SWITCH, Werdstrasse 2, P.O. Box, CH-8021 Zürich
www.switch.ch

SWITCH Innovation Lab “Comprehensible Data Quality“

Innovation Lab Partner
Swiss Academy for Engineering Sciences - SATW on behalf of SWITCH
Esther Koller-Meier
Manuel Kugler

Executive Summary
Up-to-date, freely accessible and usable research data is the most important resource for ensuring Switzerland's strengths in Research and Innovation and a catalyst for the development of new research findings. Comprehensible data quality would allow for broader accessing, sharing, re-using and consolidation of research data.

Hence, research data management in Open Science presents Swiss universities and research institutions with a variety of challenges, but also opens up numerous opportunities. A range of activities are already being conducted in these areas; however, the current state of knowledge and the respective expectations vary greatly from one institution to another.

The purpose of this SWITCH Innovation Lab is to document the current state of knowledge and implementation of measures used to promote broad-based comprehensible data quality in various fields of research using expert survey data collected by SATW. In addition, this study aims to identify national needs and problems arising in this context.

Some key results of this study point towards the need for accessibility of research data, automated processes and documented metadata to achieve necessary quality standards. Authenticity, integrity and indisputability are also fundamental aspects for data quality. Moreover, the results point towards the need of set guidelines and standards. All of these aspects are paramount for achieving the vision of a research data connectome for Switzerland.

Summary report

Comprehensible data quality

Esther Koller-Meier, Manuel Kugler

28.01.2020

Swiss Academy of Engineering Sciences (SATW)
St. Annagasse 18 | 8001 Zurich | +41 44 226 50 11 | info@satw.ch | www.satw.ch

1 About this report 3
1.1 Objectives of the SWITCH Innovation Lab «Comprehensible data quality» 3
1.2 Approach 3
2 Results from the written survey 4
2.1 Characteristics and importance of research data and data quality 4
2.2 Comprehensible data quality: important aspects and challenges 4
2.3 Effects of guidelines and standards on data quality 6
2.4 Publication of research data 7
2.5 Using research data from other domains for one's own research 8
3 Key points and recommendations for further steps 10
3.1 Key points 10
3.2 Recommendations for further steps 10
4 Appendix 11
4.1 Experts participating in the survey 11
5 Glossary 11

1 About this report

1.1 Objectives of the SWITCH Innovation Lab «Comprehensible data quality»

The issues surrounding research data and data management in open science present Swiss universities and research institutions with a variety of challenges, but also open up numerous opportunities. A range of parallel activities are already being conducted in these areas; however, the current state of knowledge and the respective expectations vary greatly from one institution to another.

SWITCH – which forms an integral part of the Swiss academic community – has identified 'comprehensible data quality' as a central issue when it comes to sharing, re-using and consolidating research data.

The purpose of this study is to document the current state of knowledge and implementation of measures used to promote broad-based comprehensible data quality in various fields of research. In addition, it aims to identify national needs and problems arising in this context.

1.2 Approach

To achieve the stated objectives, the Swiss Academy of Engineering Sciences SATW was commissioned to conduct an expert survey to gather the necessary information. SATW mobilised its internal and external networks of experts, including contacts through its umbrella organisation, the Swiss Academies of Arts and Sciences a+.

The first step was to identify experts from nationally relevant research fields, industry and the service sector who work in the field of 'comprehensible data quality'. Then, a questionnaire was used to gather expert knowledge, and the answers received were summarised in the present summary report.

2 Results from the written survey

2.1 Characteristics and importance of research data and data quality

According to the respondents, the **characteristics** of research data are very **heterogeneous** and **project-specific**. Data quality (DQ) is determined by the respective **use** of the data: depending on the application, there are differing demands regarding data quality. The **responsibility** for DQ lies with the researchers; they must comply with scientific standards and ethical guidelines.

The respondents consider DQ to be very important. Achieving the highest quality standards represents a major challenge and can usually only be accomplished with considerable effort. For this reason, it is common to define a number of different **quality levels**[1]. Nevertheless, ensuring the desired DQ still requires an investment of time, due to the possible need for curation, for instance.

It was mentioned that not all stakeholders always recognise the value of DQ, e.g. management. If the framework conditions relating to DQ are unclear, this can create additional work during data collection[2]. Depending on the source, the importance of DQ may also be secondary[3]. However, the general trend in research is towards more and earlier consideration of DQ[4].

2.2 Comprehensible data quality: important aspects and challenges

Various **aspects** have to be taken into account to ensure comprehensible data quality (DQ) (see Figure 1). How relevant these aspects are depends on the specific research question under consideration. For any given problem, only a selection of aspects will normally be defined as criteria. Typically, the criterion that is least fulfilled limits the data quality.

Access to data was mentioned most frequently as an important prerequisite for comprehensible DQ. Without access, all other characteristics can only be addressed theoretically. In research, however, only a fraction of the data produced is currently accessible. The reasons given are:

- **Research questions** are often so **specific** that researchers are unable to find or publish suitable data. Or there is a lack of **transparency** as to what information is available and accessible.
- In the case of publications, there are often **specifications**, from commissioning editors or publishers, for example.
- Access to data is often restricted by mandatory **registration**: For legal reasons, anyone wishing to use data must first agree to data usage agreements.
- To publish data, a lot of time needs to be spent on **documentation**.

[1] For example, both the EIDAS and the eCH standards (http://www.ech.ch/standards/48092 or https://www.ech.ch/standards/39992) distinguish between different quality levels for authentication and attribute verification. Typically, the user-friendliness for the authenticating person tends to become significantly worse at higher quality levels than at lower ones.
[2] For example, through more elaborate data curation or revised data acquisition due to altered DQ requirements.
[3] In surveys, for example, the priority is on acquiring truthful answers to questions.
[4] For example, as a result of demands from research funding organisations such as the Swiss National Science Foundation (SNF).

Authenticity, **integrity**[5] **and indisputability** are also fundamental to DQ and the reusability of data. However, these factors are often barely checked. **Processes** for ensuring DQ, such as establishing **quality rules, quality standards** and **data governance**, continuous **data monitoring** and **error analysis**, pose major challenges in data management. Existing processes need to be optimised and **automated** where possible.

The specific **challenges** relating to DQ are illustrated in Figure 2**Error! Reference source not found.**. The most frequently mentioned were **erroneous or duplicated data** and the **timeliness, consistency** and **relevance** of data. **Completeness** and **accuracy** were also mentioned almost as frequently. Reality can only be described approximately using data, which results in a lack of precision in further evaluation. This factor needs to be taken into account as early as the data collection stage. Furthermore, ensuring the timeliness of the data (which changes over time) is difficult and costly.

With certain methods such as machine learning, the quality and use of high-quality **labels** or **metadata** are of prime importance. The issue of **scalability** should also not be underestimated in this context: today, extremely large data sets are required to train models; the use of high-quality labels for such data sets is a major challenge.

Figure 1: The criteria for comprehensible data quality identified in the survey.

[5] Some experts did not clearly understand the difference between authenticity and integrity, and see authenticity as part of data integrity.

Figure 2: Challenges associated with comprehensible data quality, as identified by the experts surveyed.

2.3 Effects of guidelines and standards on data quality

In general, guidelines, standards and best practices have a positive impact on data quality. It is mainly researchers who are responsible for complying with scientific standards and (ethical) guidelines. In some cases, respondents indicated that data archivists monitor various standards in their organisations and implement them where appropriate. Some respondents stated that they were not aware of any metadata standards and that they do not apply them.

FAIR principles: Adhering to the FAIR principles (findable, accessible, interoperable and reusable) ensures that data can be re-used. Most respondents stated that they were familiar with the FAIR principles. Several research groups also said that they try to implement them as far as possible. Some institutions even offer courses, workshops and presentations for this purpose. However, in many cases the principles are not applied due to a lack of time. Another important issue in connection with FAIR is the long-term retention of research data.

Guidelines: In some cases, **domain-specific guidelines** are applied. Data administrators sometimes use **instruments and methodologies** from international projects as **guidelines**.

Standards: Depending on whether data is generated in the context of research projects, in administration or in industry, different **standards** apply, for example regarding the exchange of metadata. For surveys, comprehensive standards are set by industry associations[6]. However, universally accepted standards for data management have not yet been established. If data is collected using non-standardised procedures, it is of little use to third parties[7].

[6]These include the ESOMAR standards for data collection.
[7] Nonetheless, they need to be published to advance the standardisation process.

Official statistics follow the **European Statistics Code of Practice**[8i]. This way of organising and maintaining data and the Code of Practice itself could serve as a model for similar efforts in research.

The eCH association promotes, develops and adopts standards in the field of e-government for efficient electronic cooperation between public authorities, businesses and private individuals[9]. Where master data is concerned, corresponding best practices for data governance also exist[10]. As far as government bodies are concerned, the Federal Statistical Office's metadata and nomenclatures seem to be the most widely supported. However, international agreements do not necessarily guarantee a high level of usability.

2.4 Publication of research data

Raw data is usually only published in **anonymised form**. Complying with data protection requirements and correctly indicating sources plays an important role. Some researchers never publish raw data. In most cases, only **aggregated data** is published, in which the original information is no longer available in its full detail. The re-use of aggregated data can be challenging, as it is not always clear how the aggregation was carried out.

Increasingly, journals also require the cleaned raw data that the aggregated results are based on. A **requirement** to make raw data from research projects accessible has a generally positive effect on DQ. Researchers are then aware from the outset that others can use and possibly verify their data. This usually leads to better **documentation** and care in the handling of data and metadata. After all, nobody in research wants to be criticised for having manipulated data. Experiences in administration have also been similar[ii].

Aggregated and processed[11] data is often published, offered and made accessible via websites and public services[12], on archive platforms[13] or in relational databases, e.g. with a **CC BY licence**[14]. **Sensitive data** can be managed via access restrictions.

---

[8] The European Statistics Code of Practice sets the standard for the development, production and dissemination of European statistics. It is based on a unified ESS definition of quality in statistics and addresses all relevant areas in institutional settings, from statistical production processes to our output: European official statistics.

[9] For example, the administration uses standard eCH-0170, which is based on corresponding European and American standards.

[10] Data management processes according to Otto, Boris (HSG) 'Master data management'. An internal evaluation of two reference systems has shown that, while they are applied differently, they are applied fairly well overall. Improvements are now being undertaken.

[11] For example, curated data.

[12] e.g. UniProtKB, https://www.uniprot.org/help/uniprotkb

[13] For example, FORS publishes the anonymised raw data for all surveys on its archive platform https://forsbase.unil.ch/ together with all associated documentation.

[14] This form of licence is the freest possible and also allows the data to be used and processed commercially, distributed and expanded upon as long as the author of the original is named.

2.5 Using research data from other domains for one's own research

Data collected by researchers themselves is generally well documented. In the case of external data, however, the quality of the documentation varies considerably. The **context of the primary data collection** is important: What was the data collected for? This gives secondary data users an indication of the extent to which the data can be used and for what purpose. The context and rationale behind the data collection should be documented in metadata.

Important assessment criteria for the correctness of external data include its **origin**: the use of data always entails a degree of **trust** in its quality and therefore also in the data supplier concerned[15]. Certain data producers are trusted more – especially if they have a long track record of collecting data in a corresponding context (authenticity). The researchers' or producers' **sensitivity** to DQ is of central importance. In addition, specific **skills** are needed to critically examine sources, use and analyse data and interpret results.

Currently, in research still relatively little scientific data is being exchanged[16]: many researchers still only work with their own data or with data from companies. However, researchers often make their data available upon publication and in certain cases prepare it for interested parties. Humanities seem to be more advanced than other disciplines in networking and re-using research data[17]. This type of re-use can increase the visibility of their own research. In addition, it allows researchers to benefit from the priorities, expertise and assignments of their research colleagues, making more productive use of the data collected.

Some researchers also use public statistical data (Federal Statistical Office (FSO), cantons, etc.)[18]. Access to administrative data is now frequently possible, but sometimes complicated and time-consuming. The associated metadata can be very heterogeneous: some data sets are easy to find, well described, validated and versioned by the providers; others less so.

For certain research disciplines, the linkage[19] of their own data with administrative data or data from private and state providers is of particular interest[20]. The FSO can link a wide range of statistical information by means of identification numbers. This kind of **data linkage** makes it possible to expand information, apply new statistical analyses and thereby gain new insights from existing data. Duplicates are avoided, costs are reduced to a minimum and synergies can be harnessed. Links of this kind reduce the effort required for data collection, since fewer people need to be consulted directly. Data linkage is subject to strict rules regarding data protection and security: adhering to them is a top priority in this context. Provided that certain conditions are met, the FSO may link data without reference to individuals for research, planning and statistical

[15] In research, trust refers to the scientific honesty and diligence of colleagues and the origin of data sources.
[16] For example, the Swiss Personalized Health Network SPHN, which exchanges biomedical data, is considered a pioneer in data exchange.
[17] Dodis actively participates in Metagrid and histHub, for example.
[18] For example, federal geodata (e.g. from swisstopo, SFOE, FSO) and from cantons, natural risk or population figures for Switzerland
[19] One of the FSO's major problems is the lack of an interoperability platform to which all federal, cantonal and municipal authorities can upload their metadata describing the existing data. At the end of September 2019, the Federal Council commissioned the FSO to set up and manage an interoperability platform that will serve as a public metadata system for all federal offices and later for all public administrations in Switzerland.
[20] See project linhub.ch

purposes as part of a linking and data protection agreement. Federal, cantonal and municipal organisations as well as recognised research institutions such as universities and universities of applied sciences are authorised to do so.

Other commonly used data providers include other trustworthy institutions[21] and online sources[22]. Of particular interest to researchers is data published online[23] that is otherwise difficult or impossible to access[24]. In certain fields, access itself is not a problem, apart from occasional fee-based services. However, researchers often only find relevant publications after a long delay.

If no primary sources are used, correctly interpreting the documentation, e.g. metadata, can represent a challenge: a clear definition of the statistical population[25] is crucial. **Framework conditions** are always project-dependent and all factors that influence the interpretation of data must be recorded[26]. In addition, comprehensive documentation[27] is important.

It is preferable to use **validated data**[28]. Occasionally, **systematic reports**[29] are available for published data. **Standards**-based specifications for information suppliers are another way of ensuring quality. **Transparency** regarding the accessibility of data increases trust, since **frequent use** by as many different users as possible provides good protection: the more data and the more frequently that data is used, the higher the implicitly assumed DQ. Ideally, it should be possible to provide **feedback**. **Feedback**[30] on data instances and the **semantics** of published data help enormously to improve quality.

Due to the increasing use of open access and open data, the availability of research data has been steadily increasing for several years. Personalised private data[31], on the other hand, is still rarely available.

[21] e.g. MeteoSuisse
[22] e.g. websites, tweets, newspaper articles etc.
[23] This includes environmental data, geodata, medical data and 'open government' data.
[24] e.g. through archive visits.
[25] For example, sampling frames, sampling processes, exhaustion and other data generation parameters. For surveys, this could include the type of interviews, time of the survey, field duration, base population, recruitment, etc.
[26] Another example would be the usability of anonymised and pseudonymised data depending on its type and use. Interventions or omissions need to be described and documented in the metadata.
[27] In the case of technical measurements, examples would include the type of measuring device, the institution that performs the measurement, the person who supervises the measurements or the description of the calibration applied. For surveys this could be codebooks, questionnaires, or how the weightings are calculated.
[28] For example, one possible validation might be to use statistical methods to look at the distribution of data, timestamps or number of observations according to other criteria such as location.
[29] These describe the methods and procedures on which the surveys, results and analyses of the public data are based. See e.g. https://www.bfs.admin.ch/bfs/en/home/services/recherche/methodological-reports.html
[30] The most helpful feedback combines positive and negative points, and makes corrections where necessary.
[31] e.g. Facebook, Google

3 Key points and recommendations for further steps

3.1 Key points

- DQ is a high priority: the general trend in research is towards more and earlier consideration of DQ. DQ varies depending on the respective data use. The responsibility lies with the researchers who collect the data.
- Achieving the highest quality standards can usually only be accomplished with considerable effort. It is usually more appropriate to define different **quality levels**. Nevertheless, ensuring the desired DQ takes a lot of time. **Processes** to ensure DQ should be **automated** as far as possible.
- **Accessibility** is the most important prerequisite for transparent DQ. **Authenticity**, **integrity** and **indisputability** are also fundamental to DQ and the reusability of data. However, these factors are often barely checked. It is costly and difficult to ensure that data is **up to date**.
- **Guidelines and standards** have a positive effect on DQ. If the framework conditions are unclear, this can create additional work. No universally accepted standards for data management have yet been established. If data is collected using non-standardised procedures, it is of little use to third parties.
- The **FAIR principles** are well known, but they are still rarely applied due to time constraints.
- **Raw data** is usually published anonymously, but more often only **aggregated data** is made available. All factors that influence the interpretation of data must be documented in **metadata**. If, for example, the context of the primary survey or the aggregation is insufficiently documented, re-use is rendered difficult.
- In research, still relatively little scientific data is exchanged. Often, researchers might not find out about an online publication at all or only after a long delay. An **obligation** to make raw data accessible has a positive effect on **documentation** and therefore on DQ. Specific **skills** are needed to critically examine sources, use and analyse data and interpret results.
- **Data linkage** is of particular interest for some fields of research and makes it possible to gain new insights from existing data. Metadata relating to public statistics is heterogeneous: some data sets are easy to find, well described, validated and versioned, others less so.
- **Origin** is an important criterion for the correctness of data: data use involves **trust** in suppliers, whose **sensitivity** to DQ is essential. The more data and the more frequently that data is used, the higher the implicitly assumed DQ. The opportunity to provide **feedback** helps to increase quality.

3.2 Recommendations for further steps

- Gain an overview of existing initiatives in the field of open science and open data and make use of synergies. In particular, actively follow and help shape developments around FAIR.
- Identify and clarify the most important data repositories (research, administration and industry), determine which data is available and which metadata standards are applied.
- Improve the searchability of research data with automated tools and facilitate access (e.g. in the form of a research data connectome).

- Clarify the extent to which uniform (minimal) standards and frameworks for research data can be developed (documentation, metadata, aggregation etc.).
- Define and automate processes to ensure DQ and take the burden off researchers.
- Create awareness among researchers regarding DQ and promote open data.

4 Appendix

4.1 Experts participating in the survey

Prof.	Andreas	Spichiger	Bern University of Applied Sciences (BFH)
Mr	Bertrand	Loison	Federal Statistical Office (FSO)
Dr	Christiane	Sibille	Diplomatic Documents of Switzerland (Dodis)
Dr	Ursin	Lutz	Dicziunari Rumantsch Grischun (DRG)
Prof.	Georg	Lutz	FORS
Mr	Adrian	Meyer	Mobiliar
Mr	Heinz	Stockinger	Swiss Institute of Bioinformatics (SIB)
Mr	Mario	Valle	Swiss National Supercomputing Centre (CSCS)
Prof.	Philippe	Cudre-Mauroux	University of Fribourg (UNI FR)
Dr	Markus	Christen	University of Zurich (UZH)
Mr	Andreas	Fürholz	Zurich University of Applied Sciences (ZHAW)
Dr	René	Locher	Zurich University of Applied Sciences (ZHAW)

5 Glossary

Aggregated data	Aggregated data is essentially individual values combined into larger units.
Authenticity	Authenticity refers to the trueness of the data in the sense of 'found to be original'.
Best practice	The term 'best practice' refers to a tried-and-tested method of carrying out a work process. It is a technique or methodology that has been proven through experience and research to be reliable in achieving a desired result.
Data format	The data format determines how data is structured and presented and how it is to be interpreted during processing. It therefore specifies the syntax and semantics of data within a file.
Data type	The data type describes the type of data and which logical operations can be performed with it.
Data curation	The term 'curation' is used in the sense of preserving and handling. Data curation therefore describes the management activities required to maintain data in the long term so that it is available for re-use.

Data linkage	Data can be linked by the use of identification numbers in different data sets. The purpose of data links is to obtain information from existing data, avoid duplication, minimise costs and achieve synergies Safeguarding data protection is given the highest priority. For this reason, data linkage is subject to strict conditions with regard to data protection and data security
FAIR principles	The term FAIR (findable, accessible, interoperable and reusable) data was coined in 2016 by the FORCE 11 community for sustainable research data management. The main objective of the FAIR principles is to ensure the optimal processing of research data so that it is findable, accessible, interoperable and reusable.
Integrity	Integrity refers to the correctness or intactness of data, i.e. it must not be possible to make any undetected or unnoticed changes to data.
Metadata	Metadata is structured data that contains information about characteristics of other data. The data described by metadata often consists of larger data collections such as documents or files.
Open access	Open access is concerned with making access to scientific literature and other materials (including primary and metadata) freely available online.
Open data	Open data is data that can be used and distributed by anyone without restriction.
Guidelines	Guidelines are intended to provide all employees of an institution with information on which data management procedures should be used and how data should be handled.
Primary data	Primary data (also called raw data or original data) is data obtained directly from data collection.
Raw data	Raw data (also called primary data or original data) is data obtained directly from data collection.
Semantics	Semantics is concerned with the meaning of signs and sign sequences.
Standard	A standard is a comparatively uniform or unified, widely accepted and generally commonly used way of describing something.
Syntax	Syntax refers to a system of rules for combining elementary characters into compound characters in natural or artificial sign systems. When referring to languages, it describes the usual connection of words to word groups and sentences or the correct linking of linguistic units in a sentence.
Original data	Original data (also called primary data or raw data) is data obtained directly from data collection.
Accessibility	Data and metadata should be archived and made available on a long-term basis so that it can be easily downloaded and used by humans and machines.

[i] The European Statistics Code of Practice contains 16 basic principles for the production and dissemination of European official statistics and the institutional environment in which the national and community statistical authorities operate. A series of good practice indicators for each of the 16 basic principles provides guidance for monitoring the code's implementation.
The European Statistics Code of Practice was adopted by the Statistical Programme Committee on 24 February 2005 and revised by the European Statistical System Committee in September 2011 and November 2017. The European Statistical System Committee adopted the Quality Assurance Framework along with the 2011 version of the Code of Practice. It serves as a guide for the implementation of the European Statistics Code of Practice.
Eurostat has adopted a 'Protocol on impartial access to Eurostat data' to support the implementation of the code. It is aimed at Eurostat users, staff and partners in the production of European statistics. Each year, Eurostat monitors compliance with the Code of Practice throughout the ESS. The European Statistical Governance Advisory Board (ESGAB) receives summarised information for its annual report to the European Parliament and the Council. The ESGAB reports on the implementation of the Code of Practice insofar as it relates to Eurostat and includes an assessment of the implementation of the Code of Practice throughout the ESS. The ESGAB annual report, which will be published from 2009 onwards, is available on a dedicated website.

Every five to six years, the principles described in the above-mentioned Code of Practice are reviewed by international experts selected by Eurostat.

[ii] For company data, the Federal Act on the Unique Business Identification Number (UID Act) provides a gradation in the importance of the data sources. For example, data in the commercial register has the highest priority. At the time of the introduction of the UID register, it was known that the register with the highest priority does not necessarily guarantee the highest quality of data, so a second (non-public) 'last known address' field was introduced for the (public) 'address' field so that the most up-to-date information is available in case of problems. The public nature of the UID register means that this field is no longer needed, because the pressure to report corrections has increased accordingly.