelne Personen zu verhindern.

Filter: Erfüllt das FDMS, das Sie derzeit benutzen, Ihre Anforderungen?



Question type: 5 point choice. Slider layout: Slider with emoticon. Created with `https://www.limesurvey.org` *(LimeSurvey GmbH)*

Filter: Gibt es etwas, das Ihnen an Ihrem aktuellen FDMS besonders gefällt?

____

Filter: Gibt es etwas, das Ihnen an Ihrem aktuellen FDMS besonders misfällt?

____

Filter: Für welche Phasen des Forschungsdatenzyklus verwenden Sie ein FDMS?

Bitte wählen Sie die zutreffenden Antworten aus:

☐ Planung

☐ Datensammlung

☐ Datenverarbeitung und -analyse

☐ Datenarchivierung

☐ Wiederverwendung von Daten

Frage: Welche drei Softwareanwendungen sind für Ihre Forschung am wichtigsten bzw. verwenden Sie in diesem Zusammenhang am häufigsten (z.B. Excel, SPSS, Gocad, Zotero, spezifische Datenbanken, MATLAB, Eclipse, ArcGIS, OneNote etc.)?

____

____

____

Antworten werden in der Bachelorarbeit anonymisiert, falls eine Software von weniger als 10 Personen genutzt wird.

Frage: Verwahren Sie Forschungsdaten auch analog (nicht digitalisiert), wie zum Beispiel biologische Proben, handschriftliche Notizen etc.?

Question type: Yes/No. Display type: Button group. Created with `https://www.limesurvey.org` *(LimeSurvey GmbH)*

Filter (nur, wenn Frage nach analoger Verwahrung mit "Ja"beantwortet wurde): Welche Art von Forschungsdaten?

____

Filter (nur, wenn Frage nach analoger Verwahrung mit "Ja"beantwortet wurde): Warum haben Sie diese Daten nicht digitalisiert?

____

Frage: Wie groß ist eine typische digitale Ausgabe (z.B. Datenbanken, Excel-Tabellen, Interview-Aufnahmen etc.) Ihrer Forschung maximal in MB?

In dieses Feld dürfen nur Zahlen eingegeben werden.

____

Nutzerbedürfnisse 1/4

In den folgenden vier Abschnitten geht es um Ihre Wünsche und Bedürfnisse an ein Forschungsdatenmanagementsystem (FDMS). Wir würden hier gerne erfahren, wie Ihr ideales FDMS aussieht.

Frage: Wie wichtig sind Ihnen die folgenden Kriterien eines Forschungsdatenmanagementsystem (FDMS)?[a] *(Auswahl zwingend erforderlich)*

	1 - überhaupt nicht wichtig	2	3	4	5 - sehr wichtig	keine Antwort
Vorhandensein bevorzugter Anwendungsart (Web Applikation, mobile Applikation, lokale Installation)	◯	◯	◯	◯	◯	◯
Mehrsprachige Oberfläche	◯	◯	◯	◯	◯	◯
Open Source	◯	◯	◯	◯	◯	◯
API (Schnittstelle zur Programmierung von Anwendungen) existiert	◯	◯	◯	◯	◯	◯
Integration anderer Tools	◯	◯	◯	◯	◯	◯
Supportmöglichkeiten (z.B. integriertes Handbuch, Hilfe über Chat, Support-Hotline etc.)	◯	◯	◯	◯	◯	◯
Kollaborative Umgebung für Teams/Gruppen mit einem Zugriffsrechtesystem	◯	◯	◯	◯	◯	◯
Dateiverschlüsselung	◯	◯	◯	◯	◯	◯
Freitextsuche	◯	◯	◯	◯	◯	◯
Suchfilter (Facettensuche)	◯	◯	◯	◯	◯	◯
Unterstützung flexibler Forschungsabläufe	◯	◯	◯	◯	◯	◯
Assistenten, die Sie durch Prozesse führen (z.B. ein Dialogfeld, das durch den Datenimport führt)	◯	◯	◯	◯	◯	◯

a All answer options had the same size in the original questionnaire.

Nutzerbedürfnisse 2/4

In den folgenden drei Abschnitten geht es um Ihre Wünsche und Bedürfnisse an ein Forschungsdatenmanagementsystem (FDMS). Wir würden hier gerne erfahren, wie Ihr ideales FDMS aussieht.

Frage: Wie wichtig sind Ihnen die folgenden Kriterien eines Forschungsdatenmanagementsystem (FDMS)?[a] *(Auswahl zwingend erforderlich)*

Begriffe, die mit einer Ziffer versehen sind, sind in der Hilfe unten kurz erklärt.

	1 - überhaupt nicht wichtig	2	3	4	5 - sehr wichtig	keine Antwort
Funktion zur Vorschau des Datensatzinhalts	◯	◯	◯	◯	◯	◯
Visualisierung der Datensätze	◯	◯	◯	◯	◯	◯
Funktionalität zur Anonymisierung von Forschungsdaten	◯	◯	◯	◯	◯	◯
Einhaltung von Datenschutzbestimmungen (DSGVO, NDA, Ethik etc.)	◯	◯	◯	◯	◯	◯
Kontrolle über Datenstandort/Speicherort (lokal/Cloud/spezielles Land)	◯	◯	◯	◯	◯	◯
Volle Kontrolle über eigene Daten (z.B. Löschung, Bearbeitung, Export etc. der Daten)	◯	◯	◯	◯	◯	◯
Änderungsverfolgung	◯	◯	◯	◯	◯	◯
Kontrolle über Veröffentlichungstermine der Daten	◯	◯	◯	◯	◯	◯
Kontrolle über Ablaufdatum der Daten	◯	◯	◯	◯	◯	◯
Datenexport in verschiedene Formate	◯	◯	◯	◯	◯	◯
Möglichkeit der Datenzitierung	◯	◯	◯	◯	◯	◯
Transformation (1) von Daten innerhalb des FDMS	◯	◯	◯	◯	◯	◯

(1) **Transformation** der Daten umfasst Konvertierung, Reinigung und Strukturierung der Daten in ein für den weiteren Prozess nützliches Format (z.B. löschen von Duplikaten, Umwandlung von Werten, Umwandlung in andere Dateiformate etc.).

a All answer options had the same size in the original questionnaire.

Nutzerbedürfnisse 3/4

In den folgenden zwei Abschnitten geht es um Ihre Wünsche und Bedürfnisse an ein Forschungsdatenmanagementsystem (FDMS). Wir würden hier gerne erfahren, wie Ihr ideales FDMS aussieht.

Frage: Wie wichtig sind Ihnen die folgenden Kriterien eines Forschungsdatenmanagementsystem (FDMS)?[a] *(Auswahl zwingend erforderlich)*

Begriffe, die mit einer Ziffer versehen sind, sind in der Hilfe unten kurz erklärt.

	1 - überhaupt nicht wichtig	2	3	4	5 - sehr wichtig	keine Antwort
Nachverfolgung von Interaktionen externer Entitäten mit dem Datensatz (z.B. Zitierungen)	○	○	○	○	○	○
Datenherkunft (Wie wurden die Daten erstellt/geändert/ verwendet?)	○	○	○	○	○	○
Eindeutige Identifikatoren für Daten	○	○	○	○	○	○
Logische Verbindungen zwischen zusammenhängenden Daten zur Verbesserung der Auffindbarkeit	○	○	○	○	○	○
Obligatorische und optionale Metadaten-Elemente	○	○	○	○	○	○
Automatisierte Vergabe von Metadaten	○	○	○	○	○	○
Möglichkeit mehrere Metadaten-Schemata einzurichten	○	○	○	○	○	○
Domänenspezifische Metadaten-Schemata (1) verfügbar	○	○	○	○	○	○
Datenvalidierung (Erzwingen von korrekter Daten-/Metadatenstruktur (2), Dateiformat, etc.)	○	○	○	○	○	○
Datenidentifizierung mittels DOI (3) oder URL	○	○	○	○	○	○
Einhaltung des OAI-PMH (4)	○	○	○	○	○	○

(1) **Domänenspezifische Metadaten-Schemata** sind spezielle Schemata für einen wissenschaftlichen Bereich (z.B. Darwin Core für Naturgeschichte, Text Encoding Initiative für Geisteswissenschaften etc.).

(2) **Metadaten** sind Daten über Daten (z.B. Autor, Veröffentlichkeitsjahr, Version etc.) und können in beschreibende, technische, Aufbewahrungs- Rechte- und strukturelle Metadaten und Auszeichnungssprachen unterteilt werden.

(3) **DOI** steht für Digital Object Identifier und ist eine eindeutige Nummer. Damit können Objekte (z.B. Online-Publikationen) dauerhaft identifiziert und zuverlässig auf sie zugegriffen werden.

(4) **OAI-PMH** ist das Open Archives Initiative Protocol for Metadata Harvesting und dient dem Einsammeln und Weiterverarbeiten von Metadaten zur besseren Auffindbar- und Nutzbarkeit von elektronischen Publikationen im Internet.

[a] All answer options had the same size in the original questionnaire.

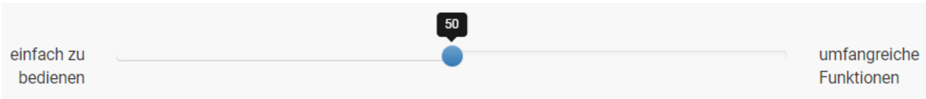
Nutzerbedürfnisse 4/4

Im folgenden Abschnitt geht es um Ihre Wünsche und Bedürfnisse an ein Forschungsdatenmanagementsystem (FDMS). Wir würden hier gerne erfahren, wie Ihr ideales FDMS aussieht.

Frage: Gibt es noch etwas, das Sie sich von einem FDMS wünschen? Bitte nennen Sie alle Funktionen, die Ihnen in den Sinn kommen.

____

Frage: Bevorzugen Sie ein einfach zu bedienendes FDMS oder lieber umfangreiche Funktionen?



Question type: Multiple numerical input. Use slider layout: On. Created with `https://www.limesurvey.org` *(LimeSurvey GmbH)*

Als einfach zu bedienen gilt für uns, wenn Sie das FDMS ohne umfangreiche Schulung oder Studium eines Nutzerhandbuchs verwenden können.

Frage: Bevorzugen Sie für ein FDMS die Benutzung über den Internet Browser (Web Applikation), über Handy/Tablet (mobile Applikation), über den Computer (lokale Installation) oder eine Kombination davon?

Bitte wählen Sie die zutreffenden Antworten aus:

☐ Web Applikation

☐ Mobile Applikation

☐ Lokale Installation

Frage: Falls Sie möchten, dass bestimmte Tools in das FDMS integriert werden, welche wären dies?

Bitte wählen Sie die zutreffenden Antworten aus:

☐ Die drei zuvor genannten Softwareanwendungen

☐ Sonstiges: ________________________

Frage: Für welche Phasen des Forschungsdatenzyklus würden Sie ein FDMS gerne verwenden?

Bitte wählen Sie die zutreffenden Antworten aus:

- ☐ Planung
- ☐ Datensammlung
- ☐ Datenverarbeitung und -analyse
- ☐ Datenarchivierung
- ☐ Wiederverwendung von Daten

Demografische Daten

Zum Abschluss möchten wir noch einige statistische Daten erheben.

Frage: Name der Institution

____

Antworten werden in der Thesis anonymisiert um Rückschlüsse auf einzelne Personen zu verhindern.

Frage: Land der Institution

Bitte wählen Sie eine der folgenden Antworten:

Auswahlliste mit allen Ländernamen lt. Tabellen 6.2 bis 6.3

Frage: Welche Position(en) haben Sie derzeit inne? *(Auswahl zwingend erforderlich)*

Bitte wählen Sie die zutreffenden Antworten aus:

- ☐ Professor(in)
- ☐ Postdoc
- ☐ Wissenschaftliche(r) Mitarbeiter(in)
- ☐ Datenkurator(in)
- ☐ Data scientist
- ☐ Fördergeber(in)
- ☐ Sonstiges: ________________________

Frage: Welcher Art von Institution gehören Sie an?
Bitte wählen Sie eine der folgenden Antworten:

- ○ Universität
- ○ Nicht-universitäre Forschungseinrichtung
- ○ Sonstiges: ____________________

Frage: In welchem wissenschaftlichen Bereich sind Sie hauptsächlich tätig? *(Auswahl zwingend erforderlich)*
Bitte wählen Sie eine der folgenden Antworten:

Auswahlliste mit wissenschaftlichen Bereichen lt. Tabelle 6.1

Frage: Welcher Altersgruppe gehören Sie an?
Bitte wählen Sie eine der folgenden Antworten:

- ○ $\leq$ 29
- ○ 30-39
- ○ 40-49
- ○ 50-59
- ○ $\geq$ 60

Frage: Welchem Geschlecht fühlen Sie sich zugehörig?
Bitte wählen Sie eine der folgenden Antworten:

- ○ männlich
- ○ weiblich
- ○ divers

Vielen herzlichen Dank, dass Sie geholfen haben, ein neues FDMS mitzugestalten. Wir wissen Ihren Beitrag sehr zu schätzen.

Falls Sie die daraus resultierende Bachelorarbeit erhalten möchten, haben Sie die Möglichkeit, hier Ihre E-Mail Adresse zu hinterlassen. Durch den separaten Link wird sichergestellt, dass diese nicht mit Ihren Antworten in Verbindung gebracht werden kann. Die Erhebung unterliegt den gleichen Datenschutzbestimmungen wie diese Umfrage.

6.5 POSITIONS

61 respondents entered a position they currently hold which was not available in the pre-defined list. Some inputs were made in addition to also selecting one of the pre-defined positions. Table 6.4 details how those other positions were grouped. Any positions that were ambiguous were grouped based on all of the data provided by the respondent.

Position entered by respondent	Grouped to position
PhD candidate PhD Student Prae Doc	PhD student
Data Analytics Student Ex student Master student Master student in the master's thesis phase Student Student intern	Student
assistant professor Educator learner Private lecturer	Professor
Research Professor	Professor + Researcher
freelancer Principal Scientist Research Associate Research Engineer Research Scientist Researcher Scientific Associate Senior Researcher Senior Scientist Staff Scientist	Researcher
Research Assistant (original: "studentische Hilfskraft")	Research assistant + Student
Technician	Research assistant
Data Manager	Data curator
Assistant Secretary General (Libraries, Open Science) Data Steward Head of scientific documentation RDM support Research support	Data steward
Advisor (same pay scale group as PostDoc) (original: "Referentin (diesselbe Tarifgruppe wie PostDoc)")	Postdoc
Job seeker	— (respondent also selected position "Professor")
Scholarship holder/Fellow Scientific online tutor (original: "Wissenschaftlicher Online Tutor")	— (respondent also selected position "Research assistant")
Clinician director Health management Humanitarian worker manager Program Coordinator Project coordinator Project Manager Quality and Environmental Management Upper management	Manager

Table 6.4: Respondent's Position

6.6 analog storage of research data

Tables 6.5 through 6.8 contain a full list of the types of research data stored in analog form, along with the reason for not having it digitalized yet. Where needed, the input was translated or anonymized. This list completes the summary in Section 3.4.2.

6.7 user need criteria lists

Tables 6.9 through 6.14 contain the detailed ratings of 35 criteria for an RDMS, pertaining to the summary in Section 3.4.4. The data is grouped by major scientific field.

Type of Research Data	Reason for not Digitalized
“Recipes” for reactions	This is how it is handled within the company
All sorts of things	No time, too much
archaeological finds, modern reference material	Physical objects/collections cannot be reasonably digitized.
Archaeometric samples, Drill cores, Thin sections, Excavation diaries	We are currently in the process of digitizing the data and making it searchable
Archival material	are digitized or cataloged
Assessment data, data collected in the stable (climate measurements, etc.)	Data is digitized (transferred to Excel), but also stored in its original form for safety reasons.
biological and other material samples	In the absence of digital twins, down to the nanoscale material level; because samples that have already been measured may need to be re-evaluated later
Biological samples	Not feasible to digitize biological samples
Biological samples in refrigerators and nitrogen storage, Laboratory books in analog form	It is possible to scan / digitize laboratory books after completion.
biological samples in various equipment; laboratory journal	According to the supervisor, you must always keep a laboratory journal when doing laboratory work and write in it with a non-erasable pen so that data is not falsified. I wanted to digitize mine, but I wasn’t allowed to because then the data would have been easier to manipulate.
Biological samples, Geological samples	Because the sample analysis (especially for geological samples) is not always consolidated at the time of collection. Sometimes questions develop later and rely on archive samples
biological samples, handwritten protocols from field experiments	The data already collected from the biological samples and the handwritten protocols have been digitized, although with the biological samples there is the possibility that further data can be collected using new methods and with the protocols there is the possibility of later analyzing larger connections.
biological samples, handwritten protocols from field experiments	The data already collected from the processing of the samples has been digitized but continues to be stored because new methods regularly enable the collection of further data; the handwritten protocols have also been digitized but are still stored in order to be able to understand certain connections later.
Biological samples. Still have some paper forms, although these all are scanned and only kept for compliance with ethics at the time. Moving away completely from paper	Can’t digitise all of the biological samples
Biological samples: histology slides, lab looks	No facilities to do so. Lab books are commonly hand written
Cell samples	Because I want to analyse them later
Clinical	No time, no system
Collection data, population and community structure data	We permanently do not use those data for publication
Complete printing of the database is carried out annually. This is a requirement from the authorities.	Is digitized.
Consent forms	They have manual signatures. Sometimes we scan these but still keep hard copies in a locked cabinet.
Conversation notes	Scanning is time-consuming and unnecessary
Copies	Effort
Data on vegetation composition, biodiversity	All data is digitized, with several backups, but the most important ones are also archived analog
DNA, strains, microscopy slides	Backup plus some things are not recoverable when digitized (living stuff)
e.g. Arabic, Greek	because, e.g. WORD is constantly “correcting” and interfering with the key assignments
Excerpts, transcripts	The relevant sources were examined in the archive and photographed accordingly and first recorded by hand. The transfer to digital usually only takes place in the second step.
Experiment analyses and some thoughts arguing them.	In the beginning it was easier to have a pen and paper on hand to put ideasa down faster.

Table 6.5: Research Data Stored in Analog Form (1/4)

Type of Research Data	Reason for not Digitalized
Experimental measurement, visual observations, processed data (hand calculation)	Flexibility of format, cost of software, reliance on third party=software provider for code maintenance, confidentiality of data storage. Also, lab notebooks (at the opposite of computerized documents) have a legal value when assessing things like prior art in patenting (hand written notes with signature, no back dating for example). You might also find useful insight by reading about the attempt of biotech industry to digitalize batch records in GMP manufacturing; there are significant obstacles that might be relevant for RDMS development.
Field notes, questionnaires	We keep both
field results from measuring forest trees	it is the next step
Hand written notes about experiments. Mostly thought processes, things that are not important if they are lost.	I have the habit of writing to help thinking, and it is difficult to not do it!
hand written notes during data collection with human participants in psychiatry	I have tried but it was not convenient because the typing makes too much noise during the tasks, whereas writing with a pen is completely silent. I could then digitalize the notes but as for now I am the only one who use them, the gain is not worth the pain.
handwritten documentation	Archive material is digitized, but analog material is also used for evaluation
Handwritten field notes/diaries etc	Sometimes in the field where I have conducted research it is not feasible to use a digital platform for data entry (e.g. security, lack of internet connectivity, no electricity, very unsual/indiscreet in the setting for a researcher to use an electronic device etc)
Handwritten notes	Notes are partly scanned, but not yet recorded digitally. Expenditure of time
handwritten notes	are only needed for a short time
Handwritten notes	Difficulty of using the platform. It is quicker to take handwritten notes
handwritten notes	I prefer handwritten notes, it helps me follow my own train of thought.
handwritten notes	it was backup data
handwritten notes	Time commitment is not always affordable
handwritten notes	unnecessary
handwritten notes	Would cost a lot of time with unclear benefits.
handwritten notes	
Handwritten notes and field notes	I digitalised only the important ones
Handwritten notes and manuals	Some data are even better read and interpreted with paper.
Handwritten notes for brainstorming ideas, drawing diagrams of physical systems or visualising code.	They are largely uninteresting (and perhaps even incoherent) to anyone else. Some important ideas or diagrams will be digitised if it is useful in a publication
handwritten notes such as changes to protocols, some research findings following assays and other techniques. These will also be captured in digital format.	All research data is digitalised but often some aspects of metadata is captured in written notes. One reason why all data is not digitalised is that there is no one standard protocol or SOP for groups. This is a key objective going forward .
Handwritten notes, biological samples	biological samples cannot be digitalized. I did not digitalize handwritten data because of lack of time.
Handwritten notes, laboratory notebook, frozen biological tissues	As a backup, for easier tracking
Handwritten notes, like notes and diary	Because in my opinion everything which is handwritting has a different value than digitalized
Handwritten survey collected from farmers	I created a database based on this data but continue to keep the original answers
Handwritten voice data	They are constantly being digitized.
Ideas and drawings	No need to do so.
Immunostained sections, labbooks	It is partially digitalized. But the physical proof, for example the sections that were imaged, has to be stored at least for 10 years, after the data is published.
Lab notes	Some people still find it more convenient

Table 6.6: Research Data Stored in Analog Form (2/4)

Type of Research Data	Reason for not Digitalized
Laboratory notebook with handwritten notes	I digitize the data from the laboratory notebook regularly. What is not digitized are entries such as: what was done on which day, which procedure was successful. The biological samples are also preserved.
Laboratory protocols	On paper, a laboratory protocol is easier to keep and use in the laboratory. There are digital laboratory protocol options, but they are rather energy-consuming
Measurement protocols	Data is subsequently digitized, but only recorded by hand when recorded and measured.
Medical research	No ideas
microbiome (stool, nasal washings) , blood, serum, cells, lung function tests, skin tests, imaging	
Mostly digitised, but some disciplines (e.g., archaeology) may have physical samples without full associated meta-data	as above
Mostly long duration storage samples for stability studies under controlled environmental conditions. Also some master batch records.	Physical samples used in stability studies required by regulatory agencies cannot be digitalized. The master batch records could be digitalized, but would require a consistent electronic signature system first (we started with DocuSign in some cases, but not for all of them).
MRI images related to clinical examination results	Clinical examination results are collected chairside on the patient, therefore initially in an analog manner
notes	Has no discernible advantage
notes	I have both but I ork better on paper so it's easier for me
notes for design methodology and analysis, journaling and notes related to results	easiness of handwriting (less time consumption), flexibility of updates (most of the time). lack of support in the digitalized tools (as per my knowledge)
Notes from field work	I digitize the handwritten notes shortly after the field work (usually on the same day) in the form of Excel spreadsheets. However, it is often difficult to carry a laptop or tablet with you in the field itself (due to weather, etc.)
Notes, excerpts, annotations in books	Tied to medium (e.g. in book).
notes, scans, digital images, biological samples	you cannot digitize biological samples. We only use these as back ups - everything has been digitized.
Notes, scripts, recordings	Too time-consuming
Notes, sketches, reflections	Long live paper! The clean copy is then texted.
Object data	Lack of staff
Observations about sample collection and also the log of what happened to the samples. Basically, I wrote down what happened in the lab to what samples.	Actual numbers are usually digitized, but it would take too much time to digitize the lab notebook after the fact and there is no support for a digital lab notebook.
Personal notes	I do this for notes that I find most relevant just to me (e.g. reminders)
Personal notes, research diary	Practicality in everyday work
Pictures.laboratory notebooks	I do not know where to store
Printouts, books, handwritten notes	Books in the library are partly analog and partly digital; handwritten notes: drafts only; printouts: only in exceptional cases
Procedures, evaluations, observations, preliminary results	I'm not required to and it works fine
Protocols for laboratory analyses, reserve samples	Reserve samples: It is possible to determine other parameters that were not in the original plan, Protocols laboratory analysis: verification and traceability
qPCR, Western Blot, IHC, etc	Paper-based record is easier for me to organize.
Questionaires	I am not sure
Questionary	It is preferable to have hard copies

Table 6.7: Research Data Stored in Analog Form (3/4)

Type of Research Data	Reason for not Digitalized
Questionnaires for face-to-face participation	Data has been digitized, serves as a backup
Questionnaires, interview protocols	Effort, lack of storage options
Quotes, source work	too much effort
research data we mean all data which was collected for scientific purposes	Busy with studies
Saliva samples, handwritten notes (score sheets)	are additionally digitized, but are stored as primary data because there is currently no legally admissable database (animal testing/score sheet)
Signed data protection declarations, hand-filled questionnaires	Transfer too time consuming, will be stored in an archive
simple experiment descriptions in lab book, also technical numbers for experiments which do not require more processing, mostly ""overview"" data	a physical lab book is quite convenient for day to day work
social science and biological data	The data is also digitized, storage in analog form is only for safety purposes, storage of samples for a certain period of time is mandatory
Sociodemographic variables - quantitative and qualitative. Observations in interviews	Not necessary
source documents for clinical trials	extra work would be involved, and cost
Specimens of insects	Insect specimens are required for verification or comparison of new catches
Temperature, pressure and other conditions of chemical processes. Process recipes, nanofarication protocols.	Ease of use, cleanliness (notebooks in cleanrooms and laboratories carry less dust than computers, as are resistant to chemical damage). Further, only parts of the recipes are relevant for reproducibility and publications. It would be good to also digitalize them, it's just not part of the working culture in my institute.
The University of [...] covers all topics of sciences so researchers store research data in analog form from different scientific fields	Do not have exact explanation from researchers of the University of [...], I expect that analog form is still happening in some scientific fields
Time points, deviations from plans, protocols, weighted samples, date and time of operations	It's a lot of data - especially milligram weights with about 6 digits it's just annoying to enter and error-prone. At best I would make a picture of it but that's very data-inefficient (several MB for something that should take a kb) and I don't like to use my phone on the lab (where the notebook has to be).
Tissue samples, rock samples, biological specimens	na
Typically, insect samples	
Unfortunately, too many, e.g. deposition protocols from sputtering systems. But we are currently implementing a RDMS.	Because my predecessor wanted to sit on "her" data and thought she could get ahead without any digitization (device was purchased 3 years ago). I would have liked at least an Excel spreadsheet. I'm currently working with my bachelor's student to find out which data actually needs to be written down because the machine writes protocols (which are not readable for "classic" experimentalists without programming knowledge)
various things, genetic samples, protein samples, tissues,	What we can digitalize is in the process of digitalizing.
Voice/ text / visual	Legacy systems
	images from microscopes

Table 6.8: Research Data Stored in Analog Form (4/4)

	#	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}	Q_{12}	Q_{13}	Q_{14}	Q_{15}	Q_{16}	Q_{17}	Q_{18}	Q_{19}	Q_{20}
Total	207	3.9	2.6	3.6	3.1	3.5	3.6	3.8	3.6	4.0	4.1	3.5	3.2	3.6	3.9	3.4	3.9	3.8	4.5	4.2	3.7
Agricultural Sciences	20	3.8	2.5	3.7	3.4	3.9	3.9	4.0	3.6	4.4	4.1	3.9	3.7	3.8	4.1	3.4	3.6	4.3	4.9	4.6	4.1
Agricultural biotechnology	2	4.5	3.0	5.0	4.0	5.0	5.0	4.0	3.0	5.0	2.5	4.0	3.5	4.0	4.5	3.0	3.5	4.0	5.0	4.5	3.5
Agriculture, forestry, and fisheries	7	3.4	2.9	3.7	4.1	4.3	4.1	3.7	3.4	4.0	3.9	3.9	4.0	4.1	4.3	3.1	2.7	3.6	4.9	4.4	3.7
Animal and dairy science	4	3.0	2.8	2.3	3.3	2.3	2.8	3.8	3.3	5.0	5.0	4.8	3.5	3.5	4.0	4.0	4.8	5.0	4.8	4.5	4.8
Veterinary science	3	4.7	1.3	4.0	1.7	4.7	3.7	5.0	4.7	4.3	4.3	4.7	3.3	4.0	4.0	3.7	3.7	4.7	5.0	5.0	4.7
Other agricultural sciences	4	4.3	2.3	4.0	3.0	3.8	4.3	4.0	3.8	4.0	4.0	2.3	3.8	3.3	3.5	3.0	4.0	4.5	5.0	4.5	4.0

	#	Q_{21}	Q_{22}	Q_{23}	Q_{24}	Q_{25}	Q_{26}	Q_{27}	Q_{28}	Q_{29}	Q_{30}	Q_{31}	Q_{32}	Q_{33}	Q_{34}	Q_{35}
Total	207	3.6	4.1	3.7	3.3	3.2	3.7	4.1	3.7	3.3	2.9	2.7	2.5	3.2	3.5	2.3
Agricultural Sciences	20	4.0	4.6	3.7	3.5	3.5	3.8	3.9	4.0	3.3	3.0	2.8	2.9	3.1	3.5	2.1
Agricultural biotechnology	2	3.5	5.0	4.0	3.5	4.5	5.0	3.5	4.0	4.5	4.5	4.0	5.0	4.5	4.0	2.5
Agriculture, forestry, and fisheries	7	3.7	4.7	2.9	3.7	2.9	3.3	3.7	3.3	2.7	3.0	3.0	3.0	3.4	2.9	2.9
Animal and dairy science	4	4.5	4.5	4.5	4.5	3.3	4.3	3.5	5.0	4.5	3.0	4.0	4.0	4.0	4.3	2.5
Veterinary science	3	3.7	5.0	3.7	1.7	4.3	4.0	5.0	4.7	1.3	1.3	0.3	0.3	0.0	2.7	0.0
Other agricultural sciences	4	4.3	4.0	4.3	3.5	3.5	3.3	4.0	3.8	3.8	3.5	2.3	2.3	3.3	4.0	1.5

1 = not at all important / 5 = very important / 0 = no answer

Q_1: Availability of preferred usage method
Q_2: Multilingual interface
Q_3: Open source
Q_4: Existence of an API
Q_5: Integration of other tools
Q_6: Support options
Q_7: Collaborative environment for teams/groups with access rights system
Q_8: File encryption
Q_9: Free-text search capability
Q_{10}: Search filters (faceted search)
Q_{11}: Support for flexible research workflows
Q_{12}: Wizards to guide through processes
Q_{13}: Feature for previewing dataset content
Q_{14}: Visualization of datasets
Q_{15}: Functionality to anonymize research data
Q_{16}: Compliance with data regulations
Q_{17}: Control over data location
Q_{18}: Full control over own data
Q_{19}: Tracking of content changes
Q_{20}: Control over data release dates
Q_{21}: Control over data expiry dates
Q_{22}: Data export in different formats
Q_{23}: Possibility of data citation
Q_{24}: Transformation of data within the RDMS
Q_{25}: Tracking of interactions of external entities with dataset
Q_{26}: Data provenance
Q_{27}: Unique identifiers for data
Q_{28}: Logical links between related data
Q_{29}: Mandatory/optional metadata elements
Q_{30}: Automated metadata tagging
Q_{31}: Possibility to set up multiple metadata schemata
Q_{32}: Domain-specific metadata schemata
Q_{33}: Data validation
Q_{34}: Data identification via DOI/URL
Q_{35}: Compliance with OAI-PMH

Table 6.9: Importance of Criteria for an RDMS - Agricultural Sciences

	#	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}	Q_{12}	Q_{13}	Q_{14}	Q_{15}	Q_{16}	Q_{17}	Q_{18}	Q_{19}	Q_{20}
Total	207	3.9	2.6	3.6	3.1	3.5	3.6	3.8	3.6	4.0	4.1	3.5	3.2	3.6	3.9	3.4	3.9	3.8	4.5	4.2	3.7
Engineering and Technology	30	3.8	3.1	3.5	3.8	3.7	3.6	3.9	3.8	3.8	4.2	3.7	3.2	3.7	4.0	3.2	3.8	3.7	4.5	4.2	3.9
Chemical engineering	1	4.0	2.0	4.0	3.0	3.0	4.0	5.0	4.0	5.0	5.0	3.0	4.0	4.0	4.0	5.0	5.0	5.0	5.0	4.0	4.0
Electrical/electronic/information engineering	8	4.1	3.8	3.5	4.1	4.1	3.6	4.1	3.4	3.5	4.3	3.5	3.4	4.3	4.0	3.5	4.3	4.3	4.5	4.5	4.3
Industrial Biotechnology	1	2.0	1.0	3.0	3.0	2.0	3.0	2.0	2.0	5.0	4.0	5.0	1.0	5.0	5.0	1.0	0.0	3.0	5.0	4.0	3.0
Materials engineering	3	4.0	3.3	5.0	5.0	4.0	3.7	3.3	3.3	3.7	4.3	4.7	2.7	2.3	3.3	1.7	3.0	2.3	4.7	5.0	4.3
Mechanical engineering	3	2.7	4.0	3.0	3.3	2.3	2.3	3.0	2.3	4.0	3.7	2.7	1.7	3.0	4.0	3.0	3.3	2.7	3.0	2.3	3.7
Medical engineering	1	3.0	4.0	2.0	4.0	4.0	4.0	5.0	4.0	3.0	5.0	3.0	5.0	4.0	5.0	5.0	5.0	3.0	4.0	2.0	2.0
Nano-technology	1	3.0	3.0	4.0	5.0	4.0	4.0	4.0	4.0	4.0	4.0	3.0	3.0	2.0	2.0	0.0	0.0	3.0	5.0	4.0	4.0
Other engineering and technologies	12	4.0	2.7	3.3	3.5	3.8	3.8	4.0	4.6	3.8	4.1	4.0	3.5	3.9	4.1	3.6	4.3	3.9	4.8	4.5	3.8

	#	Q_{21}	Q_{22}	Q_{23}	Q_{24}	Q_{25}	Q_{26}	Q_{27}	Q_{28}	Q_{29}	Q_{30}	Q_{31}	Q_{32}	Q_{33}	Q_{34}	Q_{35}
Total	207	3.6	4.1	3.7	3.3	3.2	3.7	4.1	3.7	3.3	2.9	2.7	2.5	3.2	3.5	2.3
Engineering and Technology	30	3.6	3.8	3.7	3.4	3.5	3.8	4.4	4.0	3.6	3.4	3.4	3.3	3.8	3.7	2.9
Chemical engineering	1	5.0	5.0	3.0	4.0	4.0	3.0	4.0	3.0	3.0	4.0	3.0	3.0	3.0	4.0	4.0
Electrical/electronic/information engineering	8	4.1	4.0	3.6	4.1	3.9	3.5	4.5	4.6	4.3	4.0	4.0	3.8	4.3	3.5	3.4
Industrial Biotechnology	1	1.0	4.0	5.0	2.0	3.0	4.0	4.0	5.0	4.0	5.0	4.0	5.0	4.0	5.0	0.0
Materials engineering	3	3.0	2.7	4.3	2.7	4.0	4.3	5.0	4.3	3.7	2.7	2.7	2.7	4.7	4.7	3.3
Mechanical engineering	3	3.3	3.3	3.0	2.7	3.3	3.0	3.7	3.3	3.0	3.0	3.0	2.3	2.3	3.3	3.0
Medical engineering	1	3.0	5.0	3.0	3.0	4.0	4.0	5.0	3.0	3.0	3.0	3.0	3.0	4.0	4.0	0.0
Nano-technology	1	4.0	4.0	4.0	3.0	4.0	4.0	4.0	3.0	3.0	3.0	3.0	4.0	4.0	4.0	3.0
Other engineering and technologies	12	3.5	3.8	3.8	3.3	3.2	4.1	4.3	3.8	3.3	3.2	3.3	3.2	3.8	3.6	2.9

1 = not at all important / 5 = very important / 0 = no answer

Q_1: Availability of preferred usage method
Q_2: Multilingual interface
Q_3: Open source
Q_4: Existence of an API
Q_5: Integration of other tools
Q_6: Support options
Q_7: Collaborative environment for teams/groups with access rights system
Q_8: File encryption
Q_9: Free-text search capability
Q_{10}: Search filters (faceted search)
Q_{11}: Support for flexible research workflows
Q_{12}: Wizards to guide through processes
Q_{13}: Feature for previewing dataset content
Q_{14}: Visualization of datasets
Q_{15}: Functionality to anonymize research data
Q_{16}: Compliance with data regulations
Q_{17}: Control over data location
Q_{18}: Full control over own data
Q_{19}: Tracking of content changes
Q_{20}: Control over data release dates
Q_{21}: Control over data expiry dates
Q_{22}: Data export in different formats
Q_{23}: Possibility of data citation
Q_{24}: Transformation of data within the RDMS
Q_{25}: Tracking of interactions of external entities with dataset
Q_{26}: Data provenance
Q_{27}: Unique identifiers for data
Q_{28}: Logical links between related data
Q_{29}: Mandatory/optional metadata elements
Q_{30}: Automated metadata tagging
Q_{31}: Possibility to set up multiple metadata schemata
Q_{32}: Domain-specific metadata schemata
Q_{33}: Data validation
Q_{34}: Data identification via DOI/URL
Q_{35}: Compliance with OAI-PMH

Table 6.10: Importance of Criteria for an RDMS - Engineering and Technology

	#	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}	Q_{12}	Q_{13}	Q_{14}	Q_{15}	Q_{16}	Q_{17}	Q_{18}	Q_{19}	Q_{20}
Total	207	3.9	2.6	3.6	3.1	3.5	3.6	3.8	3.6	4.0	4.1	3.5	3.2	3.6	3.9	3.4	3.9	3.8	4.5	4.2	3.7
Humanities	19	4.3	2.6	3.5	3.3	3.5	3.5	3.9	3.5	4.6	4.4	2.3	3.2	3.2	3.8	3.2	3.4	3.8	4.4	4.2	3.9
History and archaeology	10	4.1	3.2	4.0	3.3	3.4	3.5	3.9	3.5	4.7	4.4	2.4	3.6	3.0	4.0	3.1	3.6	4.1	4.5	4.4	4.3
Languages and literature	3	4.0	2.3	2.0	3.7	4.0	2.3	3.0	2.7	4.0	3.7	1.7	1.7	4.0	4.3	3.0	1.3	4.0	5.0	4.3	4.3
Philosophy, ethics and religion	5	4.6	1.6	4.0	2.8	4.0	3.8	4.2	4.2	4.8	4.6	3.0	2.8	2.8	3.0	3.0	3.8	3.4	3.8	4.0	3.6
Other humanities	1	5.0	3.0	0.0	4.0	0.0	5.0	5.0	2.0	5.0	5.0	0.0	5.0	5.0	5.0	5.0	5.0	3.0	5.0	3.0	0.0

	#	Q_{21}	Q_{22}	Q_{23}	Q_{24}	Q_{25}	Q_{26}	Q_{27}	Q_{28}	Q_{29}	Q_{30}	Q_{31}	Q_{32}	Q_{33}	Q_{34}	Q_{35}
Total	207	3.6	4.1	3.7	3.3	3.2	3.7	4.1	3.7	3.3	2.9	2.7	2.5	3.2	3.5	2.3
Humanities	19	3.4	4.1	4.1	3.4	2.4	3.4	3.7	3.7	3.3	2.5	2.3	2.2	2.9	3.8	2.5
History and archaeology	10	4.0	4.2	4.7	3.7	3.0	4.0	4.4	4.2	3.4	2.8	2.8	2.7	3.9	4.0	2.9
Languages and literature	3	4.0	4.3	3.7	2.3	1.7	2.3	3.3	4.0	3.0	3.0	2.7	2.7	3.3	4.3	3.3
Philosophy, ethics and religion	5	2.6	3.4	3.6	3.4	2.0	2.8	2.4	3.4	3.2	2.2	1.4	1.4	1.4	3.2	1.6
Other humanities	1	0.0	5.0	2.0	3.0	0.0	3.0	5.0	0.0	4.0	0.0	0.0	0.0	0.0	4.0	0.0

1 = not at all important / 5 = very important / 0 = no answer

Q_1: Availability of preferred usage method
Q_2: Multilingual interface
Q_3: Open source
Q_4: Existence of an API
Q_5: Integration of other tools
Q_6: Support options
Q_7: Collaborative environment for teams/groups with access rights system
Q_8: File encryption
Q_9: Free-text search capability
Q_{10}: Search filters (faceted search)
Q_{11}: Support for flexible research workflows
Q_{12}: Wizards to guide through processes
Q_{13}: Feature for previewing dataset content
Q_{14}: Visualization of datasets
Q_{15}: Functionality to anonymize research data
Q_{16}: Compliance with data regulations
Q_{17}: Control over data location
Q_{18}: Full control over own data
Q_{19}: Tracking of content changes
Q_{20}: Control over data release dates
Q_{21}: Control over data expiry dates
Q_{22}: Data export in different formats
Q_{23}: Possibility of data citation
Q_{24}: Transformation of data within the RDMS
Q_{25}: Tracking of interactions of external entities with dataset
Q_{26}: Data provenance
Q_{27}: Unique identifiers for data
Q_{28}: Logical links between related data
Q_{29}: Mandatory/optional metadata elements
Q_{30}: Automated metadata tagging
Q_{31}: Possibility to set up multiple metadata schemata
Q_{32}: Domain-specific metadata schemata
Q_{33}: Data validation
Q_{34}: Data identification via DOI/URL
Q_{35}: Compliance with OAI-PMH

Table 6.11: Importance of Criteria for an RDMS - Humanities

	#	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}	Q_{12}	Q_{13}	Q_{14}	Q_{15}	Q_{16}	Q_{17}	Q_{18}	Q_{19}	Q_{20}
Total	207	3.9	2.6	3.6	3.1	3.5	3.6	3.8	3.6	4.0	4.1	3.5	3.2	3.6	3.9	3.4	3.9	3.8	4.5	4.2	3.7
Medical and Health Sciences	22	4.4	2.6	3.4	3.6	3.8	4.0	4.2	4.3	4.3	4.4	4.0	3.5	3.6	4.1	4.0	4.5	4.2	4.5	4.3	4.0
Basic medicine	5	4.6	2.2	4.2	3.6	3.4	4.2	4.8	4.2	4.6	4.6	4.0	3.0	4.0	4.0	3.6	4.0	4.4	4.6	4.0	4.2
Clinical medicine	3	5.0	2.7	3.0	2.7	3.7	4.0	3.7	4.3	4.0	4.0	4.0	3.3	3.3	4.3	4.0	4.7	3.7	5.0	5.0	4.3
Health sciences	8	4.3	2.6	3.0	4.3	4.1	3.9	4.4	4.3	4.8	4.9	4.3	4.0	4.0	4.5	4.8	4.9	4.5	4.6	4.6	4.3
Other medical sciences	6	4.0	3.0	3.5	3.3	3.7	3.8	3.7	4.3	3.7	3.8	3.7	3.3	3.0	3.5	3.3	4.3	3.8	4.2	3.8	3.5

	#	Q_{21}	Q_{22}	Q_{23}	Q_{24}	Q_{25}	Q_{26}	Q_{27}	Q_{28}	Q_{29}	Q_{30}	Q_{31}	Q_{32}	Q_{33}	Q_{34}	Q_{35}
Total	207	3.6	4.1	3.7	3.3	3.2	3.7	4.1	3.7	3.3	2.9	2.7	2.5	3.2	3.5	2.3
Medical and Health Sciences	22	4.0	4.2	3.4	3.4	3.7	4.2	4.5	3.8	3.9	3.2	3.2	2.9	3.5	3.3	2.9
Basic medicine	5	3.6	3.6	3.2	2.4	3.6	4.0	4.2	3.6	4.0	4.2	4.0	3.8	3.6	3.2	3.0
Clinical medicine	3	4.3	5.0	3.7	4.3	4.0	4.7	5.0	4.3	4.3	2.7	3.3	2.7	4.0	4.3	3.7
Health sciences	8	4.3	4.4	3.8	4.1	3.5	4.4	4.6	4.4	4.5	3.1	2.9	2.5	4.1	3.6	3.4
Other medical sciences	6	3.8	4.0	3.0	2.8	3.8	4.0	4.2	3.0	2.8	2.7	2.8	2.7	2.5	2.3	1.8

1 = not at all important / 5 = very important / 0 = no answer

Q_1: Availability of preferred usage method
Q_2: Multilingual interface
Q_3: Open source
Q_4: Existence of an API
Q_5: Integration of other tools
Q_6: Support options
Q_7: Collaborative environment for teams/groups with access rights system
Q_8: File encryption
Q_9: Free-text search capability
Q_{10}: Search filters (faceted search)
Q_{11}: Support for flexible research workflows
Q_{12}: Wizards to guide through processes
Q_{13}: Feature for previewing dataset content
Q_{14}: Visualization of datasets
Q_{15}: Functionality to anonymize research data
Q_{16}: Compliance with data regulations
Q_{17}: Control over data location
Q_{18}: Full control over own data
Q_{19}: Tracking of content changes
Q_{20}: Control over data release dates
Q_{21}: Control over data expiry dates
Q_{22}: Data export in different formats
Q_{23}: Possibility of data citation
Q_{24}: Transformation of data within the RDMS
Q_{25}: Tracking of interactions of external entities with dataset
Q_{26}: Data provenance
Q_{27}: Unique identifiers for data
Q_{28}: Logical links between related data
Q_{29}: Mandatory/optional metadata elements
Q_{30}: Automated metadata tagging
Q_{31}: Possibility to set up multiple metadata schemata
Q_{32}: Domain-specific metadata schemata
Q_{33}: Data validation
Q_{34}: Data identification via DOI/URL
Q_{35}: Compliance with OAI-PMH

Table 6.12: Importance of Criteria for an RDMS - Medical and Health Sciences

	#	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}	Q_{12}	Q_{13}	Q_{14}	Q_{15}	Q_{16}	Q_{17}	Q_{18}	Q_{19}	Q_{20}
Total	207	3.9	2.6	3.6	3.1	3.5	3.6	3.8	3.6	4.0	4.1	3.5	3.2	3.6	3.9	3.4	3.9	3.8	4.5	4.2	3.7
Natural Sciences	60	4.2	2.5	3.9	3.2	3.7	3.6	4.0	3.4	4.0	4.2	3.7	3.2	3.6	3.8	2.9	3.6	3.8	4.5	4.2	3.6
Biological sciences	16	4.3	2.4	3.9	2.7	3.7	3.7	4.4	4.1	4.2	4.2	3.8	3.1	3.4	3.9	2.9	3.5	3.8	4.4	4.4	4.1
Chemical sciences	3	4.7	1.3	3.0	2.0	3.7	3.3	4.7	3.3	3.7	4.3	1.3	3.7	2.3	5.0	1.0	3.3	3.3	5.0	2.7	0.0
Computer and information sciences	22	3.8	2.5	3.9	4.0	3.9	3.3	3.9	3.4	4.0	4.0	3.9	3.3	3.6	3.5	3.6	4.1	4.2	4.5	4.2	3.6
Earth and related environmental sciences	4	4.5	4.5	4.3	2.3	4.0	4.5	4.5	4.3	4.8	4.5	4.5	3.3	4.5	4.8	4.0	4.8	4.0	5.0	5.0	4.3
Mathematics	2	4.5	1.5	2.5	2.0	4.0	3.5	4.5	2.0	4.5	4.0	3.5	2.5	2.5	3.5	1.5	2.0	3.0	4.5	4.5	4.0
Physical sciences	9	4.0	1.9	4.1	2.8	3.0	3.6	3.3	2.6	3.7	4.4	3.3	2.6	3.9	3.7	1.2	2.0	2.6	3.9	3.7	2.9
Other natural sciences	4	4.8	3.8	4.5	3.8	3.8	4.3	3.5	3.0	3.8	3.8	4.0	3.8	3.5	4.3	3.8	5.0	5.0	4.8	4.5	4.5

	#	Q_{21}	Q_{22}	Q_{23}	Q_{24}	Q_{25}	Q_{26}	Q_{27}	Q_{28}	Q_{29}	Q_{30}	Q_{31}	Q_{32}	Q_{33}	Q_{34}	Q_{35}
Total	207	3.6	4.1	3.7	3.3	3.2	3.7	4.1	3.7	3.3	2.9	2.7	2.5	3.2	3.5	2.3
Natural Sciences	60	3.6	4.2	3.8	3.4	3.2	4.0	4.2	3.7	3.3	3.1	2.8	2.6	3.4	3.6	2.2
Biological sciences	16	3.8	4.4	3.6	3.8	3.6	4.3	4.4	3.8	3.3	2.8	2.6	2.3	3.1	2.8	1.4
Chemical sciences	3	1.3	4.0	3.0	3.7	1.7	3.0	4.0	3.3	1.3	2.7	1.3	0.0	1.7	2.0	2.3
Computer and information sciences	22	3.6	3.9	4.0	2.6	3.1	3.9	4.2	3.6	3.5	3.6	3.3	3.0	3.6	3.9	2.6
Earth and related environmental sciences	4	4.5	5.0	5.0	4.0	4.0	4.5	4.5	4.3	4.5	3.0	1.8	2.8	4.0	5.0	2.5
Mathematics	2	3.5	4.0	3.5	3.5	2.5	2.5	4.0	4.5	3.0	2.5	3.0	2.5	3.5	4.0	2.0
Physical sciences	9	2.8	4.7	3.7	3.9	2.9	4.0	4.0	3.0	2.7	2.3	2.3	2.2	3.2	3.8	1.4
Other natural sciences	4	4.8	3.0	4.0	3.8	3.8	4.0	4.3	3.8	3.8	3.8	3.3	3.5	4.5	4.3	3.8

1 = not at all important / 5 = very important / 0 = no answer

Q_1: Availability of preferred usage method
Q_2: Multilingual interface
Q_3: Open source
Q_4: Existence of an API
Q_5: Integration of other tools
Q_6: Support options
Q_7: Collaborative environment for teams/groups with access rights system
Q_8: File encryption
Q_9: Free-text search capability
Q_{10}: Search filters (faceted search)
Q_{11}: Support for flexible research workflows
Q_{12}: Wizards to guide through processes
Q_{13}: Feature for previewing dataset content
Q_{14}: Visualization of datasets
Q_{15}: Functionality to anonymize research data
Q_{16}: Compliance with data regulations
Q_{17}: Control over data location
Q_{18}: Full control over own data
Q_{19}: Tracking of content changes
Q_{20}: Control over data release dates
Q_{21}: Control over data expiry dates
Q_{22}: Data export in different formats
Q_{23}: Possibility of data citation
Q_{24}: Transformation of data within the RDMS
Q_{25}: Tracking of interactions of external entities with dataset
Q_{26}: Data provenance
Q_{27}: Unique identifiers for data
Q_{28}: Logical links between related data
Q_{29}: Mandatory/optional metadata elements
Q_{30}: Automated metadata tagging
Q_{31}: Possibility to set up multiple metadata schemata
Q_{32}: Domain-specific metadata schemata
Q_{33}: Data validation
Q_{34}: Data identification via DOI/URL
Q_{35}: Compliance with OAI-PMH

Table 6.13: Importance of Criteria for an RDMS - Natural Sciences

	#	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}	Q_{12}	Q_{13}	Q_{14}	Q_{15}	Q_{16}	Q_{17}	Q_{18}	Q_{19}	Q_{20}
Total	207	3.9	2.6	3.6	3.1	3.5	3.6	3.8	3.6	4.0	4.1	3.5	3.2	3.6	3.9	3.4	3.9	3.8	4.5	4.2	3.7
Social Sciences	56	3.5	2.4	3.3	2.4	3.1	3.5	3.4	3.4	3.5	3.7	3.3	3.1	3.6	3.7	3.8	4.2	3.6	4.5	4.1	3.4
Economics and business	12	3.3	2.6	2.8	2.3	3.2	4.3	3.3	3.3	3.3	3.3	2.2	3.2	3.5	3.3	3.6	3.9	3.6	4.8	4.0	2.8
Educational sciences	2	4.0	1.5	2.0	2.5	2.0	2.0	2.0	4.5	4.5	4.0	2.0	1.5	3.0	2.5	4.5	5.0	5.0	5.0	4.5	4.5
Media and communications	5	4.4	2.6	3.4	2.4	3.0	3.8	4.2	3.2	4.4	4.4	4.4	3.0	3.6	4.0	2.8	3.8	3.6	4.8	3.8	4.0
Political Science	3	3.7	2.3	3.7	3.3	3.0	3.3	3.3	2.7	1.7	2.3	3.7	3.0	3.3	3.7	2.7	3.3	3.3	4.0	3.3	1.3
Psychology	23	3.6	2.7	4.0	2.8	3.4	3.6	3.7	3.9	3.7	3.8	3.7	3.4	3.7	3.7	4.5	4.6	3.9	4.6	4.3	4.0
Social and economic geography	2	3.0	1.0	1.0	1.0	2.0	2.5	2.0	1.0	2.5	2.5	2.5	3.0	4.5	4.5	2.0	3.0	3.0	4.5	4.5	2.0
Sociology	2	3.5	2.0	4.0	3.0	1.5	4.0	2.5	2.5	3.0	3.5	3.5	2.5	1.5	2.0	2.0	2.5	2.0	3.5	3.5	1.5
Other social sciences	7	2.7	1.9	2.7	1.4	2.7	2.4	3.6	3.0	3.4	4.3	3.9	2.4	4.1	4.7	4.0	4.9	3.4	4.0	4.1	3.6

	#	Q_{21}	Q_{22}	Q_{23}	Q_{24}	Q_{25}	Q_{26}	Q_{27}	Q_{28}	Q_{29}	Q_{30}	Q_{31}	Q_{32}	Q_{33}	Q_{34}	Q_{35}
Total	207	3.6	4.1	3.7	3.3	3.2	3.7	4.1	3.7	3.3	2.9	2.7	2.5	3.2	3.5	2.3
Social Sciences	56	3.3	4.0	3.4	3.1	2.9	3.4	3.9	3.3	2.9	2.5	2.4	1.8	2.7	3.2	1.8
Economics and business	12	2.4	4.2	3.3	3.0	2.5	3.5	4.2	2.8	2.2	1.9	1.8	1.1	2.3	3.0	0.7
Educational sciences	2	4.5	4.0	2.5	2.0	2.0	3.5	5.0	3.0	2.5	1.5	2.0	1.5	3.0	1.5	1.5
Media and communications	5	4.0	4.2	3.6	3.2	3.2	3.6	2.2	3.4	2.4	2.2	2.2	1.2	1.4	2.6	0.8
Political Science	3	3.3	3.7	3.0	1.3	3.0	3.0	3.0	3.0	2.0	2.3	2.0	2.3	3.3	2.3	0.3
Psychology	23	3.7	4.0	3.6	3.3	3.2	3.3	4.0	3.5	3.1	2.8	2.6	2.0	2.9	3.6	2.4
Social and economic geography	2	2.0	4.5	4.5	2.5	3.5	4.0	4.0	3.5	3.5	3.5	3.5	3.5	4.0	3.5	3.0
Sociology	2	1.5	4.0	3.5	4.0	2.0	4.0	4.0	4.0	3.0	1.5	1.5	1.5	2.5	2.0	2.0
Other social sciences	7	3.7	3.6	3.1	3.6	2.9	3.7	4.1	3.1	4.0	2.7	2.9	2.4	3.3	3.7	3.0

1 = not at all important / 5 = very important / 0 = no answer

Q_1: Availability of preferred usage method
Q_2: Multilingual interface
Q_3: Open source
Q_4: Existence of an API
Q_5: Integration of other tools
Q_6: Support options
Q_7: Collaborative environment for teams/groups with access rights system
Q_8: File encryption
Q_9: Free-text search capability
Q_{10}: Search filters (faceted search)
Q_{11}: Support for flexible research workflows
Q_{12}: Wizards to guide through processes
Q_{13}: Feature for previewing dataset content
Q_{14}: Visualization of datasets
Q_{15}: Functionality to anonymize research data
Q_{16}: Compliance with data regulations
Q_{17}: Control over data location
Q_{18}: Full control over own data
Q_{19}: Tracking of content changes
Q_{20}: Control over data release dates
Q_{21}: Control over data expiry dates
Q_{22}: Data export in different formats
Q_{23}: Possibility of data citation
Q_{24}: Transformation of data within the RDMS
Q_{25}: Tracking of interactions of external entities with dataset
Q_{26}: Data provenance
Q_{27}: Unique identifiers for data
Q_{28}: Logical links between related data
Q_{29}: Mandatory/optional metadata elements
Q_{30}: Automated metadata tagging
Q_{31}: Possibility to set up multiple metadata schemata
Q_{32}: Domain-specific metadata schemata
Q_{33}: Data validation
Q_{34}: Data identification via DOI/URL
Q_{35}: Compliance with OAI-PMH

Table 6.14: Importance of Criteria for an RDMS - Social Sciences

6.8 ADDITIONAL FEATURES REQUESTED FOR RDMS

Table 6.15 contains a full list of other functions or features the respondents would like an RDMS to offer, grouped by the categories "General", "Features", "Data", and "Storage". Where needed, the input was translated or anonymized, and entries covering information about more than one category were split up accordingly.

General:
(Very) simple interfaces for important programming languages so that self-written tools can store results directly in the RDMS instead of first writing CSVs or similar. Just as simple interfaces to access data, e.g. to create graphics.
An easy intuitive user interface for data entry, and responses to queries, by nurses, doctors, researchers, and CRO staff is essential
Generally user-friendly. Of course, it's okay if you have to spend a little time initially learning how to use the system, but no in-depth computer science knowledge should be necessary
user friendliness
User-friendly interface.
It should also be barrier-free. This means e.g. that entries should be possible in several ways (not only via the mouse, but also via the keyboard, for example). It should be compatible with screen readers and magnification capabilities.
The range of functions must not come at the expense of accessibility, although accessibility does not necessarily have to be equated with "easy to use".
Personalization, adaptive
Cross-institutional exchange with colleagues
Internal use for all the laboratories in the facility.
free of charge
If your system has the same framework as the one that will be used for an electronic manufacturing batch record tracking, large industries might consider this a significant asset (training staff for proper software usage is resource-consuming)
Institutional support in the way that the IT department research data personnell can be trained to establish the RDMS in the institution. It needs also support for the data storage infrastructure.
Strongly force people to input all information they have, so it's not scattered across notebooks. Have a local backup.
not regarding the RDMS itself, but maybe to raise awareness towards the usefulness and availability or RDMS
Features:
Supporting features for the data curator. Quality indicators on the data sets, e.g. traffic light system.
3D visualizations
Artificial Intelligence in visualizations
Outputs in terms of visual data management and analysis and also self generating reports.
Good data visualization tools such as word clouds or counts and author maps with interlinks.
View data by studies - we are a centre with multiple groups and multiple studies - often exceeding over 50 projects.
Linking data to each other; additionally: linking to publications etc. to the relevant data; input categories that can be set up flexibly
The most important I guess would be flexibility for different data types and being able to link different data sets together. For example, I have biweekly monitoring data for my site, an incubation study, and soil cores. Those are three sets of data that are all related but can't be put together easily into one data sheet. It would be nice to be able to link them together. For example, the ability to view Site A monitoring data, soil core information, and incubation results together somehow. In this example, there are other sites, for example B-D, and they might not all have incubation data or core data.
Offering other data based on the searched one
various filter functions, e.g. according to the survey context, among others.
Possibility to keep data private it the data is not for public research.
anonymization according to standards, confidentiality assessment according to classes
A good help function to help us use the data entry and sign off, so that if we forget, it is easy to find out how to use the database.
Essentially a best practice howto for the most common research processes within the field - e.g. standard cross-sectional survey, multi-wave survey for intervention studies, diary studies, etc. Having this as a template in the system and then being able to adapt it within the system would be attractive.
As long as I have API access where I can filter the database as I want, I'm happy.
Capable for use with qualitative data entry.
Capable for use in different field settings - for example I have conducted a lot of research in rural and informal settlements in lower and middle income countries, where there are additional challenges to data entry and use.
Compatibility with a LIMS (Laboratory Information Management System)
Integration of existing repositories such as GitHub, GitLab, Zenodo, OSF, . . .
Interoperability with R/Rstudio
Possibility of web publication, especially of digital editions
Practicality regarding tools that are coupled with it (e.g. including license management)
I don't really like working with [...] WebGUI. I wish I could do all of the protocols and lab notes in MS Word so I could import them afterwards, otherwise I'd be doing the work twice.
Feature to integrate LLMs
Handwriting possibilities
Translation of audio files (e.g. interview recordings) into text.
Digitalization of paper data via screenshots and character recognition.
Loading existing metadata schemata and offering a user interface to fill them out
Usability of the data for research applications (statistical values, grouping of certain data sets, e.g. how many publications/lectures were created, percentage of certain data in the total) -> i.e. reusability of one's own data that goes beyond the RDM
Data:
Ability to mange multiple types of data e.g. genomic, imaging, clinical, pre-clinical omics
Consider the management of image data set.
Manage models, code, scripts and workflows as data
DOI
Storage
Tamper-proof storage
Data longevity should be guaranteed
Long-term archiving
Something related to the Scalability. As data volumes grow, ensuring that the RDMS can handle increased load and complexity becomes crucial for efficient research operations.

Table 6.15: Functions/Features Requested

[Amo+17] Ricardo Carvalho Amorim, João Aguiar Castro, João Da Rocha Silva, and Cristina Ribeiro. „A comparison of research data management platforms: architecture, flexible metadata and interoperability." In: *Universal Access in the Information Society* 16.4 (2017), pp. 851–862. ISSN: 1615-5289. DOI: 10.1007/s10209-016-0475-y.

[And+07] Nicholas R. Anderson, E. Sally Lee, J. Scott Brockenbrough, Mark E. Minie, Sherrilynne Fuller, James Brinkley, and Peter Tarczy-Hornoch. „Issues in biomedical research data management and analysis: needs and barriers." In: *Journal of the American Medical Informatics Association : JAMIA* 14.4 (2007), pp. 478–488. ISSN: 1067-5027. DOI: 10.1197/jamia.M2114.

[Arn22] Witold Arndt. *A survey on research data management practices among researchers in the Helmholtz Association*. 2022. DOI: 10.3289/HMC_publ_05.

[BM21] Blend Berisha and Endrit Mëziu. *Big Data Analytics in Cloud Computing: An overview*. 2021. DOI: 10.13140/RG.2.2.26606.95048.

[BS22] Katarzyna Biernacka and Sandra Schulz. *Forschungsdatenmanagement in der Informatik*. Humboldt-Universität zu Berlin, 2022. DOI: 10.18452/25376. URL: https://edoc.hu-berlin.de/handle/18452/26061.

[Bre+19] Bela Brenger, Ania Lopéz, Stephanie Rehwald, Stefan Stieglitz, and Konstantin Wilms. *UNEKE: Vom USB-Stick zur NFDI - Entwicklung eines Kriterien-geleiteten Entscheidungsmodells für den Aufbau von Forschungsdaten-Infrastrukturen. Forschungsdatenspeicherung - Praxis und Bedarfe. Online-Survey 2019*. Aug. 2019. DOI: 10.4232/1.13321.

[BC18] Maryam Bugaje and Gobinda Chowdhury. „Identifying Design Requirements of a User-Centered Research Data Management System." In: *Maturity and Innovation in Digital Libraries*. Ed. by Milena Dobreva, Annika Hinze, and Maja Žumer. Vol. 11279. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 335–347. ISBN: 978-3-030-04256-1. DOI: 10.1007/978-3-030-04257-8_35.

[Car12] Jake Carlson. „Demystifying the data interview." In: *Reference Services Review* 40.1 (2012), pp. 7–23. ISSN: 0090-7324. DOI: 10.1108/00907321211203603.

[Com02] Committee for Scientific and Technological Policy. *Revised Field of Science and Technology (FOS) Classification in the Frascati Manual*. [Online; accessed 08-October-2023]. 26.02.2007. URL: https://www.oecd.org/science/inno/38235147.pdf.

[CH15] Constanze Curdt and Dirk Hoffmeister. „Research data management services for a multidisciplinary, collaborative research project.“ In: *Program: electronic library and information systems* 49.4 (2015), pp. 494–512. ISSN: 0033-0337. DOI: 10.1108/PROG-02-2015-0016.

[Dir11] Direktion Bevölkerung, Bereich Forschung und Digitalisierung. *Österreichische Systematik der Wissenschaftszweige 2012*. [Online; accessed 08-October-2023]. 11.11.2021. URL: https://www.statistik.at/kdb/downloads/pdf/prod/OEFOS_2012_Alphabetikum_A_20211111.pdf.

[EM15] Thomas Ebel and Alexia Meyermann. *Hinweise zur Anonymisierung von quantitativen Daten*. [Online; accessed 23-October-2023]. Frankfurt am Main, 2015. URL: https://www.forschungsdaten-bildung.de/files/fdb-informiert-nr-3.pdf.

[EMS16] Thomas Eifert, Stephan Muckel, and Dominik Schmitz. „Introducing research data management as a service suite at RWTH Aachen university.“ In: *9. DFN-Forum Kommunikationstechnologien*. Bonn: Gesellschaft für Informatik e.V., 2016, pp. 55–64. ISBN: 978-3-88579-651-0.

[FOR16] FORCE11. *The FAIR Data Principles - FORCE11*. [Online; accessed 13-July-2023]. 2016. URL: https://force11.org/info/the-fair-data-principles/.

[Fer16] Nicky Ferguson. *Report For The Proposed Research Data Shared Service On Focus Groups Held Between May And October 2016 And The Metadata Issues And Requirements Identified*. 2016. DOI: 10.5281/zenodo.193018.

[Fic+22] Urs Alexander Fichtner, Boris Brühmann, Harald Binder, and Jochen Knaus. *Investigating needs within conceptualization of research data management for the faculty of medicine at the University of Freiburg*. 2022. DOI: 10.5281/zenodo.7085977.

[Fra24] Natalie Franz. „Aufarbeitung von Standards und Methoden im Forschungsdatenmanagement.“ Regensburg, 2024.

[GKP22] Jens Garstka, Jochen Kerdels, and Gabriele Peters. *Einführung in Mensch-Computer-Interaktion: Kurseinheit 5: Entwurfsmethoden für interaktive Systeme*. Hagen, 2022.

[Kay15] John Kaye. *Jisc RDM Shared Service Pilot Initial Statement of Requirements*. [Online; accessed 07-August-2023]. 2015. URL: https://researchdata.jiscinvolve.org/wp/files/2015/11/Draft-RDMSS-Requirements-Specification-V1.0.docx.

[KBF17] John Kaye, Rachel Bruce, and Dom Fripp. „Establishing a shared research data service for UK universities.“ In: *Insights the UKSG journal* 30.1 (2017), pp. 59–70. DOI: 10.1629/uksg.346.

[Ken+08] Gregor E. Kennedy, Terry S. Judd, Anna Churchward, Kathleen Gray, and Kerri-Lee Krause. „First year students' experiences with technology: Are they really digital natives?" In: *Australasian Journal of Educational Technology* 24.1 (2008). ISSN: 1449-3098. DOI: 10.14742/ajet.1233.

[Kni15] Gareth Knight. „Building a research data management service for the London school of hygiene & tropical medicine." In: *Program: electronic library and information systems* 49.4 (2015), pp. 424–439. ISSN: 0033-0337. DOI: 10.1108/PROG-01-2015-0011.

[Kon23] Universität Konstanz. *Glossar | Praxis kompakt | Forschungsdaten und Forschungsdatenmanagement*. [Online; accessed 13-July-2023]. 2023. URL: https://forschungsdaten.info/praxis-kompakt/glossar/.

[Krä15] Esther Krähwinkel. *Forschungsdatenmanagement an der Philipps-Universität Marburg. Die Ergebnisse der Umfrage zum Forschungsdatenmanagement im November 2014*. July 2015. DOI: 10.17192/es2015.0019.

[May13] Horst Otto Mayer. *Interview und schriftliche Befragung: Grundlagen und Methoden empirischer Sozialforschung*. 6., überarb. Aufl. Berlin/Boston: De Gruyter and De Gruyter Oldenbourg, 2013. ISBN: 978-3-486-71762-4.

[Müc+07] Claus Mückschel, Jens Nieschulze, Cornelia Weist, Branislav Sloboda, and Wolfgang Köhler. „Herausforderungen, Probleme und Lösungsansätze im Datenmanagement von Sonderforschungsbereichen." In: *eZAI* 2 (2007).

[Pre01] Marc Prensky. „Digital Natives, Digital Immigrants Part 1." In: *On the Horizon* 9.5 (2001), pp. 1–6. ISSN: 1074-8121. DOI: 10.1108/10748120110424816.

[Raz11] Matthias Razum. *Handbuch Forschungsdatenmanagement*. Ed. by Stephan Büttner, Hans-Christoph Hobohm, and Lars Müller. Bad Honnef: Bock + Herchen, 2011. Chap. Systeme und Systemarchitekturen für das Datenmanagement, pp. 123–138. ISBN: 978-3-88347-283-6.

[Ril17] Jenn Riley. *Understanding metadata: What is metadata, and what is it for*. NISO Primer series. Baltimore, MD: National Information Standards Organization, 2017. ISBN: 978-1-937522-72-8. URL: http://www.niso.org/apps/group_public/download.php/17446/understanding%20metadata.

[Sam01] P. Samarati. „Protecting respondents identities in microdata release." In: *IEEE Transactions on Knowledge and Data Engineering* 13.6 (2001), pp. 1010–1027. ISSN: 10414347. DOI: 10.1109/69.971193.

[Sch+22] Rolf Schwartmann, Andreas Jaspers, Niels Lepperhoff, Steffen Weiß, and Michael Meier. *Praxisleitfaden zum Anonymisieren personenbezogener Daten*. [Online; accessed 23-October-2023]. 2022. URL: https://stiftungdatenschutz.org/fileadmin/Redaktion/Dokumente/Anonymisierung_personenbezogener_Daten/SDS_Studie_Praxisleitfaden-Anonymisieren-Web_01.pdf.

[SW21] Hans-Werner Six and Mario Winter. *Software Engineering: Kurseinheit 1: Software Engineering im Überblick*. Hagen, 2021.

[Swe02] Latanya Sweeney. „k-anonymity: a model for protecting privacy." In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 557–570. ISSN: 0218-4885. DOI: 10.1142/S0218488502001648.

[Wil] „The FAIR Guiding Principles for scientific data management and stewardship." In: *Scientific data* 3 (2016), p. 160018. DOI: 10.1038/sdata.2016.18.

[Tun09] Daniel Tunkelang. *Faceted Search*. Cham: Springer International Publishing, 2009. ISBN: 978-3-031-01134-4. DOI: 10.1007/978-3-031-02262-3.

[Voc19] Rainer Vock. *BExIS user survey 2019: Evaluation des Zentralen Datenmanagements der Biodiversitäts-Exploratorien (ZDM-BE) an der Friedrich-Schiller-Universität Jena*. 2019. DOI: 10.22032/dbt.40313.

[VKP23] Lars Vogt, Marcel Konrad, and Manuel Prinz. *Knowledge Graph Building Blocks: An easy-to-use Framework for developing FAIREr Knowledge Graphs*. 2023. DOI: 10.48550/arXiv.2304.09029.

[WM16] Christie Wiley and William H. Mischo. *Data Management Practices and Perspectives of Atmospheric Scientists and Engineering Faculty*. 2016. DOI: 10.5062/F43X84NJ.