Betancort Cabrera, Noemi et al.

Working Paper

White Paper on implementing the FAIR principles for data in the social, behavioural, and economic sciences

RatSWD Working Paper, No. 274

Provided in Cooperation with:
German Data Forum (RatSWD)

Suggested Citation: Betancort Cabrera, Noemi et al. (2020) : White Paper on implementing the FAIR principles for data in the social, behavioural, and economic sciences, RatSWD Working Paper, No. 274, Rat für Sozial- und Wirtschaftsdaten (RatSWD), Berlin, https://doi.org/10.17620/02671.60

This Version is available at:
https://hdl.handle.net/10419/229719

RatSWD Working Paper Series

www.ratswd.de

RatSWD.
German Data Forum

274

White Paper on implementing the FAIR principles for data in the Social, Behavioural, and Economic Sciences

Authored by members of the Economic and Social Sciences goINg FAIR Implementation Network (EcoSoc-IN) (in alphabetic order):
Noemi Betancort Cabrera, Elke C. Bongartz, Nora Dörrenbächer, Jan Goebel, Harald Kaluza, Pascal Siegers

December 2020

Working Paper Series of the German Data Forum (RatSWD)

The *RatSWD Working Paper Series* was launched at the end of 2007. The online series is exclusively publishing conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that appear in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the German Data Forum (RatSWD). Papers, addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, or locally specialized journals.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the German Data Forum (RatSWD). Its funding organisations have not influenced the publications.

White Paper on implementing the FAIR principles for data in the Social, Behavioural, and Economic Sciences

Authored by members of the Economic and Social Sciences goINg FAIR Implementation Network (EcoSoc-IN) (in alphabetic order):
Noemi Betancort Cabrera[1], Elke C. Bongartz[2], Nora Dörrenbächer[3], Jan Goebel[4], Harald Kaluza[5], Pascal Siegers[6]

doi: 10.17620/02671.60

[1] https://orcid.org/0000-0002-0156-3556 (last accessed: 18.11.2020).
[2] https://orcid.org/0000-0002-2136-7675 (last accessed: 18.11.2020).
[3] https://orcid.org/0000-0002-6246-1051 (last accessed: 18.11.2020).
[4] https://orcid.org/0000-0002-3243-1935 (last accessed: 18.11.2020).
[5] https://orcid.org/0000-0001-7045-3193 (last accessed: 18.11.2020).
[6] https://orcid.org/0000-0001-7899-6045 (last accessed: 18.11.2020).

Introduction

The FAIR principles formulate guidelines for the sustainable reusability of research data. FAIR stands for Findability, Accessibility, Interoperability, and Reusability of data and metadata. As Mons et al. (2020) highlight, the FAIR principles are not a new standard. Moreover, they are neither a top-down requirement nor an all-or-nothing binary state (FAIR or not FAIR) (ibid). Instead, the FAIR principles are more of a continuum, from being less FAIR to more FAIR. They are a guide and a *resource for optimal choices* to be made for data management and tool development. The application of the principles ensures interdisciplinary and international access to data and its use. The principles emerged in January 2014 at the Lorentz workshop in Leiden (NL). Following this workshop, the principles and their 15 sub-principles were posted at the FORCE11 website[7] for community comments and subsequently published in *Scientific Data* (Wilkinson et al. 2016).

Ever since their publication, the principles have sparked converging initiatives. These include the GO FAIR[8] initiative and other international and interdisciplinary efforts, such as the Commission on Data of the International Science Council (CODATA) or working groups at the Research Data Alliance (RDA). These mostly bottom-up, stakeholder-driven and self-governed initiatives embark on the challenge of guiding different domain communities toward more mature FAIR choices and infrastructures. Different implementation guidelines emerged in this process such as "the FAIR metrics" (Wilkinson et al. 2017), the "FAIR Data Maturity Model" (RDA FAIR Data Maturity Model Working Group 2020), or the "Top 10 FAIR Data & Software Things"(Martinez et al. 2019)[9] (see also Jacobsen et al. 2020). Overall, the FAIR principles are becoming an international standard and adherence is often also required from research funding organisations.

However, while there is a growing body of general implementation guidelines, so far there is a lack of specific recommendations on how to apply the FAIR principles to the specific needs of social, behavioural and economic (SBE) science data. This field faces some specific challenges. First, SBE scientists work with highly diverse data types. These include qualitative and quantitative data from interviews, process-generated data from public and private institutions (e. g., official registries, business corporations), experimental and (panel) survey data, behavioural data (e. g. derived from sensors and wearables), texts such as official documents, journal articles, qualitative interviews, etc. Moreover, much SBE data are potentially disclosive, containing confidential information on individuals, companies, or institutions. Therefore, the legal provisions on data protection and the assurance of confidentiality must be observed during data processing. Before sharing, disclosive data needs to be anonymised by coarsening or removing disclosive attributes from the data. Access to sensitive data needs appropriate organizational and technical protection.

These features of social science data pose some challenges to the useful implementation of the FAIR principles – especially regarding the machine-actionability of data and metadata that is at the core of the FAIR principles. Therefore, in November 2018, the Economic and Social Sciences goINg FAIR Implementation Network (EcoSoc-IN)[10] was founded. EcoSoc-IN is the first GO FAIR implementation network for the SBE sciences. The creation of the

[7] https://www.force11.org/fairprinciples (last accessed: 18.11.2020).
[8] https://www.go-fair.org/ (last accessed: 18.11.2020).
[9] see European Commission Directorate-General for Research and Innovation (2020).
[10] https://www.go-fair.org/implementation-networks/overview/ecosoc-in/ (last accessed: 18.11.2020).

implementation network was initiated by the German Data Forum (RatSWD)[11] and is based on its long-standing expertise, as well as the 38 Research Data Centres (RDC) that it has accredited so far (see Linne et al. 2021, in press).[12] In 2017, the FAIR principles were first compared with the RDCs' already established RDM. This evaluation revealed that many aspects are already being implemented in daily practice by a large part of the RDCs, especially with regard to the findability (e. g., data catalogues), accessibility and reusability (e. g., high documentation standards) of research data. With regard to the standardisation of metadata practices, the improvement of data interoperability and machine-actionability, however, there is still a need for action that also has to meet the different disciplinary requirements.

Filling these gaps is particularly important in light of the recently established national research data infrastructure (NFDI)[13] in Germany. The NFDI is a science-driven process of bringing together consortia acting on their own initiative to form a networked structure. In this structure, systematic indexing, sustainable safeguarding, and availability, as well as networking data in science and research (inter)nationally, are objectives of high interests. Within the NFDI, the Consortium for the Social, Behavioural, Educational, and Economic Sciences (KonsortSWD) started its activities in October 2020. KonsortSWD builds directly on the important work done by the German Data Forum (RatSWD) and is committed to the FAIR principles. KonsortSWD will therefore contribute to the implementation of FAIR standards for data management and access from within EcoSoc-IN. We intend to warrant harmonised and easy access to research data based on cross-disciplinary consensus about standards for research data management, also with regard to technological solutions.

The primary goal of this White Paper is the further dissemination and development of the FAIR principles, so that they can be used within the RDCs, KonsortSWD and beyond. This includes a vision for a FAIR compliant data infrastructure with recommendations for the implementation of FAIR data and services in the SBE sciences and respective data centres. In the following, for each of the 15 FAIR (sub)principles, this White Paper proposes minimum requirements and provides a vision for a full-implementation of the FAIR principles by repositories and data centres. For this purpose, the abstract principles are formulated for the SBE sciences and the peculiarities of the data and object types in these disciplines are taken up. Consequently, the White Paper focuses less on the FAIR principles, which are generically technical (i.e. use of persistent identifiers (PIDs) or interfaces for metadata, etc.). Instead it discusses in more depth those sub-principles which need to be filled with domain specific content of the SBE sciences. Particularly relevant are the conventions of appropriate access rules in the case of data that are not completely anonymised/anonymisable, as well as questions of metadata quality to facilitate reuse and to improve interoperability and machine-actionability of (meta)data.

[11] https://www.konsortswd.de/ratswd/ (last accessed: 18.11.2020).
[12] The RDCs guarantee at least one access path for the reuse of the research data they curate, such as sensitive research data from official statistics and official registers, social security institutions, departmental research institutions and scientific research institutes, which in particular hold the data of the large survey studies. The RDCs carry out the complex weighing process to find the balance between anonymisation and research potential, in line with the "intelligent openness" concept, endeavour to make access as open as possible, but as secure as necessary.
[13] https://www.forschungsdaten.org/index.php/Nationale_Forschungsdateninfrastruktur_-_NFDI (last accessed: 18.11.2020).

Findability

Data and metadata not only have to be readable by humans. Given the trend towards digital sciences it is even more important that they are machine-actionable. Therefore, data and metadata should be easy to find by both humans and computer systems. Much data in the SBE sciences are sensitive (e. g. disclosive or confidential) and therefore often not available open access. This makes findability of data even harder than in other disciplines. Hence, it is important to make sensitive data findable at least in a sense of being able to detect it. Here, findable metadata play a key role.

Metadata should be designed in such a way that they can be made open access even if they refer to sensitive data. To enhance findability for humans and machines, PIDs are a cornerstone of the FAIR principles. Four aspects are to be kept in mind around Findability:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

Established PID providers emerged during the long history and development of data infrastructures in SBE sciences (see also Juty et al. 2020). A large majority of data curated by RDCs and data archives are referenced by a PID (most often a Digital Object Identifier (DOI) from DataCite). The utility of PIDs is particularly high, if they are integrated into a metadata scheme facilitating the exchange of data based on the PIDs. In the context of the European Open Science Cloud (EOSC)[14], a Persistent Identifier Policy was recently published (European Commission Directorate-General for Research and Innovation 2020). It defines a set of expectations about which PIDs will be used in support of a functioning environment of FAIR research (Hellström et al. 2020).

Generally, the automated exchange and reuse of metadata could be facilitated even more by increasing the granularity of the data elements referenced by a PID. If each significant element of a study (e. g., an attribute or fragment of the data file) receives a PID, citing and reusing becomes even more targeted.

Recommendation F1:

- Minimum implementation: For each study (i. e., an entity of data collected within a common methodological frame), both data and metadata should be referenced by a common PID. The PID service that is used should meet the criteria of the EOSC PID[15] policy and domain specific PID policies[16].
- Maximum implementation: Each significant element of a study together with its corresponding metadata is referenced by a PID. The definition of significant elements depends on the data type. Examples for significant elements are variables/attributes in tabular data of quantitative research designs or fragments of recordings in qualitative research designs.

We strive for a free PID service for all European Research Data Infrastructures within the EOSC that can be combined with domain specific metadata standards. A technical

[14] https://ec.europa.eu/digital-single-market/en/european-open-science-cloud (last accessed: 18.11.2020).
[15] For current policies see: https://www.eoscsecretariat.eu/eosc-symposium2019/pid-policies (last accessed: 18.11.2020) and Hellström et al. (2020); see also the efforts under the FREYA_project to establish a PID infrastructure: https://www.project-freya.eu/en/about/mission (last accessed: 18.11.2020).
[16] For current policies see: Hausstein and Horton (2020).

harmonisation of PID services would contribute to the factual integration of research data infrastructures.

F2. data are described with rich metadata.

To make data findable the second sub principle refers specifically to rich metadata. Particularly to increase machine-actionability, it is important that metadata follow general standards. So-called Metadata for Machines (M4M) workshops[17] may be useful to generate such metadata. The M4M workshops are agile, hackathon events where experts on metadata of specific domains and technical experts develop together solutions and template for metadata that comply with the FAIR principles.
The Data Documentation Initiative (DDI) is a metadata standard, which has been developed for the documentation and sharing of quantitative data in the social sciences. This standard has been continuously extended to include other data types and to optimise the exchange of statistical data. It should be the common standard for Research Data Management (RDM) tools in the field of social sciences. Beyond DDI and in order to make metadata findable across disciplinary boundaries and the web, metadata should also rely on more generic web vocabularies, such as schema.org.[18] which optimises particularly findability on the web.

In order to enhance findability, it is also recommended that communities of practice declare their FAIR implementation practices in the so-called FAIR Convergence Matrix[19], an effort by the GO FAIR community to accelerate broad community convergence on FAIR implementation options (Pergl Sustkova et al. 2020). It makes the choices and challenges of implementing the FAIR principles findable for other communities and thus is a FAIR resource itself that is openly available.[20] While FAIR implementation guidelines such as this White Paper provide a resource that is readable by humans, the FAIR Convergence Matrix also provides machine-actionable and findable solutions. The matrix allows to identify FAIR Implementation Profiles (FIPs) of use and reuse of existing resources. It supports in this way the transparent derivation of strategies that optimally coordinate convergence on standards and technologies in the emerging Internet of FAIR Data and Services (ibid p. 159). Beyond the discipline specific DDI or general standards such as Schema.org, relevant cross disciplinary standards identified by the matrix so far are Dublin Core, DataCite or DCAT. These standards would in principle also be conceivable or desired standards in the SBE sciences as they provide metadata core elements.

[17] https://www.go-fair.org/resources/go-fair-workshop-series/metadata-for-machines-workshops/ (last accessed: 18.11.2020).
[18] Founded by Google, Microsoft, Yahoo and Yandex, Schema.org vocabularies are developed by an open community (https://www.w3.org/community/schemaorg (last accessed: 18.11.2020)) process, using the public-schemaorg@w3.org mailing list and through GitHub).
[19] https://docs.google.com/spreadsheets/d/1MUZn7uh4x5YLPjqxi-V8XubsSEEonQWvx2jBlcyyNdU/edit#gid=0 (last accessed: 18.11.2020).
[20] The Convergence Matrix is typically filled in by a Community Data Steward via a Questionnaire to list the implementation choices that correspond to each of the FAIR Principles. These choices are made on the basis of Considerations that involve numerous community-specific factors including the FAIR Requirements and various sources of Constraints endemic to the Community. These data are automatically captured in a FAIR, open and machine-actionable format (Schultes 2019, pre-print).

Recommendation F2:

- Minimum implementation: Each study should be accompanied with descriptive metadata. For findability within the discipline, the metadata should be standardised using the DDI metadata scheme or another standard that can be mapped to DDI.
- Maximum implementation: Each significant element of a study is accompanied by standardised metadata. Metadata at the study level and significant element level are disseminated through application programming interfaces. DDI is the preferred standard for metadata provision. For interdisciplinary findability the metadata should be standardised according to the most relevant general cross-disciplinary standards such as schema.org and/or relevant standards identified by the FAIR Convergence Matrix.

F3. Metadata clearly and explicitly include the identifier of the data they describe

Often data and metadata are updated continuously. In order to establish unambiguous links between the different versions of the data file and its corresponding metadata, PIDs should be included in both data and the corresponding metadata files, be it at the study level or at the level of significant elements. A technology that provides this link is FAIR Data Point (FDP)[21].·FDP is a FAIR metadata publication platform via an API. It is based on the Data Catalogue model (DCAT). FDP provide unique identifiers for potentially multiple layers of metadata (Jacobsen et al. 2020). FAIR Data Points also provide "a single, predictable, and searchable path through these layers of descriptors, down to the data object itself" (ibid p. 17). The approach of starting with a M4M workshops that lead to a metadata template which then becomes part of a FIP (see above F1) and is then implemented in an FDP, is also called the Three-Point Framework for FAIRification[22]. In the FAIR community, this framework is aimed at providing a general approach for a broad spectrum of stakeholders to demonstrate what it means to "go FAIR" in practice.

Recommendation F3:

- Minimum implementation: PIDs should be included in the data file **and** the metadata including documentation files (e. g., PDF files containing questionnaires, codebooks, or methods reports). This requires updated documentation in the metadata, if data files were modified.
- Maximum implementation: A standardised procedure using PIDs is used to link metadata with the data. Technically, a FDP could be used or the DDI element "Citation" could be used to refer to each significant element by including its PID in the metadata. Moreover, the PIDs of the significant elements are included within the data file (e. g., as a characteristic of the variable in the data file).

F4. (meta)data are registered or indexed in a searchable resource.

Finally, to make data and metadata findable, it needs to be findable by search engines. Thus, it needs to be registered or indexed in such a way that search engines can harvest the relevant information. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[23] is

[21] https://github.com/FAIRDataTeam/FAIRDataPoint (last accessed: 18.11.2020).
[22] https://www.go-fair.org/how-to-go-fair/ (last accessed: 18.11.2020).
[23] https://www.openarchives.org/pmh/ (last accessed: 18.11.2020).

a low-barrier mechanism for this purpose. Repositories that expose structured metadata via OAI-PMH allow service providers such as search engines to harvest that metadata.

Recommendation F4:

- Minimum implementation: Study level metadata should be provided for data discovery via an OAI-PMH or free exchange interface.
- Maximum implementation: DDI compatible metadata are provided for the significant elements of the data (e. g., at the measurement level), thus allowing data discovery tools to be extended to these significant elements. Discovery tools provide web-standard annotation of metadata (e. g., Schema.org) for domain specific data structures and higher visibility in general purpose search engines.

Accessibility

In the SBE sciences much data is sensitive in nature because it is collected from individuals, companies, and organisations (e. g., statistical offices, social insurances). As a consequence, this data may not always be openly available for reuse. According to the Open Research Data Pilot, research data should follow the principle "as open as possible, as closed as necessary" (European Commission Directorate-General for Research & Innovation 2016). This means that there is always a trade-off between maximising the research potential of data and protecting the rights of individuals (personal data) and organisations (confidential data) participating in data collection (see figure 1).

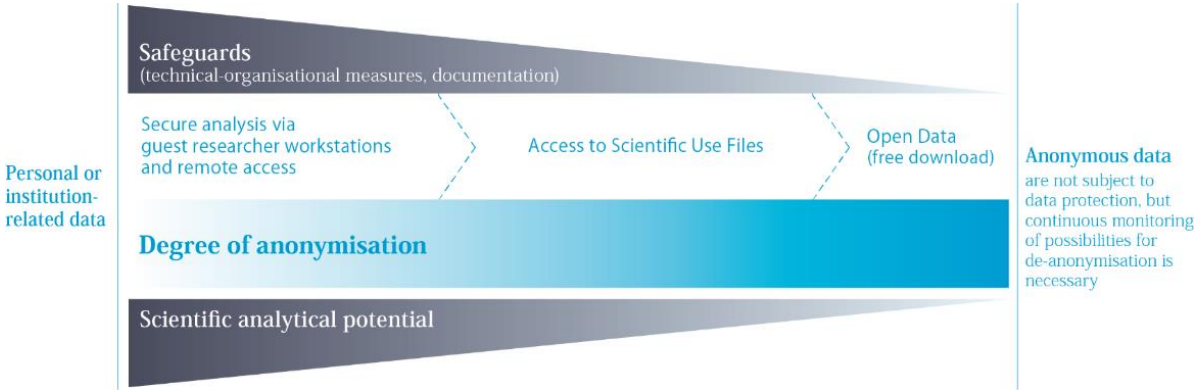

Fig. 1: Levels of anonymisation (RatSWD 2018, p.8)

The Requirements of privacy protection legislations, confidentiality, copyright, licences, or contracts between data infrastructures and data users must be observed when data are shared. The contracts are part of the organisational measures required for compliance with the General Data Protection Regulation (GDPR).[24]

Beyond such legal requirements also ethical requirements need to be considered when making research data accessible. Recently, ethical considerations in data generation have gained importance and are at least well recognised or even highly recommended (Landi et al. 2020). The approach of the CARE principles for Indigenous Data Governance is particularly

[24] European Parliament and Council of the European Union (2016): Regulation (EU) 2016/679 of The European Parliament and of The Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, Volume 59, 4 May 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC (last accessed: 13.11.2020).

noteworthy in the area of data ethics (RDA International Indigenous Data Sovereignty Interest Group 2019). CARE is an acronym for Collective Benefit, Authority for Control, Responsibility and Ethics. According to GIDA – Global Indigenous Data Alliance, the CARE Principles are complementary to the FAIR principles. For SBE sciences this addition and specification of principles may be very valuable. More generally, ethical requirements include considerations concerning the potential harm of study participation and the (formal) organisation of informed consent of study participants considering also the validity of the consents and possibly renewals of consent (see also RatSWD 2017; RatSWD 2020, 21 et seqq.).

Granting access for analysis to the data requires the development of a data protection concept, including technical and organisational measures to prevent the illegal use of data. Many RDCs in Germany and international data archives set up guest researcher work stations for access to confidential data. They provide a secure environment for data access and analysis (see, e. g., RatSWD 2018). Other technical protection measures comprise the use of passwords to avoid unauthorised use of data processing equipment, the use of protected server areas or the encryption of data, destruction of obsolete data, and deletion periods in the course of the project.

In the course of risk assessment, the probability of risk occurrence and the consequences for disclosing the data of study participants have to be addressed. If intellectual property is affected, management of copyrights and exploitation rights of third parties is necessary (see also section on "Reusability" below).

Relying on the sub-principles on accessibility the section below discusses solutions to make sensitive SBE data accessible for analysis while doing justice to data protection and ethical considerations.

A1 (meta)data are retrievable by their identifier using a standardised communications protocol.

The retrieval of data should be possible without specialised and proprietary tools or communication methods. Concerning the communication protocol, access barriers have to be avoided. Data users should get reliable and secure access to digital resources. In the case of fully machine-actionable access protocols, it is of substantial importance how data and metadata can be retrieved using their identifiers and loaded into a web browser. If open access publication is permitted, data, details to data (conditions of accessibility, PID), documentation, and (open) software code should be made accessible via a certified repository that is committed to open access. Citing data is also an important component to facilitate access by humans and machines to the data itself and to associated metadata, documentation, code, and other related materials for reasons of informed use of referenced data.

Depending on the sensitivity of the data there are currently different anonymisation levels and different access levels offered by RDC and social science data archives: namely on-site or off-site (Scientific Use Files (SUF), Public Use Files (PUF), Campus Use Files (CUF)). Additionally, remote access options are possible (see RatSWD 2019). For all access levels well-defined access conditions need to apply. That means, roles, rights, and responsibilities of all agents involved in data processing and the data access process have to be documented and communicated transparently. Registration of users is recommended, if data use is restricted to scientific purposes. A registration requires that data users acknowledge the terms of use

prohibiting them to perform operations on the data, potentially deanonymising study participants.

Generally, automated data access using technical interfaces are currently only possible for anonymised SUF with low usage restriction. For sensitive data with higher usage restriction, non-mechanised access protocols, like personal requests and usage contracts are currently needed. When such non-mechanised access protocols are applied they should be compatible with the FAIR principles. This means for example, that contact information of the person or institution responsible for the data collection (e. g. via email or telephone number) is provided and included in the metadata (see Jacobsen et al. 2020).

Also, for sensitive data with high usage restrictions, automated access protocols are currently developed and may gain importance in the context of newly emerging data types. Increasingly in the field of administration, consumer behaviour, and social media new data types are produced that are characterised for example by linkages with several data sources, high granularity, or large quantity of unstructured information (big data). These micro data types require special treatment concerning data protection, because they are often private and confidential in nature. In this context, the Annodata framework (Bender et al. 2019) as a standard for transparent technical and legal data access and data sharing focusing on machine-actionability aims at filling a gap in existing metadata standards. At the same time, Annodata contributes to implementing the FAIR principles in practice. By now, a Python-programmed contract generator with free available code was developed. In combination with the contract generator the Annodata scheme facilitates access to different data from different service providers in different countries. The International Network for Exchanging Experience on Statistical Handling of Granular Data (INEXDA) – an international cooperative project of central banks – is based on Annodata and has established several working groups.[25] The objective is, inter alia, to promote harmonisation activities among INEXDA members, expansion of common metadata sharing, and possible data exchange between involved network institutions in the future. Furthermore, the intention is to lay the foundations of exchange of metadata on publicly available granular datasets with external researchers. The INEXDA data access working group develops solutions in the field of unified data access procedures, best practices on granting access to data, and open software solutions. Developments within INEXDA should be followed further and possible application within the RDCs could be considered.

The considerations concerning standardised communications protocols for sensitive data in the SBE sciences as discussed above lead to the following specific recommendations concerning access protocols:

A1.1 the protocol is open, free, and universally implementable.
To exploit the potential of data reuse, the protocol should be offered at no-cost (free), be open-sourced and globally implementable, this is to say that technology is available on a non-discriminatory basis for all data users. Modern communication protocols like http(s) and ftp or smtp are well-defined and good examples (Jacobsen et al. 2020).

[25] https://www.inexda.org/activities/ (last accessed: 18.11.2020).

Recommendation A1.1:
For metadata:

- Minimum and maximum implementation: Metadata define accessibility conditions, which should after a check be modifiable upon request. Metadata should be accessible without restrictions. Metadata should be designed in such a way that no access restrictions are needed. If not the data, then at least the metadata should be accessible by computer and internet connection and, for example, via free universal protocols for information retrieval like OAI-PMH, REST API or network protocols that are common in library such as Z.39.50 and its successor SRU (Search / Retrieve via URL).

For data:

- Minimum implementation: Data owners or publishers should define conditions for data access, rules for data use, and data sharing. The rules specify how researchers can access data and who decides about data requests. If applicable, the rules inform about restrictions regarding (1) criteria for exclusion from data access, (2) authorised research purposes, and (3) authorised research operations. The rules should be publicly available, transparent, and universally applicable as part of the metadata.
- Maximum implementation: A standardised access protocol is developed to harmonise all steps of data access. This requires, in advance, the development of a responsible, controlled access model or access policy with specification of all agents involved in the process (Landi et al. 2020). For process efficiency reasons, human and machine-readable formats of ethical and regulatory and legal requirements are highly recommended. Not only access to data but already allowing or denying access to data is fully automated. Developments such as Annodata within INEXDA should be followed further and possible application could be considered.

A1.2 the protocol allows for an authentication and authorisation procedure, where necessary.

The principle of accessibility does not imply free or open access to research data but the provision of exact, transparent, and well-defined information on access conditions. Even sensitive data can be FAIR. Accessibility should be specified in a way that a computer system can understand the requirements, execute the requirements automatically, or, at least, alert the user to the requirements. The https protocol is a well-known and compliant protocol for machine-actionable data access (Jacobsen et al. 2020).

Recommendation A1.2:
For user identification, we recommend building on a networked infrastructure for authentication and authorisation services.

- Minimum implementation: Multi-sign-on-solutions should continue to exist. This applies particularly as long as participation in a networked infrastructure does not yet exist. Moreover, multi-sign-on solutions should continue to exist as a permanent, additional alternative. Controlled access to the data download can be achieved by using an account managed by the data publisher with their own user ID, password, and time restriction. Data users can manage log-in data with the help of an individual password manager. Finally, authentication and authorisation by telephone is also compatible with this FAIR sub-principle.

- Maximum implementation: For user identification, we recommend the use of the pan-European Authentication and Authorisation Infrastructure (AAI) integrated into the

EOSC or national solutions as part of the NFDI. AAI's systems control access of data users, manage access rights, and offer different access levels. This is done in accordance with privacy and data protection legislation. In the context of the ELIXIR AAI[26], identification does not require a password. An identification of data users is simplified by a single sign-on procedure using existing external accounts from external authentication providers. These may be universities or research institutions meeting special Identity Provider server requirements, community (e. g. ORCID[27] for PIDs for individual researchers, or eduroam that authenticated users in research, higher education and further education to provide them with international roaming) or commercial identities (e. g. Google, LinkedIn).

As for PID services, common AAI services are a cornerstone for factual integration of distributed research infrastructures. AAI services should include information about research domains and status of users (e. g. student, faculty member) to facilitate management of access rights to data.

A2 metadata are accessible, even when the data are no longer available.
It might occur that data is lost. Reasons for data loss include excessive cost of storage, insolvency or liquidation of companies or dissolution of institutions. Particularly in the SBE sciences data may also be removed as a consequence of data protection requirements. Also, loss by accident, technical failure or criminal undertaking can occur. In such cases at least the metadata should remain available and inform about reasons for unavailability of data. Even if the original data are no longer available, metadata can be highly useful in the research process because in this way the scientific community knows at least what has already been researched, even if details are no longer known. Also, removal of metadata would itself contradict the Open Science criteria or the FAIR principles as they apply at least to the metadata. Moreover, if the data were once available, they may have been cited. If the metadata were also removed, the research work related to them would no longer be comprehensible. The researcher would possibly run the risk of having worked dishonestly or used ghost data. Last but not least, the data could perhaps be made accessible again at a later date (reconstruction, clarification of rights, etc.). An assignment to existing, preserved metadata can thus, be made and prevents that metadata have to be created anew or made accessible again.

Recommendation A2:

- Minimal implementation: Repositories should provide a concept for digital long-term preservation for data and its metadata. FAIR data providers certify trustworthiness of their repository for long-term digital preservation of research data and metadata. Data repositories should aim for FAIR-aligned certification (e. g. CoreTrustSeal[28]) (Koers et al. 2020; Mokrane and Recker 2019). Even if the data is lost or may no longer be made available, the metadata should remain available and inform about reasons for unavailability of data.
- Maximum implementation: The loss of data and metadata is generally prevented. For this purpose, we recommend reciprocal agreements between data infrastructures about transferring data and metadata files in case of functional failure or outage of a service provider. Large scale research infrastructures like EOSC and NFDI should require data preservation policies, a concept for digital long-term preservation, and a certificate of

[26] ELIXIR AAI went productive in late 2016. Available at: https://elixir-europe.org/services/compute/aai (last accessed: 18.11.2020). See also Linden et al. (2018).
[27] https://orcid.org/ (last accessed: 18.11.2020).
[28] https://www.coretrustseal.org/ (last accessed: 18.11.2020).

trustworthiness and FAIR-compliance from data provider. Taking this a step further, these large-scale research infrastructures could also strive to harmonise the services or even offer such services themselves.

Interoperability

Interoperability concerns the issue of identifying the content of data correctly, which opens up for linking data across different data sources. The objective is to enhance the analytic value of existing data by creating new linked data sets. Interoperability has several dimensions ranging from technical and syntactical to semantic interoperability. All dimensions are highly relevant for all disciplines particularly in terms of machine-actionability. However, the following section focuses particularly on semantic interoperability. This choice is based on the fact that semantic interoperability poses some discipline-specific challenges to the SBE sciences.

Whether or not the term is new, the concept of interoperability has had a long history in social sciences. Measurements of core concepts like educational achievement differ across regional and national contexts. Therefore, the social sciences developed specialised ontologies for the harmonisation of measurements like the International Standard Classification of Education (ISCED) (UNESCO Institute for Statistics 2012) or the International Standard Classification of Occupation (ISCO[29]). Moreover, DDI contains many controlled vocabularies for data documentation covering for example data collection modes.

Currently, these resources are not provided in a FAIR format which complicates their use in data documentation and – even more importantly – the content description are often not provided in a machine-actionable format. Moreover, for many concepts used in social sciences no formal description is available at all that could be used for documentation by researchers and research infrastructures.

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

Rich metadata must contain a description of the contents included in the data. This description must be provided in a way (1) that a generic user can understand what information the attribute in the data contains and (2) that is standardised across data sets (Jacobsen et al. 2020). For example, a measurement of personality traits should always be documented using the same vocabulary. Only then, similar or identical contents can be found and linked across studies. Currently, an encompassing ontology of concepts and vocabularies for their description is not available.

Existing thesauri for social sciences and adjacent disciplines were developed for indexing publications and are not suited for data (Friedrich and Siegers 2016). Only the CESSDA Topic Classification and the European Language Social Science Thesaurus (ELSST)[30] were systematically developed for (multilingual) data indexing. However, they are not yet systematically used for data documentation.

[29] https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=CL_ISCO08&StrLanguageCode=DE&IntPcKey=&StrLayoutCode=HIERARCHIC&language=DE (last accessed: 18.11.2020).

[30] https://elsst.ukdataservice.ac.uk/ (last accessed: 18.11.2020).

The existing substantive vocabularies (for example ISCED and ISCO) are frequently used especially in survey research and should be used systematically when applicable. When used appropriate references to the documentation vocabularies should be given (see recommendation for minimal implementation). To facilitate the use of standardised resources in data documentation, research infrastructures have to provide technologies for managing vocabularies and thesauri including references for the automatic retrieval of metadata.

These considerations refer primarily to domain specific resources for documentation. The FAIR principles require a general representation of data contents independent of domain specific standards. The development of interoperability technologies for documentation resources should include perspectives of a general knowledge representation of the data content using the Resource Description Framework (RDF) from the World Wide Web Consortium (W3C).[31.]

Recommendation I1:

- Minimal implementation: Available ontologies and controlled vocabularies should be used for data documentation (e. g. the DDI controlled vocabularies). Metadata should report which ontologies and controlled vocabularies (e. g. ISCED, ISCO) were used for data documentation (e. g. CESSDA topic classification, ELSST).
- Maximum implementation: Relevant ontologies for SBE sciences are retrievable to link them with data. This requires the development of (1) one or several information systems for management and retrieval of ontologies and controlled vocabularies (and their different versions) used for data documentation or measurements (e. g. ISCED etc.), (2) an authority that curates ontologies and vocabularies (e. g. a concept registry), and (3) PIDs for ontologies for automated reuse of the metadata. The development of ontologies for SBE sciences constructs is extended. Furthermore, consideration should be made to whether a general knowledge representation of data contents based on the Resource Description Framework (RDF) may be established. The implementation of this recommendation will require a collective effort of service providers.

I2. (meta)data use vocabularies that follow FAIR principles.

The technologies used to implement I1 have to be open for reuse by all researchers. They have to be well documented (e. g. the versioning of the ontologies) and contain PIDs for referencing to and retrieval of resources. The realisation of I2 will stand in connection with the prior implementation of I1 max.

Recommendation I2:

Free access to metadata and systematic use of PIDs for ontologies will be sufficient to fulfil this criterion. Following the implementation of the recommendations of I1, this principle will be fulfilled.

I3. (meta)data include qualified references to other (meta)data.

Data are generally created and used for predefined purposes. Often elements from existing data are reused or re-purposed in the research process. For example, data are harmonised and linked to achieve larger coverage, data are updated with new data in longitudinal projects, or data are enhanced by adding data from other sources (e. g. spatial data, data from media, etc.). The data documentation should reveal existing links between different data sources even if the link was created subsequent to data collection. Operations appending and enhancing data are often

[31] https://www.w3.org/ (last accessed: 18.11.2020).

performed by scholars reusing data and documented in publications. Therefore, a first step to reveal usage contexts of data is, therefore, to systematically link publications to the data used for analysis.
From a more encompassing perspective, relationships between data should be systematically included in the documentation making use of the existing metadata standards.

Recommendation I3:

- Minimal implementation: Publications should be linked to the data that they use or refer to.
- Maximum implementation: The relation type between two (meta)datasets, like the element "relationType" of the DataCite Scheme for cross-referencing different metadata sources is systematically specified.

Reusability

The principle of reusability is closely related to the quality of the data documentation. That is to say, having access to a data file does not yet allow researchers the optimised reuse of data for their own purposes. Instead, researchers need extensive information about data generation, data processing and semantic content of the data. In addition, the use of the data sometimes requires authorisation from or agreements with the data owners and publishers of the data and in any case the respect of intellectual property.

R1. meta(data) have a plurality of accurate and relevant attributes.
Currently, the metadata provided by the RDC catalogues do not offer detailed information about the process of data generation and preparation. The metadata focuses on the description of the final data product (i. e., the file published for data reuse) with some information on methods used for data collection but few information about data collection, processing of the data and the steps which were done prior to the release of the data (provenance).

The peculiarity in the social sciences is that information on data preparation is potentially disclosive and, therefore, cannot be published alongside with the data. This would apply, for example, when using interviewers' names as metadata about the data collection as one would need a signed informed consent for this purpose, or the release of data processing code which includes confidential information.

Even the procedures used for the anonymisation of data are, to some extent, subject to confidentiality, so anonymisation protocols must be kept separate from the published data.
However, a better understanding of these barriers can help data producers to take appropriate actions to make their data reusable, thus avoiding administrative or bureaucratic hurdles later on. However, if the objective of a "data usage" can be reached without confidential provenance data, limiting the supply of this information does not limit usage purposes. Documentation of data collection and/or processing is crucial for the reuse of data but not every piece of information is needed – "where certain limitations or restrictions (e. g. anonymisations) would not hinder analyses, these would need to be implemented" (Schönberg 2019)

A very important aspect for the reuse of data is to specify what users are allowed to do with the data. As common licenses focus on questions of authorship, intellectual property, and how users can manipulate or (re-)distribute the data, an intermediate step is often necessary in SBE sciences as confidentiality and legal and ethical issues play an important role in data sharing (see also section on accessibility). This intermediate step would define a series of requirements

as: what (pieces of) data can be used by whom (researchers of a certain field) and for what purpose (for specific research projects, for teaching purposes or for projects without a follow-up intention) or in what way (only some of the data or a certain type of data, like text-files, limiting data linkage...), all of which would be embodied in licenses/usage contracts.

Regardless of these problems and assuming that they are solved (or minimised), if data producers and data users don't speak the same language, the data won't be understood and therefore neither interpreted nor reused. The consideration of community standards or practices as a common language will help to make data more understandable and exchangeable.

The following recommendations are proposed to address the difficulties related to data reusability.

R1.1. (meta)data are released with a clear and accessible data usage license.
For data sharing in general the adoption of CC BY and CC BY-SA licenses or CC0 (with which the depositor waives all rights to the data) is the most appropriate. The Plan S Principles number 1 suggests: "Authors or their institutions retain copyright to their publications. All publications must be published under an open license, preferably the Creative Commons Attribution license (CC BY), in order to fulfil the requirements defined by the Berlin Declaration."[32] In the case of SBE sciences this recommendation should be applied when possible.

However, there are no standard research data licenses that can cover all usage scenarios for SBE sciences data. On the premise that the authors of the data still retain copyright and ownership of data, customised licenses agreements between the creator of the data set, who are required to affirm that they have the right to publish and redistribute the given material, and the data repository should be created and released alongside with the data and documentation.[33] It should clearly specify what a user is allowed to do with the data in terms of research purposes, analysis procedures, and linking with other data sources.[34]

Recommendation R1.1:
Regarding principle R1.1 we have to differentiate its two possible applications as licensing of metadata is far more uncomplicated than for the reuse of data.

For metadata:

- Minimal and maximal implementation: we recommend using CC0 "No Rights Reserved". Metadata should be of public domain to allow users to harvest, exchange and reuse them without restrictions. CC0 allows creators and data repositories to waive interests in the descriptive information about datasets and therefore place it completely in the public domain, so that others may freely build upon, enhance and reuse that metadata for any purposes without restrictions under copyright or database law.

By removing all restrictions on the reuse of data that describes data (metadata), data repositories or data archives create opportunities for developers, content aggregators, search engines and other digital innovators to create applications, services and websites that visualise

[32] https://www.coalition-s.org/plan_s_principles/ (last accessed: 18.11.2020).
[33] https://www.openicpsr.org/openicpsr/aea (last accessed: 18.11.2020).
[34] https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives/Licensing-your-data (last accessed: 18.11.2020).

and represent the archives' diverse collections of datasets, thereby improving the findability of those datasets (s. also F4).[35]

Moreover, when generating new data from existing research outputs, sources of measurement instruments should be cited where appropriate, as it is standard practice for publications. In this way, an appropriate research assessment by means of data citation, impact measurements will be possible and the potential of reuse will grow.

For data, a second distinction has to be made: in the case of data that can be disseminated without usage restrictions, the use of CC-Licenses is recommended. In general, we encourage, like CESSDA recommends, to "choose a licence which makes data available to the widest audience possible" and "makes the widest range of uses possible."[36]
For data with restricted usage options consider the following recommendations.

- Minimal recommendation: The terms of data reuse should be specified in clearly defined and structured usage contracts that will be available along with the metadata and the data. They should give information about which uses (e. g., for educational purposes, follow-up studies), which persons or groups of persons are authorised to work with these data or for which specific research purposes they are intended. Since the majority of these terms of data usage are based on the informed consent of the persons involved in the data (creators and data subjects), informed consents must always be archived as documentary evidence of the possibilities and/or restrictions on usage.
- Maximum recommendation: Machine-actionable license models are used that regulate and standardise restrictions in data usage considering privacy and confidentiality issues. That is the case of the ANNODATA Framework (Bender et al. 2019) which is a machine-actionable schema that allows specifying data usage terms on a sufficiently granular level, as research increasingly uses granular and linked data (see above). It suggests three important factors for data reuse: clear access rules for individual datasets; rules for the combination of datasets with different rules (different sets of legal and access restrictions); and information on the requesting party, e. g. a researcher or analyst. Thereby the application of this "common annodata taxonomy would facilitate the standardisation of processes and wherever possible automation of tasks and decisions within the data management process" (ibid, 6).

R1.2. (meta)data are associated with their provenance.
Information on provenance is not yet being systematically published with the data. For users, it is difficult to understand the workflow and procedures underlying the data products based only on available metadata. Although pieces of provenance information are already part of the methodological documentation of data collection (sampling methods, data collection methods, etc.), that doesn't cover all provenance aspects. Details on the work done between data collection and data publications are normally not part of the metadata describing data. Such information includes: Who prepared the data and how? Were changes made in the data (e. g. correction of errors etc.)? Which transcription method was used? Which other data sources were employed? What kind of modifications were performed?, etc.

A central step in data preparation is anonymisation of the data. Legal, ethical and confidentiality considerations limit the possibilities of sharing confidential data, such as

[35] https://creativecommons.org/share-your-work/public-domain/cc0/ (last accessed: 18.11.2020).
[36] https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives/Licensing-your-data (last accessed: 18.11.2020).

person-related information on participants in data collections. In this case, anonymisation is the only legal and ethical possibility of sharing data.

Anonymisation is therefore one of the main differentiating features of SBE sciences. Furthermore, when reusing anonymised data, it should be considered that this anonymisation may have affected the analysis potential of those data. In this way, users should be able to understand how anonymisation modified the data and how it may interfere in their analysis. The use of standardised anonymisation systems and the documentation of the anonymisation process are the requirements to improve reusability.

Recommendation R1.2

- Minimum recommendation: Terms from controlled vocabularies (such as those from DDI and CESSDA) should be used for the description of data sets as they allow us to provide standardised documentation about the data generation practices. Additionally, with regards to methods information contained in methods reports, we recommend that information be provided on data processing operations that transform the "raw" data into the published data, including information on transcription and anonymisation procedures, even in the form of general descriptions.
- Maximum recommendation: information on data processing requires the publication of the annotated, machine-actionable code that was used for the preparation of the data, whenever this is possible and as long as no confidential data is revealed with its publication. If the machine-actionable code also contains non-anonymised data, a general description of the code, the purpose of its use, and what results were expected, should be provided.

In general, the use of metadata standards such as DDI, CESSDA CMM, da|ra or the INEXDA metadata schema to document provenance is strongly encouraged.

R1.3. (meta)data meet domain-relevant community standards.

The Data Documentation Initiative is the most used community standard for the SBE sciences, which is being continually updated by and for the community. However, its flexibility and the different approaches of its specifications, which offer the possibility of using the same standard for different needs, increase the complexity in its implementation and generate different interpretations or uses of the same elements.

Recommendation R1.3:

Since standards should match (meta)data needs and practices of the research communities and the definition of relevant information in terms of data reusability depends on research designs and data types, we assume that not all information is equally relevant for all communities. Within the SBE sciences, different ‘data cultures’ exist.

We recommend two specific measures.

- Minimum and maximum implementation: In order to avoid burdensome processes of metadata generation in overly complex tools, research data infrastructures should cooperate with research communities to develop domain-specific versions of generally accepted standards. Communities should specify how they make use of a certain standard by creating a profile of that standard. This would simplify and narrow down the elements of a standard to those that are relevant to the community and avoid potential semantic ambiguities. Using the example of the DDI, the harmonisation of its use through the establishment of specific profiles for specific user groups and the

generalisation of the use of DDI-compliant documentation tools are both key issues for the implementation of this principle. This applies also to semantic artefacts like terminologies and ontologies: its design, standardisation and use by the community as common languages helps to ensure unambiguity by naming the same concepts with the same terms and URIs.

- Developing of Domain Data Protocols which define which provenance information are required for data reuse.[37] We encourage DDPs to define how FAIR data sharing is allowed by designing the data collection in an appropriate manner (e. g., in relation to informed consent and documentation standards).

Final Remarks

With the establishment of the model of Research Data Centres (RDCs) and the archives from the Consortium of Social Science Data Archives (CESSDA), the social, behavioural and economic (SBE) sciences have already laid the groundwork for a research data infrastructure that complies with the FAIR principles. So far this applies particularly regarding the findability of research data via PIDs and data catalogues as well as regarding accessibility and reusability via high documentation standards. With the newly emerging KonsortSWD within the National Research Data Infrastructure (NFDI) of Germany and the initiatives planned within this consortium, the conditions are particularly promising that aspects such as the standardisation of metadata practices, the improvement of data interoperability and machine-actionability will also soon be tackled to make the research data infrastructure in the SBE sciences FAIRer. This will also facilitate the integration of national data centres into the European Open Science Cloud (EOSC).

Generally, the FAIRness and the success of the research data infrastructure will depend to a large extent on cooperation between the data centres and the agreements that are reached also within larger coordinating institutions, such as EcoSoc-IN or within KonsortSWD and EOSC. Only with coordination it can be secured, that costly and non-interoperable, isolated applications or redundant problem solving is prevented. A useful tool to strengthen mutual learning and to help best practice sharing is the FAIR Convergence-Matrix. Through the matrix, individual RDC but also research data managers in smaller research projects get an insight of how others tackled similar challenges and what standards the community expects for data reusage.

The goal of this paper was to show, that the FAIRness of research data is a continuum and the underlying principle of all the steps taken should be to optimise the reusability of research data. Consequently, the FAIR principles should not only be realised on a general technical level. Instead, within the SBE sciences, special demands regarding data protection, confidentiality and research ethics need to be considered. Moreover, the diversity of data types used in these disciplines affects how research data can and should be managed by humans and machines. Thus, when implementing the FAIR principles, the different disciplinary standards and specific disciplinary challenges need to be considered. This White Paper opens a discussion about the implementation of the FAIR principles in SBE. It shall serve as a guidance by outlining possible solutions but it shall also invite RDCs and other stakeholders to discuss the ongoing process to FAIRify the research data infrastructure in the SBE sciences.

[37] In this project, connectable Domain Data Protocols for empirical educational research are currently being developed: https://www.gesis.org/en/research/external-funding-projects/overview-external-funding-projects/ddp-bildung (last accessed: 18.11.2020).

Notes on contributors

Noemi Betancort Cabrera
Systems librarian - Qualiservice metadata manager
Staats- und Universitätsbibliothek Bremen
Digitale Dienste
Bibliothekstraße 9
28359 Bremen
noemi.betancort@suub.uni-bremen.de

Dr. Elke C. Bongartz
Team Leader Library
German Institute for Adult Education (DIE)
Leibniz Centre for Lifelong Learning
Heinemannstraße 12-14
53175 Bonn
bongartz@die-bonn.de

Dr. Nora Dörrenbächer
Research Associate Business Office German Data Forum (RatSWD)
Wissenschaftszentrum Berlin (WZB)
Reichpietschufer 50
10785 Berlin
ndoerrenbaecher@ratswd.de

Dr. Jan Goebel
Board of Directors SOEP & Division Head Data Operation and Research Data Center in the German Socio-Economic Panel
German Institute for Economic Research (DIW)
Mohrenstrasse 58
10117 Berlin
jgoebel@diw.de

Harald Kaluza
Research Associate
German Institute for Adult Education (DIE)
Leibniz Centre for Lifelong Learning
Heinemannstraße 12-14
53175 Bonn
kaluza@die-bonn.de

Dr. Pascal Siegers
Head of Research Data Center German General Social Survey
GESIS - Leibniz-Institute for the Social Sciences
Unter Sachsenhausen 6-8
50667 Köln
pascal.siegers@gesis.org

References

Bender, Stefan; Jannick Blaschke, Hendrik Doll, Andrew Gordon, Christian Hirsch et al. (2019): The Annodata Framework: Putting FAIR data into practice. Technical Report 2019-03. Deutsche Bundesbank, Research Data and Service Centre. https://www.bundesbank.de/resource/blob/826340/f229ee339355e9a263ab9f1c1424238d/mL/2019-03-annodata-framework-data.pdf (last accessed: 18.11.2020).

European Commission Directorate-General for Research & Innovation (2016): H2020 Programme Guidelines on FAIR Data Management in Horizon 2020. Version 3.0. 26 July 2016. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (last accessed: 18.11.2020).

European Commission Directorate-General for Research and Innovation (2020): Six Recommendations for Implementation of FAIR Practice. By the FAIR in Practice Task Force of the European Open Science Cloud FAIR Working Group. DOI: 10.2777/986252.

Friedrich, Tanja and Pascal Siegers (2016): The Ofness and Aboutness of Survey Data: Improved Indexing of Social Science Questionnaires. In: Adalbert F.X. Wilhelm and Hans A. Kestler (eds.): Analysis of Large and Complex Data. Studies in Classification, Data Analysis, and Knowledge Organization, 629-638. Cham, Springer International Publishing Switzerland. DOI: 10.1007/978-3-319-25226-1_54.

Hausstein, Brigitte and Laurence Horton (2020): CESSDA ERIC Persistent Identifier Policy 2019. Version 2.0. Zenodo. DOI: 10.5281/zenodo.3611327.

Hellström, Maggie, André Heughebaert, Rachael Kotarski, Paolo Manghi, Brian Matthews et al. (2020): Second draft Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC). Version 2.0. Zenodo. DOI: 10.5281/zenodo.3780423.

Jacobsen, Annika; Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles et al. (2020): FAIR principles: Interpretations and implementation considerations. Data Intelligence 2(1-2), 10–29. DOI: 10.1162/dint_r_00024.

Juty, Nick; Sarala M. Wimalaratne, Stian Soiland-Reyes, John Kunze, Carole A. Goble at al. (2020): Unique, persistent, resolvable: Identifiers as the foundation of FAIR. Data Intelligence 2(1-2), 30–39. DOI: 10.1162/dint_a_00025.

Koers, Hylke; Daniel Bangert, Emilie Hermans, René van Horik, Maaike de Jong et al. (2020): Recommendations for Services in a FAIR Data Ecosystem. Patterns 1(5), 100058. DOI: 10.1016/j.patter.2020.100058.

Landi, Annalisa; Mark Thompson, Viviana Giannuzzi, Fedele Bonifazi, Ignasi Labastida et al. (2020): The “A” of FAIR – As open as possible, as closed as necessary. Data Intelligence 2(1-2), 47–55. DOI: 10.1162/dint_a_00027.

Linden, Mikael; Michal Procházka, Ilkka Lappalainen, Dominik Bucik, Pavel Vyskocil et al. (2018): Common ELIXIR Service for Researcher Authentication and Authorisation [version 1; peer review: 3 approved, 1 approved with reservations]. F1000Research 2018, 7(ELIXIR), 1199. DOI: 10.12688/f1000research.15161.1.

Linne, Monika; Ines Drefs, Nora Dörrenbächer, Pascal Siegers and Mathias Bug (2021, in press): GO FAIR und GO CHANGE: Chancen für das deutsche Wissenschaftssystem. In: Markus Putnings, Heike Neurothand Janna Neumann (eds.): Praxishandbuch Forschungsdaten-management. Berlin, Boston, De Gruyter Saur. https://www.degruyter.com/view/title/554542 (last accessed: 18.11.2020).

Martinez, Paula Andrea; Christopher Erdmann, Natasha Simons, Reid Otsuji, Stephanie Labou et al. (2019): Top 10 FAIR Data & Software Things. Zenodo. DOI: 10.5281/zenodo.3409968.

Mokrane Mustapha and Jonas Recker (2019): CoreTrustSeal–certified repositories: Enabling Findable, Accessible, Interoperable, and Reusable (FAIR). 16th International Conference on Digital Preservation (iPRES). DOI: 10.17605/OSF.IO/9DA2X.

Mons, Barend; Erik Schultes, Fenghong Liu and Annika Jacobsen (2020): The FAIR principles: First generation implementation choices and challenges. Data Intelligence 2(1-2), 1-9. DOI: 10.1162/dint_e_00023.

Pergl Sustkova, Hana; Kristina Maria Hettne, Peter Wittenburg, Annika Jacobsen, Tobias Kuhn et al. (2020): FAIR convergence matrix: Optimizing the reuse of existing FAIR-related resources. Data Intelligence 2(1-2), 158-170. DOI: 10.1162/dint_a_00038.

RatSWD [German Data Forum] (2017): Forschungsethische Grundsätze und Prüfverfahren in den Sozial- und Wirtschaftswissenschaften. RatSWD Output 9(5). Berlin, German Data Forum (RatSWD).

RatSWD [German Data Forum] (2018): The German Data Forum (RatSWD) and Research Data Infrastructure: Status Quo and Quality Management. RatSWD Output 1(6). Berlin, German Data Forum (RatSWD). DOI: 10.17620/02671.30.

RatSWD [German Data Forum] (2019): Remote Access to data from official statistics agencies and social security agencies. English Summary. RatSWD Output 5(6). Berlin, German Data Forum (RatSWD). DOI: 10.17620/02671.48. Full Report available in German. DOI: 10.17620/02671.42.

RatSWD [German Data Forum] (2020): Data collection using new information technology. Recommendations on data quality, data management, research ethics, and data protection.. RatSWD Output 6(6). Berlin, German Data Forum (RatSWD). DOI: 10.17620/02671.51.

RDA FAIR Data Maturity Model Working Group (2020): FAIR Data Maturity Model: specification and guidelines. Research Data Alliance (RDA). DOI: 10.15497/rda00050.

RDA International Indigenous Data Sovereignty Interest Group (2019): CARE Principles for Indigenous Data Governance. The Global Indigenous Data Alliance. https://www.gida-global.org/care (last accessed: 18.11.2020).

Schönberg, Tobias (2019): Data Access to Micro Data of the Deutsche Bundesbank. Technical Report 2019-02, Deutsche Bundesbank, Research Data and Service Centre. https://www.bundesbank.de/resource/blob/801044/6484d9e4aa2be3610b7378a48a1916de/mL/2019-02-data-access-data.pdf (last accessed: 18.11.2020).

Schultes, Erik (2019, pre-print): Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. Version 2.0, 11 December 2019. https://osf.io/8sv5f/ (last accessed: 18.11.2020).

UNESCO Institute for Statistics (2012): International Standard Classification of Education (ISCED) 2011. http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf (last accessed: 18.11.2020).

Wilkinson, Mark D.; Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018. DOI: 10.1038/sdata.2016.18.

Wilkinson, Mark D.; Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos et al. (2017): A design framework and exemplar metrics for FAIRness. DOI: 10.1101/225490.

Appendix

The FAIR principles[38]

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-actionable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

[38] For more detail see: https://www.go-fair.org/fair-principles/

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards