DEUTSCHE BUNDESBANK
EUROSYSTEM

The Annodata Framework: Putting FAIR data into practice

Technical Report 2019-03

S. Bender, J. Blaschke, H. Doll, A. Gordon, C. Hirsch, D. Hochfellner, J. Lane

Disclaimer
The views expressed in this technical report are personal views of the authors and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

Citation
Bender, S., J. Blaschke, H. Doll, A. Gordon, C. Hirsch, D. Hochfellner, J. Lane (2019). The Annodata Framework: Putting FAIR data into practice. Technical Report 2019-03, Deutsche Bundesbank, Research Data and Service Centre.

The Annodata Framework: Putting FAIR data into practice

Stefan Bender[1], Jannick Blaschke[1], Hendrik Doll[1], Andrew Gordon[2], Christian Hirsch[1], Daniela Hochfellner[3], Julia Lane[3]

Abstract:

Empirical data usage increasingly penetrates all fields of social science research. Large journals react by introducing data sharing requirements. The FAIR principles capture such best practices for research data use. However, the fraction of publications exempt from data sharing requirements increases. This puzzle arises, because research increasingly uses granular and linked data. Such data offers enhanced possibilities to identify research issues of interest. Yet, with increasing granularity and linkages, data protection and privacy protection issues get complex. As a solution, we introduce the annodata framework. By establishing a granular set of descriptive items pertaining to technical and legal data access, we enable a practical implementation of FAIR data access for confidential microdata. The annodata framework benefits empirical research through improved reproducibility by rendering granular data used in publications findable and accessible. Data users benefit from a level playing field with clear data usage terms and efficient data access. Data owners benefit through reduced redundancy in data governance processes and clear compliance with legal and audit norms

Keywords:

Annodata, administrative data, FAIR data, digital exhaust, microdata usage, metadata, machine-readable, data stewardship, confidential data, data management, data description, big data

[1] Deutsche Bundesbank
[2] Columbia University
[3] New York University

1. Introduction

The approach that we describe in this paper stems from our journey to finding a metadata schema for describing data access workflows to highly sensitive granular data in a secure facility. Highly sensitive granular data refers to data on the level of individual households or businesses where de-identification is a risk and access procedures need to ensure highest privacy protection standards. The aim of our project was to adhere to FAIR[2] data principles and maximize discovery, and reuse by researchers of these datasets.

We start by evaluating the suitability of established metadata schemas such as DDI for our aim. We find that they generally provide one (or more) metadata items that deal with data access.[3] However, we conclude that the limited number of items and the lack of controlled vocabulary make these approaches inadequate[4] for the data that we wanted to describe for three reasons. First, existing approaches run the risk of misinterpretation by data stewards.

Most of the surveyed metadata schemas have data access items as free text format. For example, a free text statement could read, “Researchers in general only have access to anonymized data”. Interpreting free text information may very likely differ with data stewards making access to data potentially depended on the person processing the request fundamentally disagreeing with FAIR data access principles. In the example above, it is unclear whether the word “general” indicates the existence of unspecified exemptions.

Second, existing approaches run the risk of being incomplete. Even if everybody processes the information in the same way information may be incomplete as important parts may reside in places outside of the metadata schema (for example as tacit knowledge of data stewards) further contradicting FAIR principles. Research projects involving highly sensitive microdata routinely link several granular datasets together for their analysis further discombobulating access procedures.

Third, existing approaches based on few items and free text are not machine-readable. We propose an approach inspired by the use of paradata in survey methodology (West, 2011) which captures auxiliary information about the interview process, including interviewer and

[2] The FAIR principles are an acronym for Findability, Accessibility, Interoperability, and Reusability, see for example Wilkinson et al. 2016.

[3] Items describing access to data in DDI include AccessConditions, AccessPermissions, accessRights, and TypeoOfAccess.

[4] Note that we do not intend to criticize established metadata schema fundamentally. Rather, our assessment stems from our aim of describing a very specific type of data.

respondent behaviors. The approach, called annodata, refers to a set of information needed to describe FAIR[5] data access workflows for sensitive granular data. Just as paradata, annodata are designed with the intent to complement extant metadata schemas not to supersede them.

We proceed as follows. First, we provide a discussion of prevalent gaps in existent metadata schema that retrain their usefulness to facilitate efficient sharing and reusability of data. We proceed by introducing the annodata schema and highlighting how the annodata schema helps promoting the FAIR principles. As the FAIR principles are written with an explicit emphasis on machine-action ability, the annodata schema to is designed to inform the design of software packages, harmonizing, standardizing, and automatizing data access.

2. Why can't we rely on existing metadata standards?

To contextualize our work, we look at existing metadata approaches. It is important to note here that the overall purpose of metadata is to serve the goals of the community that uses and the organizations that provide them (Willis, Greenberg & White, 2012). Different communities and organizations have different goals that guide their collection, usage, and sharing of data.

Many existing data repositories and archives discuss their work in creating, organizing, and disseminating descriptive metadata about datasets such that these datasets might be discovered, shared, understood, and reused (Hancock, 2017; Dietrich, 2010; White, 2014; Moss et al., 2016; Rücknagel et al., 2015). Thus, with the rise of generally used and flexible metadata schemas (schema.org, DataCite, Dublin Core, etc.), datasets can nowadays be described in a flexible and generally understood way.

However, information on how data are released, protected, controlled, and accessed is less well defined in the literature, particularly issues of tracking usage auditing. There is no single, easy solution for this perennial problem. Standards like the Metadata Encoding & Transmission Standard (METS) provide a scaffolding by which to define administrative metadata pertaining to intellectual property and copyrights, how objects are created and stored and original source information, but might not be granular enough to use for

[5] The FAIR principles are an acronym for Findability, Accessibility, Interoperability, and Reusability, see for example Wilkinson et al. 2016.

designating complicated usage restrictions by file, table, column, or specific fields in datasets.

Standards such as PROV-O and PREservation Metadata: Implementation Strategies (PREMIS), another metadata standard stewarded by the Library of Congress, provide guidance for defining lineage and provenance around the creation and maintenance in preserving digital objects. Gunia and Sandusky (2010) provide a detailed description of how PREMIS can be used to preserve Earth Science data, but are not aimed at the complicated chain of transformations that occur over the lifetime of a datasets usage.

Chao, Cragin, and Palmer (2014), however, have proposed a standard data curation vocabulary, which incorporates not only a data description but also practical steps of how data is used by the researchers that produce and share it. The International Rights Statements Working Group (2015) introduced a standardized vocabulary to describe usage terms and copyright status of intellectual work, which they coin “rights statements”.

This shows that an approach to characterize data beyond the classical metadata schema is highly needed as classical schemata rely on the assumption that data are a “made” entity, describing data from the production side only. We tackle this challenge by enhancing the classical metadata concept to include data administration as well. The annodata framework conceptualizes this combination.

3. The annodata framework

We define annodata as all information on the process for providing access to data, i.e. information pertaining to the set of legal requirements that have to be considered when making data available for research and analysis. The point of this chapter is to show, the level of granularity necessary when describing data access in order to derive deterministic and transparent data access descriptions. In the appendix, we provide a detailed list of annodata items.

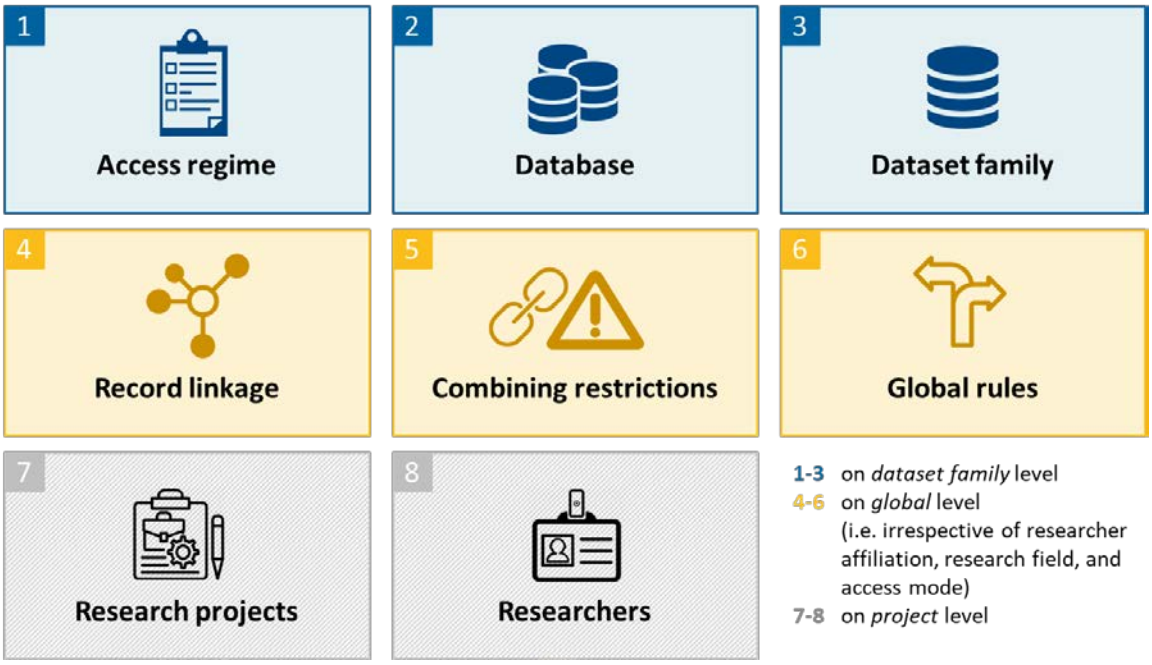
Note, that the annodata concept is not about manually producing new information when performing activities described in the data lifecycle. Rather, the contribution of the annodata concept is to provide a framework to make previously unstructured information, which has only been available as digital exhaust available for reuse. This feature of annodata is in accordance with the paradata concept in the survey data literature, where information already exists but is not curated and made available for reuse.

Therefore, to be able to transform information into reusable knowledge the annodata concept needs to have the following two features. First, annodata need to be machine-

readable to support automation of processes and decisions. Second, the efficient governance of micro data requires a clear and machine-readable set of rules, which are consistent across different datasets and potentially across different types of facilities and repositories. A common annodata taxonomy would facilitate the standardization of processes and wherever possible automation of tasks and decisions within the data management process.

Based on these thoughts, we distinguish three dimensions of annodata: First, clear access rules for individual datasets; Second, rules for the combination of datasets with different rules (different sets of legal and access restrictions); And third, information on the requesting party, e.g. a researcher or analyst, which need to be collected. Figure 1 illustrates the annodata schema's three dimensions, further split into eight sets of attributes.

Figure 1: Eight main types of annodata for the example of a research data center



The main purpose of annodata is to advance from a labor-intensive manual assessment of data governance rules, to a rule-based attribution of data to existing rules of governance. An access regime describes means to access and work with data of a particular confidentiality level. While there are many datasets, there is a fairly limited number of different legally or technically necessary access regimes.

Thereby, annodata also improves legal certainty, since newly incoming data only has to be assigned to a pre-existing established access regime. By mapping administrative information to access regimes, many datasets can share the same governance attributes.

Access regimes typically contain a number of different modes to access individual datasets. An access mode is a mode via which access to the data can be granted. Examples include download of data or secure on-site access at the premises of the data providing institution.

Each access mode in turn may have a number of different access protocols attached to it. Access protocols describe the criteria that have to be fulfilled to be granted access under a specific access mode. The protocol-criteria often times are imposed by a combination of the legal basis, the affiliation of the researcher (e.g. internal vs. external) and the degree of anonymization (non-anonymized vs. fully anonymized) of the requested data.

The annodata framework distinguishes between a database and a dataset family. Following this definition, in our example, the national credit register would be the database whereas an extract of the national credit register covering variables 1-n would be a dataset family. For the sake of reproducibility, we also need the concept of individual datasets, i.e. variables 1-n of the national credit register for the period 1992 to 2017, which are fixed over time (frozen slices in time and variable coverage) and assigned to persistent identifiers such as e.g. Digital Object identifiers (DOI).

For practical reasons, data access is typically granted on the level of a dataset family, which contains multiple datasets that only differ in their time coverage. This stems from the fact, that the research process often covers up to several years from project proposal to revise and resubmits, in the process of which the necessity for temporal updates of the data arises and is generally granted by data-providing institutions.

This implies that rules for data access should generally be the same within a dataset family. Thus, access regimes are assigned to dataset families. One may also tie access regimes to the higher level of databases. While this reduces manual effort in assigning access regimes to lower-level dataset families, often times different parts of a database might be governed by different access regimes. For example, a database may source information both from commercial data vendors and proprietary data collection. When assigned on a higher level, individual datasets inherit available attributes from dataset families and databases.

The second annodata dimension describes access to multiple datasets (families) within a single project. When combining several datasets one might face questions regarding the technical feasibility of combining the data, the legal questions, and the resulting combined access protocol that results from datasets with different access regimes.

Besides information on the technical feasibility, the item “record linkage” includes information on the methods applied and the underlying assumptions used when creating

this link. The latter is especially important to put researchers into the position to gauge the quality of the link and take a decision whether they want to use this link in own research.

Information on any legal restrictions applying to specific dataset combinations is collected in the "combining restrictions" element. These arise, since combining information sources can de-anonymize otherwise less confidential observations. Rules under this banner contain restrictions on which IDs to combine (e.g. to prohibit de-anonymizing internal identifiers) or restrictions on which attributes to combine (e.g. personal data containing sensitive information on criminal history and credit scoring).

To allow the possibility of linking dataset families from different access regimes, annodata then needs to define unambiguous "decision rules" which access protocol applies in these cases. For example, different access regimes may call for different contracts that the researcher has to sign before being granted data access. For these cases, the decision rule's element may specify a dedicated protocol that contain a rule to which contract applies under the given circumstances. Examples of such decision rules could constitute "stronger regime wins", "additive rules", or "new combined regime".

While the first two annodata dimensions link to dataset families, the third dimension covers the other side, i.e. the data requesting party, and is defined on the level of a project. Access regimes are divided in different access modes whose availability not only depends on the degree of anonymization and the type of access mode (e.g. secure on-site, download, remote access) but also on the type of researcher (e.g. internal to the respective organization or external). The availability of structured information on both the data as well as the requestor side is essential to fully leverage on the automation potential of annodata.

4. How annodata helps close the gap in extant metadata schemata

We argue that the annodata schema significantly increases transparency on access rules for non-public data and therefore facilitates both reusability and reproducibility of previous outcomes. Reproducibility refers to the closeness of the agreement between the different outcomes conducted using the same data and methodology. In this way, the annodata concept enables the application of the FAIR principles.

The FAIR principles define accessibility as "Once the user finds the required data, she/he needs to know how can be accessed, possibly including authentication and authorization". We argue that the annodata schema is ideally suited to facilitate access to non-public data as it was developed in a community of research data centers (RDCs) that are responsible for making confidential microdata available to internal and external researchers.

The origin of the concept ensures that these privacy principles can be attached to a given dataset. The resulting annodata can be used by data-governing institutions to automate privacy protection mechanisms at all times during a research lifecycle.

Thereby processes are sped up and audit requirements are ensured, enabling data owners to report at all times, who uses data under which legal basis and restrictions. Such data owners can be official institutions governing data access to confidential administrative data sets, commercial data vendors and other private businesses trying to leverage on their confidential data resources (confidential often because of personal data).

Annodata that contain machine-readable and detailed information on dataset-related access rules and restrictions can be used to design workflows providing access to confidential micro data and support data governance in general. We argue that this would support administrative tasks of data stewards in providing internal and external researchers access to data while simultaneously safeguarding the confidentiality of the submitted information.

Second, the annodata concept facilitates the reusability of data. The FAIR principle R.1.1 states that “(Meta)data are released with a clear and accessible data usage license”. The benefit of the annodata framework is the comprehensive important information needed to provide access to data, for the first time on a sufficiently granular level.

Access regimes attached to dataset families provide detailed criteria under which researchers can access datasets. Access protocols, in turn, provide detailed workflows on what needs to be done in order to provide researchers with data access.

A consequence of implementing annodata principles is a large reduction of legal uncertainty by standardizing data governance. Current data access procedures develop toward tiered access, which is to base access on clearly defined user criteria, which yields the decision who can use which data. This is facilitated by the standardized data governance in the annodata schema.

Similarly important for reusability is information about the legal and technical feasibility of linkages between datasets. This is especially important to prevent the disclosure of information concerning an individual person or business entity in confidential datasets. This is a severe problem as simple anonymization of datasets may not lead to the desired result of fully anonymous data when allowing combining datasets (e.g. De Montjoye, Radaelli, Singh, and Pentland, 2015).

In the context of FAIR Interoperable is defined as “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". While the first sentence refers to the technical and legal feasibility of combining data from different sources, the second sentence refers to the interoperability of processes and workflows from different data providers.

Annodata creates a common taxonomy. By introducing a standardized and specific data governance language for communication between and within institutions, annodata provides the possibility to learn from and quickly adapt best practices and industry standards, harmonize, automate and share data governance procedures across institutions. This gives data producing and using institutions the opportunity to collaborate.

Along with such standardization comes cost reduction through automation. Standardized processes can be built on standardized taxonomies. Both internally (disseminating data, supporting analysts), as well as externally (exchanging, combining data between institutions). Through automatic "consulting" of staff, analysts, and users (recommending data, recommending data applications), further quality improvements and efficiency gains can be obtained.

5. Conclusions

Currently many potential ways forward to promote the reusability and reproducibility of private and potentially confidential data are being discussed. We contribute to this discussion by introducint the annodata concept which may help simplify the practice of data access and data sharing. We argue in this article that this concept will help make data comply with the FAIR principles, which in turn will help facilitate reusability of data.

By introducing a standard for transparent confidential data access, benefits for data owners and data users arise. Data users benefit from a level playing field with clear and accessible data usage terms and fast and efficient data access. Data owners benefit through efficient data handling processes with less redundancy in data governance and clear compliance with legal and audit norms. Empirical research benefits through improved reproducibility by rendering data used in publications findable, and accessible.

6. References

American Economic Association. (2008). "Data Availability Policy." https://www.aeaweb.org/journals/policies/data-availability-policy (accessed January 27, 2020).

Chao, T. C., Cragin, M. H. and Palmer, C. L. (2015), Data Practices and Curation Vocabulary (DPCVocab). J Assn Inf Sci Tec, 66: 616-633. doi:10.1002/asi.23184

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. Journal of Economic Literature, 56(3), 920-80.

De Montjoye, Yves-Alexandre, Laura Radaelli, Vivek Kumar Singh, Alex “Sandy” Pentland, (2015). Unique in the shopping mall: On the reindentifiability of credit card metadata, Science, Vol. 347, Issue 6221, pp.536-539.

Dietrich, Dianne, (2010). Metadata Management in the Data Staging Repository, Journal of Library Metadata, Vol. 10 Issue 2, pp.79-98 DOI: 10.1080/19386389.2010.506376

Gunia, Betsy and Sandusky, Robert J. (2010), Designing metadata for long-term data preservation: DataONE case study. Proc. Am. Soc. Info. Sci. Tech., 47: 1-2. doi:10.1002/meet.14504701435

Hancock, Andrew, (2017). The Modernisation of Statistical Classifications to Knowledge and Information Management Systems, The Electronic Journal of Knowledge Management Vol. 15, Issue 2. pp. 126-144

International Rights Statements Working Group, (2015) Recommendations for Standardized International Rights Statements.

Frauke Kreuter (ed.), (2013). Improving Surveys with Paradata: Analytic Uses of Process Information.

Moss, Elizabeth, Christin Cave, and Jared Lyle, (2015). "Sharing and citing research data: A repository's perspective." West Academic Publishing.

Rücknagel, J., Vierkant, P., Ulrich, R., Kloska, G., Schnepf, E., Fichtmüller, D., Reuter, E., Semrau, A., Kindling, M., Pampel, H., Witt, M., Fritze, F., van de Sandt, S., Klump, J., Goebelbecker, H.-J., Skarupianski, M., Bertelmann, R., Schirmbacher, P., Scholze, F., Kramer, C., Fuchs, C., Spier, S., Kirchhoff, A. (2015): Metadata Schema for the Description of Research Data Repositories: version 3.0, 29 p. DOI: http://doi.org/10.2312/re3.008

Taylor, Sean J., (2013). “Real Scientists Make Their Own Data.” Sean J. Taylor Blog, January 25. Available at http://bit.ly/15XAq5X

Vilhuber, Lars, (2019). “Report by the AEA Data Editor,” AEA Papers and Proceedings, vol. 109, pp. 718-29.

White, Hollie C., (2014). Descriptive Metadata for Scientific Data Repositories: A Comparison of Information Scientist and Scientist Organizing Behaviors, Journal of Library Metadata, Vol. 14, Issue 1, pp. 24-51, DOI: 10.1080/19386389.2014.891896

Willis, Craig, Greenberg, Jane and White, Hollie., (2012), Analysis and synthesis of metadata goals for scientific data. J Am Soc Inf Sci Tec, 63: 1505-1520. doi:10.1002/asi.22683

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., & Bouwman, J., (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3.