Publishing Fine-Grained Standardized Metadata: Lessons Learned from Three Research Data Centers

PRACTICE PAPER

KNUT WENZIG
ANDREAS DANIEL
DOMINIQUE HANSEN
TOBIAS KOBERG
MIHAELA TUDOSE

*Author affiliations can be found in the back matter of this article

ABSTRACT

FAIRness of research data, meaning that data are managed according to the principles of being Findable, Accessible, Interoperable, and Reusable, has become a ubiquitous requirement in research data policies as well as in general guidelines for research data management. Meeting this requirement largely depends on the availability of rich and standardized DDI-metadata—based on the Data Documentation Initiative family of metadata standards—which is of particular importance for tabular data resulting from surveys and other structured observations, is often lacking (Wenzig and Han, 2024). The lack of such metadata can largely be attributed to the absence of lightweight approaches that integrate its creation into existing data preparation workflows. Against this background, a project funded by KonsortSWD-NFDI4Society brought together three research data centers (SOEP, LIfBi, and FDZ-DZHW) to investigate the requirements for converting existing metadata into a standardized DDI format and publishing it using a common protocol (OAI-PMH). The results demonstrate that generating fine-grained DDI-metadata is easy to implement, even with limited resources. Contrary to expectations, OAI-PMH did not prove to be a straightforward approach for publishing metadata. Based on these findings, the authors evaluate FAIR signposting as an alternative approach, which shows considerable potential. The results of this study may therefore serve as a best-practice example for institutions, especially from survey-based research domains seeking to implement fine-grained standardized DDI metadata, as well as a starting point for further research on approaches to publishing fine-grained standardized metadata.

CORRESPONDING AUTHOR:
Knut Wenzig
SOEP, German Institute for Economic Research (DIW) Berlin, Germany
kwenzig@diw.de

KEYWORDS:
Metadata; DDI-Codebook; OAI-PMH; Metadata publication; FAIR Data Principles

TO CITE THIS ARTICLE:
Wenzig, K., Daniel, A., Hansen, D., Koberg, T. and Tudose, M. (2026) 'Publishing Fine-Grained Standardized Metadata: Lessons Learned from Three Research Data Centers', *Data Science Journal*, 25: 13, pp. 1–13. DOI: https://doi.org/10.5334/dsj-2026-013

1. MOTIVATION

The FAIR principles (Wilkinson *et al.*, 2016) are becoming increasingly important for the policies of research institutes that collect and provide research data (European Commission, 2018; for an overview of policy elements that enable FAIRness see Davidson *et al.*, 2022).[1] However, the focus on the FAIR principles should not remain a merely latent policy requirement; instead, it must be operationalized at a level that is meaningful for day-to-day practice with real benefits. Among the various components required for FAIR data provision, the FAIRness of metadata is particularly important, as metadata are essential for finding, understanding, and using the data (see Davidson *et al.*, 2022, pp. 5–6).[2] The importance of high-quality metadata (for both humans and machines) is also emphasized by the Global Cooperation on the FAIR Data Policy and Practice in its policy recommendations:

> "*Sufficiently detailed, Standardised and Interoperable Metadata: The sine qua non to greater automation of cross-domain data combination and analysis and fine-grained and responsive access control is sufficiently detailed, standardised and interoperable metadata. There are no short cuts: data and metadata are hard. Support is necessary for the development, adoption and implementation of standards and for the increased use of tools for automated metadata management.*" (Hodson and Gregory, 2023, p. 12)

Against this background, this article focuses on the standardization and standardized provision of fine-grained metadata for tabular data, typically produced by surveys or other standardized observational methods. By describing data with metadata at the variable level, the article aims to improve overall FAIRness and to facilitate reuse across disciplines such as the social sciences, economics, psychology, public health, and education research. In this article, we focus on the DDI-standard[3] as it represents the major standard for documenting surveys and other standardized observational data, particularly with respect to fine-grained variable metadata. Other standards, such as Statistical Data and Metadata eXchange (SDMX), schema.org, Data Catalog Vocabulary (DCAT), or DataCite have notable limitations in this context (see Section 5 for a more detailed discussion).

While metadata at higher levels (e.g., study or data collection) are often well standardized, variable metadata are still rarely published using DDI. As a result, this specific layer contributes only marginally to fulfilling the FAIR requirements articulated in research data policies. Even when standardized metadata exists, it is frequently not published via standardized protocols. This gap between metadata production and metadata publication is not merely anecdotal. Wenzig and Han (2024) provide empirical evidence showing that fine-grained standardized DDI metadata for tabular research data is rarely accessible. A look at the global registry of research data repositories—re3data.org—further illustrates the situation. In 2024, more than 3,300 repositories were listed on re3data.org, and about 300 reported using DDI standards. However, only a small subset (29) exposes DDI metadata resources via standardized harvesting protocols. These repositories collectively provide more than 250,000 metadata resources in DDI-Codebook format. Among these programmatically accessible DDI metadata records, only one repository exposed metadata via the OAI-PMH protocol, and only 0.7% of the records contain metadata on the fine-grained variable level. Section 6 provides a more detailed rationale for focusing on the OAI-PMH approach.

The challenge we address in this paper is not the absence of standards, but incomplete integration of a key standard (DDI) into our institutional workflows for producing and publishing fine-grained metadata. (Section 3 discusses the selection of DDI, particularly DDI-Codebook, in comparison to alternative metadata standards.) This integration is particularly important for

1 The institutions represented in this paper also commit to adhering to the FAIR principles (e.g., excerpts from the mission statements of the FDZ-DZHW and SOEP): 'The FDZ-DZHW follows international standards to make research data findable, accessible, interoperable, and reusable in accordance with FAIR principles.' https://fdz.dzhw.eu/en/about-the-fdz 'SOEP data are made available free of charge in line with the FAIR principles for scientific data management. User support is provided through a comprehensive range of information and services that reflect the technical and methodological state of the art.' https://www.diw.de/en/diw_01.c.952867.en/mission_statement_of_the_soep.html.

2 Since survey data in the social and economic sciences (which are the subject of the practical application presented here) are in most cases sensitive personal data that may not be shared openly, the publication of FAIR metadata is even more important (see Davidson *et al.*, 2022, pp. 5–6).

3 The standards of the DDI-Family are developed and published by the Data Documentation Initiative (also DDI-Alliance): https://ddialliance.org/.

survey data used across a wide range of disciplines for improving overall FAIRness. However, assessing the need for such integration requires considering whether there is an explicit demand for standardized variable metadata. In practice, researchers rarely articulate a direct demand for improved FAIRness of fine-granular data. Instead, such expectations tend to manifest as general requirements placed on the data infrastructure itself, such as improved search results for theory-based concepts, concept-based filtering, machine actionability of metadata, and ability to merge datasets. Meeting these requirements typically depends on the availability of fine-grained standardized metadata. Consequently, enabling the systematic production of such metadata increases the likelihood that the overall FAIRness of research data will improve, directly benefiting researchers and other users of the research data. Our contribution therefore is to support the provision of standardized variable-level metadata within our institutions to better meet institutional requirements for FAIR metadata and, in doing so, to provide researchers with data that are more closely aligned with their needs. We further introduce a lightweight and practical approach for generating and publishing fine-grained, standardized metadata, which can be readily adopted and implemented by the broader community, thereby improving FAIRness at scale.

2. HOW FINE-GRAINED STANDARDIZED METADATA CONTRIBUTE TO THE FAIR-PRINCIPLES

While many research data centers (RDCs) already provide high-quality data and metadata services, there are compelling reasons to consider more standardized approaches, such as DDI, also at the granular level:

- In principle, standardization can improve the scalability of business processes and free up resources.
- Standardized metadata can be re-used within the Open Data Format (Han, Hartl, and Wenzig, 2024) to reduce the cost of managing data files for different software platforms and to deliver multilingual documentation directly to the user's statistical software.
- Artificial intelligence (AI), especially large language models (LLMs), can make use of fine-grained standardized metadata to contextualize the structure and meaning of a domain-specific data structure (e.g., in the social sciences) more accurately. This contextualization improves the performance of semantic search by enabling the model to align variable and value labels, concepts, and question texts correctly, thereby reducing misinterpretations and the likelihood of hallucinations in the search results generated by the model.

The publication of fine-grained standardized metadata would also contribute to the creation of FAIR research data and facilitate the FAIR principles:

- They increase findability, not only for LLMs but generally, because they provide more content and deeper insights for each kind of search.
- Since datasets in the social sciences are, in most cases, not freely accessible, it is even more important that relevant metadata on the variables—the core units of analysis—are openly available. Furthermore, if the data is no longer accessible at all, there are good reasons (e.g., more open licenses on metadata facilitate harvesting and storage by third parties) to expect that metadata are still available and, with them, valuable information, especially at the variable level.
- Providing the metadata at a variable level in a formal language would contribute to the interoperability of the data, by improving its comprehensibility for both humans and machines. Using an established standard, like DDI, for the metadata would further enhance interoperability by ensuring consistency and compatibility across different systems.
- Distributing information in domain-relevant standards will result in a better-informed re-use, as field definitions and general descriptions remain comprehensible over time.

As a first conclusion, we can state that providing fine-grained standardized metadata offers many advantages, as it enhances machine actionability and the overall FAIRness of the data. Therefore, it should be considered best practice.

3. CHALLENGES IN PUBLISHING FINE-GRAINED STANDARDIZED METADATA

Despite the advantages we outlined, published standardized metadata at the variable level in DDI remains scarce. This raises the question of which factors present significant obstacles to data providers in generating such metadata.

The problem is not the lack of availability of DDI-standards: Metadata standards from the DDI family (DDI-Codebook and DDI-Lifecycle, and DDI-CDI—DDI Alliance, 2012; DDI Alliance 2020; DDI Alliance 2025) offer a comprehensive solution for describing research data at a fine granular level, including information about data sets, variables, and their characteristics. Until 2025, these DDI standards only specified metadata in the form of XML files. (This changed with the model-driven approach of DDI-CDI and will continue with the next version of DDI-Lifecycle.)

Also, OAI-PMH (Open Archives Initiative, 2015) provides an established protocol that enables metadata sharing and low-level access. It is built for XML metadata and allows for harvesting metadata from a repository by the URL of the OAI-PMH endpoint.

Widely used statistical analysis software packages, such as Stata and SPSS, embed metadata at variable level, including variable names, variable labels, and categories/value labels.

The main challenges in implementing the policy requirement of FAIR metadata at a fine-granular level are not the absence of standards, access protocols, or statistical software capable of managing metadata. The main obstacles concern the mapping of the metadata to an existing standard (here DDI), integrating this mapping into data processing workflows, and implementing an infrastructure for the subsequent provision of accessible metadata. These steps require resources, both expertise and financial means, that many infrastructures often do not anticipate having at their disposal. As a result, even when fine-grained standardized metadata are technically feasible, their dissemination often falls short due to perceived resource constraints and competing operational priorities. This is further reinforced by the absence of lightweight approaches for standardizing and publishing fine-grained metadata. By proposing an approach that is simple and easy to implement, our work aims to support infrastructures of different sizes and resource capacities, offering especially a solution relevant for smaller repositories that cannot afford extensive development efforts.

In the following, we outline our approach by addressing the challenges described based on the metadata of three data centers.

4. DATA TO BE INTEGRATED

All three partners already manage metadata at a variable level. These are presented to end-users via different portals: www.paneldata.org, variablesearch.lifbi.de, and metadata.fdz.dzhw.eu. The database for these portals is as heterogenous as the metadata management philosophies:

- The Socio-Economic Panel (SOEP) manages its metadata in CSV tables on a GitLab server with standard office software (Wenzig, 2019).
- At the Leibniz Institute for Educational Trajectories (LIfBi), an MS SQL database is the backbone for the metadata, which are managed using a self-developed metadata editor (Wenzig *et al.*, 2016).
- Research Data Center for Higher Education Research and Science Studies, based at the German Centre for Higher Education Research and Science Studies (FDZ-DZHW), uses a NoSQL database (MongoDB) to store metadata in JSON files. Metadata ingestion is facilitated through metadata editor with input masks for general metadata fields, while variable and question metadata are imported using a dedicated JSON importer.

Prior to this project, the partners used this metadata for documentation on custom-developed web portals but did not officially publish metadata at the variable level in a standardized manner.

During the first project phase, the partners committed to targeting a payload on each of the relevant conceptual levels because this information is already (somehow) available in the three RDCs:

- On the study level, the information that is already used for registering the datasets' DOIs could be re-used. This means that, as a first approach, the digital object for which the DOI is registered should be the one to publish metadata in DDI-Codebook.
- On the dataset level, the file names (with or without suffix), a label, and a description should be recorded.
- On the variable level, the name of the variable, its label, and an optional description should be available. Pairs of value and labels build the categories (the term used in DDI-Codebook) or value labels (Stata or SPSS terms). It should also be possible to include question texts and all kinds of keywords or overarching concepts.
- Each metadata field containing natural language should be assigned a language code. It should be possible to provide alternative translations for those entries.

5. MAPPING TO DDI-CODEBOOK

We choose DDI for our approach because it is a widely used metadata standard for survey and other observational data in a tabular format across a broad range of disciplines. Thus, it can be considered a domain-specific metadata standard, rather than a discipline-specific standard tied to a single field. As mentioned above, other major standards exhibit significant limitations with respect to variable-level metadata, which are here addressed: SDMX primarily targets aggregated statistical data, schema.org provides only very generic support for variable-level descriptions, DCAT focuses on datasets and distributions rather than explicit variable metadata, and DataCite does not offer a dedicated model for describing variables. In contrast to these approaches, DDI is specifically designed to support rich, structured metadata at the variable level of tabular data. This made DDI a particularly suitable choice for our project, as it enables consistent, machine-actionable descriptions of variables. Within the DDI-Family, several sub-standards with varying levels of complexity exist. Since our aim was to keep the approach simple and ensure that it can be easily implemented by our institutions as well as others, we decided to use DDI-Codebook instead of more complex DDI-versions, such as DDI-Lifecycle. DDI-Codebook offers a standard that is particularly well-suited for publishing variable metadata, without requiring full lifecycle or process modeling. Consequently, the plain and straightforward mapping described in the following section should not be understood as a limitation, but a deliberate design choice to promote an easy adoption.

An XML file in DDI-Codebook standard (DDI Alliance, 2012) can have five elements below the root-element `<codeBook>`: `<docDscr>`, `<stdyDscr>`, `<fileDscr>`, `<dataDscr>`, and `<otherMat>`.

- `<docDscr>` contains bibliographic information describing the XML file itself. This is not addressed in the following as it should only contain information about the metadata itself, i.e., who created it, creation date, etc.
- The study description in `<stdyDscr>` consists of information about the data collection, study, or compilation that this DDI-compliant documentation file describes.
- Within `<fileDscr>`, information about the data file(s) that comprises a collection can be stored.
- The section `<dataDscr>` contains the descriptions of the variables, the fine-grained metadata that are in the center of this project.
- Finally, the element `<otherMat>` allows for the inclusion of other materials that are related to the study.

The tree structure (DDI Alliance, n.d.) provides a good overview of the DDI-Codebook standard, and the field-level documentation (DDI Alliance, 2014) serves as the reference book for each element.

The use case we show here is an idealized example derived from the actual holdings of the three RDCs. Figure 1 gives an impression of the material we want to document in a DDI-Codebook XML file: Within a study, we have two tables (in Stata format) tab1.dta and tab2.dta that are packed into a ZIP file. Each of the two tables contains two variables.

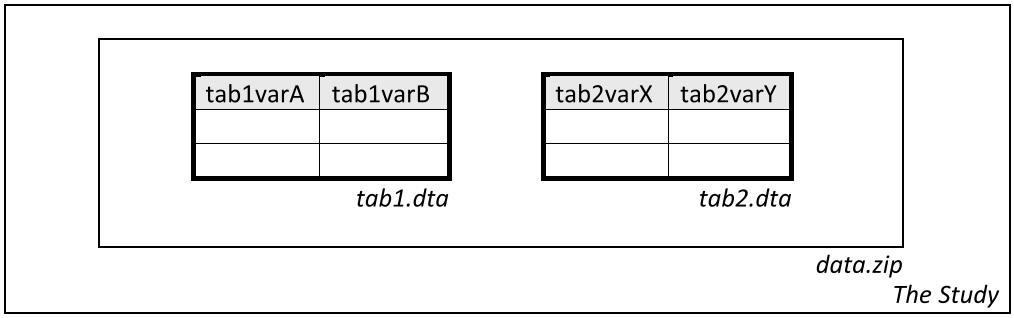

Figure 1 Use case with two Stata-files, each containing two variables, in one ZIP-file within a study.

As the study level is of minor importance for this project, the XML code for the element `<stdyDscr>` in Figure 2 only shows also the single mandatory field of the DDI-Codebook standard `<titl>` (which implies `<citation>`, `<titlStmt>`, and `<stdyDscr>` are also present). `<parTitl>` shall be used to provide a translated title. Element `<IDno>` contains the registered DOI. So, in our scenario, the `<stdyDscr>` (Figure 1) should contain information on the DOI level, which means for the three RDCs that metadata used for the DOI registration (like study title, author information, or the scope of the study) can be re-used.[4]

```
<stdyDscr>
  <citation>
    <titlStmt>
      <titl xml:lang="de">TheStudyTitleGerman</titl>
      <parTitl xml:lang="en">TheStudyTitleEnglish</parTitl>
      <IDNo agency="DOI“>studyDOI</IDNo>
    </titlStmt>
  </citation>
</stdyDscr>
```

Figure 2 Content of element `<stdyDscr>`.

The attribute xml:lang is used to specify the language of the field's content. The ISO two-letter codes, like 'de' for German or 'en' for English, should be used.

The element `<fileDscr>` (Figure 3) should contain information on file level and can be repeated for collections with multiple files. The attribute ID can be used for the filename; in our case, it could be data.zip. In the next section of the DDI-Codebook XML file, this ID can be used to assign a variable to a file. If a repository provides files in two different formats, the element should be repeated to describe the two separate files; it also could make sense to use the file name without extension.

```
<fileDscr ID="Filename">
  <fileTxt>
    <fileName xml:lang="de">FileLabelGerman</fileName>
    <fileName xml:lang="en">FileLabelEnglish</fileName>
    <fileCont xml:lang="de">FileDescriptionGerman</fileCont>
    <fileCont xml:lang="en">FileDescriptionEnglish</fileCont>
  </fileTxt>
</fileDscr>
```

Figure 3 Content of element `<fileDscr>`.

Within the element `<fileTxt>`, the elements `<fileName>` and `<fileCont>` contain a label and a description of the dataset. Again, the attribute 'xml:lang' is used to indicate the language of the descriptive text.

It is not always possible to fully describe the structure of a collection containing multiple files across different folders or bundled in a ZIP file. However, if ZIP files (or similar archives) are used, one option is to name all files explicitly using the `<fileCont>` element for both the ZIP file and its contents to indicate their relationship, at least in a human-readable way.

The specification of DDI-Codebook does not locate the information on variables within the element `<fileDscr>` but introduces the element `<dataDscr>` (Figure 4) that holds information

4 The three research data centers obtain their DOIs from DataCite (https://datacite.org/).

on each of the variables. The element `<dataDscr>` is repeatable, and if one would want to use the attribute ID for the file name, one could organize the variables of multiple files.

```
<dataDscr>
  <varGrp type="file" var="var1ID var2ID ..." name="tab1.dta">
    <labl xml:lang="en">tab1.dta-LabelEnglish</labl>
    <labl xml:lang="de"> tab1.dta-LabelGerman</labl>
    <txt xml:lang="en"> tab1.dta-DescriptionEnglish</txt>
    <txt xml:lang="de"> tab1.dta-DescriptionGerman</txt>
  </varGrp>
  <varGrp type="file" var="var3ID var4ID ..." name="tab2.dta">
    ...
  </varGrp>

  <var ID="var1ID" name="tab1varA" files=xs:IDREFS>
    <location fileid="FileName"/>
    ...
  </var>
  ...
  <var ID="var4ID" name="tab2varY">
    <location fileid="FileName"/>
    ...
  </var>
</dataDscr>
```

Figure 4 Content of element `<dataDscr>`.

The element `<varGrp>` can be used to group the variables in datasets. Therefore, the attribute 'type' must be 'file', the attribute 'name' can be used to provide the name of the dataset and the elements `<labl>` and `<txt>` can be used to store a label and a description of the dataset. Multilingual content can be accommodated using the attribute xml:lang.

Each variable has its own element `<var>` within `<dataDscr>`. The name of the variable is encoded in the attribute 'name'. The element `<location>` with its attribute 'fileid' is used to assign a variable to a file/dataset. This 'fileid' should relate to the content of attribute 'ID' in element `<fileDscr>`. There can also be attribute files in `<var>`, but it is not clear when this attribute should be used and when the element `<location>` should be used.

Overall, there appear to be too many ways to organize the variables of multiple data tables in DDI-Codebook. Although the `<varGrp>` element is generally considered the preferred mechanism for grouping variables, users also adopt other approaches. These include repeated instances of the `<dataDscr>` element, use of the `<location>` element to associate variables with files, and deployment of the files attribute within the `<var>` element. From the perspective of potential harvesters this variety of options seems not to be optimal.

The element `<var>` also is the container for more information on the variables (Figure 5).

```
<var name="VariableName">
  <location fileid="..."/>
  <labl xml:lang="en">VariableLabelEnglish</labl>
  <labl xml:lang="de">VariableLabelGerman</labl>
  <qstn xml:lang="en">QuestionTextEnglish</qstn>
  <qstn xml:lang="de">QuestionTextGerman</qstn>
  <txt xml:lang="en">VariableDescriptionEnglish</txt>
  <txt xml:lang="de">VariableDescriptionGerman</txt>
  <catgry>
    <catValu>Value</catValu>
    <labl xml:lang="de">ValueLabelGerman</labl>
    <labl xml:lang="en">ValueLabelEnglish</labl>
  </catgry>
  <catgry>...</catgry>
  <concept xml:lang="en">ConceptLabelEnglish</concept>
  <concept xml:lang="de">ConceptLabelGerman</concept>
</var>
```

Figure 5 Content of element `<var>`.

The element `<labl>` holds the variable label as known from statistical software like Stata or SPSS. If available, the element `<qstn>` can be used to store the precise wording of a question. Any other descriptive texts can go into the element `<txt>`. The element `<catgry>` holds a pair of the elements `<catValu>` and `<labl>`, the latter with optional alternatives in different languages. It must be repeated for each labeled value of the variable.

The FAIRness of metadata benefits from the inclusion of community-developed terminologies. As variable labels and question texts can hardly be standardized using terminologies, the repeatable element `<concept>` can contain information like topics, terms from thesauri, or other keyword systems.[5] Precisely, the `<concept>` element can be used to link the labels and questions to theory-based concepts drawn from community-developed thesauri like the European Language Social Science Thesaurus (ELSST, https://elsst.cessda.eu/), Thesaurus for Economics (STW, https://zbw.eu/stw/version/latest/about.en.html), or the Thesaurus for the Social Sciences (TheSoz, https://data.gesis.org/thesoz/term_10037415_de). This enables researchers to search for theory-based terms and identify variables linked to those concepts. Although we implemented an attribute to enable such linkages, connecting question and variable metadata to discipline-specific concepts is still a complex task (see Daniel *et al.*, 2024 for a pilot study on a Linked Open Data (LOD)–based concept registry).

The three project partners exported the information from their metadata systems. Alternatively, most of the information could have been derived from the Stata or SPSS dataset files: file names, variable names, variable labels, and value labels. It is easy to extract them and convert this information to a DDI-Codebook file, for example, by using the R package DDIwR (Dusa, 2024). The proposed mapping could serve as a guideline for the specific implementation within the data processing workflow.

Here is an overview of the initial results: For the SOEP-data, one big XML file was constructed, which contains information on 550 datasets and 107,210 variables. At LIfBi, six XML files sized 12 to 27 MB with information on 204 datasets and 38,730 variables were created. At DZHW, an export feature for 28 studies with 31,448 variables was developed. This feature will be integrated into the DZHW portal metadata.fdz.dzhw.eu. This will allow users to directly download one DDI-Codebook XML file for all datasets and variables of the latest version for every data package listed on the portal, as shown in Figure 6.

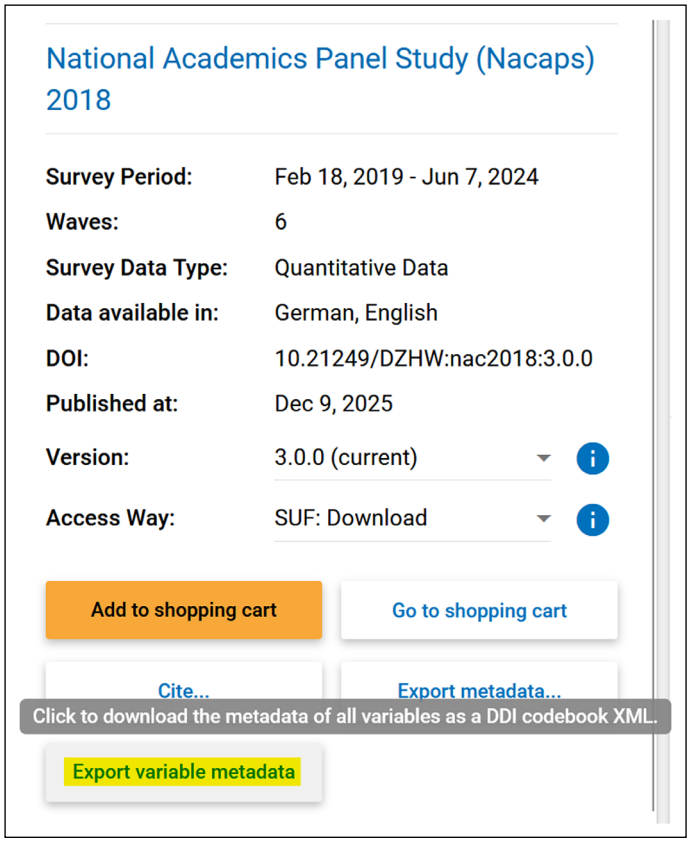

Figure 6 Introducing the option for downloading DDI-Codebook metadata from metadata.fdz.dzhw.eu.

5 Again, alternatives can be provided by the use of the attribute xml:lang.

DOI: 10.5334/dsj-2026-013

6. EVALUATION OF OAI-PMH

Besides evaluating the possibilities of extracting DDI-Codebook metadata from proprietary metadata systems, the other aim of the project was to publish metadata in a standardized way using the OAI-PMH protocol. The use of the OAI-PMH protocol provides a straightforward way of harvesting metadata, as both the interface behavior and the structure of the exposed metadata are well defined and standardized. In contrast, alternative access mechanisms such as RESTful APIs or SPAQRL endpoints, while technically powerful, are often implemented in a highly heterogenous manner.

The usage scenario could be as follows: The global registry of research data repositories re3data.org lists all repositories that publish DDI-Codebook metadata via the OAI-PMH protocol. As part of the listing, the OAI-PMH endpoint of each repository is provided. Any interested user can access the OAI-PMH endpoint and query available metadata formats. If metadata are available in DDI-Codebook format, the user can request them in the specified format. With only minor restrictions, this approach is already functional today. Most importantly it can operate completely automatically. Using this approach, it is already possible to get access to more than 250,000 DDI-Codebook XML files and analyze them.

The project partners evaluated four different methods to publish DDI-Codebook metadata via the OAI-PMH protocol:

- The 'Simple OAI-PMH 2.0 Data provider' (Meyer, 2024) does not require a database, as it reads files from folders in a specified folder structure on the server's file system. It would then nest the XML files content as-is into the OAI-PMH response. While the approach is very intuitive and it was easy to set up the software on a test system, it turned out that the XML files are too large, leading to time-out issues.
- The server software 'oai-pmh2' (Meyer, 2025) depends on an SQLite database (or other), and it is not clear how to import the metadata into the SQLite database.
- Another option would be the metadata server Kuha2 (FSD, 2024). However, it is unclear if the server could process the large metadata files produced in the previous step and if the quite complex installation would pay out.
- We also evaluated the potential of writing our own server software in Python based on the Flask framework, as described by Hallet (2019). It turns out that although initial results were rapidly available, it was not possible to finalize them within the scope of this project.

Besides the implementational obstacles described above, OAI-PMH seems to be increasingly misaligned with modern data exchange practices (see Section 7). While it is still widely deployed, it shows signs of technical aging, particularly with respect to scalability, flexibility, and integration with modern web technologies. Thus, the three project partners chose to take different paths: SOEP published the metadata at Zenodo (Wenzig, 2025), DZHW persistently integrated a download option in its portal (similar to the option Dataverse repository software offers), and LIfBi decided to publish their scientific use file also in the Open Data Format, which is based on DDI-Codebook. In the following section, we evaluate alternative approaches for publishing FAIR metadata.

7. EVALUATION OF ALTERNATIVE APPROACHES FOR PUBLISHING FAIR METADATA

While OAI-PMH relies heavily on XML, it seems that there is a shift away from this file format. From the perspective of software development, JSON (in particular, JSON-LD) would be more efficient and clearer as it uses less markup and is natively supported by JavaScript. In 2025, the DDI-Alliance published the new standard DDI-CDI, where CDI stands for 'cross-domain integration.' While DDI-CDI metadata can be expressed in XML, other formats are also supported, as DDI-CDI defines only the data model, not the metadata file format. This model-driven approach will also be used for DDI-Lifecycle 4, where an RDF representation is also expected—and it is often stated that RDF will become more important over time as it is better suited for including linked data. However, XML might offer other benefits, especially for long-term archiving. Future work should evaluate whether newer formats, especially JSON in the context of DDI-CDI, could serve as an alternative for metadata storage. In this context, the methods for distributing metadata also need to be reassessed. In the following, we therefore examine different approaches to exposing metadata in a way that takes into account the need for openness to other formats.

From an RDC perspective, it would be easy to provide metadata as a downloadable resource. FAIR signposting thus seems to be a viable option for indicating access to the metadata.

Within FAIR signposting (Van de Sompel et al., 2023), a standardized qualified link to the metadata would be published on the landing page of the digital object for which a DOI has been registered. This approach ensures that the metadata would then be easily accessible via HTTP and OAI-PMH would no longer be necessary to harvest the metadata. This seems promising as HTTP is a globally established standard, whereas OAI-PMH is a more specialized protocol for distributing XML metadata primarily used in the archival community. By using HTTP-based metadata discovery, FAIR signposting aligns with common web technologies, enhancing especially interoperability and accessibility.

Enriching a landing page with machine-readable information works well for encoding simple bibliographic details, such as author names. LIfBi enriches the HTML code of landing pages with a linked JSON-LD file (shown in Figure 7): with this relatively little amount of information, they achieve a score of 83% in the fully automated FAIR data assessment tool F-UJI (Devaraju and Huber, 2020). This approach demonstrates that machine-readable metadata can be linked and optimized with minimal effort. However, in this example, the optimization is restricted to study-level metadata, since the F-UJI tool does not evaluate metadata below the study level, such as variable metadata.

```
<link rel="type" href="https://schema.org/Dataset">
<link rel="type" href="https://schema.org/AboutPage">
<link rel="describedby" type="application/ld+json"
href="https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/signposting/DOI_10.5157_NEPS_SC1_D_10-1-0.jsonld">
```

Figure 7 Part of the landing page (LIfBi, 2025) for https://doi.org/10.5157/NEPS:SC1:10.1.0.

As there is no standardized way to indicate that 'DDI-Codebook-compliant metadata are available at this link', we demonstrate two approaches to solve this, particularly with regard to enabling machine readability.

Option 1 would be the HTML element `<link>`. It can be `<link rel = "describedby" type = "application/xml" href = "https://url.to/codebook.xml">`

The only available option seems to be: 'Here is an XML file describing the digital object.' This is not ideal, as users should not have to download an XML file only to realize it does not follow the expected standard.

An alternative solution would be to mint a specific 'rel' type for Codebook metadata (think of it as a sub property of 'describedby'). Since the 'rel' attribute accepts either a registered keyword or an arbitrary URI, you can use an IRI for that and still be fully HTML5 compliant. Naturally, other people would have to recognize this URI.

Another long-term solution would be to register a MIME type for codebook XML, for example, application/codebook+xml, to enable explicit and machine-actionable identification of codebook-based metadata.

Option 2 would rely on JSON-LD and make use of the Schema.org property 'distribution' (Figure 8).

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Dataset",
  "distribution": {
    "@type": "DataDownload",
    "encodingFormat": "https://ddialliance.org/ddi-codebook_v2.5",
    "contentUrl": "https://url.to/codebook.xml",
    "description": "XML metadata document using the DDI Codebook 2.5 standard."
  }
}
</script>
```

Figure 8 An approach to link to DDI-Codebook metadata which makes use of the Schmea.org property 'distribution'.

In this case, the distribution property points to an xml file containing the complete DDI-Codebook metadata set. This stretches the semantics of 'distribution', but the Schema.org (2024) specification references, at this point, the W3C Data Catalog Vocabulary (Albertoni *et al.*, 2024), which states in a note 'Nevertheless, the question of whether different representations can be understood to be distributions of the same dataset, or distributions of different datasets, is application specific. Judgment about how to describe them is the responsibility of the provider, taking into account their understanding of the expectations of users, and practices in the relevant community.' Given that in widely used data formats (e.g., Stata, SPSS) or the Open Data Format, metadata are tightly coupled with the data itself, the DDI Codebook can reasonably treated as a distribution-level representation.

The property 'isBasedOn' would be another option to specify some kind of relation of the digital object in schema.org, described by the landing page and the linked metadata. The semantic fit of this solution is far from being perfect, as it primarily denotes derivation rather than representation. For this reason, the approach described above should be preferred in the context of schema.org, as it provides a clearer representation of the semantic relationship.

As all three data centers participating in this paper mint their DOIs via DataCite, the obvious question arises as to why DataCite is not used as a referencing approach instead of FAIR signposting. While Data Cite provides a rich metadata schema, it operates at a different layer than FAIR signposting. FAIR signposting enables resource-based discovery by guiding directly from landing pages to the specific resources, whereas DataCite exposes metadata (including the references to the related resources) via the registry service. Although it would be possible to reference the metadata in the DataCite schema using the property 'HasMetadata', the use of DataCite properties presupposes access to and processing of DataCite records. The approach outlined in this paper instead investigates how the availability of DDI-Codebook metadata can be made visible directly at the level of study landing pages (or API endpoints), independent of external (registry) services. In this context, FAIR signposting appears to be the more promising method. However, this does not imply that referencing detailed Metadata in the DataCite record is misleading, as it fundamentally enriches the metadata provided in the context of the DOI registration. In conclusion, DataCite and FAIR signposting complement each other within the data infrastructure, operating at different layers.

Ultimately, future research must evaluate what option is best for establishing a machine-actionable link from a landing page to more complex metadata, rather than being limited to individual metadata elements. The approaches described could serve as a starting point for further discussions.

8. DISCUSSION

Given that the FAIRness of research data has become a ubiquitous requirement in research data policies in general and in our own institutions, we examined how the FAIRness of fine-grained metadata can be improved by implementing the DDI metadata standard at the variable-level of tabular data for our institutions. We chose DDI because it is one of the major standards for data generated from surveys and other observational methods across a broad range of disciplines (e.g., economics, psychology, sociology). With respect to the FAIR principles, the use of an established domain-specific yet interdisciplinary standard such as DDI enhances interoperability and reusability of the metadata for survey-based data and other structured observational data. We demonstrated that it is possible to export fine-grained standardized metadata from diverse data sources with only modest resource requirements using DDI Codebook, while also incorporating potential linkages to discipline-specific terminologies in the metadata specification.

While it turned out that publishing DDI-Codebook metadata via OAI-PMH is not as straightforward as initially expected, we discussed FAIR signposting techniques as potential alternatives to improve findability and accessibility of the metadata. These techniques show promising capabilities; however, further work is required with regard to standardization of access mechanisms. Future work should further explore standardized approaches to metadata publishing beyond OAI-PMH, for example, through standardized RESTful APIs, building on the signposting methods outlined above as an initial starting point. In this context, the registration of dedicated MIME types for DDI metadata would be beneficial in facilitating machine-actionable identification of DDI-based metadata.

The selection of DDI-Codebook offers the advantage of a very straightforward mapping and implementation. However, its simplicity comes at the cost of limited features. Therefore, future work should examine how the outlined approach could be applied to other standards within the DDI family, such as DDI-CDI, thereby addressing an even broader range of data types and disciplines. In this context, it should be evaluated whether XML remains the most appropriate format for publishing and (long-term) archiving metadata or whether JSON (or JSON-LD) would be more suitable.

With the project described in this paper, we not only improved the FAIRness of our own metadata but also provided a best-practice approach for other institutions, which can benefit the FAIRness of survey-based research data in a variety of disciplines.

ACKNOWLEDGEMENTS

We gratefully acknowledge the anonymous reviewers for their invaluable feedback and insightful comments, which significantly enhanced the quality of this manuscript. We would also like to thank Pierre-Antoine Champin for the exchange on signposting and Adam Lederer for proofreading. Any remaining errors are solely our responsibility. We thankfully acknowledge NFDI-funding by DFG, project no. 442494171 (KonsortSWD, https://www.konsortswd.de), which financed part of this work.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Knut Wenzig orcid.org/0000-0002-2259-0203
SOEP, German Institute for Economic Research (DIW), Berlin, Germany

Andreas Daniel orcid.org/0000-0002-0111-8858
FDZ-DZHW, German Centre for Higher Education Research and Science Studies (DZHW), Hannover, Germany

Dominique Hansen orcid.org/0009-0001-0387-3419
SOEP, German Institute for Economic Research (DIW), Berlin, Germany

Tobias Koberg
Leibniz Institute for Educational Trajectories (LIfBi), Bamberg, Germany

Mihaela Tudose
Leibniz Institute for Educational Trajectories (LIfBi), Bamberg, Germany

REFERENCES

Albertoni, R. *et al*. (2024) *Class Distribution – Data Catalog Vocabulary (DCAT) – Version 3*. Available at: https://www.w3.org/TR/2024/REC-vocab-dcat-3-20240822/#Class:Distribution.

Daniel, A. *et al*. (2024) *A pilot study for "Linked Open Research Data" (LORDpilot): A LOD-based concept registry for social science research data*. Available at: https://doi.org/10.5281/zenodo.11047523

Davidson, J. *et al*. (2022) *FAIR-enabling Data Policy Checklist (Version 1.0)*. Available at: https://doi.org/10.5281/zenodo.6225775 (Accessed: 31 December 2025).

DDI Alliance (n.d.) *XML Schema Outline – DDI-Codebook Version 2.1 – Tree Structure*. Available at: https://ddialliance.org/hubfs/Specification/DDI-Codebook/2.1/DTD/DDI2-1-tree.html (Accessed: 31 December 2025).

DDI Alliance (2012) *DDI-Codebook v2.5*. Available at: https://ddialliance.org/ddi-codebook_v2.5 (Accessed: 31 December 2025).

DDI Alliance (2014) *DDI-Codebook v2.5 – Linked Field-Level Documentation*. Available at: https://docs.ddialliance.org/DDI-Codebook/2.5/xmlschema/index.html (Accessed: 31 December 2025).

DDI Alliance (2020) *DDI-Lifecycle v3.3*. Available at: https://ddialliance.org/ddi-l_v3.3 (Accessed: 31 December 2025).

DDI Alliance (2025) *DDI-CDI v1.0*. Available at: https://ddialliance.org/ddi-cdi_v1.0 (Accessed: 31 December 2025).

Devaraju, A. and **Huber, R.** (2020) *F-UJI – An Automated FAIR Data Assessment Tool*. Available at: https://doi.org/10.5281/zenodo.6361400 (Accessed: 31 December 2025).

Dusa, A. (2024) *DDIwR: DDI with R (R Package)*. Available at: https://cran.r-project.org/package=DDIwR (Accessed: 31 December 2025).

European Commission, Directorate-General for Research and Innovation. (2018) *Turning FAIR into reality: Final report and action plan from the European Commission Expert Group on FAIR Data*. (Publication No. KI-06–18–206-EN-N). Publications Office of the European Union. Available at: https://doi.org/10.2777/1524 (Accessed: 31 December 2025).

FSD (2024) *Kuha2 (Software bundle)*. Available at: https://kuha2.readthedocs.io/ (Accessed: 31 December 2025).

Hallet, R. (2019) *OAI-PMH Service Updates*. Available at: https://doi.org/10.5438/ppth-pz62 (Accessed: 31 December 2025).

Han, X., Hartl, T. and **Wenzig, K.** (2024) *Introducing Open Data Format: A Platform-Independent, Non-Proprietary, Metadata-Enriched, Multilingual Data Format and its Implementation in R and Stata*. KonsortSWD Working Paper 10/2024. Available at: https://doi.org/10.5281/zenodo.14215268 (Accessed: 31 December 2025).

Hodson, S. and **Gregory, A.** (2023) *WorldFAIR Project (D1.3) First policy brief (Version 1)*. Available at: https://doi.org/10.5281/zenodo.7853170 (Accessed: 31 December 2025).

LIfBi (2025) *JSON-LD file connected linked from landing page*. Available at: https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC1/10-1-0/DOI_10.5157_NEPS_SC1_10.1.0.jsonld (Accessed: 31 December 2025).

Meyer, S. (2024) *Simple OAI-PMH 2.0 Data Provider v1.8*. Available at: https://github.com/openculturecconsulting/simple-oai-pmh/releases/tag/v1.8 (Accessed: 31 December 2025).

Meyer, S. (2025) *oai-pmh2*. Available at: https://github.com/openculturecconsulting/oai-pmh2 (Accessed: 31 December 2025).

Open Archives Initiative (2015) *The Open Archives Initiative Protocol for Metadata Harvesting, Protocol Version 2.0 of 2002-06-14*. Available at: https://www.openarchives.org/OAI/openarchivesprotocol.html (Accessed: 31 December 2025).

Schema.org (2024) *distribution – A Schema.org property v28.1*. Available at: https://schema.org/distribution (Accessed: 31 December 2025).

Van de Sompel, H. *et al*. (2023) *FAIR Signposting Profile*. Available at: https://signposting.org/FAIR/ (Accessed: 31 December 2025).

Wenzig, K. (2019) *Longitudinal Metadata at the Socio-Economic Panel*. Available at: http://doi.org/10.5281/zenodo.3554859 (Accessed: 31 December 2025).

Wenzig, K. (2025) *Metadata from SOEP v40.1 in DDI-Codebook v2.5*. Available at: https://doi.org/10.5281/zenodo.17856791 (Accessed: 31 December 2025).

Wenzig, K. *et al*. (2016) 'Management of metadata: An integrated approach to structured documentation', in H.-P. Blossfeld *et al*. (eds.) *Methodological issues of longitudinal surveys*. Wiesbaden: Springer VS, pp. 627–647. Available at: https://doi.org/10.1007/978-3-658-11994-2_35

Wenzig, K. and **Han, X.** (2024) 'State of DDI cloud', *IASSIST Quarterly*, 48(4). Available at: http://doi.org/10.29173/iq1116

Wilkinson, M. *et al*. (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3, p. 160018. Available at: https://doi.org/10.1038/sdata.2016.18

TO CITE THIS ARTICLE:
Wenzig, K., Daniel, A., Hansen, D., Koberg, T. and Tudose, M. (2026) 'Publishing Fine-Grained Standardized Metadata: Lessons Learned from Three Research Data Centers', *Data Science Journal*, 25: 13, pp. 1–13. DOI: https://doi.org/10.5334/dsj-2026-013

Submitted: 20 May 2025
Accepted: 10 March 2026
Published: 24 March 2026

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.