

11th Biennial IFC Conference on “Post-pandemic landscape for central bank statistics”

BIS Basel, 25-26 August 2022

Update on the work of the International Network for Exchanging Experiences on Statistical Handling of Granular Data (INEXDA)¹

Stefan Bender, Jannick Blaschke and Christian Hirsch,
Deutsche Bundesbank

¹ This presentation was prepared for the conference. 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 event.

Update on the work of the International Network for Exchanging Experiences on Statistical Handling of Granular Data (INEXDA)

Prepared by the chair of INEXDA¹

Abstract

INEXDA provides a platform for administrative data producers to exchange practical experiences on the accessibility of granular data, metadata, and data protection techniques. In this paper, we provide an update on the progress that INEXDA has made since 2018. This includes creating a set of machine-readable metadata items describing access procedures and conducting a survey of INEXDA members to take stock of technical, administrative, and organisational features of procedures for accessing granular data for research purposes. The results provide a comprehensive picture of existing procedures and may serve to provide insight for institutions on the cusp of setting up similar facilities.

Keywords: INEXDA, data sharing, research data centre, data access, research, granular data, exchange

Content

1. Introduction	2
2. Working Group on Data Access.....	3
2.1 Workshop and the need for unified terminology	3
2.2 The resulting metadata schema – the annodata schema	4
3. INEXDA Survey on Access Modes	5
3.1 INEXDA Working Group on Modes of Data Provision	5
3.2 Main results on access modes	6
3.3 Main results on provided data.....	7
3.4 Main results on organisational and technical access modalities	7
3.5 Main results on output checking	8

¹ As of July 2022 Stefan Bender of Bundesbank is the current chair of INEXDA. Jannick Blaschke and Christian Hirsch, both Bundesbank, contributed to this report. The authors would like to thank our colleagues of the INEXDA network, particularly those who participated in the working groups, for their valuable suggestions and feedback. All views expressed in this report are personal views of the authors and do not necessarily reflect the views of INEXDA and its members, Deutsche Bundesbank or the Eurosystem.

4. Conclusion and future plans	8
References	9

1. Introduction

This paper provides an update on the International Network for Exchanging Experiences on Statistical Handling of Granular Data (INEXDA). We presented the network in a similar paper in 2018.² In this paper, we would like to describe the work that has been accomplished since 2018. The network has seen its membership increase in both number³ and diversity, as it now includes members from outside of Europe as well as beyond national statistical institutes.

INEXDA was founded to provide a platform for interested parties to exchange their experiences in the statistical handling of granular data. The idea originated from the global financial crisis of 2008-09, which showed that certain problems may only be detected using very granular data on current economic developments. The network was originally set up to promote the G20 Data Gaps Initiative II, particularly Recommendation 20, which addressed the accessibility of granular data. Moreover, INEXDA acknowledges and supports the work on data sharing by the Irving Fisher Committee on Central Bank Statistics.

The idea of the network has always been to provide a platform for exchange. However, we also wanted to provide hands-on practical solutions intended for use in institutions that provide access to granular data or are planning to do so. One example of these solutions already existing in 2018 was the INEXDA metadata schema for describing granular data (Bender, Hausstein, and Hirsch, 2019). This schema builds on an existing metadata standard data documentation initiative (DDI)⁴ and has been modified to better accommodate the particularities of granular data.

However, are there any differences in describing macro data and aggregated data? While there are indeed differences, it is possible to take metadata designed for macrodata as a reasonable starting point. One difference is that some metadata items become more important when moving from granular to (aggregated) macro data. For granular data, for example, metadata items describing identifiers are very important, as granular data usage typically requires multiple datasets to be combined for analysis.

Due to its increasing member size, INEXDA uses specialised working groups to advance its agenda. We will therefore describe recent developments largely on the basis of the output of these working groups. First is the Working Group on Access

² Members of INEXDA (2018). INEXDA – the Granular Data Network, Irving Fisher Committee Working Paper No 18: https://www.bis.org/ifc/publ/ifcb49_31.pdf

³ As of July 2022, INEXDA members include, in alphabetical order, the Banca d'Italia, Banco Central de Chile, Banco de España, Banco de México, Banco de Portugal, Bank of England, Bank of Russia, Banque de France, Central Bank of the Republic of Turkey, Deutsche Bundesbank, European Central Bank, Eurostat, and the Office of National Statistics (ONS). In addition, the Bank for International Settlements participates as a permanent guest.

⁴ For more information on DDI see <https://ddialliance.org/>

Modes. Access modes define the ways in which access to data is provided, which could, for example, be achieved via download or on-site access at the premises of the institution providing the data.

Providing access to granular data also involves facing significant legal and technical challenges related to, amongst other aspects, safeguarding statistical confidentiality. As a result, INEXDA decided to organise a workshop bringing together senior experts from central banks, national statistical institutes and international organisations to share their experiences on real-world approaches to facilitating access to granular data for analytical purposes. This workshop served as a starting point for the group's work.

Building on the outcome of the workshop and the output of the working group, INEXDA released a second metadata schema aimed at describing access procedures in a machine-readable way (INEXDA Working Group on Data Access, 2020). This schema, known as the "annodata schema", works in much the same way as metadata schemas that describe data, but has instead been designed to describe the procedures for accessing the data. The annodata schema is thus both compatible with and complementary to the INEXDA metadata schema described above.

Building on the annodata schema a working group conducted a survey among INEXDA members on currently available modes of data provision for external scientific researchers. The focus of the survey was to generate insights on the practical approaches employed by each participating institution. While generating useful information on the many particularities this very detailed approach also hampered the comparison of the responses of individual members.

This paper is organised as follows: Section 2 describes the annodata schema in more detail. We put special emphasis on access modes, which represent the smallest set of information needed to describe access to granular data. Section 3 describes the results of a survey of INEXDA members on the access modes used by member organisations to provide access to granular data for scientific research purposes. Section 4 presents the plans for the coming years and concludes.

2. Working Group on Data Access

2.1 Workshop and the need for unified terminology

Providing access to granular data for scientific research and analytical purposes requires developing a set of procedures that define how authorised personnel ought to use the data. Spanning a project's life cycle from request to publication, these procedures must ensure ongoing compliance with the legal framework and also satisfy the technical, organisational, and administrative requirements set out by the data provider. The INEXDA Working Group on Data Access focused on (i) taking stock and (ii) identifying common features and differences among these procedures.

As one of its first activities after being established, the group organised a workshop aimed at bringing together experts from central banks, national statistical institutes and international organisations to share their experiences on real-world approaches to providing access to granular data. Presentations revolved around the topics of (i) "data discovery centres" (i.e. information on where to find the data), (ii)

“data access centres” (i.e. how to grant access to the data), and (iii) “data hubs” (i.e. access to multi-source remote data).

As a result of the workshop, it became clear that the working group could not immediately start tackling its original task of taking stock of and examining existing access procedures, as it still had to complete some preliminary work. This included, first, overcoming the challenges posed by differences in legal, organisational and technical conditions as well as different terminology that complicated the discussion on data access procedures. The working group therefore started by defining unambiguous terminology, including definitions for “access protocol”, “access mode” and “access regime”.

Second, the working group’s tasks required documenting and processing a wealth of institutional and data-specific requirements immanent in procedures for accessing granular data. Drawing on existing metadata schemas for this task proved insufficient, as they typically focus on the question of “How are the data produced?”⁵ rather than help to answer the question of “How is access to the data managed?”. The working group therefore used its newly established common taxonomy to describe the set of legal, technical, and organisational rules that make up the procedures for accessing granular data.⁶

To our knowledge, this is the first machine-readable description of data access procedures and it stands in clear contrast to the predominantly free-text descriptions used in existing metadata schemas. This new metadata schema is therefore not intended to replace any existing metadata schemas, but rather to complement them with well-structured information on data access procedures.

2.2 The resulting metadata schema – the annodata schema

This metadata schema, which INEXDA has named the “annodata schema” (derived from “annotation to data”), comprises three interrelated dimensions. The first dimension covers the dataset aspect of access procedures.⁷ This dimension collects all information needed to unambiguously identify a dataset and its origins, and connects datasets to access modes (e.g. on-site access, remote execution). The second dimension contains all information regarding the aggregation of granular data, which is especially important, as access to granular data generally involves combining different datasets. Finally, the third dimension comprises information on users and projects.

A pivotal concept in the annodata schema is the “access mode”. An access mode connects inputs to outputs. In the example below, a function with three inputs produces one output. The inputs are:

1. “Who?” – Information on the user (e.g. internal or external to the organisation)
2. “What?” – Information on the dataset (e.g. anonymised or perturbed)

⁵ See Bender et al. (2021) for a more thorough discussion on this point.

⁶ For more information on the motivation and rationale behind the annodata schema, see the final report of the Working Group (INEXDA Working Group on Data Access, 2020).

⁷ A complete version of the annodata schema, including a description of all metadata items, can be found in the final report of the Working Group (INEXDA Working Group on Data Access, 2020).

3. “Why?” – Information on the purpose of use (e.g. scientific research)

As the output, these three inputs determine “how” a specific user requesting access to a specific type of dataset for a specific type of use can access the data. Examples of “how” could include “secure on-site access” in a research data centre or “remote access”. Of course, the inputs could also produce a nil return as their output if no corresponding access mode exists.

The decision regarding which inputs lead to which output comes directly from the legal framework or its interpretation by the institution providing the data. For example, the legal framework could state that external researchers may use anonymised data for scientific research if they access the data within a secure environment. Here, the contribution of the annodata schema is to convert these rules from text form into machine-readable metadata.

With the common terminology set out, the Working Group on Data Access provided the basis for future discussions within INEXDA and facilitated the survey on access modes that was conducted by the subsequent working group. The next section presents the main results of this survey.

3. INEXDA Survey on Access Modes

3.1 INEXDA Working Group on Modes of Data Provision

In 2020, INEXDA launched its third working group, which focused on “Modes of Data Provision”. The group was chaired by the Banque de France and its main objective was to conduct a survey of INEXDA members on the existing modes of data provision for external scientific researchers. The survey results would then allow common features and differences with regard to access procedures amongst INEXDA members to be identified. The survey was conducted in 2020 and 2021 with the participation of eight countries⁸ as well as both the European Central Bank and Eurostat.

The survey builds on the INEXDA annodata schema and the terminology established therein. For this reason, the survey itself was conducted at the level of access modes. The overarching goal of the survey was to obtain insights into the practical approaches employed by each participating institution. The focus on access modes provided the detailed information needed to draw conclusions from the responses of individual members.

This very detailed approach reflected all particularities of the practical approaches to access to granular data stemming from the legal, organisational and technical frameworks. However, while useful for obtaining insights, these particularities also make it more difficult to compare the responses of individual members. For example, differences in the legal framework might lead to different interpretations of an access mode. This section therefore focuses on presenting the main results and lessons learned from the survey. In the future, INEXDA will conduct an updated survey to capture the effects of the COVID-19 pandemic.

⁸ The Deutsche Bundesbank (Germany), Banco de España (Spain), Banque de France (France), Banca d'Italia (Italy), Banco de México (Mexico), Banco de Portugal (Portugal), Bank of Russia (Russia), and Central Bank of the Republic of Turkey (Turkey).

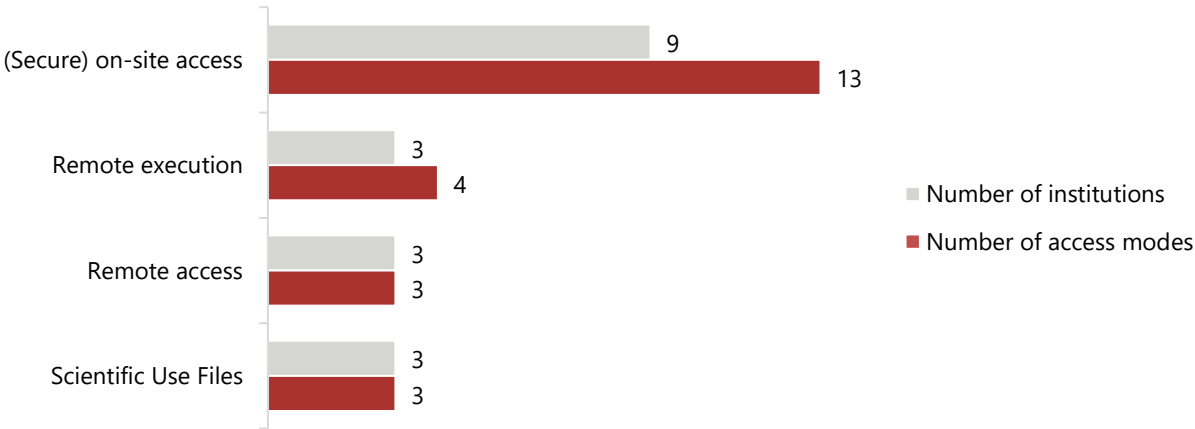
3.2 Main results on access modes

The survey found that most respondents provided access to data in several ways, often combining the provision of Public Use Files (PUFs) with on-site solutions. Among the ten participating institutions, eight offered multiple ways to access data.⁹ Access modes differed with respect to the available dataset or degree of anonymisation, with formal anonymisation being the most commonly used method. However, all institutions granted access to their data for scientific research in some form.

Distribution of access modes

Number of institutions and access modes per access mode type*

Figure 1



* Ten INEXDA members participated with 1-n possible access modes each. Public Use Files were excluded from this figure.

Source: INEXDA Working Group on Modes of Data Provision.

As Figure 1 shows, on-site access and secure on-site access, where researchers work in a secure environment on the premises of the institution providing the data, were the most frequent types of access. Three institutions offered remote execution, remote access and Scientific Use Files.

The number of researchers per access mode varied between institutions, but, in general, more researchers accessed data through the two download modes, although the actual usage of Public Use Files is hard to determine. Across institutions and access modes, the average number of researchers per project was between one and two. Access was granted primarily to researchers affiliated to universities or research institutions, while financial and commercial firms were able to access data solely through PUFs.

⁹ **Secure on-site access:** Provision on the premises of the institution in a dedicated secure environment (data room). **Remote execution:** Researcher sends a code and the RDC sends back results. **Remote access:** Researcher can access data stored in another location remotely from their own institution in a controlled environment. **Download:** Files are available for download and copy and are used according to the terms of use (e.g. Scientific Use Files or Public Use Files).

3.3 Main results on provided data

As nine of the ten respondents of the survey were central banks, most of the data made available to researchers were on the topic of "Economy and finance", followed by "Industry, trade and services". This is depicted in Figure 2. Within the "Economy and finance" topic area, access was most often granted to data on "Firms and banks".

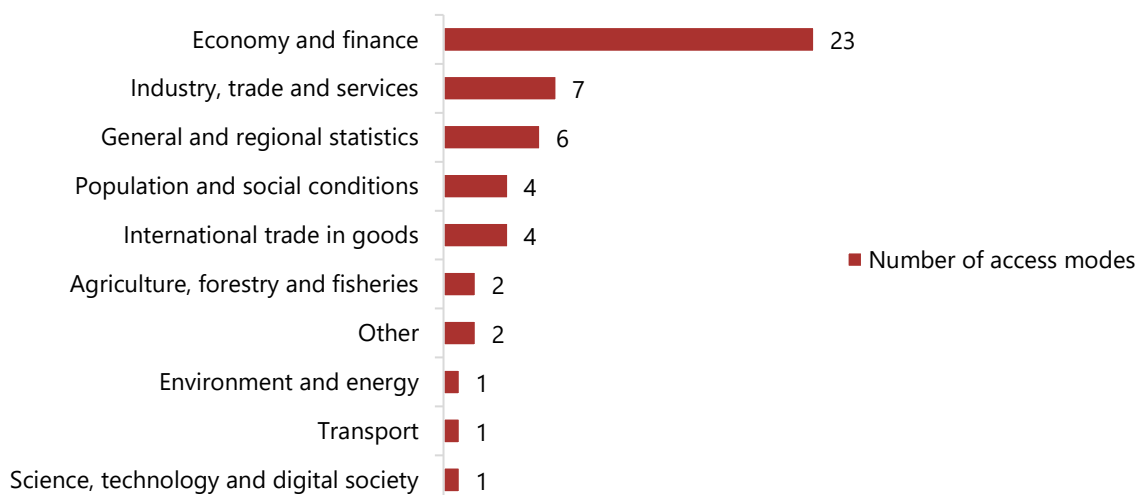
The majority of these datasets were provided in standardised form, so they were identical across research projects. The two biggest advantages of standardised datasets as compared to customised datasets are the economy of resources when preparing the data and the significantly easier reproducibility of research results. The latter is achieved, in particular, by the options for tailoring metadata and access rules perfectly to the dataset and uniquely identifying and citing the datasets with a Digital Object Identifier (DOI). However, some datasets were also provided in customised form, for example due to legal constraints.

When asked about the handling of external data brought in by the researchers themselves, a predominantly uniform approach could be observed. External researchers were usually allowed to bring their own data, but these were checked by the institution providing the data before being made available in the project. One respondent even offered automatic import of external data.

Data source by topic

Number access modes per topic*

Figure 2



* Ten INEXDA members participated with 1-n possible access modes each.

Source: INEXDA Working Group on Modes of Data Provision.

3.4 Main results on organisational and technical access modalities

The survey showed that there were different approaches to the governance of data for research access. Depending on the institution, different decision bodies were involved in the process and there were a variety of administrative measures to ensure

that no confidential data was ever disclosed. All respondents declared that, for all data other than Public Use Files, researchers needed to sign some form of confidentiality agreement.

Since the technical environment for data access is very much related to the respective access mode, the working group decided to limit the scope at this point to on-site access. Currently, data access is granted mainly via file systems, although some institutions have already switched to more advanced systems, such as dedicated servers or Hadoop platforms. Researchers can usually access the data via specific workstations or virtual machines. Regarding the available software, researchers can mostly choose between Stata, Python, R, and, in some cases, also Matlab and SAS.

3.5 Main results on output checking

All respondents of the survey agreed that, for data that is not accessed via download, output control is mandatory. This is usually performed by the researchers themselves and then checked by the institution providing the data. Only results that are compliant with the output rules can be taken out of the institution's environment. The output check is usually carried out manually with some semi-automated tools providing assistance. These tools may, for example, calculate whether the release of a specific result, such as a summary table or regression, would lead to a disclosure problem.

4. Conclusion and future plans

This paper provides an update on the work that INEXDA has accomplished. Since 2018, INEXDA has experienced growth in its number of members, which has led to the composition of its membership becoming more diverse. The network devised the annodata schema, a machine-readable set of items for describing data access procedures intended to complement existing metadata schemas that predominantly utilise free-text descriptions. Finally, a survey conducted among members provided information on the many particularities in practical approaches to granular data access.

For the future, INEXDA plans to use the outcomes of the annodata schema and the survey to discuss possible potential for harmonisation with regard to access procedures within INEXDA. In addition, the results facilitate the development of common best practices for data sharing and the handling of granular data. This would be extremely helpful for any institution planning to establish a new access mode, for example, and looking for reference. Of course, these considerations need to acknowledge the limits set by the prevailing legal, technical and procedural frameworks in which INEXDA members operate.

Finally, INEXDA will continue supporting international initiatives aimed at strengthening micro data access. For instance, INEXDA will support the new G20 Data Gaps Initiative, for which the IMF, the Financial Stability Board and the Inter-agency Group on Economic and Financial Statistics are currently developing a workplan.

References

Bender, S., B. Hausstein, and C. Hirsch (2019). An introduction to INEXDA's metadata schema, Proceedings from the Ninth IFC Conference in Basel, Switzerland, 30-31 August 2018.

Bender, S., J. Blaschke, H. Doll, C. Hirsch, and F. Ritchie (2021). The Annodata Framework: Putting FAIR data into practice. Deutsche Bundesbank, Research Data and Service Centre.

INEXDA Working Group on Data Access (2020): Final Report, https://www.inexda.org/wp-content/uploads/2020/05/FINAL_INEXDA_WG_DA_Final_Report.pdf

Members of INEXDA (2018). INEXDA – the Granular Data Network, Irving Fisher Committee Working Paper No 18.



Update on the work of the International Network for Exchanging Experiences on Statistical Handling of Granular Data (**INEXDA**)

Stefan Bender

INEXDA chair, Deutsche Bundesbank

Christian Hirsch

INEXDA secretary, Deutsche Bundesbank



International **N**etwork of **E**xchanging Experiences
on Statistical Handling of Granular **D**ata

General mission

- General mission is to **promote data sharing and data access**
- Promoting the **G20 Data Gaps Initiative II**, in particular recommendation 20, addressing the accessibility of granular data. INEXDA is mentioned in a G20 paper
- Acknowledging and supporting the work on data sharing of the **Irving Fisher Committee** on Central Bank Statistics

Organisation

Current chair of INEXDA is Stefan Bender of the Deutsche Bundesbank.

Current members



BANCO
CENTRAL
DE CHILE



DEUTSCHE
BUNDESBANK
EUROSYSTEM



BANK OF ENGLAND



Office for
National Statistics

BANCO DE ESPAÑA
Eurosistema



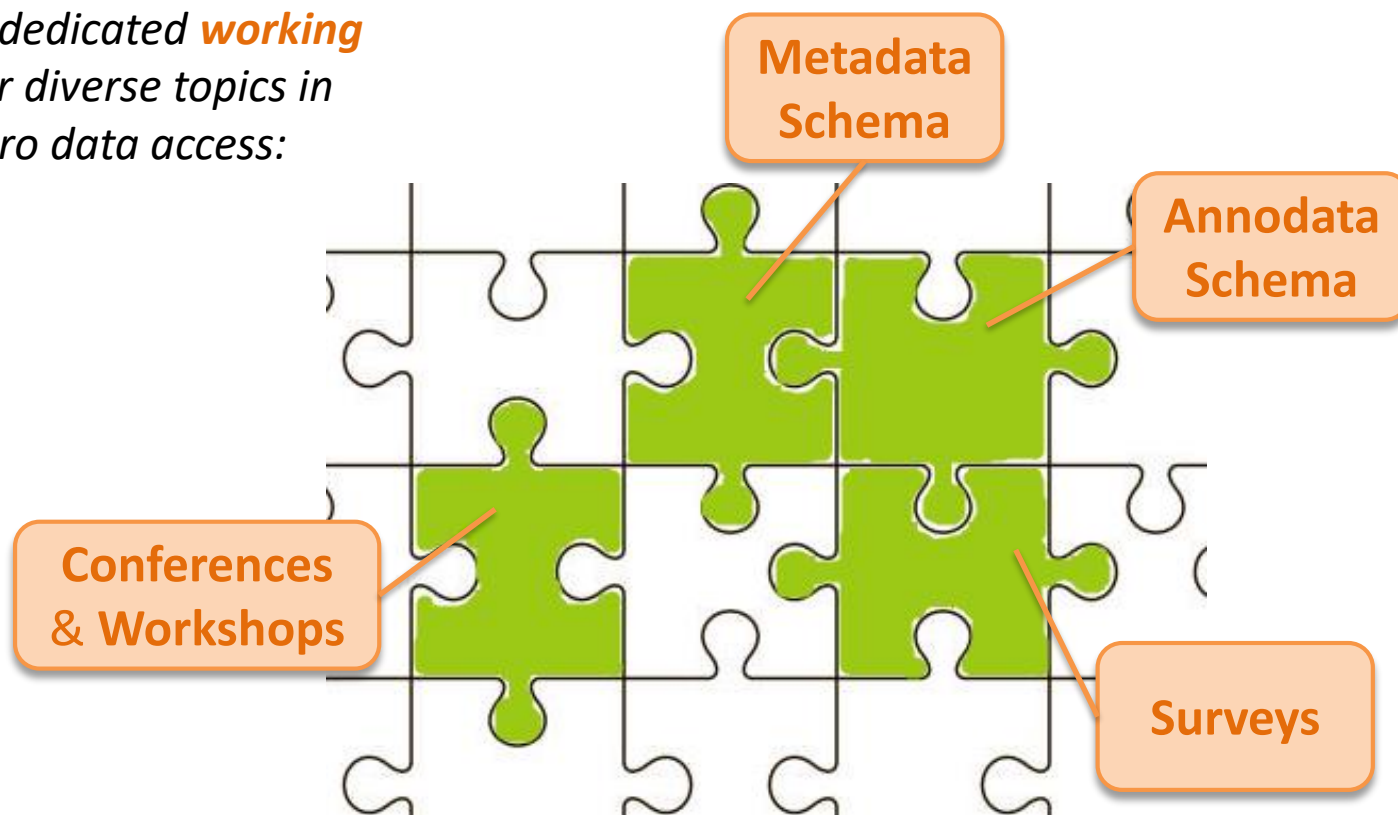
eurostat



INEXDA's working arrangement

Micro data access is a multifaceted field. INEXDA's focus is on **developing**, **applying** and **disseminating** new ideas.

For this it uses dedicated **working groups** to cover diverse topics in the field of micro data access:



Eight types of annodata

1

**Access regime**

2

**Database**

3

**Dataset family**

4

**Record linkage**

5

**Combining restrictions**

6

**Global rules**

7

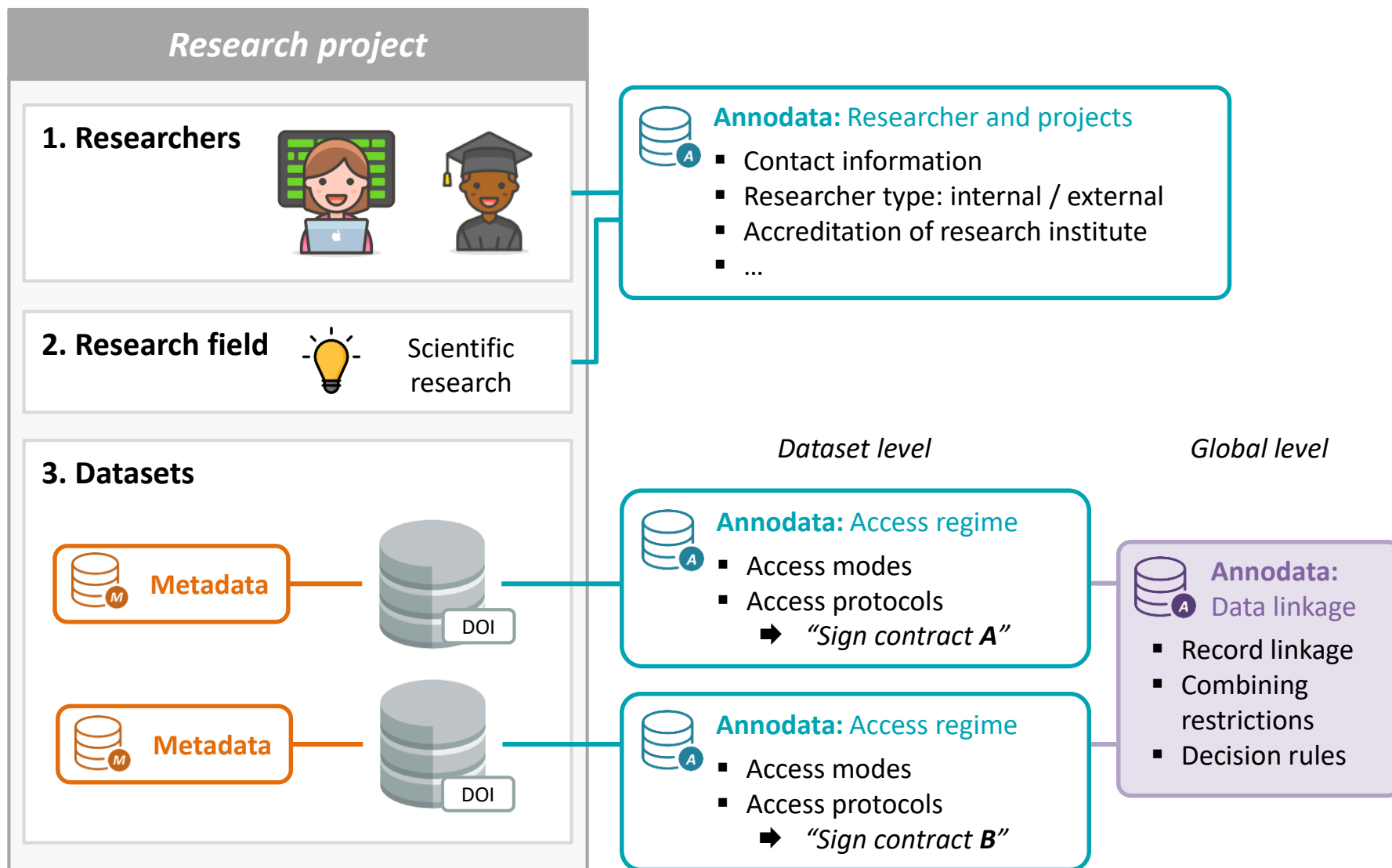
**Researchers**

8

**Research projects**

1-3 at *dataset family* level
4-6 at *global* level
(i.e. irrespective of researcher
affiliation, research field, and
access mode)
7-8 at *project* level

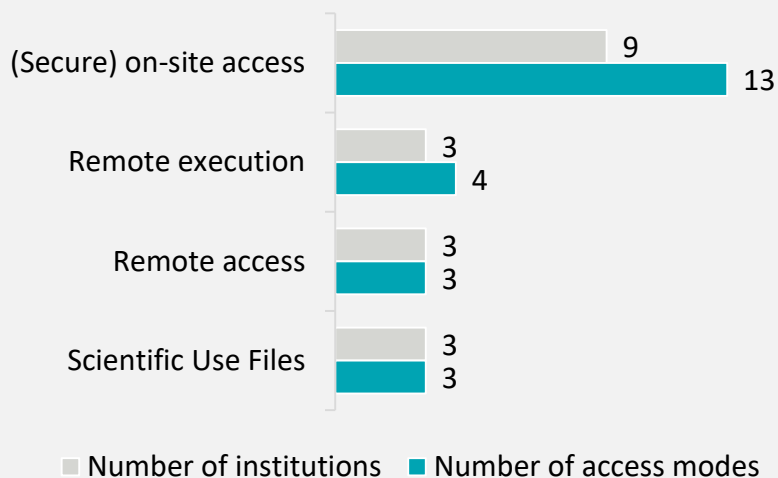
How micro data access could be automated



Working Group on “Modes of Data Provision”

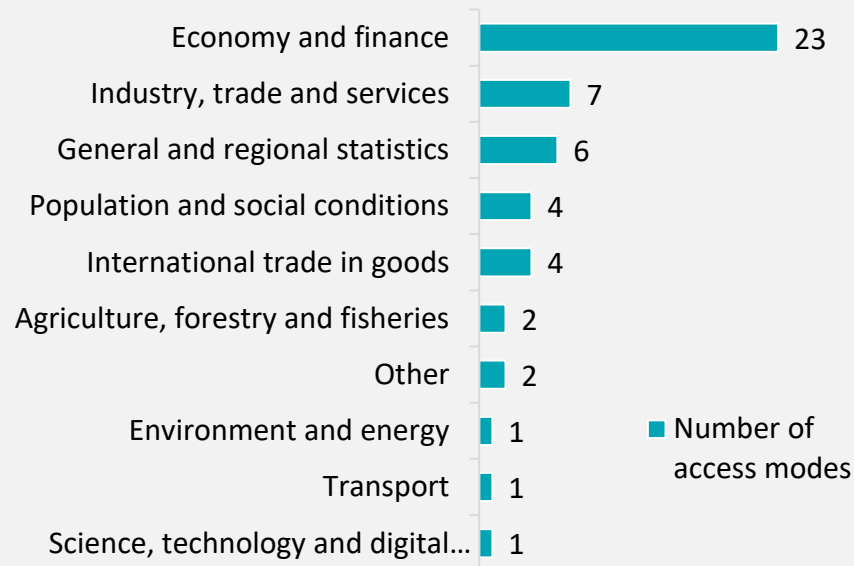
*The working group, which was chaired by **Banque de France**, built on the **annodata** schema and conducted a survey on the very granular level of **access modes**.*

Distribution of Access Modes



* 10 INEXDA members participated with 1-n possible access modes each. Public Use Files were excluded from this figure.

Data source by topic



* 10 INEXDA members participated with 1-n possible access modes each.

Conclusion and future plans



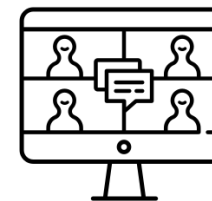
Metadata schema



Annodata schema



Survey



Conferences

In the past, INEXDA worked on data access procedures (e.g. annodata schema, survey).

For the future, INEXDA plans to

- launch **new working groups** on anonymization and output control.
- use the outcomes of the working groups to develop **common best practices for data sharing** and the handling of granular data. This would be extremely helpful for any institution, which plans to establish e.g. a new access mode and is looking for reference.



These considerations obviously need to **acknowledge the limits** set by extant legal, technical or procedural frameworks in which INEXDA members operate.



**Thank you for your
attention!**

Website:

inexda.org/

Contact at INEXDA:

Chair
Secretary

Stefan Bender, Deutsche Bundesbank
Inexda@Bundesbank.de



How to become an INEXDA member

Other central banks, national statistical institutes and international organizations are **encouraged to join INEXDA**.

*The following **procedure** has been established to admit new members:*

	A representative of the applying institution has to attend at least one INEXDA meeting in person before the formal application process is initiated.
	Send an application letter signed by the head of the statistical department of the respective institution (or, in case of national statistical institutes, by the head of the responsible statistical department) to the chair of INEXDA.
	All INEXDA members must agree to any application by a new institution.
	Sign the MoU*. Congratulations, you are now a member of INEXDA.

* The INEXDA Memorandum of Understanding (MoU) can be accessed [here](#).