



**Data reuse in the Social Sciences and Humanities:  
Project report of the SWITCH Innovation Lab  
“Repositories & Data Quality”**

Nicolai Hauf, Andreas Fürholz, Vanessa Christina Klaas, Jennifer Morger, Elena Šimukovič, Martin Jaekel

ZHAW Zurich University of Applied Sciences  
with support from SWITCH

March 2021

DOI: [10.21256/zhaw-2404](https://doi.org/10.21256/zhaw-2404)



This is an open access publication under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

**Principal:**

SWITCH  
Werdstrasse 2  
CH-8004 Zürich

**Lab Partner:**

ZHAW Zurich University of Applied Sciences  
Research and Development Unit  
Gertrudstrasse 15  
CH-8401 Winterthur



Project lead for ZHAW: Nicolai Hauf – University Library

Project team and contributors:

From ZHAW

- Nicolai Hauf  
University Library  ORCID: [0000-0003-0665-7786](https://orcid.org/0000-0003-0665-7786)  
Contact: [hauo@zhaw.ch](mailto:hauo@zhaw.ch)
- Andreas Fürholz  
Research and Development Unit  ORCID: [0000-0002-2224-7544](https://orcid.org/0000-0002-2224-7544)  
Contact: [fueh@zhaw.ch](mailto:fueh@zhaw.ch)
- Vanessa Christina Klaas  
Research and Development Unit  ORCID: [0000-0001-5484-4354](https://orcid.org/0000-0001-5484-4354)  
Contact: [klav@zhaw.ch](mailto:klav@zhaw.ch)
- Jennifer Morger  
University Library  ORCID: [0000-0002-5730-1408](https://orcid.org/0000-0002-5730-1408)  
Contact: [morj@zhaw.ch](mailto:morj@zhaw.ch)
- Elena Šimukovič  
University Library  ORCID: [0000-0003-1363-243X](https://orcid.org/0000-0003-1363-243X)  
Contact: [simv@zhaw.ch](mailto:simv@zhaw.ch)
- Martin Jaekel  
Research and Development Unit  ORCID: [0000-0001-8445-949X](https://orcid.org/0000-0001-8445-949X)  
Contact: [jaek@zhaw.ch](mailto:jaek@zhaw.ch)

From SWITCH

- Sebastian Sigloch  ORCID: [0000-0002-3047-6048](https://orcid.org/0000-0002-3047-6048)  
Contact: [sebastian.sigloch@switch.ch](mailto:sebastian.sigloch@switch.ch)
- Andrea Bertino  ORCID: [0000-0002-5080-036X](https://orcid.org/0000-0002-5080-036X)  
Contact: [andrea.bertino@switch.ch](mailto:andrea.bertino@switch.ch)

© 2021 ZHAW Zurich University of Applied Sciences with support from SWITCH

The underlying data and additional materials in connection with this publication are available at:  
<https://doi.org/10.5281/zenodo.4609834>

Recommended citation:

Hauf, Nicolai; Fürholz, Andreas; Klaas, Vanessa Christina; Morger, Jennifer; Šimukovič, Elena; Jaekel, Martin: Data reuse in the social sciences and humanities : project report of the SWITCH Innovation Lab “Repositories & Data Quality”. Winterthur: ZHAW Zurich University of Applied Sciences, 2021. Available at: <https://doi.org/10.21256/zhaw-2404>

## CONTENTS

Executive summary .....	01
1. Introduction and background .....	02
2. Approach .....	02
2.1 Scoping .....	02
2.2 Survey design .....	03
2.3 Desk research on relevant data sources .....	04
2.4 Analysis and presentation of results .....	04
2.5 Assessment and limitations .....	05
3. Scoping and review of existing studies .....	05
4. Survey development and description .....	07
5. Results and analysis .....	09
6. Desk research into relevant data sources .....	21
7. Additional not covered topics .....	23
8. Main findings and recommended next steps .....	24
Acknowledgements .....	25
Bibliography .....	26
List of tables and figures .....	27
Appendices .....	28

## EXECUTIVE SUMMARY

This report is the result of the SWITCH Innovation Lab “Repositories & Data Quality”, a project that ran from October 2020 until February 2021 as a collaboration between SWITCH and ZHAW Zurich University of Applied Sciences. Its aim was to complement previous studies on research data management issues (conducted in part as earlier SWITCH Innovation Labs) and to identify relevant data sources for researchers in the social sciences and humanities (SSH) in Switzerland. More particularly, the project focused on the reuse of existing data sets by SSH researchers and the criteria they applied when choosing a suitable data source for their work and research. Some of the steps in this task consisted of finding the locations where valuable data is shared, published and accessed as well as conducting a more specific investigation into data availability, modes of accessibility and aspects related to assessing data quality.

For this purpose, the project team designed and carried out an online survey targeted specifically at active SSH researchers in Switzerland. To disseminate the survey questionnaire towards this target audience, mailing lists of several research organizations in these fields were utilized. The survey ran for about 8 weeks until early February 2021 and received responses from 260 participants. Some of the main findings include a generally high number of researchers making use of existing data for their own work. Central data providers such as FORSbase, FSO and the GESIS data archive are the most frequently named sources. Trust in these data sources and sufficient additional materials like documentation and methodologies are key criteria for selecting data for reuse.

Some data sources could provide desired data sets but are hardly accessible and reusable for researchers (if at all). This mainly includes administrative data and records of (Swiss) public authorities and offices, as well as historic assets from archives, libraries and museums. Furthermore, qualitative research data like interviews, surveys, questionnaires and observations were often highlighted among valuable yet usually inaccessible data sources. At the same time, the case of qualitative data such as interview recordings and ethnographic fieldnotes illustrate well a certain reusability dilemma. To enable “reusability” of such materials from a legal perspective (i.e. to protect personal identities of research participants), much of sensitive context-related details need to be removed. Yet, it is exactly those details that are necessary to contextualize and reuse these data in a proper way from a qualitative researcher’s point of view.

Finally, the list of relevant data sources in the SSH contains a large number of individual studies, data sets and resources. This fact supports the idea to connect and link this data, as repeatedly voiced by survey respondents. This demand presents a particular opportunity for future efforts in this area that directly align with the broad objectives at SWITCH. More specifically, the vision of the SWITCH Research Data Connectome is to interconnect research data from different sources, which contributes to the current development of a knowledge graph. Building on this knowledge base that documents and links metadata to enable a more effective search for and reuse of data, new specialized services could be employed in the future. The results of the SWITCH Innovation Lab “Repositories & Data Quality” project shall help lay the groundwork for such future client-oriented services, by providing more detailed information about the handling and reuse of data in Switzerland.

## 1. INTRODUCTION AND BACKGROUND

The Innovation Lab “Repositories & Data Quality” is a collaboration between SWITCH and ZHAW Zurich University of Applied Sciences. Its purpose is to lay the groundwork for the implementation of thematic innovation activities in the field of Open Science and research data management, specifically the SWITCH Research Data Connectome. The project started in October 2020 and ran until February 2021. The tasks were carried out by a team of research data management professionals at ZHAW consisting of members from ZHAW President’s Office, Research and Development Unit and the university library in close cooperation with SWITCH.

The main purpose of this lab was to identify the relevant data for researchers in Switzerland in the disciplines of social sciences and humanities (SSH). The first step consisted of determining the **locations** where valuable data is shared, published and accessed. This enables a specific investigation about data **availability**, modes of **accessibility** and **data quality**. The objective was to complement existing studies and provide new insights into the dealing with research data. Our focus therefore concerned the reuse of existing data and the criteria researchers use to select data. A reasonable assessment of relevance can only be achieved by considering the researcher’s own views on the data they utilize. Hence, our questions were directed directly towards SSH researchers in Switzerland. Data is regarded as “relevant” when it is valuable for researchers and creates scientific and societal impact.

This project is part of a series of labs addressing promising topics related to the overall strategic orientation of SWITCH. The aim of this study was to acquire detailed information about the handling of data within certain research fields to lay the groundwork for future, client-oriented services, especially the **SWITCH Research Data Connectome**. The Connectome is an ecosystem initiative coordinated by SWITCH, to make data from various disciplines, stored in different locations, easily findable, accessible, interoperable and reusable (FAIR). It supports the idea of Open Science to make research data widely available, connected and sustainable by providing a single point of access to find and reuse linked data from different repositories and other sources.<sup>1</sup> To this end, the Connectome will harvest data repositories and other relevant sources for metadata describing data sets published there. This metadata will then be harmonized, enriched, and mapped to a standardized schema. The sum of all this information leads to the development of a knowledge graph, a knowledge base that documents and links metadata to enable a more effective search for and discovery of data. In the future, this graph should also enable the exploitation of linked open data by different kinds of service.

## 2. APPROACH

Structurally this lab consists of five objectives with specific deliverables. During the **scoping phase**, the project team defined the focus of this project more closely by examining prior studies targeting data handling among researchers in the SSH. In a second step we paved the way to reach the researchers for our investigation. We **identified and acquired important multipliers** in the Data and Service Center for the Humanities (DaSCH), the Swiss Center of Expertise in the Social Sciences (FORS) and the Schweizerische Akademie der Geistes- und Sozialwissenschaften (SAGW) as central players in the field of SSH. To answer the research questions, we **designed a survey** in an active exchange and feedback loop with SWITCH and researchers of the considered fields. The **survey was disseminated** through the multipliers and results were collected by using REDCap. We normalized and processed incoming responses to prepare **data analysis**. The project is summarized in this **final report** and key results will be presented to the Connectome partners and their stakeholders.

Below we give a comprehensive overview of our approach and methods employed during the different phases of this project.

### 2.1 SCOPING

To complement previous analyses regarding data management, the scope of our investigation was closer defined by analyzing existing literature. This included the most relevant and recent studies regarding research data management in Switzerland and in the considered disciplines of SSH. We considered the key findings of publications recognized as important for this lab. This list included:

- SNSF Workshop on Open Data in Science (2015)<sup>2</sup>
- SNSF Open Research Data: Landscape and cost analysis (2019)<sup>3</sup>
- SNSF Open Research Data monitoring report 2017-2018 (2020)<sup>4</sup>

<sup>1</sup> SWITCH Open Science. Available at: <https://www.switch.ch/about/open-science/>

<sup>2</sup> SNSF (2015)

<sup>3</sup> SNSF (2019)

<sup>4</sup> Milzow et al. (2020)

- SCNAT MAP Open Data Survey (2020)<sup>5</sup>
- BAKOM Digital Switzerland Strategy (2018)<sup>6</sup>
- NICT White Paper: Research Data Management Landscape in Switzerland (2019)<sup>7</sup>
- Swissuniversities ORD survey (2020)<sup>8</sup>

During this project the list of relevant publications was extended to include:

- OPERAS Survey on SSH Scholarly Communication<sup>9</sup> (ongoing)
- DARIAH-EU European survey on scholarly practices and digital needs in the arts and humanities (2016)<sup>10</sup>
- Linkhub.ch / FORS Accessing and linking data for research in Switzerland (2020)<sup>11</sup>

The main findings of our literature review and important lessons learned for our study are summarized in chapter 3.

## 2.2 SURVEY DESIGN

The scoping revealed an information gap concerning the reuse of data by researchers. Therefore, the main objective of this survey was to identify relevant data sources in the SSH. This encompasses an overview of data sources that are currently utilized and an outlook on those that might be of interest in the future. The goal was a **systematic collection of data sources** used in the different research fields. The question regarding the sources was accompanied by specifications about the type of data obtained there and the purpose for the data reuse. It was considered essential to analyze all responses primarily in the context of the specific research field the participants belong to. Of special interest were sources other than the major repositories and platforms (such as FORSbase, Zenodo), including non-academic platforms. Following the identification, a strong focus lied on the **criteria** researchers in different fields have for selecting individual sources. The survey aimed to answer the question, what researchers need to choose a relevant source. This includes issues like availability and accessibility as well as data quality. Requirements will likely vary from research field to research field and between different kinds of data.

The survey was kept short (11 questions in 5-15 minutes) to encourage more responses and to stay on target. The questionnaire was developed in an iterative process involving SWITCH, DaSCH, FORS and SSH researchers. It was tested with the help of researchers in the relevant disciplines and further personal contacts to verify general comprehensibility and operability.

We used **REDCap**<sup>12</sup> (Version 10.6.0) for data collection and storage. REDCap is a web-based application for managing online databases and surveys. As a standard tool in clinical research, it focuses on data security and integrity and provides a rich set of features such as secure data access through its API and building surveys in multiple languages. ZHAW is running its own instance wherefore all collected data is stored on ZHAW infrastructure. REDCap provides a public URL to the published survey for participants to answer.

**Researchers in Switzerland within the SSH were our target audience.** Of primary interest were active researchers that had previous experience in using existing data. We reached potential respondents by using the networks of leading Swiss data and service providers in these fields, DaSCH, FORS and SAGW. The announcement of the survey was spread through mailing lists of DaSCH and FORS to contact associated researchers personally. The reference and information to this survey was part of the SAGW newsletter<sup>13</sup> in December 2020. Furthermore, some additional channels like personal contacts, relevant subject specific mailing lists<sup>14</sup> and the social media presence of the Swiss portal for the historical sciences (infoclio.ch)<sup>15</sup> were utilized as well. Although there was no intention to lean towards certain research fields, the channels through which the survey was communicated might have influenced the representative nature of our results.

The survey was launched on the 14<sup>th</sup> of December 2020 and was intended to be active until the 24<sup>th</sup> of January 2021. In the end, this period was extended until the 5<sup>th</sup> of February 2021 to enable further circulation. DaSCH started to announce the survey to their contacts in December (with a reminder in the

<sup>5</sup> SCNAT (2020)

<sup>6</sup> BAKOM (2018)

<sup>7</sup> Brüwer (2019)

<sup>8</sup> Swissuniversities (2020)

<sup>9</sup> OPERAS (2020)

<sup>10</sup> Dallas et al. (2016)

<sup>11</sup> Linkhub.ch, FORS (2020)

<sup>12</sup> <https://www.project-redcap.org/>

<sup>13</sup> SAGW Newsletter. Available at: <https://sagw.ch/sagw/aktuell/publikationen/newsletter/>

<sup>14</sup> [Humanistica](#), [DHd](#), [AICUD](#) for the digital humanities

<sup>15</sup> <https://www.infoclio.ch/en>

middle of January), the SAGW newsletter was published and sent out in late December too. Additionally, the survey link was published in the referenced three mailing lists for the digital humanities at the start of January. FORS disseminated the survey details at the end of January. The general timing of the survey (with Christmas and the start of the new year) was not ideal to trigger responses but was unavoidable due to the project schedule.

In total, **21.483 researchers** were reached by these means. This number does not consider possible other ways to notice the survey like social media postings<sup>16</sup>, website news or personal contacts and interpersonal exchange among researchers. It is possible and even likely that researchers received our invitation several times through different channels. We obtained **263 responses** to our questionnaire. Three responses were incomplete, thus rejected from further analysis. The response rate was 1,21%.

Participants were able to select one of three languages in which the survey was available (German, English, French).

## 2.3 DESK RESEARCH ON RELEVANT DATA SOURCES

We finish this study with a desk research on identified relevant data sources focusing on aspects not covered in our survey. This includes the **modes of access** to the provided data sets, the **description and presentation of data sets** and further measures to secure **data quality**. Our findings are based solely on information being available on their websites (self-declaration). Due to time constraints we could not incorporate other meta studies or publications regarding these sources, as well as meetings to discuss the issues personally.

## 2.4 ANALYSIS AND PRESENTATION OF RESULTS

### Analysis

#### *Data access and used technology*

For all data processing (cleaning, analysis, plotting) we used the statistical language R (version 4.0.2) with API access to the survey data stored in REDCap through the package redcapAPI (version 2.3).

#### *Data cleaning and preparation*

Since the question for the data source was an open text field, we had to normalize the names that were given as responses, e.g. both answers "FORSbase" and "FORS" occurred in the dataset and were normalized to "FORS" to have unique names for identical sources. In addition, answers were grouped according to data providers or collective terms where applicable. All sources were assigned general categories. Since survey participants specified several sources in one response, we treated this question as a multiple response variable. For this variable we provided one type of categorization that was agreed upon with SWITCH. We also added a categorization to the variable "research field" based on the SNSF P<sup>3</sup> database<sup>17</sup>.

The survey structure is nested, giving the opportunity to specify up to five sources, each with their corresponding kind of data and reuse purpose. Depending on the question combination, we considered them either as single response or multiple response variables.

Question		
Research field		
Data Reuse		
Experience		
...		
Kind 1	Source 1	Purpose 1
Kind 2	Source 2	Purpose 2
...		
...		

Table 1: Nested survey structure

For example, purpose is considered a multiple response variable when tested in combination with research field.

<sup>16</sup> For instance the posting on the [Twitter](#) and [Facebook](#) account of infoclio.ch in the middle of January.

<sup>17</sup> SNSF P<sup>3</sup> database. Available at: <http://p3.snf.ch/>

### *Statistical analysis*

For statistical analysis of the survey questions, we performed Pearson's Chi-squared tests for single response variables and used the *MRCV* package (version 0.3-3)<sup>18</sup> to test for independence of multiple response variables. We used the Bonferroni correction method to test for multiple marginal independence (MMI) or simultaneous pairwise marginal independence (SPMI) between variables, leading to Bonferroni-adjusted p-values for each 2x2 contingency table resulting from all possible response combinations of the two questions analyzed. The R-function from the package then provides an overall adjusted p-value indicating if MMI or SPMI exists and allows a more detailed insight into specific associations between responses through the individual 2x2 p-values. We reported the overall significances of the tests. A result is considered statistically significant if  $p < 0.05$ . This method provides a more conservative result, especially in the case of questions with many different available responses. We preferred this method because it considers that an individual can contribute to a contingency table with multiple responses.

### **Presentation of results**

We used the package *ggplot2* (version 3.3.3) to create plots and visualizations in combination with the *viridis* package (version 0.5.1) to use a color scheme that is perceptually uniform and perceivable by viewers with common forms of color blindness. We describe and visualize our data set through bar plots to show the distribution of responses and we use balloon plots to visualize the contingency tables of question pairs. These balloon plots do not distinguish between single response variables and multiple response variables.

## **2.5 ASSESSMENT AND LIMITATIONS**

At multiple stages of this evaluation certain limitations must be considered when looking at the results. In our sampling of respondents our focus was on researchers in Switzerland associating with the SSH. Participants were contacted through DaSCH, FORS and SAGW, so mainly researchers subscribed and connected with these institutions were reached. This could exclude specific research fields in the SSH that are not highly represented in this network, like architecture, archaeology, philosophy and law. Generally, our results are likely not informative and indicative for these research fields with a low number of responses. A significant number of survey recipients were reached by FORS, possibly leading to a higher representation of the social sciences. Furthermore, the distribution through language-specific mailing lists in the humanities reached researchers in other countries, and this may have resulted in the inclusion of individual records that were not intended to be a part of the survey. We have no information about recipients of this survey that declined our invitation participate.

### **Limitations concerning data processing and analytics**

Considering the variety of responses and variance in abstraction levels of responses to free-text questions such as the question for the data source, normalization and categorization as described above was a manual processing step and therefore not completely stringent. Analytics focused on descriptive statistics and correlation analysis. Since the combination of purpose vs. kind does not correspond to true multiple response variables, we considered each response as a single data record, hence the variables purpose and kind were treated as single response variables. In that case, we ignored that a person could have provided several answers.

## **3. SCOPING AND REVIEW OF EXISTING STUDIES**

We started by analyzing existing studies concerning research data management in Switzerland and the SSH. Therefore, our work is complementary to these studies carried out by other stakeholders including swissuniversities, SNSF and SCNAT. A brief overview of important studies and their impact on our general focus and survey design follows.

A prior **SWITCH Innovation Lab with the Swiss Academy of Engineering Sciences (SATW)** in 2020<sup>19</sup> examined data quality in the context of research. Data quality is a very important factor for the reuse of data. Different data quality layers include for example the data set itself, its description through metadata or the information infrastructure it can be accessed by. The study proposes that common guidelines and standards will enhance data quality. A connection of data within and between disciplines will further increase value and initiate new discoveries.<sup>20</sup> These findings lead us to investigate the criteria for why researchers in the SSH select particular data sets and data providers.

---

<sup>18</sup> Koziol et al. (2014)

<sup>19</sup> Koller-Meier et al. (2020)

<sup>20</sup> Koller-Meier et al. (2020)



**The platform of mathematics, astrology and physics (MAP) of SCNAT** conducted a survey among Swiss researchers focusing on their view on the transition towards open research data in 2020. The public questionnaire gave us valuable information on what kind of general questions to adopt in our own survey (academic position, being changed to research experience, kind of institution). Another important detail noted was the importance of concise terminology and the use of simple, easy to understand questions and answers.

In the same year, **swissuniversities** conducted a recent ORD survey among Swiss higher education institutions (HEI) and research institutions. The inquired topics included the availability of policies, monitoring activities, institutional e-infrastructure (like repositories) and services. This examination provided a valuable overview of the landscape and helped sharpening our own survey. In the end, the resulting list of institutional data sources could be used in connection with our own list of data sources in the SSH to check for overlap.

**NICT** published a white paper in 2019 summarizing their study among members of ICT service organizations. Its main focus was to give a perspective on research data management from a technical, infrastructural and organizational point of view. It contained an overview of currently available support services for researchers in Switzerland, as well as an analysis of existing gaps. One of which is a partly lack of visibility, governance and sustainability in existing services, hampering accessibility for all researchers. A comprehensive, domain specific catalog of existing services (that includes data providers and other data sources) was recommended as a valuable future effort. Consequently, the deliverable of such a list of relevant data sources was firmly established as a priority and need.

**OPERAS** launched a detailed survey in 2020 targeting European researchers within our scoped disciplines of the SSH. The target was to gain an overview of current practices, habits and issues of scholarly communication. Summarizing, the survey addresses the dimensions of a) publishing and communicating, b) reading, writing and collaborating and c) search, access and discovery. The focus lied on publications. The upcoming results of the OPERAS survey regarding the search, access and discovery of data can possibly be compared to some of our own findings about data sources. To stay complementary, questions already covered by the OPERAS survey were omitted in our study.

In 2016, **DARIAH-EU** investigated a similar topic in (digital) scholarly practices as part of their European study directed at researchers in the humanities. One specific profile was created for Switzerland. Aspects of scholarly communication were covered more generally, focusing on the sharing and publication of data, means to discover data and how to access it. Improved findability and access to existing digital resources was one of the key results as an important need for researchers.

After the development of the survey, a report of **linkhub.ch and FORS** was published about the accessing and linking of research data in Switzerland. It focuses on current practices and the legal basis of these aspects for administrative and sensitive data. The report provides a basis for the development of a research-friendly institutional and regulatory framework. The publication lists challenges for linking and using administrative and private data, namely that this data is often not compliant with the FAIR principles. In addition, structured metadata and documentation are often not available. A lack of standards and best practices in processing, linking and sharing is noted. Furthermore, the unwillingness of companies to share data as well as restrictions concerning the reuse (data protection) impedes research. The report emphasizes the importance of metadata as a cornerstone for accessing and linking data. The existence of data and its context, structure and quality is often unclear. Therefore, common standards are necessary to make metadata useful and interoperable. Our survey contains metadata (and data-accompanying materials) as one selection criteria for data sets and data sources. Analysis will show, how researchers value this information over other quality criteria for research data reuse. In exploring the most commonly cited data sources in the SSH, we will also take a look at their practice to describe data thoroughly (see chapter 6).

In summary, a big emphasis concerning research data management has been put into finding the information infrastructure researchers use to archive and publish their research data and the reasons for their choice. Scholarly communication and discovery of data is currently examined as well. Yet, an aspect that has been largely neglected thus far is the sources scientists consult for obtaining existing data for their own new work. The identification of relevant data providers in specific research fields is the basis for connecting them and making data centrally accessible to provide added value for researchers. Additionally, the criteria for selecting sources is an important aspect to gain insights into why certain sources are used and what is needed for building a trusted central data reference collection.

#### 4. SURVEY DEVELOPMENT AND DESCRIPTION

In the following we give a comprehensive overview of the questionnaire and its development. The questionnaire presented to the participants is included in Appendix I.

The first part (A) consists of three general questions describing our participants. The survey was anonymous and socio-demographic factors were not of interest, hence no questions in that direction were included.

<b>A.1 What type of institution do you belong to?</b>	<b>Options</b> (single-choice, mandatory)	
	- University - University of Applied Sciences - University of Teacher Education	- ETH-Domain - Other institution - No institutional affiliation
<b>Description:</b> The list of institution types was based on the general higher education landscape in Switzerland, supplemented by an option for other institutions (which aimed to include private institutions and businesses) and for no institutional affiliation.		

<b>A.2 What is your field of research?</b>	<b>Options</b> (multiple-choice, mandatory)		
	- Archaeology - Architecture - Art studies, musicology, theatre and film studies - Economics - Educational studies - Ethnology - Geography	- Health - History - Law - Linguistics - Literature - Media and Communication studies - Philosophy	- Political sciences - Psychology - Sociology, social work - Theology and religious studies - Other, please specify
<b>Description:</b> We based the options on the research fields named in the SNF P <sup>3</sup> database <sup>21</sup> attributed to the section for humanities and social sciences.			

<b>A.3 How many years of experience do you have in research?</b>	<b>Options</b> (single-choice, mandatory)	
	- 0 – 5 years - 6 – 15 years	- 16 – 30 years - 31+ years
<b>Description:</b> Information on the amount of research experience by our participants was interesting to us to possibly discover correlations between expertise and data reuse. We decided on time periods rather than academic positions to avoid uncertainty about terminology (e.g. associate, junior professor, senior scientist). In retrospect, it would have been more accurate to choose time periods of equal extend, despite this leading to more options to select from.		

The second part of the survey (B) addresses the (past) reuse of data. The section starts with the definitions of the relevant terms “data” and “reuse”.

We understand **data** ... as all digital materials, sources and results scholars collect, generate, evaluate and use. This can include text, digitized works, audio, video, surveys, interviews, etc.

By **reuse of data** ... we mean any use of already existing data for new endeavors. This can include including certain parts in your own research, teaching or replicating or verifying previous results.

**This was essential since the term “data” or “research data” is not commonly used und uniformly understood in the SSH.** In many discussions and strategy documents concerning Open Research Data (ORD) or Research Data Management (RDM) data is mainly described from a STEM point of view, meaning quantitative and experimental data sets. Researchers from our considered disciplines thus may not identify with this concept and were reminded of the wider definition of the term and what it comprises of in their context. Since the term “research data” might resonate more strongly with prevailing impressions, we decided to use the more general term “data”. Additionally, not all existing data is a result of the research process. Another notable reason for giving a clear definition is the broad spectrum of data types being managed in the SSH, ranging from physical objects and their digital 3D scans, historic books and digitized images to annotated texts, geodata, statistics up to complete text corpora and large collections of audio and video materials. The main specification is our focus on digitally available data. The use of consistent und uniformly understood vocabulary is essential in dealing with heterogenic

<sup>21</sup> SNSF P<sup>3</sup> database. Available at: <http://p3.snf.ch/>

disciplines and data types. **Our definition was predominantly based on DARIAH-DE’s explanation of the term “research data”<sup>22</sup>.** ALLEA was consulted complementarily<sup>23</sup>.

Like the term “data”, “**reuse**” does need some further explanation. It is presumed to be a seldomly used phrase describing the secondary use of existing data for new purposes. Examples again try to give a more precise picture of this concept. It was agreed upon not to use the term “secondary use” because of additional complexity and to avoid misunderstandings. Secondary data, secondary sources or secondary analysis might refer to checking and reusing existing research. However, data might be available for further use without being utilized in prior analysis (e.g. digitized materials). We tried to avoid possible conflicting horizons of meaning. The more general term “use” was discarded for not being too concise for our inquiry.

<b>B.1 Did you reuse data in the past?</b>	<b>Options</b> (single-choice, mandatory)	
	- Yes	- No
<b>Description:</b> When participants did choose “no”, the next questions in B were skipped and the questionnaire continued with part C. Further differentiation into possible response options (such as “Occasionally”, “Often”) was discarded because these terms are up to individual interpretation and vary greatly.		

<b>B.2 Please describe kind, source and main purpose of the data reused.</b>	<b>Options</b> (single choice and free text field, first row was mandatory)		
	<b>Kind</b> - Scholarly publications (books, papers, ...) - Digital artefacts (text editions, pictures, audio, video, ...) - Numerical data - Surveys and interviews - Other	<b>Source</b> Free text field	<b>Purpose</b> - Inspiration for new research questions (e.g. for proposals) - Integration of data into your own research - Teaching - Verification and cross-checking of your own data/results - Reproduction and replication studies of others - Other
<b>Description:</b> This question is presented as a table with three separate questions as columns. Five rows are provided for participants to describe individual data sets with information about the kind of data (column 1), the source of the data (column 2), and the purpose for which it was needed (column 3). Our aim was to describe instances of data reuse as precisely as possible. Main options for kind and purpose were presented as a predefined list. We used a free text field for data sources to obtain specific information. The aim was to receive the names of certain digital platforms or data providers that make data sets publicly available. In hindsight however the term “source” was often also interpreted as the specific data material reused (e.g. certain studies, certain publications or authors). We received mentions of concrete data providers (FORS, GESIS) or their data repository. Simultaneously, respondents named specific data sources and more generic source types (“Psychinfo” as one specific publication database, but also “Literaturdatenbanken” in general). In some cases, a clear inference could be made from one statement to a specific data provider (e.g. the Selects <sup>24</sup> study carried out by FORS and published in FORSbase). We did not provide examples to not influence participants in their answers, especially since lesser-known sources were of primary interest for this project.			

<b>B.3 Which are the most important criteria for you in considering data for reuse?</b>	<b>Options</b> (multiple-choice with up to three answers, mandatory)	
	- Comprehensive metadata (e.g. keywords) - Additional materials (e.g. methodology) - Availability of raw data	- Normalized and clean data - Usability with familiar tools - Ease of access - Transparent licencing - Trustworthiness of the source
<b>Description:</b> We wanted to deliver a ranking of (quality) criteria considered for selecting relevant data sets and data sources. Participants could choose up to three answers out of the options as their main criteria, thus establishing a ranking over all responses. The answering options were selected to cover different layers of quality assessment. Metadata and additional materials like documentation target the		

<sup>22</sup> Research data in the context of DARIAH-DE. Available at: <https://de.dariah.eu/en/weiterfuehrende-informationen>

<sup>23</sup> ALLEA (2020), p, 8

<sup>24</sup> FORS: Selects Swiss Election Study. Available at: <https://forscenter.ch/projects/selects/>

quality of the **data description**. Normalized and raw data as well as usability with familiar tools cover the state and technical properties of the **data itself**, while also encompassing the dimension of what data can be made available by data providers. The ease of access, clear licensing and trustworthiness cover quality aspects of repositories and other **data providers**.

<b>B.4 These other criteria are important for me in selecting data and sources.</b>	<b>Options</b> (free text field, voluntary) <i>Free text field</i>
<b>Description:</b> Participants had the option to name additional criteria to the options in question B.3.	

The survey continues in part C with the goal of obtaining information about data or data sources of interest that are not currently available for reuse.

<b>C.1 Which data sources are you interested in, but do not offer reusable data yet?</b>	<b>Options</b> (free text field, voluntary) <i>Free text field</i>
<b>Description:</b> This question complements the overview of valuable data sources in the SSH.	

The questionnaire ends in part D with a free text field for respondent’s general comments and questions. The availability for follow-up personal interviews and the participation in a prize raffle could be selected by checking appropriate boxes and providing an e-mail address.

## 5. RESULTS AND ANALYSIS

We present our results according to the structure of the survey. Each question constitutes a separate subchapter. Every subchapter visualizes the responses in a chart and provides a detailed description and interpretation of the chart. Wherever relevant, possible correlations and possible links to other questions were included.

### A.1 What kind of institution do you belong to?

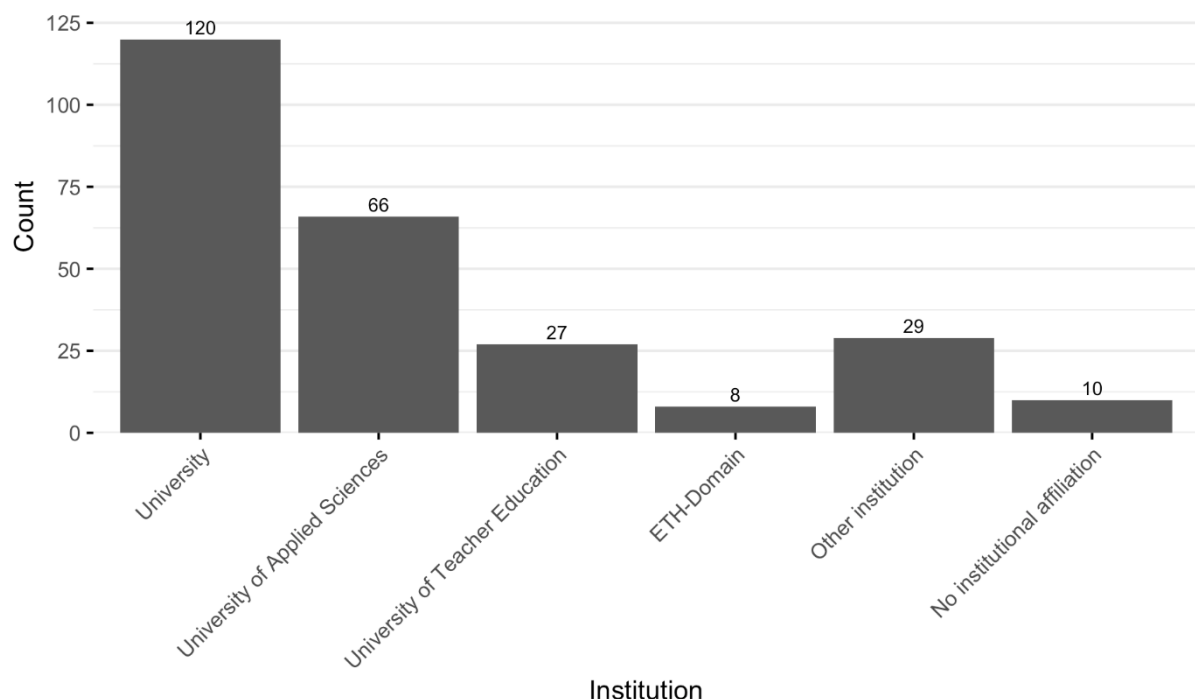


Figure 1: Participant’s distribution across institution types (absolute count)

46,15% of respondents belong to a University, 25,38% to a University of Applied Sciences, 10,38% to a University of Teacher Education and only 3,08% to the ETH-Domain. The frequent selection of “Other institution” (N=29, 11,15%) shows that several participating researchers are not affiliated with a common higher education institution (HEI). This can mean non-university research institutions, societies and service providers from the public and private sector (like EHB, FORS).

The results are expected since research in the SSH is predominantly conducted at classical research universities, whereas the ETH-Domain focuses more on research in science and technology.

## A.2 What is your field of research?

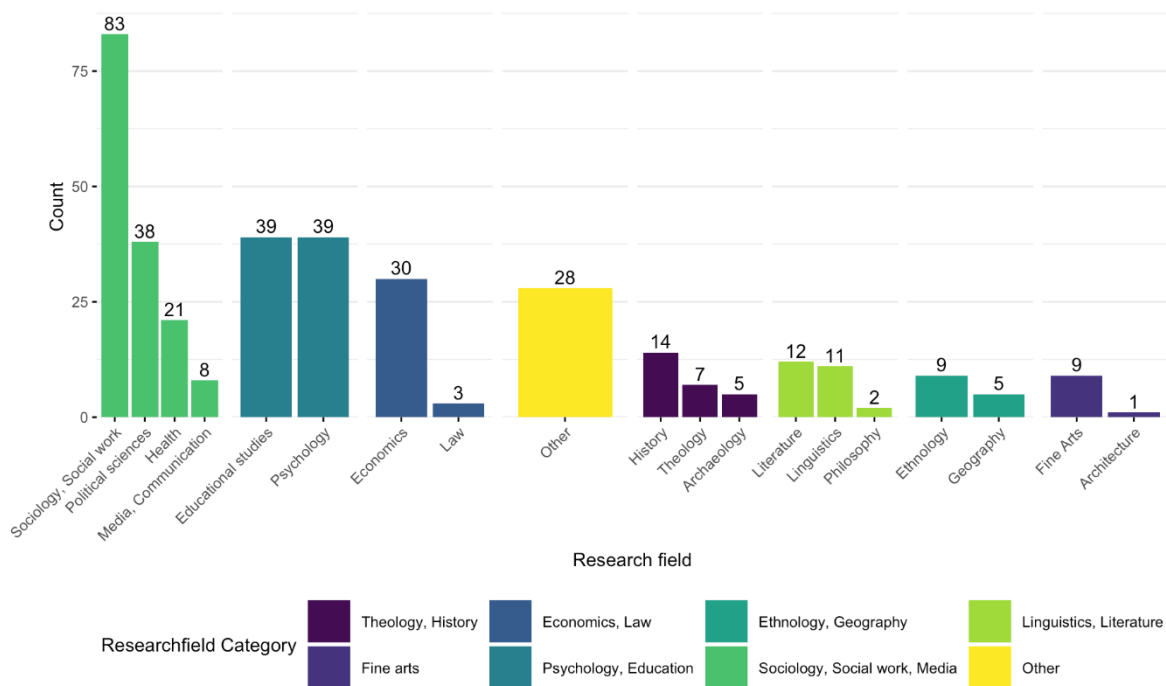


Figure 2: Participant's distribution across research fields (absolute count, grouped and sorted by category)

The distribution of research fields is summarized in Figure 2. The absolute count of responses for each field is displayed with bars. The research fields are grouped into categories and sorted from left to right by the cumulative frequency for the categories in descending order.

Most participants selected Sociology and Social work as their research field (N=83). As distant second and third followed Educational studies and Psychology (each N=39). Researchers from Political sciences (N=38), Economics (N=30) and Health (N=21) completed the main five research fields represented in this survey. Many participants (N=28) took advantage of the possibility to self-declare their not listed specialty by choosing "Other" (alone or in addition to a listed option) and filling in a free text field. A majority of answers name a research field that is still part of the overarching umbrella of SSH (e.g. Digital Humanities, Demography, Criminology). In some cases, a high-level interdisciplinary approach across different fields was mentioned (e.g. a listed answer in combination with medicine, environmental science or sports science).

## A.3 How many years of experience do you have in research?

Participants in this survey have a diverse background regarding their experience in research. The time intervals vary in duration. This can explain higher numbers of researchers in the two categories in the middle (6-15 years and 16-30 years). However, we think that our sample in general equally includes researchers from early, middle and late stages of a scientific career. Hence, our observations and results are not strongly influenced by the skewed distribution.

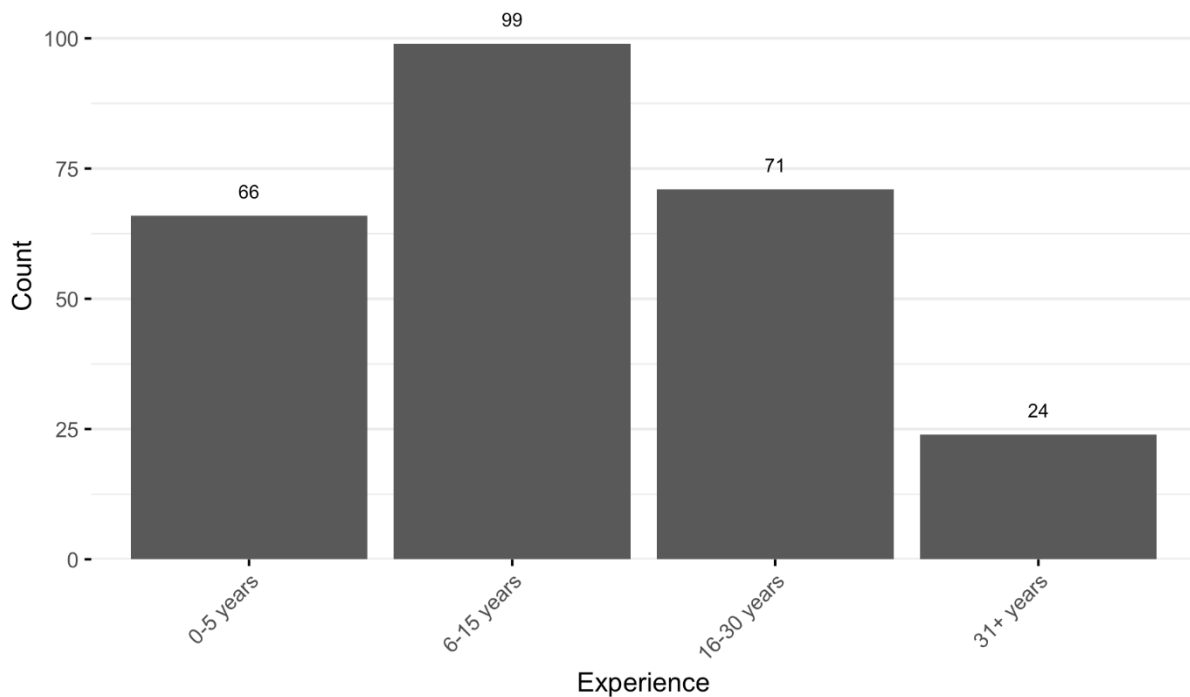


Figure 3: Participant's distribution across research experience (absolute count)

### B.1 Did you reuse data in the past?

Overall, 60 participants (23,08%) indicated that they have not reused data in the past by selecting “No”. Consequently, 200 respondents (76,92%) selected “Yes”.

The definition of data was chosen very broadly, including scholarly publications. The definitions of both terms, “data” and “reuse” were provided with this question to support the participants. One possible explanation for lack of experience in reusing data can be that researchers in some of our considered research fields do not regard their research assets as “data”.<sup>25</sup> Perception of the term “reuse” might also have been diverse, as shown by some of the participant’s general comments at the end of the survey. Regarding scholarly publications, the intensive research with primary sources might be associated with “reuse”, the study and citation of secondary sources like articles and books however not.

#### Correlation between data reuse and research field

Figure 4 shows the answers regarding the data reuse in combination with the selected research fields, the size of the balloon indicating the number of responses and the color the classification of single research fields to our chosen categories. Our calculations indicate a correlation between research fields and the reuse of data in general (SPMI  $X^2 = 104,33$ ;  $p < 0,01$ ). More specifically, the research area of psychology is the only discipline in which a narrow majority of participants indicated that they had never reused data.

This summary might point towards research fields in which (based on our responses) data reuse is unanimously common (e.g. Archaeology, Fine Arts, Linguistics, Literature). Of interest might also be further investigations into research fields in where a reasonable number of researchers have not reused data in the past (e.g. Psychology, Sociology/Social work, Educational studies). Some of these fields may rely more intensively on collecting own data, especially when context and individual characteristics of an examination are of utmost importance. Sometimes sensitive data might be anonymized in a way that impedes further reuse. Data might also generally not be available because anonymization is not possible without affecting the data and its significance. Furthermore, a closer look can be taken into individual records selecting “No” and their further responses concerning interesting data sources that do not offer reusable data yet (question C.1). Finally, their answers might also be contrasted with other responses from their research field selecting “Yes” here.

<sup>25</sup> See also ALLEA (2020), p. 8

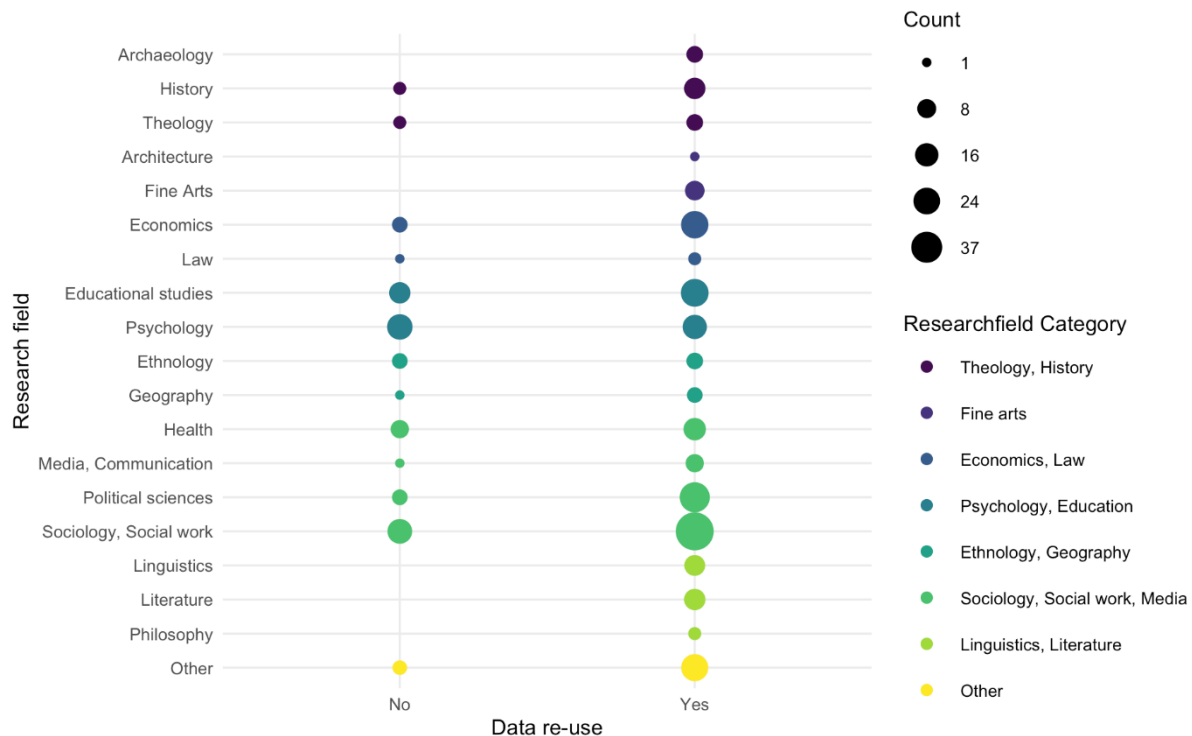


Figure 4: Participant's data reuse in combination with research fields (absolute count, grouped by category)

### Correlation between data reuse and research experience

Another noteworthy observation is that data reuse proportionally becomes more common with increasing research experience. The connection between these aspects appear to be statistically significant (Chi-squared test,  $\chi^2 = 9,77$ ;  $p = 0,02$ ). It is generally more likely not to reuse data in the early years in research. At the start of one's career, the need to collect own data is possibly also more common than to build on existing data. Reuse becomes more probable after some time.

### B.2 Please describe kind, source and main purpose of the data reused

Our participants were asked to describe certain scenarios in which they reused data in the past. The question consists of the three aspects kind of data, source of the data and their purpose for utilizing it.

#### Kind of data

Within the confines of our choices, numerical data was the most utilized type of data among researchers of the SSH (N=115, 33,10%). This includes statistics and other quantitative data from existing collections. Scholarly publications follow in second place (N=96, 27,70%). Another relevant kind of data are surveys and interviews (N=76, 21,90%). Digital artefacts like text (editions), pictures and audio-visual materials are named less often (N=46, 13,30%). Our results are summarized in Figure 5 below.

#### Correlations between kind and research field

The distribution can be explained by considering the research fields of participants in this survey, which showed peaks in within the social sciences (especially Sociology/Social work) that more commonly use numerical data, surveys and interviews. Scholarly publications are most likely relevant for every research field, meanwhile digital artefacts might be of higher relevance within the humanities. In general, our examination points towards a relevant correlation between selected research fields and the kind of data being reused (SPMI  $\chi^2 = 290,77$ ;  $p < 0,01$ ). More specifically, the combination of digital artefacts with the research fields archaeology, art, history, and literature shows statistically significant dependencies (all  $p < 0,01$ ). However, the test does not reveal the kind of dependency.

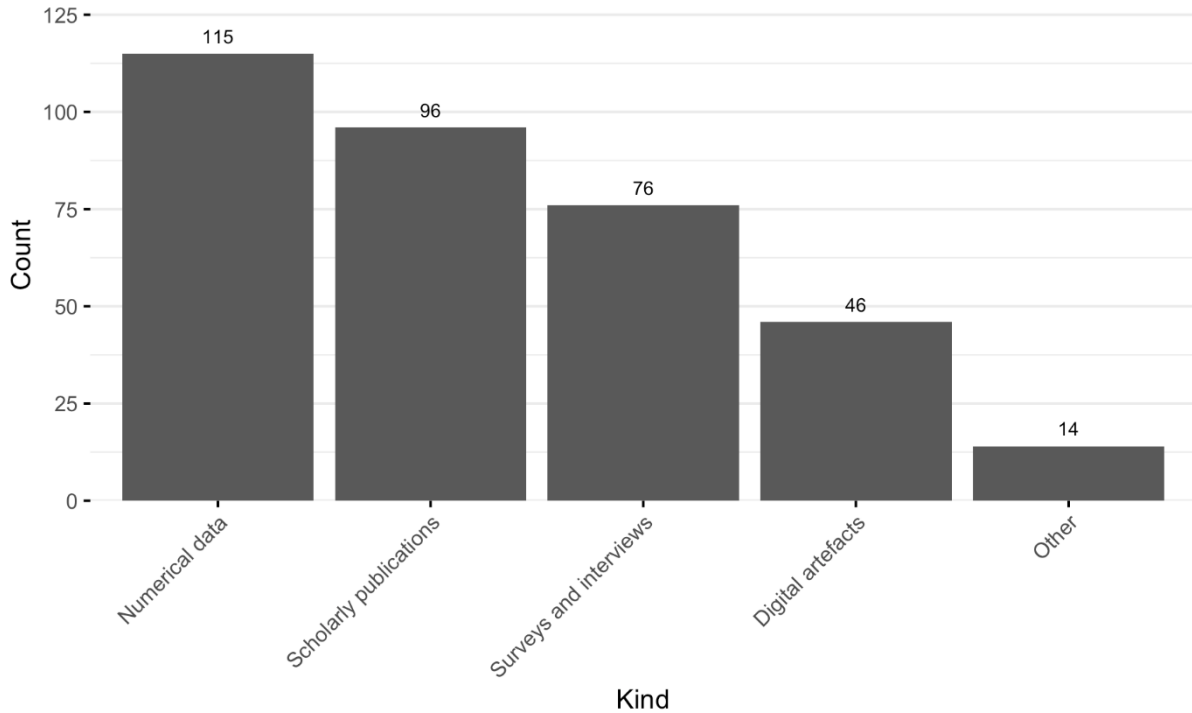


Figure 5: Kind of data reused (absolute count)

**Data source**

This diversity of answers made it difficult to provide a genuine picture of relevant data sources in the SSH for researchers in Switzerland. The comprehensive list of sources (more than 200 different names) might however still be of interest to future purposes, especially when it comes to addressing lesser-known sources that are not accessible through popular repositories and similar platforms. It is provided in Appendix II.

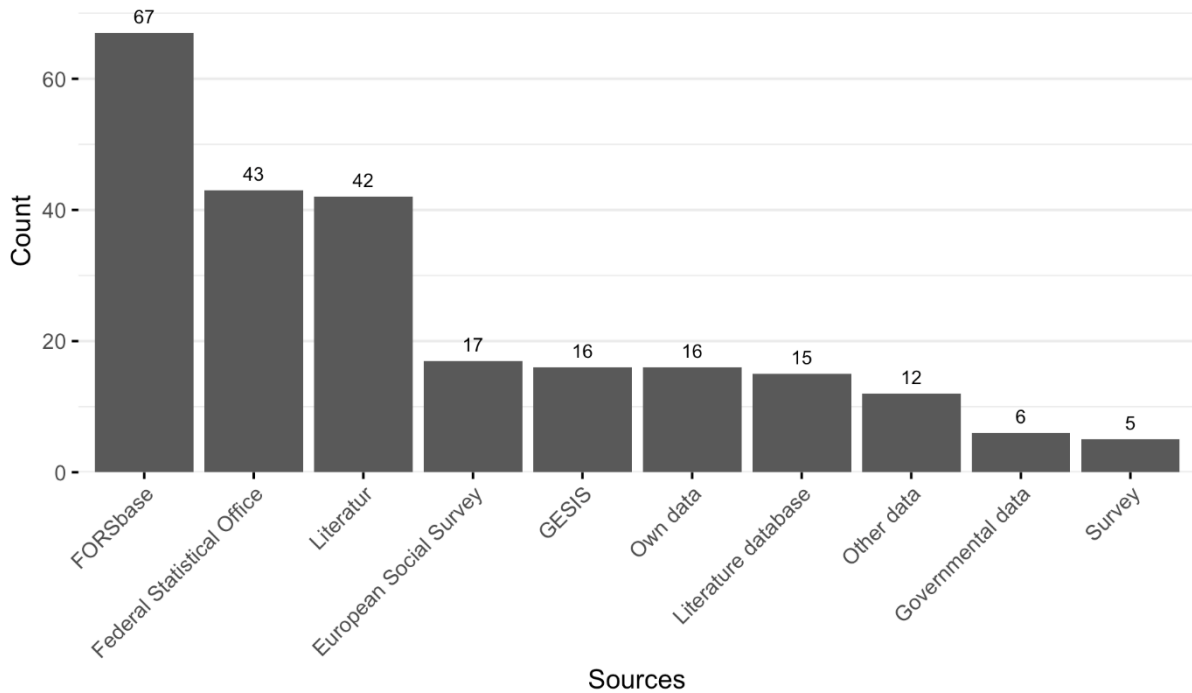


Figure 6: Ten most named sources (absolute count)



Figure 6 provides a summary of the ten most named data sources by our participants. The count refers to the total number of times one source was quoted. One respondent can name the same source theoretically up to five times describing different scenarios for data reuse. Yet, the distribution is very similar considering single records.

FORS with their digital repository FORSbase was named 67 times, followed by the Federal Statistical Office (FSO) with 43 mentions. The next individual source is the European Social Survey (ESS) stated 17 times. GESIS with their digital archive follows suit with 16 mentions. Our list contains several collective categories. Literature in general and works of specific authors were given regularly (“Articles”, “books”, “publications digitales”, “Journals”, “Kafka”) (N=42). Own data is referenced as well (N=16). The summary is completed by publication databases (N=15), other data sets (“Forschungsdatensatz”, “fremde datensätze”, “Studies from colleagues/students”) (N=12), general administrative or government data (N=6) and non-specified surveys (N=5).

The term “publications” or “literature” might be fuzzy, meaning primary sources for text-based research and/or generally secondary and research literature. For our categorization, those sources were all grouped to “Literature” – except for complete editions and corpora that were concretely named. “Publication databases” subsumes all statements towards aggregators of published work, normally secondary and research literature.

As an additional method to characterize our results we attempted a general classification of all sources to the following categories: Generic data sources (no specialist focus), subject-specific data sources (divided into individual data sources and collective data sources, like repositories and collections of various materials) and other data sources (general and not assignable terms).

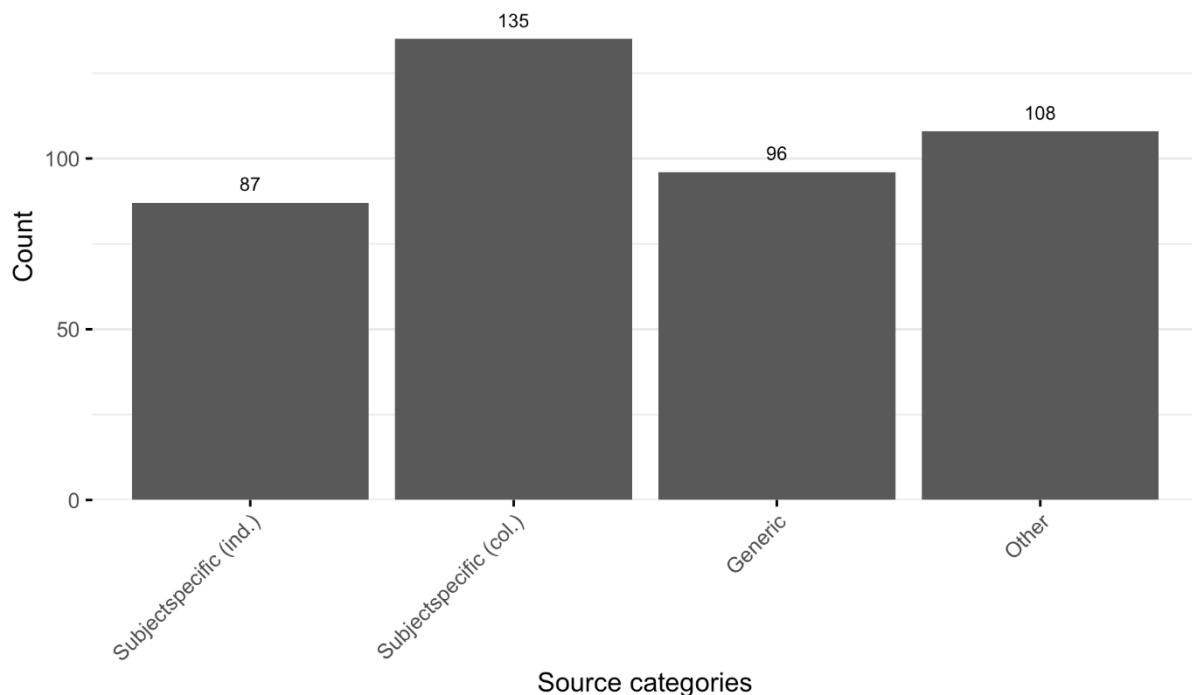


Figure 7: Categorization of all data sources (absolute count)

This classification shows that a major part of the sources mentioned were subject-specific (N=87+135). Within this category, the number of collective sources was higher. This includes some of the central subject-specific repositories (like FORSbase, GESIS data archive), but also many independent sources that provide a variety of data to certain themes. The number of sources that provide data without a specific focus on a subject is still significant. Included are popular mentions like the FSO and the UK Data Service, but also other general repositories. The large number of “Other” sources illustrates our challenges in dealing with answers on many different levels and of different informative value.

#### Correlation between data source and kind of data

Certain data sources were given in connection with certain kinds of data. According to our responses, FORSbase was consulted mostly for numerical data, but also for surveys and publications (the later possibly referencing the same type of item). FSO was accessed for numerical data and surveys. GESIS was mentioned only in connection to numerical data. Unsurprisingly, ESS provided numerical data

unanimously. The SPMI test for independence reveals a significant correlation between the kind of data and the data sources (SPMI  $X^2 = 1269,22$ ;  $p < 0,01$ ). However, the statistical conspicuities are limited to the combinations of the data source “Literature” with the data types “Scholarly publications” and “Digital artefacts” (which includes digitized texts). This connection is self-explanatory.

### Purposes of data reuse

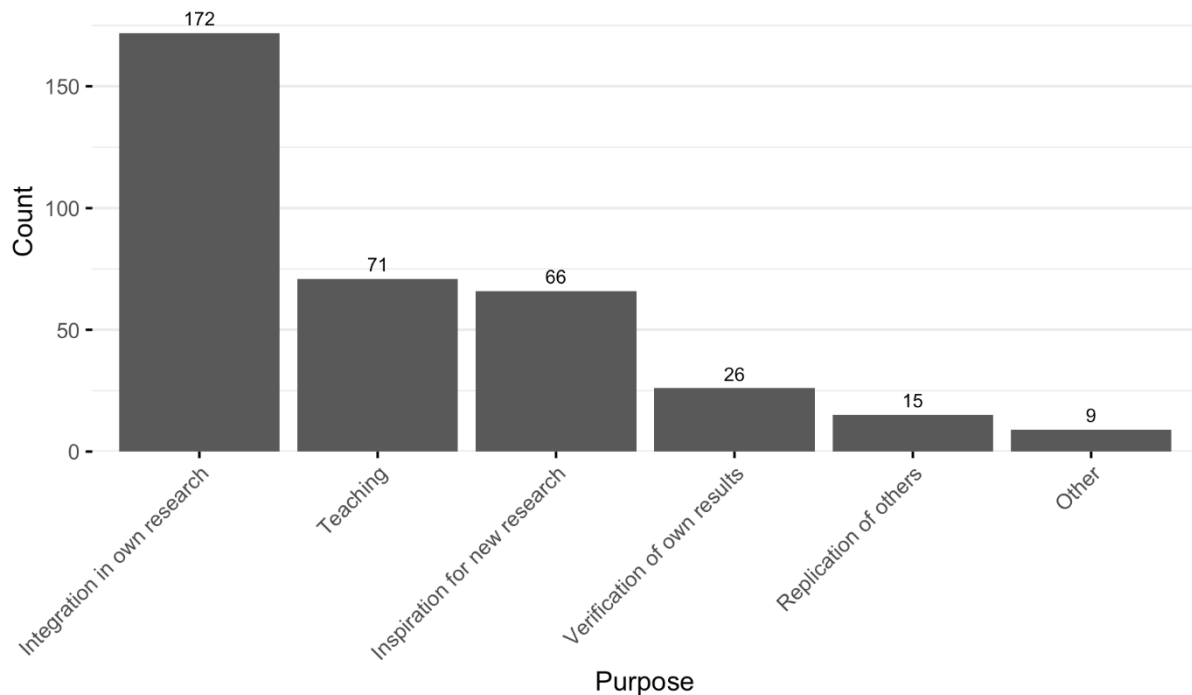


Figure 8: Purpose of data reuse (absolute count)

The by far most common reason to utilize existing data (according to our responses) was the direct integration in one’s own research (N=172, 47,90%). We intended this choice to imply using parts of information or the whole asset for new research issues. Available data was also used for teaching purposes (N=71, 19,80%), closely trailed by existing data offering inspiration for new research questions (N=66, 18,40%). In rarer cases, previous data was used to verify one’s own research (N=26, 7,20%) or replicate the results of others (N=15, 4,20%).

### Correlation between purpose and data source

Testing for correlations between the data sources and the purpose for data reuse showed significant results mainly for the purpose of replication (SPMI  $X^2 = 1272.38$ ;  $p < 0,01$ ). Relevant sources for replication include OSF<sup>26</sup>, DODIS<sup>27</sup>, Varieties of Democracy<sup>28</sup> and Chesdata<sup>29</sup>. However, the small number of mentions for each of these sources (< 5) might influence this analysis. Further investigations can continue from this point.

### B.3 Which are the most important criteria for you in considering data for reuse?

The trustworthiness of the data source can be identified as the prime criterion in selecting data (N=117, 20,17%). “Source” might here again reference a certain data provider (repository), but also the authors of a data set. Trust has to be seen as a consequence of other boxes being checked, like the reputation of the source or the author, the prominence of a data source within a certain community, prior usage of the data, transparency in methods as well as the consistency, completeness and lack of error in the data itself.<sup>30</sup> To enable reuse of high-quality data, trust has to be established. It is therefore particularly important for data sources and data providers to define a complete list of factors that enable trust and to meet these criteria. The same holds true for initiatives like the Connectome trying to offer comprehensive evidence and access to relevant data sets and connecting them, which requires a selection of high-quality and trusted data sources.

<sup>26</sup> Open Science Framework. Available at: <https://osf.io/>

<sup>27</sup> Diplomatiscche Dokumente der Schweiz. Available at: <https://www.dodis.ch/>

<sup>28</sup> Varieties of Democracy. Available at: <https://www.v-dem.net/en/>

<sup>29</sup> Chesdata. Available at: <https://www.chesdata.eu/>

<sup>30</sup> Gregory et al. (2020), p. 40 f.

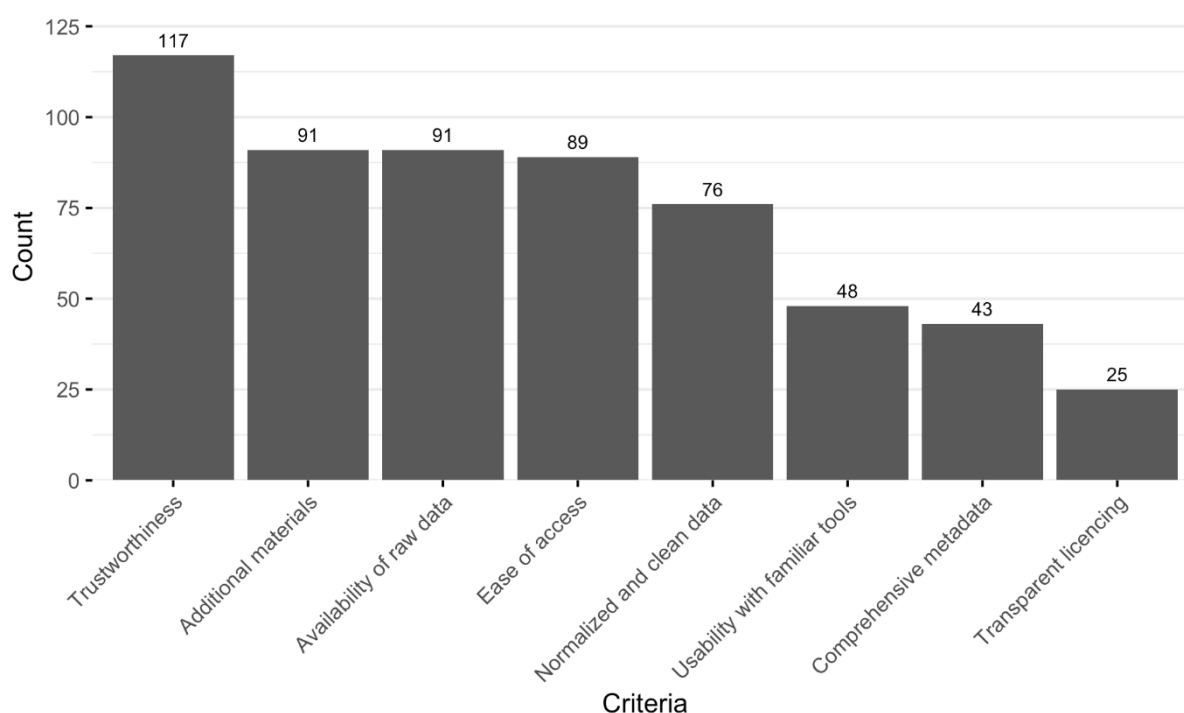


Figure 9: Criteria for selecting data sets and data sources (absolute count)

As the other main criteria for selecting data follow the existence of additional materials (like documentation, methodologies) (N=91, 15,69%), the availability of raw data (N=91, 15,69%) and the ease of gaining access to data sets (N=89, 15,34%). It is interesting that raw data is valued slightly higher than normalized and clean data (N=76, 13,10%), despite the fact that the validation of own results or those of others was not noted as an important purpose of data reuse. These are instances where work with raw data to reproduce existing results is essential. Presumably, raw data is valued highly for the integration of data in one's own research as well, being free from additional tempering. Furthermore, raw data can also contain information not relevant for the original analysis that is however important for reuse.

One conclusion is that **both kinds of data (raw and clean) are similarly important for researchers**, thus the availability of both states of data sets is desirable. Their individual usefulness is likely based on specific use cases and circumstances. Consequently, researchers themselves should be motivated and enticed to share, publish and archive data in different states of processing, together with sufficient documentation and explanations. Repositories and other data sources should offer the opportunity to provide and publish both raw and clean research data.

A second interesting observation is that the **access to additional materials is valued significantly higher than comprehensive metadata** (N=43, 7,41%) when it comes to describing data. Different kinds of data are usually described differently, strongly depending on the place they are made available as well. Repositories and directories of cultural institutions like archives and libraries utilize metadata to a large extent to describe data within, meanwhile data provided on individual websites or portals (like statistical data from governmental institutions in many cases, but also individual landing pages for surveys) often lack comprehensive and standardized metadata. The collection of numerical data and surveys often requires additional documentation and explanations how data was acquired and processed. Digital objects like digitized text and pictures otherwise are commonly described with metadata. Hence the similar distribution (high mentions to numerical data and surveys, important criteria of additional materials; lesser numbers in publications and digital objects, less important criteria of metadata) could be explained. Additionally, metadata is most relevant for finding data in the first place, and this aspect is not so important for then selecting data for reuse. Compared to other aspects, comprehensive metadata might also just be less relevant.

#### Correlation between criteria and purposes

Certain correlations between selected criteria and the purposes for utilizing existing data appears to exist (SPMI  $\chi^2 = 288,55$ ;  $p < 0,01$ ). Significant correlations could be found between the purpose of integration into one's research and the five most important criteria "Additional materials", "Normalized

and clean data”, “Availability of raw data”, “Ease of access” and “Trustworthiness of the source” (all  $p < 0,01$ ). Trust is very relevant in connection with reusing materials for teaching ( $p < 0,01$ ) as well.

**Correlation between criteria and data kind**

The distribution of selection criteria across various kinds of data is (proportionally) very consistent, our calculations hint towards existing correlations (SPMI  $\chi^2 = 193,39$ ;  $p < 0,01$ ).

**Correlation between criteria and research fields**

The data suggests a connection between certain research fields and the criteria for selecting data sets or data sources (SPMI  $\chi^2 = 238,12$ ;  $p < 0,02$ ). The distribution of criteria among research fields is rather consistent (see Figure 10). Only the combination of archaeology and license shows an unexpected presentation in the collected data ( $p < 0,02$ ).

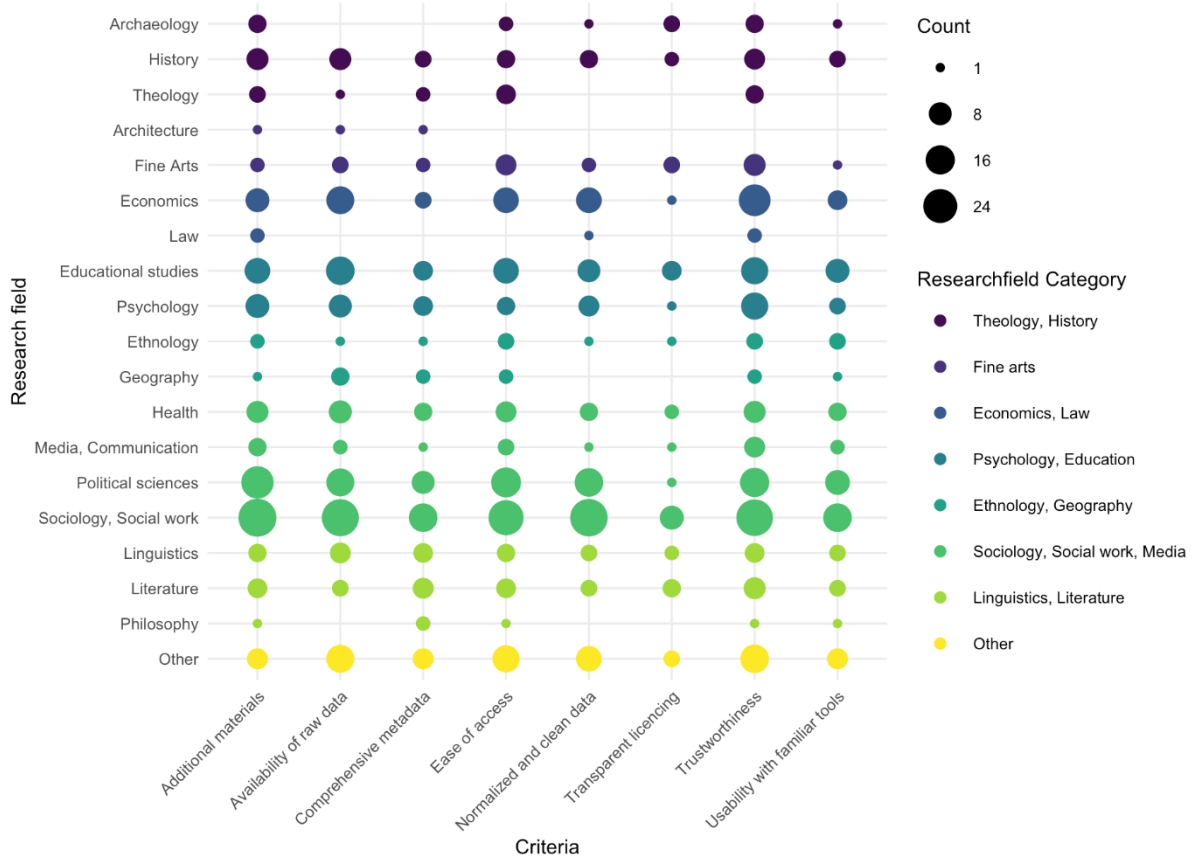


Figure 10: Criteria for selecting data in combination with research fields (absolute count, grouped by category)

**B.4 These other criteria are important for me in selecting data and sources**

In addition to the initial selection of prevalent quality criteria in question B.3 we wanted to explore further important aspects for choosing data sets and data sources. We provided a free text field for the participants to name those criteria. Our analysis is based on identifying reoccurring, similar answers. In some cases, answer options given in B.3 were repeated here, perhaps suggesting that this aspect is not within the top three most important criteria being able to be selected, but still worth mentioning here. Since these answers offer a more detailed look at certain criteria, they are nevertheless included in this overview.

96 participants answered this voluntary question. Answers sometimes were addressing multiple aspects and were then counted in multiple categories. Figure 11 displays the most pertinent answers congregated to general topics. Light-blue boxes indicate that these criteria aspects were already provided in question B.3. Dark-blue boxes point towards new criteria for selecting data sets and data sources.



Figure 11: Additional criteria for selecting reusable data named by participants (congregated to general topics)

The **content-related fit and relevance** of the data regarding the research issue and field was the most important new criteria given by the participants (n = 21). This is not surprising, as data reuse often involves integrating that data into one's own research (see purpose of data reuse, Figure 8). Therefore, a strong connection can be expected. This aspect was considered for admissions to our predefined list of criteria but was omitted in the end because it was implicitly taken for granted. This however does not have to be the case, especially for reusing data for teaching and getting to know new research methods. Relevance for a research issue can be a quality criterion for data sources and should be included in following studies.

A second important new criterion for selecting data sets and data sources is the **quality of the data** itself. This can include comprehensiveness and validity as well as technical properties of data files. Our predefined set of criteria was aimed at finding the most relevant aspects that define the quality of data. Necessary limits constrained the extend of this list. Like the relevance for the research field it is not surprising that the quality of data plays a big part in assessing data sets for reuse. This reassurance is valuable, data quality however can have miscellaneous manifestations concerning different data types and research fields.

The analysis of this question seems to support the selection of criteria for our list given in question B.3 to some degree as well. Many aspects were repeated here in high numbers, for example the availability of additional materials like a detailed documentation of the data, methodology and research methods used in collecting, processing and evaluating the data. Furthermore, the ease of access to data sets was repeated and characterized in more detail, naming costs as an impeding issue and wishing for free, easy and fast availability and access to data. Defined terms of use for data, the usability with familiar tools and the availability of different expressions of data (the general state of the data set: raw data, cleaned data; the granularity of the data) were mentioned as well. Scientific respectability and trust in data is another important mention, considering mainly the data in general, but also the repositories where they can be found and the authors who published the data.

### C.1 Which data sources are you interested in, but do not offer reusable data yet?

The main interest of this analysis concerns the data sources currently used by researchers in Switzerland and the criteria they base their choice on. Still, a look at data sources that do not offer reusable data yet does contribute to a more comprehensive picture and gives valuable pointers towards addressing deficiencies in data availability for research. In response to this question, 129 participants answered the voluntary free text field, 17 answers were not considered (“None”, “I don’t know any”, etc.) (N = 112). We group the answers into overarching categories and provide an overview. Some participants provided information on multiple issues. These answers were counted several times to all appropriate categories.

The answers were focused on referencing certain types of data and data collection contexts, more so than concrete data providers. The answers are summarized in Figure 12.

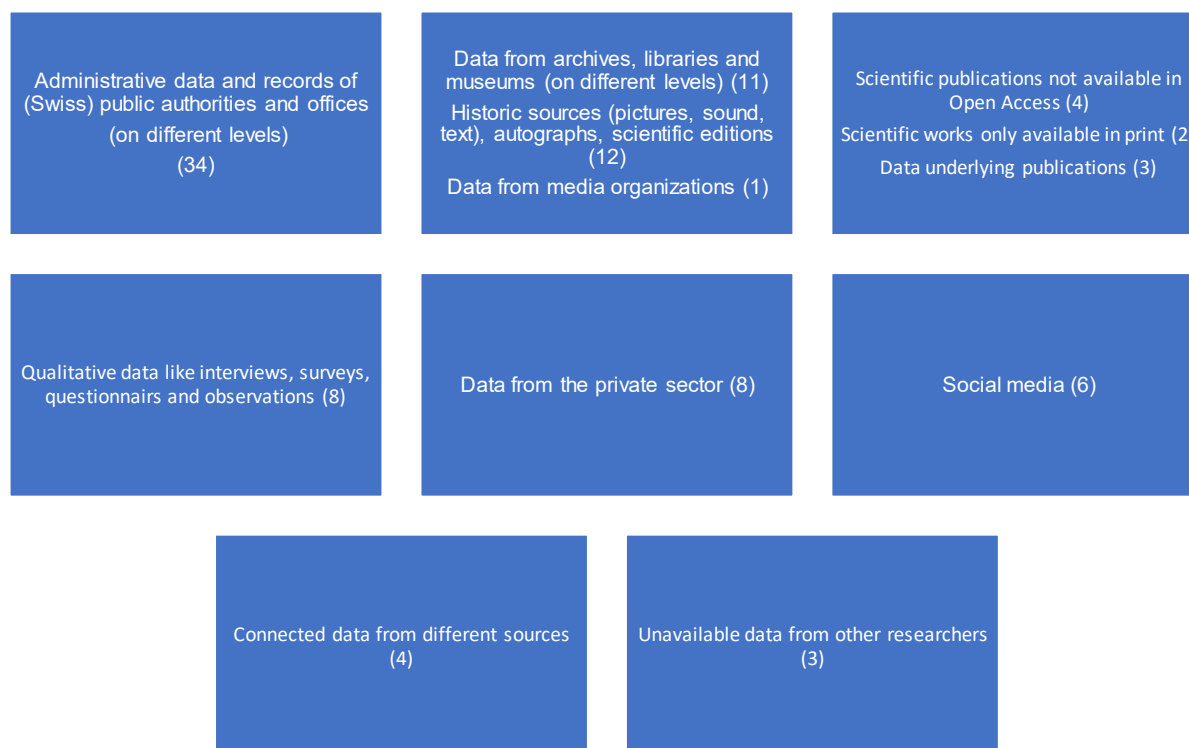


Figure 12: Interesting sources without reusable data named by participants (congregated to general topics)

Many participants (34) noted the unsatisfactory situation concerning **administrative data** gathered by public authorities and offices in Switzerland. Data from all levels of governmental institutions (municipal, cantonal, federal) are of high interest but can often not be used at this point in the desired way (according to respondents in this survey). Accessing available data is often fraught with many difficulties (bureaucratic burdens, data protection agreements, costs, etc.). Respondents also addressed the need for more digitally available information provided by these bodies. Several agencies report some of their collected data to the Swiss Federal Statistical Office (FSO) but underlying raw data and additional information is not accessible. According to participants, in many cases no data is available for reuse at all or the lack of data limits possibilities of utilizing the present information. All this makes easy reuse of this data arduous and forms an obstacle for conducting new research based on these unexhausted data treasures.

Named in connection with administrative data was (among others) data on education (4), court-, justice-, law enforcement- and penitentiary data (4), data on taxation (3) as well as health-related data (children and youth welfare & medicine, clinical psychology, obstetrics, public health institutions and insurance, missing health registries). Others mentioned register data, information on migration, economic power, personnel, recruiting and management. Missing cantonal archaeologies, social statistics, data on elections and direct democracy add to a diverse picture. Explicitly mentioned sources include the Swiss National Cohort (2), PISA, FOSS, FWS Job-Stress Analysis and EU surveys (EU-MIDIS and EU-LGBT).

A second major potential source can be identified in **data from archives, libraries and museums**. These cultural institutions (on different governmental levels) offer a vast and rich treasure of objects needed for research. This includes many historical and unique records (text, photographs and paintings, audio and video recordings). Researchers are interested in originals like autographs as well as edited



works like scientific editions. Despite considerable efforts in the past, not all desired data is digitized and made available by the institutions or via digital platforms (like Europeana<sup>31</sup>). **Digitized works may not be reusable according to survey respondents because of insufficient quality or data formats, technical properties and supported standards (like IIF)**. High-level processing, structuring, indexing and analysis of these works is desired – and currently often missing – as well. Participants named the Digital Image Archive of Medieval Music (DIAMM)<sup>32</sup>, the Thesaurus Musicarum Latinarum (TML)<sup>33</sup> and The Dictionary of Old Norse Prose (ONP)<sup>34</sup> as interesting, but currently not reusable data sources. For museums, providing high-grade digital copies of paintings or 3D-scans of physical objects in a way that researchers can use them for their own work can be constrained by technical possibilities and the law. Archives cannot provide data on files under term of protection. Legal restrictions pose a considerable constraint on the availability of cultural data as well as data in general. Reuse might be enhanced for available digitized works with an increased usage of standard free licenses (like Creative Commons<sup>35</sup>).

In addition to primary sources in the form of digitized works from cultural institutions, **scientific publications** are another significant type of data in the SSH. Some scholarly works are published in Open Access to be freely available and reusable. Other publications are made accessible through university libraries and consortia for students and researchers at Swiss institutions. However, still a quite significant number of scientific publications is not easily available. This is especially the case for the SSH, since publications in the form of (printed) books and chapters is a highly regarded form of publishing research results. The Open Access publication of books gained traction only in the most recent years by publishers and researchers alike<sup>36</sup>. Some publications are also not available in digital form, which can impede research. The fact that data sets underlying published works such as books, articles and reports are not made available by default was criticized as well.

Data from the **private sector** generates interest for our surveyed researchers too. Among the mentions were Google, credit card firms, commercial job portals and networking sites like LinkedIn and Xing or the pharmaceutical industry. Researchers are furthermore interested in **social media data**. Interactional behavior, mobility data and possible personality profiles can be useful for research.

Some participants voice the need for **interconnected data** from various sources and across disciplines. The development of the SWITCH Connectome and other specific services in the future could support researchers in this regard.

### Comments and questions

Participants of the survey were offered a free text field to share additional comments and questions about the issues raised in this questionnaire. Among some 50 responses, several major themes emerged. One of the most frequent comments concerned the **problematic notion of "data"** as well as struggling with the particular understanding thereof in this survey. For instance, several respondents commented that it was difficult for them to resonate with the idea of replicability of research results. Although it is often expected in research fields dominated by quantitative methods, that is not given in qualitatively oriented research in the same way. In many cases, materials used or collected in qualitative research are highly context-sensitive and could be (re-)interpreted in a different way not only from a new angle by another research team, but also by the same researcher at a later point in time.

Moreover, a close relationship between researchers and their research subjects particularly in respect to ethnographic research methods was highlighted. As noted by one respondent:

In ethnology, we speak of co-producing the data between researchers and research participants. In this way, we can explicitly address the active role played by researchers themselves in generating the data and reflect on their own role and the context of data genesis. For this reason, I cannot see myself making use of other researchers' observations and anonymized (!) interviews. Either these are not really anonymized, or the necessary contextual information is missing, thus, preventing further research.

This quote illustrates several issues faced by qualitative researchers in social sciences and beyond wishing to reuse others' research data or even to repurpose their own earlier datasets. On the one hand, any personal data usually needs to be removed or at least anonymized comprehensively (including the names and references of people and places) from the legal point of view. Yet then again, interpreting this data is often highly dependent on such information since the questions discussed might be substantially shaped by a particular historical or geographical context. Therefore, a certain dilemma

<sup>31</sup> <https://www.europeana.eu/>

<sup>32</sup> <https://www.diamm.ac.uk/>

<sup>33</sup> <https://chmtl.indiana.edu/tml/>

<sup>34</sup> <http://onp.ku.dk/onp/>

<sup>35</sup> <https://creativecommons.org/licenses/>

<sup>36</sup> Grimme et al. (2019)

arises: to enable "reusability" of data in legal terms, much of its context and sensitive details need to be removed. At the same time, it is exactly those details that are necessary in order to contextualize and "reuse" these data in a proper way.

Another major thread addressed some issues with the **term "reusability"** itself as well as different possible ways and levels at which a certain dataset might be reused. For example, several respondents reported to be reusing not only the results, but also the methodology (such as codebooks and/or the questionnaires) of a given survey. This use case was named particularly often for teaching purposes such as training students to work with different methods. Coupled with the desire for better accessibility of surveys and interviews as a type of data from qualitative research (see previous survey question C.1 above), future efforts could be channeled towards increasing reusability in this area. However, due to the **reusability dilemma** described above, several participants expressed their concerns over the limited potential to reuse interview recordings from other research projects, if personal and location-related details were to be removed. Overall, the comments in this free text field confirm once again the difficulties with the notion of "data" among certain respondents as well as a broadly felt uneasiness when attempting to translate it to their own research practices. Although a particular effort was put into a broad definition thereof in order to encompass the possible meaning of data in the SSH fields (see particularly our explanations on survey development on pp. 7-8 in this report), the comments suggest that some doubts still remained with regard to the apparently science-driven understanding thereof as common in the STEM fields.

## 6. DESK RESEARCH INTO RELEVANT DATA SOURCES

Our survey delivered a comprehensive list of relevant sources that contain valuable data for reuse. Certain aspects like a research field specific correlations and selection criteria were addressed within the survey. Besides the location, this project aimed to provide information about the availability of this relevant data (including ways to access it) and its quality. Therefore, a short desk research covering these topics for our four main (individual) sources is given below.

### FORSbase

FORSbase is the repository from FORS, the “Swiss Centre of expertise in the social sciences”. In addition to maintaining its national social science data archive FORSbase, FORS produces national and international survey data, does thematic and methodological research in empirical social sciences, and offers consulting services for social science researchers. The archive offers online access to study descriptions of social science projects carried out mostly since the early 1990s, including datasets where available.

Metadata and public documentation can be accessed immediately without prior registration. To access data a **registration form** must be completed including information on name, address, and e-mail (preferentially institutional e-mail). Additionally, a description of the planned research must be provided, and the user contract<sup>37</sup> accepted. In some cases, data access has to be approved by the author. After expiration of the contract, data must be deleted or the contract renewed. Data can only be used for scientific research and/or education purposes. Furthermore, any publications stemming from this data has to be communicated to FORS.

FORSbase has a clear user interface, and FORS offers numerous **guides**<sup>38</sup> as well as **personal support** for researchers. For **quality control**, every submitted dataset is assessed by FORS archivists before publication. FORSbase offers guidelines to advise researchers in how to prepare data for publication and specifies which file formats<sup>39</sup> are accepted and preferred. For metadata, FORSbase is using the **Data Documentation Initiative (DDI)**, an international standard for describing data from the social sciences. DDI allows for detailed documentation of the study performed, including information on data type (qualitative vs quantitative), methods and instruments used, and keywords. Advanced search allows not only for searching by discipline, but also for filtering by methods used (e.g. laboratory experiment) or by study type (e.g. doctoral thesis). The platform is multilingual but not the search for keywords. Search queries should therefore currently be done in all languages to be exhaustive. FORSbase was awarded the **CoreTrustSeal**<sup>40</sup>, a certificate for trustworthy data repositories.

<sup>37</sup> FORSbase: User contract. Available at: [https://forsbase.unil.ch/media/general\\_documentation/en/User\\_contract\\_E.pdf](https://forsbase.unil.ch/media/general_documentation/en/User_contract_E.pdf)

<sup>38</sup> FORS: Help & Resources. Available at: <https://forscenter.ch/data-services/help-resources/>

<sup>39</sup> FORSbase: List of accepted file formats. Available at: <https://forsbase.unil.ch/supported-file-formats/>

<sup>40</sup> <https://www.coretrustseal.org/>



FORSbase will transition in 2021 to **SWISSUbase**, which will be based on the same workflow, but will be scaled up and used as well by other institutions in Switzerland, specifically the University of Lausanne and the University of Zurich.

### Federal Statistical Office (FSO)

According to our survey, obtaining data from the Federal Statistical Office (FSO) can be attributed a high importance (N=43). In principle, the FSO has the task of collecting the necessary data on the state and development of the population, society, education, research and environment in Switzerland (compare Art. 65, Federal Constitution<sup>41</sup>). Some of the data flow together decentrally from various bodies, which means that the FSO also has a coordinating function.

The FSO is also home to the **Data Science Competence Centre (DSCC)**<sup>42</sup>, from which various services are offered. The competence centre also develops quality standards, guidelines for data protection and basic infrastructures. Furthermore, there is an “**Open Government Data Office**”<sup>43</sup> that is helping to shape the implementation of the national Open Government Data Strategy 2019-2023<sup>44</sup>. Part of this strategy is the availability of open data on the [opendata.swiss](https://opendata.swiss) portal<sup>45</sup>.

The data sets used by researchers and mentioned in this survey are very diverse and cover a wide range of topics. For example, data from the Swiss Labour Force Survey (SLFS <sup>46</sup>) or data from the Swiss Health Survey (SGB<sup>47</sup>) were obtained several times. According to our research, the studies are easy to find, for example via the FSO search mask or via the [opendata.swiss](https://opendata.swiss) portal. Comprehensive documentation is easily accessible in each case. The data itself can be obtained via catalogues and databases<sup>48</sup>, interactive tables (STAT-TAB<sup>49</sup>) or programming interfaces (API<sup>50</sup>). Some data are only available for scientific research projects and on request (e.g. SLFS, SGB data).

Due to the large number of FSO datasets available, as well as different data access, it is difficult to characterise the re-usability of the data. The trustworthiness of the FSO as a data provider and the appropriate consideration of aspects for the use of data can be assessed as very high.

### European Social Survey (ESS)

The European Social Survey (ESS)<sup>51</sup> was identified as a very relevant and valuable data source for researchers in the SSH. It is an **academically driven cross-national survey** within Europe that has been established in 2001 and conducted ever since. It measures attitudes, beliefs and behavior patterns in different population, gathering information on the social, political and moral fabric of Europe with standardized and approved indicators. The ESS is managed by the ESS European Research Infrastructure Consortium (ESS ERIC).<sup>52</sup>

The data sets, including transparent and intensive documentation, are freely available for non-commercial purposes and can be downloaded from the website. The data is available for different years, countries and themes. Access requires a **registration** at no charge containing personal information, your country, discipline, institution and intention. The site offers an online tool for data analysis as well to process and export customized data files.<sup>53</sup> In addition, it is possible to access **customized subsets** with cumulative and harmonized data from selective countries, rounds or variables. Some of the data sets are made available in the GESIS data archive and FORSbase as well.

The data is not described with additional metadata on the website. Applicable data sets can be accessed by using top-level filters (country, year, theme). A certain data set contains the data file, specific additional data and all available documentation. This includes questionnaires, contact forms, showcards, fieldwork and interview instructions and letters to respondents. The general **methodology** of the ESS is described on the website too, entailing (among others) survey specifications, source

<sup>41</sup> <https://www.bfs.admin.ch/bfs/en/home/fso/official-statistics/legal-underpinnings/confederation.html>

<sup>42</sup> <https://www.bfs.admin.ch/bfs/en/home/dsc/dsc.html>

<sup>43</sup> <https://www.bfs.admin.ch/bfs/en/home/services/ogd/office.html>

<sup>44</sup> <https://www.bfs.admin.ch/bfs/en/home/services/ogd/strategy.html>

<sup>45</sup> <https://opendata.swiss/en>

<sup>46</sup> <https://www.bfs.admin.ch/bfs/en/home/statistics/work-income/surveys/slfs.html>

<sup>47</sup> <https://www.bfs.admin.ch/bfs/de/home/statistiken/gesundheit/erhebungen/sgb.html#346123120>

<sup>48</sup> <https://www.bfs.admin.ch/bfs/en/home/statistics/catalogues-databases.html>

<sup>49</sup> <https://www.bfs.admin.ch/bfs/en/home/services/recherche/stat-tab-on-line-data-search.html>

<sup>50</sup> <https://www.bfs.admin.ch/bfs/de/home/dienstleistungen/forschung/api/api-diffusion.html>

<sup>51</sup> <https://www.europeansocialsurvey.org/>

<sup>52</sup> Prospectus: European Social Survey, European Research Infrastructure Consortium (2017). Available at:

[https://www.europeansocialsurvey.org/docs/about/ESS\\_prospectus.pdf](https://www.europeansocialsurvey.org/docs/about/ESS_prospectus.pdf)

<sup>53</sup> <http://nesstar.ess.nsd.uib.no/webview/>

questionnaire, translation, sampling, data collection, data processing and archiving and data quality assessment.<sup>54</sup>

The ESS does not only offer content-related data, but also provides high-quality insights into methodology, questionnaire design, sampling and other aspects of data collection and processing. Therefore, it facilitates the training of researchers in comparative quantitative measurements and analysis, including training courses and other materials on their website.

**Data quality** is assured by various activities in the survey lifecycle and across ESS rounds. That means a continuous evaluation of the measurement instruments regarding quality and comparability, the assessment of the sample composition as well as other external benchmark data.<sup>55</sup>

All ESS data are licensed under **CC BY-NC-SA 4.0**, the documentation under **CC BY-SA 4.0**. Other uses are possible upon request. This clear declaration of reuse conditions is positive.

### GESIS data archive

GESIS – Leibniz Institute for the Social Sciences operates a data archive preserving quantitative social research data from national and international studies and makes them available to researchers for secondary analysis. Currently, there are more than 6500 datasets available.

As with FORSbase, metadata and public documentation can be accessed immediately, and the download of data usually requires a **registration** including information on name, research field, and an e-mail, but some data can be downloaded without registration. There are four different **access categories** from category 0 that allows the reuse by everyone without specific purpose to the most restrictive category C where data is only released for academic research and teaching after the data depositor’s written authorization. If not indicated otherwise, data and documents are available only for research and teaching and for a limited time.<sup>56</sup> Furthermore, the archive has an **OAI-PMH interface**.

For the use of disclosive data subject to special access requirements and restrictions GESIS offers the Secure Data Center (SDC) and their Safe Room.

Guidelines for data preparation are available and all data is briefly checked by the archive before making them available for download.<sup>57</sup> Additionally, GESIS offers fee-based services including quality assurance and documentation of survey data.<sup>58</sup> The data archive has implemented the **DataCite Metadata Schema** and the **Data Documentation Initiative (DDI)**. The data archive was awarded the **CoreTrustSeal** in 2017.

## 7. ADDITIONAL NOT COVERED TOPICS

Since the main aim of this lab was to identify relevant data sources for researchers in Switzerland and to investigate access possibilities and selection criteria, many other interesting research questions were omitted to keep the survey concise – thus encouraging responses. To contribute to a complete picture, we list additional topics not covered in our survey here. This might be useful for future efforts to investigate data management in the SSH more comprehensively.

One question might address the ways researchers **search and discover data** in the first place. Answers could include a general search engine (like Google), specific search engines for scholarly publications and data, literature, general digital repositories, subject specific digital repositories, colleagues, social media, library catalogues and others. In addition, repositories and other platforms could be named in a subsequent inquiry, thus specifying data search and interesting data sources.

In addition to discovering data and reusing data, it is interesting to look at how researchers **publish and share their own results** as well. Answers may include not sharing or publishing your own data at all, sharing upon personal request, publishing in a general repository, publishing in a subject specific repository, publishing data as part of a publication as supplementary material, publishing data on one’s personal website, using an institutional website or other ways. Consequently, certain answers can be followed up by inquiring specific names. This aspect was omitted since many studies already cover the publication of research data. However, the point of interpersonal sharing of data and the focus of data publication in certain research fields might be very interesting for further investigations. Specifying which

<sup>54</sup> <https://www.europeansocialsurvey.org/methodology/>

<sup>55</sup> [https://www.europeansocialsurvey.org/methodology/ess\\_methodology/data\\_quality.html](https://www.europeansocialsurvey.org/methodology/ess_methodology/data_quality.html)

<sup>56</sup> GESIS: Benutzungsordnung. Available at:

[https://www.gesis.org/fileadmin/upload/dienstleistung/daten/umfragedaten/bgordnung\\_bestellen/2018-05-25\\_Benutzungsordnung\\_GESIS\\_DAS.pdf](https://www.gesis.org/fileadmin/upload/dienstleistung/daten/umfragedaten/bgordnung_bestellen/2018-05-25_Benutzungsordnung_GESIS_DAS.pdf)

<sup>57</sup> GESIS: Preparing Data for Submission. Available at: <https://www.gesis.org/en/services/archiving-and-sharing/sharing-data/data-archiving/preparing-data-for-submission>

<sup>58</sup> GESIS Datenservices. Available at: <https://www.gesis.org/datenservices/unsere-servicepakete>

types of data are published or shared in which way might be of interest too. At this point, sharing research data is not yet widely established in the SSH as something important and valuable. Thus, looking at possible and necessary incentives and requirements to promote sharing in the first place might be of interest.

Our question regarding the quality criteria used for selecting data sets and data sources covers the positive aspects that enable successful data reuse. Complementary, the examination of **problems encountered by researchers in reusing data** is intriguing. Answers could include not knowing how to search for data, not finding any data for one’s purpose, uncertainty about available documentation, insufficient metadata or documentation, no data available in the right state of processing or granularity, no license attached, too much effort to gain access to data, language barrier or no available preview of the data set.

Another interesting question to ask researchers about is whether in their opinion their own data is **reusable for other researchers**. A distinction can be made into whether data is considered valuable for other researchers and whether data can practically be successfully reused by others.

The **inter- and multidisciplinary nature of research data** can be a fascinating aspect as well. This can include what data from other disciplines are interesting for researchers in the SSH, but also what data of your own research fields can be interesting for outsiders. Requirements for a successful reuse and findability of data sets are just two of many further points that can be deepened here.

Finally, a closer look at **researcher’s tools** and the possible **need for data conversion** in order to actually make use of data can yield important insights.

## 8. MAIN FINDINGS AND RECOMMENDED NEXT STEPS

### Data reuse

It is noticeable that a high number of participating researchers do reuse existing data for their work. Sharing and publishing data to make it more openly accessible is important to support research. A significant number of participants however has not yet reused data (23%). This might to some extent be explained with difficulties in resonating with the terms “data” and “data reuse” for SSH researchers.

### Main data sources

Central data providers are identified as important sources for researchers in the SSH. FORSbase and GESIS data archive are prominent data repositories allowing researchers to publish and access subject specific data. The FSO as a general data provider is named second most, illustrating the importance of open governmental data.

### Additional sources

Besides the main mentions we gathered a list of more than 200 individual sources and data sets researchers utilized (see Appendix II). This data is not centrally recorded and made accessible. Therefore, the idea of creating a Connectome for research data is underlined. Connecting these various sources, however, is challenging.

### Metadata

To connect data sources in the Connectome, metadata is of high importance to describe data records. From our list of relevant sources, only the main repositories (FORSbase, GESIS data archive) currently describe their data sets with comprehensive metadata. This is however not the case for other data providers (like FSO) and the many other individual data sources. Gathering (meta)data is likely very difficult and not possible with a standardized interface or protocol.

### Kinds of data

Numerical data was the most reused kind of data in our survey, followed by scholarly publications and surveys and interviews. Digital artefacts were given less often.

### Main purpose for reuse

The main purpose for utilizing existing data is the integration into one’s own research.

### Main selection criterion

Trustworthiness of the data source is the most relevant criterion for selecting data sets and data sources. Additional materials like documentation and methodologies are of higher importance than comprehensive metadata.

### **Availability of raw and cleaned data**

Raw data and clean/normalized data are similarly important for researchers regarding the reuse of data. This should encourage researchers to also publish their own data in different states of processing and granularity. Data sources should enable the sharing of these data sets.

### **Additional selection criteria**

Relevance of the data for a specific research issue and the general quality of the data itself are important other criteria for selecting data not covered in our initial question.

### **Desiderata for data reuse**

Researchers want to reuse governmental data and digital assets from cultural institutions (like archives, libraries and museums). However, some of these data providers currently do not offer reusable data to the desired extent or quality.

These results lead us to propose the following actions as **recommended next steps**.

### **Deeper analysis of statistical results**

Due to time constraints, only a preliminary examination for potential correlations was possible. A closer look at specific results may be useful to deepen the insights gained here. Furthermore, our data could also be analyzed separately for individual research fields. This information may be helpful in targeting specific communities.

### **Deepening interviews with interested participants**

82 participants of our survey indicated their availability for follow-up interviews. A personal exchange or joint workshop can examine some of our results and analyzed topics in more detail.

### **Evaluate identified sources for the Connectome**

The identified main sources for researchers in the SSH should be analyzed in more detail concerning the integration into the SWITCH Research Data Connectome. Reviewing the comprehensive list of mentioned data sources can supply additional names to be checked for the Connectome.

### **Extend the scope of this survey**

The thematic scope of this survey can be expanded to contrast researcher's practices in *publishing* their own data and *reusing* existing data. Topics can include the relevant means and sources used, as well as the decision criteria. Another direction would be to analyze challenges and mitigating factors that impede data reuse.

Our approach can be applied to study data reuse in other disciplines as well. Some adjustments in terminology and answer options should be considered to account for differences in the disciplines. These additional studies will identify more relevant data sources for the Connectome. Furthermore, additional references would allow interdisciplinary comparisons to be made regarding the reuse of data.

## **ACKNOWLEDGEMENTS**

First and foremost, we express our great gratitude to all participants answering the survey and supplying us with important answers and comments concerning data reuse in their disciplines.

We want to thank SWITCH for funding this study and supporting us closely during the whole project with useful insights, constructive feedback and important contact mediation.

This endeavor would not have been possible without the help of central service providers in the SSH in Switzerland. DaSCH, FORS and SAGW gave impactful feedback during the survey design and later distributed the questionnaire through their communication channels directly to the researchers. We want to particularly thank Daniela Subotic and Ivan Subotic (DaSCH); Brian Kleiner, Stefan Buerli, Bojana Tasic and Monika-Michaela Vettovaglia (FORS); Heinz Nauer (SAGW).

Furthermore, we are very thankful for having had the opportunity to discuss the survey design with researchers in the fields we examined and profit from their specialist knowledge. Many thanks to Claire Clivaz, Sylvie Johner-Kobi, Roberta Padlina and Melanie Röthlisberger.

## BIBLIOGRAPHY

### ALLEA (2020):

ALLEA (2020). Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities. Edited by Natalie Harrower, Maciej Maryl, Beat Immenhauser, and Timea Biro. Berlin. DOI: <https://doi.org/10.7486/DRI.tq582c863>

### BAKOM (2018):

Bundesamt für Kommunikation BAKOM (2018): Digital Switzerland Strategy. Available at: [https://www.bakom.admin.ch/dam/bakom/en/dokumente/informationsgesellschaft/strategie/Strategie\\_DS\\_Digital\\_2-EN-interaktiv.pdf.download.pdf/Strategie\\_DS\\_Digital\\_2-EN-interaktiv.pdf](https://www.bakom.admin.ch/dam/bakom/en/dokumente/informationsgesellschaft/strategie/Strategie_DS_Digital_2-EN-interaktiv.pdf.download.pdf/Strategie_DS_Digital_2-EN-interaktiv.pdf)

### Brüwer (2019):

Brüwer, Michael (2019): White Paper – A view of Swiss universities' Network of ICT Experts on the Research Data Management Landscape in Switzerland. Available at:

[https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Organisation/SUK-P/SUK\\_P-2/RDM\\_Landscape\\_CH\\_WhitePaper\\_v1.1\\_final.pdf](https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Organisation/SUK-P/SUK_P-2/RDM_Landscape_CH_WhitePaper_v1.1_final.pdf)

### Dallas et al. (2016):

Dallas, Costis et al. (2016): European survey on scholarly practices and digital needs in the arts and humanities. Available at: <https://doi.org/10.5281/zenodo.260101>

### Gregory et al. (2020):

Gregory, Kathleen; Groth, Paul; Schamhorst, Andrea; Wyatt, Sally (2020): Lost or Found? Discovering Data Needed for Research. Harvard Data Science Review, 2(2). Available at: <https://doi.org/10.1162/99608f92.e38165eb>

### Grimme et al. (2019):

Grimme, Sara; Holland, Cathy; Potter, Peter; Taylor, Mike; Watkinson, Charles (2019): The State of Open Monographs: An analysis of the Open Access monograph landscape and its integration in the digital scholarly network. London : Digital Science, 2019. Available at: <https://doi.org/10.6084/m9.figshare.8197625>

### Koller-Meier et al. (2020):

Koller-Meier, Esther; Kugler, Manuel (2020): SWITCH Innovation Lab «Nachvollziehbare Datenqualität» Available at:

[https://www.switch.ch/export/sites/default/about/innovation/galleries/files/SWITCHInnovationLab\\_SATW\\_Ergebnisse.pdf](https://www.switch.ch/export/sites/default/about/innovation/galleries/files/SWITCHInnovationLab_SATW_Ergebnisse.pdf)

### Koziol et al. (2014):

Koziol, Natalie; Bilder, Christopher (2014): MRCV: A Package for Analyzing Categorical Variables with Multiple Response Options. The R Journal. 6. 144-150. Available at: <https://doi.org/10.32614/RJ-2014-014>

### Linkhub.ch, FORS (2020):

Linkhub.ch; FORS (2020): Accessing and linking data for research in Switzerland. Available at: [https://linkhub.ch/wp-content/uploads/2020/11/Report\\_Data\\_Access\\_and\\_Linking\\_11\\_2020.pdf](https://linkhub.ch/wp-content/uploads/2020/11/Report_Data_Access_and_Linking_11_2020.pdf)

### Milzow et al. (2020):

Milzow, Katrin; von Arx, Martin; Sommer, Cornélia; Cahenzi, Julia; Perini, Lionel (2020): Open Research Data: SNSF monitoring report 2017-2018. Available at: <https://doi.org/10.5281/zenodo.3618123>

### OPERAS (2020):

OPERAS (2020): Survey on SSH Scholarly Communication. Available at: <https://operas.hypotheses.org/operas-survey-on-ssh-scholarly-communication>

### SCNAT (2020):

MAP Open Data Survey (2020). SCNAT. Available at: [https://map.scnat.ch/en/duties/open\\_data\\_survey](https://map.scnat.ch/en/duties/open_data_survey)

### SNSF (2015):

SNSF Workshop on Data Science (2015). Available at: [http://www.snf.ch/SiteCollectionDocuments/Workshop\\_Minutes.pdf](http://www.snf.ch/SiteCollectionDocuments/Workshop_Minutes.pdf)

### SNSF (2019):

SNSF Open Research Data: Landscape and cost analysis of data repositories currently used by the Swiss research community, and requirements for the future (2019). Available at: <https://doi.org/10.5281/zenodo.2643460>

### Swissuniversities (2020):

Swissuniversities (2020): National Open Research Data Strategy. Analysis Report based on Survey and Workshop Panels. Available at:

[https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Hochschulpolitik/Open\\_Science/20201123\\_ORD\\_Grunlagenbericht\\_final\\_swu.pdf](https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Hochschulpolitik/Open_Science/20201123_ORD_Grunlagenbericht_final_swu.pdf)

**LIST OF TABLES AND FIGURES****Tables:**

Table 2: Nested survey structure .....	04
--	----

**Figures:**

Figure 1: Participant's distribution across institution types (absolute count) .....	09
Figure 2: Participant's distribution across research fields (absolute count, grouped and sorted by category) .....	10
Figure 3: Participant's distribution across research experience (absolute count) .....	11
Figure 4: Participant's data reuse in combination with research fields (absolute count, grouped by category) .....	12
Figure 5: Kind of data reused (absolute count) .....	13
Figure 6: Ten most named sources (absolute count) .....	13
Figure 7: Categorization of all data sources (absolute count) .....	14
Figure 8: Purpose of data reuse (absolute count) .....	15
Figure 9: Criteria for selecting data sets and data sources (absolute count) .....	16
Figure 10: Criteria for selecting data in combination with research fields (absolute count, grouped by category) .....	17
Figure 11: Additional criteria for selecting reusable data named by participants (congregated to general topics) .....	18
Figure 12: Interesting sources without reusable data named by participants (congregated to general topics) .....	19

## APPENDIX I: QUESTIONNAIRE

# Reuse of data in the Social Sciences and Humanities

This survey wants to identify the relevant data and sources for researchers in Switzerland in the Social Sciences and Humanities.

The gathered data and information contribute to the development of the Connectome, an ecosystem initiative coordinated by SWITCH, to make data from various disciplines, stored in different locations, easily findable, accessible, interoperable and reusable (FAIR) - thus connected.

Among all completed questionnaires we raffle [Sony WH-1000XM3 headphones](#) and a [Yohann Laptop Stand!](#)

Thank you very much for contributing!

**REMARK:** Participation in this survey is voluntary. All data is collected anonymously and treated in strict confidence. The results will be published in the form of reports, which can contain the anonymized data set. There is no intent or possibility to identify individuals through their answers. Personal data provided for follow-up interviews and for participation in the prize raffle will only be used for these specific purposes and will be deleted afterwards. This information is not analyzed in combination with the answers given in the survey. By participating in this survey, you accept these terms and conditions. For any questions and comments connected to this survey, please contact us at [researchdata@zhaw.ch](mailto:researchdata@zhaw.ch).

## Part A: General information

### A.1 What kind of institution do you belong to?\*

Please select one suitable answer.

\*must provide value

- University
- University of Applied Sciences
- University of Teacher Education
- ETH-Domain
- Other institution
- No institutional affiliation

### A.2 What is your field of research?\*

Please select all suitable answers.

\*must provide value

- Archaeology
- Architecture
- Art studies, musicology, theatre and film studies
- Economics
- Educational studies
- Ethnology
- Geography
- Health
- History
- Law
- Linguistics
- Literature
- Media and communication studies
- Philosophy
- Political sciences
- Psychology
- Sociology, social work
- Theology & religious studies
- Other, please specify



**A.3 How many years of experience do you have in research?**

\*

Please select one suitable answer.

\*must provide value

- 0 - 5 years
- 6 - 15 years
- 16 - 30 years
- 31+ years

[Reset](#)

**Part B: Please describe specific data sets you reused in the past.**

**We define data as ...** all digital materials, sources and results scholars collect, generate, evaluate and use. This can include text, digitized works, audio, video, surveys, interviews, etc.

**By reuse of data ...** we mean any use of already existing data for new endeavours. This can for example mean including certain parts in your own research, teaching or replicating or verifying previous results.

**B.1 Did you reuse data in the past?\***

\*must provide value

- Yes
- No

[Reset](#)

**B.2 Please describe kind, source and main purpose of the data reused.**

We would appreciate it if you describe several data sets!

**Kind\***

\*must provide values for the first row

▼
Scholarly publications (books, papers, ...)
Digital artefacts (text editions, pictures, audio, video, ...)
Numerical data
Surveys and interviews
Other

Please select one suitable answer.

**Source name\***


Please name the source.

**Purpose\***

▼
Inspiration for new research questions (e.g. for proposals)
Integration of data into your own research
Teaching
Verification and cross-checking of your own data/results
Reproduction and replication studies of others
Other

Please select one suitable answer.

**B.3 Which are the most important criteria for you in considering data for reuse?\***

Please select your **three** most important answers.

\*must provide value

- Comprehensive metadata (e.g. keywords)
- Additional materials (e.g. documentation)
- Normalized and clean data
- Availability of raw data
- Usability with familiar tools
- Ease of access
- Transparent licencing
- Trustworthiness of the source

**B.4 These other criteria are important for me in selecting data and sources.**

--



## Part C: Data needs

**C.1 Which data sources are you interested in, that do not offer reusable data yet?**

Expand

## Closing remarks

**Your room for comments and questions.**

Expand

**I am available for an interview.**

Yes

**I want to take part in the raffle for prizes.**

Yes

**Submit**

## APPENDIX II: COMPREHENSIVE LIST OF DATA SOURCES

Academia.edu	FORSbase (Democratic Governance and Citizenship Survey)	National Center of Competence in Research Democracy
Addiction Monitoring in Switzerland	FORSbase (MOSAiCH)	National Centre for Longitudinal Data (Household, Income and Labour Dynamics in Australia)
American Economic Association	FORSbase (Novizinnen und Novizen im Schreibunterricht)	National Longitudinal Study of Adolescent to Adult Health
American Economic Association (American Economic Review)	FORSbase (Optimus)	Net-Metrix (new: Mediapulse)
American National Election Studies	FORSbase (Programme for International Student Assessment)	New Testament Virtual Manuscript Room
American Visionary Art Museum	FORSbase (Swiss Election Study)	Norwegian Centre for Research Data
Anforderungsprofile.ch	FORSbase (Swiss Federal Surveys of Adolescents)	not known anymore
Answer not clear	FORSbase (Swiss Household Panel)	Notes
Archive	FORSbase (Swiss Information and Data Archive for the Social Sciences)	Open Science Framework
Archive for Spoken German	FORSbase (Swiss Job Market Monitor)	Organisation for Economic Co-operation and Development
Atlas	FORSbase (Transitions from initial education to working life)	Orsay museum database
Austrian Social Science Data Archive	FORSbase (VOTO studies (Swiss Popular Vote))	Other data
British Election Study	Frantext	Own data
Bundesamt für Sozialversicherungen	Gallica	own data (survey)

Bureau of Justice Statistics (National Crime Victimization Survey)	GDELT Project	Peace Research Institute Oslo
Cantonal Individual Tax Files	General Social Survey	Pittsburgh Youth Study
Carl Maria von Weber Gesamtausgabe	GESIS	police data
CESAR	GESIS (comperative study of electoral systems)	Publication database
Chapel Hill expert surveys	GESIS (European Values Study)	Quality of Government Data
CIOTS database	GESIS (German General Social Survey)	RAND (Indonesian Family Life Survey)
Collection of Swiss Law Sources	GESIS (German Longitudinal Election Study 2017)	Recordings
Comédie-Française Registers	Gesis (Politbarometer)	Repositories
Community Research	Getty provenance Index	Repositories (Software)
c-surf	GIS Server	Research Data Centre for Higher Education Research and Science Studies
DAB	GitHub (Swissparl)	Research Data Centre of the Socio-Economic Panel
data from colleagues	Global Burden of Disease	Research Project Lavater
Data from governmental sources (canton of Zurich)	Google	ResearchGate
Data Journal (Nature Scientific Data)	Google Scholar	SILC
Database (Literature: jstor)	Governmental data	Social Policy Indicators ( Social Assistance and Minimum Income Protection)
Database (Literature: Proquest - Social Services Abstracts)	Hamburg Center for Language Corpora	SSRN
Database (Literature: Proquest - Sociological Abstracts)	handrit.is	SSSAJ
Database (Literature: PsycInfo)	Harvard Dataverse (Quarterly Journal of Economics)	Stapfer-Enquête
Database (Literature: pubmed)	Health Behaviour in School-aged Children	State Secretariat for Economic Affairs
Database (Literature: ScienceDirect)	Health institution	Student texts/responses
Database (Literature: Scopus)	Herausfordernde Verhaltensweisen von Erwachsenen	Study on costs non-communicable diseases
Database (Literature: web of science)	HETSL (Study on family normativity)	Study on spatial planning
Database (Literature: WISO)	HETSL (Study on needs and conditions of home care)	Study plans of the Canton Ticino
Database for Spoken German	Historical Dictionary of Switzerland	Survey
Database from Museum	Historical Statistics of Switzerland	Survey of Health, Ageing and Retirement in Europe
Datenbanken	Icelandic Saga Map	Surveys on the Ticino educational system
Deutsches Text Archiv	IMPUS Current Population Survey	Swiss Multicenter Adolescent Survey on Health 2002
Diplomatic Documents of Switzerland	Income data (IK)	Swiss National Cohort
diverse literature in open access	International Inventory of Musical Sources	Swiss National Science Foundation
Diverse Quellen	International Music Score Library Project	Swiss Society for the Common Good (Swiss Volunteering Survey)
Document collection Intercantonal authority	international organisations	Swiss StudentLife Study
DVD	International Social Survey Programme	Swissvotes
e-codices	Internet Archive	TÁRKI
Education Statistics Canton of Zurich	interview of discipline expert	Théâtres de société
EFS	interview transcripts	Trends in International Mathematics and Science Study
EHA	Interviews	Turbücher des Kanton Berns
election studies	Job advertisements (online and newspaper)	Twitter
E-periodica	Journal	U.S. Bureau of Labor Statistics
ETH (Ethnic Power Relations)	Journal of Applied Econometrics Data Archive	U.S. Bureau of Labor Statistics (NLSY79)
Eurobarometer	kaggle	U.S. Bureau of Labor Statistics (NLSY79-YA)

European Election Studies	Leaders for Equality	U.S. Census Bureau
European Social Survey	Library of congress	U.S. Census Bureau (Current Population Survey)
Eurostat (European Union Labour Force Survey)	LISS Panel Data	Überprüfung des Erreichens der Grundkompetenzen
Eurostat (European Union Statistics on Income and Living Conditions)	Literalität im Alltag und Beruf	UK Data Service (British Household Panel Survey)
Fachportal Pädagogik Deutschland	Literature	UK Data Service (Millennium Cohort Study)
Federal Department of Finances FDF	Literature (Articles)	UK Data Service (UK Household Longitudinal Study)
Federal Office for Spatial Development (national passenger transport model)	Literature (Books)	underwater archeology - city of Zurich
Federal Office of Public Health	Literature (Corpora)	Universities
Federal Office of Public Health (SOMED)	Literature (divers)	Universities (University of Bern)
Federal Statistical Office	Literature (economics and statistics journals)	Universities (University of Geneva)
Federal Statistical Office (Business and Enterprise Register)	Literature (Edition)	Universities (University of Zurich)
Federal Statistical Office (Business demographics statistics)	Literature (Google books)	universities of applied sciences
Federal Statistical Office (Material deprivation)	Literature (Joyce)	universities of applied sciences (University of Applied Science and Arts Northwestern Switzerland)
Federal Statistical Office (Police crime statistics)	Literature (Kafka)	University of Peking Open Research Data (China Family Panel Studies)
Federal Statistical Office (Population and Households Statistic)	Literature (Murakam)	Varieties of Democracy
Federal Statistical Office (Social Assistance)	Literature (Musil)	various election data
Federal Statistical Office (STAT-TAB)	Literature (Nietzsche)	various image databases
Federal Statistical Office (Swiss Health Survey)	Literature (own literature)	various questionnaires
Federal Statistical Office (Graduate Survey)	Literature (review)	various quantitative surveys
Federal Statistical Office (Longitudinal Analysis in Education)	Literature (springer)	Victimization survey
Federal Statistical Office (Population data)	Manifesto Project / Comperative Manifestos Project	Videos
Federal Statistical Office (Statistics on Diploma)	Marenzio Online Digital Edition	VoxIt: Enquêtes post-votations standardisées
Federal Statistical Office (studend survey)	Master theses	Website
Federal Statistical Office (Swiss Earnings Structure Survey)	Mediapulse	WEMF AG für Werbemedienforschung
Federal Statistical Office (Swiss Labour Force Survey)	Meguid (2005)	World Bank Indicators
Federal Statistical Office (Touris accommodation statistic)	Mendeley	World Health Organization
Flickr	Menota data services	World Values Survey
FORSbase	Mexican Family Life Survey	Zurich school survey
FORSbase (Comparative Candidate Survey)	Mutual Information System on Social Protection	
FORSbase (Concon: Cohort 1)	My Personality Project	