An introduction to INEXDA's metadata schema[1]

Stefan Bender, Brigitte Hausstein and Christian Hirsch,
Deutsche Bundesbank

---

[1] This paper was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

An introduction to INEXDA's metadata schema[1]

Stefan Bender,[2] Brigitte Hausstein,[3] and Christian Hirsch[4]

This paper introduces the metadata schema used by the international network INEXDA to describe granular datasets from different countries. The schema, agreed on by all members, facilitates a comprehensive inventory of existing granular datasets conducted in the member institutions. This inventory, in turn, will foster harmonisation activities between INEXDA members, broaden metadata sharing and potentially future data sharing between institutions represented in the network, and pave the way for metadata on publicly available granular datasets to be shared with external researchers. The INEXDA metadata schema was developed to be easily adaptable for non-INEXDA institutions.

Keywords: Metadata, Microdata, International Network

1. Introduction

Metadata is essential for documenting data, citing it and finding it in catalogues. Metadata can be defined simply as information about data, ie a description of data. It is "…structured or semi-structured information which enables the creation, management, and use of records [i.e. data] through time and within and across domains in which they are created. Recordkeeping metadata can be used to identify, authenticate, and contextualize records; and the people, processes and systems that create, manage and maintain and use them" (Wallace, 2001, p.255). Metadata provides a means for visibility and presentation of data as well as discovery. It supports the re-use, management, exchange and long-term preservation of data. "Data without metadata is just stuff. Nobody needs more stuff today" (Recker, 2014).

The use of a metadata schema goes through various standardisation phases, from which more or less binding metadata standards develop. If such a standard is specified, documented and legally recognised, it is referred to as a standard. In addition, there are de facto or quasi standards based on discipline-specific practical experience and recognised rules of a community. Meanwhile, there is a large array of metadata standards focusing on a particular subject domain, content type, function or application, such as:

- Data Documentation Initiative (DDI)

---

[1] The views expressed here are those of the authors and do not necessarily reflect those of the Deutsche Bundesbank, GESIS, or the INEXDA Network.

[2] Head of Research Data and Service Centre, Deutsche Bundesbank.

[3] GESIS Leibniz Institute for the Social Sciences.
[4] Research Data and Service Centre, Deutsche Bundesbank.

- Dublin Core Metadata Initiative (DCMI)

- DataCite Metadata Schema

- da|ra Metadata Schema

- Metadata Encoding and Transmission Standard (METS)

- Preservation Metadata Maintenance Activity (PREMIS)

- Statistical Data and Metadata Exchange (SDMX)

The significance of a metadata standard lies in the uniform language, which makes it possible to record information about data in an understandable and structured manner. In the case of domain-specific standards, metadata often constitutes very granular – semantically rich – statements about an object. In order to describe data uniformly and unambiguously, standardised terminologies (eg controlled vocabulary such as thesaurus or keywords) are used for certain metadata elements to describe particular metadata with uniform information. This should be distinguished from semi-structured or unstructured metadata, which describe important contextual information on the creation of the data.

Since its foundation in January 2017, the members of the International Network for Exchanging Experiences on Statistical Handling of Granular Data (INEXDA) have, amongst other things, collaborated to harmonise metadata structures and complete an extensive stock-take of available datasets in member institutions. This stock-taking exercise has three aims.

- First, it is designed to give member institutions an overview of the available and possible comparable granular datasets.

- Second, it makes it easier for data users to discover and use datasets appropriate for research and analysis by using harmonised metadata.

- Third, it provides a framework to facilitate a possible harmonisation of datasets in the (near) future.

Because the description of the data should be comparable, all members agreed on a metadata schema for the granular data. This paper presents the metadata schema underlying the stock-taking exercise by INEXDA members. We also explain practical considerations when implementing a metadata schema for microdata, which can easily be adopted by other institutions. Naturally, the INEXDA metadata schema was designed with certain goals in mind. Therefore, some choices may not translate to other situations. For example, the INEXDA schema allows for information about the relation of different datasets to each other, which is essential for the possible harmonising of datasets.

2. The INEXDA metadata schema

2.1 Scope

The INEXDA metadata schema is designed to provide metadata for microdata on the dataset level. Since microdata are data at the level of individual reporters, the data may allow re-identification of individual reporting units. Because of its sensitive nature, microdata is therefore always subject to protection of confidentiality for

individual observations. Metadata about microdata also has to adhere to the same high standards when it comes to protecting confidentiality.

The INEXDA metadata schema is based on the da|ra metadata schema (Version 4.0) which was jointly developed by GESIS – Leibniz Institute for the Social Sciences and ZBW – Leibniz Information Centre for Economics (Koch 2017). da|ra operates a registration service for social science and economic data which allocates a Digital Object Identifiers (DOI) to different types of datasets, allowing a unique identification of resources (https://www.da-ra.de). da|ra maintains a cooperation with DataCite (https://www.DataCite.org).

Adapting an existing metadata schema to fit the purpose of INEXDA provides a level of standardisation for microdata coming from different countries, institutions, and collection purposes. In addition, the interoperability with the da|ra schema allows for a seamless transition between the INEXDA and da|ra databases, which makes it easier to obtain a DOI identifiers in the future. However, ensuring interoperability may also result in restrictions on altering metadata items (see section 2.2).

2.2 Metadata items

All INEXDA members agreed on a metadata schema, which, on the one hand, describes the datasets in a sufficient way for the purposes mentioned above while, on the other hand, being easy to handle for potential users or data producers. Therefore, the INEXDA network came up with 21 items (see Table 1), which we describe below in more detail. For each metadata item, we also provide an example drawn from Bundesbank datasets to illustrate its use. Please note that you will find the example after the description of the item. These examples are written in italics.

1	Resource Type
2	Resource Identifier
3	Name of Dataset
4	Creator
5	DOI Proposal
6	URL
7	Language of Resource
8	Publication Date
9	Availability
10	Sampled Universe
11	Sampling
12	Temporal Coverage
13	Time Dimension
14	Collection Mode
15	Unit Descriptions
16	Descriptions
17	Geographical Coverage
18	Keywords
19	Alternative Identifiers
20	Relations
21	Publications

Table 1: INEXDA metadata schema: items

Resource Type

A description of the underlying resource for which metadata is compiled. In case of INEXDA, the resource type is always a dataset. This item is included in the INEXDA metadata schema to facilitate interoperability with the da|ra schema.

Resource Identifier

The "Resource Identifier" is a unique value to disambiguate between resources. A version number can be provided as a reference that changes have been made between versions.

Name of Dataset

This item displays the name of the dataset.
Example: Securities Holdings Statistics – Base (name of one Bundesbank dataset)

Creator

The item "Creator" refers to the name(s) of the institution, and/or division, and/or department responsible for developing, collecting and/or managing the dataset. Names of individual persons should not be provided.
Example: Deutsche Bundesbank, Statistics Department

DOI Proposal

The schema allows suggesting a DOI name for the dataset. A Digital Object Identifier (DOI) is a permanent, persistent identifier used for citing and linking electronic resources (texts, research data or other content).
Example: 10.12757/Bbk.SHSBase.05121603 (for one version of the Securities Holding Statistics)

URL

URL is the address of a webpage which displays information about the dataset, if available. Alternatively, the landing page of the research data centre or the institution could be used if no specific webpage exists for the dataset.
Example:
https://www.bundesbank.de/Redaktion/EN/Standardartikel/Bundesbank/Research_C entre/research_data_micro_data_securities_holdings_statistics.html?nsc=true&https =1

Language of Resource

This item refers not the language in which the metadata elements are expressed but instead refers to the language in which the dataset is available. Default is English. List a language other than English only if dataset is not available in English.
Example: English

Publication Date

The date on which the dataset was released internally or made publicly available needs to be entered here in ISO 8601 format. If a DOI has been assigned, the convention in the INEXDA network is to use the date of DOI registration.

Example: 2016-08-01

Availability

Providing information on procedures under which data are being made available to data users may help to better understand which datasets they might be able to access and how to access those datasets. Availability procedures are best described by controlled vocabulary, i.e. by a pre-defined list detailing possible procedures which avoids confusion of data users due to the same content being described by different language.

This item lists the procedures under which data is being made available to researchers. Entries allowed are restricted to the following predefined list:

- Download – researchers can directly download data from the website;
- Delivery – researchers can receive the data set or access data from any location via remote access or send codes to the data owner and receive controlled output;
- On-site – researchers have to come to the premises of the data owner to see or work with the data;
- Not available – researchers cannot access / use the data by any means;
- Unknown – other type of availability, please specify in free text.

Example: On-site[5]

Sampled Universe

The item "sampled universe" provides information on elementary units about which inferences are to be drawn and to which analytic results refer when analysing the dataset.

Example: The SHS-Base is a full census (no reporting thresholds apply), i. e. all financial institutions domiciled in Germany report any securities they are holding for domestic and foreign customers. In addition, domestic banks provide information about their own holdings, irrespective of where the securities are held. The financial institutions who are obliged to report comprise domestic banks (monetary financial institutions excluding money market funds), domestic investment companies and "other" domestic financial companies.

The data collection involves holdings of debt securities, shares and investment fund shares or units 3, irrespective where the securities were issued, in what currency they are denominated or if they are listed or not. Only securities which are in circulation and which can be assigned to an investor are included in the Securities Holdings Statistics.

A basic set of information is required to be reported on a security-by-security level. This includes the International Securities Identification Number (ISIN), the nominal amount or number of units held, and the sectoral classification and residency of the holder. For securities quoted as a percentage (eg bonds and debt securities), the

---

[5] Most Bundesbank data are available for non-commercial research via the Research Data and Service Center (https://www.bundesbank.de/Navigation/EN/Bundesbank/Research/RDSC/rdsc.html).

nominal value is stated in the relevant nominal currency and for securities quoted as a number of units (eg equities, mutual fund shares) the number of units held is reported. As the significance of securities repurchase and securities lending transactions has increased strongly in recent years, securities holdings which are passed on or acquired as part of such contracts are to be flagged separately. This information is mandatory for the own holdings of domestic banks only. Since reference month January 2014 the monetary financial institutions have to report the book values of their own securities holdings. Securities that are attributed to the trading portfolio must be labelled.

Sampling

This item describes the type of the sample and sample design used to select the observations to represent the population. As an INEXDA convention, the value of the item will be set to "Total Population" if dataset contains no sampling, i.e. dataset is the population.
Example: Total Population

Temporal Coverage

The item "Temporal Coverage" refers to the sample period of the dataset, which is the time period during which the data was collected or observations were made. The time period should be expressed in ISO 8601 format. As an INEXDA convention, information on month and quarter may be omitted if frequency of the dataset is annual.
Example: 2005-12 to 2016-03

Time Dimension

Describe the time dimension of the data collection. In the INEXDA schema, the following three time dimensions are allowed,

- "Panel" datasets contains information collected from the same (or almost the same) set of entities over time.
- "Time Series" data is collected repeatedly over time to study changes in observations. The "Time Series" category is appropriate for almost all macro datasets, e.g. data on GDP or unemployment rates.
- "Cross-section" dataset are data about a population collected only once.

These items are further broken down by the frequency (e.g. monthly, quarterly) with which data about entities is collected in the dataset.

Furthermore, the following INEXDA conventions apply to this metadata item.

- In case time intervals between consecutive data collections are not equally-spaced (e.g. datasets collected from event-driven reporting), the highest available frequency in the dataset is selected, that allows for a meaningful analysis of the data.

- In case the frequency of data collection changes over time (e.g. because of structural breaks following a change in regulation governing the collection of the dataset) the most recent frequency available in the dataset should be used.

Example: Panel: monthly

Collection mode

This item provides information on the method used to collect the data.
Example: Reporting agents file their reports electronically to Deutsche Bundesbank.

Unit descriptions

Microdata are data on the level of the individual observation. Microdata regularly contain numbers or strings attached to individual observations with the aim to make identification easier. These identifiers may exist on the company level (e.g. country tax number, LEI) or security level (e.g. ISIN). For analysis and research purposes it is often necessary to combine microdata from different sources or policy domains using these identifiers.

The item "Unit Descriptions" provides information about available identifier(s) in the dataset. More than one identifier is allowed, separated by a comma. Identifiers from external data vendors are allowed.
Example: Bank ID, ISIN

Descriptions

The purpose of the metadata item "Descriptions" in the INEXDA schema is twofold.

1. It provides a short description of the dataset.
2. It is used to share information about the scope of structural breaks in the dataset, where structural breaks are defined as major events and revisions that have impacted the dataset.

Designing a metadata schema for microdata needs to employ a procedure documenting significant changes made to a dataset over time stemming from a change in underlying reporting requirement rules. These structural breaks, if they occur, affect metadata items in that the appropriate documentation becomes time dependent. For example, the reporting frequency of a dataset changes in 2012 from quarterly to monthly. In the INEXDA metadata schema the item "Temporal Coverage" will now depend on the time period: before 2012 the appropriate value is quarterly, after the structural break in 2012 the value is monthly.

Besides changes to the time frequency with which data is collected, other examples of structural breaks include:

- Changes to the set of collected variables.
- Changes in the population or sampling.

Example:

The SHS-Base is the core module of the Deutsche Bundesbank's Securities Holdings Statistics (WpInvest). The motive behind creating the SHS-Base is to be able to answer the question: "Who holds what from whom and how much?". Financial institutions domiciled in Germany report securities which they hold for domestic or foreign customers ("custodian-approach"). In addition, domestic banks provide information about their own holdings, irrespective of where the securities are held. Reporting agents are domestic banks (monetary financial institutions excluding money market funds), domestic investment companies and "other" domestic investment companies. The SHS-Base is collected by means of a full census, i.e. every reporting agent has to send a report (if no securities are held in safe custody, a nil report is filed). The

reporting agents provide information on securities holdings broken down by the holder's economic sector and country of origin on a security-by-security basis. The reporting scheme comprises the holdings of debt securities, shares and investment fund shares or units. Holders are classified according to the ESA sectoral classification scheme and the amounts held in safe custody are transmitted to the Deutsche Bundesbank. Securities data are indispensable for monetary analysis as any shifts in financing between the banking system and the securities markets may affect the transmission of monetary policy. In addition, data are essential to monitor the development of amounts held and the distribution among different (groups of) investors. Financial stability analysis makes use of the data to measure the risks associated with different types of instruments and exposures to individual counterparties. Information on the composition and quality of the portfolios of holders is used to better understand investor behavior.

• *From 200512 to 201212: end-of-quarter collection*

• *From 201301 to 201603: end-of-month collection*

Geographical Coverage

The item "Geographical Coverage" provides information on the region where the data was gathered or on which the data is focused.
Example: DE (for Germany)

Keywords

Please enter the keywords describing the general content of the dataset. You may choose up to 10 keywords. Entries allowed are restricted to a predefined list.
Example: Banks, Securities, Financial markets, Debt securities, Equity and investment fund shares or units, Financial derivatives

Alternative identifier

This metadata item is used to indicate that the information about the dataset provided here belongs to the metadata collection of INEXDA.
Example: INEXDA

Relations

The item "Relations" will be used in the future to describe relations between datasets and databases in the INEXDA metadata database. Examples of relations will include:

- Different versions of a dataset.
- Relation between dataset and database (in a given country)
- Relation between datasets containing similar units (in different countries)
- Dataset feeds into an ECB dataset

Publications

Allowing descriptions of publications be associated with related datasets provides additional information which is complementary to existing metadata. Publications include information on how a dataset is being used (e.g. variable transformation), what a dataset is being used for (e.g. topics or methodology), and who has used the

dataset. All this information is useful to help data user better understand and discover datasets relevant for their research.

Therefore, the INEXDA metadata schema contains an item that provides information on scientific publications relating to the registered dataset. Publications listed here may also include descriptions of datasets (e.g. technical reports, data reports, or user guides) which are publicly available.

Example: Bade, M., Flory, J. and T. Schönberg (2016). SHS-Base, Data Report 2016-02 - Metadata Version 1-1. Deutsche Bundesbank Data and Service Centre.

3. Conclusion

The relevance of metadata also becomes clear in connection with the re-use and corresponding citation of the data. Bibliometrical methods, which should make the performance of data production measurable similar to the impact factors in text publications, are based on metadata. In addition to the bibliographic and content information, these metadata should also include a persistent identifier (so called PID) that enables the identification and localization of the used data uniquely and permanently. There are now a number of services that offer the allocation of PIDs in the form of Digital Object Identifiers (DOI) for research data. The German Leibniz Institutes GESIS and ZBW are offering a special service for the assignment of DOI names for social and economic data.

References

Day, M., (2005). Metadata, DCC Digital Curation Manual. In: S. Ross, M. Day (eds), Retrieved 10 July 2018, from http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata

Harzenetter, K., Borschewski, K.: da|ra: Solutions to the Challenges of Data Registration, Access & Exchange. CNI Spring 2016 Membership Meeting; San Antonio (Texas); April 4-5, 2016. Retrieved 10 July 2018, from https://www.cni.org/topics/information-access-retrieval/dara-solutions-to-the-challenges-of-data-registration-access-and-exchange

Jensen, U., Wasner, C., Zenk-Möltgen, W. (2018): Metadatenstandards im Kontext sozialwissenschaftlicher Daten. In: U. Jensen, S. Netscher, K. Weller (eds.), Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten. Opladen. (in print)

Koch, Ute and Akdeniz, Esra and Meichsner, Jana and Hausstein, Brigitte and Harzenetter, Karoline: da|ra Metadata Schema: Documentation for the Publication and Citation of Social and Economic Data. Version 4.0. GESIS Papers 2017/25.)

Recker, A.: Was sind Daten | Forschungsdaten | Metadaten? Forschungsdatenmanagement in den Geowissenschaften. Vortrag. Berlin, 3-4 July 2014. Retrieved 10 July 2018, from

Unrestricted

https://www.slideshare.net/GESISarchivetraining/was-sind-daten-forschungsdaten-metadaten

Wallace, D. (2001). Archiving metadata forum: report from the Recordkeeping Metadata Working Meeting, June 2000. Archival Science, 1(3), p 253-269.

An introduction to INEXDA's metadata schema

An introduction to INEXDA's metadata schema[1]

Stefan Bender, Brigitte Hausstein and Christian Hirsch,
Deutsche Bundesbank

---

[1] This presentation was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

INEXDA
The Granular Data Network

1

Introduction to INEXDA's Metadata Schema

Christian Hirsch

(Deutsche Bundesbank)

July 2018

The views expressed here do not necessarily reflect the opinion of the Deutsche Bundesbank, the INEXDA network, or the Eurosystem.

Motivation

- Metadata is data about data, i.e. a description of data

- Examples of existing metadata schema
  - Statistical Data and Metadata Exchange (SDMX)
  - Data Documentation Initiative (DDI)
  - da|ra Metadata Schema

- Important is the purpose
  1. Overview of available and possible comparable granular datasets
  2. Help data users discover and use datasets appropriate for research and analysis by using harmonised metadata.
  3. Provide framework to facilitate a possible harmonisation of datasets in the (near) future.

INEXDA
The Granular Data Network

INEXDA Metadata Database

- The database serves
    1. as an information tool for INEXDA members, researchers and analysts
    2. as the basis for the harmonisation activities of INEXDA (e.g. item relation)

- Current contributions to the database come from

    - Deutsche Bundesbank   30 (12)
    - European Central Bank   (6)
    - Banca d'Italia   26 (6)
    - Banco de España   21 (2)
    - Banco de Portugal   16 (2)

INEXDA The Granular Data Network

Part 1: Basic Information

1	Resource Type
2	Resource Identifier
3	Name of Dataset
4	Creator
5	DOI Proposal
6	URL
7	Language of Resource
8	Publication Date
9	Availability
10	Sampled Universe
11	Sampling
12	Temporal Coverage
13	Time Dimension
14	Collection Mode
15	Unit Descriptions
16	Descriptions
17	Geographical Coverage
18	Keywords
19	Alternative Identifiers
20	Relations
21	Publications

- *Creator* is a mandatory item in da|ra. May be used to provide more granular information on the data compiler

- *URL* refers to the webpage which displays information about the dataset

- *Availability (controlled)* describes the procedure under which the data can be accessed (eg download or on-site)

- *DOI Proposal* provides the suggested DOI name of the dataset. A Digital Object Identifier (DOI) is a permanent, persistent identifier used for citing and tracking datasets

INEXDA
The Granular Data Network

Part 2: Methods

1	Resource Type
2	Resource Identifier
3	Name of Dataset
4	Creator
5	DOI Proposal
6	URL
7	Language of Resource
8	Publication Date
9	Availability
10	Sampled Universe
11	Sampling
12	Temporal Coverage
13	Time Dimension
14	Collection Mode
15	Unit Descriptions
16	Descriptions
17	Geographical Coverage
18	Keywords
19	Alternative Identifiers
20	Relations
21	Publications

- *Sampling* displays the type of sample design used to select the observations to present the population

- *Time Dimension* provides information on
  - frequency of observations.
  - whether dataset structure is panel, time-series or cross-sectional

- Structural breaks are defined as major events and revisions that have impacted the dataset

- Examples of structural breaks include:
  - changes to the time frequency with which data is collected
  - changes to the set of collected variables
  - changes in the population or sampling

INEXDA
The Granular Data Network

Part 3: Descriptions

1	Resource Type
2	Resource Identifier
3	Name of Dataset
4	Creator
5	DOI Proposal
6	URL
7	Language of Resource
8	Publication Date
9	Availability
10	Sampled Universe
11	Sampling
12	Temporal Coverage
13	Time Dimension
14	Collection Mode
15	Unit Descriptions
16	Descriptions
17	Geographical Coverage
18	Keywords
19	Alternative Identifiers
20	Relations
21	Publications

- *Unit Description* provides information on the entities that are being observed in the dataset

-

- Datasets may contain more than one unit of observation. For example, in a credit register information on the following units are collected:
    - Banks
    - Companies
    - Governments
    - Loans

- *Descriptions* also contains detailed information on structural breaks in the dataset

INEXDA
The Granular Data Network

Part 4: Relations and Publications

1	Resource Type
2	Resource Identifier
3	Name of Dataset
4	Creator
5	DOI Proposal
6	URL
7	Language of Resource
8	Publication Date
9	Availability
10	Sampled Universe
11	Sampling
12	Temporal Coverage
13	Time Dimension
14	Collection Mode
15	Unit Descriptions
16	Descriptions
17	Geographical Coverage
18	Keywords
19	Alternative Identifiers
20	Relations
21	Publications

- Describes relations between datasets and databases in the INEXDA metadata database...
  - ... in a given country
  - ... across countries

- Examples of use cases in INEXDA context
  1. Relation between datasets containing similar units (in different countries).
  2. Dataset feeds into a ECB dataset.

- *Publications* provides information on scientific publications related to the dataset.

INEXDA
The Granular Data Network

Conclusion

- INEXDA's metadata schema is based on the GESIS DOI registration service da|ra (GESIS is cooperating with DataCite)

  https://www.da-ra.de/en/home

- Purpose is to foster harmonisation between INEXDA members and broaden metadata sharing within INEXDA and possibly outside

- Name of metadata items closely follows da|ra conventions to enable seamless DOI registration, if desired later in the project

- Using DOI for datasets may help
  - … facilitate datasets citations
  - … foster reproducibility of results

INEXDA
The Granular Data Network