State of the DDI Cloud

Knut Wenzig[1] and Xiaoyao Han[2]

Abstract

As the DDI community continues to grow, an increasing number of repositories are providing their metadata in various DDI formats. However, the current landscape of DDI metadata standards usage is not well understood. Understanding this landscape is crucial as it helps identifying usage patterns, improve interoperability, and guide future developments. To address this research gap, we investigated the availability and comprehensive element usage of DDI standards across 29 repositories registered on the platform re3data.org, using the OAI-PMH API. By analyzing approximately, a quarter of a million metadata records in DDI-Codebook format, we summarized statistics on the usage of popular DDI elements and their distribution across repositories. Our findings may have implications for the deployment of DDI metadata and the further development of these standards. They also inform researchers and data stewards about how DDI-Codebook is utilized by the community. Overall, this investigation underscores the value of openly available metadata in supporting research and achieving the goals of the FAIR data movement.

Keywords

DDI, Metadata Harvesting, OAI-PMH, DDI-Codebook, Data Catalogues

Introduction

Over the years the diagram for the LOD cloud (McCrae 2024) shows the success story of Linked Open Data: When it started in 2007 the LOD cloud reported the existence of 12 interconnected datasets (resolvable and accessible RDF Data with at least 1000 triples). In 2023, 16 years later, 1,314 datasets build the LOD cloud showing that the idea of publishing interconnected data is attracting an increasing number of users.

Driven by the goals of standardization, linking and re-use, the standards of the DDI Alliance follow a similar mission in the world of research data: From DDI-Codebook (DDI Alliance 2014) to DDI-Lifecycle (DDI Alliance 2020), the DDI standards are constantly becoming more extensive and complex. We are interested in which parts of the standards are used to inform users and development. Therefore, we collected and analyzed DDI metadata in the wild. As a result, we created an overview of the DDI cloud: sources where DDI metadata can be found, the structure of published metadata itself, as re-use is intended (there can and should be links between the various entities) so the term “cloud” is appropriate here, too.

DOI: https://doi.org/10.29173/iq1116

```
<codeBook[1]>
    <docDscr[0-n]>
        ... bibliographic information describing the DDI document itself ...
    </docDscr>
    <stdyDscr[1-n]>
        ... information about the data collection, study, or compilation ...
        <citation[1-n]>
            <titleStmt[1]>
                <titl[1]>Text of the Title</titl>
            </titleStmt>
        </citation>
    </stdyDscr>
    <fileDscr[0-n]>
        ... information about the data file(s) ...
    </fileDscr>
    <dataDscr[0-n]>
        ... description of variables ...
    </dataDscr>
    <otherMat[0-n]>
        ... other materials that are related to the study ...
    </otherMat>
</codeBook>
```

Figure 1: DDI-Codebook metadata as pseudo-XML-code with some remarks on the payload information. The cardinality in the square brackets is not part of the XML code: [1] stands for mandatory and non-repeatable element, [1-n] means mandatory and repeatable, [0-1] means optional and non-repeatable, [0-n] is optional and repeatable.

Version 2.5 of the standard DDI-Codebook is the latest version of DDI-Codebook and has been last published with some backward compatible bug corrections in 2014. Version 2.6 is currently under development and about to be released soon.

A DDI-Codebook compliant XML file (see Figure 1) can cover 5 main areas, which are located on level 2 within the element <codeBook> on level 1:

1. The optional and repeatable description of the XML file itself is located within the element <docDscr>, where for example bibliographic information for the file can be stored in an element <citation>.
2. The mandatory and repeatable element <stdyDscr> describes the study and holds information about the data collection (or compilation) and general information including title, abstracts or keywords.
3. The element <fileDscr> is used to describe the files that comprise the collection documented by the DDI-Codebook XML. It is optional and can be repeated for multiple files.
4. The optional and repeatable element <dataDscr> holds information about variables and their categories.
5. Other materials that are related to the collection/study can be described using the element <otherMat>.

As Figure 1 shows, at least the title of the study in the element <titl> is mandatory. It is the only mandatory element within the standard. All together DDI-Codebook 2.5 specifies 252 different elements: 243 global and 9 local ones.

In social sciences tabular data are widely used. The most prominent but proprietary data formats in social sciences (like SPSS, Stata, SAS) already contain per default different metadata like variable labels and value labels, often combined with multilingual features. Providing those metadata would be straightforward, relatively inexpensive and undoubtedly in line with each aspect of the four FAIR principles (Wilkinson, Dumontier, Aalbersberg, et al. 2016), which deal in the end with the availability of metadata: (1) If the metadata, which typically form the basis for data catalogues, are richer and contain information on variables, *findability* would increase; (2) The *availability* of fine-grained metadata would be possible, even if the data are no longer available; (3) More metadata would be published using a formal, accessible, shared, and broadly applicable language for knowledge representation, which would increase *interoperability*; (4) Data are described with more relevant attributes, which would contribute to *reusability*. This kind of metadata can be stored below element <dataDscr>.

While DDI-Codebook covers the needs of social science data archives, DDI-Lifecycle for example can describe questionnaires and introduces better options for metadata re-use.

Methodology

To obtain a comprehensive overview of the usage of DDI metadata elements, we decided to use re3data.org (https://www.re3data.org) as our primary platform for data collection. As a global registry of research data repositories, re3data.org records a growing number of data providers covering a wide array of academic disciplines. These data providers offer research data and its metadata (ideally exposing metadata via interfaces) and/or service providers (e.g., a portals) that harvest the metadata of research data from data providers as a basis for building value-added services. (See Table 1 for numbers and a comparison to 2017.)

The registry went live in autumn 2012 and has been funded by the German Research Foundation (DFG). Since the end of 2015 re3data.org is managed under the auspices of DataCite.

The family of DDI standards is ranked third of all reported metadata standards, after Dublin Core, and the Data Cite metadata schema. As of 2024, re3data.org lists 280 registered repositories that report using DDI metadata standards. Only DDI standards are suited for tabular data since they enable storing fine granular metadata on variable level like variable and value labels or descriptive statistics.

Table 1: Overview of the number of repositories listed in re3data.org by type, metadata standard used and provided API for 2017 and 2024. Source: Own research and Wenzig (2017)

	2017	2024
Research data repositories on re3data.org	1,989	3,179
Among data providers	1,794	2,916
Among service providers	765	1,075
Repositories that report use of Dublin Core	175	561
Repositories that report use of Data Cite	78	386
Repositories that report use of DDI	116	280
Among report SWORD	20	116
Among report REST-API	28	115
Among report DDI	14	85

Inclusion criteria

To make the metadata readily available, re3data.org allows for reporting various API options. SWORDS, REST, and OAI-PMH are the most common APIs under these repositories. We compared these APIs for our research purposes and found that SWORD and REST are not as feasible or convenient as OAI-PMH:

SWORD (Cottage Labs 2021) is in the first place a deposit protocol, but one could also retrieve metadata from an object. None of the repositories (besides one, which does not use DDI) that offer SWORD access provide a valid URL with details for the SWORD access, therefore, we could not check whether DDI metadata could be retrieved this way. The single repository, which provides actual access via this protocol, requires users to have credentials – which makes sense as SWORD is used to automate deposit processes.

The usage of a REST API (Fielding 2000) would need customized access for each repository. The responses to these queries would structurally need to be harmonized, as they typically will not result in a standardized output. Nevertheless, we checked manually the REST APIs of the few repositories, that reported using DDI and some software other than Dataverse (which also allows metadata access via OAI-PMH, see below). In the documentation of those APIs, we did not find any indication that DDI metadata will be available via REST. For these reasons, the two most frequently mentioned protocols are not suitable for harvesting and analyzing DDI codebook metadata.

OAI-PMH (Open Archives Initiative 2015) stands out as a straightforward and efficient protocol for harvesting metadata from repositories. Its standardized approach proves particularly advantageous for large-scale metadata harvesting from multiple repositories, aligning perfectly with our research objectives. Consequently, we chose OAI-PMH as the method to acquire the required metadata. Out of the 280 registered repositories that report using DDI as a metadata, only 85 report using DDI as a metadata standard and support OAI-PMH. The repositories do not report whether they make DDI metadata available via OAI-PMH, however, they state independently whether they use DDI as a metadata standard and/or whether they offer an OAI-PMH access to their metadata. Table 1 gives an

Table 2: Status of the repositories, that report on re3data.org to use DDI as a metadata standard and OAI-PMH as an API

Status of the Repository	Number of Repositories
URL to OAI-PMH in re3data.org is obviously wrong or a duplicate	36
DDI metadata unavailable (DDI metadata is not available via OAI-PMH)	9
The answer of OAI-PMH query contains permanent errors	3
Merge to upper level of repository	8
OAI-PMH with DDI Codebook 2.5 up and running (one of them offers also DDI Lifecycle 3.2)	29

overview of repositories on re3data.org by type, top-3-metadata standards used and API for the years 2017 and 2024.

Before embarking on metadata harvesting through the API, we conducted essential preparatory work. Initially, we manually scrutinized the validity of the API addresses provided for each repository listed on re3data.org. The inclusion criteria for repositories in our research are as follows:

1. The repository is registered on re3data.org.
2. The repository provides DDI as one of the metadata standards.
3. The metadata is available through the OAI-PMH API.
4. There is a valid link pointing to the API endpoint.

Finally, we identified 29 repositories that satisfied these inclusion criteria, forming the basis for further processing. We reported issues such as incorrect links on websites and inaccuracies in DDI metadata standards categorization to re3data.org. They either rectified the inaccuracies or provided updated information. Table 2 shows how the number of 85 repositories reporting the usage of DDI and offering an OAI-PMH API breaks down to the 29 repositories we analyzed.

Data collection

Once the valid repositories were identified, we employed a Python script to conduct metadata harvesting. The script interacted with the API interfaces of each repository, enabling us to traverse the XML metadata tree. To harvest metadata from the OAI-PMH API, two essential pieces of information were required: the API address and the metadata prefix. After manually validating links and collecting prefixes, the information underwent further processing through a Python script for metadata collection.

The OAI-PMH protocol does not require that every item should be available in all formats supported by the repository. Since we received information that different items might actually be accessible when using different prefixes to retrieve records, we decided to include all records we could access using the various DDI-style prefixes. This includes the prefixes with different styles and languages, like ddi25, ddi-c, ddi25-en, oai_ddi 25-nl. Table 3 shows a complete list of those metadata prefixes we found and identified as related to DDI standards. Because only two repositories publish metadata in DDI-Lifecycle, we excluded DDI-Lifecycle from our study. (The prefixes in brackets, which can be found in Table 3 indicate, that there are metadata in DDI-Lifecycle.)

Three Python libraries were utilized for this task: requests, xml.etree, and pandas. The requests library, a web-scraping tool, facilitated HTTP requests and handled responses, allowing us to query online APIs and retrieve XML metadata information. xml.etree was employed to parse XML elements and prepare the data for extraction. Pandas provided convenient data structures and functions for manipulating and analyzing structured data. Once the XML file was parsed through xml.etree, the metadata elements were passed to pandas for the results, which enabled us to store the records as a CSV file. This process resulted in data from 29 different repositories, encompassing 259,606 DDI-Codebook entries. The data were collected in July/August 2024.

During the research we observed significant variability over time, as new repositories were added, entries were updated, repositories were temporarily not available or changed their publication strategy and did not provide OAI-PMH access any more. New repositories can be suggested via https://www.re3data.org/suggest.

Table 3: Number of records, DDI related prefixes and OAI-PMH-Endpoint (linked to identification) of repositories found on re3data.org, which use DDI as a metadata standard and provide an OAI-OMH-endpoint to this metadata

Repository	Records	DDI Related Prefixes	OAI-PMH-Endpoint (Link to Identification)
Harvard Dataverse	92,267	oai_ddi	https://dataverse.harvard.edu/oai
Open Forest Data	79,762	oai_ddi	https://dataverse.openforestdata.pl/oai
CESSDA Data Catalogue	30,289	oai_ddi25	https://datacatalogue.cessda.eu/oai-pmh/v0/oai
GESIS Data Archive	20,689	oai_ddi25, oai_ddi25-de, oai_ddi25-en, (oai_ddi32)	https://dbkapps.gesis.org/dbkoai/
EASY	9,804	oai_ddi25_en, oai_ddi25_nl	https://easy.dans.knaw.nl/oai
DataverseNL	7,446	oai_ddi	https://dataverse.nl/oai
Finnish Social Science Data Archive	3,934	oai_ddi25, ddi_c	https://services.fsd.tuni.fi/v0/oai
Swedish National Data Service	2,943	ddi25, (ddi33)	https://snd.se/oai-pmh
Texas Data Repository	2,139	oai_ddi	https://dataverse.tdl.org/oai
DaRUS	1,578	oai_ddi	https://darus.uni-stuttgart.de/oai
Austrian Social Science Data Archive	1,540	oai_ddi	https://data.aussda.at/oai
Repository for Open Data	1,458	oai_ddi	https://repod.icm.edu.pl/oai
SWISSUbase	979	oai_ddi25	https://www.swissubase.ch/oai-pmh/v1/oai
CORA. Repositori de Dades de Recerca	902	oai_ddi	https://dataverse.csuc.cat/oai
Social Science Japan Data Archive	867	oai_ddi25	https://ssjda.iss.u-tokyo.ac.jp/Direct/oai2/
heiDATA	586	oai_ddi	https://heidata.uni-heidelberg.de/oai
ICRISAT	446	oai_ddi	https://dataverse.icrisat.org/oai
Social Data Repository (RDS)	436	oai_ddi	https://rds.icm.edu.pl/oai
Macromolecular Xtallography Raw Data Repository	433	oai_ddi	https://mxrdr.icm.edu.pl/oai
UCLA	314	oai_ddi	https://dataverse.ucla.edu/oai
LibraData	243	oai_ddi	https://dataverse.lib.virginia.edu/oai
Trolling	171	oai_ddi	https://dataverse.no/oai
Repositório de Dados de Pesquisa Unifesp	76	oai_ddi	https://repositoriodedados.unifesp.br/oai
ASU Library Research Data Repository	74	oai_ddi	https://dataverse.asu.edu/oai
Debreceni Egyetem Adattár	59	oai_ddi	https://adattar.unideb.hu/oai
University of Warsaw Research Data Repository	59	oai_ddi	https://danebadawcze.uw.edu.pl/oai
UNB Libraries Dataverse Research Data Repository	56	oai_ddi	https://dataverse.lib.unb.ca/oai
osnaData	34	oai_ddi	https://osnadata.ub.uni-osnabrueck.de/oai
Repositorio de Datos Académicos Universidad Nacional de Rosario	22	oai_ddi	https://dataverse.unr.edu.ar/oai
Total	259,606		

Result: How DDI-Codebook 2.5 is used?

From the 252 different DDI-Codebook 2.5 elements 119 elements were used in our sample. DDI-Codebook also imports other schemas like Dublin Core, XHTML, or XML. But this feature is rarely used, and only one element from Dublin Core (<coverage>) is found[3] in 64 records we analyzed.

Within the DDI-Codebook payload, every DDI-Codebook 2.5 compliant XML file has the element <codebook> on level 1 as shown in the introduction. This element is mandatory and non-repeatable. On level 2 there are the elements <docDscr>, <stdyDscr>, <fileDscr>, <dataDscr>, and <otherMat>: <codebook> is used in every record once. Nearly all records use <docDscr>, only some use it twice. <studyDscr> is used once within all records. <fileDscr> appears in 20.9% of the records. Over 90% of

these records contain it only once, while the remainder include it up to 8 times. <dataDscr> is only used in 0.7% of all records. Finally, 47.9% of the records use <otherMat> (an element which can contain itself and therefore can exist in multiple locations), 50% of these records use it once (Median equals 1), 90% use it up to 16 times (90th percentile equals 16), the highest usage per record is 64,491 (see Table 4).

These results and information on the usage of all level 3 elements in DDI-Codebook 2.5 can also be found in Table 4: Like <otherMat> also the elements <citation> and <notes> can be used on different locations within the tree structure of a DDI-Codebook XML. As we did not account for the location in our analysis, we reported it for the first possible occurrence and then referenced this line.

Within the element <fileDscr>, information like file name, file type, fingerprint, dimensions, or a file description can be stored. Only 5 repositories provide information at this level. Therefore, only 20.9% of the analyzed records use this element (and the elements within). Only one repository uses this element two or more times, for example, to document a collection of multiple files.

Within <dataDscr>, only the element <var> is used (if one disregards <notes>). It is used only in 0.7% of the records. If used, the mean use is 116.3 times, the median use is 119, the 90th percentile of usage numbers is 337.3, the maximal usage is 3,188. Only one repository deploys this feature of DDI-Codebook and delivers information not only on dataset but on variable level.

The rare use of <fileDscr> and <dataDscr> is surprising: Information on files should be easily available for the repositories and it would make sense to inform the data users in advance about the file dimensions they can expect. If the data files are available as Stata, SPSS, or SAS files, the metadata for the area within <dataDscr> would be very inexpensive to extract, because most of the work has been done during the data curation process.

A complete list of all found elements and their usage statistics can be found in the Appendix: Table 5 and the complete dataset is also available online (Wenzig and Han, 2024). The dataset includes one repository that used the element <conOps>, which is not part of DDI-Codebook. The same Dataverse driven repository also used in one record <ConOps>, which is part of the standard.

Table 4: Usage statistics (usage in records, mean/median/90th percentile/max of use per record, number of repositories with element found) for all elements in DDI-Codebook up to level 3

Line	Element			Usage in Records	Mean	Median	90th Per-centile	Max	Used in Repos
	Level 1	Level 2	Level 3						
0	codeBook			100.0%	1.0	1	1.0	1	29
1		docDscr		99.7%	1.0	1	1.0	2	28
1.1			citation	100.0%	3.1	2	3.0	421	29
1.2			guide	0.0%					
1.3			docStatus	0.0%					
1.4			docSrc	0.0%					
1.5			controlledVocabUsed	0.0%					
1.6			notes	78.4%	14.4	2	9.0	64,491	24
2		stdyDscr		100.0%	1.0	1	1.0	1	29
2.1			citation			see line 1.1			
2.2			studyAuthorization	0.0%					
2.3			stdyInfo	100.0%	1.0	1	1.0	2	29
2.4			studyDevelopment	0.0%					
2.5			method	95.8%	1.0	1	1.0	2	27
2.6			dataAccs	99.9%	1.0	1	1.0	2	29
2.7			othrStdyMat	91.7%	1.0	1	1.0	2	26
2.8			notes			see line 1.6			
3		fileDscr		20.9%	1.0	1	1.0	8	5
3.1			fileTxt	11.6%	1.2	1	2.0	192	5
3.2			locMap	0.0%					
3.3			notes			see line 1.6			
4		dataDscr		0.7%	1.0	1	1.0	1	1
4.1			varGrp	0.0%					
4.2			nCubeGrp	0.0%					
4.3			var	0.7%	163.3	119	337.3	3,188	1
4.4			nCube	0.0%					
4.5			notes			see line 1.6			
5		otherMat		47.9%	22.1	1	16.0	64,491	22
5.1			labl	48.6%	39.6	1	20.0	64,491	23
5.2			txt	47.4%	22.2	1	16.0	64,491	22
5.3			notes			see line 1.6			
5.4			table	0.0%					
5.5			citation			see line 1.1			
5.6			otherMat			see line 5			

Recommendations

1. *Repositories:* The repositories should enrich the published metadata with information from the datasets by supplementing the metadata with information that is typically already encoded in Stata, SPSS, or SAS files, e.g., variable labels and value labels. Figure 2 shows a possible re-use of that information in a DDI-Codebook file.

2. *Dataverse Developers:* None of the records containing the element <fileDscr> have been published by a repository that reports using Dataverse software. We recommend that Dataverse should expose this preexisting information about files via OAI-PMH.

3. *re3data.org:* When we tried to access the different APIs using the given link found on re3data.org, we had to learn that the quality of data is often poor. Instead of an endpoint or a website with detailed information about the repository's API, there is too often only a link of the generic documentation of the software's APIs and no server specific information. re3data.org should consider ensuring that only

```
<codeBook>
    ...
    <fileDscr>
        <fileTxt>
            <fileName>FILENAME</fileName>
        </fileTxt>
    </fileDscr>
    <dataDscr>
        <var name="VARIABLENAME" files="FILENAME">
            <labl>VARIABLELABEL</labl>
            <catgry>
                <catValu>VALUE</catValu>
                <labl>VALUELABEL</labl>
            </catgry>
            ...
        </var>
        ...
    </dataDscr>
    ...
</codeBook>
```

Figure 2: Additional codebook information, that can easily be extracted from Stata, SPSS oder SAS files.

endpoints of registered OAI-PMH providers (https://www.openarchives.org/Register/BrowseSites) would be specified.

4. *DDI Alliance:* Obviously, there are use-cases for multiple metadata prefixes related to a single standard when providing access to the metadata of the repository. The specification of the OAI-PMH protocol does not allow to qualify the multiple usage of a single standard, apart from encoding the use-cases in the name of the metadata prefix. The specification document also states: "Communities should adopt guidelines for sharing metadataPrefixes, metadata schema and XML namespace URIs of metadata formats." (Open Archives Initiative 2015, section 3.4) The DDI community should consider providing guidance on which metadata prefixes should be used and what should be done, if one standard is used by more than one prefix. We recommend that the DDI Alliance starts a discussion about the use and structure of metadata prefixes.

5. *Repositories:* While trying to access the OAI-PMH server, we encountered several issues with data providers. First, repositories should try to improve server stability. Occasionally, a repository can respond very slowly or even disconnect, while at other times it works fine. Second, the metadata quality is not consistent, and it may not always be pre-checked by the repositories. As a results data collection may fail due to small errors in the XML.

Limitations

- In DDI-Codebook some elements are allowed on multiple locations. For example, the element <notes> can be used within all five elements on level 2 and 16 other locations. The element <otherMat> even can contain itself, theoretically infinite number of times. While the location

of usage may be of interest, in this first approach we only counted the usage numbers of the elements independently of their location.

- While all elements in DDI-Codebook support a basic set of attributes (ID, xml:lang, source, elementVersion, elementVersionDate, DDILifecycleUrn, DDICodebookUrnNo), the element <var> supports more than 30 attributes. The analysis of attribute usage has been out of the scope of this analysis but may be of interest in future research.
- We did not perform any content analysis or quality checks. Data professionals who harvest DDI metadata to provide it aggregated in catalogues, often report inconsistent usage of the different elements. However, the usage of elements we miss in nearly all records (e.g., those that describe variables in datasets) should be more straightforward.

Summary

We analyzed 259,606 metadata records in DDI-Codebook format from 29 different repositories. While all records contained information at the study level, and almost half of the records described other material, only 5 repositories (20.9% of the records) provide information on files and only one repository (0.7% of the records) provided detailed information at the variables level. While we did not collect information about where in the schema elements are used (if allowed on multiple locations) and did not analyze the usage of attributes, insights on the use of the standard might be valuable for developers and users of the standard.

There is a lack of availability of fine-grained metadata, however: “The sine qua non to greater automation of cross-domain data combination and analysis and fine-grained and responsive access control is sufficiently detailed, standardized, and interoperable metadata. There are no short cuts: data and metadata are hard.” (Hodson/Gregory 2023, p. 12)

The high correlation between the usage of Dataverse as a repository software and providing DDI metadata draws attention to Dataverse, because the efforts to include options to edit variable metadata in Dataverse (Lubitch 2023) will be relevant for the community.

Acknowledgements

We gratefully acknowledge the anonymous reviewers and Tom Hartl for their invaluable feedback and insightful comments, which significantly enhanced the quality of this manuscript. Any remaining errors are solely our responsibility.

References

All Links checked on August 10, 2024.

Cottage Labs (2021). SWORD 3.0 Specification. https://swordapp.github.io/swordv3/swordv3.html

DDI Alliance (2014). DDI-Codebook 2.5. https://ddialliance.org/Specification/DDI-Codebook/2.5/

DDI Alliance (2020). DDI Lifecycle 3.3. https://ddialliance.org/Specification/DDI-Lifecycle/3.3/

Fielding, R.T. (2000). Architectural Styles and the Design of Network-based Software Architectures – Dissertation. https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation_2up.pdf

Hodson, S., & Gregory, A. (2023). WorldFAIR Project (D1.3) First policy brief (Version 1). https://doi.org/10.5281/zenodo.7853170

Lubitch, V. (2023). Enhancing DDI Support in the Open Source Dataverse Repository Software [presentation]. https://doi.org/10.5281/zenodo.10636123

McCrae J.P. (2024). LOD cloud diagram. https://lod-cloud.net/

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

Wenzig, K. (2017). Next Time Try Recycling - What Reusable Metadata (Should) Look Like. https://doi.org/10.5281/zenodo.1084106

Wenzig, K. and Han, X. (2024). State of the DDI Cloud - Additional Datasets (Information on 259606 DDI-Codebook Records). https://doi.org/10.5281/zenodo.13255674.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Appendix I

Table 5: List of all found DDI-Codebook elements. Also available as CSV file in Wenzig and Han (2024).

Element	Usage in Records	Mean	Median	90th Percentile	Max	Used in Repos
abstract	100.0%	1.2	1	2.0	13	29
accsPlac	0.4%	1.8	2	2.0	2	6
actMin	0.0%	1.0	1	1.0	1	5
altTitl	0.7%	1.2	1	2.0	2	17
anlyInfo	73.6%	1.0	1	1.0	2	23
anlyUnit	16.9%	1.6	1	3.0	24	16
AuthEnty	100.0%	2.1	1	4.0	338	29
avlStatus	6.0%	1.1	1	1.0	2	7
biblCit	73.6%	1.2	1	2.0	379	23
caseQnty	5.1%	1.0	1	1.0	1	1
catgry	0.6%	886.0	641	1739.0	13567	1
citation	100.0%	3.1	2	3.0	421	29
citReq	4.4%	1.5	2	2.0	2	15
cleanOps	0.0%	1.0	1	1.0	1	9
codeBook	100.0%	1.0	1	1.0	1	29
collDate	56.8%	2.1	2	2.0	162	25
collectorTraining	0.0%	1.0	1	1.0	1	7
collMode	21.6%	1.7	1	2.0	74	17
collSitu	0.0%	1.0	1	1.0	1	9
collSize	0.1%	1.0	1	1.0	1	6
complete	0.0%	1.0	1	1.0	1	3
concept	14.7%	6.0	5	10.0	86	5
conditions	8.5%	1.0	1	1.0	2	11
confDec	0.1%	1.0	1	1.0	1	10
ConOps	0.0%	1.0	1	1.0	1	4
contact	73.2%	1.0	1	1.0	12	23
copyright	12.2%	1.7	1	4.0	5	4
dataAccs	99.9%	1.0	1	1.0	2	29
dataAppr	0.0%	1.0	1	1.0	1	1
dataColl	95.8%	1.0	1	1.0	2	27
dataCollector	5.5%	1.0	1	1.0	1	11
dataDscr	0.7%	1.0	1	1.0	1	1
dataKind	28.5%	1.2	1	2.0	35	25
dataSrc	1.2%	1.5	1	1.0	67	16
depDate	69.1%	1.0	1	1.0	1	22
depositr	64.5%	1.0	1	1.0	2	23
deposReq	3.6%	1.6	2	2.0	2	7
deviat	0.0%	1.0	1	1.0	1	5
dimensns	5.1%	1.0	1	1.0	1	1
disclaimer	0.9%	1.0	1	1.0	1	6
distDate	92.5%	1.9	2	2.0	419	28
distrbtr	96.2%	2.3	2	3.0	9	28
distStmt	100.0%	2.1	2	2.0	418	29
docDscr	99.7%	1.0	1	1.0	2	28
eastBL	41.3%	1.0	1	1.0	18	12
EstSmpErr	0.0%	1.0	1	1.0	1	3
ExtLink	12.1%	1.2	1	1.0	55	24
fileDscr	20.9%	1.0	1	1.0	8	5
fileName	4.2%	1.4	1	2.0	192	3
fileTxt	11.6%	1.2	1	2.0	192	5
fileType	5.1%	1.0	1	1.0	1	1
frequenc	0.1%	1.0	1	1.0	1	10
fundAg	4.4%	1.6	1	2.0	23	3
geoBndBox	41.2%	1.0	1	1.0	19	13
geogCover	55.4%	2.1	2	2.0	638	22
geogUnit	13.0%	1.0	1	1.0	28	15
grantNo	6.3%	1.3	1	2.0	19	21
holdings	59.8%	2.4	1	4.0	421	22
IDNo	100.0%	2.5	2	4.0	423	29
keyword	84.4%	7.5	3	13.0	721	27
labl	48.6%	39.6	1	20.0	64491	23
method	95.8%	1.0	1	1.0	2	27
nation	53.3%	2.3	1	2.0	390	24

Element	Usage in Records	Mean	Median	90th Percentile	Max	Used in Repos
northBL	41.3%	1.0	1	1.0	19	13
notes	78.4%	14.4	2	9.0	64491	24
origArch	0.3%	1.0	1	1.0	1	5
otherMat	47.9%	22.1	1	16.0	64491	22
othId	10.0%	2.7	2	4.0	130	16
othRefs	5.3%	1.2	1	2.0	14	12
othrStdyMat	91.7%	1.0	1	1.0	2	26
parTitl	12.0%	1.3	1	2.0	3	6
prodDate	41.1%	1.1	1	1.0	2	22
prodPlac	48.5%	1.0	1	1.0	1	20
prodStmt	97.0%	1.2	1	2.0	2	28
producer	49.9%	1.1	1	2.0	14	24
qstn	0.7%	163.3	119	337.3	3188	1
qstnLit	0.7%	220.3	133	468.0	6376	1
relMat	7.1%	6.4	4	14.0	201	18
relPubl	20.4%	2.8	1	3.0	417	26
relStdy	0.7%	2.1	1	4.0	23	18
resInstru	1.0%	1.0	1	1.0	1	9
respRate	0.4%	1.9	2	2.0	2	9
restrctn	21.3%	1.7	2	2.0	6	15
rspStmt	100.0%	1.1	1	1.0	2	29
sampleSize	0.2%	1.0	1	1.0	1	7
sampleSizeFormula	0.0%	1.0	1	1.0	1	2
sampProc	19.9%	1.7	1	2.0	14	17
serInfo	1.7%	1.9	2	2.0	4	16
serName	3.3%	1.4	1	2.0	4	17
serStmt	3.7%	1.4	1	2.0	4	19
setAvail	77.6%	1.0	1	1.0	2	23
software	1.2%	1.4	1	2.0	12	14
sources	73.1%	1.0	1	1.0	1	22
southBL	41.2%	1.0	1	1.0	18	12
specPerm	0.5%	1.0	1	1.0	1	7
srcChar	0.0%	1.0	1	1.0	1	8
srcDocu	0.1%	1.0	1	1.0	1	9
srcOrig	0.6%	1.0	1	1.0	1	12
stdyDscr	100.0%	1.0	1	1.0	1	29
stdyInfo	100.0%	1.0	1	1.0	2	29
subject	100.0%	1.0	1	1.0	2	29
subTitl	0.9%	1.0	1	1.0	2	18
sumDscr	100.0%	1.0	1	1.0	2	29
targetSampleSize	0.2%	1.0	1	1.0	1	7
timeMeth	16.3%	1.5	1	2.0	10	14
timePrd	34.8%	2.0	2	2.0	9	21
titl	100.0%	2.9	2	3.0	421	29
titlStmt	100.0%	3.0	2	3.0	421	29
topcClas	30.1%	4.8	3	10.0	128	24
txt	47.4%	22.2	1	16.0	64491	22
universe	15.4%	1.3	1	2.0	6	16
useStmt	98.7%	1.0	1	1.0	1	24
var	0.7%	163.3	119	337.3	3188	1
varQnty	5.1%	1.0	1	1.0	1	1
verResp	5.1%	1.0	1	1.0	1	1
version	91.8%	1.1	1	1.0	4	27
verStmt	95.8%	1.0	1	1.0	2	27
weight	5.3%	1.0	1	1.0	3	6
westBL	41.2%	1.0	1	1.0	19	13

Endnotes

[1] DIW Berlin/SOEP, kwenzig@diw.de

[2] DIW Berlin/SOEP, xhan@diw.de

[3] Example: https://easy.dans.knaw.nl/oai/?verb=GetRecord&identifier=oai:easy.dans.knaw.nl:easy-dataset:115768&metadataPrefix=oai_ddi25_en

DOI: https://doi.org/10.29173/iq1116