Open Research Data: Landscape and cost analysis of data repositories currently used by the Swiss research community, and requirements for the future

Dr. Markus von der Heyde – InformationsTechnologe (vdH-IT)
ORCID: 0000-0002-6026-082X
Weimar, April 2019

Abstract

In light of the SNSF's decision to make sharing data from funded projects mandatory in 2016, this study examined the sharing and reuse behaviour of researchers in the Swiss community in 2018. Since it was to be conducted across all disciplines throughout Switzerland, the range of the questions was very broadly designed and questions from earlier international studies were used for comparability. Additionally, a second questionnaire addressed international repositories in order to learn about their perspectives and plans for future development. The results were analyzed using statistical methods and can be regarded as representative.

Generally, the motivation and concerns for sharing data and reuse in the Swiss community are not different from other scientific communities. Differences in sharing and reuse behaviour are found according to the disciplines of the researchers, which were assessed using the bepress taxonomy. Different methods used by the researchers did not result in different sharing behaviour, but in where the data was shared. While the sharing is done equally in general repositories and smaller disciplinary repositories, of which a great number exist, the researchers prefer to use disciplinary repositories if they want to reuse data.

Overall, about a third of the Swiss research community share data in repositories. The main reason for not sharing was researchers' plans to publish their results first. Also, many participants claimed to have a different concept of data; while we tried to define terms carefully, apparently there is a need for more discipline-specific information and discussion on the topic. Future requirements for services from the Swiss community are not yet met by the international repositories' plans. Several recommendations on the future SNSF governance on data sharing are proposed to conclude the study.

Please cite this work as:
von der Heyde, M. (2019). Open Research Data: Landscape and cost analysis of data repositories currently used by the Swiss research community, and requirements for the future [Report to the SNSF]. Retrieved from https://doi.org/10.5281/zenodo.2643460

Contents

1 Introduction

Research is based on many different sources, including historical artefacts, simulations, empirical research data, concepts, and primary literature. However, every discipline of science[1] produces results and makes them accessible through publication. Often, background information is shared within the discipline for the sake of projects or collaborations.

One way to share the results of scholarly production is to upload publications, in conjunction with the research data underlying them, into an Open Data Repository. The Swiss National Science Foundation (SNSF) and swissuniversities are encouraging this and plan to support the research community with appropriate funding. They therefore mandated a survey to establish a broad overview of the current data sharing and repository situation. This multi-layered description and analysis aimed to explore the overall landscape of data repositories already in use, their future development and the services required from them, as well as the current needs of various disciplines and scientists within the whole range of scientific methods.

The analysis consisted of three main parts, each of which added one perspective necessary for the overview. Existing databases and research outcomes on both national and international communities were collected and act as a baseline. The landscape survey across the complete Swiss research community collected information from 2,384 scientists about their data sharing practices and data reuse via an online questionnaire. The repository survey added the perspective of 208 international repositories in terms of their genesis, provided services and use, cost and finance structure, and self-assessment of the degree of FAIR principle implementation.

All three perspectives provide a complex picture of the overall Swiss research community and its needs, objections to policies, highly diverse attitudes on data, and perceptions of the value of data as well as the need to share and reuse them. This paper summarizes the results of many statistical analyses on an abstract level and provides a set of recommendations for the SNSF regarding future changes to policies on open data.

2 Methodology

2.1 Terms and definitions

The research was carried out with the intention of applying standard procedures as often as possible. However, crucial terms like “data” and “sharing” are known to have different meanings throughout the scientific community. The project therefore adopted with care a definition which is both close to those of the contracting authorities (SNSF, swissuniversities) and used in previous work in the area.

Data: We define data using the NIH definition of ‘Final Research Data’, as follows: “Recorded factual material commonly accepted in the scientific community as necessary to document and support research findings. This does not mean summary statistics or tables;

[1] We use the terms “science” and “research” as synonyms. Social sciences, humanities, life sciences, natural sciences, engineering and all potential other fields of research are considered to be equally relevant for this project.

rather, it means the data on which summary statistics and tables are based. For the purposes of this policy[2], final research data do not include laboratory notebooks, partial datasets, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as gels or laboratory specimens." (National Institutes of Health (NIH), 2003).

Repository: A repository offers archival functionality to publicly share data used in scientific publications. A repository contains data packages, which are described by meta-data to allow search by humans and machines. The size of data packages is measured in kilo-, mega-, giga- or terabytes.

FAIR: The FAIR principles were collected and published by a distinct group of stakeholders (Wilkinson et al., 2016). Their aim was to establish a baseline for the requirements on repositories to find entries not only by human scholars, but also support machine-based queries in automatic ways.

Sharing: The spectrum between simple personal data sharing and FAIR / open research data can be described as analogous to the Curation Lifecycle Model[3] and with reference to the Data Continuum Model (Treloar, Groenewegen, & Harboe-Ree, 2007). Therefore, all forms of data sharing — from personal connection and exchange via email to open data platforms and published work in journals — are included.

Reuse: Scientific data which is collected for one purpose by one group of scientists and used for other purposes by other scientists is considered to be reused.

Open Science: The word "open" is very general and can be seen from many different perspectives. The SNSF expects that data generated by funded projects are publicly accessible in digital databases provided there are no legal, ethical, copyright or other issues[4]. The survey's use of "open" focused on aspects of free access and accessibility within the scientific community. The access often has to be managed by means of an authorisation infrastructure, which in itself is not the focus of this research. For all terms concerning Open Science, we refer to the Foster taxonomy of open science (Knoth & Pontika, 2015). The most relevant terms in this project are "Open Data Use and Reuse" and "Open Repositories".

Scientific Disciplines: Generally the scientific community gathers knowledge about different topics. In structuring the overall landscape, we divided the community into disciplines which share a common understanding of a special topic (see section 2.3.1).

Scientific Methods: Independent from the topic, methods are used throughout the scientific community to learn about something. In this project, we limited ourselves mostly to methods on the level of "Frameworks for Research and Research Designs" following (Beissel-Durrant, 2004; Luff, Byatt, & Martin, 2015) as described in section 2.3.2.

[2] The term "policy" was replaced by "survey" in the introduction text to the surveys.
[3] See http://www.dcc.ac.uk/resources/curation-lifecycle-model.
[4] See http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/default.aspx#.

2.2 Data sources and tools

The data in both surveys were collected in SurveyMonkey[5], and reviewed and tested for plausibility in Excel (Version 10). Detailed descriptions are given in (von der Heyde, 2019b, 2019a). The complete data records are available in various forms (raw, checked for plausibility and processed) on Zenodo[6].

The data analysis for the Principal Component Analysis (PCA) and Factor Analysis (FA) statistical tests was performed with JMP[7] Version 13.0. JMP was also used for categorical tests, including Chi² (Chi Square) values.

Further data records from other sources were included in the analysis and are referred to in the appropriate chapters. We would like to thank the open science platforms re3data, FAIRsharing, p³, openDOAR, and openAIRE for their generous support and open APIs. Please refer to "Appendix B: Data sources" for additional information.

2.3 Methods

The respective methods for gathering data in the landscape and repository surveys are described in the data papers. Here we focus on how the data has been analysed and interpreted.

Since most researchers use more than one method and do research in more than one discipline, they were able to name all those applicable to them. To compare data, we have to group those methods and disciplines again into a schema useful for comparison. The two key filters we have applied are the mapping of scientific disciplines and the mapping of scientific methods. They provide the independent variables to perform statistical comparisons.

2.3.1 Mapping of "Scientific Disciplines"

The main scientific disciplines, according to (DFG, 2017), are

- Humanities and Social Sciences
- Life Sciences
- Natural Sciences
- Engineering Sciences

This project followed the definition of the 14 **scientific areas** and **scientific subjects** used by the German Research Foundation[8], which is also used for the re3data classification. This level of abstraction is used for most analyses in this report on differences between disciplines. The survey itself was conducted using the bepress discipline taxonomy, which includes 1,243 terms (Warner, 2018). Therefore, the dataset can be mapped onto any other constellation in which disciplines of bepress are pooled together in one abstract description of a "discipline". Data in the report are also mapped onto systematic approaches of the SNSF

[5] Software for online questionnaires developed by the SVKM Inc. See description at https://en.wikipedia.org/wiki/SurveyMonkey.
[6] See SNSF Community at https://zenodo.org/communities/snsf/.
[7] Software for statistical analysis developed by the SAS Institute. See description at https://en.wikipedia.org/wiki/JMP_(statistical_software).
[8] Deutsche Forschungsgemeinschaft (DFG)

(SNSF, 2016) and FSO[9] (BFS, 2018), as they are most relevant to the target audience. Additional scientific subjects (e.g., ethics, sports sciences, military studies, and gender studies) have been classified within the most appropriate scientific area without extending the DFG classification or have been grouped into other areas depending on the purpose.

2.3.2 Mapping of Scientific Methods

The various methods used in science are often described in the context of a specific publication. To compare literature systematically, a number of disciplines have started to collect and classify typical methods. However, an overall catalog of scientific methods with a structure that allows most disciplines to easily find their specific terms could not be located. Therefore, a systematic collection of methods from various fields was conducted.

For the social sciences, a typology from the NCRM (Beissel-Durrant, 2004; Luff et al., 2015) distinguished the main categories (or hierarchies) of the overall typology. Within this approach, we limited ourselves to methods fitting the category "1. Frameworks for Research and Research Designs". We **excluded** terms from the following hierarchy levels unless they were essential for other disciplines:

2. Data Collection
3. Data Quality and Data Management
4. Data Handling and Data Analysis
5. ICT, Software and Simulation
6. Research Management and Application of Research
7. Research Skills, Communication and Dissemination

For the other scientific disciplines, we tried to match the level of abstraction given by the updated version of the NCRM taxonomy (Luff et al., 2015). Vessey et al. collected computational methods suitable for most quantitative and engineering disciplines (Vessey, Ramesh, & Glass, 2005). Pickard and Dixon's work on philosophical methods was used to derive central abstract terms for the humanities (Pickard & Dixon, 2004). To enhance the catalog further, a cross check on Wikipedia was performed.

To reasonably limit the number of methods and still have a sufficiently complete collection, the list was reviewed and synonyms or rare terms were excluded. The final collection offered 50 methods.

To provide a grouping across all methods for analytical purposes, we defined the following levels of abstraction:

- Qualitative Methods
- Quantitative Methods
- Meta Methods
- Analytical Methods
- Critical Methods
- Speculative Methods
- Creative Methods

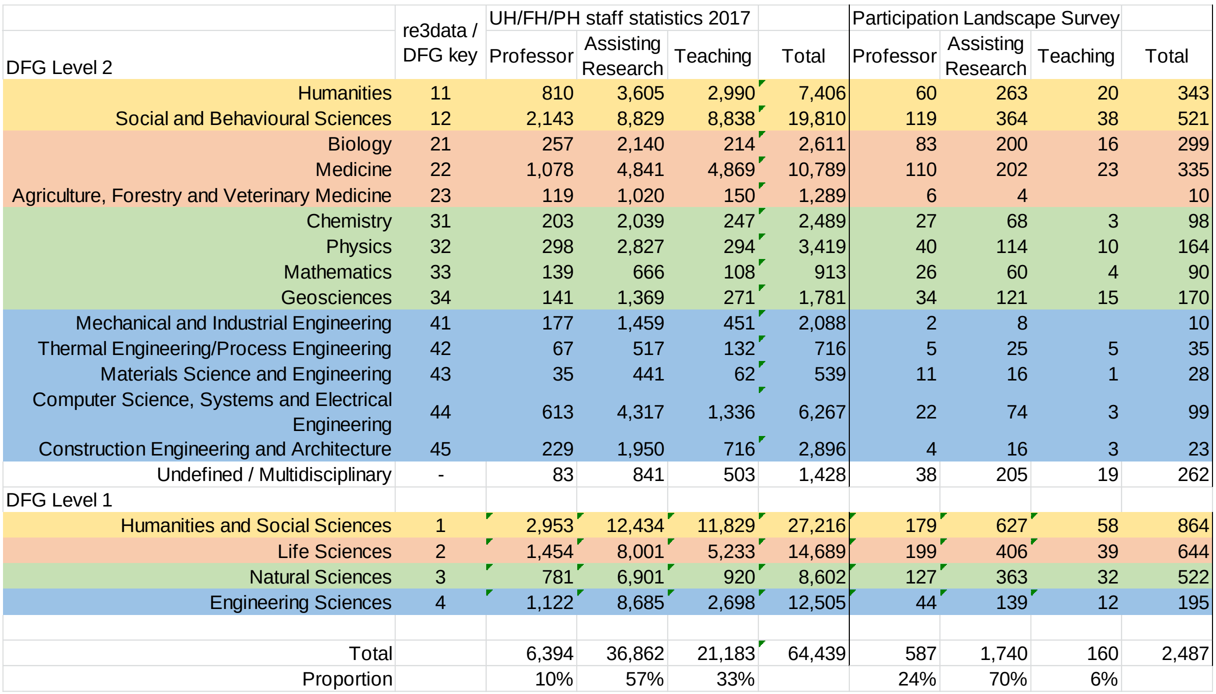
[9] Swiss Federal Statistical Office (FSO); German: Bundesamt für Statistik, BFS.

The methods from which researchers could select and their mapping into these seven categorical classes can also be found in Appendix D: Mapping of Scientific Methods.

2.4 Validation

Is the sample of the landscape survey representative of the Swiss research community?

In Table 1, the number of researchers, research assisting and teaching personnel in Switzerland's higher education sector is summarized according to the DFG's classification. All participants in the landscape survey are grouped according to the same system. The overall participation rate of the staff class 'professor' was highest (9.2% = 587 of the 6,394 professors in Switzerland). The assisting research personnel group (mostly doctoral students) had the highest absolute number of participants (1,740), but the survey reached only 4.7% of all people in this group. As teaching staff were not excluded but not explicitly targeted, the proportion of 6% (participation rate 0.8% overall) is very low.



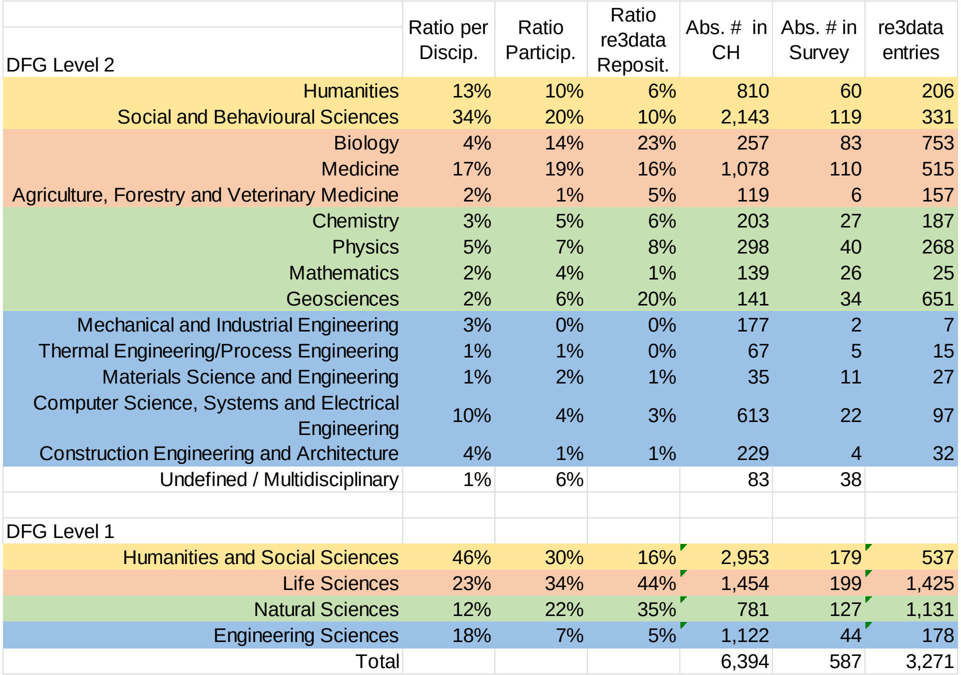
	re3data / DFG key	UH/FH/PH staff statistics 2017				Participation Landscape Survey			
DFG Level 2		Professor	Assisting Research	Teaching	Total	Professor	Assisting Research	Teaching	Total
Humanities	11	810	3,605	2,990	7,406	60	263	20	343
Social and Behavioural Sciences	12	2,143	8,829	8,838	19,810	119	364	38	521
Biology	21	257	2,140	214	2,611	83	200	16	299
Medicine	22	1,078	4,841	4,869	10,789	110	202	23	335
Agriculture, Forestry and Veterinary Medicine	23	119	1,020	150	1,289	6	4		10
Chemistry	31	203	2,039	247	2,489	27	68	3	98
Physics	32	298	2,827	294	3,419	40	114	10	164
Mathematics	33	139	666	108	913	26	60	4	90
Geosciences	34	141	1,369	271	1,781	34	121	15	170
Mechanical and Industrial Engineering	41	177	1,459	451	2,088	2	8		10
Thermal Engineering/Process Engineering	42	67	517	132	716	5	25	5	35
Materials Science and Engineering	43	35	441	62	539	11	16	1	28
Computer Science, Systems and Electrical Engineering	44	613	4,317	1,336	6,267	22	74	3	99
Construction Engineering and Architecture	45	229	1,950	716	2,896	4	16	3	23
Undefined / Multidisciplinary	-	83	841	503	1,428	38	205	19	262
DFG Level 1									
Humanities and Social Sciences	1	2,953	12,434	11,829	27,216	179	627	58	864
Life Sciences	2	1,454	8,001	5,233	14,689	199	406	39	644
Natural Sciences	3	781	6,901	920	8,602	127	363	32	522
Engineering Sciences	4	1,122	8,685	2,698	12,505	44	139	12	195
Total		6,394	36,862	21,183	64,439	587	1,740	160	2,487
Proportion		10%	57%	33%		24%	70%	6%	

Table 1: Number of participants in the landscape survey in comparison to the overall Swiss scientific community (UH/FH/PH = all Universities and Universities of Applied Sciences; data provided by the BFS), sorted by the DFG discipline schema.

Across all disciplines, both the number of professors and the proportion of them vary considerably. Two main factors are important to understand the (self-)selection process. The first is the distribution of disciplines across Switzerland, as shown in Table 2. The second is the proportion of disciplines targeted by specific repositories cataloged by the re3data base, also in Table 2. The sum of databases indicated by re3data is presumably correlated with the overall importance of open data to the discipline (Kindling et al., 2017).

The rate of participation almost always lies in between those two values. This is in accordance with the notion of a self-selection bias. Scientists were more willing to participate in cases where Open Research Data is an important topic for them and/or for their discipline.

The collected sample is therefore representative to an acceptable level in both views: It is close to the distribution of scientific disciplines, but also reflects the distribution of disciplines within the open data repository landscape to a high degree. There seems no better way to match both factors at the same time.



DFG Level 2	Ratio per Discip.	Ratio Particip.	Ratio re3data Reposit.	Abs. # in CH	Abs. # in Survey	re3data entries
Humanities	13%	10%	6%	810	60	206
Social and Behavioural Sciences	34%	20%	10%	2,143	119	331
Biology	4%	14%	23%	257	83	753
Medicine	17%	19%	16%	1,078	110	515
Agriculture, Forestry and Veterinary Medicine	2%	1%	5%	119	6	157
Chemistry	3%	5%	6%	203	27	187
Physics	5%	7%	8%	298	40	268
Mathematics	2%	4%	1%	139	26	25
Geosciences	2%	6%	20%	141	34	651
Mechanical and Industrial Engineering	3%	0%	0%	177	2	7
Thermal Engineering/Process Engineering	1%	1%	0%	67	5	15
Materials Science and Engineering	1%	2%	1%	35	11	27
Computer Science, Systems and Electrical Engineering	10%	4%	3%	613	22	97
Construction Engineering and Architecture	4%	1%	1%	229	4	32
Undefined / Multidisciplinary	1%	6%		83	38	
DFG Level 1						
Humanities and Social Sciences	46%	30%	16%	2,953	179	537
Life Sciences	23%	34%	44%	1,454	199	1,425
Natural Sciences	12%	22%	35%	781	127	1,131
Engineering Sciences	18%	7%	5%	1,122	44	178
Total				6,394	587	3,271

Table 2: Comparison of the rate of participation of professors in the landscape survey, the proportion of research-oriented professors in these disciplines in the overall scientific community, and the number of disciplinary repositories in the re3data database.

The self-selection bias in the form of potential order effects is discussed in the data papers. In addition, we observed order effects for the independent variable of scientific discipline (grouped for DFG and SNSF catalogues): In the landscape survey, participants from biology were among the first to respond after receiving the invitation and reminder. In contrast, the majority of social scientists, especially from the education disciplines, participated rather towards the end of the survey. This corresponds with the results in the data paper.

However, the comparison of dependent variables between early and late participants did not produce any strong effects. Neither did the grouping for methods. When both groups are very similar, the sample is considered to be representative for the overall Swiss scientific community.

3 Limitations

Overall, 755 of the total 2,384 participants commented on the Landscape survey. The comments were evaluated and taken into consideration during the statistical analysis and the compilation of the final reports. However, the shortcomings of the survey should be considered during future research.

The landscape survey data paper shows the distribution of the comments across various categories (von der Heyde, 2019b). In terms of limitations of the methodology, we evaluated the 52 comments specifically on this topic. 81 comments on the specific goals of the survey were also taken into account.

The number of comments on the methodological approach was considerably lower in the repository survey (von der Heyde, 2019a). Again, the data paper presents the complete distribution of the 210 comments. The majority were comments on the given answers in the standardized choices (64%). Only 10 participants criticised certain aspects of the survey or it goals.

3.1 Methodology

Several participants proposed having a “not applicable (n.a.)” option for most of the questions. Skipping of a question or page was considered misleading since it could also suggest "I don't know". The sliders (chosen to reduce answering time) were judged to make answers somewhat arbitrary.

The general criticism of too extensive a catalog for the disciplines and methods was expressed several times. However, the bepress catalog was chosen for its mapping potential.

3.2 Goal and scope of the survey

The very general approach of the survey led to quite opposite comments. Some participants valued the specific aspects and design of the survey, while others thought their specific needs were not properly covered. Open data in general is viewed in a variety of ways and the need for more discussion was expressed. Some participants declared the survey to be biased since the problems of data sharing were not mentioned.

Standard scientific principles, such as the reproducibility of scientific results, are standard in science in Switzerland just as everywhere else. Since no specific results for the Swiss community were to be expected, these categories were not included in questions concerning purpose of sharing and reuse. However, these standards are often addressed by other research (Eynden et al., 2016, Chapter 6.5; Pasquetto, Randles, & Borgman, 2017), thus some participants expected them to be included.

Researchers from disciplines in the humanities and social sciences remarked on the lack of specific aspects of qualitative research. This might be a result of our understanding of the terms “science” and “research” as synonymous, which was not stated in the introduction. Others noted that the survey only focused on hard sciences. An additional survey focusing on the humanities was suggested.

Participants also noted the sharing of computer source code to be of importance. As this is included in the NIH definition, it could have been part of the survey. The same applies to data sharing within research consortia.

The mentioning of commercial platforms (figshare, github, and others) was criticized due to the potential benefit to the platforms by advertising these options.

3.3 Time

Several participants felt they spent too much time on the survey due to its length, although it actually took less than 15 minutes for some of them. Others complained about repetitions and redundancies. 50% of all participants completed 17 to 25 pages, which on average took 21 minutes.

3.4 Terms and definitions

The common understanding of terms across all disciplines was an inherent challenge of the project. Understanding of data repository terms is apparently based on experience. Scientists not using repositories often doubted the validity of their answers, while others claimed the questions could not be answered without further distinction between raw, processed and interpreted data.

The definition of data (NIH) was not well received by at least 41 participants (= 1.6%). In the comments they referred to other concepts and stated uncertainty in definition. The definition also could lead to “blind spots”, a participant noted. It is possible that some of the difficulties were based on the confusion around the term “open”, which was defined in the introduction, but not specifically reflected with respect to the access modes to the data. Controlled and managed access had been taken for granted, but some researchers’ comments implied an understanding of unrestricted access.

47 participants (= 1.9%) stated they had general difficulties in the application of the “concept of data” to their discipline. A total of 32 participants (= 1.3%) indicated not having any data due to their disciplines (e.g. law, pure mathematics).

3.5 Technical issues

During the implementation of the web based questionnaire, one item of the bepress catalog was lost during copy and paste actions. The item “Vocational Education” was nevertheless used by participants and noted in “other”. This was discovered during quality control and a category was added into the data set accordingly.

4 Sharing and reuse of research data

Previous research has looked at data sharing from various perspectives. The motivation to share data can be summarized in four rationales: “(1) to reproduce or to verify research, (2) to make results of publicly funded research available to the public, (3) to enable others to ask new questions of existing data, and (4) to advance the state of research and innovation” (Borgman, 2012). The reasons why scientists share their data are highly diverse: Individual differences, disciplinary traditions, policies of the funding agencies, requirements of the journals and many more have been identified to be part of the complex situation in which research data is effectively shared. Recent literature reviews can be found in (Fecher, Friesike, & Hebing, 2015) and (Perrier et al., 2017). However, many publications have looked only at specific aspects of the problem and thus fail to provide a complete picture. The present project is trying to form an overview including as many perspectives as possible without over-simplification. It is focused on the relevance for the overall Swiss research community.

Results and concepts from various publications which also offer the reuse of their datasets were selected for the baseline of the landscape analysis. The broad survey on figshare users, although suffering from a selection bias, offered an extensive data report in addition to the complete datasets (Hahnel et al., 2017; Treadway et al., 2016). Kim and Stanton extended their qualitative interview-based approach (Kim & Stanton, 2012) in their survey on STEM disciplines 2012/2013 (Kim, 2016) and provided a detailed multilevel analysis of the combination of individual and institutional factors based on the behavioral model of data sharing activity (Kim & Stanton, 2016). Kim and Zhang extended this model further by including the concept of attitude towards data sharing (Kim & Zhang, 2015). In the context of data reuse, the attitude towards reuse was linked to the intention to reuse (Yoon & Kim, 2017). To enable a symmetric analysis, this step was applied to the context of data sharing in the current research. Further factors were motivated by the work of Linek et al. (Linek, Fecher, Friesike, & Hebing, 2017). They found links between the individual researchers' personalities and effective data sharing practises. Although being limited to researchers funded by the Wellcome Trust, the survey by Eynden et al. offered a good, validated collection of factors which contribute to data sharing and reuse (Eynden et al., 2016).

4.1 Why data are shared

Most relevant factors from previous works were selected on the basis of their contribution to theoretical frameworks and overall relevance for the quantitative explanatory power of data sharing. Often the original references to the variables used in other surveys were kept to enable easy identification. The complete set of variables used is shown in Table 3. The domains refer to the work of Kim and colleagues in their framework on factors, attitudes, and intentions. Factors from other authors were added to the domains to enable a comparison between the newly collected data and all references.

Domain - factors for data sharing		
	Reference / Variable	Rated statement on scale 0%=disagree ... 100%=agree
Motivation and data sharing behavior		
	Altruism1	I am willing to help other researchers by sharing data.
	A2-GreatContribution	Freely available research data is a great contribution to scientific progress.
	Altruism5	Sharing data contributes to better scientific research.
Perceived career benefit		
	ShareBenefit2	Data sharing would enhance my academic recognition.
	E6-Quotation	I would share my data if I were cited in publications using my data.
Perceived career risk		
	B1-BeforePublishing	I would share my data even if other researchers could use my data to publish before me.
Perceived effort		
	ShareEffort2	I need to make a significant effort to share data.
Attitude toward data sharing		
	ShareAttitude1	Sharing data is valuable.
Normative pressure		
	ShareNorm2	In my discipline, researchers care a great deal about data sharing.
	ShareNorm3	In my discipline, researchers share data even if not required by policies.
Metadata		
	ProvideMetadata2	In my discipline, researchers provide metadata when they share data.
Perceived availability of data repositories		
	ShareRepository2	In my discipline, data repositories are available for researchers to share data.
Perceived pressure by funding agencies		
	Funding3	Public funding agencies require researchers to share data.
Perceived pressure by journals		
	Journal3	Journals require researchers to share data.
Resources		
	ShareResource4	In my organization (e.g., university), information technologies are available to support my data sharing.
Intention to share data		
	ShareIntention1	I am likely to share my data from future research.

Table 3: Variables and corresponding statements used to assess why data was shared.

Note: A selection of these factors was also used to evaluate the willingness to share data which has not yet been shared, but is considered by the scientist as a potential "hidden treasure" (see section 8.3).

Participants were asked to rate the statements given in Table 3 on a scale between ‘I disagree’ and ‘I agree’. Even though some questions would generally prompt a yes / no answer, the sliders were set to a value between the extremes indicating a “level of agreement” most of the time. Therefore, the ratings could be used as a continuous rating which could be further analysed in PCA and FA (see also Appendix G: Principal component and factor analyses).

The findings of (Kim & Stanton, 2016) and (Kim & Zhang, 2015) were confirmed. As well, the hypotheses they built upon the findings of (Kim & Stanton, 2012) were confirmed, both on the level of intention and attitude, and the level of self-reported sharing. Multiple indicators were positively tested. The only negative correlation was observed between effort and data sharing behavior (intended and actual), again confirming the previous findings. The negative correlation of career risks and attitude towards data sharing was observed as a positive correlation due to a change in the phrasing of the statement; here the statement from Linek et al. was used instead of the original statement of Kim and colleagues.

The additional factors ShareResource4, A2-GreatContribution, E6-Quotation, and ProvideMetadata2 showed positive correlations on all levels of data sharing behaviour as expected; ShareAttitude1, ShareIntention1 and actual data sharing (‘Sharing frequency’ and ‘Published work within the last two years having been shared’) were significantly affected.

Overall, the Swiss research community was found to not be different from other communities investigated by other research teams.

4.2 Why data are reused

Without the reuse of research data, sharing would be a waste of energy and time. Again, most researchers agree with the general notion of shared knowledge and collaboration. Particularly in certain fields where data acquisition is expensive, complex or unique in other ways, the joint efforts involving shared datasets is long-established scientific praxis (Borgman, 2012; Pasquetto et al., 2017; Wallis, Rolando, & Borgman, 2013). At the same time, the notion of freely available research data and the ability to reuse others’ data without concerns is seen as critical by the community, as summarized by the same authors. The quality of research data, the research climate of the specific discipline, and the effort to adapt to others’ systematic and learned standards (if they exist) might all be linked to the attitude towards data reuse, the intention to reuse data and the actual occurring reuse.

In a survey symmetrical to the ones on data sharing, Kim and colleagues also searched for enabling factors (Kim, 2017; Kim & Yoon, 2017; Yoon & Kim, 2017). Mainly in reference to this survey, the set of variables shown in Table 4 was chosen. The aim again was to determine if the Swiss research community is different from other communities in the reuse of data.

Again, as in section 4.1, the participants were asked for their ratings and many made use of the possibility to not simply answer yes or no. The scale of 0-100 was again used by most respondents.

Domain - factors for data reuse		
	Reference / Variable	Rated statement on scale 0%=disagree ... 100%=agree
Motivation and data sharing behavior		
	ReuseAltruism1	I am willing to reuse others' data for my research.
	ReuseAltruism5	Reusing others' data contributes to better scientific research.
Perceived Usefulness		
	ReuseUsefulness1	Reusing other researchers' data improves the quality of my research.
Perceived Concern		
	ReuseConcerns1	If I reuse other researchers' data I worry that I might misinterpret the data.
Perceived Effort		
	ReuseEfforts1	Reusing other researchers' data requires time and effort to locate data sets.
Attitude towards data use		
	ReuseAttitude1	Reusing other researchers' data is valuable.
Subjective Norm		
	ReuseNorm1	In my discipline, it is expected that researchers reuse other researchers' data.
Availability of data repositories		
	ReuseRepository2	In my discipline, researchers can easily access data repositories to reuse data.
Organizational Resources		
	ReuseResources2	In my organization (e.g., university) information technologies are available to support my data reuse.
Disciplinary Climate		
	ReuseClimate1	In my discipline, researchers cooperate well.
Intention to Reuse Other Researchers' Data		
	ReuseIntention1	I am likely to reuse other researchers' data for my future research.

Table 4: Variables and corresponding statements used to assess why data was reused.

The ratings showed various correlations as expected (see Appendix F: Correlations for estimates of pairwise method), which are not discussed in detail. Instead, the main findings of the baseline from previous work were replicated. Some effects which were possibly too small for (Yoon & Kim, 2017) to detect, but were hypothesised due to the findings in (Kim & Yoon, 2017), were confirmed by our data.

The hypothesised negative effect of perceived effort in connection with the reuse intention was not found. Neither for the complete set nor for the reduction to DFG discipline 12 (Social and Behavioral Sciences) could we see any correlation of the variables ReuseEfforts1 and ReuseIntention1 (see Table 4 for the rated statements). Moreover, we could see a positive correlation of the ReuseEfforts1 and ReuseAttitude1 variables both for DFG discipline 12 and the complete dataset. The negative effect as in hypothesis H3 by (Yoon & Kim, 2017) was not confirmed. An alternative interpretation could be: If people have had positive experiences, they might rate the effort as adequate.

Regarding reuse of data, the Swiss research community was again not different from others.

4.3 Ways data are shared

The ways in which data are shared were assessed in reference to the Wiley study (L. Ferguson, 2014). The overall comparison is given in Figure 1. The range of options participants could choose from was extended based on feedback during a pilot phase of structured interviews. Most prominent was the addition of making data sharing visible outside of the organized institutions: personal contact between scientists is still the most prominent way of exchanging data.

The category “supplementary” still exists, but an additional option to indicate publication in data journals was offered. This potentially made a difference in the distribution of the participants’ answers between the Wiley study and the landscape survey.

The participants were also asked in a semi-symmetrical way if there had been projects from which no data was shared. This is different from other surveys, which tried to assess whether no data was shared by subtracting the “sharers” from the total. However, about 25% of the participants indicated that they had not shared some data. About 14% of the participants did not answer this question at all, which adds to the uncertainty.

Table 5 summarises the different ways of sharing. Since multiple answers could be given, different sub-proportions can be calculated. In comparison with the Wiley survey, the rate of sharing is considerably higher (1,272/2,031 = 63% vs 52% in Wiley’s survey).

14%	278	Participants skipped the question									
3%	61	Indicated others ways of sharing, when no other option was used									
7%	143	Indicated not to have shared data for some projects, but did not gave answers to any of the other sharing options									
76%	1,549	Indicated sharing in any form	18%	277	Share by personal request only						
			82%	1,272	Shared in an open way (journals, web, repositories)	54%	685	Explicit usage of repositories	34%	230	Repositories only
									38%	263	Rep + Journal
									15%	105	Rep + Web + Journal
									13%	87	Rep + Web
						33%	421	Journals only	58%	244	Supplemental
									19%	78	Data journals
									24%	99	Both
						6%	78	Webpage only			
						7%	88	Journals & webpages, but no repositories			
Sums	2,031	Total amount of answers (excluding people who skipped the question)		1,549	Indicated sharing in any form		1,272	Shared in an open way (journals, web, repositories)			

Table 5: Summary of the ways of sharing. Here the sub-groups of different ways of sharing data are combined to give an overview.

The number of participants indicating an explicit use of data repositories (institutional, discipline specific and general purpose taken together) is 685/2,031 = 33.7%, much lower than the rate for open sharing (63%) or sharing overall (76%). Comparing sharing rates has to be done with caution, since counting cases is done in different forms across surveys.

Evaluating comments in “other” across the overall dataset, the options “No data exist” and “Difficulties with data sharing concept” were constructed (see chapter 3). Those two options were not chosen by the participants directly, but added afterwards in the analysis to identify criticism in proportion to the overall positive responses.

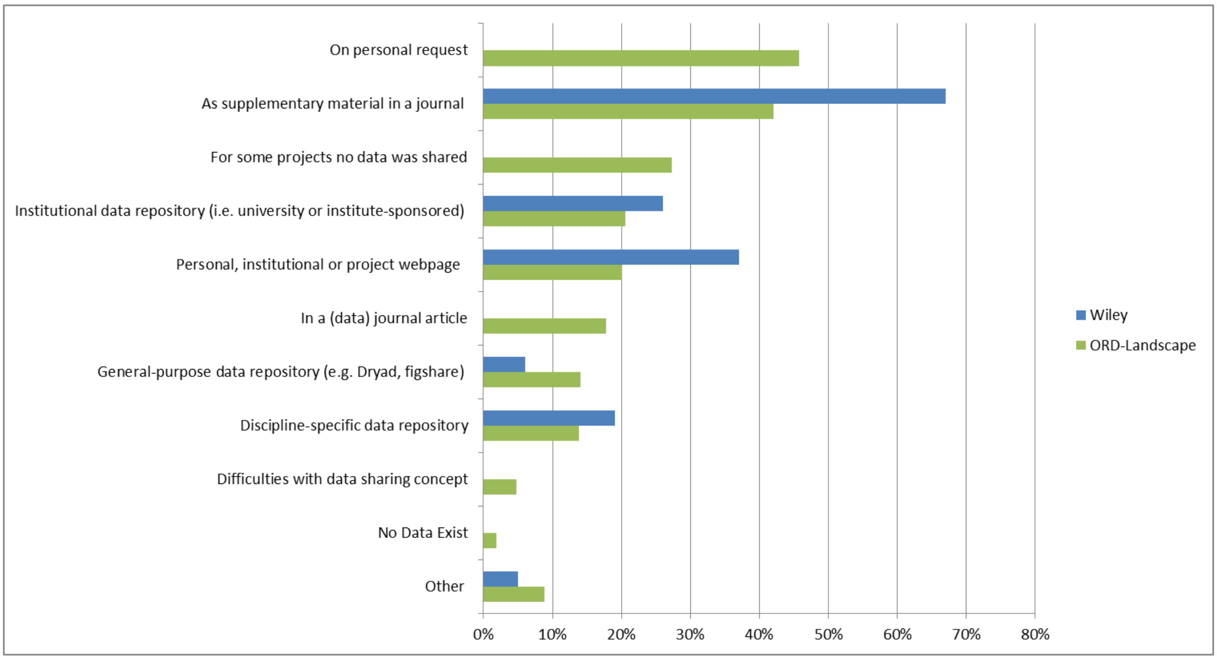


Figure 1: Comparison to Wiley study (L. Ferguson, 2014) on how data sharing is performed.

In sum, the rate of scientists sharing their data is higher than the baseline found by the Wiley survey. The proportion of participants indicating that they publish their research data openly (journal, web and repositories) is 63%.This means about 1/3 of the participants do not share openly. This is consistent with the notion of a visible change in scientists’ behaviours.

4.4 *Why data is not shared*

The most common reason for not sharing was not generating any data. In total, 32 scientists indicated in their comments that they had no data. About 47% of them belong to the Mathematics discipline (n=15), and about 22% belong to Social and Behavioral Science. The largest number of participants in any other discipline who said they did not generate any data was three. Since this option was previously regarded as rare, it was not offered during the survey, which prompted criticism about the methodology.

Overall, 453 participants gave answers to specific reasons why data sharing was not performed during some of their recent projects. About 40% of those chose as their number one answer the current plan to publish the work first. Intellectual property or confidentiality issues were mentioned by about 1/3 of the participants as the second most important reason not to publish. See Figure 2 for all reasons in descending order of importance.

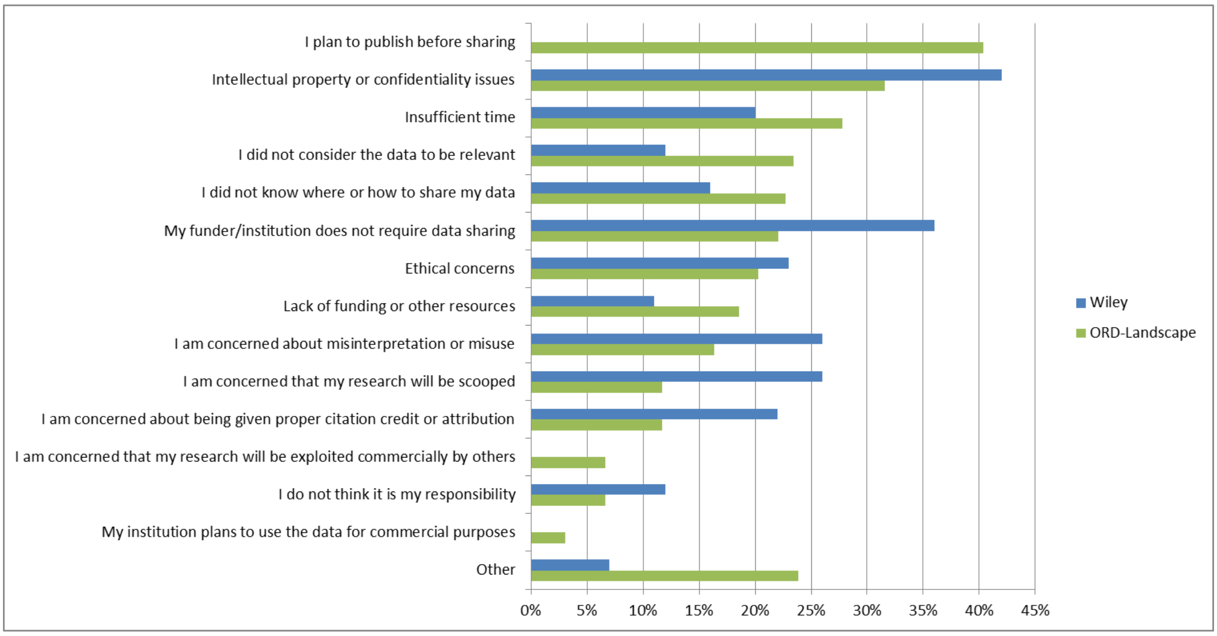


Figure 2: Comparison of reasons not to share data between the Wiley study (Ferguson, 2014) and the current Landscape survey in Switzerland.

Some of the items were rated to be not as important as had been suggested by (L. Ferguson, 2014): scooping of research and the lack of a requirement for sharing by the funding agency show the most prominent differences. The latter is consistent with the SNSF policy which we assumed to be known. However, chapter 9 shows that researchers in Switzerland are not overly familiar with policies.

Conversely, participants judged the data to not be relevant more often in the current survey than in the Wiley study. Participants in our survey rated not to know where to share higher than in the earlier survey. This might be due to a potential self-selection bias in the Wiley survey since participants were potentially using Wiley services.

In many respects, the participants in our survey seem to have other reasons for not sharing data than the participants of the Wiley study (see Figure 2). We therefore assume that either the situation in Switzerland is different for unknown (or not yet evaluated) reasons, the selections by participants in the Wiley survey were different due to a selection bias in the surveys, or the reasons for not sharing have changed over the past four years.

4.5 Differences between disciplines

From common experience, we know that scientists from different scientific disciplines vary considerably in their culture, habits, beliefs, language, concepts, and so on. Previous work has shown also differences for the disciplines in the context of data sharing and reuse (Borgman, 2012; Dallmeier-Tiessen et al., 2014; Eynden et al., 2016; Fecher, Friesike, Hebing, Linek, & Sauermann, 2015; L. Ferguson, 2014; Pasquetto et al., 2017; Tenopir et al., 2011, 2015; Wallis et al., 2013).

However, only a few surveys have actually looked into the differences on a detailed level and at the overall scientific community at the same time. Our focus within the Swiss community was aiming precisely at both views simultaneously: i.e., performing analyses on all disciplines at once and also on a very detailed level.

4.5.1 The bepress catalogue of disciplines

As described in the data paper (von der Heyde, 2019b), the landscape survey assessed the disciplines along the bepress taxonomy of disciplines (Warner, 2018).

This three-tiered taxonomy offers 10 categories (f=10) on the top level, which served as entry points to the second (s=363) and third levels (t=881) in the survey. Since not all s-categories offer sub-categories, the true number of leaves (l=1,049) in the graph is not equivalent to the third level. In other words, the leaves of the graphs consist of level two and level three categories, depending on the existence of sub-categories of level two.

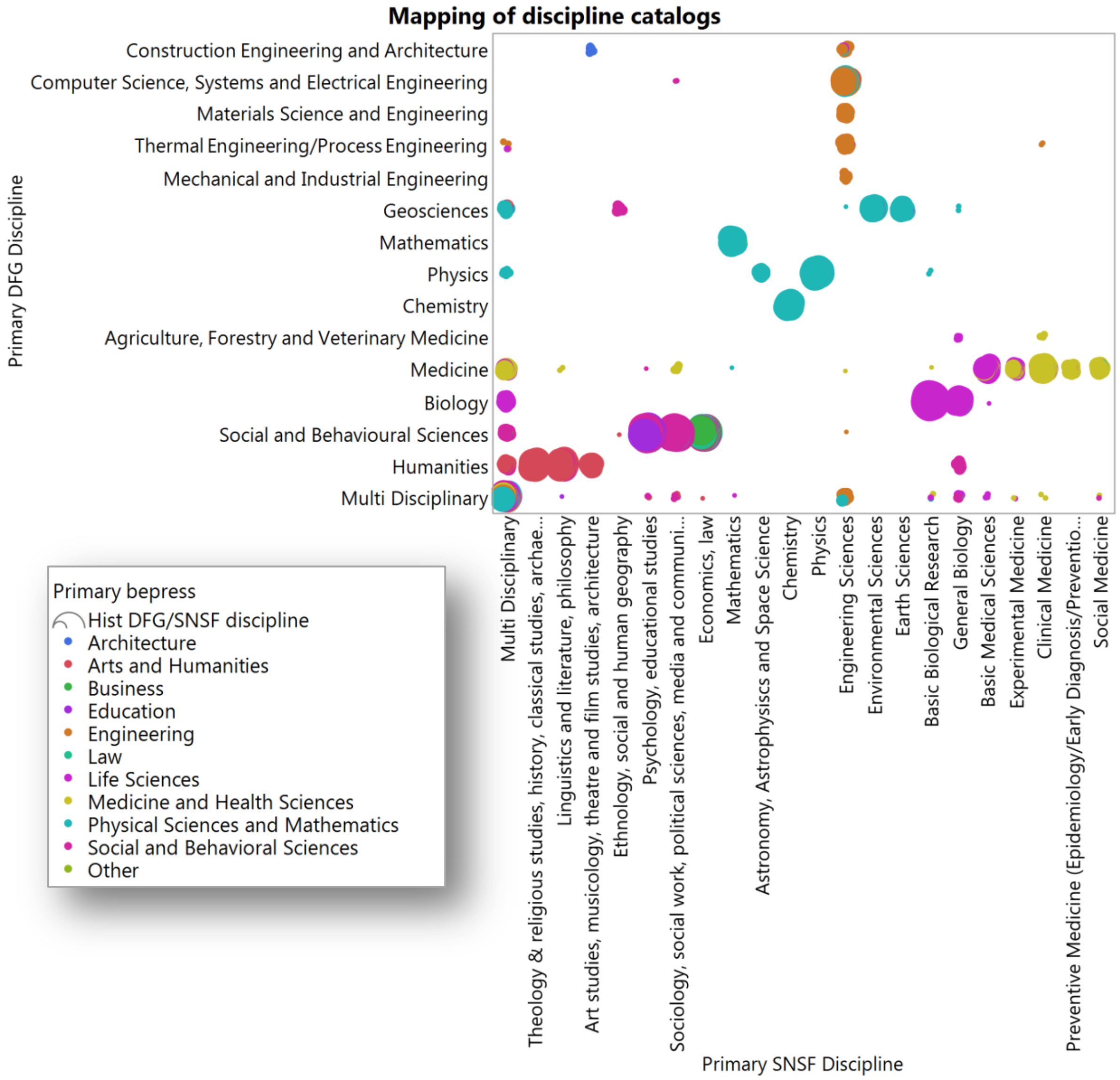


Figure 3: Mapping of bepress categories to DFG and SNSF disciplines (also see Table 6).

4.5.2 Mapping of disciplinary catalogs

As described earlier, the need to group disciplines remains even in our approach. Due to statistical demands, we needed to gather groups of at least 100 participants to ensure significant results. Groups between 30 and 100 already suffer by a loss of statistical power; below 30, we should not infer any results. The highly diverse bepress taxonomy did enable us to map the survey to multiple constellations matching other demands, like the re3data base system (equivalent to the DFG system) or the SNSF system.

Both discipline mappings - DFG and SNSF - are not optimal, as can be seen in Figure 3. For example, the DFG category “Medicine” is mapped into five sub-categories at SNSF. Conversely, the category “Engineering Sciences” of the SNSF catalog is split into five categories in the DFG mapping. The mapping of the “Humanities” as well as the “Social and Behavioral Sciences” also differs considerably between the two systems.

The number of participants with their primary discipline mapping to the DFG or SNSF category is given in Table 6. Most categories contain a sufficient number of cases, confirming the effectiveness of the mapping procedure. The described mapping is used in the following sections to categorize the findings across the different disciplines.

		PrimDFGDiscipline																
PrimSNSFDiscipline	SNSF Key - Level 2	Mixed Disciplines	Humanities	Social and Behavioural Sciences	Biology	Medicine	Agriculture, Forestry and Veterinary Medicine	Chemistry	Physics	Mathematics	Geosciences	Mechanical and Industrial Engineering	Thermal Engineering/Process Engineering	Materials Science and Engineering	Computer Science, Systems and Electrical	Construction Engineering and Architecture	not used	Sum
DFG Key - Level 2			11	12	21	22	23	31	32	33	34	41	42	43	44	45		
Mixed Disciplines		103	29	22	30	47			8		23		4					266
Theology & religious studies, history, classical studies, archaeology, prehistory and early history	10100		119															119
Linguistics and literature, philosophy	10200	1	119			2												122
Art studies, musicology, theatre and film studies, architecture	10300		55													7		62
Ethnology, social and human geography	10400			1							13							14
Psychology, educational studies	10500	3		178		1												182
Sociology, social work, political sciences, media and communication studies, health	10600	4		166		7									2			179
Economics, law	10700	1		149														150
Mathematics	20100	1				1				90								92
Astronomy, Astrophysiscs and Space Science	20200								26									26
Chemistry	20300							97										97
Physics	20400								128									128
Engineering Sciences	20500	19		1		1					1	10	29	28	97	16	1	203
Environmental Sciences	20700										75							75
Earth Sciences	20800										55							55
Basic Biological Research	30100	3			182	1			2									188
General Biology	30200	9	16		86		5				2							118
Basic Medical Sciences	30300	3			1	75												79
Experimental Medicine	30400	2				37												39
Clinical Medicine	30700	2				92	5						2					101
Preventive Medicine (Epidemiology/Early Diagnosis/Prevention)	30800					26												26
Social Medicine	30900	2				40												42
not used			4			4											12	20
Sum		153	342	517	299	334	10	97	164	90	169	10	35	28	99	23	13	2,383

Table 6: Number of participants per DFG and SNSF discipline mapping (also see Figure 3).

4.5.3 Different ways of sharing

To statistically evaluate if the ways of sharing are significantly different between disciplines, a categorical analysis was performed (the complete analysis is part of Appendix E: Categorical analyses). In addition, a filtered analysis with only the disciplines containing more than 100 datasets was performed. Both reach the same result: Only ‘use of institutional repositories’ was not significantly different across all disciplines. This holds for both DFG and SNSF systems. Pooling the data across methods (see section 4.6) excluded further significant differences. The ‘personal sharing’, ‘sharing no data’ and ‘sharing of data in data journals’ options were not significant if seen from the pooling by primary methods. On the other hand, ‘sharing in discipline-specific repositories’, ‘sharing in general purpose repositories’, ‘sharing on webpages’, ‘providing supplementary material’, and having no data or having a different concept of data are significantly different across all selections of the overall dataset.

When we first distribute the shares of all mentioned scientific disciplines per person and then pool all data across all disciplines, and do not limit ourselves to the “primary” ones in either of the systems, we can derive a new perspective on the overall dataset.

In this constellation, it is important to look at the pattern of the differences, especially for the variables which were prominent in the categorical analysis. In the following graphs, the differences between disciplines for the variables describing the sharing and reuse behaviours are depicted for Y=disciplinary and X=general purpose repositories.

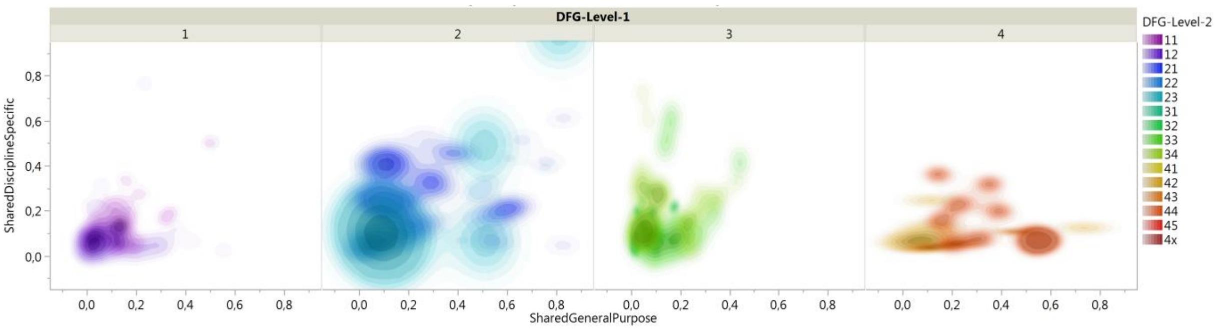


Figure 4: Ways of sharing depicted for the frequency of sharing in disciplinary and general purpose repositories, sorted by the DFG classification. Disciplines with zero or one participants were excluded to remove the risk of extreme outliers. Please use Table 6 as reference for the DFG-Level-2 names. Additional data within a Level-1 category is marked with “x”.

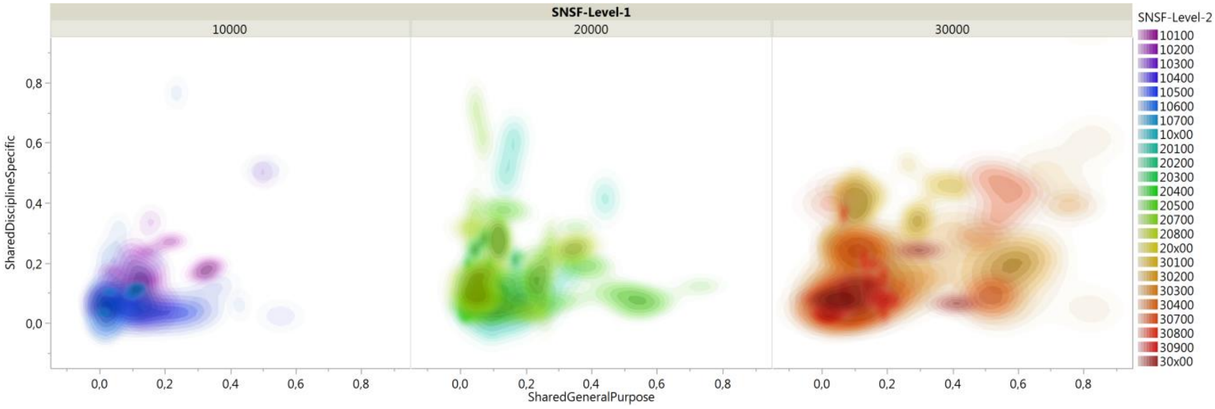


Figure 5: Ways of sharing depicted for the frequency of sharing in disciplinary and general purpose repositories, sorted by the SNSF classification. Please use Table 6 as reference for the SNSF-Level-2 names.

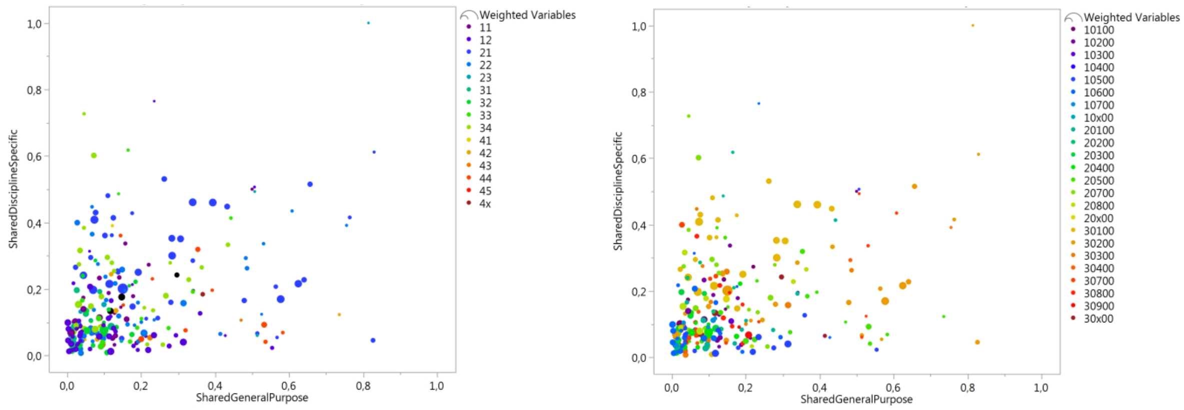


Figure 6: Raw data of all disciplines scaled by the number of participants having indicated working in this discipline. On the left, we show the DFG mapping; on the right, the SNSF mapping. Please use Table 6 as reference for the SNSF- and DFG-Level-2 names.

The overall pattern of the natural sciences (in green, DFG 3 and SNSF 20000) is quite similar in Figure 4 and Figure 5. The differences in the disciplinary mapping are most prominent for the engineering sciences: The DFG 4 disciplines are mixed into the SNSF 20000.

Please note: As we cannot visualise actual differences within the variables in the categorical analysis, which only takes the primary discipline into account, these graphs show all data of the ~900 disciplines (1,243 – 338 excluded (zero or one participant) = 905). The visual grouping of the DFG classification helps for the general understanding of the data.

4.5.4 Some do and some don't[10]

Taking the same disciplinary mappings from section 4.5.3, we can look at the variables of sharing frequencies across disciplines. Again, we limit ourselves to disciplines having at least two participants.

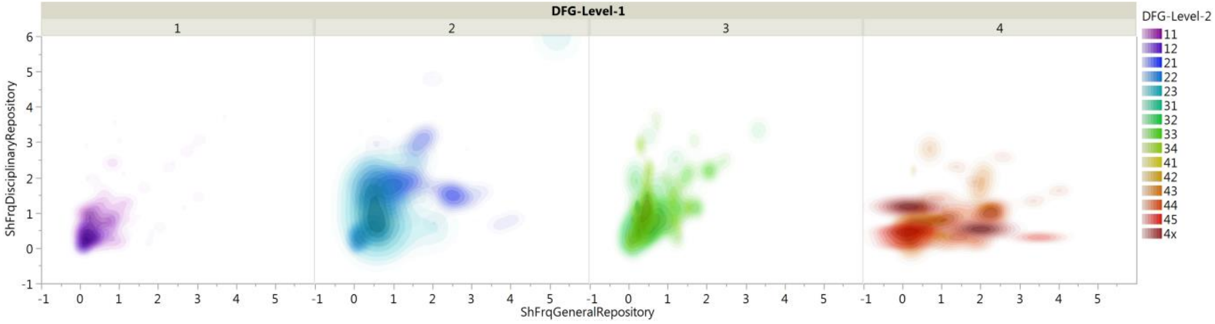


Figure 7: Sharing behavior across disciplines. Shown is the frequency of sharing in disciplinary repositories over the frequency of sharing in general purpose repositories. The mapping is shown here for the DFG catalog. Please use Table 6 as reference for the DFG-Level-2 names.

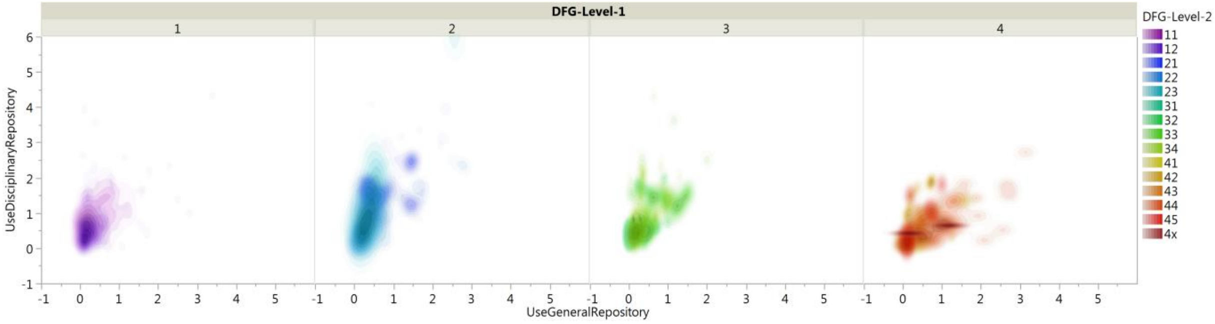


Figure 8: Reuse behavior across disciplines. Shown is the frequency of reusing other researchers' data from disciplinary repositories over the frequency of reusing data from general purpose repositories. The mapping is again shown for the DFG catalog. Please use Table 6 as reference for the DFG-Level-2 names.

Both graphs show a clear difference between the disciplines in the pattern for sharing frequency in disciplinary and general purpose repositories (see Figure 7) as well as the same schema for the reuse of data (see Figure 8).

In sum, it can be observed that sharing in the humanities and social sciences (DFG1) is not yet as common as in the other disciplines. In all cases, the tendency to share in general purpose repositories is not mirrored by the reuse of data from these repositories: Here the pattern is elongated in the direction of the disciplinary repositories. This indicates that even if disciplines tend to share in general purpose repositories, the actual reuse happens within the disciplinary repositories. This is most prominent for the life sciences (DFG2), but also clearly visible for the engineering sciences (DFG4). A reason might be that staff in data specific repositories more often curates data (see section 7.1).

[10] A.A. Milne: *Winnie-the-Pooh*, Chapter V.

4.5.5 Are scientists aware of the repositories?

One factor in sharing and reusing is the perceived readiness to share. Often technical barriers appear to disable sharing and amplify other reasons (those shown in Figure 2) in their perceived effect.

Would an average researcher read more journals than they publish in? Analogously, would a researcher use data from more repositories than she or he publishes data in?

Currently, researchers perceive a higher availability of repositories for sharing than for reuse. Figure 9 shows considerably more participants below the diagonal of the diagram.

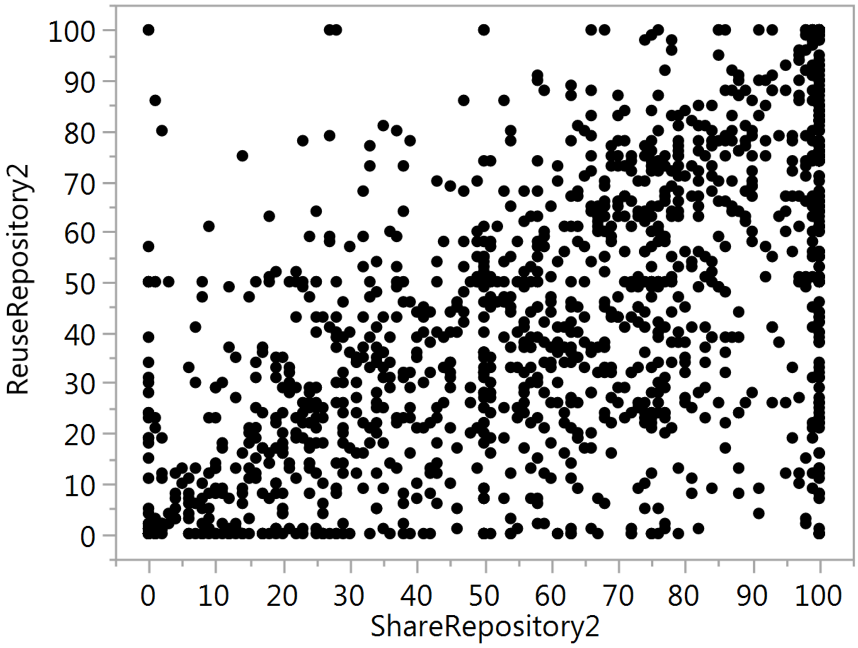


Figure 9: Judgements on the availability of repositories for sharing and reuse.

In Figure 10, scientific disciplines' points of view can be seen to differ considerably. Interestingly, the difference between sharing and reuse for Astronomy or Basic Biological research is very much smaller than the ones for Social Medicine or Psychology. Reasons why disciplines differ like this could be subject to further study.

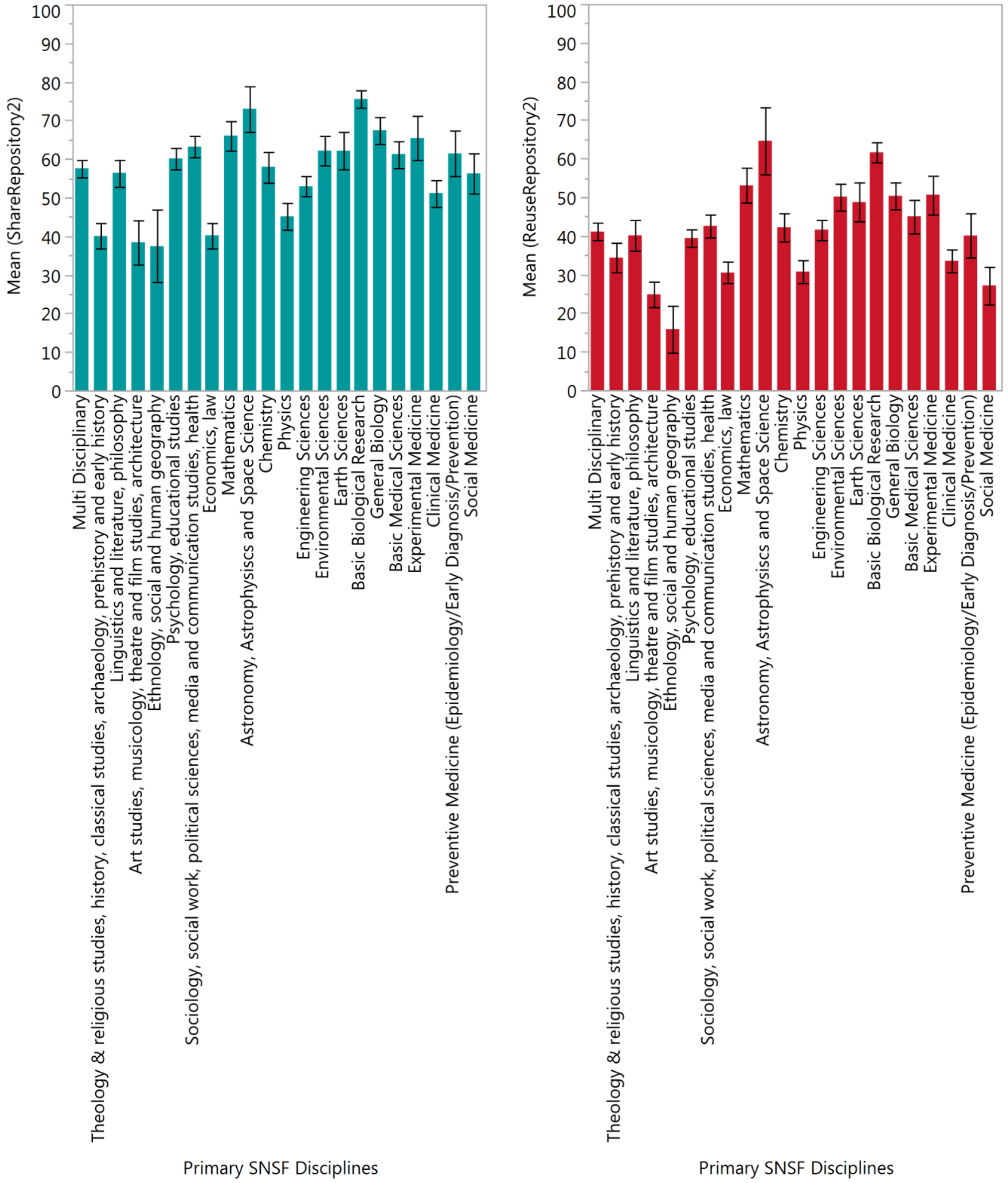


Figure 10: The perceived availability of repositories known within specific disciplines for sharing and reuse. Sharing is depicted on the left and reuse on the right.

4.6 Differences between scientific methods used

The project hypothesized that the differences seen between the disciplines could be better understood on the basis of the methods used. Therefore, the survey assessed the methods in a detailed but standardized schema. Participants were asked to select their primary methods from a catalog of methods. The overall number of answers per method is depicted in Figure 11.

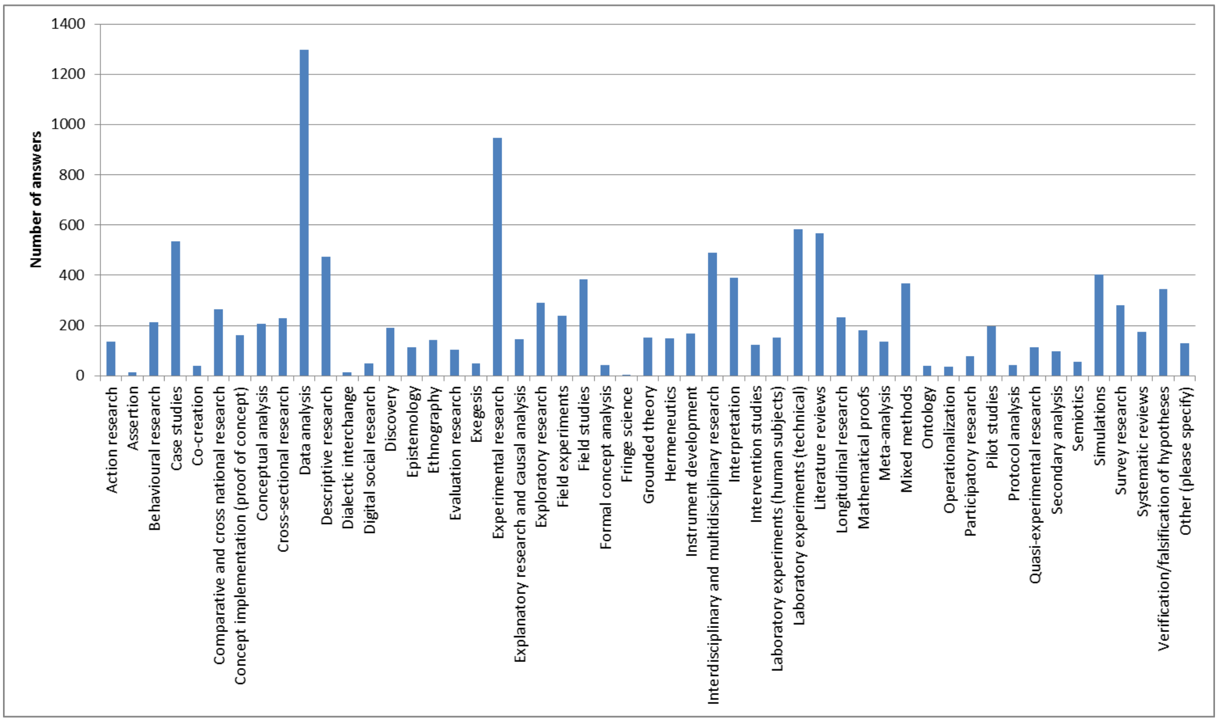


Figure 11: Overview of all selected methods in the landscape survey.

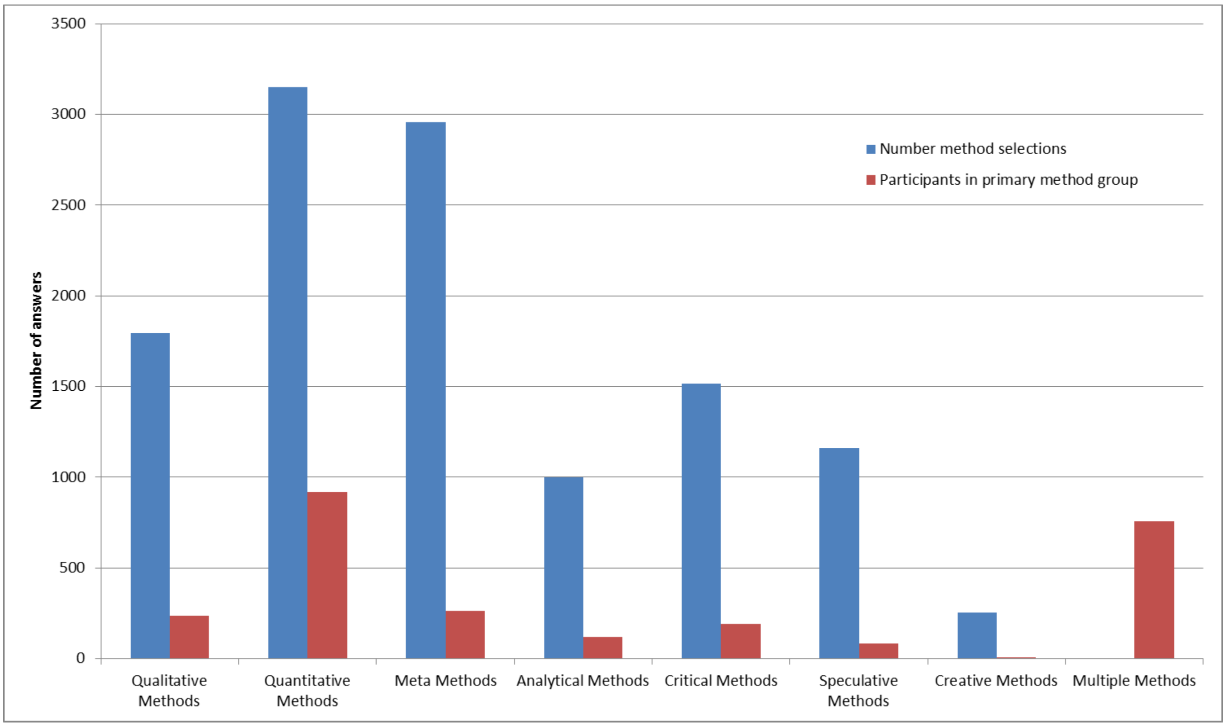


Figure 12: Grouping of the scientific methods according to abstract principles, and grouping of the participants by their primary research method.

After grouping the methods in an abstract classification, we derived the constellation depicted in Figure 12. Since the use of 'data analysis' was most prominent, all attempts to pool data would be imbalanced.

After this grouping, the primary method of each participant was determined on the basis of the major proportion of the selected methods in the abstract classes. Since 'mixed methods' is also a category in the Meta Methods, any balanced case was applied to a new virtual class (Multiple Methods).

Taking the primary method as a categorical variable, we can again perform a statistical analysis, details of which are given in Appendix E: Categorical analyses. The statistical results confirm the homogeneous use of institutional repositories across methods. Also, the differences in 'personal sharing', 'sharing no data', and 'sharing of data in specific data journals' proved to be not significant across methods. Differences remain between methods concerning sharing via 'supplementary material', 'general-purpose repositories', 'webpages', 'discipline-specific repositories', 'institutional repositories', 'personal requests', and 'Journal articles'. The variables describing 'other', 'no data exist', 'not applicable' and 'shared no data' were also different across the method groups.

5 Mapping of repositories

Using data repositories for data sharing is not yet widespread throughout the Swiss research community. Less than 10% of survey participants named repositories they were using. This also corresponds with insights from the text evaluation of the general comments. Some people admitted to having to learn more about data sharing and reuse to be competent. The list and frequency distribution of repositories is published in the data paper for the Landscape survey (von der Heyde, 2019b). The distribution of participants naming repositories was reflective of the number of participants in each discipline. No discipline contained an overly large proportion of the people naming repositories (see Figure 13).

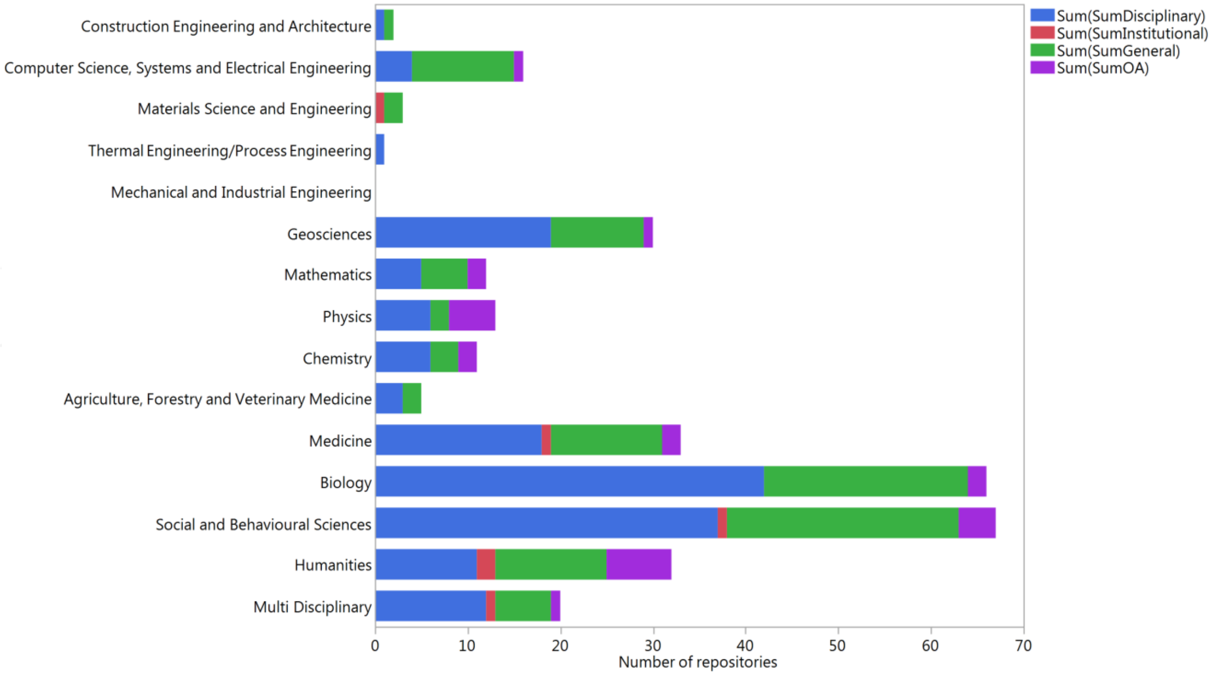


Figure 13: Distribution of repositories named by each primary DFG discipline. Repositories are categorized by type.

The distribution of the frequencies across DFG disciplines is depicted in Figure 14 for all repositories mentioned at least twice in the survey. The repositories mentioned once are summarized as 'others'. The equivalent for the SNSF system is shown in Figure 15. The repositories used by participants are shown according to naming frequency; general purpose repositories and disciplinary ones are mixed. To gain a complete overview, a display of all repositories would be necessary. The other 97 repositories were only mentioned once by 67 participants in total.

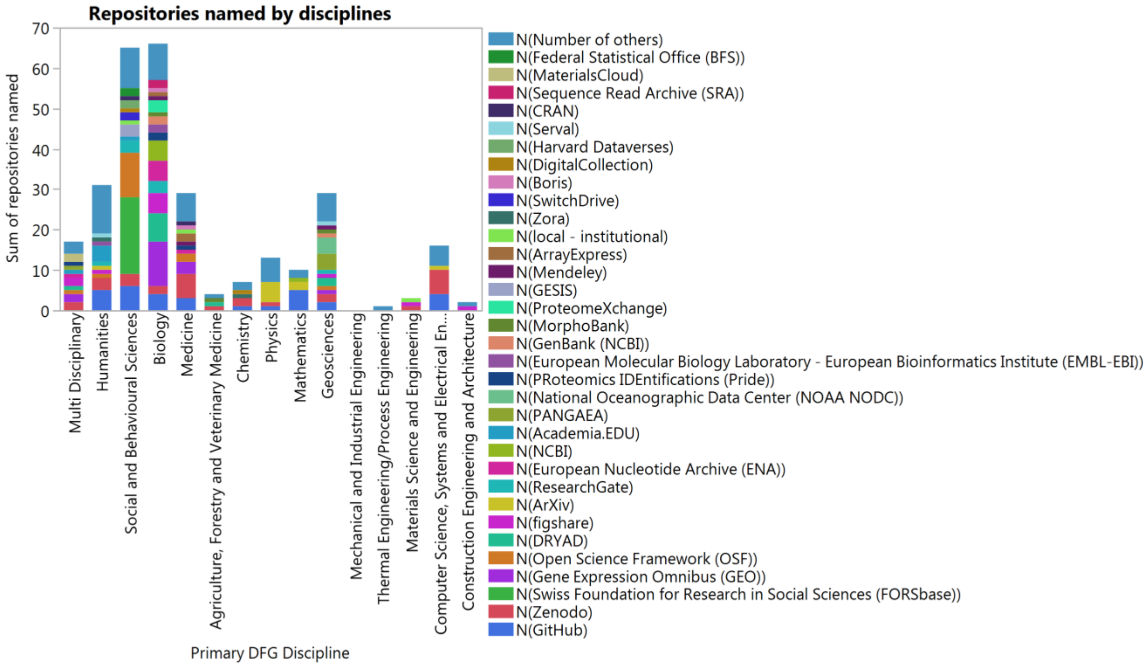


Figure 14: Distribution of repositories named by participants across the DFG disciplines.

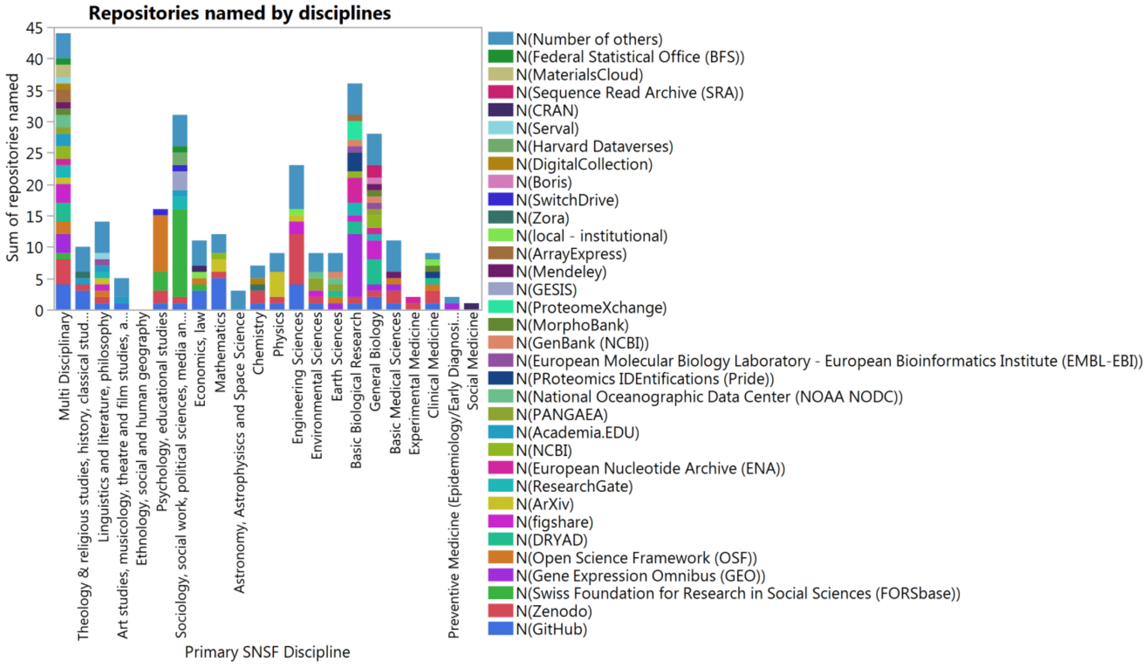


Figure 15: Distribution of the repositories named by participants across the SNSF system.

Further analysis and mapping to known repositories in re3data revealed the naming of library repositories for Open Access (OA) as well as several websites which provide data for one research group without a formal repository. In Figure 16 the types of repositories and other data sharing options are shown in relation to the number of different names given. For readability, the term “repository” is used in the figures despite some answers not referring to a repository.

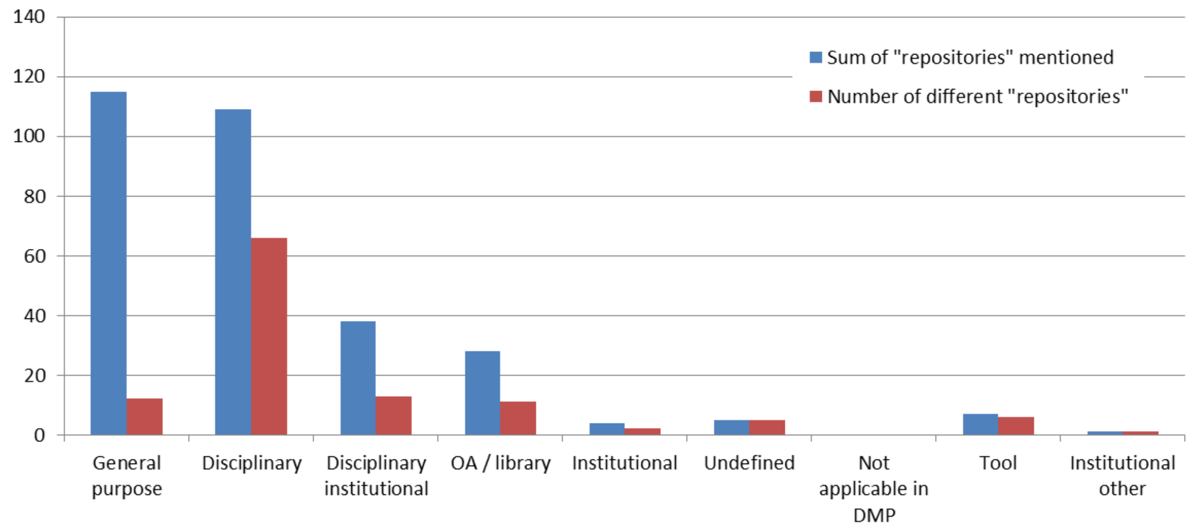


Figure 16: Frequency of mentioned repository types or other places for data sharing in the landscape survey. The left-hand bars display the number of times an entry was mentioned. The right-hand bars show how many different repositories were named.

As expected, the general repositories are far more frequently named per repository than the disciplinary ones. The same analysis was performed on the repositories mentioned in the data management plans (DMPs) when institutions or projects wrote their funding applications to the SNSF (Figure 17).

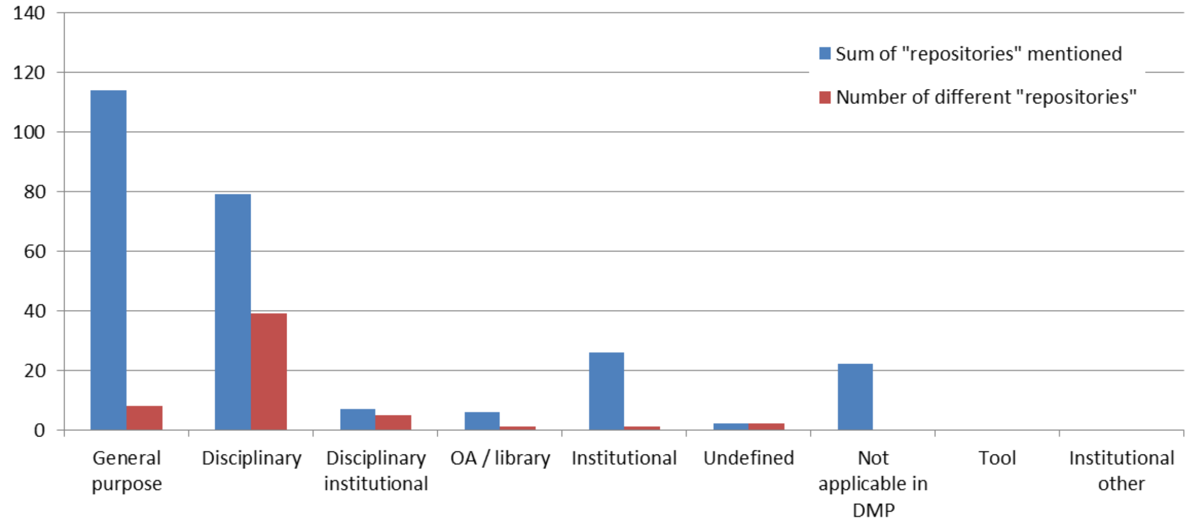


Figure 17: Classification of the repositories named in the DMPs submitted the SNSF in the last two years.

A complete list of the three biggest groups (disciplinary repositories, general repositories and OA libraries) can be seen in Figure 18. Due to their localization in the Swiss community, Zenodo and FORSbase are mentioned more frequently.

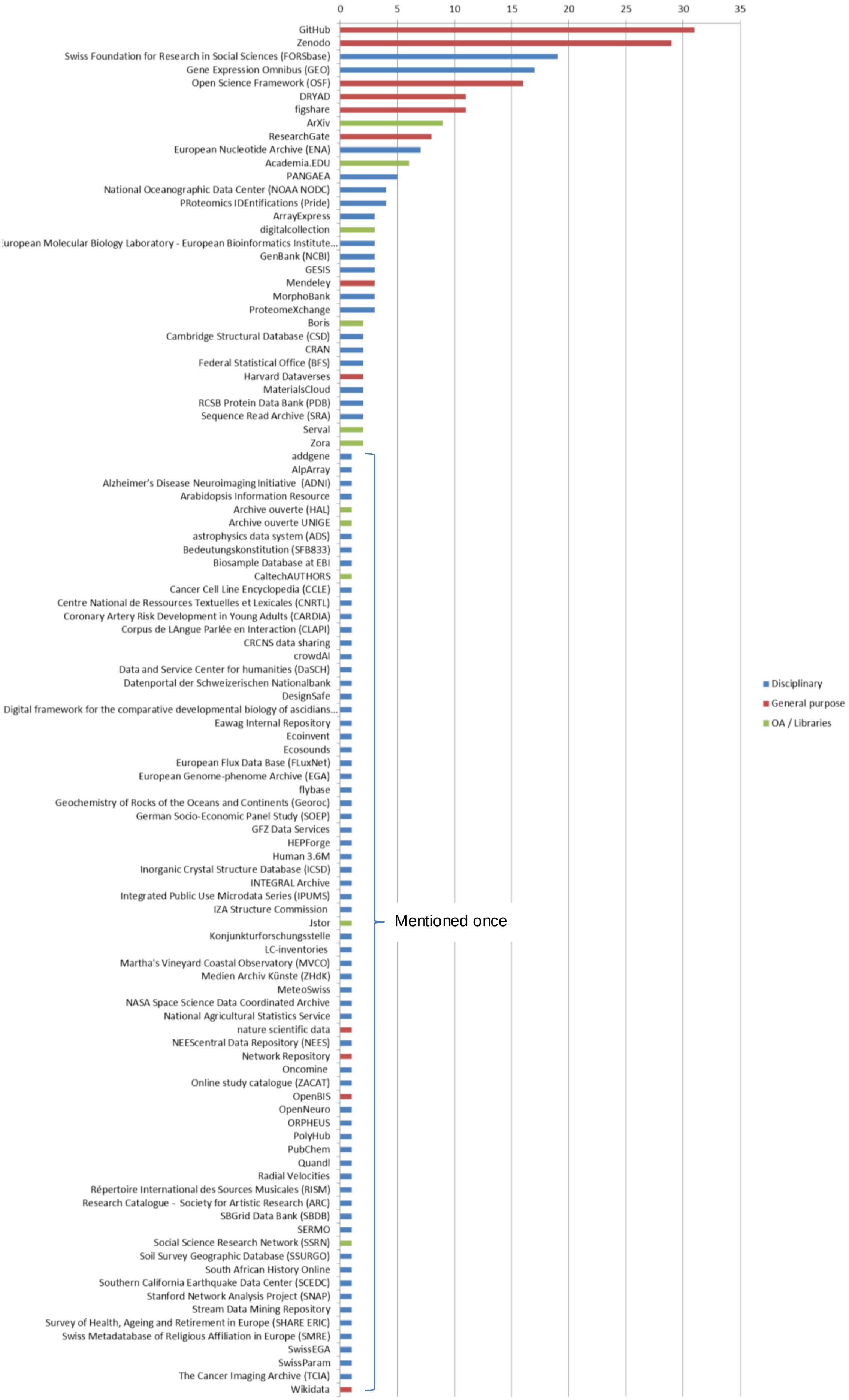


Figure 18: Frequency distribution of repositories and OA libraries in the landscape survey.

Using re3data, we collected the number of repositories connected to institutions located in Switzerland. The analysis reached the following results:

- 61 repositories are connected to Swiss institutions, including 21 without international participation. Six are based only in one institution.
- 32 repositories refer to 27 Swiss organisations for funding (including 5 universities).
- 53 repositories refer to 42 Swiss organisations not for funding, but for general or technical support.

5.1 Combination of DMP results and landscape survey

In comparing the general results from the repositories in the DMPs and the repositories mentioned in the landscape survey, we observed a similar pattern of responses, but also a number of repositories which were only mentioned in one of the two sources (see Figure 19). We therefore combined both sets for a more complete picture (Figure 20).

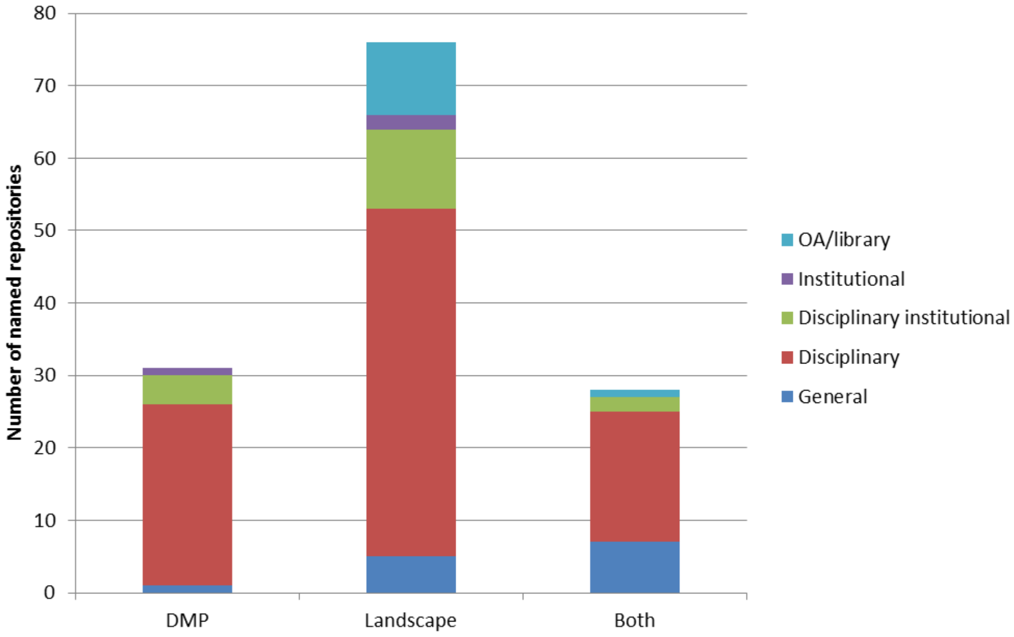


Figure 19: Number of repositories named in the landscape survey and in the DMPs; overlap is summarized as "both".

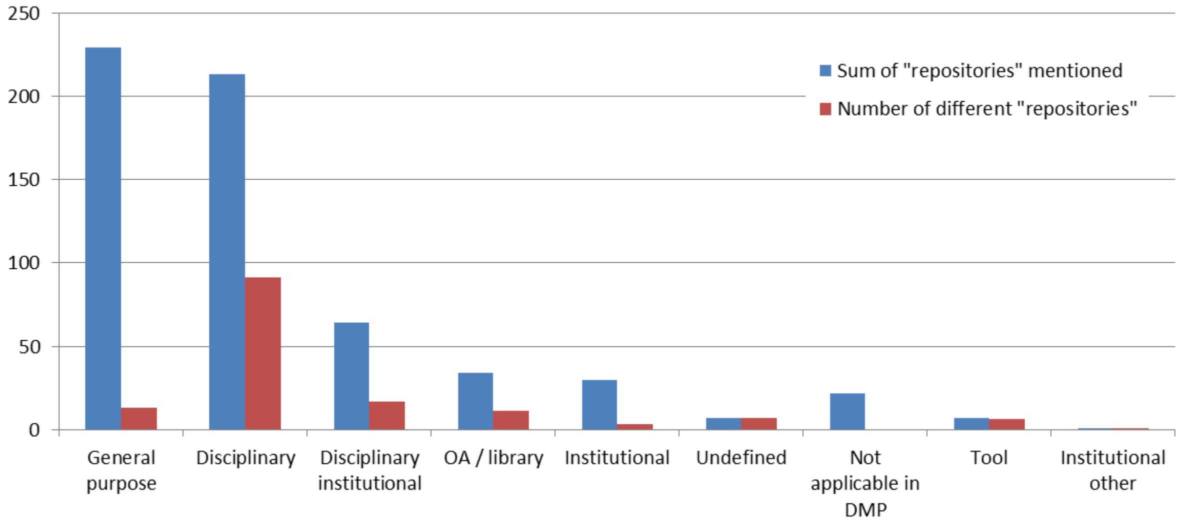


Figure 20: Combined dataset of repositories from the DMPs and the landscape survey.

	Sum of repositories mentioned	Number of different repositories	Ratio of sum and number
General	229	13	17.6
Disciplinary	213	91	2.3
Disciplinary institutional	64	17	3.8
OA / library	34	11	3.1
Institutional	30	3	10.0
Not applicable in DMP	22		
Other	15	14	1.1
Sum	607	149	4.1

Table 7: Summary of all repositories in DMPs and the landscape survey, split into types of repositories following re3data classification.

In Table 7 we have summarized the count and sums of both analyses. The ratio between the sum of mentioned repositories and the number of times each repository was mentioned serves as a rough estimate of the use intensity. Therefore, on average in every disciplinary repository only two to four Swiss scientists share or reuse data. General purpose repositories have a much higher use intensity. In sum, the fragmentation of the repository landscape is pronounced.

5.2 *Use of international repositories*

While the institutional and disciplinary institutional repositories used by the Swiss community are naturally located in Switzerland, nearly all general and many disciplinary ones are located elsewhere (Figure 21).

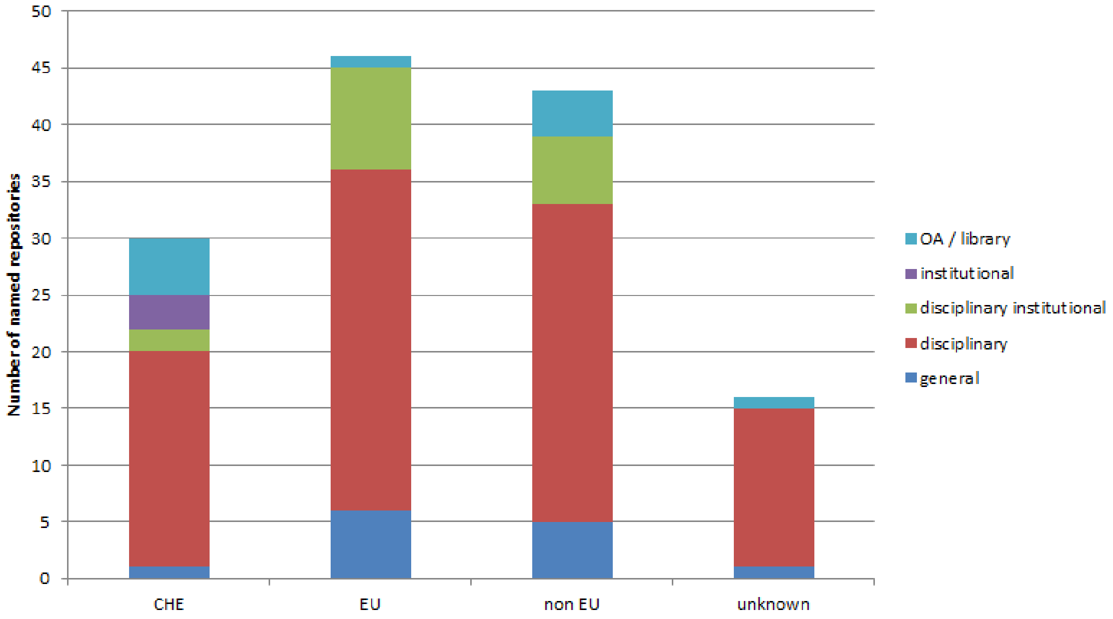


Figure 21: Location of the 131 repositories named in the landscape survey and in the DMPs. Repositories having at least one institutional responsibility in Switzerland are classified as CHE (accordingly for EU).

5.3 *Summary of repository mapping*

We summarize the repository usage for the Swiss research community as follows:

- In comparison to the low number of researchers who actually mentioned a specific repository, the diversity of repositories is surprisingly high. About 200 scientists named

100 different repositories. Some repositories were mentioned several times, resulting in 300 responses. In short, nearly every scientist has their favorite repository and the overlap is limited to the general purpose repositories. The exception is FORSbase.

- The use intensity of general purpose repositories is much higher than the use intensity of institutional or disciplinary repositories.
- The Swiss research community uses international repositories extensively: Switzerland is one of the institutional partners for only 25% of the repositories mentioned. The other 75% are represented within the EU or internationally.
- Switzerland provides institutional or financial support for 13% of all repositories mentioned.

6 Future requirements

Following the ideas of (Goldstein, 2017), we allowed participants to select areas in which future repository services should be improved. Since the categories gathered by Goldstein and colleagues were rather abstract, a high level of competency was required to translate the everyday demands of scientists into the choices offered. Consequently, the rate of answers was low since no detailed description was offered during the surveys.

6.1 Service requirements in the landscape survey

About 2/3 of the participants rated the offered service categories on a continuous scale between “not important=0” to “very important=100”. For further information on what the categories refer to, see details in (Goldstein, 2017). The average rating per discipline (DFG) across all services is depicted in Figure 22.

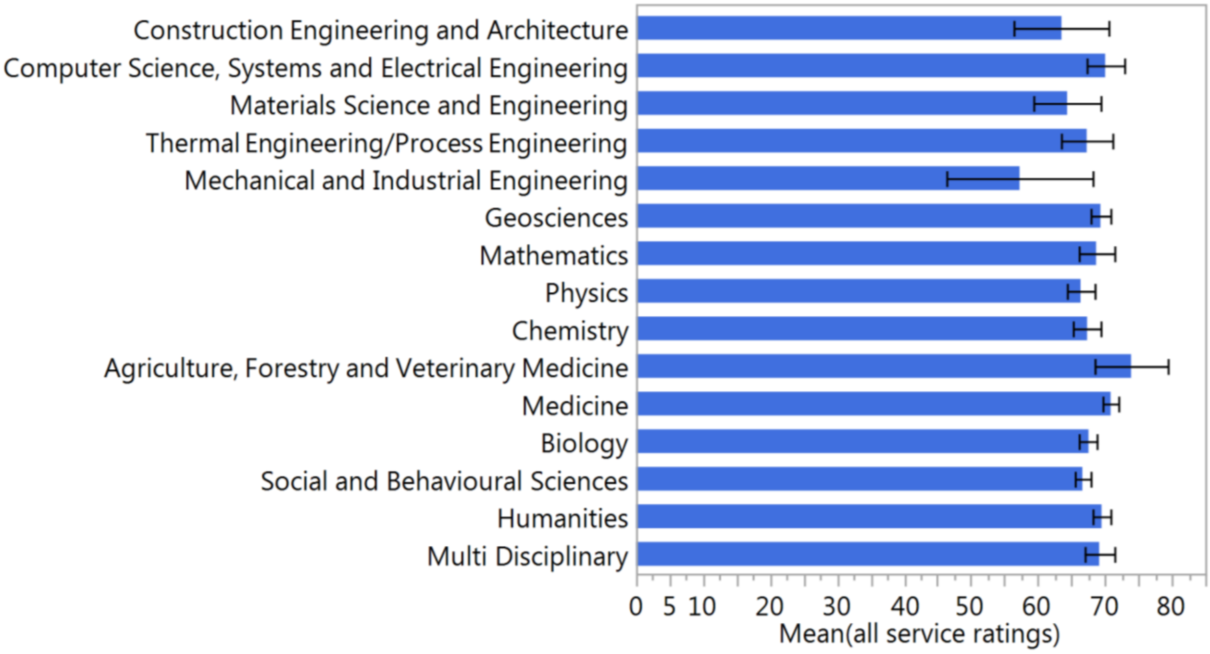


Figure 22: Average ratings of the need for future service per DFG discipline.

As expected, we observe quite some variation between participants from different disciplines. See Figure 23 for a display of the ratings for particular services following the DFG categories of disciplines.

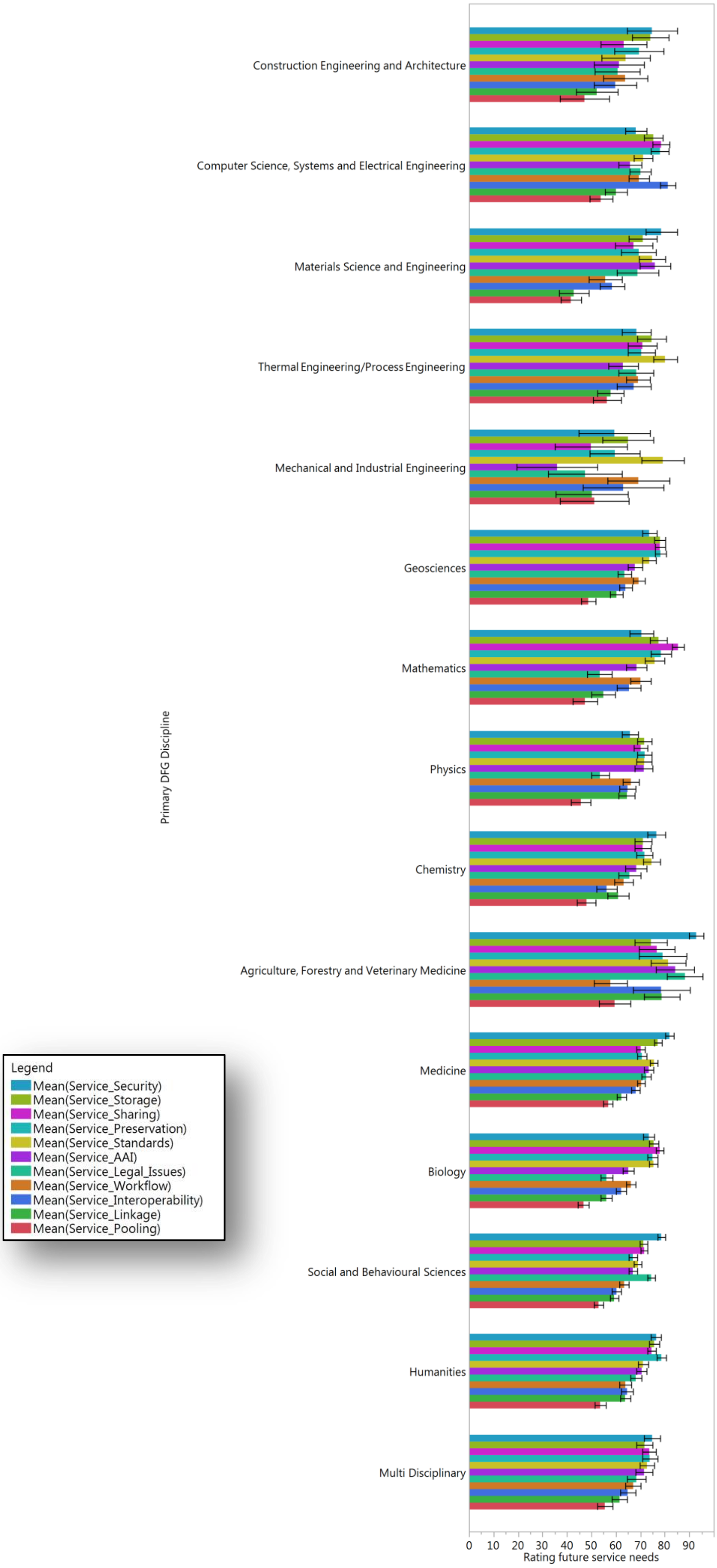


Figure 23: Service requirements for categories rated by participants from different disciplines.

Since the order of services is sorted by increasing means in Figure 23, we can easily detect differences of the individual discipline from the mean. For example, the demand for 'interoperability' is very pronounced for Computer Science and Electrical Engineering disciplines. Other examples are the demand from the Social and Behavioral Sciences for 'legal support', as well as the demand from Chemistry for 'linkage'. Statistical tests can support the observed findings if required.

Further, we observed an interesting difference between the ratings from men and women. In Figure 24 we show the same service ratings split for gender. The number of participants in 'other' or 'prefer not to disclose' was below 5%; the error bars reflect this low number. The main difference between men's and women's ratings can be observed in the categories of legal issues and security.

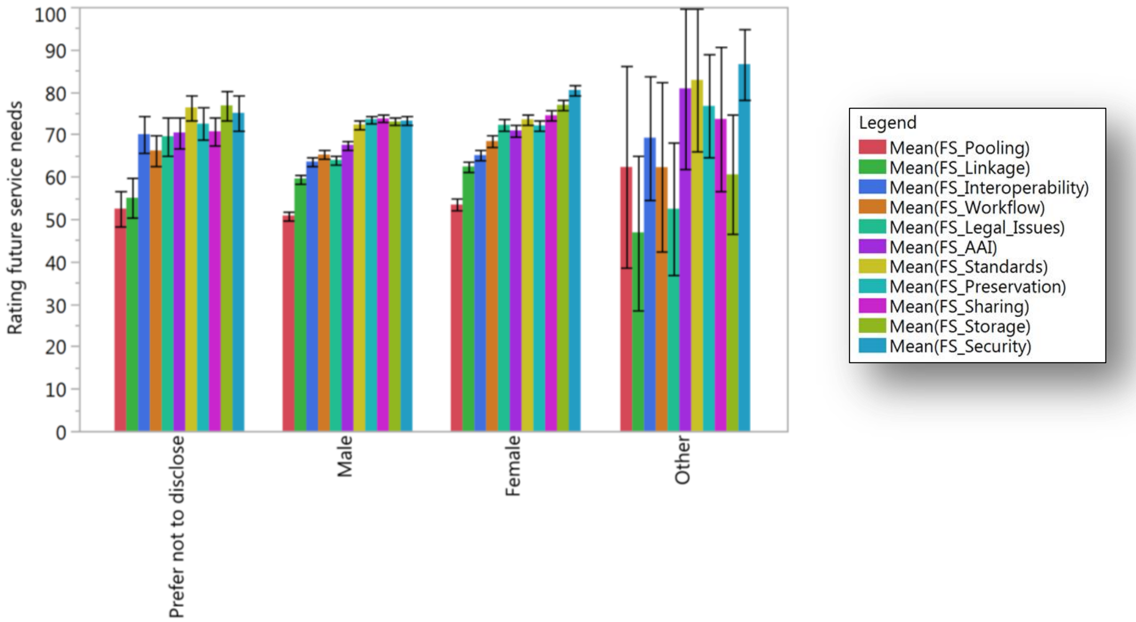


Figure 24: Future service requirements split for gender.

A deeper analysis shows a multi-level system of variables contributing to the situation. In Figure 24, the differences in rating of importance for the categories 'legal issues' and 'security' are depicted by gender, age and method.

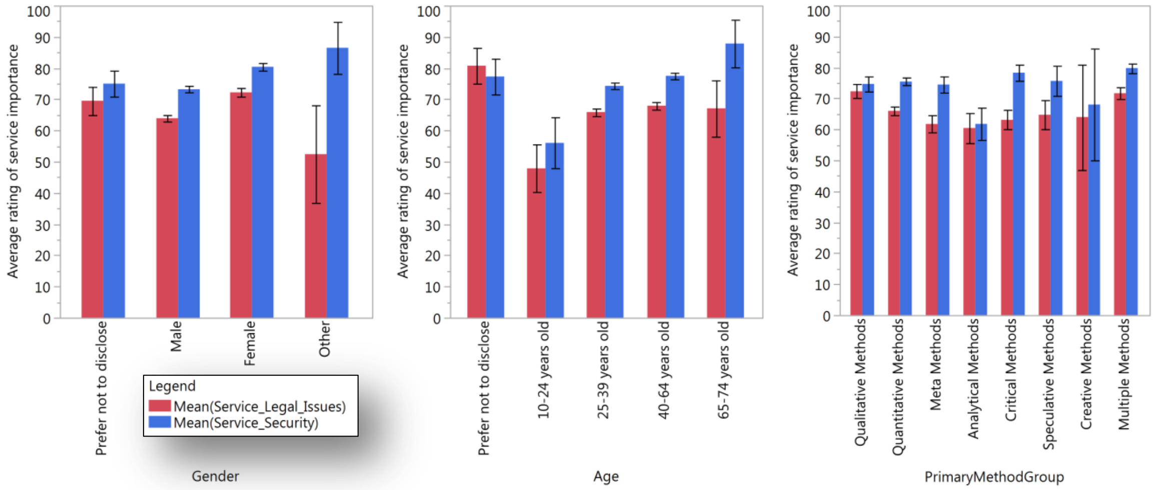


Figure 25: Graphs for the observations for two of the future service demands, 'legal issues' and 'security', split by different groupings. From left to right: gender, age, and primary research method.

Whereas female participants rated the future service for security and legal issues generally higher than their male colleagues, there is also a clear difference across age groups: The demand for both services increases with seniority. As expected, we observed main differences between the research methods (see section 4.6 for the description of categorization): qualitative methods show the least difference, but a high level of requirement for both categories. The demand decreases for the disciplines using quantitative-, meta-, and analytical methods for the legal issues, but maintains a high level for the security rating. Overall, the highest consistent ratings for security are in the critical methods group.

6.2 Service requirements in the repository survey

The repository survey asked the maintainers of repositories to rate the importance of the topics for the future development of their service; the phrasing was slightly different, but the same scale was used.

In Figure 26 the expected development of the repository services depending on their funding status is shown. The major differences for the funded repositories are seen in legal issues, AAI, storage and security.

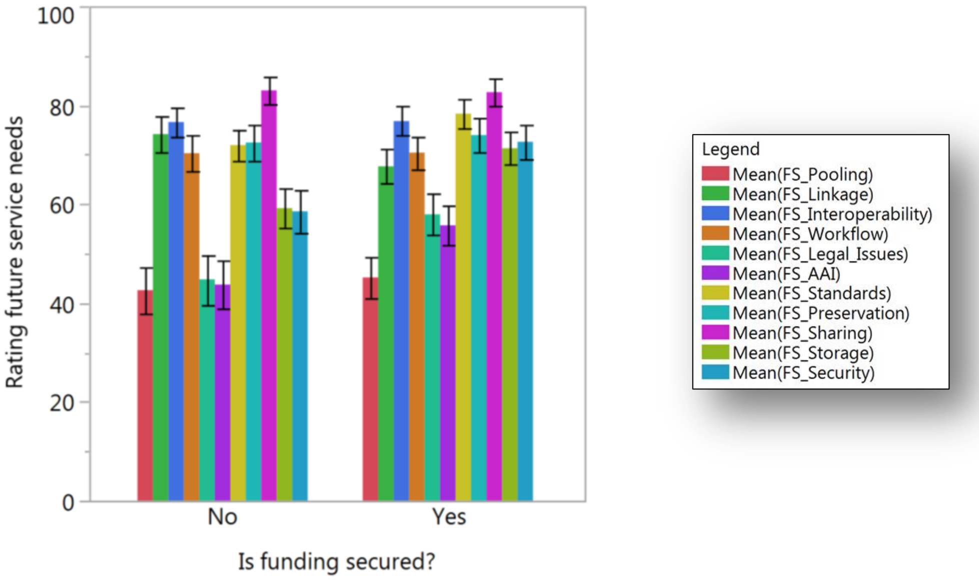


Figure 26: The future service development for the repositories depends partly on secured funding.

The service topics are in the same order as in Figure 23 and Figure 24. Therefore, we can easily see discrepancies between the users' expectations and the repositories' view on services. The categories of linkage and interoperability appear most prominent, and are rated as more important than workflow and ahead of the users' requirements. The ratings for legal issues, AAI, storage and security are below those from the user requirements.

In correspondence to the disciplinary approach to distinguishing service needs, we plotted service development across the bepress primary selection for the disciplinary repositories only. Figure 27 shows the future service needs of 91 repositories grouped by their primary bepress discipline.

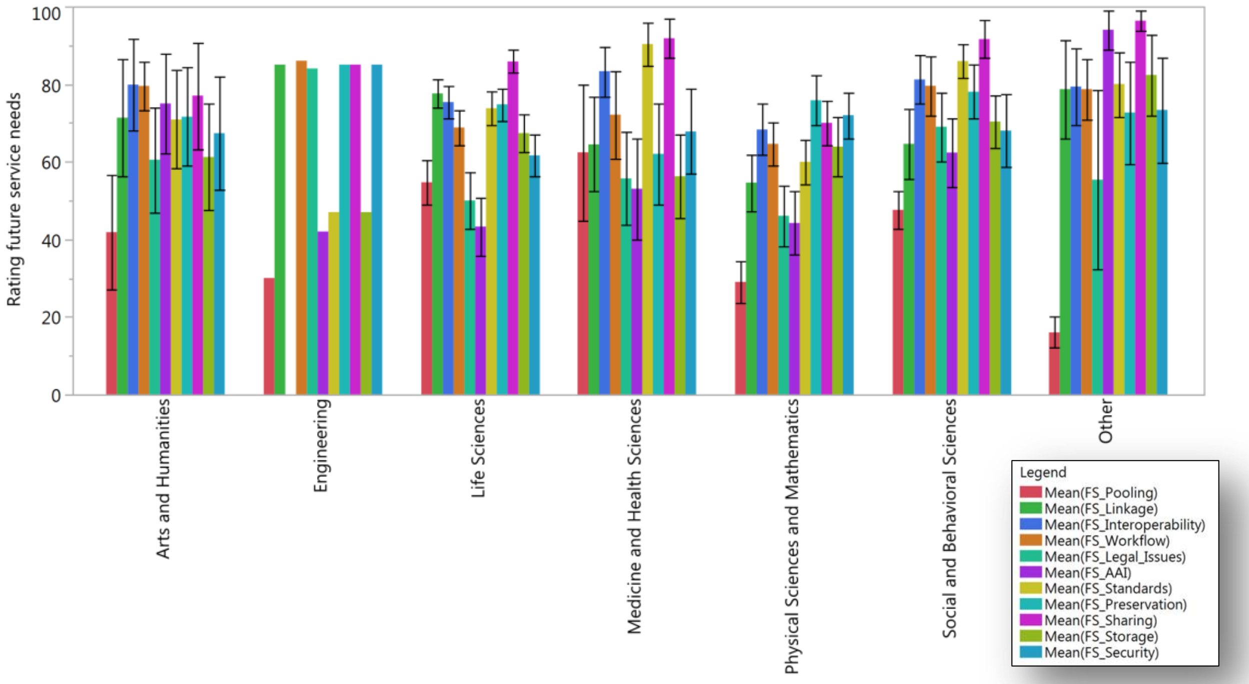


Figure 27: Future service development as seen by the repositories, grouped by bepress disciplines.

While many of the values follow the general pattern, the differences do not form a clear pattern. Some major users' demands are reflected to a certain degree; e.g., the rating for future services in the category of legal issues is more prominent for the humanities and social sciences.

7 Perspective of the repositories

The repository survey (see data paper (von der Heyde, 2019a)) offered additional perspectives not only about the future services, but on the overall status and required resources to provide FAIR data services. General parameters to characterize the repositories were also assessed to enable context-specific analyses.

7.1 Data complexity and curation

One of the core questions in the distribution of workload is the work done either by the scientists or the repository staff. It is commonly accepted that staff capacity cannot and will not scale with the overall scientific community. The specific view of the repositories' staff is nonetheless important for the estimates of future finance demands for staffing.

From the perspective of the repository, curation by the scientists does not correlate with data complexity or the data specificity of the repository. However, both variables do correlate with curation done by the repository staff. Figure 28 depicts the data in conjunction with the corresponding linear fits.

Even though the correlations between the curation done by scientists and the data pattern for the complexity of data and data specific services of the repositories are not significant, we observe the following difference: With increasing complexity of the data, curation is regarded necessary – either done by the repository staff or the scientist; whereas data specific curation is done mostly by the repositories' staff.

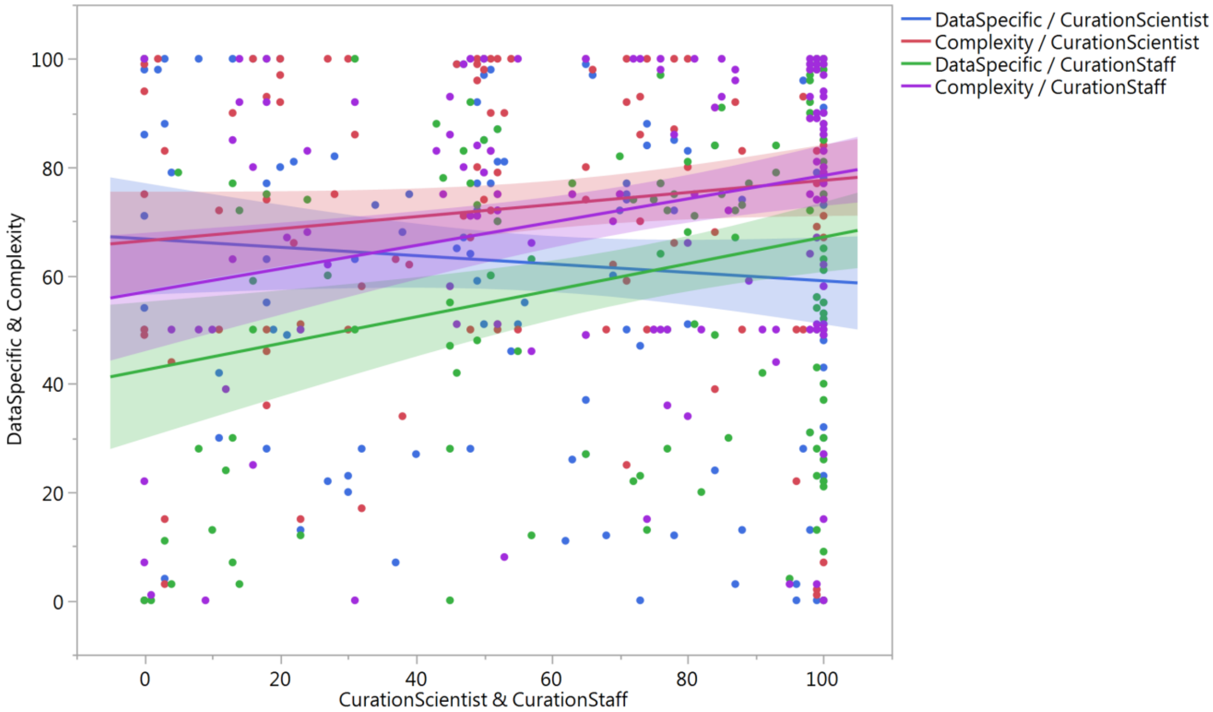


Figure 28: Display of judgements on stored data complexity and specific design of repository vs. curation done by scientists or staff. Values were assessed by sliders between 0 and 100.

7.2 Maturity and FAIRness

The maturity of the repository was assessed using a slider between three stages, described in the survey as follows:

- Initial project [slider value 0]: Building the repository and establishing major parts of its functionality.

- Establishment phase [slider value 50]: After the initial phase, the repository needs to secure additional funds. A business model needs to be established during the second stage of the project. The repository has a clearer focus on who and what its services and customers are.

- Mature institution [slider value 100]: The repository is mature with respect to its type. Sustainable, stable funding supports the repository, which is seen as an institution by its users. Naturally, changes never end and the repository slowly adapts to changes in the "market".

FAIRness, as defined by (Wilkinson et al., 2016), is one of the central demands of scientists, funding agencies and governments when it comes to the future development of repositories.

The representatives of the repositories were asked to judge both maturity and FAIRness. Figure 29 shows the correlation of the maturity rating with a self-assessment of the FAIR principles.

Apparently, adoption of FAIR principles does not depend on secure funding. Few repositories have started to implement the FAIR principles in the initial project phase. Accordingly, established repositories have implemented the FAIR principles more often.

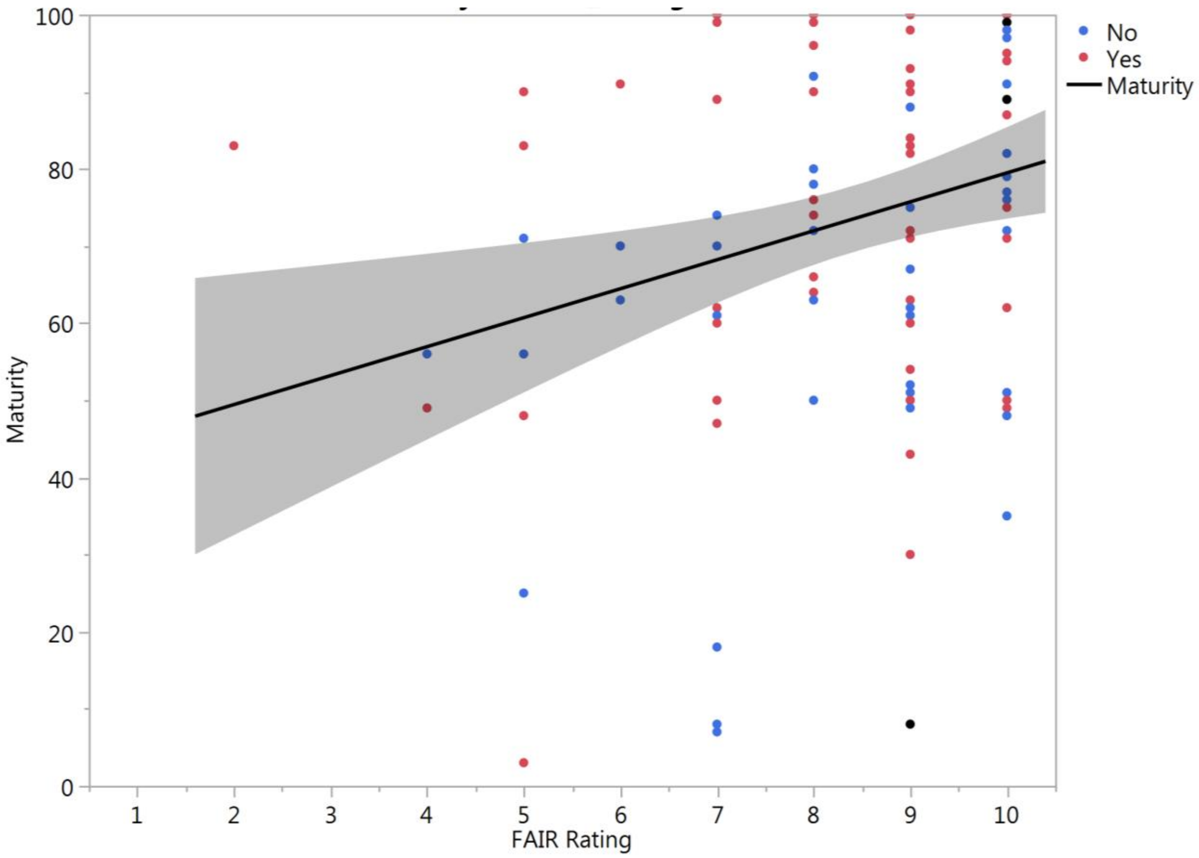


Figure 29: Maturity rating of the repositories is correlated to the self-rating on FAIR principles. The colors of the dots refer to the question of whether the funding of the repository is stable.

8 Cost analysis

The ultimate goal of this project was to determine the amount of resources necessary to support the Swiss community by 2025. Since absolute numbers in millions of CHF would be hard to derive from individual scientists' perspectives, we tried to aggregate multiple sources of information and derive relative estimates.

The baseline for the estimate of future costs was set by (Ember et al., 2013) on page 11:

"The percentage of the total research budget needed to support this approach is likely to be domain specific. We estimate that successful domain repositories can be operated at funding levels of less than 5% of the total research budget (Some fields might be as low as 1%; the cost might rise to 10% in fields with high data rates or particularly diverse and complex metadata). These are modest costs to assure a strong return on public investments in the research and to enable uses of data unanticipated by the original investigators."

The RDA puts this into the context of EU-wide funding and confirmed this hypothesis in 2014 ('The Data Harvest Report – sharing data for knowledge, jobs and growth', 2014) on page 33:

"Our informal estimate is that the infrastructure and operation of a truly effective data-sharing system could cost on the order of 5 per cent of total research budgets. For the Commission, which spends over €10 billion a year through its Horizon 2020 programme, that would amount to half a billion euros."

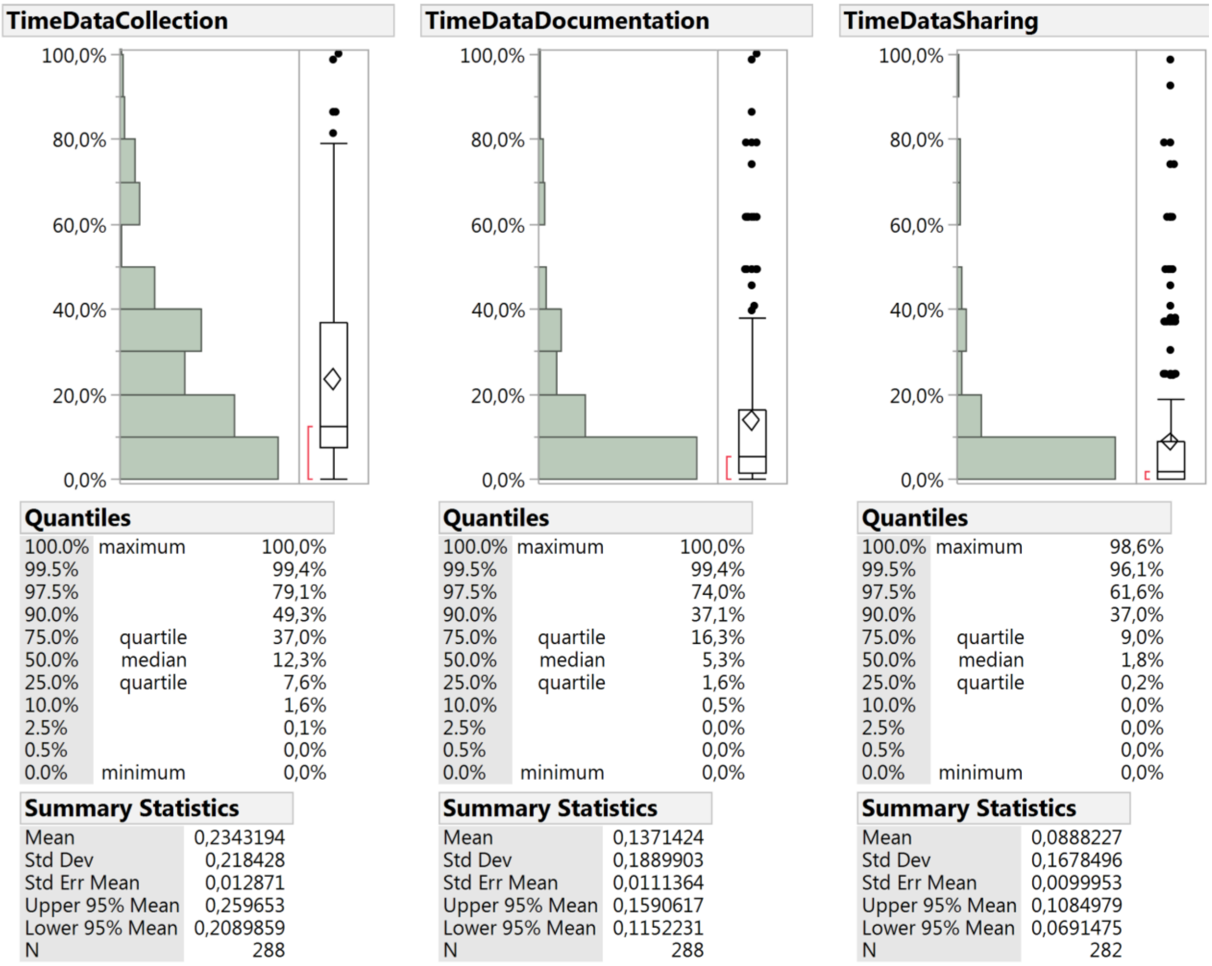
To more reliably estimate the actual cost of data repositories across and specific to scientific disciplines, we have combined many sources of information. The following paragraphs briefly describe the key factors used to estimate the overall resource amount required in the end.

8.1 Scientists spend time on data curation

About 335 scientists answered our question concerning the amount of time spent on data collection, data documentation and sharing their data per project. In addition, we assessed the number of parallel projects and their average duration.

The data was converted from the multiple choices, using the mean of the offered time ranges as workload. The average time spent on a task is equal to the product of parallel projects and task time divided by the average duration of projects.

Some of the data records contained implausible time frames, resulting in time spent on tasks indicating a workload of > 100%. We excluded less than 40 records for those reasons, and continued with about 290 ratings. The statistical analysis of the time spent on data management tasks is shown in Figure 30.



TimeDataCollection

Quantiles

100.0%	maximum	100,0%
99.5%		99,4%
97.5%		79,1%
90.0%		49,3%
75.0%	quartile	37,0%
50.0%	median	12,3%
25.0%	quartile	7,6%
10.0%		1,6%
2.5%		0,1%
0.5%		0,0%
0.0%	minimum	0,0%

Summary Statistics

Mean	0,2343194
Std Dev	0,218428
Std Err Mean	0,012871
Upper 95% Mean	0,259653
Lower 95% Mean	0,2089859
N	288

TimeDataDocumentation

Quantiles

100.0%	maximum	100,0%
99.5%		99,4%
97.5%		74,0%
90.0%		37,1%
75.0%	quartile	16,3%
50.0%	median	5,3%
25.0%	quartile	1,6%
10.0%		0,5%
2.5%		0,0%
0.5%		0,0%
0.0%	minimum	0,0%

Summary Statistics

Mean	0,1371424
Std Dev	0,1889903
Std Err Mean	0,0111364
Upper 95% Mean	0,1590617
Lower 95% Mean	0,1152231
N	288

TimeDataSharing

Quantiles

100.0%	maximum	98,6%
99.5%		96,1%
97.5%		61,6%
90.0%		37,0%
75.0%	quartile	9,0%
50.0%	median	1,8%
25.0%	quartile	0,2%
10.0%		0,0%
2.5%		0,0%
0.5%		0,0%
0.0%	minimum	0,0%

Summary Statistics

Mean	0,0888227
Std Dev	0,1678496
Std Err Mean	0,0099953
Upper 95% Mean	0,1084979
Lower 95% Mean	0,0691475
N	282

Figure 30: Statistics for time spent on data management tasks.

The mean ratings of the percentage of time spent on each data management task in Figure 30 are actually quite low (23% on data collection, 14% on data documentation and 9% on data sharing).

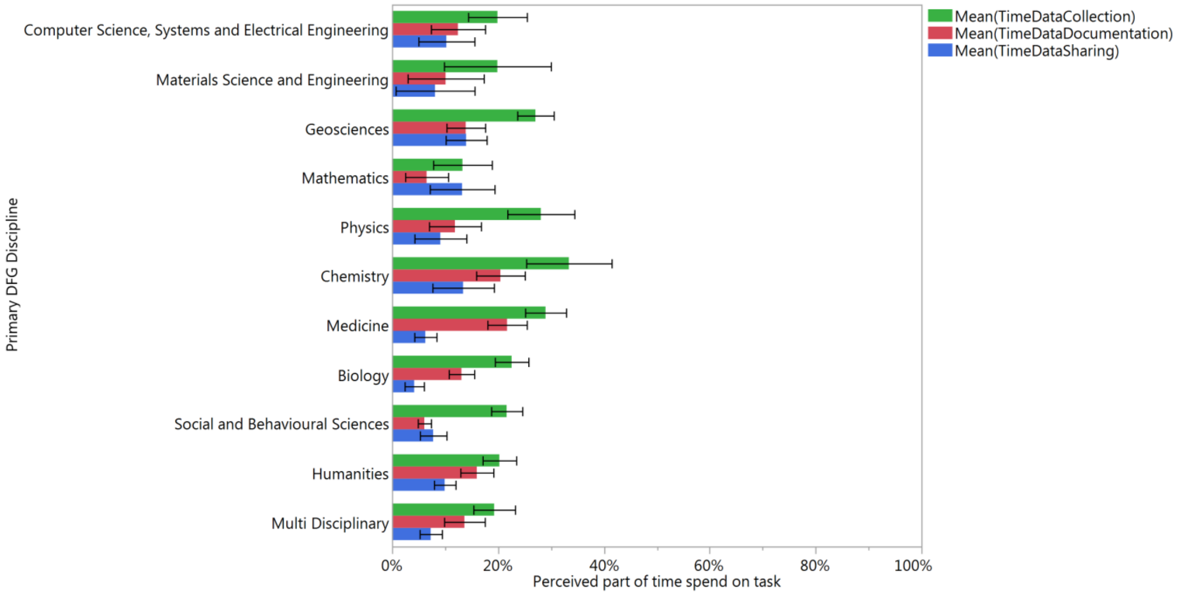


Figure 31: Perceived time scientists spend on data management tasks. Across DFG classified disciplines, data collection is most time consuming, while data sharing is least time demanding.

The average time spent on tasks by disciplines, according to the DFG classification, is displayed in Figure 31. The disciplines Thermal Engineering, Construction Engineering Mechanical Engineering and Agriculture, Forestry and Veterinary medicine were excluded as they suffered from noisy data. The remaining disciplines mostly show the same general pattern varying due to the differences between the disciplines. Only in Mathematics does the effort necessary for data documentation seem to be rated very low.

8.2 Size of datasets

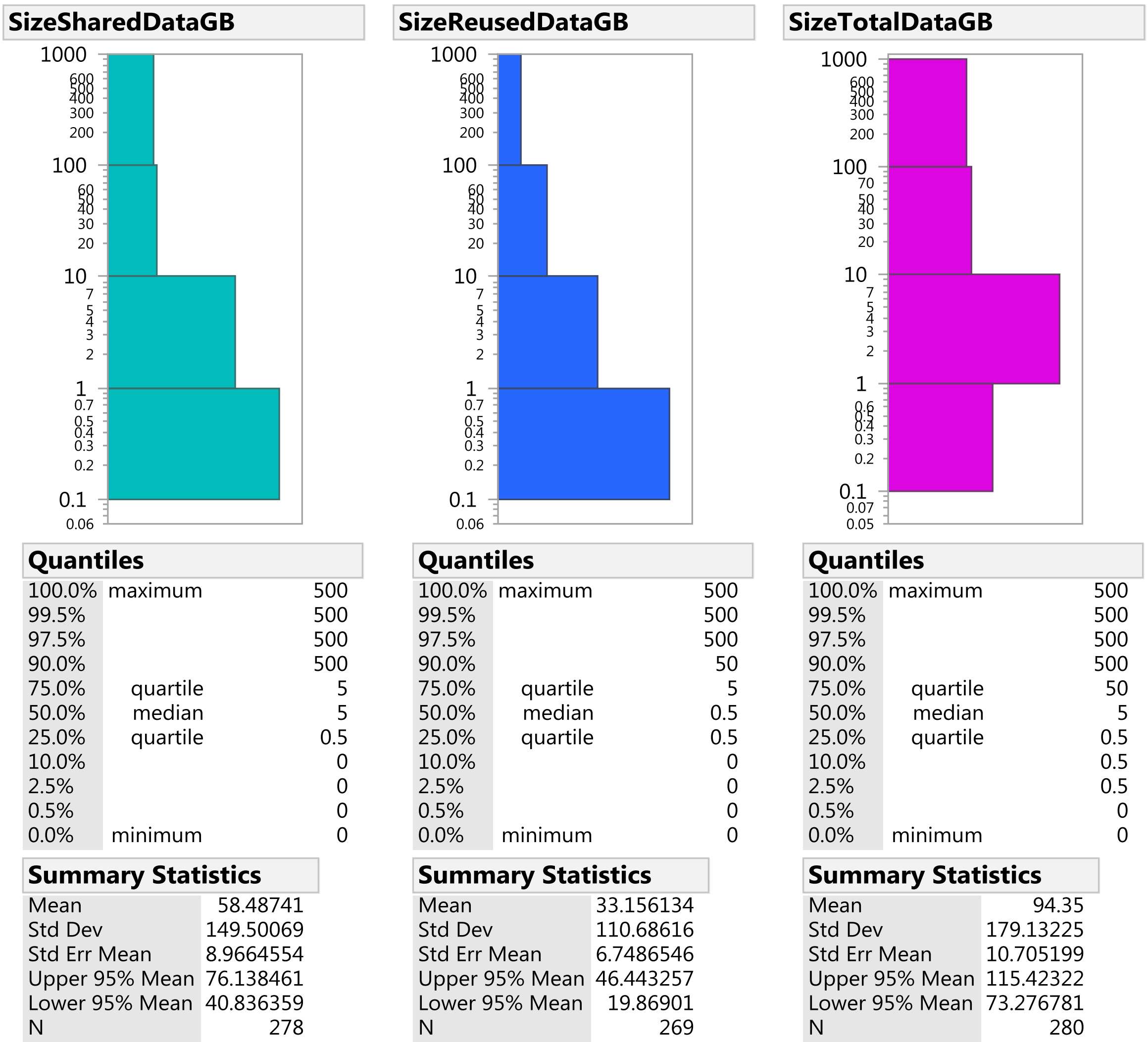
Using the dataset compiled by (Nature Research, 2016) corresponding to the report by (Treadway et al., 2016), we derived the following basic statistical values:

- We observe a very imbalanced distribution of data:
  - About 11 to 12% of the participants in the survey did not know suitable answers to the questions about data set sizes, or skipped this section for other reasons.
  - About 2 to 3% of the participants own 92% of the overall data.
  - The remaining 85% of people provide only an additional 8% of data.
- We found that overall data size did not depend on / correlate with the file types in the sense of technical format.
- Data size and file count were slightly correlated.
- In addition, we found a correlation between the file type (content wise) and the speed and frequency with which data is generated.
  - Slow: Questionnaires, transcripts, codebooks
  - Fast: Laboratory notebooks, field notebooks, diaries, photographs, films, slides, artifacts, specimens, samples
  - Medium: Rest of the types (e.g., text documents, database content, models, analysis, workflows)
- In total, each person reported sharing about 250 GB (it remains unclear if this only refers to the figshare platform). As said before, this average is misleading, since the

"average" person shares only 8% of this (~ 20GB). The "rest" is shared by the 2.5% of researchers who are power-users.

Overall, this pattern suggests the need to refrain from scaling factors across data sizes, as they introduce a factor of 10 into the equations. As 20GB seems reasonable for the average "free" platform on the web, there is no real question about scaling sizes and volumes.

The landscape survey collected responses from about 280 of the participants on data size. Figure 32 shows the distribution, mean and median values of dataset sizes.



SizeSharedDataGB

Quantiles

100.0%	maximum	500
99.5%		500
97.5%		500
90.0%		500
75.0%	quartile	5
50.0%	median	5
25.0%	quartile	0.5
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	58.48741
Std Dev	149.50069
Std Err Mean	8.9664554
Upper 95% Mean	76.138461
Lower 95% Mean	40.836359
N	278

SizeReusedDataGB

Quantiles

100.0%	maximum	500
99.5%		500
97.5%		500
90.0%		50
75.0%	quartile	5
50.0%	median	0.5
25.0%	quartile	0.5
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	33.156134
Std Dev	110.68616
Std Err Mean	6.7486546
Upper 95% Mean	46.443257
Lower 95% Mean	19.86901
N	269

SizeTotalDataGB

Quantiles

100.0%	maximum	500
99.5%		500
97.5%		500
90.0%		500
75.0%	quartile	50
50.0%	median	5
25.0%	quartile	0.5
10.0%		0.5
2.5%		0.5
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	94.35
Std Dev	179.13225
Std Err Mean	10.705199
Upper 95% Mean	115.42322
Lower 95% Mean	73.276781
N	280

Figure 32: Distribution statistics for data sizes. While Shared and Reused data show the same pattern on the logarithmic scale, the total size of data differs. This supports the hypothesis of unpublished material, which forms the "long tail of sciences" (see section 8.3).

Confirming the results we derived from the figshare data, we additionally observed the median to be below 10% of the average data size. This supports the notion of a highly unbalanced usage of data storage.

8.3 *Results from hidden treasures*

Overall, 316 participants indicated whether they had a "hidden treasure" in terms of data or not (see Figure 33). Those participants indicating no (n=148) skipped the following questions about specific attributes of the (potential) treasure; 168 participants continued with questions concerning the data, which could have potential value if shared with others.

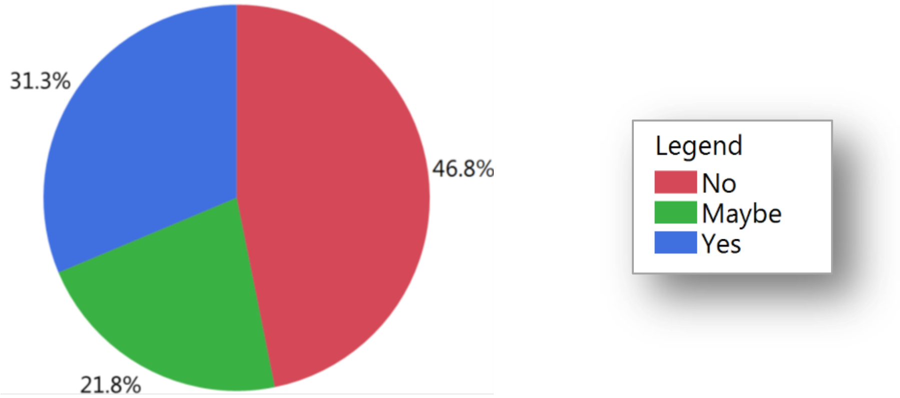


Figure 33: Proportion of the participants indicating having data which has not yet been shared and might be of value for others. This potential "treasure" was assessed with further questions, when participants answered maybe or yes.

For further planning by the SNSF and swissuniversities, the sizes of these potential treasures are of interest. The average size of the hidden treasures as estimated by the participants, measured in gigabytes, is shown in Figure 34. Due to the low participation rate, the error bars are quite large.

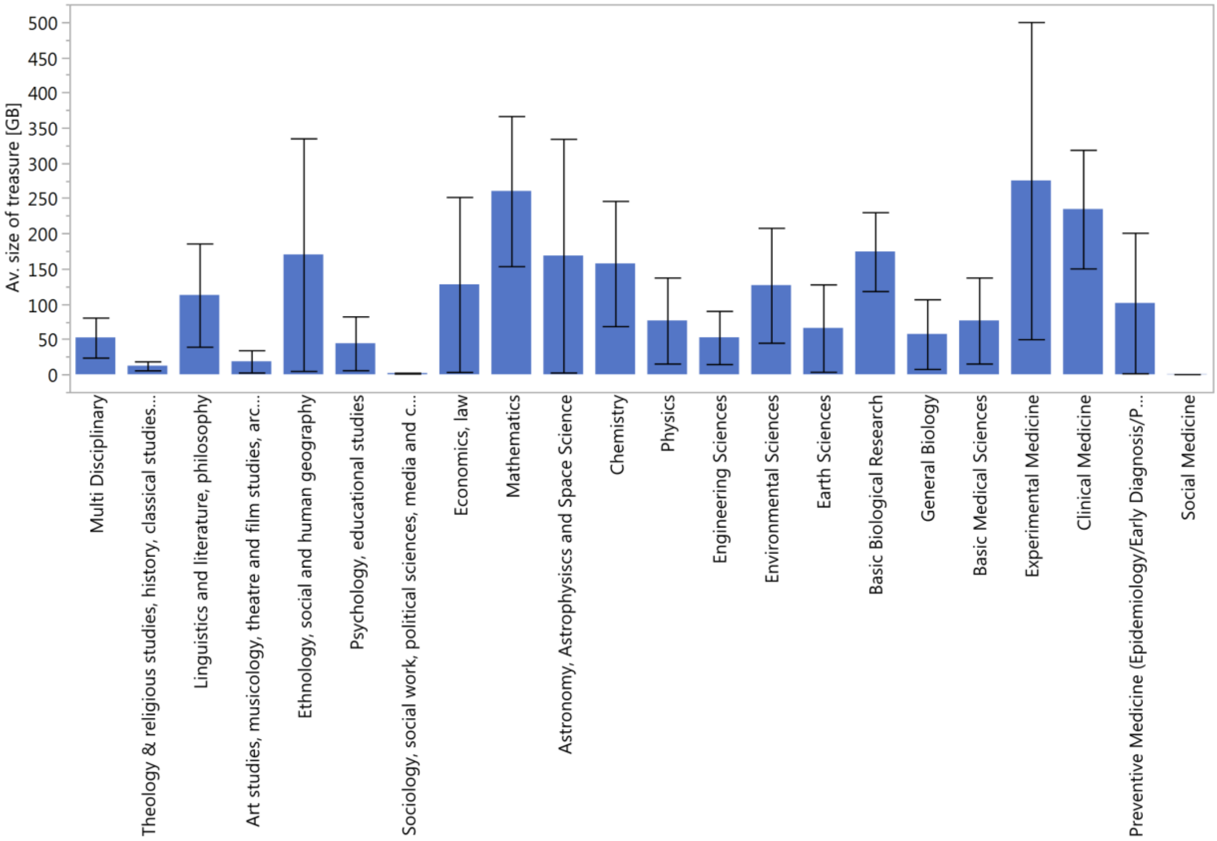


Figure 34: Average 'treasure' size in gigabytes across SNSF disciplines.

Adding up the sizes of all the treasures in each SNSF discipline leads to the very rough estimates shown in Figure 35. The totals of the estimated treasures in the Basic Biology and Clinical Medical disciplines were the largest.

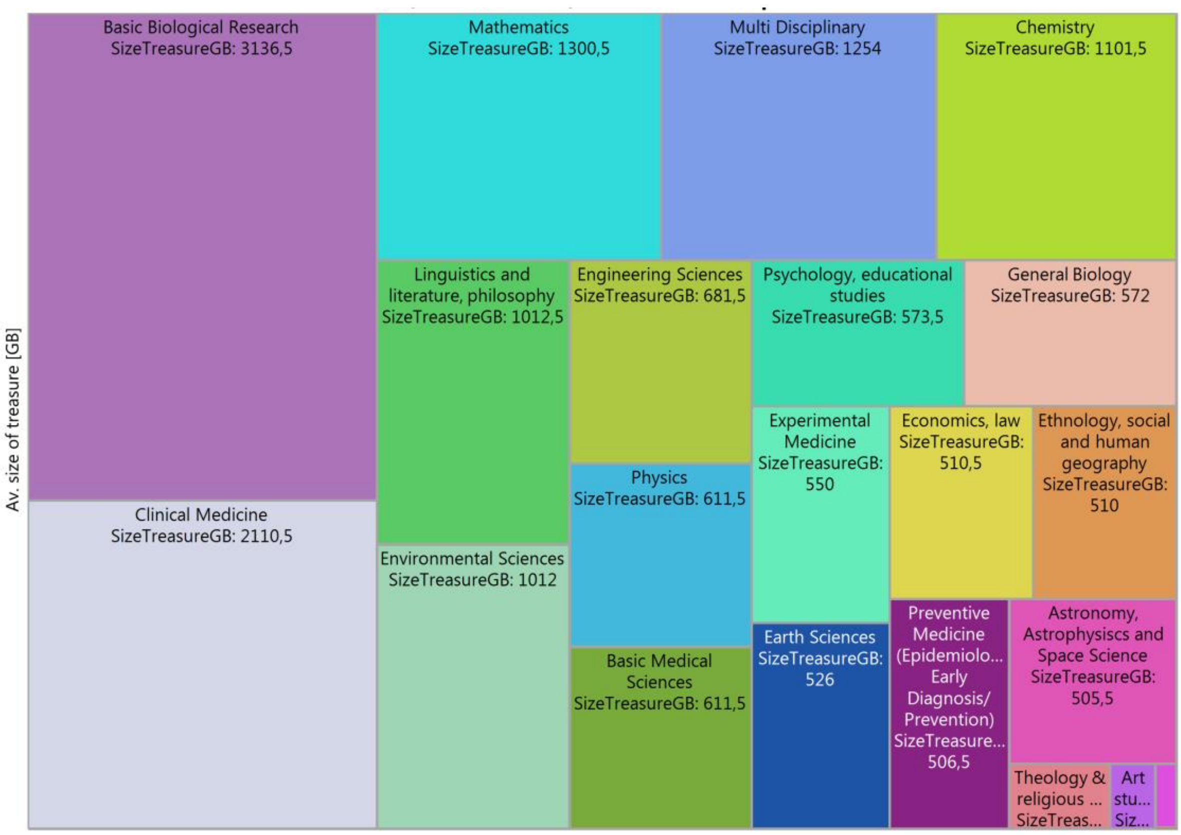


Figure 35: Sum of the hidden treasures by SNSF discipline.

The overall amount of hidden treasure data must be put into to the context of the other data amounts given by the participants. Figure 36 shows the amounts of shared and reused data in comparison to overall existing data and the sizes of the hidden treasures of data.

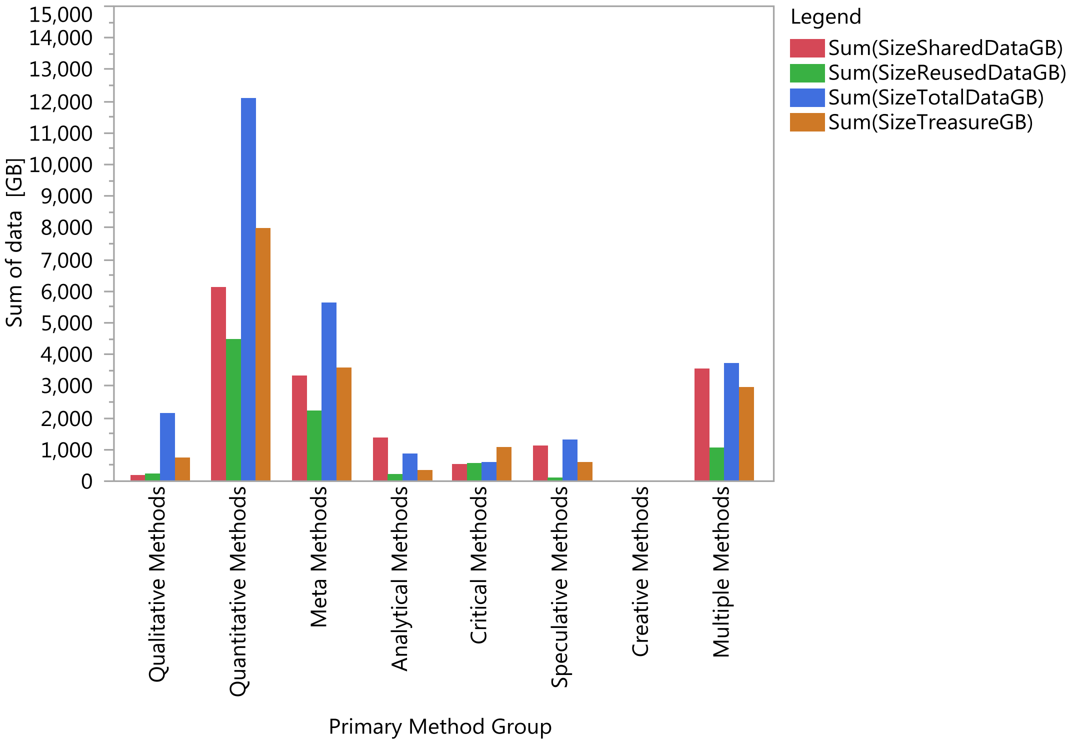


Figure 36: Sizes of existing shared, reused, 'hidden treasure', and total data, split by the primary method used by the 280 responding scientists.

Perhaps unsurprisingly, the amount of data is biggest for the quantitative methods. Due to a mixture of methods in the Meta and Multiple groups, those represent a mixture of all disciplinary fields and so show larger amounts of data. In comparison, Qualitative, Critical and Analytical methods display data volumes in the same order of magnitude, which is about 1/10 of the amounts in the quantitative methods group.

The comparison between not yet shared data (sum = 17.4 TB) and shared data (sum = 16 TB) confirms the results of other research, which shows that about 50% of the data worth sharing has been shared. Looking at the "long tail of sciences", we can only speculate whether those data mentioned by Ferguson et al. (A. R. Ferguson, Nielson, Cragin, Bandrowski, & Martone, 2014) are already included in the overall data indicated by our participants. We also speculate that those additional treasures are seen as beyond reach, since most of the material is not yet digital by nature.

The overall data size amounts to about 68 TB which, split across 280 scientists, results in a reasonable size of 250 GB per person. Following the results from (Treadway et al., 2016) mentioned in section 8.2, the average scientist would have about 20 GB and 7 of the 280 would work with about 9 TB each. Storing 20 GB per scientist falls within the range of standard storage capacity; on the other hand, offering 9TB of high performance storage is costly. Storing data for the broad majority seems not to be a technological or financial problem; storing for power users is demanding. In conclusion, the challenges for repositories are apparently not the funding of storage technologies.

8.4 Repository storage

Only a few of the repositories provided data on storage sizes (n=44, ~21%). The overall size amounted to about 8.5 PB, of which 6 PB are located in low-end storage (on tapes).

The average sizes per storage group were 6 TB, 59 TB, and 251 TB in high-end (SSDs), mid-range (hard discs) and low-end storage respectively (see Figure 37).

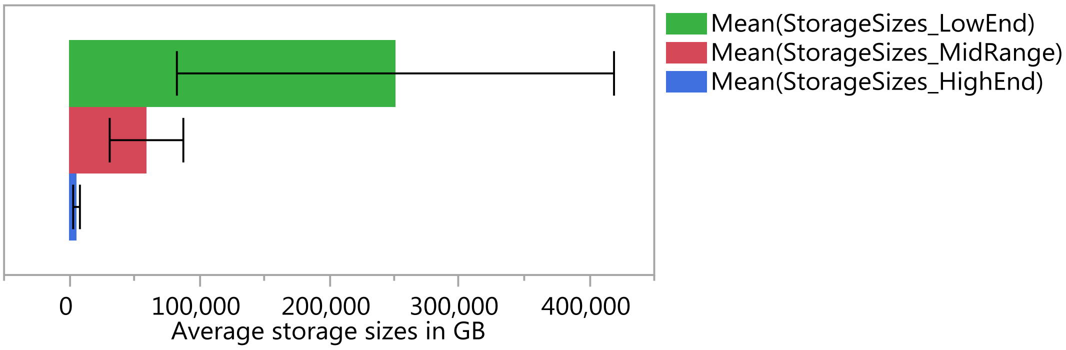
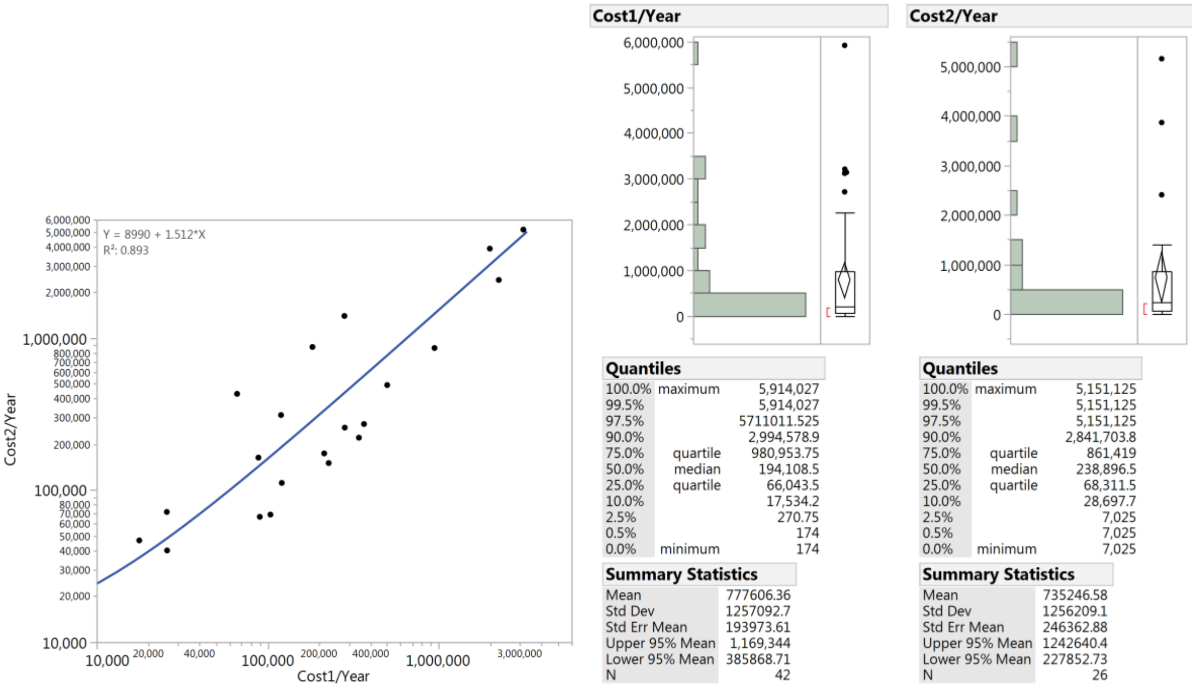


Figure 37: Average storage sizes by type of storage. Due to the low number of repositories providing data, the error bars are quite large.

8.5 Repository costs

The repository representatives were asked for the total project cost during the first two phases of the repository project. Even though some participants complained about this rigid structure of financial questions, 46 gave their numbers for the first project phase and 33 for the second project phase. Those numbers were converted to current value by the OECD purchasing power index (PPP) for the appropriate year and currency (OECD, 2018). Next, the USD equivalent sum was scaled for each year (since the funding was given) using a virtual 1.5% inflation rate.

The average costs per year could be calculated for 42 projects for the first phase and 26 for the second. The linear fit of both distributions is shown in Figure 38. The average increase of the costs between first and second phase is a factor of 1.5. The average costs per year of all projects remained the same: around 735,000 USD (corrected for PPP and inflation).



Cost1/Year

Quantiles		
100.0%	maximum	5,914,027
99.5%		5,914,027
97.5%		5711011.525
90.0%		2,994,578.9
75.0%	quartile	980,953.75
50.0%	median	194,108.5
25.0%	quartile	66,043.5
10.0%		17,534.2
2.5%		270.75
0.5%		174
0.0%	minimum	174

Summary Statistics	
Mean	777606.36
Std Dev	1257092.7
Std Err Mean	193973.61
Upper 95% Mean	1,169,344
Lower 95% Mean	385868.71
N	42

Cost2/Year

Quantiles		
100.0%	maximum	5,151,125
99.5%		5,151,125
97.5%		5,151,125
90.0%		2,841,703.8
75.0%	quartile	861,419
50.0%	median	238,896.5
25.0%	quartile	68,311.5
10.0%		28,697.7
2.5%		7,025
0.5%		7,025
0.0%	minimum	7,025

Summary Statistics	
Mean	735246.58
Std Dev	1256209.1
Std Err Mean	246362.88
Upper 95% Mean	1242640.4
Lower 95% Mean	227852.73
N	26

Figure 38: Left: Linear fit of costs per year between the first and second funding phases of repository projects. Right: Baseline statistics of both variables are shown.

8.6 Repository cost structure

About 75% of the funding of the repositories can be attributed to the higher education or public funding in general (see Figure 39). Private funding and generating a revenue stream is uncommon.

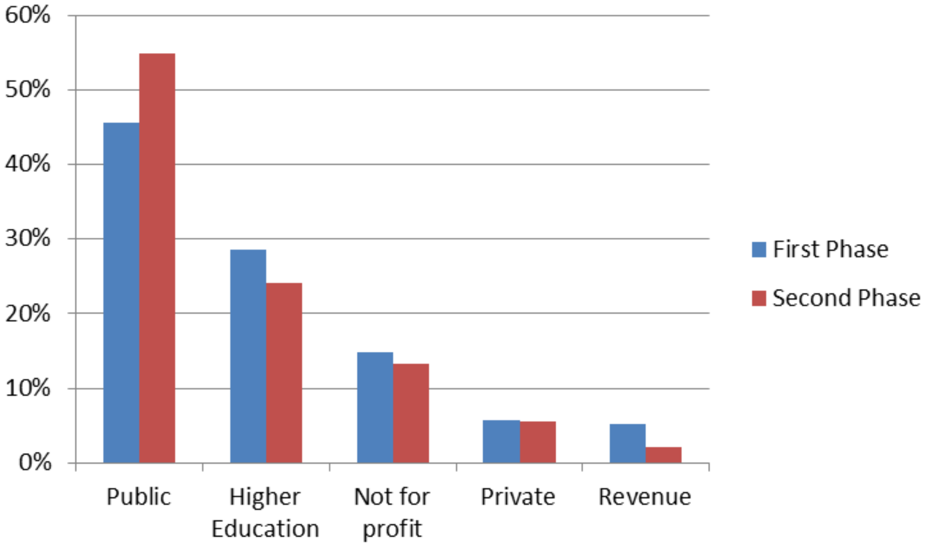


Figure 39: Funding source during the first and potential second phases of establishing a repository.

Overall, the structure of funding cost distribution among account groups and tasks reflects typical projects running mostly in publicly-funded environments (see Figure 40).

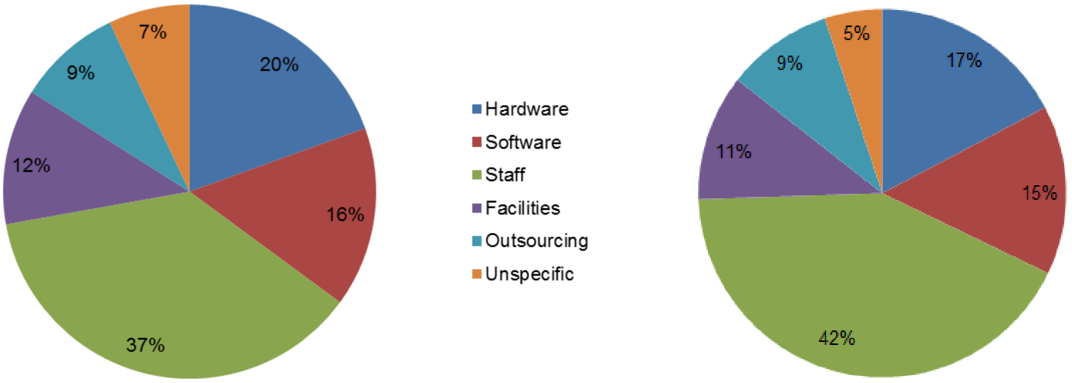


Figure 40: Average shares of spending in first (left) and second (right) project phase of the repositories.

Repositories seem to shift the expenditures between the first and second funding phases slightly towards changes and new services. However, development and running the repository are the biggest tasks, adding up to about 50% of the activities (see Figure 41). Since this analysis derived the expenditure on a coarse level from Likert scale ratings by normalizing each repository, the relative index cannot be attributed to EUR spent or other absolute scales.

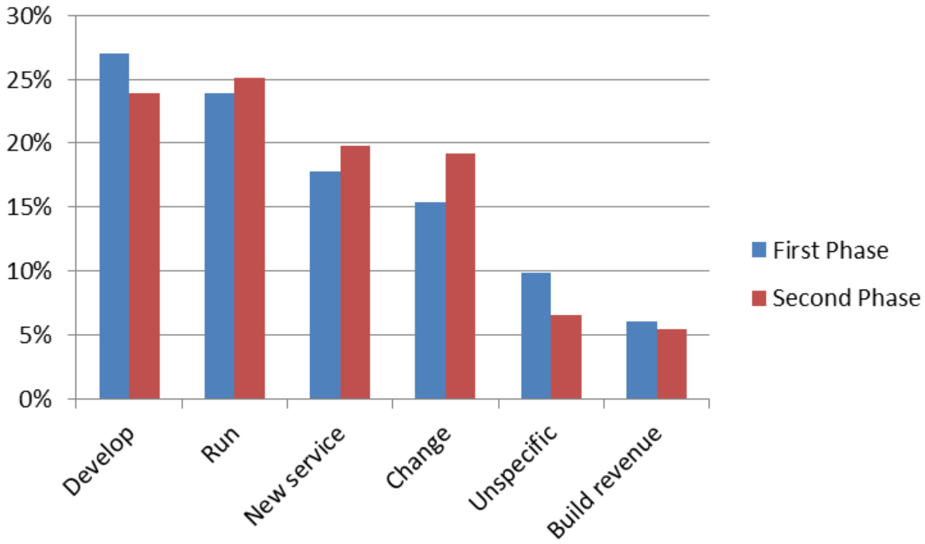


Figure 41: Expenditure of the repositories during the first and second project phases.

Most variables showed very similar distributions when comparing the first and second project phases. However, the expenditure on new services differed between the phases (see Figure 42).

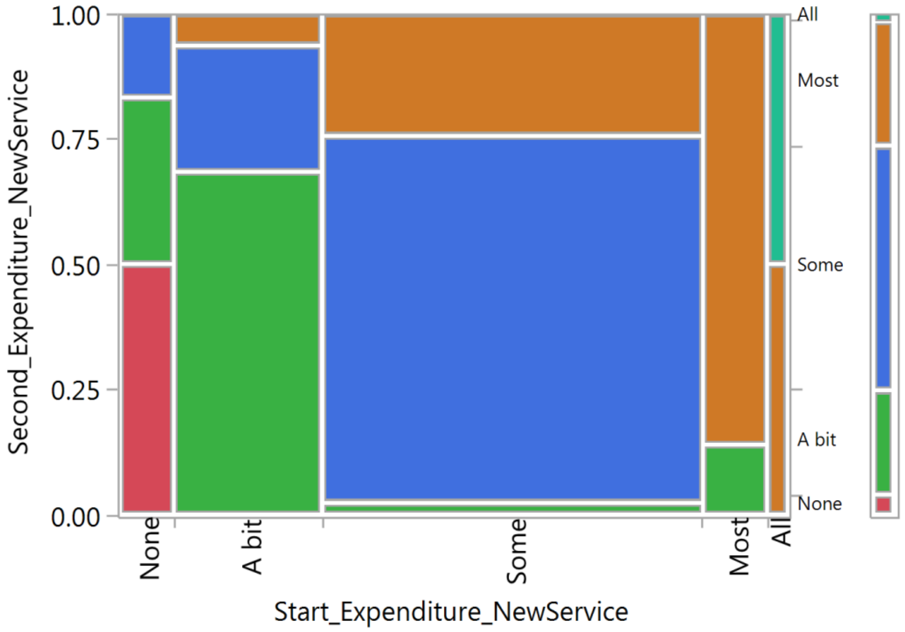
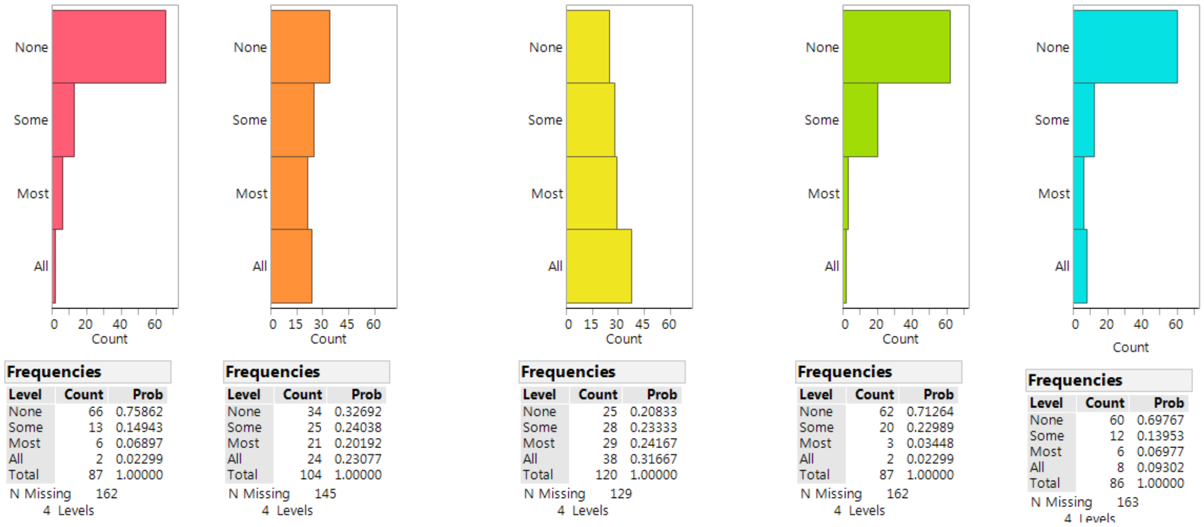


Figure 42: A mosaic plot allows the comparison between the first and second funding phases. Depicted is the expenditure by 72 repositories on new services.

After the initial phases, the project should reach a stable state, where the funding is secured and service delivery is the main focus. Current funding sources are depicted in the histograms in Figure 43.



Frequencies

Level	Count	Prob
None	66	0.75862
Some	13	0.14943
Most	6	0.06897
All	2	0.02299
Total	87	1.00000

N Missing 162
4 Levels

Frequencies

Level	Count	Prob
None	34	0.32692
Some	25	0.24038
Most	21	0.20192
All	24	0.23077
Total	104	1.00000

N Missing 145
4 Levels

Frequencies

Level	Count	Prob
None	25	0.20833
Some	28	0.23333
Most	29	0.24167
All	38	0.31667
Total	120	1.00000

N Missing 129
4 Levels

Frequencies

Level	Count	Prob
None	62	0.71264
Some	20	0.22989
Most	3	0.03448
All	2	0.02299
Total	87	1.00000

N Missing 162
4 Levels

Frequencies

Level	Count	Prob
None	60	0.69767
Some	12	0.13953
Most	6	0.06977
All	8	0.09302
Total	86	1.00000

N Missing 163
4 Levels

Figure 43: Repository funding as of today. Most repositories are either funded by higher education or other public funding.

Further observations were made reviewing the cost and funding structure:

- Most repositories are publicly funded and spend their money on staff that develops software for the repository; this increases in the second project phase. On the other hand, investments in “change” decreased towards the second project phase.
- As building up revenue streams is not a focus for most repositories, alternative business models are therefore out of the question.
- Funding from the higher education sector decreases from the first to the second funding phase, whereas the general public funding increases.

8.7 Summary

Extrapolating from these data to the Swiss research community overall based on these few numbers is very difficult. Assuming from all the other data that there are no fundamental differences between Switzerland and the rest of EU and the world in general, there hardly seems to be an alternative to public funding, if repositories should be stable and provide a long-term (public) service. Alternative scenarios as discussed in e.g., (von der Heyde, Hartman, Auth, & Erfurth, 2018) will drive external business models without public funding eventually. However, paying for publications which are then sold back to the scientific community should not be repeated for open data.

9 Influences of policies

How effective are policies? To be effective, a policy has be known and acknowledged by the parties involved. Naturally, a policy also has to contain effective measures.

The landscape survey asked the participants to indicate their knowledge about policies of special interest for open data, data sharing, and data reuse. Overall, 1,423 researchers gave at least one answer to the overall block of questions and in this sense participated in that part of the survey. From the text comments we derived general uncertainty and little knowledge about the current policy situation: surprisingly few scientists are aware of the policies.

Most participants claimed to have heard about a specific policy; many fewer actually know the content or even comply with the policy. The ratio of participants who stated any level of knowledge about the SNSF policy is about 96% (1,367/1,423) (see Figure 44). When asked for potential reasons not to share, about 22% stated ‘funder does not require to do so’ (Figure 2); thus, we conclude that only up to 78% of all the scientists are actually aware of the SNSF requirements.

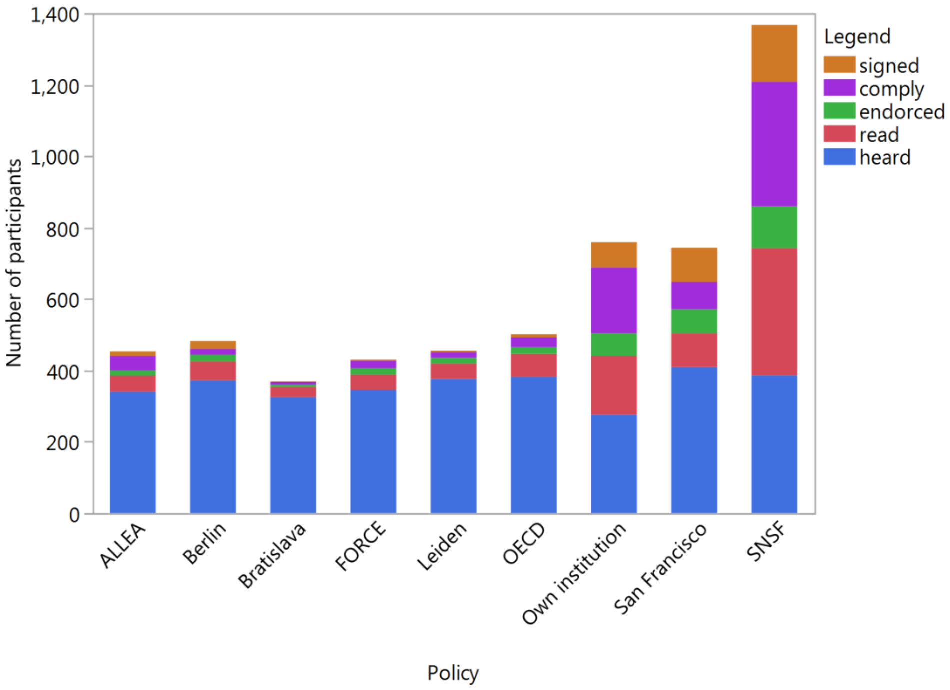


Figure 44: Scientists in the landscape survey have heard about certain policies more often than others. Depicted is the number of participants who indicated any connection (heard of, comply with, signed etc.) to the specific policy.

Beside the best-known SNSF policy, the local policy and San Francisco Declarations (DOAR) are also known by >50%. The other 6 policies only differ slightly in being known by 25 to 35 percent of the participants. This corresponds with the 60% of participants who had 'never heard of FAIR' in (Hahnel et al., 2018).

The average participant knows three to four of the total nine policy choices offered (see Figure 45). As the averages are not very different between disciplines, this overall situation of the Swiss community does not depend on discipline-specific factors or cultural issues.

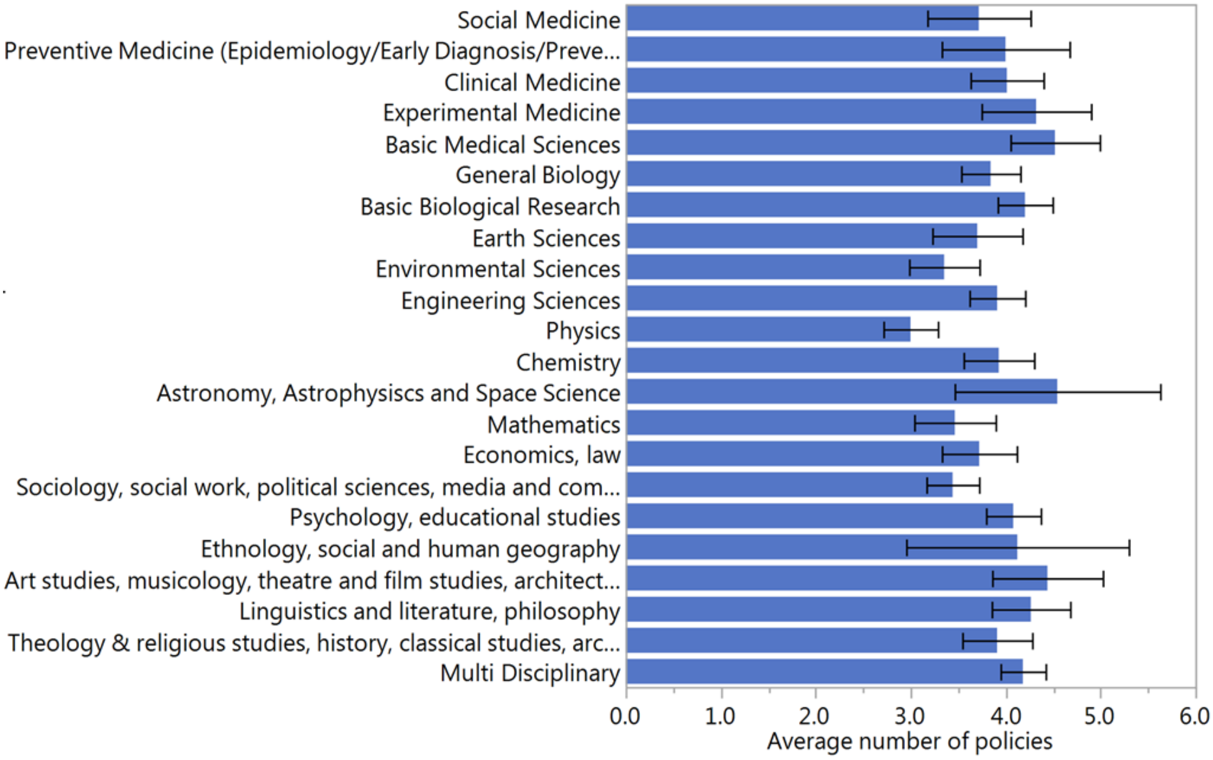


Figure 45: Average number of policies known, read, endorsed or complied with. We only included participants responding to at least one policy.

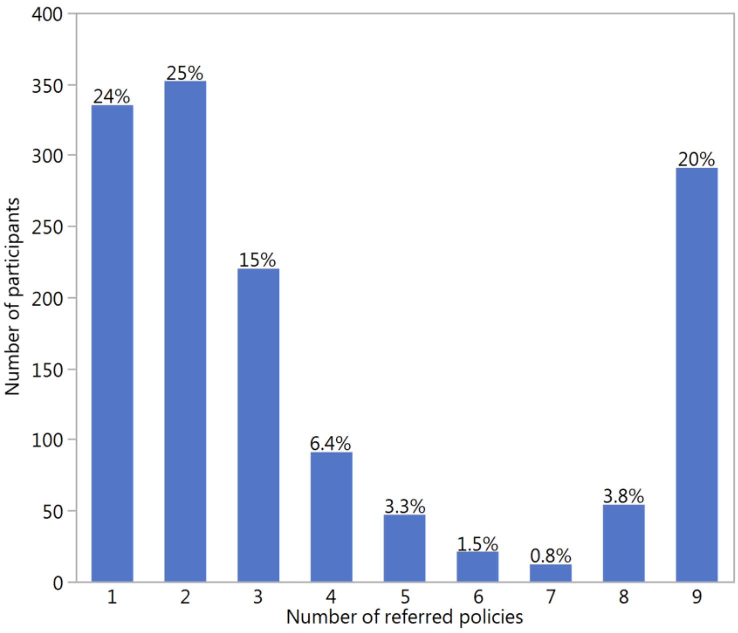


Figure 46: Number of policies the participants referred to.

The histogram of the number of policies referred to (see Figure 46) shows that the majority actually has basic knowledge for three or fewer policies. In combination with Figure 44, this indicates that most scientists have simply heard about some policies, but have not yet gained deeper knowledge about them.

Since two of the nine policy choices are highly relevant for all participants (the SNSF policies and the local policy of the institution), the results document a certain amount of ignorance towards “external guidance”. In sum, funding agencies and international organizations need to advertise the existence and the details of their policies.

10 Recommendations

The overall situation of researchers in Switzerland is in many aspects comparable to the EU or worldwide level. The surveys often replicated effects reported in earlier work (see 4.1 to 4.4). Since the project combined several methods, the overall picture seems to be more complete (see 4.5, 4.6 and 5). It therefore seems possible to derive specific recommendations for Switzerland. It remains arguable if these in turn might even apply to the overall scope of open data in the EU and beyond.

Disciplines are different in their habits, language, and concepts. This project was meant to shed a systematic light upon those differences in relation to data sharing and reuse practices. Some of the factors are related to the methods scientists apply, others to the factors described in previous work. However, the overall similarity of certain problems is also obvious. To address both disciplinary topics as well as general issues, we identify seven main areas of recommendations which can be derived from the overall project:

1. Research is shared by 3/4 of all scientists. However, the use of data repositories as one option for data preservation and sharing is not yet widespread across the Swiss research community (33% of all researchers use repositories; see Table 5). General repositories are more commonly known and used (see Table 7). The overall frequency for sharing and reuse of data in general purpose repositories or disciplinary repositories is about the same. However, the intensity of usage of general purpose repositories is much higher, because the usage of the disciplinary repositories is split between a great number of researchers (Figure 18). Almost everybody has a favorite repository (see section 5.1).
   Recommendation: The high fragmentation should not be pushed further by funding calls addressing small research groups. Small groups should get funded if a visible community with highly shared data concepts and methods supports the projects and no international repository overlaps with the initiative.
2. Aside the general purpose repositories, there are hardly any repositories used by a substantial number of scientists from Switzerland (see section 5.1). The key exception to this general view is FORSbase, which has been established with a solid user community (see Figure 18).
   Recommendation: Asking the scientific community to establish method-based repositories might create a more homogenous landscape than the current situation, as demonstrated by the success of FORSbase.
3. The future services demanded by the users are not met sufficiently by the plans of the repository providers (comparison of Figure 24 and Figure 26). In particular, services for the support of legal issues and security in general have to be more in focus, as this has the highest overall demand.
   Recommendation: Repositories in the second stage of maturity should be motivated to

adapt more quickly to the needs of their specific users. This should help them to grow and establish a reliable community.

4. The knowledge of policies around the topic of open data needs improvement. This situation is nearly identical in all disciplines (see chapter 9). There seems to be a lack of discussion between the funding agencies releasing the policies and the scientific community (see also section 3.2).
   Recommendation: Before changing, adding or enhancing policies, ways to make them better known should be considered. Acceptance will come from discussion and trust.
5. Scientists spend their time on research very effectively (see section 8.1). Therefore, every additional step or formal requirement is perceived as a distraction (see chapter 3). Without the benefits of additional steps being immediately visible, changing the data publishing culture takes far too long.
   Recommendation: Future policy changes should be derived from a broad consultation with the scientific community. Following overall strategic goals is part of its governance. Agreement on how to reach goals needs to be supported by the community.
6. Several participants remarked on the complexity of the overall topic and referred to upcoming major changes in their disciplines (see sections 3.2 and 3.4). In addition, we observed that some participants had difficulties in providing answers to questions which required background knowledge in data management. Finally, the corresponding local, national, and international policies are far from being known by the majority (see chapter 9). On average, people know only three out of nine data sharing policies, not taking into account those who skipped these questions altogether.
   Recommendation: Community-based programs to facilitate the use of data repositories should offer advanced training for scientists of all ages. Projects which do not offer services to build up potential end users' knowledge should provide evidence for the broad acceptance of their standards within the respective community.
7. The observed imbalance of data sizes and overall size demands are not unusual (see section 8.2).
   Recommendation: Funding agencies have to decide either to support a few scientists very well or a broader spectrum of scientists for the common good. The national policies were written with this broad concept in mind. Supporting only the scientists with 'big data' fails (see section 8.7).

11 Acknowledgements

This project would not have been possible without the help of many people:

First of all, we'd like to thank all the scientists, repository managers and repository staff who invested their time in answering our survey or participating in the pilot interviews for the preparation of the survey. Overall about 2,750 people provided their knowledge.

Thanks to the members of the Scientific Advisory Committee, who supported the project with comments on reviews, methods, and scientific questions:

- Ms. Geraldine Clement-Stoneham, Head of Knowledge Management and Scholarly Communication, Medical Research Council GB
- Mr. Benedikt Fecher, Forschungsprogrammleiter: Lernen, Wissen, Innovation am Alexander von Humboldt Institut für Internet und Gesellschaft
- Mr. Hans Pongratz, CIO TU München, Themenpate "Neue Geschäftsmodelle, Technologien & Lebenslanges Lernen" im Hochschulforum Digitalisierung
- Mr. Robert Ulrich, Karlsruhe Institute of Technology, kit Library (IT-Dienste), re3data
- Ms. Eveline Wandl-Vogt, Austrian Academy of Sciences, workgroup Chair in DARIAH-EU
- Mr. Ramin Yahyapour, CIO der Universität Göttingen, CIO Universitätsklinik Göttingen, Geschäftsführer der GWDG, Mitglied der KfR und RFII

Additional comments were provided in discussions by scientific colleagues Mr. Andreas Breiter, Mr. Marcel Hebing, and Mr. Jörg Hofmann.

We also would like to acknowledge the help and feedback of Ms. Gabriele Kloska concerning the re3data database quality control and the mapping of the DFG disciplines. We'd further like to thank the maintainers of the databases (see Appendix B: Data sources) for their work on quality assurance and the open APIs which were used to speed this project up. For providing special formatted data from the BFS, we thank Ms. Elena Zafarana.

We also thank the advisory group at the SNSF and swissuniversities, namely Ms. Ayşim Yılmaz, Mr. Martin von Arx, and Ms. Gabi Schneider. As representatives of the contracting authorities, they provided valuable input and guidance throughout the process. Without the work of Ms. Sandra Samsoodeen-Ueltschi, the invitations to all scientists in Switzerland would not have been possible.

Finally, we thank the authors of all the previous work (see Appendix A: References) for their excellent research on the matter and their willingness to provide their data, either openly or after personal contact.

Appendix A: References

Beissel-Durrant, G. (2004). A Typology of Research Methods Within the Social Sciences.

BFS. (2018, September 27). Anzahl Personalressourcen HS nach Fachbereichsgruppe und Personalkategorie, 2017. Retrieved from https://www.bfs.admin.ch/bfs/de/home/statistiken/bildung-wissenschaft/bildungsindikatoren/bildungssystem-schweiz/bildungsstufen/hochschulen/personal-hs.assetdetail.6186647.html

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059–1078. https://doi.org/10.1002/asi.22634

Dallmeier-Tiessen, S., Darby, R., Gitmans, K., Lambert, S., Matthews, B., Mele, S., … Wilson, M. (2014). Enabling Sharing and Reuse of Scientific Data. New Review of Information Networking, 19(1), 16–43. https://doi.org/10.1080/13614576.2014.883936

DFG. (2017, February 2). DFG Classification of Scientific Disciplines, Research Areas, Review Boards and Subject Areas (2016-2019). Retrieved from http://www.dfg.de/download/pdf/dfg_im_profil/gremien/fachkollegien/amtsperiode_2016_2019/fachsystematik_2016-2019_en_grafik.pdf

Ember, C., Hanisch, R., Alter, G., Berman, H., Hedstrom, M., & Vardigan, M. (2013). Sustaining Domain Repositories for Digital Data: A White Paper. https://doi.org/10.3886/sustainingdomainrepositoriesdigitaldata

Eynden, V. V. den, Knight, G., Vlad, A., Radler, B., Tenopir, C., Leon, D., … Corti, L. (2016, October 31). Survey of Wellcome researchers and their attitudes to open research. https://doi.org/10.6084/m9.figshare.4055448.v1

Fecher, B., Friesike, S., & Hebing, M. (2015). What Drives Academic Data Sharing? PLOS ONE, 10(2), 25. https://doi.org/10.1371/journal.pone.0118053

Fecher, B., Friesike, S., Hebing, M., Linek, S., & Sauermann, A. (2015). A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing (DIW Berlin No. 1454). Retrieved from Deutsches Institut für Wirtschaftsforschung e.V. website: http://hdl.handle.net/10419/107687

Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E., & Martone, M. E. (2014). Big data from small data: data-sharing in the ‘long tail’ of neuroscience. Nature Neuroscience, 17(11), 1442–1447. https://doi.org/10.1038/nn.3838

Ferguson, L. (2014). How and why researchers share data (and why they don’t). Retrieved from Wiley website: http://exchanges.wiley.com/blog/2014/11/03/how-and-why- researchers-share-data-and-why-they-dont/

Goldstein, S. (2017). The evolving landscape of Federated Research Data Infrastructures - Final report on the situation in the six Knowledge Exchange partner countries (p. 40). Retrieved from Knowledge Exchange, Jisc website: www.informall.org.uk

Hahnel, M., Fane, B., Treadway, J., Baynes, G., Wilkinson, R., Mons, B., … Osipov, I. (2018). State of Open Data 2018. Retrieved from figshare website: https://doi.org/10.6084/m9.figshare.7195058.v1

Hahnel, M., Treadway, J., Fane, B., Kiley, R., Peters, D., & Baynes, G. (2017, October 23). The State of Open Data Report 2017 (Digital Science, Ed.). https://doi.org/10.6084/m9.figshare.5481187.v1

Kim, Y. (2016). Scientists’ Data Sharing Behaviors [Data set]. https://doi.org/10.3886/E100087V7

Kim, Y. (2017). Scientific Data Reuse Survey [Data set]. https://doi.org/10.3886/E100404V1

Kim, Y., & Stanton, J. M. (2012). Institutional and individual influences on scientists' data sharing practices. Journal of Computational Science Education, 3(1), 47–56. https://doi.org/doi:10.1234/12345678

Kim, Y., & Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. Journal of the Association for Information Science and Technology, 67(4), 776–799. https://doi.org/10.1002/asi.23424

Kim, Y., & Yoon, A. (2017). Scientists' data reuse behaviors: A multilevel analysis. Journal of the Association for Information Science and Technology, 68(12), 2709–2719. https://doi.org/10.1002/asi.23892

Kim, Y., & Zhang, P. (2015). Understanding data sharing behaviors of STEM researchers: The roles of attitudes, norms, and data repositories. Library & Information Science Research, 37(3), 189–200. https://doi.org/10.1016/j.lisr.2015.04.006

Kindling, M., Pampel, H., van de Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., … Scholze, F. (2017). The Landscape of Research Data Repositories in 2015: A re3data Analysis. D-Lib Magazine, 23(3/4). https://doi.org/10.1045/march2017-kindling

Knoth, P., & Pontika, N. (2015, September 2). Open Science Taxonomy. https://doi.org/10.6084/m9.figshare.1508606.v3

Linek, S. B., Fecher, B., Friesike, S., & Hebing, M. (2017). Data sharing as social dilemma: Influence of the researcher's personality. PLOS ONE, 12(8), e0183216. https://doi.org/10.1371/journal.pone.0183216

Luff, R., Byatt, D., & Martin, D. (2015). Review of the Typology of Research Methods within the Social Sciences [Working Paper]. Retrieved from National Centre for Research Methods website: http://eprints.ncrm.ac.uk/3721/

National Institutes of Health (NIH) (Ed.). (2003, March 5). NIH Data Sharing Policy and Implementation Guidance. Retrieved from https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

Nature Research (Ed.). (2016, October 14). Open Data Survey 2016 [Dataset]. Retrieved from https://figshare.com/articles/Open_Data_Survey/4010541

OECD (Ed.). (2018). PPPs and exchange rates. Retrieved from https://stats.oecd.org/Index.aspx?DataSetCode=SNA_TABLE4#

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the Reuse of Scientific Data. 16, 8. https://doi.org/10.5334/dsj-2017-008

Perrier, L., Blondal, E., Ayala, A. P., Dearborn, D., Kenny, T., Lightfoot, D., … MacDonald, H. (2017). Dataset for: Research data management in academic institutions: a scoping review [Data set]. https://doi.org/10.5281/zenodo.557043

Pickard, A., & Dixon, P. (2004). The applicability of constructivist user studies: how can constructivist inquiry inform service providers and systems designers? Information Research, 9(3), paper 175. Retrieved from http://www.informationr.net/ir/9-3/paper175.html

SNSF. (2016, January). Research Domains and Disciplines. Retrieved from http://www.snf.ch/SiteCollectionDocuments/allg_disziplinenliste.pdf

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. PLOS ONE, 6(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., … Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLOS ONE, 10(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

The Data Harvest Report – sharing data for knowledge, jobs and growth. (2014, December 3). Retrieved 11 August 2018, from RDA website: https://www.rd-alliance.org/data-harvest-report-sharing-data-knowledge-jobs-and-growth.html

Treadway, J., Hahnel, M., Leonelli, S., Penny, D., Groenewegen, D., Miyairi, N., … Hook, D. (2016, October 25). The State of Open Data Report. https://doi.org/10.6084/m9.figshare.4036398.v1

Treloar, A., Groenewegen, D., & Harboe-Ree, C. (2007). The Data Curation Continuum: Managing Data Objects in Institutional Repositories. D-Lib Magazine, 13(9/10). https://doi.org/10.1045/september2007-treloar

Vessey, I., Ramesh, V., & Glass, R. L. (2005). A unified classification system for research in the computing disciplines. Information and Software Technology, 47(4), 245–255. https://doi.org/10.1016/j.infsof.2004.08.006

von der Heyde, M. (2019a). International Open Data Repository Survey: Description of collection, collected data, and analysis methods [Data paper]. Retrieved from https://doi.org/10.5281/zenodo.2643450

von der Heyde, M. (2019b). Open Data Landscape: Repository Usage of the Swiss Research Community: Description of collection, collected data, and analysis methods [Data paper]. Retrieved from https://doi.org/10.5281/zenodo.2643430

von der Heyde, M. (2019c, April). Data and tools of the landscape and cost analysis of data repositories currently used by the Swiss research community. Retrieved from https://doi.org/10.5281/zenodo.2643495

von der Heyde, M. (2019d, April). Data from the International Open Data Repository Survey. Retrieved from https://doi.org/10.5281/zenodo.2643493

von der Heyde, M. (2019e, April). Data from the Swiss Open Data Repository Landscape survey. Retrieved from https://doi.org/10.5281/zenodo.2643487

von der Heyde, M., Hartman, A., Auth, G., & Erfurth, C. (2018). Forschung in der disruptiven Digitalisierung von Hochschulen - Faktoren der Skalierung und ein Zukunftsszenario. Informatik Spektrum, 41(6), 359–368. https://doi.org/10.1007/s00287-018-01126-1

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLOS ONE, 8(7), e67332. https://doi.org/10.1371/journal.pone.0067332

Warner, A. (2018). Disciplines: Digital Commons Three-Tiered Taxonomy of Academic Disciplines (p. 28). Retrieved from bepress website: https://www.bepress.com/reference_guide_dc/disciplines/

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016, March 15). The FAIR Guiding Principles for scientific data management and stewardship [Comments and Opinion]. https://doi.org/10.1038/sdata.2016.18

Yoon, A., & Kim, Y. (2017). Social scientists' data reuse behaviors: Exploring the roles of attitudinal beliefs, attitudes, norms, and data repositories. Library & Information Science Research, 39(3), 224–233. https://doi.org/10.1016/j.lisr.2017.07.008

Appendix B: Data sources

We thank the various data providers for their help and access to the data bases. Often we were able to perform database queries using API definitions referred to in Table 8. The respective formats were analysed locally using custom-made tools.

All data were harvested using custom-made `bash` scripts, adjusting for the individual differences in the API. Lowlevel tools as `curl`, `xmlstarlet` and `xsltproc` were used to acquire and examine the raw xml data and translate them into flat files via XSLT-File definitions.

First frequency analytics were performed on the XML XPATH level, offering a quick overview across the schema and contained information. No flexible tool independent from the actual `xmlns` was found; therefore, a `bash` script (<130 lines) was developed.

Further data analytics were performed in *JMP* and in easy cases in `Excel`.

Database	API documentation	Export format	Date of retrieval and number of records used
re3data	https://www.re3data.org/api/doc	XML	2018-07-29: • 2,136 datasets
openAIRE	http://api.openaire.eu/overview.html	XML	1st to 5th of Aug. 2018: • projects [~2.5 million] • datasets [~807 K] • organizations [~127 K] • Registry of Research Data Repository [6,741]
OpenDOAR (provided by Jisc)	http://www.opendoar.org/tools/api.html	XML, JSON	2018-07-27: • 3,519 datasets
FAIRsharing	https://fairsharing.org/api	JSON	2018-10-02: • 1,673 datasets

Table 8: List of data providers, and their APIs and formats they offer. The number of datasets used is stated with reference to the time of harvest.

Appendix C: Supplementary Material

Extensive supplementary material for the landscape survey is provided within the SNSF community[11] on Zenodo (von der Heyde, 2019e) in conjunction with the data paper (von der Heyde, 2019b):

- Questionnaire as PDF: All questions in the appearance of the online survey.
- Questionnaire as JSON: Export from SurveyMonkey cloud platform including all options.
- Questionnaire in xlsx format: this includes references to other surveys using identical or near-identical questions.
- Anonymized raw data (CSV, xlsx).
- Final plausibility-checked data (CSV, xlsx).
- Additional analytical data sheets for text, biometrics and disciplinary mapping.

Additional supplementary material for the repository survey is provided within the SNSF community on Zenodo (von der Heyde, 2019d) in conjunction with the data paper (von der Heyde, 2019a):

- Questionnaire as PDF: All questions in the appearance of the online survey.
- Questionnaire as JSON: Export from SurveyMonkey cloud platform including all options.
- Questionnaire in xlsx format: this includes references to other surveys using identical or near-identical questions.
- Anonymized raw data (CSV, xlsx).
- Final plausibility-checked data (CSV, xlsx).
- Additional analytical data sheets.

Additional material for this report is also provided on Zenodo (von der Heyde, 2019c), in the following forms:

- JMP scripts for generating statistical analyses and graphs.
- bepress taxonomy mapping to DFG and SNSF disciplines in xlsx format.
- A list of scientific methods and their mapping to categories in xlsx format.

[11] See https://zenodo.org/communities/snsf/.

Appendix D: Mapping of Scientific Methods

Key	Label	Qualitative Methods	Quantitative Methods	Meta Methods	Analytical Methods	Critical Methods	Speculative Methods	Creative Methods
M01	Action research							1
M02	Assertion				1			
M03	Behavioural research	1						
M04	Case studies	1						
M05	Co-creation							1
M06	Comparative and cross national research		1					
M07	Concept implementation (proof of concept)						1	
M08	Conceptual analysis						1	
M09	Cross-sectional research		1					
M10	Data analysis			1				
M11	Descriptive research	1						
M12	Dialectic interchange					1		
M13	Digital social research	1						
M14	Discovery						1	
M15	Epistemology					1		
M16	Ethnography	1						
M17	Evaluation research		1					
M18	Exegesis					1		
M19	Experimental research		1					
M20	Explanatory research and causal analysis				1			
M21	Exploratory research						1	
M22	Field experiments		1					
M23	Field studies	1						
M24	Formal concept analysis				1			
M25	Fringe science						1	
M26	Grounded theory				1			
M27	Hermeneutics					1		
M28	Instrument development			1				
M29	Interdisciplinary and multidisciplinary research			1				
M30	Interpretation					1		
M31	Intervention studies		1					
M32	Laboratory experiments (human subjects)		1					
M33	Laboratory experiments (technical)		1					
M34	Literature reviews					1		
M35	Longitudinal research		1					
M36	Mathematical proofs				1			
M37	Meta-analysis			1				
M38	Mixed methods			1				
M39	Ontology				1			
M40	Operationalization				1			
M41	Participatory research							1
M42	Pilot studies						1	
M43	Protocol analysis				1			
M44	Quasi-experimental research						1	
M45	Secondary analysis			1				
M46	Semiotics					1		
M47	Simulations			1				
M48	Survey research		1					
M49	Systematic reviews					1		
M50	Verification/falsification of hypotheses				1			
M51	Other (please specify)							

Table 9: List of scientific methods and their mapping to abstract classes.

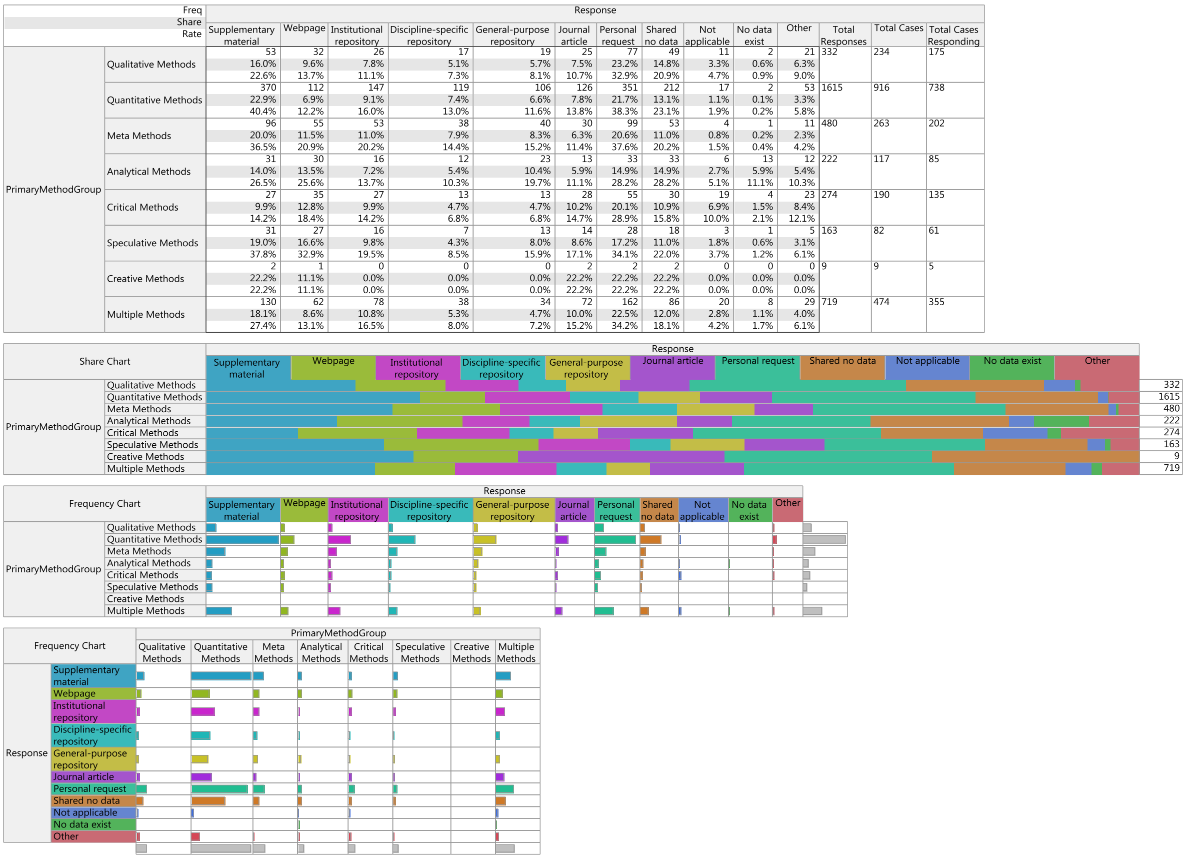
The participants in the landscape survey were invited to select all of the methods that applied to their scientific work from those shown in the “Label” column.

Appendix E: Categorical analyses

The categorical analyses are presented in the following order:

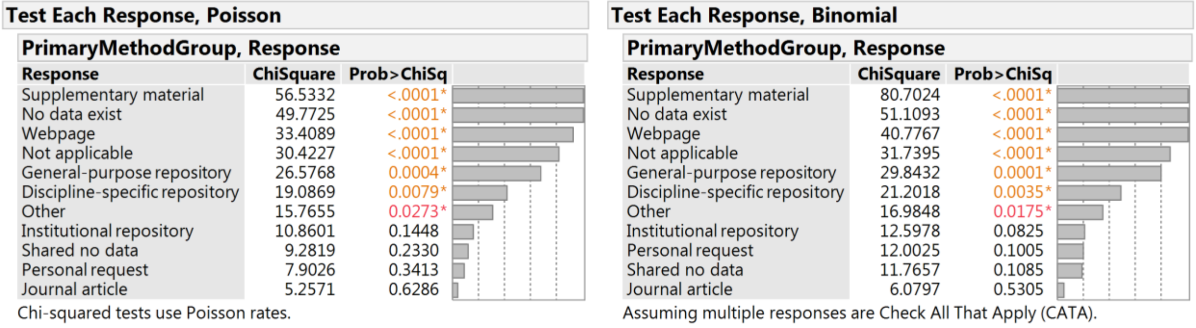
1. Ways of sharing
    a. primary method mapping
    b. primary DFG discipline Level-2 mapping
    c. primary SNSF discipline Level-2 mapping

In addition to each categorical analysis, we performed a Poisson count test and a binomial homogeneity test on these distributions.



		Freq Share Rate	Response													
			Supplementary material	Webpage	Institutional repository	Discipline-specific repository	General-purpose repository	Journal article	Personal request	Shared no data	Not applicable	No data exist	Other	Total Responses	Total Cases	Total Cases Responding
PrimaryMethodGroup	Qualitative Methods		53	32	26	17	19	25	77	49	11	2	21	332	234	175
			16.0%	9.6%	7.8%	5.1%	5.7%	7.5%	23.2%	14.8%	3.3%	0.6%	6.3%			
			22.6%	13.7%	11.1%	7.3%	8.1%	10.7%	32.9%	20.9%	4.7%	0.9%	9.0%			
	Quantitative Methods		370	112	147	119	106	126	351	212	17	2	53	1615	916	738
			22.9%	6.9%	9.1%	7.4%	6.6%	7.8%	21.7%	13.1%	1.1%	0.1%	3.3%			
			40.4%	12.2%	16.0%	13.0%	11.6%	13.8%	38.3%	23.1%	1.9%	0.2%	5.8%			
	Meta Methods		96	55	53	38	40	30	99	53	4	1	11	480	263	202
			20.0%	11.5%	11.0%	7.9%	8.3%	6.3%	20.6%	11.0%	0.8%	0.2%	2.3%			
			36.5%	20.9%	20.2%	14.4%	15.2%	11.4%	37.6%	20.2%	1.5%	0.4%	4.2%			
	Analytical Methods		31	30	16	12	23	13	33	33	6	13	12	222	117	85
			14.0%	13.5%	7.2%	5.4%	10.4%	5.9%	14.9%	14.9%	2.7%	5.9%	5.4%			
			26.5%	25.6%	13.7%	10.3%	19.7%	11.1%	28.2%	28.2%	5.1%	11.1%	10.3%			
	Critical Methods		27	35	27	13	13	28	55	30	19	4	23	274	190	135
			9.9%	12.8%	9.9%	4.7%	4.7%	10.2%	20.1%	10.9%	6.9%	1.5%	8.4%			
			14.2%	18.4%	14.2%	6.8%	6.8%	14.7%	28.9%	15.8%	10.0%	2.1%	12.1%			
	Speculative Methods		31	27	16	7	13	14	28	18	3	1	5	163	82	61
			19.0%	16.6%	9.8%	4.3%	8.0%	8.6%	17.2%	11.0%	1.8%	0.6%	3.1%			
			37.8%	32.9%	19.5%	8.5%	15.9%	17.1%	34.1%	22.0%	3.7%	1.2%	6.1%			
	Creative Methods		2	1	0	0	0	2	2	2	0	0	0	9	9	5
			22.2%	11.1%	0.0%	0.0%	0.0%	22.2%	22.2%	22.2%	0.0%	0.0%	0.0%			
			22.2%	11.1%	0.0%	0.0%	0.0%	22.2%	22.2%	22.2%	0.0%	0.0%	0.0%			
	Multiple Methods		130	62	78	38	34	72	162	86	20	8	29	719	474	355
			18.1%	8.6%	10.8%	5.3%	4.7%	10.0%	22.5%	12.0%	2.8%	1.1%	4.0%			
			27.4%	13.1%	16.5%	8.0%	7.2%	15.2%	34.2%	18.1%	4.2%	1.7%	6.1%			

Figure 47: Categorical test of 'ways of sharing' by primary research method.



Test Each Response, Poisson

PrimaryMethodGroup, Response

Response	ChiSquare	Prob>ChiSq
Supplementary material	56.5332	<.0001*
No data exist	49.7725	<.0001*
Webpage	33.4089	<.0001*
Not applicable	30.4227	<.0001*
General-purpose repository	26.5768	0.0004*
Discipline-specific repository	19.0869	0.0079*
Other	15.7655	0.0273*
Institutional repository	10.8601	0.1448
Shared no data	9.2819	0.2330
Personal request	7.9026	0.3413
Journal article	5.2571	0.6286

Chi-squared tests use Poisson rates.

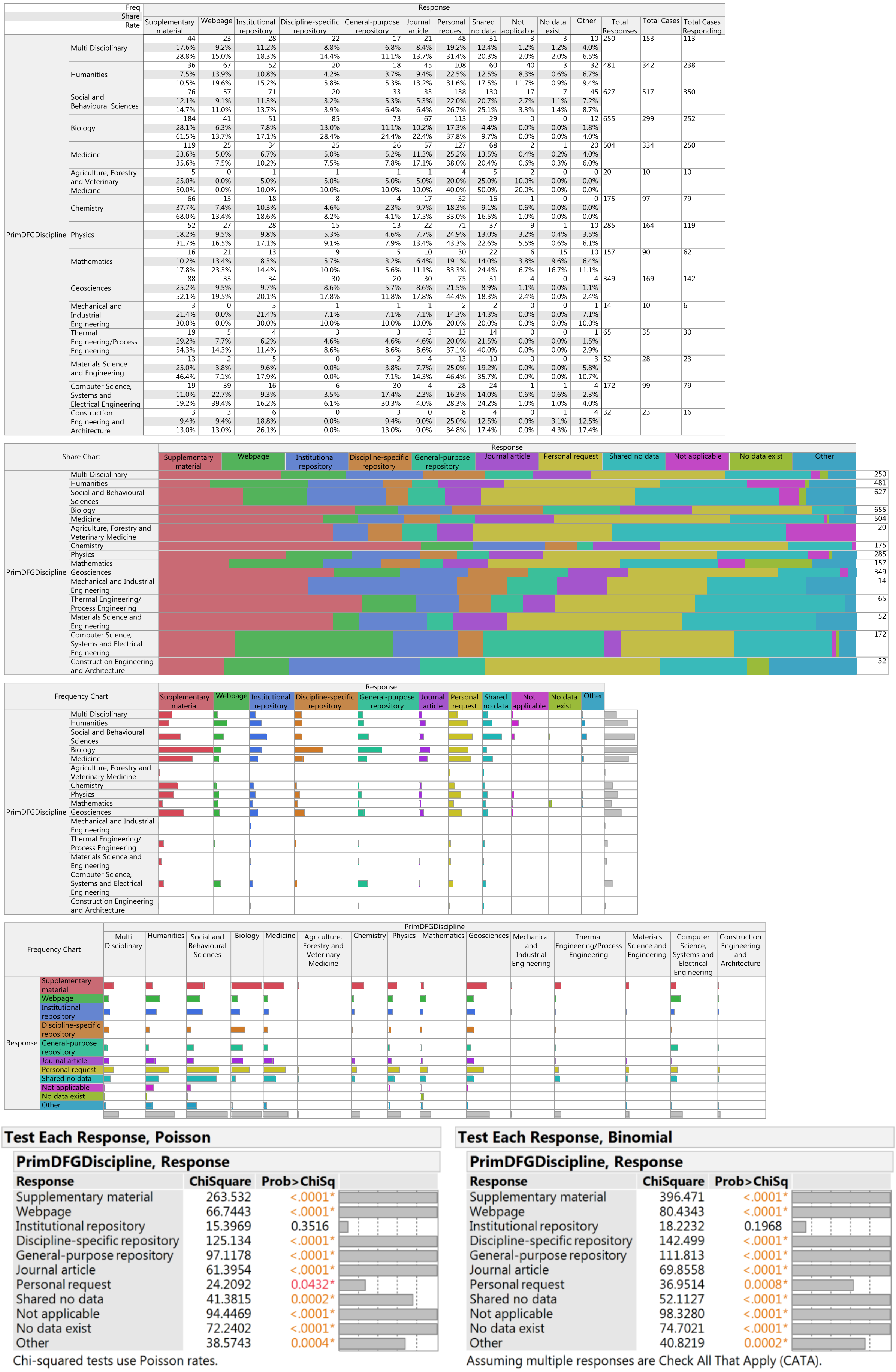
Test Each Response, Binomial

PrimaryMethodGroup, Response

Response	ChiSquare	Prob>ChiSq
Supplementary material	80.7024	<.0001*
No data exist	51.1093	<.0001*
Webpage	40.7767	<.0001*
Not applicable	31.7395	<.0001*
General-purpose repository	29.8432	0.0001*
Discipline-specific repository	21.2018	0.0035*
Other	16.9848	0.0175*
Institutional repository	12.5978	0.0825
Personal request	12.0025	0.1005
Shared no data	11.7657	0.1085
Journal article	6.0797	0.5305

Assuming multiple responses are Check All That Apply (CATA).

Figure 48: Tests for homogeneous distributions of 'ways of sharing' across the primary research methods. Significant results are expected for violations of the homogeneity assumption.



Freq Share Rate		Response													
		Supplementary material	Webpage	Institutional repository	Discipline-specific repository	General-purpose repository	Journal article	Personal request	Shared no data	Not applicable	No data exist	Other	Total Responses	Total Cases	Total Cases Responding
PrimDFGDiscipline	Multi Disciplinary	44	23	28	22	17	21	48	31	3	3	10	250	153	113
		17.6%	9.2%	11.2%	8.8%	6.8%	8.4%	19.2%	12.4%	1.2%	1.2%	4.0%			
		28.8%	15.0%	18.3%	14.4%	11.1%	13.7%	31.4%	20.3%	2.0%	2.0%	6.5%			
	Humanities	36	67	52	20	18	45	108	60	40	3	32	481	342	238
		7.5%	13.9%	10.8%	4.2%	3.7%	9.4%	22.5%	12.5%	8.3%	0.6%	6.7%			
		10.5%	19.6%	15.2%	5.8%	5.3%	13.2%	31.6%	17.5%	11.7%	0.9%	9.4%			
	Social and Behavioural Sciences	76	57	71	20	33	33	138	130	17	7	45	627	517	350
		12.1%	9.1%	11.3%	3.2%	5.3%	5.3%	22.0%	20.7%	2.7%	1.1%	7.2%			
		14.7%	11.0%	13.7%	3.9%	6.4%	6.4%	26.7%	25.1%	3.3%	1.4%	8.7%			
	Biology	184	41	51	85	73	67	113	29	0	0	12	655	299	252
		28.1%	6.3%	7.8%	13.0%	11.1%	10.2%	17.3%	4.4%	0.0%	0.0%	1.8%			
		61.5%	13.7%	17.1%	28.4%	24.4%	22.4%	37.8%	9.7%	0.0%	0.0%	4.0%			
	Medicine	119	25	34	25	26	57	127	68	2	1	20	504	334	250
		23.6%	5.0%	6.7%	5.0%	5.2%	11.3%	25.2%	13.5%	0.4%	0.2%	4.0%			
		35.6%	7.5%	10.2%	7.5%	7.8%	17.1%	38.0%	20.4%	0.6%	0.3%	6.0%			
	Agriculture, Forestry and Veterinary Medicine	5	0	1	1	1	1	4	5	2	0	0	20	10	10
		25.0%	0.0%	5.0%	5.0%	5.0%	5.0%	20.0%	25.0%	10.0%	0.0%	0.0%			
		50.0%	0.0%	10.0%	10.0%	10.0%	10.0%	40.0%	50.0%	20.0%	0.0%	0.0%			
	Chemistry	66	13	18	8	4	17	32	16	1	0	0	175	97	79
		37.7%	7.4%	10.3%	4.6%	2.3%	9.7%	18.3%	9.1%	0.6%	0.0%	0.0%			
		68.0%	13.4%	18.6%	8.2%	4.1%	17.5%	33.0%	16.5%	1.0%	0.0%	0.0%			
	Physics	52	27	28	15	13	22	71	37	9	1	10	285	164	119
		18.2%	9.5%	9.8%	5.3%	4.6%	7.7%	24.9%	13.0%	3.2%	0.4%	3.5%			
		31.7%	16.5%	17.1%	9.1%	7.9%	13.4%	43.3%	22.6%	5.5%	0.6%	6.1%			
	Mathematics	16	21	13	9	5	10	30	22	6	15	10	157	90	62
		10.2%	13.4%	8.3%	5.7%	3.2%	6.4%	19.1%	14.0%	3.8%	9.6%	6.4%			
		17.8%	23.3%	14.4%	10.0%	5.6%	11.1%	33.3%	24.4%	6.7%	16.7%	11.1%			
	Geosciences	88	33	34	30	20	30	75	31	4	0	4	349	169	142
		25.2%	9.5%	9.7%	8.6%	5.7%	8.6%	21.5%	8.9%	1.1%	0.0%	1.1%			
		52.1%	19.5%	20.1%	17.8%	11.8%	17.8%	44.4%	18.3%	2.4%	0.0%	2.4%			
	Mechanical and Industrial Engineering	3	0	3	1	1	1	2	2	0	0	1	14	10	6
		21.4%	0.0%	21.4%	7.1%	7.1%	7.1%	14.3%	14.3%	0.0%	0.0%	7.1%			
		30.0%	0.0%	30.0%	10.0%	10.0%	10.0%	20.0%	20.0%	0.0%	0.0%	10.0%			
	Thermal Engineering/Process Engineering	19	5	4	3	3	3	13	14	0	0	1	65	35	30
		29.2%	7.7%	6.2%	4.6%	4.6%	4.6%	20.0%	21.5%	0.0%	0.0%	1.5%			
		54.3%	14.3%	11.4%	8.6%	8.6%	8.6%	37.1%	40.0%	0.0%	0.0%	2.9%			
	Materials Science and Engineering	13	2	5	0	2	4	13	10	0	0	3	52	28	23
		25.0%	3.8%	9.6%	0.0%	3.8%	7.7%	25.0%	19.2%	0.0%	0.0%	5.8%			
		46.4%	7.1%	17.9%	0.0%	7.1%	14.3%	46.4%	35.7%	0.0%	0.0%	10.7%			
	Computer Science, Systems and Electrical Engineering	19	39	16	6	30	4	28	24	1	1	4	172	99	79
		11.0%	22.7%	9.3%	3.5%	17.4%	2.3%	16.3%	14.0%	0.6%	0.6%	2.3%			
		19.2%	39.4%	16.2%	6.1%	30.3%	4.0%	28.3%	24.2%	1.0%	1.0%	4.0%			
	Construction Engineering and Architecture	3	3	6	0	3	0	8	4	0	1	4	32	23	16
		9.4%	9.4%	18.8%	0.0%	9.4%	0.0%	25.0%	12.5%	0.0%	3.1%	12.5%			
		13.0%	13.0%	26.1%	0.0%	13.0%	0.0%	34.8%	17.4%	0.0%	4.3%	17.4%			

Test Each Response, Poisson

PrimDFGDiscipline, Response

Response	ChiSquare	Prob>ChiSq
Supplementary material	263.532	<.0001*
Webpage	66.7443	<.0001*
Institutional repository	15.3969	0.3516
Discipline-specific repository	125.134	<.0001*
General-purpose repository	97.1178	<.0001*
Journal article	61.3954	<.0001*
Personal request	24.2092	0.0432*
Shared no data	41.3815	0.0002*
Not applicable	94.4469	<.0001*
No data exist	72.2402	<.0001*
Other	38.5743	0.0004*

Chi-squared tests use Poisson rates.

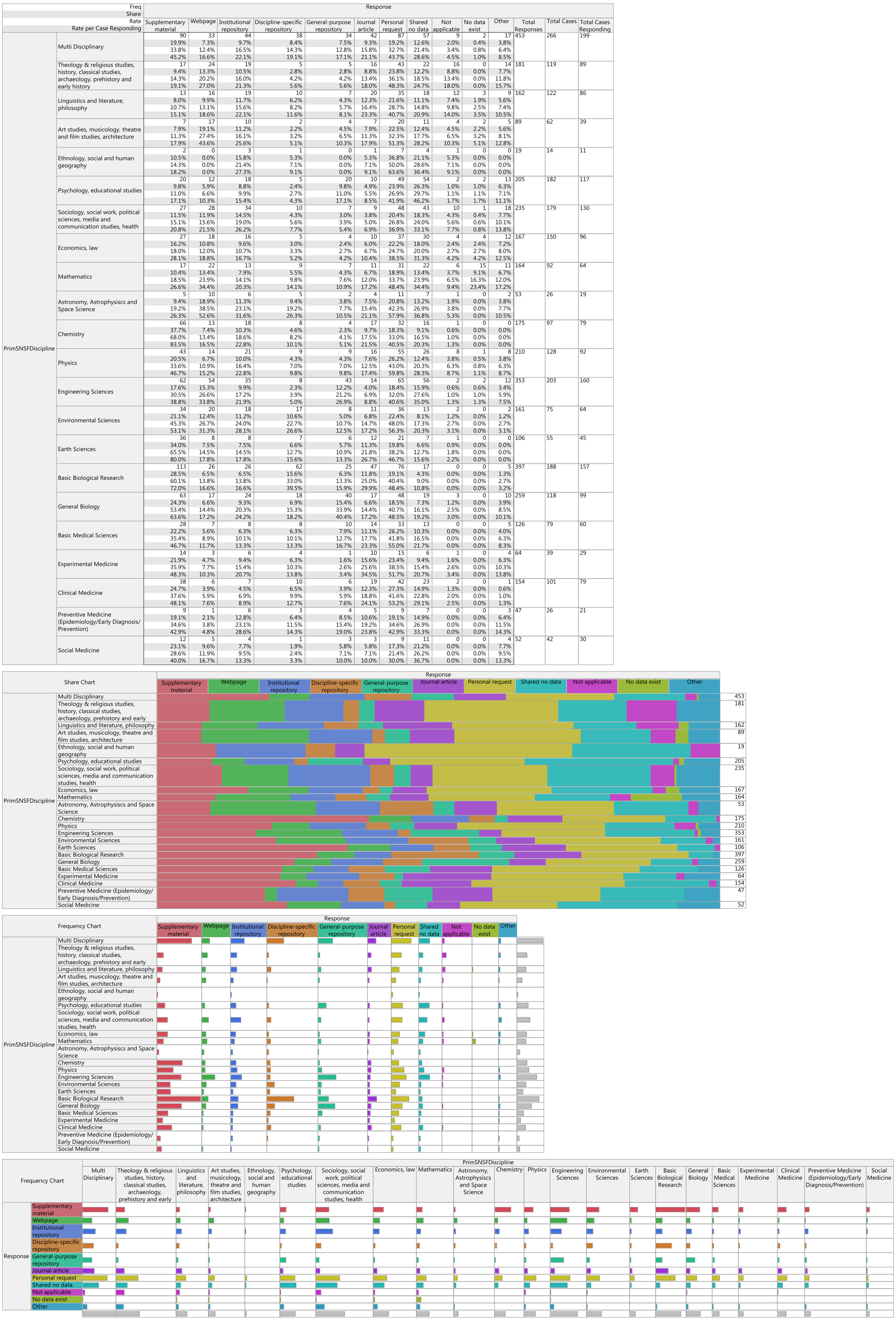
Test Each Response, Binomial

PrimDFGDiscipline, Response

Response	ChiSquare	Prob>ChiSq
Supplementary material	396.471	<.0001*
Webpage	80.4343	<.0001*
Institutional repository	18.2232	0.1968
Discipline-specific repository	142.499	<.0001*
General-purpose repository	111.813	<.0001*
Journal article	69.8558	<.0001*
Personal request	36.9514	0.0008*
Shared no data	52.1127	<.0001*
Not applicable	98.3280	<.0001*
No data exist	74.7021	<.0001*
Other	40.8219	0.0002*

Assuming multiple responses are Check All That Apply (CATA).

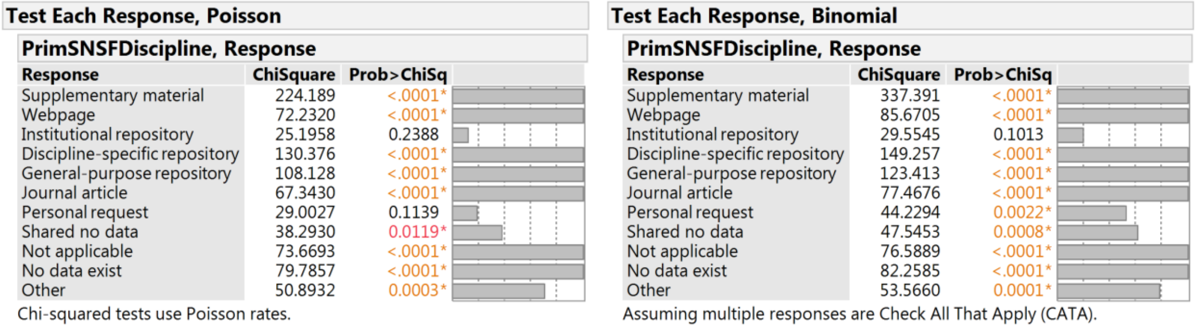
Figure 49: Categorical tests of 'ways of sharing' by primary DFG discipline, including tests for homogeneous distributions.



		Freq Share Rate Rate per Case Responding	Response												
			Supplementary material	Webpage	Institutional repository	Discipline-specific repository	General-purpose repository	Journal article	Personal request	Shared no data	Not applicable	No data exist	Other	Total Responses	Total Cases	Total Cases Responding
PrimSNSFDiscipline	Multi Disciplinary		90	33	44	38	34	42	87	57	9	2	17	453	266	199
			19.9%	7.3%	9.7%	8.4%	7.5%	9.3%	19.2%	12.6%	2.0%	0.4%	3.8%			
			33.8%	12.4%	16.5%	14.3%	12.8%	15.8%	32.7%	21.4%	3.4%	0.8%	6.4%			
			45.2%	16.6%	22.1%	19.1%	17.1%	21.1%	43.7%	28.6%	4.5%	1.0%	8.5%			
	Theology & religious studies, history, classical studies, archaeology, prehistory and early history		17	24	19	5	5	16	43	22	16	0	14	181	119	89
			9.4%	13.3%	10.5%	2.8%	2.8%	8.8%	23.8%	12.2%	8.8%	0.0%	7.7%			
			14.3%	20.2%	16.0%	4.2%	4.2%	13.4%	36.1%	18.5%	13.4%	0.0%	11.8%			
			19.1%	27.0%	21.3%	5.6%	5.6%	18.0%	48.3%	24.7%	18.0%	0.0%	15.7%			
	Linguistics and literature, philosophy		13	16	19	10	7	20	35	18	12	3	9	162	122	86
			8.0%	9.9%	11.7%	6.2%	4.3%	12.3%	21.6%	11.1%	7.4%	1.9%	5.6%			
			10.7%	13.1%	15.6%	8.2%	5.7%	16.4%	28.7%	14.8%	9.8%	2.5%	7.4%			
			15.1%	18.6%	22.1%	11.6%	8.1%	23.3%	40.7%	20.9%	14.0%	3.5%	10.5%			
	Art studies, musicology, theatre and film studies, architecture		7	17	10	2	4	7	20	11	4	2	5	89	62	39
			7.9%	19.1%	11.2%	2.2%	4.5%	7.9%	22.5%	12.4%	4.5%	2.2%	5.6%			
			11.3%	27.4%	16.1%	3.2%	6.5%	11.3%	32.3%	17.7%	6.5%	3.2%	8.1%			
			17.9%	43.6%	25.6%	5.1%	10.3%	17.9%	51.3%	28.2%	10.3%	5.1%	12.8%			
	Ethnology, social and human geography		2	0	3	1	0	1	7	4	1	0	0	19	14	11
			10.5%	0.0%	15.8%	5.3%	0.0%	5.3%	36.8%	21.1%	5.3%	0.0%	0.0%			
			14.3%	0.0%	21.4%	7.1%	0.0%	7.1%	50.0%	28.6%	7.1%	0.0%	0.0%			
			18.2%	0.0%	27.3%	9.1%	0.0%	9.1%	63.6%	36.4%	9.1%	0.0%	0.0%			
	Psychology, educational studies		20	12	18	5	20	10	49	54	2	2	13	205	182	117
			9.8%	5.9%	8.8%	2.4%	9.8%	4.9%	23.9%	26.3%	1.0%	1.0%	6.3%			
			11.0%	6.6%	9.9%	2.7%	11.0%	5.5%	26.9%	29.7%	1.1%	1.1%	7.1%			
			17.1%	10.3%	15.4%	4.3%	17.1%	8.5%	41.9%	46.2%	1.7%	1.7%	11.1%			
	Sociology, social work, political sciences, media and communication studies, health		27	28	34	10	7	9	48	43	10	1	18	235	179	130
			11.5%	11.9%	14.5%	4.3%	3.0%	3.8%	20.4%	18.3%	4.3%	0.4%	7.7%			
			15.1%	15.6%	19.0%	5.6%	3.9%	5.0%	26.8%	24.0%	5.6%	0.6%	10.1%			
			20.8%	21.5%	26.2%	7.7%	5.4%	6.9%	36.9%	33.1%	7.7%	0.8%	13.8%			
	Economics, law		27	18	16	5	4	10	37	30	4	4	12	167	150	96
			16.2%	10.8%	9.6%	3.0%	2.4%	6.0%	22.2%	18.0%	2.4%	2.4%	7.2%			
			18.0%	12.0%	10.7%	3.3%	2.7%	6.7%	24.7%	20.0%	2.7%	2.7%	8.0%			
			28.1%	18.8%	16.7%	5.2%	4.2%	10.4%	38.5%	31.3%	4.2%	4.2%	12.5%			
	Mathematics		17	22	13	9	7	11	31	22	6	15	11	164	92	64
			10.4%	13.4%	7.9%	5.5%	4.3%	6.7%	18.9%	13.4%	3.7%	9.1%	6.7%			
			18.5%	23.9%	14.1%	9.8%	7.6%	12.0%	33.7%	23.9%	6.5%	16.3%	12.0%			
			26.6%	34.4%	20.3%	14.1%	10.9%	17.2%	48.4%	34.4%	9.4%	23.4%	17.2%			
	Astronomy, Astrophysiscs and Space Science		5	10	6	5	2	4	11	7	1	0	2	53	26	19
			9.4%	18.9%	11.3%	9.4%	3.8%	7.5%	20.8%	13.2%	1.9%	0.0%	3.8%			
			19.2%	38.5%	23.1%	19.2%	7.7%	15.4%	42.3%	26.9%	3.8%	0.0%	7.7%			
			26.3%	52.6%	31.6%	26.3%	10.5%	21.1%	57.9%	36.8%	5.3%	0.0%	10.5%			
	Chemistry		66	13	18	8	4	17	32	16	1	0	0	175	97	79
			37.7%	7.4%	10.3%	4.6%	2.3%	9.7%	18.3%	9.1%	0.6%	0.0%	0.0%			
			68.0%	13.4%	18.6%	8.2%	4.1%	17.5%	33.0%	16.5%	1.0%	0.0%	0.0%			
			83.5%	16.5%	22.8%	10.1%	5.1%	21.5%	40.5%	20.3%	1.3%	0.0%	0.0%			
	Physics		43	14	21	9	9	16	55	26	8	1	8	210	128	92
			20.5%	6.7%	10.0%	4.3%	4.3%	7.6%	26.2%	12.4%	3.8%	0.5%	3.8%			
			33.6%	10.9%	16.4%	7.0%	7.0%	12.5%	43.0%	20.3%	6.3%	0.8%	6.3%			
			46.7%	15.2%	22.8%	9.8%	9.8%	17.4%	59.8%	28.3%	8.7%	1.1%	8.7%			
	Engineering Sciences		62	54	35	8	43	14	65	56	2	2	12	353	203	160
			17.6%	15.3%	9.9%	2.3%	12.2%	4.0%	18.4%	15.9%	0.6%	0.6%	3.4%			
			30.5%	26.6%	17.2%	3.9%	21.2%	6.9%	32.0%	27.6%	1.0%	1.0%	5.9%			
			38.8%	33.8%	21.9%	5.0%	26.9%	8.8%	40.6%	35.0%	1.3%	1.3%	7.5%			
	Environmental Sciences		34	20	18	17	8	11	36	13	2	0	2	161	75	64
			21.1%	12.4%	11.2%	10.6%	5.0%	6.8%	22.4%	8.1%	1.2%	0.0%	1.2%			
			45.3%	26.7%	24.0%	22.7%	10.7%	14.7%	48.0%	17.3%	2.7%	0.0%	2.7%			
			53.1%	31.3%	28.1%	26.6%	12.5%	17.2%	56.3%	20.3%	3.1%	0.0%	3.1%			
	Earth Sciences		36	8	8	7	6	12	21	7	1	0	0	106	55	45
			34.0%	7.5%	7.5%	6.6%	5.7%	11.3%	19.8%	6.6%	0.9%	0.0%	0.0%			
			65.5%	14.5%	14.5%	12.7%	10.9%	21.8%	38.2%	12.7%	1.8%	0.0%	0.0%			
			80.0%	17.8%	17.8%	15.6%	13.3%	26.7%	46.7%	15.6%	2.2%	0.0%	0.0%			
	Basic Biological Research		113	26	26	62	25	47	76	17	0	0	5	397	188	157
			28.5%	6.5%	6.5%	15.6%	6.3%	11.8%	19.1%	4.3%	0.0%	0.0%	1.3%			
			60.1%	13.8%	13.8%	33.0%	13.3%	25.0%	40.4%	9.0%	0.0%	0.0%	2.7%			
			72.0%	16.6%	16.6%	39.5%	15.9%	29.9%	48.4%	10.8%	0.0%	0.0%	3.2%			
	General Biology		63	17	24	18	40	17	48	19	3	0	10	259	118	99
			24.3%	6.6%	9.3%	6.9%	15.4%	6.6%	18.5%	7.3%	1.2%	0.0%	3.9%			
			53.4%	14.4%	20.3%	15.3%	33.9%	14.4%	40.7%	16.1%	2.5%	0.0%	8.5%			
			63.6%	17.2%	24.2%	18.2%	40.4%	17.2%	48.5%	19.2%	3.0%	0.0%	10.1%			
	Basic Medical Sciences		28	7	8	8	10	14	33	13	0	0	5	126	79	60
			22.2%	5.6%	6.3%	6.3%	7.9%	11.1%	26.2%	10.3%	0.0%	0.0%	4.0%			
			35.4%	8.9%	10.1%	10.1%	12.7%	17.7%	41.8%	16.5%	0.0%	0.0%	6.3%			
			46.7%	11.7%	13.3%	13.3%	16.7%	23.3%	55.0%	21.7%	0.0%	0.0%	8.3%			
	Experimental Medicine		14	3	6	4	1	10	15	6	1	0	4	64	39	29
			21.9%	4.7%	9.4%	6.3%	1.6%	15.6%	23.4%	9.4%	1.6%	0.0%	6.3%			
			35.9%	7.7%	15.4%	10.3%	2.6%	25.6%	38.5%	15.4%	2.6%	0.0%	10.3%			
			48.3%	10.3%	20.7%	13.8%	3.4%	34.5%	51.7%	20.7%	3.4%	0.0%	13.8%			
	Clinical Medicine		38	6	7	10	6	19	42	23	2	0	1	154	101	79
			24.7%	3.9%	4.5%	6.5%	3.9%	12.3%	27.3%	14.9%	1.3%	0.0%	0.6%			
			37.6%	5.9%	6.9%	9.9%	5.9%	18.8%	41.6%	22.8%	2.0%	0.0%	1.0%			
			48.1%	7.6%	8.9%	12.7%	7.6%	24.1%	53.2%	29.1%	2.5%	0.0%	1.3%			
	Preventive Medicine (Epidemiology/Early Diagnosis/ Prevention)		9	1	6	3	4	5	9	7	0	0	3	47	26	21
			19.1%	2.1%	12.8%	6.4%	8.5%	10.6%	19.1%	14.9%	0.0%	0.0%	6.4%			
			34.6%	3.8%	23.1%	11.5%	15.4%	19.2%	34.6%	26.9%	0.0%	0.0%	11.5%			
			42.9%	4.8%	28.6%	14.3%	19.0%	23.8%	42.9%	33.3%	0.0%	0.0%	14.3%			
	Social Medicine		12	5	4	1	3	3	9	11	0	0	4	52	42	30
			23.1%	9.6%	7.7%	1.9%	5.8%	5.8%	17.3%	21.2%	0.0%	0.0%	7.7%			
			28.6%	11.9%	9.5%	2.4%	7.1%	7.1%	21.4%	26.2%	0.0%	0.0%	9.5%			
			40.0%	16.7%	13.3%	3.3%	10.0%	10.0%	30.0%	36.7%	0.0%	0.0%	13.3%			

Share Chart
Response
Supplementary material
Webpage
Institutional repository
Discipline-specific repository
General-purpose repository
Journal article
Personal request
Shared no data
Not applicable
No data exist
Other
PrimSNSFDiscipline
Multi Disciplinary
Theology & religious studies, history, classical studies, archaeology, prehistory and early
Linguistics and literature, philosophy
Art studies, musicology, theatre and film studies, architecture
Ethnology, social and human geography
Psychology, educational studies
Sociology, social work, political sciences, media and communication studies, health
Economics, law
Mathematics
Astronomy, Astrophysiscs and Space Science
Chemistry
Physics
Engineering Sciences
Environmental Sciences
Earth Sciences
Basic Biological Research
General Biology
Basic Medical Sciences
Experimental Medicine
Clinical Medicine
Preventive Medicine (Epidemiology/Early Diagnosis/Prevention)
Social Medicine
453
181
162
89
19
205
235
167
164
53
175
210
353
161
106
397
259
126
64
154
47
52
Frequency Chart
Response
PrimSNSFDiscipline
Frequency Chart
PrimSNSFDiscipline
Response

Figure 50: Categorical test of 'ways of sharing' by primary SNSF discipline.



Test Each Response, Poisson

PrimSNSFDiscipline, Response

Response	ChiSquare	Prob>ChiSq
Supplementary material	224.189	<.0001*
Webpage	72.2320	<.0001*
Institutional repository	25.1958	0.2388
Discipline-specific repository	130.376	<.0001*
General-purpose repository	108.128	<.0001*
Journal article	67.3430	<.0001*
Personal request	29.0027	0.1139
Shared no data	38.2930	0.0119*
Not applicable	73.6693	<.0001*
No data exist	79.7857	<.0001*
Other	50.8932	0.0003*

Chi-squared tests use Poisson rates.

Test Each Response, Binomial

PrimSNSFDiscipline, Response

Response	ChiSquare	Prob>ChiSq
Supplementary material	337.391	<.0001*
Webpage	85.6705	<.0001*
Institutional repository	29.5545	0.1013
Discipline-specific repository	149.257	<.0001*
General-purpose repository	123.413	<.0001*
Journal article	77.4676	<.0001*
Personal request	44.2294	0.0022*
Shared no data	47.5453	0.0008*
Not applicable	76.5889	<.0001*
No data exist	82.2585	<.0001*
Other	53.5660	0.0001*

Assuming multiple responses are Check All That Apply (CATA).

Figure 51: Tests for homogeneous distributions of ‘ways of sharing’ across the primary SNSF disciplines. Significant results are expected for violations of the homogeneity assumption.

Appendix F: Correlations

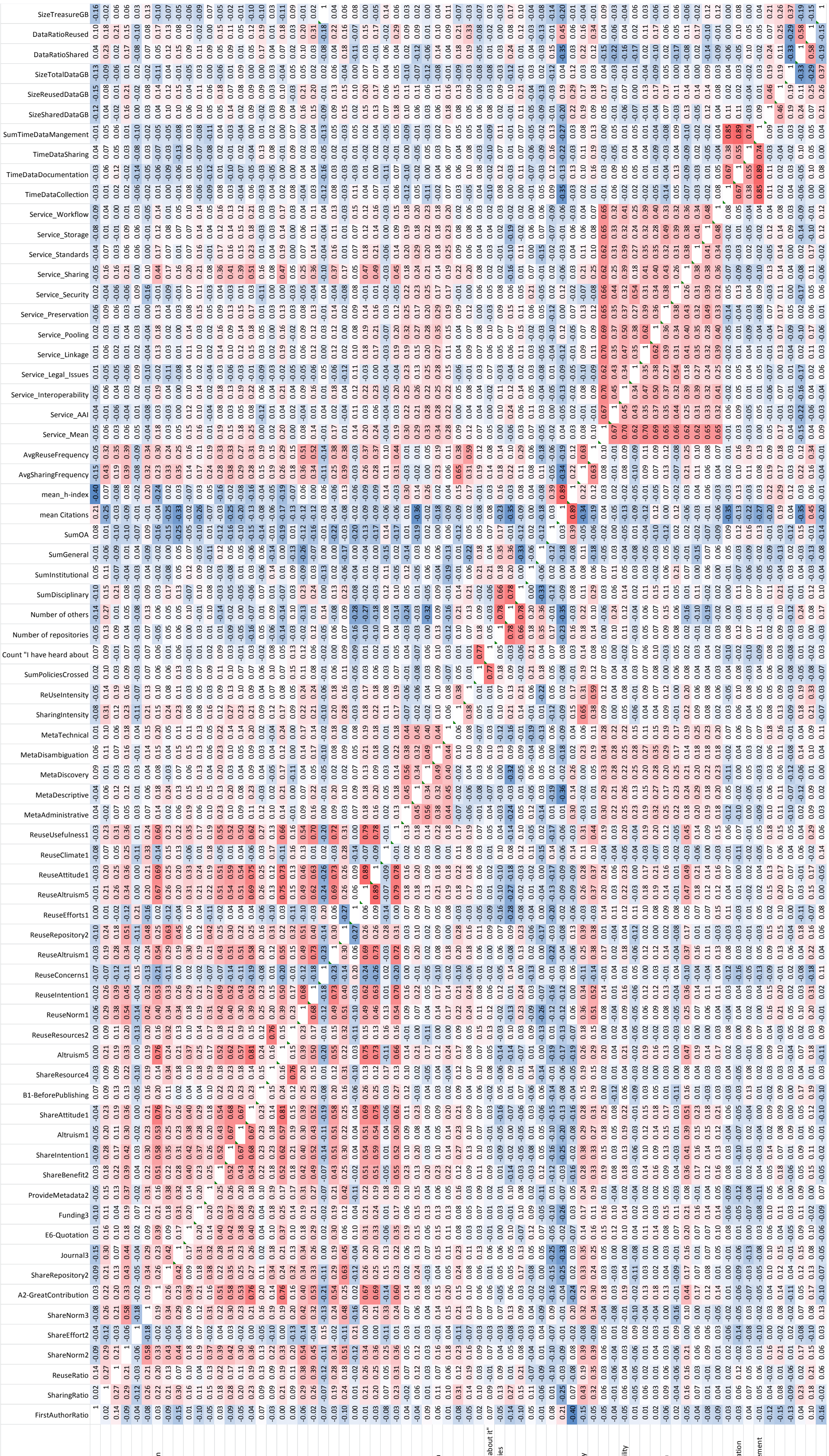


Table 10: Correlations of most ordinal variables. The PCA in the next appendix is based on those correlations. Different parts of the Landscape survey are clearly visible as blocks of self-correlation.

Appendix G: Principal component and factor analyses

A principal component analysis (PCA) was performed on all variables of sharing and reuse (see Table 3 and Table 4). Our results in general are very similar to the results of the previous literature, as discussed in sections 4.1and 4.2. The Factor Analysis (FA) was performed on 5 components, due to the selection in the Scree plot.

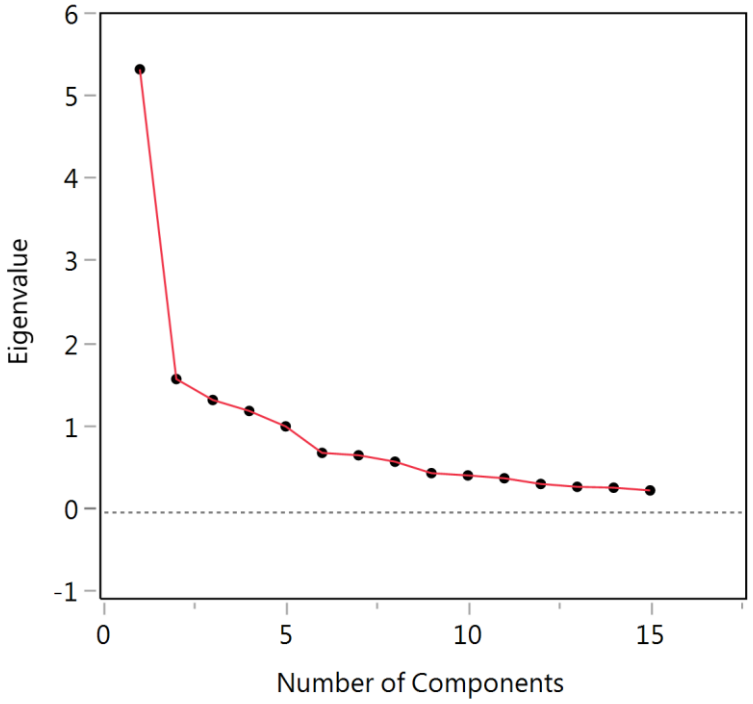
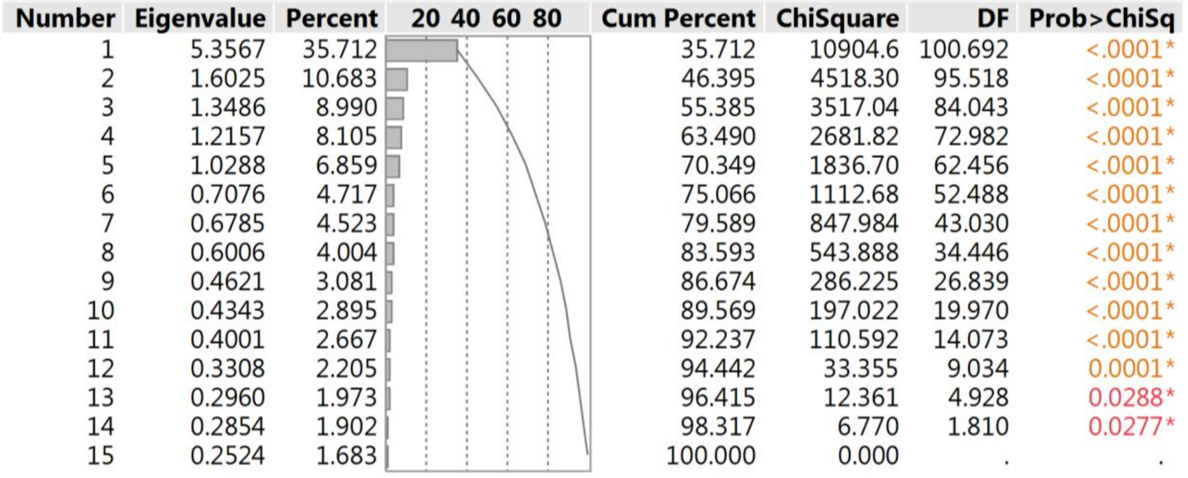
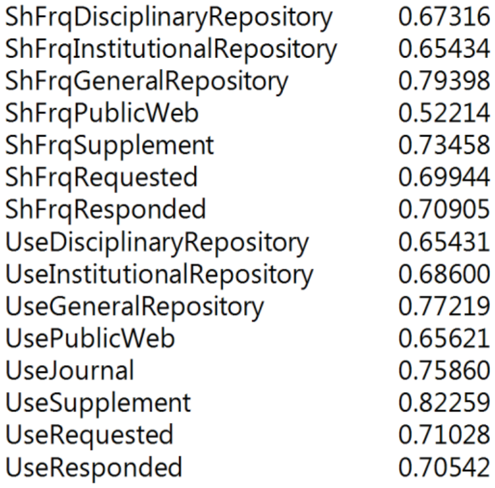


Figure 52: Scree plot principal components of sharing and reuse.



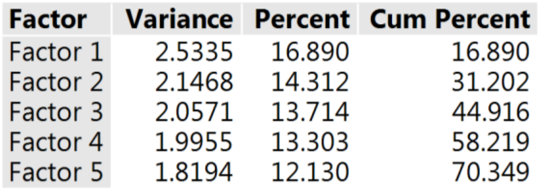
Number	Eigenvalue	Percent	20 40 60 80	Cum Percent	ChiSquare	DF	Prob>ChiSq
1	5.3567	35.712		35.712	10904.6	100.692	<.0001*
2	1.6025	10.683		46.395	4518.30	95.518	<.0001*
3	1.3486	8.990		55.385	3517.04	84.043	<.0001*
4	1.2157	8.105		63.490	2681.82	72.982	<.0001*
5	1.0288	6.859		70.349	1836.70	62.456	<.0001*
6	0.7076	4.717		75.066	1112.68	52.488	<.0001*
7	0.6785	4.523		79.589	847.984	43.030	<.0001*
8	0.6006	4.004		83.593	543.888	34.446	<.0001*
9	0.4621	3.081		86.674	286.225	26.839	<.0001*
10	0.4343	2.895		89.569	197.022	19.970	<.0001*
11	0.4001	2.667		92.237	110.592	14.073	<.0001*
12	0.3308	2.205		94.442	33.355	9.034	0.0001*
13	0.2960	1.973		96.415	12.361	4.928	0.0288*
14	0.2854	1.902		98.317	6.770	1.810	0.0277*
15	0.2524	1.683		100.000	0.000	.	.

Figure 53: Eigenvectors of the principal components of sharing and reuse.



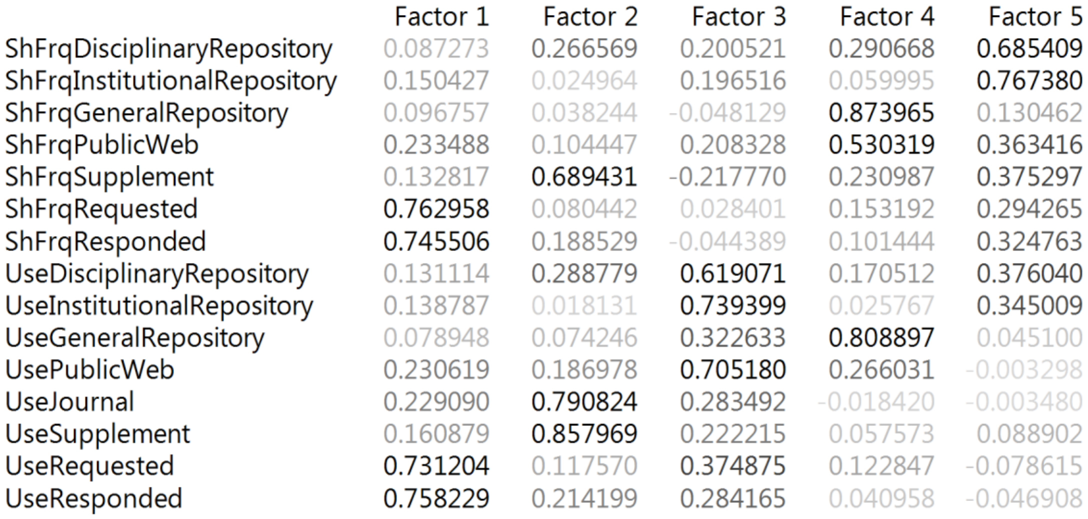
ShFrqDisciplinaryRepository	0.67316
ShFrqInstitutionalRepository	0.65434
ShFrqGeneralRepository	0.79398
ShFrqPublicWeb	0.52214
ShFrqSupplement	0.73458
ShFrqRequested	0.69944
ShFrqResponded	0.70905
UseDisciplinaryRepository	0.65431
UseInstitutionalRepository	0.68600
UseGeneralRepository	0.77219
UsePublicWeb	0.65621
UseJournal	0.75860
UseSupplement	0.82259
UseRequested	0.71028
UseResponded	0.70542

Figure 54: Final communality estimates of the FA for five factors.



Factor	Variance	Percent	Cum Percent
Factor 1	2.5335	16.890	16.890
Factor 2	2.1468	14.312	31.202
Factor 3	2.0571	13.714	44.916
Factor 4	1.9955	13.303	58.219
Factor 5	1.8194	12.130	70.349

Figure 55: Variance explained by each factor.



	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
ShFrqDisciplinaryRepository	0.087273	0.266569	0.200521	0.290668	0.685409
ShFrqInstitutionalRepository	0.150427	0.024964	0.196516	0.059995	0.767380
ShFrqGeneralRepository	0.096757	0.038244	-0.048129	0.873965	0.130462
ShFrqPublicWeb	0.233488	0.104447	0.208328	0.530319	0.363416
ShFrqSupplement	0.132817	0.689431	-0.217770	0.230987	0.375297
ShFrqRequested	0.762958	0.080442	0.028401	0.153192	0.294265
ShFrqResponded	0.745506	0.188529	-0.044389	0.101444	0.324763
UseDisciplinaryRepository	0.131114	0.288779	0.619071	0.170512	0.376040
UseInstitutionalRepository	0.138787	0.018131	0.739399	0.025767	0.345009
UseGeneralRepository	0.078948	0.074246	0.322633	0.808897	0.045100
UsePublicWeb	0.230619	0.186978	0.705180	0.266031	-0.003298
UseJournal	0.229090	0.790824	0.283492	-0.018420	-0.003480
UseSupplement	0.160879	0.857969	0.222215	0.057573	0.088902
UseRequested	0.731204	0.117570	0.374875	0.122847	-0.078615
UseResponded	0.758229	0.214199	0.284165	0.040958	-0.046908

Figure 56: Rotated factor loadings for sharing and reuse.

An additional principal component analysis (PCA) based on correlations of most variables using eight factors for the FA is displayed in the following figures.

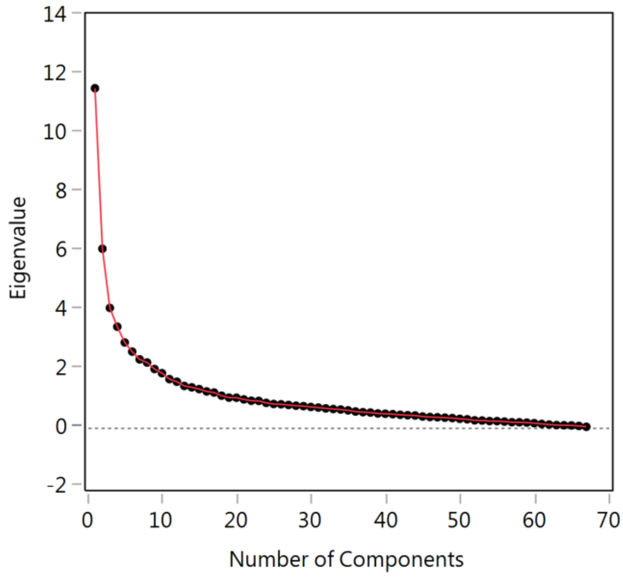


Figure 57: Scree plot of the PCA.

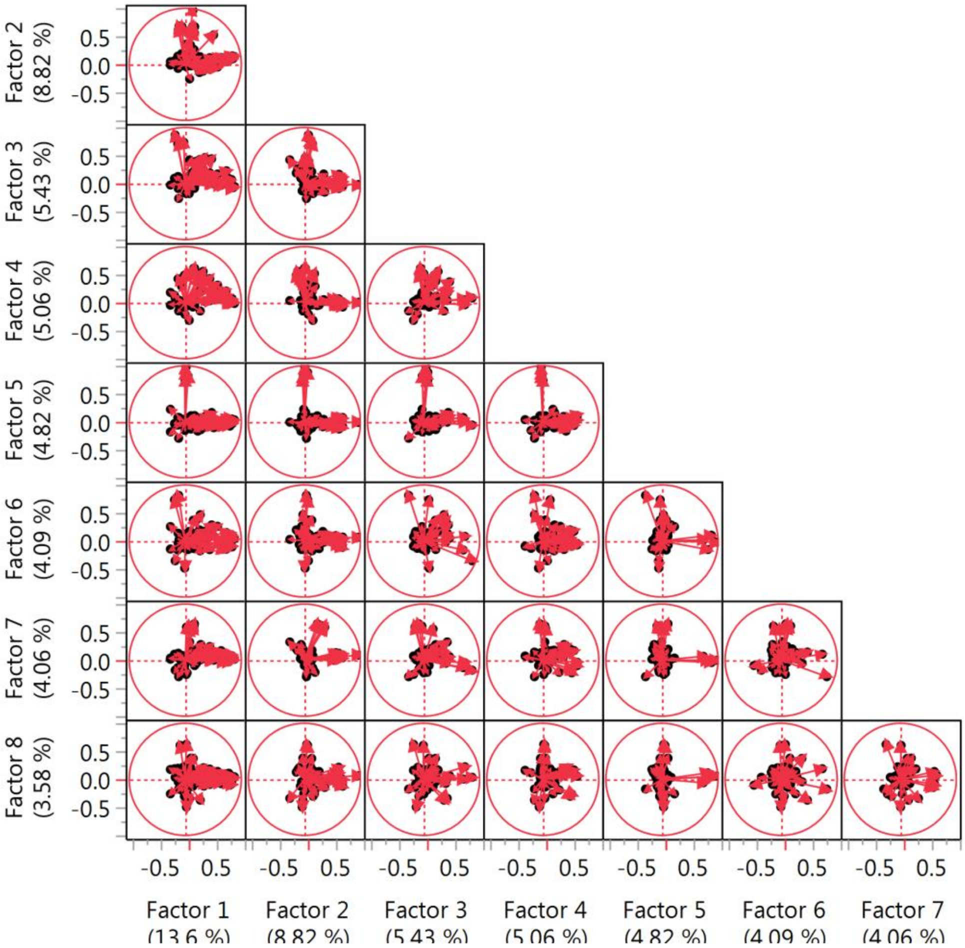
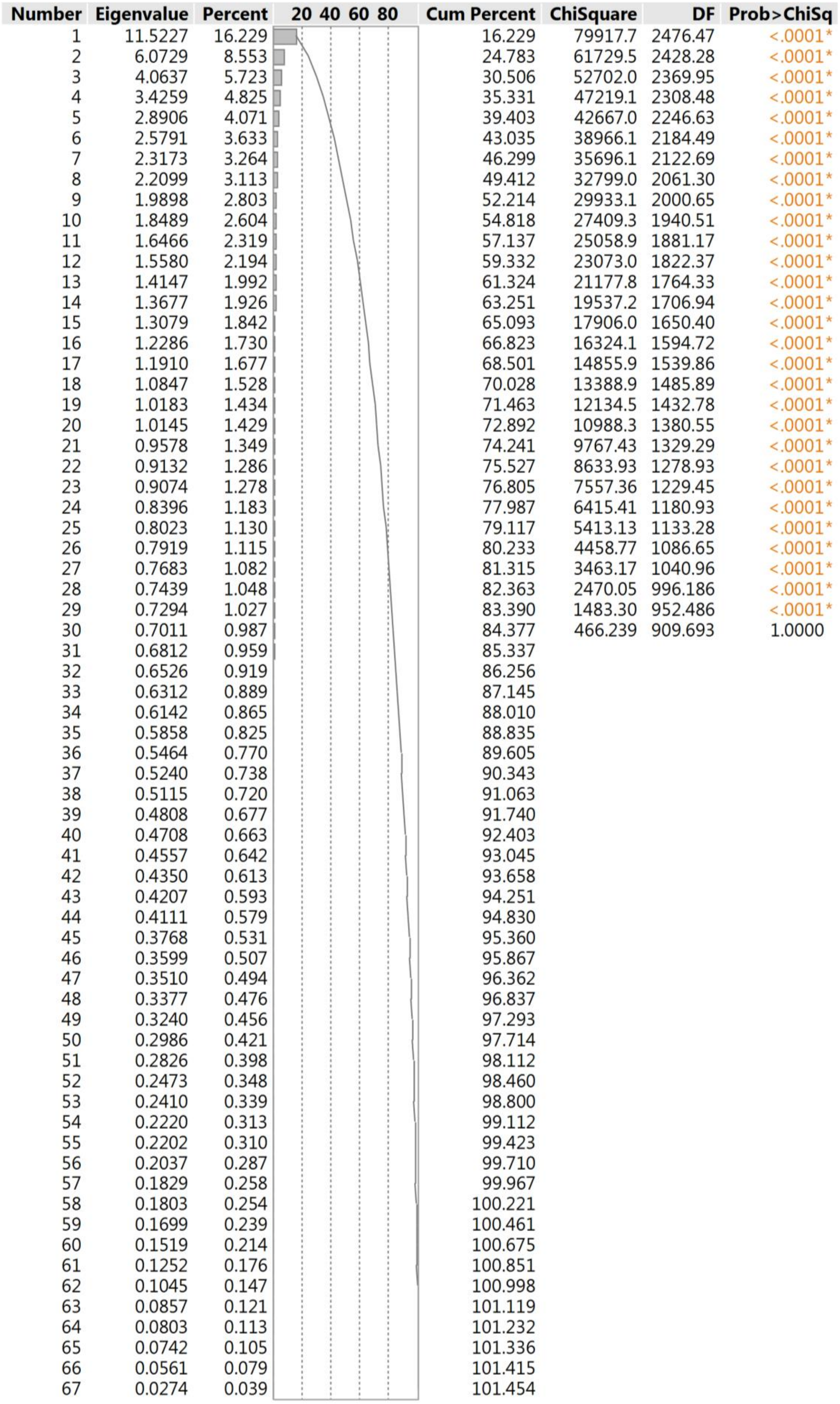
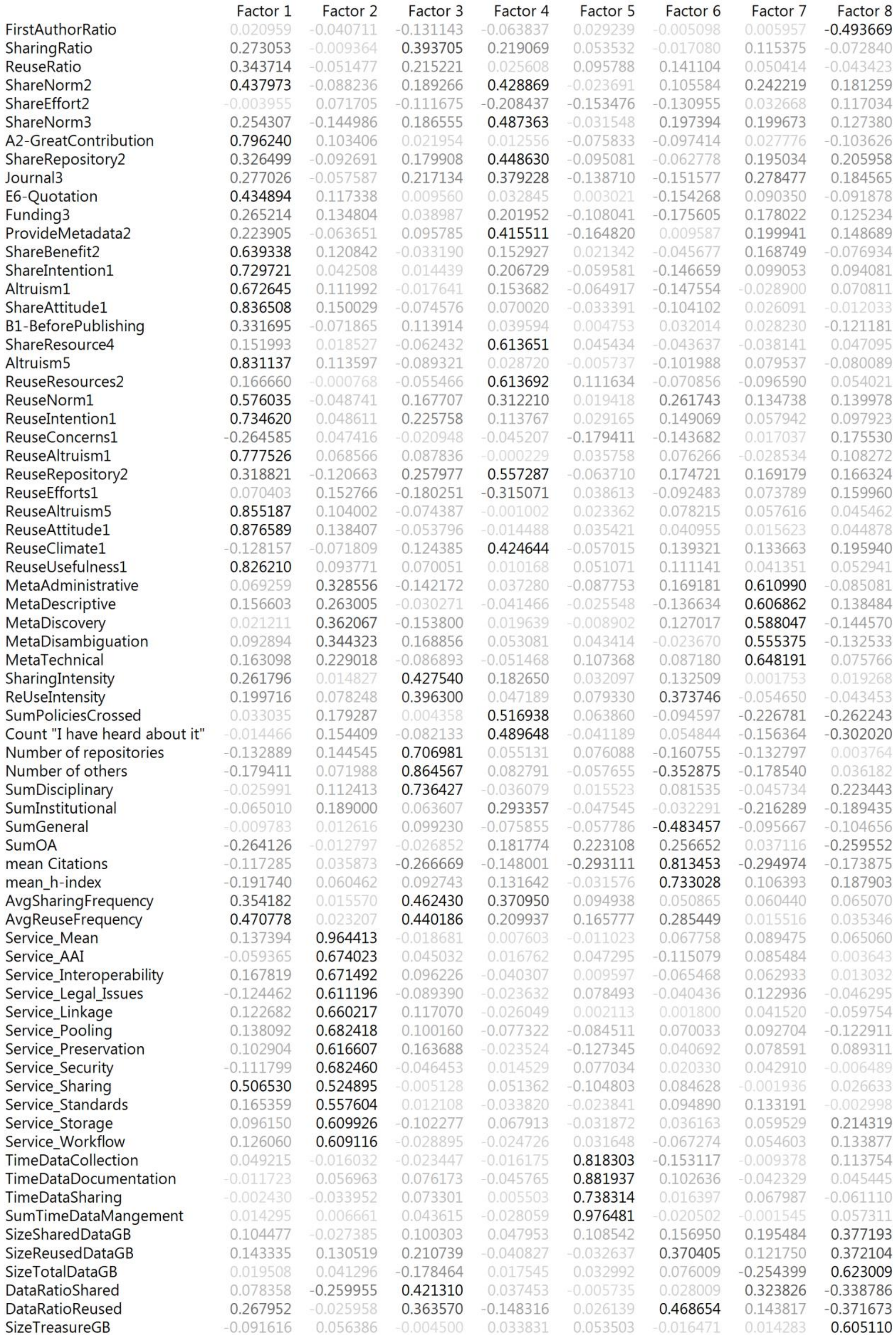


Figure 58: Factor loading plot of the PCA.



Number	Eigenvalue	Percent	20 40 60 80	Cum Percent	ChiSquare	DF	Prob>ChiSq
1	11.5227	16.229		16.229	79917.7	2476.47	<.0001*
2	6.0729	8.553		24.783	61729.5	2428.28	<.0001*
3	4.0637	5.723		30.506	52702.0	2369.95	<.0001*
4	3.4259	4.825		35.331	47219.1	2308.48	<.0001*
5	2.8906	4.071		39.403	42667.0	2246.63	<.0001*
6	2.5791	3.633		43.035	38966.1	2184.49	<.0001*
7	2.3173	3.264		46.299	35696.1	2122.69	<.0001*
8	2.2099	3.113		49.412	32799.0	2061.30	<.0001*
9	1.9898	2.803		52.214	29933.1	2000.65	<.0001*
10	1.8489	2.604		54.818	27409.3	1940.51	<.0001*
11	1.6466	2.319		57.137	25058.9	1881.17	<.0001*
12	1.5580	2.194		59.332	23073.0	1822.37	<.0001*
13	1.4147	1.992		61.324	21177.8	1764.33	<.0001*
14	1.3677	1.926		63.251	19537.2	1706.94	<.0001*
15	1.3079	1.842		65.093	17906.0	1650.40	<.0001*
16	1.2286	1.730		66.823	16324.1	1594.72	<.0001*
17	1.1910	1.677		68.501	14855.9	1539.86	<.0001*
18	1.0847	1.528		70.028	13388.9	1485.89	<.0001*
19	1.0183	1.434		71.463	12134.5	1432.78	<.0001*
20	1.0145	1.429		72.892	10988.3	1380.55	<.0001*
21	0.9578	1.349		74.241	9767.43	1329.29	<.0001*
22	0.9132	1.286		75.527	8633.93	1278.93	<.0001*
23	0.9074	1.278		76.805	7557.36	1229.45	<.0001*
24	0.8396	1.183		77.987	6415.41	1180.93	<.0001*
25	0.8023	1.130		79.117	5413.13	1133.28	<.0001*
26	0.7919	1.115		80.233	4458.77	1086.65	<.0001*
27	0.7683	1.082		81.315	3463.17	1040.96	<.0001*
28	0.7439	1.048		82.363	2470.05	996.186	<.0001*
29	0.7294	1.027		83.390	1483.30	952.486	<.0001*
30	0.7011	0.987		84.377	466.239	909.693	1.0000
31	0.6812	0.959		85.337			
32	0.6526	0.919		86.256			
33	0.6312	0.889		87.145			
34	0.6142	0.865		88.010			
35	0.5858	0.825		88.835			
36	0.5464	0.770		89.605			
37	0.5240	0.738		90.343			
38	0.5115	0.720		91.063			
39	0.4808	0.677		91.740			
40	0.4708	0.663		92.403			
41	0.4557	0.642		93.045			
42	0.4350	0.613		93.658			
43	0.4207	0.593		94.251			
44	0.4111	0.579		94.830			
45	0.3768	0.531		95.360			
46	0.3599	0.507		95.867			
47	0.3510	0.494		96.362			
48	0.3377	0.476		96.837			
49	0.3240	0.456		97.293			
50	0.2986	0.421		97.714			
51	0.2826	0.398		98.112			
52	0.2473	0.348		98.460			
53	0.2410	0.339		98.800			
54	0.2220	0.313		99.112			
55	0.2202	0.310		99.423			
56	0.2037	0.287		99.710			
57	0.1829	0.258		99.967			
58	0.1803	0.254		100.221			
59	0.1699	0.239		100.461			
60	0.1519	0.214		100.675			
61	0.1252	0.176		100.851			
62	0.1045	0.147		100.998			
63	0.0857	0.121		101.119			
64	0.0803	0.113		101.232			
65	0.0742	0.105		101.336			
66	0.0561	0.079		101.415			
67	0.0274	0.039		101.454			

Figure 59: Eigenvalues of the PCA factors.



	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
FirstAuthorRatio	0.020959	-0.040711	-0.131143	-0.063837	0.029239	-0.005098	0.005957	-0.493669
SharingRatio	0.273053	-0.009364	0.393705	0.219069	0.053532	-0.017080	0.115375	-0.072840
ReuseRatio	0.343714	-0.051477	0.215221	0.025608	0.095788	0.141104	0.050414	-0.043423
ShareNorm2	0.437973	-0.088236	0.189266	0.428869	-0.023691	0.105584	0.242219	0.181259
ShareEffort2	-0.003955	0.071705	-0.111675	-0.208437	-0.153476	-0.130955	0.032668	0.117034
ShareNorm3	0.254307	-0.144986	0.186555	0.487363	-0.031548	0.197394	0.199673	0.127380
A2-GreatContribution	0.796240	0.103406	0.021954	0.012556	-0.075833	-0.097414	0.027776	-0.103626
ShareRepository2	0.326499	-0.092691	0.179908	0.448630	-0.095081	-0.062778	0.195034	0.205958
Journal3	0.277026	-0.057587	0.217134	0.379228	-0.138710	-0.151577	0.278477	0.184565
E6-Quotation	0.434894	0.117338	0.009560	0.032845	0.003021	-0.154268	0.090350	-0.091878
Funding3	0.265214	0.134804	0.038987	0.201952	-0.108041	-0.175605	0.178022	0.125234
ProvideMetadata2	0.223905	-0.063651	0.095785	0.415511	-0.164820	0.009587	0.199941	0.148689
ShareBenefit2	0.639338	0.120842	-0.033190	0.152927	0.021342	-0.045677	0.168749	-0.076934
ShareIntention1	0.729721	0.042508	0.014439	0.206729	-0.059581	-0.146659	0.099053	0.094081
Altruism1	0.672645	0.111992	-0.017641	0.153682	-0.064917	-0.147554	-0.028900	0.070811
ShareAttitude1	0.836508	0.150029	-0.074576	0.070020	-0.033391	-0.104102	0.026091	-0.012033
B1-BeforePublishing	0.331695	-0.071865	0.113914	0.039594	0.004753	0.032014	0.028230	-0.121181
ShareResource4	0.151993	0.018527	-0.062432	0.613651	0.045434	-0.043637	-0.038141	0.047095
Altruism5	0.831137	0.113597	-0.089321	0.028720	-0.005737	-0.101988	0.079537	-0.080089
ReuseResources2	0.166660	-0.000768	-0.055466	0.613692	0.111634	-0.070856	-0.096590	0.054021
ReuseNorm1	0.576035	-0.048741	0.167707	0.312210	0.019418	0.261743	0.134738	0.139978
ReuseIntention1	0.734620	0.048611	0.225758	0.113767	0.029165	0.149069	0.057942	0.097923
ReuseConcerns1	-0.264585	0.047416	-0.020948	-0.045207	-0.179411	-0.143682	0.017037	0.175530
ReuseAltruism1	0.777526	0.068566	0.087836	-0.000229	0.035758	0.076266	-0.028534	0.108272
ReuseRepository2	0.318821	-0.120663	0.257977	0.557287	-0.063710	0.174721	0.169179	0.166324
ReuseEfforts1	0.070403	0.152766	-0.180251	-0.315071	0.038613	-0.092483	0.073789	0.159960
ReuseAltruism5	0.855187	0.104002	-0.074387	-0.001002	0.023362	0.078215	0.057616	0.045462
ReuseAttitude1	0.876589	0.138407	-0.053796	-0.014488	0.035421	0.040955	0.015623	0.044878
ReuseClimate1	-0.128157	-0.071809	0.124385	0.424644	-0.057015	0.139321	0.133663	0.195940
ReuseUsefulness1	0.826210	0.093771	0.070051	0.010168	0.051071	0.111141	0.041351	0.052941
MetaAdministrative	0.069259	0.328556	-0.142172	0.037280	-0.087753	0.169181	0.610990	-0.085081
MetaDescriptive	0.156603	0.263005	-0.030271	-0.041466	-0.025548	-0.136634	0.606862	0.138484
MetaDiscovery	0.021211	0.362067	-0.153800	0.019639	-0.008902	0.127017	0.588047	-0.144570
MetaDisambiguation	0.092894	0.344323	0.168856	0.053081	0.043414	-0.023670	0.555375	-0.132533
MetaTechnical	0.163098	0.229018	-0.086893	-0.051468	0.107368	0.087180	0.648191	0.075766
SharingIntensity	0.261796	0.014827	0.427540	0.182650	0.032097	0.132509	0.001753	0.019268
ReUseIntensity	0.199716	0.078248	0.396300	0.047189	0.079330	0.373746	-0.054650	-0.043453
SumPoliciesCrossed	0.033035	0.179287	0.004358	0.516938	0.063860	-0.094597	-0.226781	-0.262243
Count "I have heard about it"	-0.014466	0.154409	-0.082133	0.489648	-0.041189	0.054844	-0.156364	-0.302020
Number of repositories	-0.132889	0.144545	0.706981	0.055131	0.076088	-0.160755	-0.132797	0.003764
Number of others	-0.179411	0.071988	0.864567	0.082791	-0.057655	-0.352875	-0.178540	0.036182
SumDisciplinary	-0.025991	0.112413	0.736427	-0.036079	0.015523	0.081535	-0.045734	0.223443
SumInstitutional	-0.065010	0.189000	0.063607	0.293357	-0.047545	-0.032291	-0.216289	-0.189435
SumGeneral	-0.009783	0.012616	0.099230	-0.075855	-0.057786	-0.483457	-0.095667	-0.104656
SumOA	-0.264126	-0.012797	-0.026852	0.181774	0.223108	0.256652	0.037116	-0.259552
mean Citations	-0.117285	0.035873	-0.266669	-0.148001	-0.293111	0.813453	-0.294974	-0.173875
mean_h-index	-0.191740	0.060462	0.092743	0.131642	-0.031576	0.733028	0.106393	0.187903
AvgSharingFrequency	0.354182	0.015570	0.462430	0.370950	0.094938	0.050865	0.060440	0.065070
AvgReuseFrequency	0.470778	0.023207	0.440186	0.209937	0.165777	0.285449	0.015516	0.035346
Service_Mean	0.137394	0.964413	-0.018681	0.007603	-0.011023	0.067758	0.089475	0.065060
Service_AAI	-0.059365	0.674023	0.045032	0.016762	0.047295	-0.115079	0.085484	0.003643
Service_Interoperability	0.167819	0.671492	0.096226	-0.040307	0.009597	-0.065468	0.062933	0.013032
Service_Legal_Issues	-0.124462	0.611196	-0.089390	-0.023632	0.078493	-0.040436	0.122936	-0.046295
Service_Linkage	0.122682	0.660217	0.117070	-0.026049	0.002113	0.001800	0.041520	-0.059754
Service_Pooling	0.138092	0.682418	0.100160	-0.077322	-0.084511	0.070033	0.092704	-0.122911
Service_Preservation	0.102904	0.616607	0.163688	-0.023524	-0.127345	0.040692	0.078591	0.089311
Service_Security	-0.111799	0.682460	-0.046453	0.014529	0.077034	0.020330	0.042910	-0.006489
Service_Sharing	0.506530	0.524895	-0.005128	0.051362	-0.104803	0.084628	-0.001936	0.026633
Service_Standards	0.165359	0.557604	0.012108	-0.033820	-0.023841	0.094890	0.133191	-0.002998
Service_Storage	0.096150	0.609926	-0.102277	0.067913	-0.031872	0.036163	0.059529	0.214319
Service_Workflow	0.126060	0.609116	-0.028895	-0.024726	0.031648	-0.067274	0.054603	0.133877
TimeDataCollection	0.049215	-0.016032	-0.023447	-0.016175	0.818303	-0.153117	-0.009378	0.113754
TimeDataDocumentation	-0.011723	0.056963	0.076173	-0.045765	0.881937	0.102636	-0.042329	0.045445
TimeDataSharing	-0.002430	-0.033952	0.073301	0.005503	0.738314	0.016397	0.067987	-0.061110
SumTimeDataMangement	0.014295	0.006661	0.043615	-0.028059	0.976481	-0.020502	-0.001545	0.057311
SizeSharedDataGB	0.104477	-0.027385	0.100303	0.047953	0.108542	0.156950	0.195484	0.377193
SizeReusedDataGB	0.143335	0.130519	0.210739	-0.040827	-0.032637	0.370405	0.121750	0.372104
SizeTotalDataGB	0.019508	0.041296	-0.178464	0.017545	0.032992	0.076009	-0.254399	0.623009
DataRatioShared	0.078358	-0.259955	0.421310	0.037453	-0.005735	0.028009	0.323826	-0.338786
DataRatioReused	0.267952	-0.025958	0.363570	-0.148316	0.026139	0.468654	0.143817	-0.371673
SizeTreasureGB	-0.091616	0.056386	-0.004500	0.033831	0.053503	-0.016471	0.014283	0.605110

Figure 60: Rotated factor loadings of the PCA.