

Statistical Commission  
Fifty-fourth session  
28 February – 3 March 2023  
Item 3(c) of the provisional agenda  
**Items for discussion and decision: Household surveys**

Background document  
Available in English only

## **Standards and Good Practice for Survey Data Documentation**

Prepared by

World Bank

for

Inter-Secretariat Working Group on Household Surveys

<b>Contents</b>	
<b>Executive summary</b> .....	4
<b>Acknowledgments</b> .....	4
<b>Background and rationale</b> .....	5
<b>The case for rich metadata</b> .....	7
<b>Benefits for data users</b> .....	7
<b>Benefits for data producers and data curators</b> .....	7
<b>Scope of rich metadata</b> .....	8
Cataloguing material.....	9
Contextual information.....	9
Explanatory material.....	9
Licensing and terms of use.....	10
<b>Augmenting metadata</b> .....	10
<b>The case for structured metadata</b> .....	12
<b>Metadata standards and schemas</b> .....	13
<b>Controlled vocabularies</b> .....	14
<b>Benefits of structured metadata</b> .....	15
Completeness of the metadata.....	15
Usability of metadata.....	16
Data discoverability.....	16
Interoperability of applications.....	21
Visibility of data.....	21
Question banks and harmonization of data collection.....	23
<b>Recommended and suggested standards and schemas</b> .....	25
<b>Metadata standards and schemas by data type</b> .....	25
<b>A note on the SDMX standard</b> .....	26
<b>The DDI Codebook standard</b> .....	27
Purpose and versions.....	27
Scope and structure.....	27
Development and maintenance of the DDI Codebook metadata standard.....	30
DDI Codebook vs DDI Lifecycle.....	30
Users' community.....	31

<b>Metadata formats and tools</b> .....	31
<b>The JSON and XML formats</b> .....	31
<b>Tools for the production of standard-compliant metadata</b> .....	32
Multi-standard metadata editor .....	32
R packages, Python libraries, other utilities.....	34
DDI Codebook in software applications .....	35
<b>Tools for the exploitation of standard-compliant metadata</b> .....	35
Features enabled by rich and structured metadata .....	35
Filters / facets.....	35
Advanced search .....	35
Variable-level search and comparison .....	35
Data and metadata API .....	36
Recommendations.....	36
Efficiency and affordability of catalog maintenance .....	38
<b>Capacity building and support</b> .....	39
<b>A 10-tasks action plan</b> .....	39

## Executive summary

---

Statistical agencies and other data producers face an increasing demand for their microdata and a growing expectation for transparency in their data collection and production processes and methods. The responsible dissemination of microdata can increase the use and value of data, while consolidating the trust stakeholders have in these products. But it must be done responsibly, in accordance with legal and ethical rules and principles, and following technical best practice. This note focuses on a technical aspect of microdata dissemination: the production, dissemination, and use of metadata which are critical to ensure the visibility and discoverability of the data, their usability, and their credibility.

The note recommends the adoption of the Data Documentation Initiative (DDI) Codebook metadata standard. It provides a justification to produce rich and structured metadata, and advocates for the establishment of a community of practice around open standards and tools. The note also recommends a research program on data discoverability (recommender systems, semantic searchability). An action plan is proposed. These recommendations align with priorities expressed by many countries in the 2021 Survey on the Implementation of the Cape Town Global Action Plan for Sustainable Development Data. The note was produced for the Inter-Secretariat Working Group on Household Surveys (ISWGHS) Task Force on Metadata. The goal of the Task Force is to improve the quality and availability of survey metadata generated and published by national, regional, and international organizations, with a specific objective to formulate recommendations on survey data documentation standards and best practices, and to propose a common, structured framework to organize the content, presentation, transfer, and preservation of metadata.

## Acknowledgments

---

This note was prepared by Olivier Dupriez and Mehmood Asghar, draws on a draft document titled *Metadata Standards and Schemas for Improved Data Discoverability and Usability* by the World Bank Data Group.<sup>1</sup> It was submitted for review to the organizations represented in the ISWGHS and was discussed at a meeting of the ISWGHS held in Washington, DC on 15 September 2022. Comments and suggestions were received from multiple organizations. Input from Babatunde Abidoye (UNDP), Haoyi Chen (UNSD), Pietro Gennari (FAO), Luis Gerardo Gonzalez Morales (UNSD), Abdulla Gozalov (UNSD), Yves Jaques (UNICEF), and Matthew Welch (World Bank) is gratefully acknowledged.

---

<sup>1</sup> <https://mah0001.github.io/schema-guide/>

## Background and rationale

---

In most countries, data producers are faced with an expanding demand for access to the underlying microdata on which published statistics are based. Access to microdata enables new and more diverse research. It also allows the development of innovative ways of using, processing, and displaying information, and the generation of new datasets by combining data from multiple sources such as satellite imagery.<sup>2</sup>

Responsible microdata dissemination is guided by legal and ethical rules and principles and comes with technical constraints and requirements. To be responsibly more openly shared and used, microdata must not only be anonymized and accessible, but also made more visible, discoverable, understandable, and usable. For all these purposes, the production and dissemination of good metadata is essential. This note builds the case for the adoption of metadata standards and schemas to generate, publish, and exploit richer metadata.

Many statistical agencies have endorsed the Generic Statistical Business Process Model (GSBPM) which defines the production of metadata as an overarching requirement. “Good metadata management is essential for the efficient operation of statistical business processes. Metadata are present in every phase, either created or carried forward from a previous phase. In the context of this model, the emphasis of the over-arching process of metadata management is on the creation, use and archiving of statistical metadata, though metadata on the different sub-processes themselves are also of interest, including as an input for quality management. The key challenge is to ensure that these metadata are captured as early as possible and stored and transferred from phase to phase alongside the data they refer to. Metadata management strategy and systems are therefore vital to the operation of the model (...).”<sup>3</sup>

But investments in the production, dissemination, and use of metadata, and in enabling and promoting the secondary use and re-purposing of data, have not been in par with the attention and resources devoted to data collection and production. A survey of national statistical organizations (NSOs) conducted in 2021<sup>4</sup> indicated a strong interest in addressing this issue. Most NSOs (86 percent in low and lower-middle income countries) identified strengthening the compilation and dissemination of metadata as the top priority to support the adoption of open data principles and practices (Figure 1). Almost nine in ten NSOs (97 percent in low and lower-middle income countries) identified the strengthening of online data dissemination platforms and tools as a top priority to enhance data dissemination capacity (Figure 2).

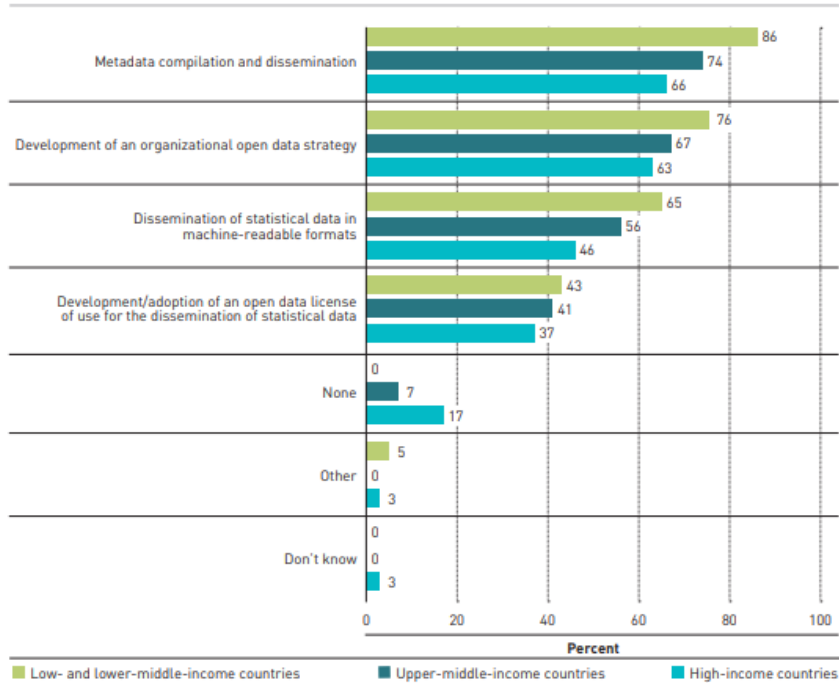
---

<sup>2</sup> Dupriez, Olivier and Ernie Boyko. 2010. “Dissemination of Microdata Files. Formulating Policies and Procedures”, International Household Survey Network, IHSN Working Paper No 005

<sup>3</sup> UNECE 2017; see <https://statswiki.unece.org/display/GSBPM>

<sup>4</sup> See <https://documents1.worldbank.org/curated/en/826351643712794722/pdf/Survey-on-the-Implementation-of-the-Cape-Town-Global-Action-Plan-for-Sustainable-Development-Data.pdf>

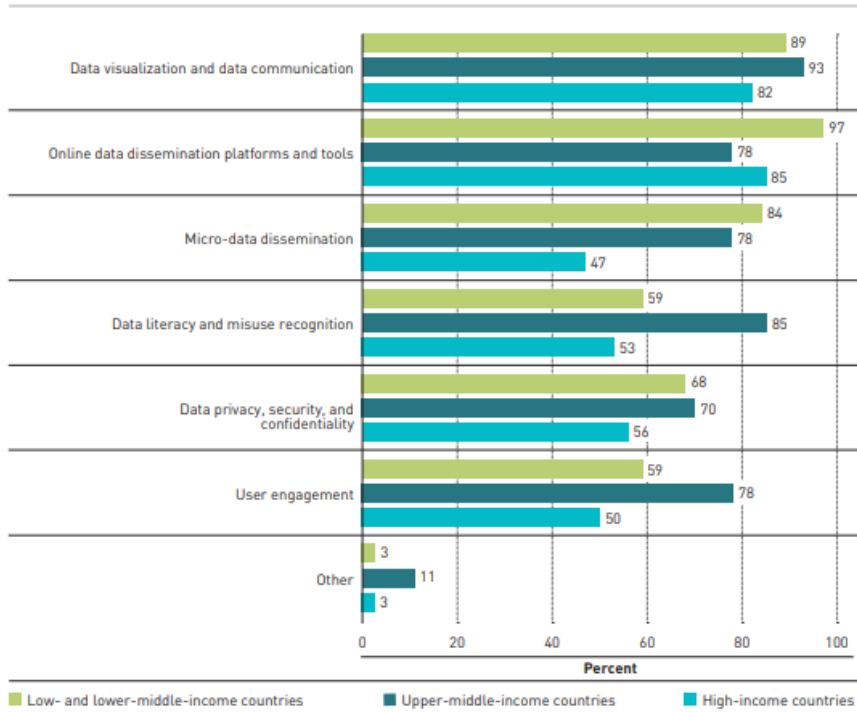
**Question: Please select priority areas that need to be strengthened in order to enable the adoption of open data principles and practices in your organization**



**N=99**

*Figure 1 – 2021 NSO survey results on priority areas for open data implementation*

**Question: What are the areas where the NSO wants to further strengthen their data dissemination capacity over the next three years?**



**N=98**

*Figure 2 – 2021 NSO survey results on priority areas for data dissemination capacity*

In this note, we focus on metadata related to survey microdata and on their production by national statistical agencies and international organizations. But the call and recommendation for improved practice and for the adoption of international metadata standards applies to other types of data and other data producers and curators. This will require the further development of a **toolset** to facilitate the adoption of metadata standards and best practice by statistical agencies. This toolset includes software applications, metadata standards, and training materials, all published as open public goods maintained by a **community of practice**. To ensure global relevancy and suitability for different environments (in terms of skills, IT infrastructure, and financial resources) the toolset must be modular and flexible. A dedicated training and technical assistance program would support the implementation of the tools and practice in resource-constrained agencies.

In the note, we describe the benefits and advocate for rich and structured metadata and for the **adoption of the DDI Codebook metadata standard for the documentation of survey microdata**. We provide practical information on their implementation and propose an action plan of 10 tasks that would allow the NSO community to achieve fast progress in this area.

## **The case for rich metadata**

---

By *rich metadata*, we mean metadata that are **detailed** and **comprehensive**. Metadata must cover the whole life cycle of the data product, not just the final product. For a survey dataset, metadata should at least cover the phases of questionnaire design, sampling, data collection, data processing and editing, tabulation and analysis, and must provide a detailed description of all available data files and variables (a data dictionary).

### **Benefits for data users**

Rich metadata help data users to:

- **Find and access data of interest.** Data catalogs and resource location systems (including search engines like Google) index metadata to make the datasets discoverable. The richer the metadata, the more the search engine will be able to identify and return relevant resources. Incomplete metadata results in a loss of visibility and discoverability.
- **Understand and use the data.** Metadata help users understand what the data are measuring, and how and for what purposes they have been created. When provided with incomplete documentation, data users may misunderstand—and possibly misuse—the data. Some users may (wisely) decide not to use the data at all if the accompanying metadata is insufficient.
- **Assess the quality of the data.** Detailed metadata allows data users to evaluate the reliability and fitness for purpose of a dataset, and to assess their consistency with other datasets when the data must be integrated with other datasets.

### **Benefits for data producers and data curators**

Rich metadata help data producers and data curators to:

- **Build trust in the data** by ensuring **transparency, auditability, and credibility** of the data and of products derived from the datasets. Principle 3 of the Fundamental Principle of Official Statistics states that “To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics”. Rich metadata, including the

documentation of limitations of the data, will strengthen the credibility of the data producer and of their products.

- Increase the **visibility** and **discoverability** of data. The volume and diversity of data available to secondary users is growing fast. Most official data producers publish their data on-line, but often with limited or no strategy to maximize their visibility and accessibility. Queries on search engines like Google or Bing show that official data producers often perform poorly compared with non-official sources in terms of visibility (see section “The case for structured metadata”). Data users are not always guided to the most appropriate source of data. Even when searches are performed directly on the statistical agencies’ respective websites, they often return no information, or non-relevant information, or insufficient information. Search engines need to be optimized to bring the most relevant data to the attention of users; this optimization requires rich (and structured) metadata.
- **Increase the use and value of data.** Data with good documentation will be more visible, discoverable, and usable. This will result in increased use and value of the data.
- **Reduce the risk of accidental misuse of the data.** By making data more usable and their production process more transparent, data producers reduce the risk that their data may be accidentally mis-interpreted and analyzed.
- **Reduce the cost of data dissemination.** Detailed metadata can be generated and disseminated at a small cost. It is an investment that may significantly reduce the cost of having to respond to users’ requests for information.
- **Harmonize and integrate data** across sources and types, and over time. Detailed metadata are needed both for ex-post harmonization (integrated use of data from diverse sources that are not compatible by design) and for ex-ante harmonization (to help identify where classifications, questions, concepts, and methods used by an agency or statistical system could or should be harmonized). Rich metadata provide the meta-database needed to harmonize data collection methods and instruments.<sup>5</sup>
- **Improve the quality of future data collection.** Metadata help data producers identify weaknesses and/or inconsistencies within and across data sources. They are a necessary input to data quality assessment and to the enhancement and harmonization of data collection instruments and methods.
- **Preserve institutional memory.** The process of data production and analysis is changing at a fast pace. Many statistical agencies have specific programs of innovation and are incubating new practice, often involving new modes of data acquisition and the exploitation of new data sources. Good documentation of the data and of the processes involved in their production is critical to preserve institutional memory, train new staff, and share experience with peers in other statistical agencies, especially those who do not have resources to lead innovation and incubation.

### Scope of rich metadata

What makes metadata *rich* is specific to each data type (e.g., the metadata for a survey dataset will differ from the metadata for a geographic dataset). Typically, the metadata will include

---

<sup>5</sup> A compelling case for rich metadata for transparency and harmonization was made by Gordon Priest in “*The Struggle for Integration and Harmonization of Social Statistics in a Statistical Agency - A Case Study of Statistics Canada*” ([www.ihsn.org/sites/default/files/resources/IHSN-WP004.pdf](http://www.ihsn.org/sites/default/files/resources/IHSN-WP004.pdf)).



cataloguing material, contextual information, explanatory materials, and information on the data terms of use.

### Cataloguing material

Cataloguing material are elements such as the dataset title and unique identifier (such as a digital object identifier - DOI), a version number, as well as information related to the data curation (who generated the metadata and when). This information allows the dataset to be uniquely identified within a collection and to be properly cited in publications.

### Contextual information

Contextual information describes the context in which the data were collected. It enables secondary data users to understand the background and processes behind the data production.

Contextual information should cover topics such as:

- The justification for the data collection (objectives, mandate of the data producer).
- The population or universe of the study (the persons or entities covered by the data collection).
- The geographic and temporal coverage of the data.
- Changes and developments that may have occurred over time in the data collection and processing methodology. For cross-section or panel surveys, this may include information describing changes in the question text, sampling procedures, and others.
- The output of the data collection (publications and reports, list of cross-tabulations and indicators, etc.)
- A description of problems encountered in the implementation of the study, and a description of known weaknesses and limitations of the data.
- Other useful information on the life cycle of the dataset.

### Explanatory material

Explanatory materials ensure the usability of the dataset, and include:

- Information about the data collection and processing methodology: survey instruments used, sampling design and sampling frames, mode of data capture (paper-and-pen, CAPI, phone interviews, web-based interviews), procedures of data quality control and editing, etc.
- Information about the data sources: when the source consists of responses to survey questionnaires, each question and the related interviewer's instructions and universe should be part of the documentation.
- Information about the structure of the dataset, including a detailed data dictionary. The data dictionary should include variable and value labels, an identification of key variables (used to merge data files) and of variables that define the sample design (stratification, primary sample units, sample weights), the number of valid and missing observations and other summary statistics (like frequencies for categorical variables) for each variable.
- Detailed information about derived and imputed variables (recoding instructions or description of imputations and derivations methods).
- Confidentiality and anonymization: if perturbative or non-perturbative statistical disclosure control methods have been applied to prevent identification of respondents,

some information should be provided on how this affects the data (taking care of not providing information that would enable a reverse-engineering of the procedure).

### Licensing and terms of use

- Microdata must be published with clear, formal terms of use that define who can access and use the data, under what conditions, and for what purpose. Ideally, a standard license will be used.<sup>6</sup>

### **Augmenting metadata**

Most of the metadata related to a survey will be provided by the data producers and data curators, ideally as a continuous process during the survey life cycle to guarantee that the best-informed contributors provide the required information. Some information may be extracted programmatically from the data files (like the list of variables, variable and value labels, or variable-level summary statistics). Once ready, these metadata can be *augmented*. **Metadata augmentation** consists of adding information obtained from an external source, and/or generated using machine learning tools.

Adding information from an external source will typically aim to provide search engines with terms and phrases (stored as keywords, topics, tags, or others) that are not found in the contextual information or exploratory materials. For example, the dataset of a UNICEF Multiple Indicator Cluster Survey (MICS) will often include anthropometric variables “age in month”, “weight”, and “height” for children aged 0 to 59 months. These variables are intended to calculate malnutrition indicators such as the percentage of wasting and stunting children. Adding keywords like “malnutrition”, “wasting”, and “stunting”, which are not found in the data dictionary, will increase the discoverability of the data. Including a list of key indicators that can be generated from the dataset can easily be implemented for surveys like the MICS, Demographic and Health Surveys (DHS), or labor force surveys which are designed to generate indicators. It is more difficult to implement on multi-topic surveys like the Living Standards Measurement Study (LSMS) which cover many topics and are intended to provide input to research and analytical work.

Another option to augment metadata is to exploit natural language processing (NLP) algorithms to automatically extract lists of relevant topics or keywords and/or to generate *embeddings* for the metadata. Topics/keywords can be extracted by applying models like the Latent Dirichlet Allocation (LDA) to the survey metadata (and possibly to the survey reports and publications). *Embeddings* are numeric representations (in the form of a large-dimension vector) of the semantic content of a document (in this case of the survey metadata). Embeddings are used to **implement semantic searchability** (by measuring the distance between a numeric vector representing the survey, and one representing the user query). The approach is extensively used by search engines like Google or Bing, which exploit machine learning solutions to “understand” user queries and to optimize the ranking of the results of the queries. It is not yet prevalent in data catalogs. Currently, most catalogs rely on keyword-based search (full-text matching), which perform poorly in many situations. As an example, Figure 3 shows the most relevant results returned for a query for data on “dutch disease” (an economic concept) in the World Bank Development Data Hub (the Bank’s central data catalog). The reliance on a keyword-based search makes the search engine assume that the query is related to health and to the Netherlands and returns entries unrelated to the actual concept of dutch disease.

---

<sup>6</sup> For example, Attribution 4.0 International — CC BY 4.0 - Creative Commons 9  
<https://creativecommons.org/licenses/by/4.0/>

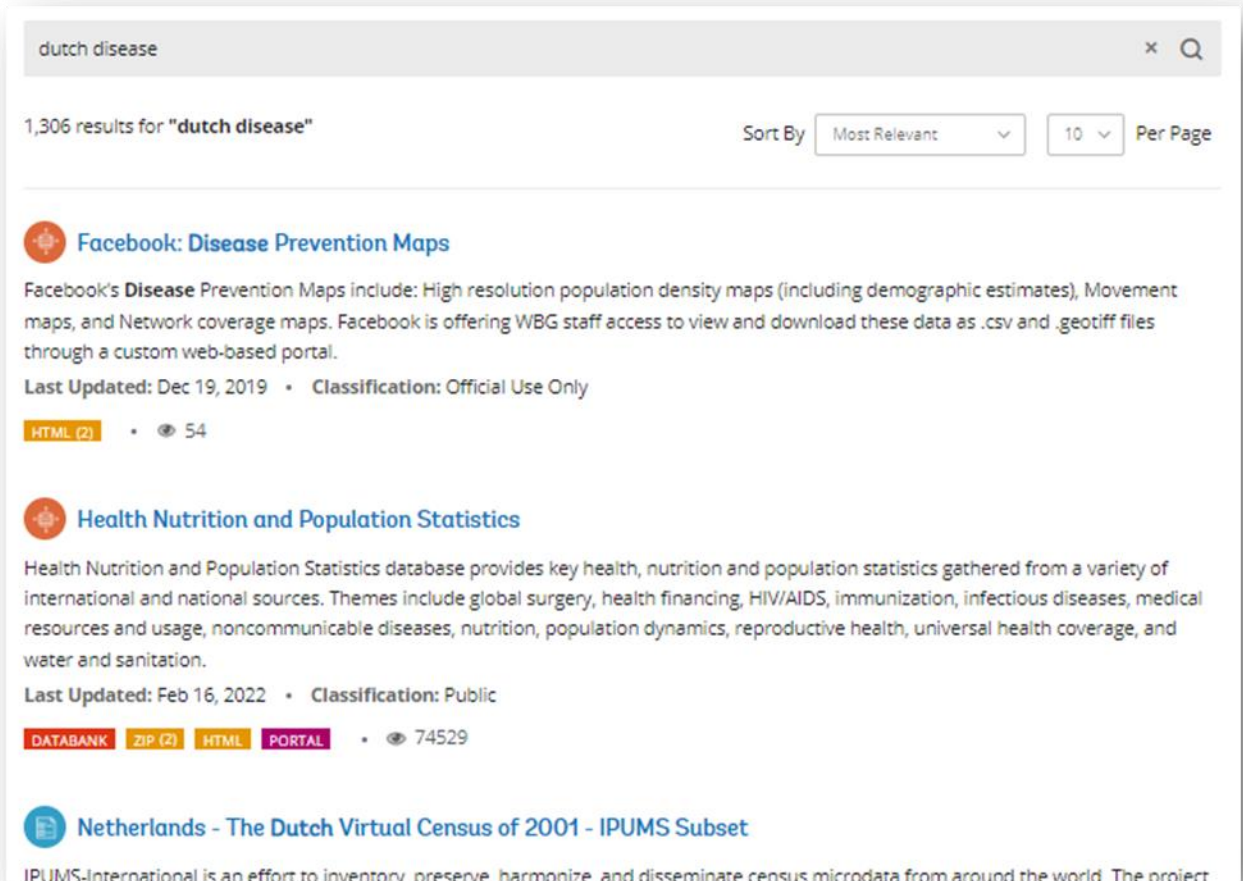


Figure 3 - Searching for "dutch disease" in the World Bank Development Data Hub

A search engine that exploits word embeddings will understand the economic nature of the concept and automatically associate it with closely related concepts (Figure 4).<sup>7</sup> Semantic closeness will in this case provide more relevant results than a keyword search. Embeddings will be particularly useful to build recommender systems that can associate concepts (like “economic growth”, or “demographic transition” for example) with the name or labels of indicators or variables found in survey metadata.

<sup>7</sup> The example is from the NLP Explorer, an application developed by the World Bank Data Group to explore the potential of NLP for improving data discoverability. The project trained LDA and embedding models on a corpus of about 350,000 documents, related to social and economic development issues. A larger corpus of > 1 million documents has been collected to train a new version of the models.

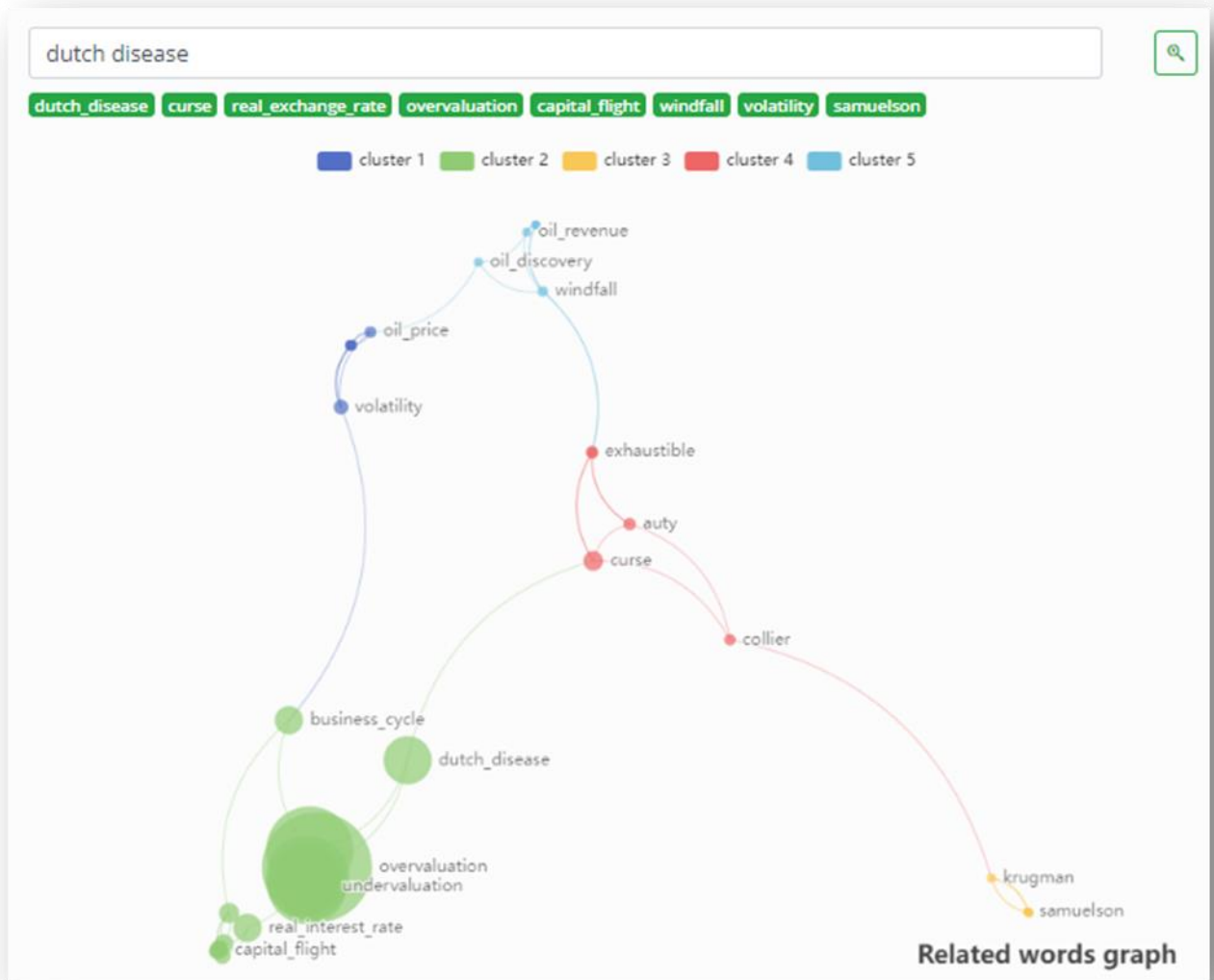


Figure 4 - Related word graph obtained from a machine learning embedding model

To provide users with the best search experience, the respective advantages of semantic searchability and full-text search must be combined. But training machine learning models and optimizing search engines for data discoverability are not trivial exercises. LDA and embedding models must be trained on large corpora of documents to be accurate and relevant. And the automatic mapping between concepts and statistical indicators or variables in datasets is a challenge that still requires research and exploratory work. This should be the objective of a concerted, multi-agency partnership.

Trained models, when available, can be made openly accessible to data catalogs developers and administrators (including via APIs), allowing them to implement semantic searchability, keyword suggestions, and eventually transform their catalogs into advanced recommender systems.

### The case for structured metadata

Metadata should not only be comprehensive and detailed, but also organized in a *structured* manner, preferably using a *standardized* structure. A standard structure will foster the

completeness of the metadata, help application developers build inter-operable metadata production and dissemination systems, and increase the discoverability and visibility of the (meta)data by enabling advanced indexing and search engine optimization (SEO) solutions. It will also allow data producers to build and maintain question banks.

## Metadata standards and schemas

**Structured** metadata means that the metadata are generated and stored in specific fields (or elements) organized in a metadata standard or schema<sup>8</sup>. **Standardized** means that the list and description of elements are not specific to an agency but are agreed by a community of practice. Standards and schemas must be specific to each of the main data types. The metadata elements needed to document microdata would be different from the standards used to document geographic datasets, or publications, or images, etc. For a statistical agency, the most relevant data types are microdata (for surveys, censuses, and possibly administrative datasets), time series/indicators, documents and reports, geographic data (vector and raster datasets), and statistical tables. Other types (images, audio recordings, video files, and programs and scripts) may be relevant too.

**Metadata standards and schemas** consist of commonly agreed and well-documented lists of elements to be used to document a dataset of a certain type. These lists are typically provided in XML or JSON format, as their content is intended to be stored in databases and exploited by computer applications. Each element in a standard or schema has a name, a type, a description, and can be set as repeatable or non-repeatable, as required or optional, and can contain sub-elements. For example, the title of a survey dataset would be stored in an element title of type *string* (i.e., it contains text), made *required* and *non-repeatable* (a survey must have one, and only one, official title). For that same survey, the contribution of sponsoring agencies could be documented in an *optional* element named *funding\_agencies*, *repeatable* (as a survey may have more than one sponsor) and comprising four sub-elements applicable to each sponsor: name, abbreviation, *grant\_number*, and role.

Metadata compliant with standards and schemas will typically be stored as a JSON dictionary or an XML file. Both are plain text files, non-proprietary and platform independent. The example below shows how a simple free-text content could be structured and stored in JSON:

### **Free text version:**

*The Child Mortality Survey (CMS) was conducted by the National Statistics Office of Popstan from July 2010 to June 2011, with financial support from the Child Health Foundation (trust fund TF123\_456) and from the Ministry of Health.*

### **Structured, machine-readable (JSON) version:**

```
{
  "title"      : "Child Mortality Survey 2010-2011",
  "alternate_title" : "CMS 2010-2011",
  "nation"     : [{"name":"Popstan", "abbreviation":"POP"}]
  "authoring_entity": [{"name":" National Statistics Office ", "abbreviation":"NSO"}],
  "funding_agencies": [{"name":"Child Health Foundation (CHF)", "grant":"TF123_456"},
                       {"name":"Popstan Ministry of Health"}],
  "coll_dates"  : [{"start":"2010-07", "end":"2011-06"}],
}
}
```

---

<sup>8</sup> Metadata “standards” are “schemas”; we consider that a schema is a standard when its development and maintenance is governed by a recognized organization (e.g., ISO19115 for geospatial by ISO, DDI by the DDI Alliance).

*Although* both versions contain the same information, the structured version is more suitable for publishing in a meta-database. Organizing, storing, and publishing metadata in a machine-readable and structured format enables all kinds of applications. It becomes straightforward for example to apply filters (e.g., a filter by country using the nation name or abbreviation element), or to enable targeted searches to answer questions like “What data are available for year 2010?” or “What surveys did [sponsor X] finance?” or “What dataset contains variables on disability”? Information contained in structured metadata can be made accessible not only in web interfaces (data catalogs) but also via API.

The number of elements in a metadata standard or schema can be large, as an element must be provided for every possible piece of information that may be available. In practice, data curators will only make use of a subset of the available elements and develop their own templates to capture metadata. Generating complete and detailed metadata may be seen as a burden by some organizations, and some may be tempted to make use of an overly simplified template. But generating detailed metadata will typically represent a very small fraction of the time and budget invested in the production of the data; it represents a small investment that adds much value to the data.

Some metadata standards have originated from the academia, like the Data Documentation Initiative (DDI), maintained by the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan. Others found their origins in specialized communities of practice (like the ISO 19139 for geospatial resources). The private sector also contributes to the development of standards, like the International Press Telecommunications Council (IPTC) standard developed by and for news media to document collections of images, or the more generic schema.org developed jointly by Google, Microsoft, Yahoo and Yandex for a related but somewhat different purpose.<sup>9</sup>

Some types of data are not (yet) covered by formally established metadata standards. This is the case for time series/indicators, statistical tables, and reproducible analytics. The World Bank and IHSN have developed a set of JSON schemas for documenting and cataloguing such data, by compiling and organizing metadata elements found in multiple sources (such as the World Bank Development Indicators database and the United Nations SDG database, for the time series metadata schema). These schemas have not been subject to a formal review process, and their maintenance is not subject to formal governance mechanisms, so they cannot be qualified as *standards*.

National statistical agencies and international organizations who have the mandate and expertise to produce and disseminate data and metadata should be active in the development and maintenance of metadata schemas and standards.

### **Controlled vocabularies**

Metadata standards and schemas provide structured lists of elements to be used to store and organize metadata. But they do not (with a few exceptions) dictate what these elements should contain. Controlled vocabularies can be powerful complements to the standards and schemas. Controlled vocabularies are pre-defined lists of options for the content of a metadata element. They enable filtering options in data catalogs. For example, a pre-defined list of country names can be provided as a controlled vocabulary for the metadata element “nation/name” of the DDI standard. Imposing (when relevant) constraints on the content of metadata elements helps foster

---

<sup>9</sup> [Schema.org](http://Schema.org) is not specifically intended to document datasets to ensure their transparency and usability; the schema(s) focus on the on-line discoverability of resources.

compatibility across data catalogs and solves issues of inconsistent metadata. For example, the name of a same country should not vary in a data catalog; using a controlled vocabulary, one can ensure that the Democratic Republic of Congo would not be referred to as “Congo, DR”, “Congo, Dem. Rep.”, “RD Congo”, etc. depending on the data curator.

Controlled vocabularies should be developed following the FAIR principles - Findable, Accessible, Interoperable, Reusable<sup>10</sup>. “*The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.*”<sup>11</sup>

Controlled vocabularies can be established at an organization level, or more globally. For interoperability of data catalogs, adopting standard vocabularies has much value (e.g., topics classification). Agencies from the United Nations<sup>12</sup>, the ISO organization<sup>13</sup>, and others are the custodians of the main classifications and controlled vocabularies.

When the use of a controlled vocabulary is relevant or recommended, metadata standards provide a sub-element to identify the name and source of one or multiple vocabularies (Figure 5). For example, to document a list of topics that a dataset covers, one may use an agency-specific taxonomy of topics and the topic classification proposed by the Consortium of European Social Science Data Archives (CESSDA)<sup>14</sup>.

```
- "topics": [  
  - {  
    "topic": "string",  
    "vocab": "string",  
    "uri": "string"  
  }  
],
```

Figure 5 - Structure of the Topics element in the DDI metadata standard (repeatable element)

## Benefits of structured metadata

The use of metadata standards and schemas offers multiple advantages: it fosters **completeness and quality of the metadata**, it enables **inter-operability of software and systems**, it facilitates **metadata exchange** across organizations, it contributes to **data usability, visibility, and discoverability**, and it facilitates the **harmonization of data collection** instruments.

### Completeness of the metadata

When they document datasets, data curators who do not make use of metadata standards and schemas tend to focus on the readily available documentation. They will often omit some information that secondary data users—and search engines—may need. Metadata standards and schemas operate as checklists of what information could or should be provided. These checklists

<sup>10</sup> See <https://www.go-fair.org/fair-principles/i2-metadata-use-vocabularies-follow-fair-principles/> and <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009041>

<sup>11</sup> Source: <https://www.go-fair.org/fair-principles/>

<sup>12</sup> For example, the International Standard Classification of Occupations (ISCO, by ILO), the International Statistical Classification of Diseases and Related Health Problems (ICD, by WHO), or the International Standard Classification of Education (ISCED, by UNESCO).

<sup>13</sup> Such as ISO 3166 for country codes, or ISO 639-1 for languages.

<sup>14</sup> See <https://vocabularies.cessda.eu/vocabulary/TopicClassification?v=3.0>



are developed by experts and are regularly updated or upgraded based on feedback received from a large community of practitioners. They help data curators think of all relevant information they should include in the metadata. They can also be used to assess the completeness of metadata (e.g., to identify variables or values with no labels, sample survey with no information on sample frame, etc.) With structured metadata and controlled vocabularies, customized diagnostic tools can easily be developed.

Documenting a dataset should not be seen as a last and independent step in the implementation of a data collection or production project. Ideally, and to foster accuracy and completeness, metadata will be captured continuously and in quasi-real time during the entire life cycle of the data collection/production and contributed by those who have the best knowledge of each phase of the data production process.

### Usability of metadata

Fully understanding a dataset before conducting analysis should be a pre-requisite for all researchers and data users. But this will only be possible when the data documentation is easy to access and use. Structured metadata stored in XML or JSON format provide such convenience, as they can be transformed into bookmarked PDF documents, searchable websites, machine-readable codebooks, etc. API accessibility to structured metadata can also facilitate other types of applications, such as automatic reporting on survey catalog content, harvesting of metadata by “aggregators”, and others.

### Data discoverability

The introduction statement of the European Union’s INSPIRE Directive<sup>15</sup> states that “The loss of time and resources in searching for existing resources (...) or establishing whether they may be used for a particular purpose is a key obstacle to the full exploitation of the data available”. This statement refers to spatial datasets and services. But it is equally valid for many other data types including survey microdata. Locating, accessing, and using survey microdata remains a significant challenge for data analysts. We illustrate this with a simple example<sup>16</sup>, of a researcher looking for data on “*data on disability in Nigeria*” (Figure 6). A search on Google will return links to blogs, documents, and other types of resources, but no link to microdata.

---

<sup>15</sup> INSPIRE - Infrastructure for Spatial Information in Europe, Recommendations for INSPIRE Spatial Data Services, 2011 (available at [https://inspire.ec.europa.eu/documents/Spatial\\_Data\\_Services/Spatial%20Data%20Services%20Working%20Group%20Recommendations%20v1.1.pdf](https://inspire.ec.europa.eu/documents/Spatial_Data_Services/Spatial%20Data%20Services%20Working%20Group%20Recommendations%20v1.1.pdf))

<sup>16</sup> Searches performed on 24 March 2022.



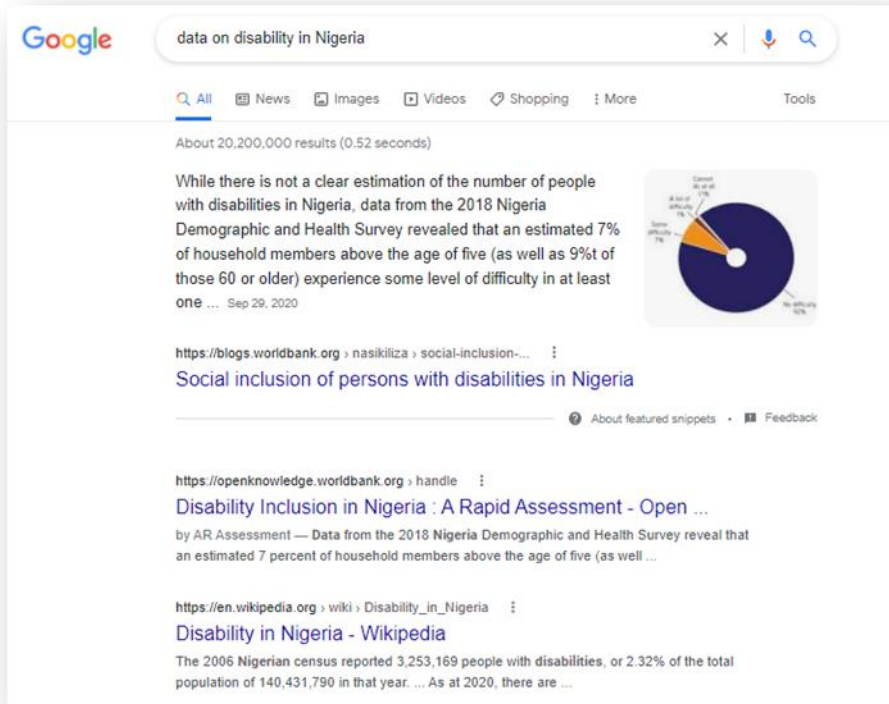


Figure 6 - Searching for "data on disability in Nigeria" on Google

The researcher may then submit a more specific query for “survey microdata on disability in Nigeria”. This will return more relevant results, listing three potentially useful datasets (Figure 7).<sup>17</sup>

<sup>17</sup> The three datasets have been documented using the DDI metadata standard and their metadata published respectively by ILO and the World Bank in a catalog that embeds search engine optimization procedures.

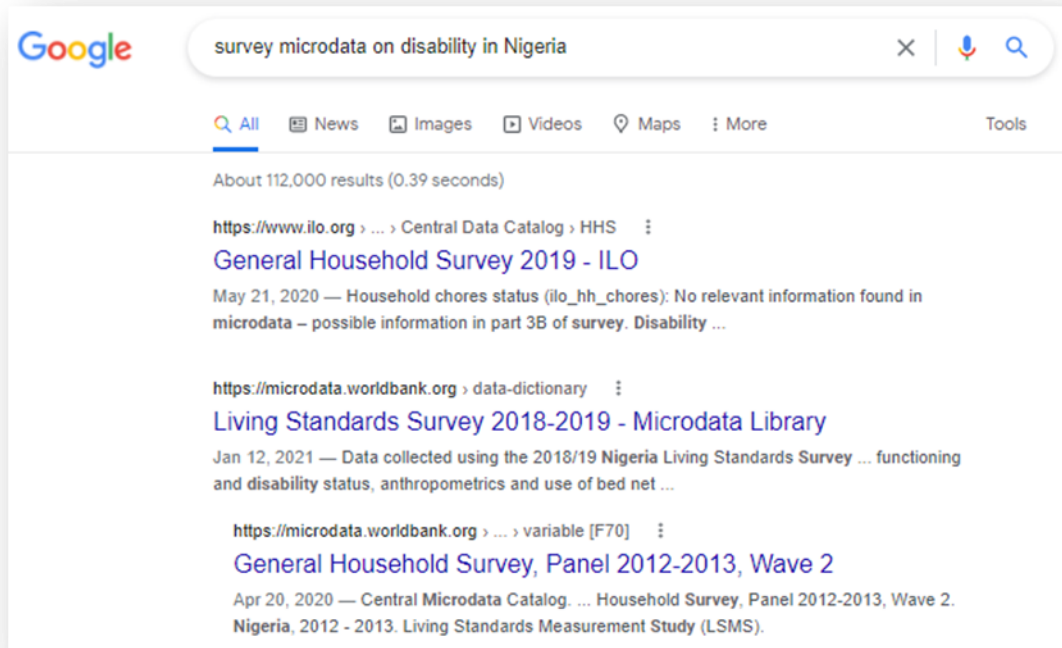


Figure 7 - Searching for “survey microdata on disability in Nigeria” on Google

Alternatively, the researcher may search for data on **Google Dataset Search**, Google’s search engine dedicated to finding data. The query “*microdata disability in Nigeria*” will return two datasets: a 2020 COVID-19 survey and the 2018 Demographic and Health Survey (Figure 8). Note that for some reason, a query for “*microdata disability Nigeria*” only returns one of these two surveys, which further illustrates the somewhat unpredictable behavior of search engines (Figure 9).

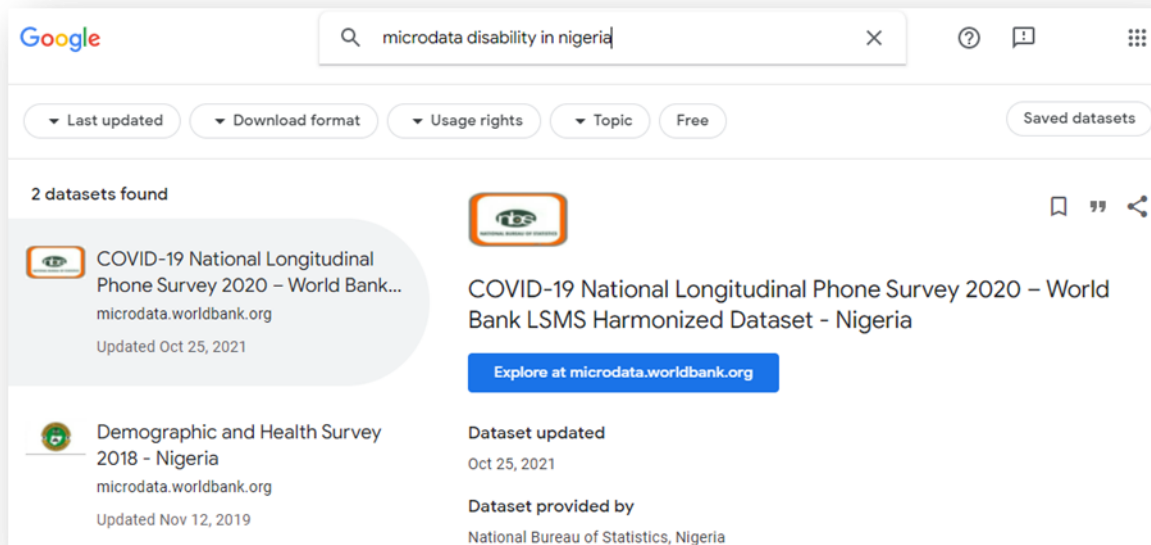


Figure 8 - Searching for “microdata disability in Nigeria” on Google Dataset Search

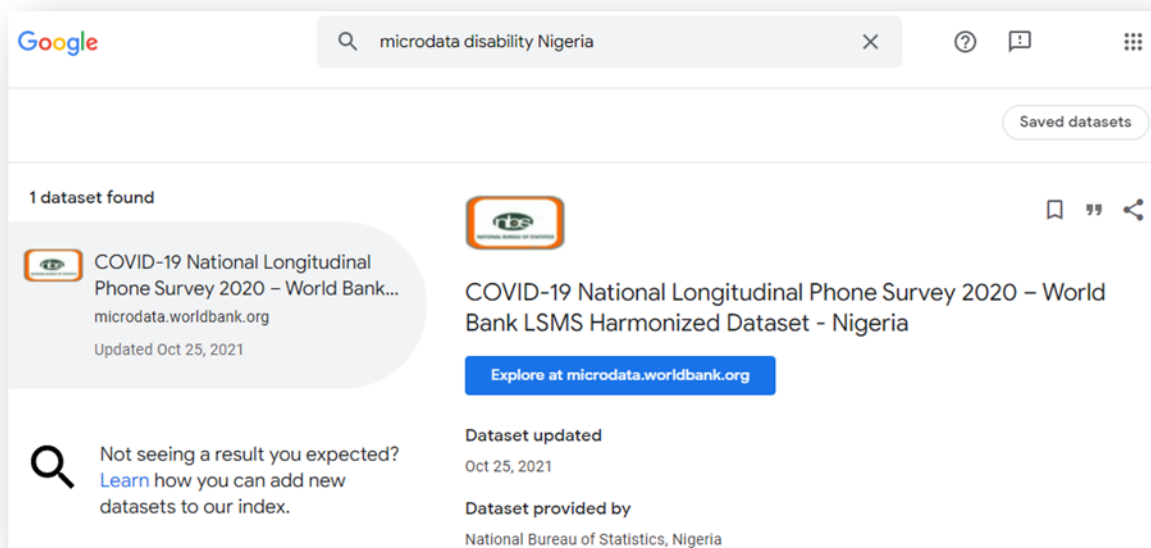


Figure 9 - Searching for “microdata disability Nigeria” on Google Dataset Search

If the researcher is aware of the existence of the World Bank Microdata Library<sup>18</sup>, s/he may use the Microdata Library’s tool to filter datasets by country and search for keywords in the description of variables. A filter on “country = Nigeria” and search for keyword “disability” will return 14 datasets. A similar search in the IHSN catalog<sup>19</sup> returns 27 datasets. This diversity and unpredictability of results shows the importance of addressing the issue of discoverability of data.

<sup>18</sup> See <https://microdata.worldbank.org/>

<sup>19</sup> See <https://catalog.ihsn.org/home>

Finding a keyword in a dataset is not enough to assess the relevancy of the data for a specific purpose. Detailed and comprehensive metadata will help users assess the relevancy of a specific dataset to their purpose. Taking the example of “disability in Nigeria”, what will make the data fit for purpose may be the geographic coverage of the survey (e.g., national coverage, or sub-national representativeness), the year of data collection, the comparability of the relevant variable with data from another source, the nature of the information collected, the sample size, and more. This information should be available in the published metadata.

This example shows that finding data can be a challenging and time-consuming experience, that luck plays a part, and that the search results may be highly sensitive to the formulation of the search queries. Ideally, data users should be provided with:

- A search experience that does not expect them to have prior knowledge of where the data may be found, or to know what specific keywords should be used to locate data of interest (in our example, a search for keyword “*handicap*” instead of “*disability*” would not have identified the variables on “*disability*”). This requires that semantic searchability be implemented in data catalogs. Few data catalogs have such feature; most are limited to lexical/full-text search.
- The possibility to start their search for data in generic search engines like Google or Bing. This requires that specialized data catalogs implement better search engine optimization (SEO) to increase their visibility and ranking (see section “Visibility of data” below).
- Inter-operable data catalogs, or “aggregators” of the content of data catalogs.
- Rich metadata to allow them to assess the fitness-for-purpose of any dataset.

To understand the value of structured metadata to address these issues, we need to take into consideration how search engines ingest, index, and exploit metadata. In brief, metadata need to be acquired, augmented, analyzed, transformed, and indexed before they can be made searchable.<sup>20</sup>

- **Acquisition:** Search engines like Google and Bing acquire metadata by crawling billions of web pages using web crawlers (or bots), with an objective to cover the entire web. Guidance is available to webmasters on how to optimize websites for visibility<sup>21</sup>. Search engines in specialized data catalogs have a much simpler task, as they only process content that catalog administrators and curators generate or control. The acquisition/extraction of metadata must preserve the structure of the metadata. This will be critical for optimizing the performance of the search tool and the ranking of query results.
- **Augmentation** or enrichment: the acquired metadata can be augmented or enriched in multiple ways, by extracting information from an external source or using machine learning algorithms.
- **Analysis and transformation:** The metadata will mostly consist of text. For the purpose of discoverability, some of the text has no value; words like “the”, “a”, “it”, referred to as *stop words*, will be excluded from the indexed metadata. The remaining words may be submitted to spell checkers and will be stemmed or lemmatized (stemming or lemmatization consist of converting words to their stem or root). Last, the transformed metadata will be tokenized, i.e., split into a list of terms (*tokens*). To enable semantic searchability, a numeric representation of the metadata can also be generated using a natural language processing embedding model.

---

<sup>20</sup> The process is described in detail in D. Turnbull and J. Berryman (2016)

<sup>21</sup> See for example Google’s Search Engine Optimization Starter Guide available at <https://developers.google.com/search/docs/beginner/seo-starter-guide>

- **Indexing:** The last phase of metadata processing is the indexing of the tokens. The index of a search engine is an inverted index, which will contain a list of all terms found in the metadata, with the following information (among other) attached to each term:
  - The document frequency, i.e., the number of metadata documents where the word is found (a metadata document is the metadata related to one dataset).
  - The identification of the metadata documents in which the term was found.
  - The term frequency in each metadata document.
  - The term positions in the metadata document, i.e., where the term is found in the document. This is important to identify collocations. When a user submits a query for “demographic transition” for example, documents where the two terms are found next to each other (i.e., as a “phrase”) will be considered more relevant than documents where both terms appear but in different parts of the document.

This process is embedded and automated in data cataloguing applications. But a good understanding of its mechanisms helps understand how and why rich and structured metadata matters.

Once the metadata have been acquired, transformed, and indexed, they can be used via user interfaces (UI). A data catalog UI will typically include a search box and facets (filters). The search engine underlying the search box can be simple (out-of-the-box full text search, looking for exact matches of keywords), or advanced (with semantic search capability and optimized ranking of query results). Rich and structured metadata, combined with advanced search optimization tools and machine learning solutions, allow catalog administrators to tune the search engine, and implement advanced solutions that improve data discoverability.

#### Interoperability of applications

Data catalogs that adopt common metadata standards and schemas can share information through automated harvesting and synchronization procedures. This allows them to increase their visibility, and to publish their metadata in hubs/aggregators. Interoperability between data catalogs is improved when common controlled vocabularies are used. Recommendations and guidelines for improved inter-operability of data catalogs are provided by the Open Archives Initiative (OAI).

The adoption of metadata standards by software developers also contributes to the easy transfer of metadata across applications. For example, Survey Solutions by the World Bank and CsPro by the US Census Bureau offer options to export metadata compliant with the DDI Codebook standard, which can then be edited or published in DDI-compliant metadata editors or catalogs.

#### Visibility of data

Many data users will start their search for data not in specialized data catalogs, but in Google or other commercial search engine. Google dominate the search engine use with an estimated 87 percent market share of desktop searches as of September 2021.<sup>22</sup> User behavior data (2020) also showed that “only 9% of Google searchers make it to the bottom of the first page of the search results”, and that “only 0.44% of searchers go to the second page of Google’s search results”.<sup>23</sup> Data users may thus not find—and therefore not use—data resources that are ranked low in Google results. The example below (Figure 10) shows the results of a search for “*GDP of India*

<sup>22</sup> Source: <https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/>

<sup>23</sup> Source: <https://www.smartinsights.com/search-engine-marketing/search-engine-statistics/>

2020” on Google. The website of the Ministry of Statistics of India (MOSPI), which is the official source of that information, only appears in page 7 of the results (in 63rd position).

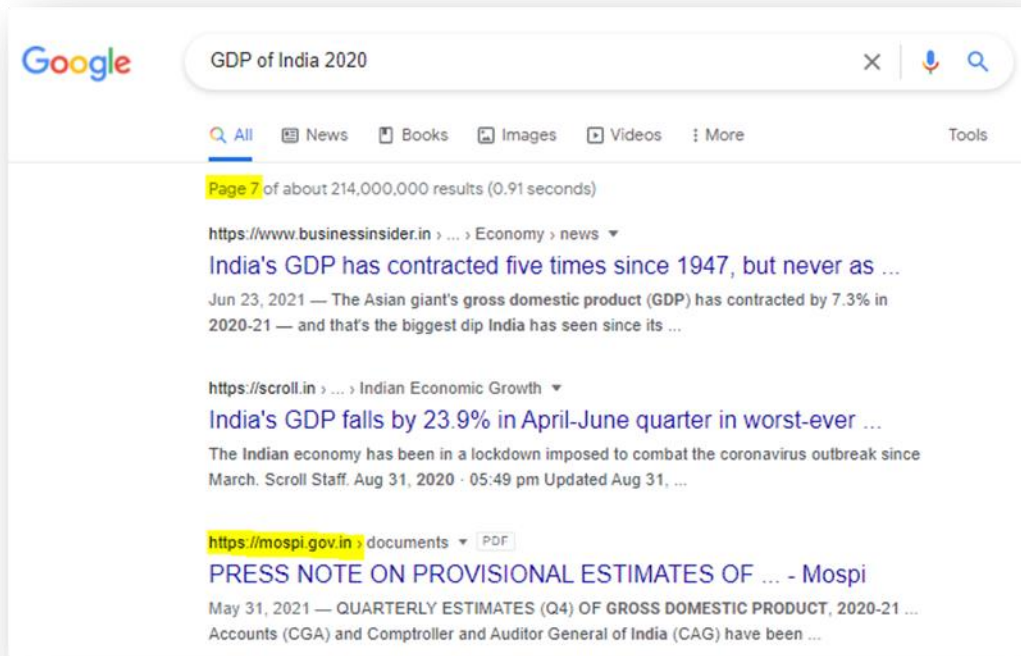


Figure 10 - Searching "GDP of India 2020" on Google (queried on 25 March 2022)

Administrators of data catalogs should pursue a double objective to (i) optimize the discovery and accessibility of metadata in their own platform, and (ii) optimize the accessibility of their metadata by web crawlers and other metadata harvesters to maximize visibility. This second objective involves active search engine optimization (SEO). SEO is “*the process of improving the quality and quantity of website traffic to a website or a web page from search engines. SEO targets unpaid traffic (known as “natural” or “organic” results) rather than direct traffic or paid traffic. (...) As an Internet marketing strategy, SEO considers how search engines work, the computer-programmed algorithms that dictate search engine behavior, what people search for, the actual search terms or keywords typed into search engines, and which search engines are preferred by their targeted audience. SEO is performed because a website will receive more visitors from a search engine when websites rank higher on the search engine results page.*”<sup>24</sup> “*Because search engines crawl the web pages that are generated from databases (rather than crawling the databases themselves), your carefully applied metadata inside the database will not even be seen by search engines unless you write scripts to display the metadata tags and their values in HTML meta tags. It is crucial to understand that any metadata offered to search engines must be recognizable as part of a schema and must be machine-readable, which is to say that the search engine must be able to parse the metadata accurately. (For example, if you enter a bibliographic citation into a single metadata field, the search engine probably won’t know how to distinguish the article title from the journal title, or the volume from the issue number. In order for the search engine to read those citations effectively each part of the citation must have its own*

<sup>24</sup> Source: [https://en.wikipedia.org/wiki/Search\\_engine\\_optimization](https://en.wikipedia.org/wiki/Search_engine_optimization)



*field. (...) Making sure metadata is machine-readable requires patterns and consistency, which will also prepare it for transformation to other schema. (...)*<sup>25</sup>

Guidelines for implementing SEO are provided by Google Search, Google Dataset Search, and other lead search engines. An important element is the provision of structured metadata that can be exploited directly by the crawlers and indexers of search engines. This is the purpose of a set of schemas known as schema.org.<sup>26</sup> In 2011 Google, Microsoft, Yandex, and Yahoo! created a common set of schemas for structured data markup on web pages with the aim of helping search engines to better understand websites. An alternative to schema.org is the DCAT (Data Catalog Vocabulary) metadata schema recommended by the W3C, also recognized by Google. “*DCAT is a vocabulary for publishing data catalogs on the Web, which was originally developed in the context of government data catalogs such as data.gov and data.gov.uk (...)*”<sup>27</sup> Mapping (selected) elements from structured metadata to the schema.org and/or DECAT standard is a critical element of SEO that will contribute significantly to the visibility of on-line data and metadata. This process does not have to be manual; it can be automated in data cataloguing applications that make use of metadata standards like the DDI.

#### Question banks and harmonization of data collection

The adoption of a structured metadata standard facilitates the development and maintenance of *question banks*, which in turn facilitates the process of harmonizing data collection instruments. For example, the documentation and cataloguing of survey microdata using the DDI Codebook standard (which includes variable-level metadata elements including variable and value labels, literal questions, interviewers’ instructions, skip instructions, universe, definitions, and more) allows users to compare variables across sources in a convenient manner. The user can be a researcher who needs to assess the comparability of data, or a data producer who needs to identify where concepts, methods and classifications may need to be harmonized or standardized. The example below (Figure 11) shows how specific variables can be located, then compared across datasets using the NADA cataloguing application developed for the International Household Survey Network. A search for “drinking water” in datasets from Bangladesh will return (in this catalog) a list of variables from 6 datasets. These variables can be compared (Figure 12). Comparisons will be more useful when detailed metadata is available (including literal questions, value labels, universe, interviewer’s instructions, and more – all of which will be displayed).

---

<sup>25</sup> From “Metadata, Schema.org, and Getting Your Digital Collection Noticed”, a blog post by Patrick Hogan available at <https://www.ala.org/tools/article/ala-techsource/metadata-schemaorg-and-getting-your-digital-collection-noticed-3>

<sup>26</sup> See <https://schema.org/>

<sup>27</sup> Source: <https://www.w3.org/TR/vocab-dcat-2/>

drinking water x Search

Years ▼ Showing 1-6 of 6 Study view Variable view Relevance ▼

Data Classifications ▼

Countries ▲  
 1 selected x Clear  
 Filter...  
 Argentina (1007)  
 Armenia (102)  
 Aruba (7)  
 Australia (57)  
 Austria (59)  
 Azerbaijan (52)  
 Bahamas (18)  
 Bahrain (28)  
 Bangladesh (124)  
 Barbados (19)  
 Belarus (45)

License ▼

**WASH KAP Survey Rohingya Cox's Bazar, 2018**  
 Bangladesh, 2018  
 UNHCR  
 ID: BGD\_2018\_WASH-CB\_v01\_M Last modified: Oct 14, 2021  
 Data available from external repository  
 Keyword(s) found in 50 variable(s) out of 380

Compare	Name	Label
<input type="checkbox"/>	s2c1	May I have a small sample of drinking water?
<input checked="" type="checkbox"/>	s2b1a	What is the principal source of drinking water for members of your household?
<input type="checkbox"/>	s2b9b	How do you clean your drinking water containers?
<input type="checkbox"/>	s2c6_2	How do you know water is safe for drinking? Taste
<input type="checkbox"/>	s2c6_3	How do you know water is safe for drinking? Smell
<input type="checkbox"/>	s2c6_4	How do you know water is safe for drinking? Color

Compare variables 1 variables selected from 1 studies

**Multiple Indicator Cluster Survey 2012-2013**  
 Bangladesh, 2012-2013  
 United Nations Children's Fund, Bangladesh Bureau of Statistics  
 ID: BGD\_2012\_MICS\_v01\_M Last modified: Mar 29, 2019 Views: 13392 Citations: 15  
 Data available from external repository  
 Keyword(s) found in 59 variable(s) out of 632

Figure 11 - Searching for "drinking water" in a NADA catalog and selecting variables for comparison



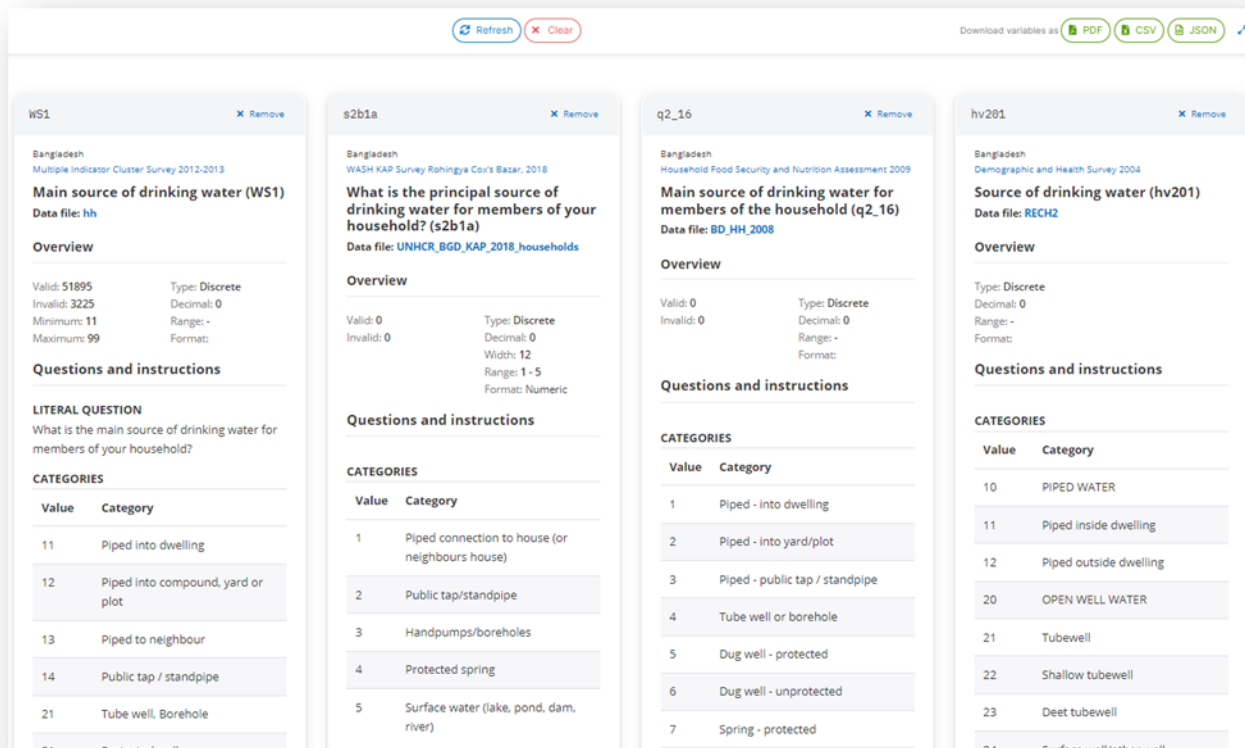


Figure 12 - Variable comparison in the NADA catalog (showing all variable-level metadata for the selected variables)

## Recommended and suggested standards and schemas

### Metadata standards and schemas by data type

To maximize the quality of the metadata, a specific metadata standard should be used for each relevant type of data. The following standards and schemas form the core set of recommended and suggested standards. For the documentation of survey microdata, the DDI Codebook (most recent version available) is recommended. Other standards and schemas listed in the table below are suggested. Detailed information on the schemas, and a justification for their selection, is available in the “*Metadata Standards and Schemas for Improved Data Discoverability and Usability*” document.<sup>28</sup>

Data type	Standard or schema
Microdata	Data Documentation Initiative 2.5 (DDI Codebook) – Version 2.6 forthcoming.  DDI LifeCycle may be considered by agencies with a high level of expertise in metadata management and capacity to build and maintain advanced applications.

<sup>28</sup> Olivier Dupriez and Mehmood Asghar, 2022; *draft* (as of October 2022) available at <https://mah0001.github.io/schema-guide/>

Time series, Indicators	<p>A schema was developed by the World Bank Data Group to document time series and the databases they belong to. The schema was developed by compiling metadata elements found in the World Bank World Development Indicators (WDI), in the metadata files published by the UN SDG Global Database, and in other indicators databases like ILOStat, FAOStat, the Demographic and Health Survey (DHS) StatCompiler, the OECD Metastore, and others.</p> <p>Note: The SDMX standard provides a data exchange format which may be considered in addition to a metadata standard.</p>
Statistical tables	<p>A schema was developed by reviewing a diverse set of statistical tables (from statistical yearbooks, census publications, and others) and deriving a set of metadata elements that accommodates the structure and components of these tables.</p>
Geographic datasets and services	<p>ISO <a href="#">19110/19115/19119</a> and their XML representation <a href="#">ISO19139</a>. For advanced users and organizations that specialize in geographic data dissemination, the SpatioTemporal Asset Catalogs (STAC) specification should be considered. STAC is “a common language to describe geospatial information, so it can more easily be worked with, indexed, and discovered.” (<a href="https://stacspec.org/en">https://stacspec.org/en</a>)</p>
Documents	<p><a href="#">Dublin Core</a> Metadata Initiative (DCMI) with some elements from the <a href="#">MARC21</a> (format for bibliographic data by the US Library of Congress), and bibliographic elements from BibTex.</p>
Photos / Images	<p>International Press Telecommunications Council (<a href="#">IPTC</a>) or Dublin Core (augmented with elements from ImageObject from <a href="#">schema.org</a>)</p>
Audio files	<p>Dublin Core augmented with elements from AudioObject from <a href="#">schema.org</a></p>
Videos	<p>Dublin Core augmented with elements from VideoObject from <a href="#">schema.org</a></p>
Programs and scripts	<p>A schema was developed to document research and analytics projects and the related scripts.</p>

For search engine optimization (SEO) and on-line visibility purpose, all standards and schemas listed in the table (or at least a subset of the elements of each schema) should be mapped to the DCAT metadata schema and/or to the dataset schema proposed by [schema.org](#). This mapping can be automatically implemented in cataloguing applications (see section “Visibility of data” above). DECAT and [schema.org](#) do not provide the necessary “specialized” solution to document datasets, but they serve a critical purpose by making data and metadata more visible and discoverable on-line.

#### **A note on the SDMX standard**

The [Statistical Data and Metadata eXchange](#) (SDMX) standard, sponsored by a group of international organizations and published as an ISO standard (ISO 17369), is not listed here as a metadata standard. SDMX is primarily intended to support the automation of machine-to-machine data and metadata exchange, not to support the production of comprehensive metadata based on a standard structure of elements. The standards we propose in the list above are all intended to be used to document data independently of the mode of dissemination of the data. But data shared

using SDMX will always be provided with machine-readable format. There is thus a close relationship to be established between the standards and schemas proposed in this note and SDMX, which can complement each other. Although SDMX provides much flexibility on what metadata elements should be attached to a dataset, it recognizes the value of standardized metadata for usability and comparability.<sup>29</sup> Metadata in SDMX are reported according to Metadata Structure Definitions (MSDs). MSDs can make use of metadata elements available in the proposed standards and schemas, and/or map elements used in MSDs to these standards and schemas, to foster inter-operability.

## The DDI Codebook standard

### Purpose and versions

To document microdata, the Data Documentation Initiative (DDI) Alliance has developed the DDI metadata standard. *“The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences. DDI is a free standard that can document and manage different stages in the research data lifecycle, such as conceptualization, collection, processing, distribution, discovery, and archiving. Documenting data with DDI facilitates understanding, interpretation, and use – by people, software systems, and computer network.”*<sup>30</sup>

The DDI standard comes in two versions: **DDI Codebook** and **DDI Lifecycle**.

- **DDI-Codebook** is a light-weight version of the standard. Its elements include descriptive content for variables, files, source material, and study level information. The standard is designed to support the discovery, the preservation, and the informed use of data.
- **DDI Lifecycle** is designed to document and manage data across the entire life cycle, from conceptualization to data publication, analysis and beyond. It encompasses all the DDI-Codebook specification and extends it.

The DDI Codebook is free software published under the terms of the GNU General Public License as published by the Free Software Foundation (version 3 or any later version). The DDI standard is published as an XML standard (but can be “translated” into a JSON schema). Version 2.5 of the DDI Codebook was published in January 2012. Version 2.6 is forthcoming and will be backward compatible with earlier versions.

### Scope and structure

The DDI Codebook contains many metadata elements grouped into five main sections:

- **Document description** (doc\_desc): “document” refers to the XML metadata; this section contains the elements used to describe the metadata, not the data. It is used mostly for catalog administration purposes.
- **Study description** (study\_desc): this section contains all elements needed to describe the survey: title, producer, sampling, methodology, objectives, process, and much more.
- **Data files** (data\_files): contains the elements used to describe each data file.
- **Variables** (variables): contains the elements used to describe each variable (name, type, variable and value labels, universe, type, literal question, interviewer’s instructions, definitions, derivation and imputation, summary statistics, skip instructions, and more).

---

<sup>29</sup> See a presentation by Elena De Jesús, United Nations Statistics Division, [https://open-sdg.org/assets/documents/webinar\\_17-June-2020/Open-SDG-webinar-UN-metadata-template-SDMX-MSD-slides.pdf](https://open-sdg.org/assets/documents/webinar_17-June-2020/Open-SDG-webinar-UN-metadata-template-SDMX-MSD-slides.pdf)

<sup>30</sup> Source: <https://ddialliance.org/>, accessed on 7 June 2021

- **Variable groups** (`variable_groups`): a section used (optionally) to organize variables into groups (thematic or others) other than the data files they belong to.

A detailed description of the standard is available from the DDI Alliance website<sup>31</sup> (the official source, describing the XML specification of the standard) and from the IHSN<sup>32</sup> (describing the JSON interpretation of it). The screenshots below (Figure 13 and Figure 14) provide an overview of the content of the *study description* and of the *variable* sections of the standard.

```

+ "doc_desc": { ... },
- "study_desc": {
  + "title_statement": { ... },
  + "authoring_entity": [ ... ],
  + "oth_id": [ ... ],
  + "production_statement": { ... },
  + "distribution_statement": { ... },
  + "series_statement": { ... },
  + "version_statement": { ... },
  "bib_citation": "string",
  "bib_citation_format": "string",
  + "holdings": [ ... ],
  "study_notes": "string",
  + "study_authorization": { ... },
  + "study_info": { ... },
  + "study_development": { ... },
  + "method": { ... },
  + "data_access": { ... }
},
+ "data_files": [ ... ],
+ "variables": [ ... ],
+ "variable_groups": [ ... ],

```

Figure 13 - Overview of the structure of the DDI Codebook metadata standard (JSON version)

Each element in the DDI documentation is described with a label, description, and more. For example, in the variables section, the element `labl` would contain the variable label, `var_qstn_qstnlit` would contain the full formulation of the question as printed in the questionnaire or asked by the interviewer, `var_qstn_ivuinstr` contains the text of the interviewer's instructions, `var_catgry` contains elements used to document the codes and value labels for categorical variables, etc.

<sup>31</sup> See <https://ddialliance.org/Specification/DDI-Codebook/2.5/>

<sup>32</sup> See <https://ihsn.github.io/nada-api-redoc/catalog-admin/#tag/Survey>

```

+ "study_desc": { ... },
+ "data_files": [ ... ],
- "variables": [
  - {
    "file_id": "string",
    "vid": "string",
    "name": "string",
    "labl": "string",
    "var_intrvl": "discrete",
    "var_dcml": "string",
    "var_wgt": 0,
    "loc_start_pos": 0,
    "loc_end_pos": 0,
    "loc_width": 0,
    "loc_rec_seg_no": 0,
    "var_imputation": "string",
    "var_derivation": "string",
    "var_security": "string",
    "var_respunit": "string",
    "var_qstn_preqtxt": "string",
    "var_qstn_qstnlit": "string",
    "var_qstn_postqtxt": "string",
    "var_forward": "string",
    "var_backward": "string",
    "var_qstn_ivulnstr": "string",
    "var_universe": "string",
    + "var_sumstat": [ ... ],
    "var_txt": "string",
    + "var_catgry": [ ... ],
    + "var_std_catgry": { ... },
    "var_codinstr": "string",
    + "var_concept": [ ... ],
    + "var_format": { ... },
    "var_notes": "string"
  }
],
+ "variable_groups": [ ... ],

```

Figure 14 - Content of the "variable description" section of the DDI Codebook standard (JSON version)

## Development and maintenance of the DDI Codebook metadata standard

The content of this section was extracted from the DDI website.<sup>33</sup>

*“Established in 2003, the Data Documentation Initiative Alliance (DDI Alliance) is an international collaboration dedicated to establishing metadata standards and semantic products for describing social science data, data covering human activity, and other data based on observational methods.*

*The DDI Alliance shares a commitment to meet worldwide requirements for publicly available standards and semantic products supporting the documentation and integration of social science data and other data for understanding the human condition. The Alliance’s purposes are to:*

- *Oversee the continued development of DDI standards and semantic products, including revisions, corrections, and new releases.*
- *Promote the adoption of DDI metadata standards and semantic products by stakeholders such as data producers, data distributors, data libraries, data archives, data users, researchers, and software developers and vendors.*
- *Support the development of training programs that encourage the use of these standards for all skill levels of potential adopters.*
- *Insofar as possible, ensure compatibility of DDI standards with emerging metadata standards in other fields.*
- *Balance the interests of a diverse community of stakeholders through a process that is open, transparent, consensus-driven, and open to recourse.*

*The membership of the Alliance consists of all stakeholder organizations in good standing that assume responsibility for the development and stewardship of DDI metadata standards and semantic products. Members remain in good standing through payment of annual member dues, provision of in-kind contributions, and adoption of DDI standards and products as appropriate. Membership is open to organizations from around the globe regardless of discipline or sector. The Alliance is a contractual, unincorporated collaboration of member institutions governed by its Bylaws. An Executive Director and Executive Board manage the operations of the Alliance, while its scientific and technical work is conducted under the guidance of the Scientific Board. The Alliance maintains a small Secretariat at the Host Institution to administer its day-to-day operations. The Alliance is financially self-supporting through Membership dues; license fees; workshop, symposia, and publication fees; and through external research or training grants and contracts with the Host or a Member Institution.”*

## DDI Codebook vs DDI Lifecycle

The development and maintenance of the DDI metadata standard is led by academic entities. Although the official statistics community has not been directly involved in this work, the DDI Alliance has been supportive of its needs, showing a concern to ensure relevance and applicability of the standard in different environments, including settings with constrained technical and financial capacity. Advanced users are encouraged to explore the features of DDI Lifecycle (DDI 3) and of the forthcoming DDI 4. But the version we recommend for adoption by the global community of official data producers is the DDI Codebook. It provides a solution that will allow rapid and significant progress in data archiving and cataloguing, transparency, usability, and discoverability. Tools (open-source applications and freeware), guidelines, and training materials on the use of DDI Codebook are available and are being further developed. The set of tools available for DDI Lifecycle is more limited. The DDI Lifecycle is significantly more

---

<sup>33</sup> Source: <https://ddialliance.org/about-the-alliance> ; extracted on 24 March 2022.

demanding in terms of infrastructure and expertise and serves purposes that resource-constrained organizations may not see as priorities.

The DDI Alliance justified as follows their decision to maintain two branches (Codebook and Lifecycle) of the standard:

*“DDI has shifted its view over time in terms of what the DDI specification encompasses. Originally, we had DDI in various versions. With version 3.0 a major change was introduced in order to support changes in technology, implementation requirements, and to address the coverage areas not addressed by versions 1 through 2. There was a general assumption that DDI users would switch to the newest version in the same way that software users shift to new versions. However, DDI was and is used for archival purposes meaning that large amounts of metadata were already captured in earlier versions of DDI and were supported by existing software. In addition, the new version of DDI had higher infrastructure requirements and many of the new features were not required by several of the current user groups. Instead, version 3.0 brought in a new range of DDI users whose needs were not met by the earlier versions. Therefore, earlier versions of DDI were used not only by those who had not switched to the newer version, but by a large group of new users supported by software from the World Bank. The requirements of this user group were met by this earlier version and the costs of using the newer version were too high in terms of infrastructure.*

*DDI determined that both versions of the standard would be maintained and named them DDI-Codebook (versions 1.0-2.x) and DDI-Lifecycle (version 3.x). Now DDI has determined that rather than treat these as continuations of the same specification with a single versioning stream, to treat them as two separate products with separate versioning streams. This approach also recognizes the DDI products that have been developed in the periphery of Codebook and Lifecycle. DDI now recognizes a suite of published products (DDI-Codebook, DDI-Lifecycle, Controlled Vocabularies, and XKOS) and products that are under development (DISCO, DDI-Cross-Domain Integration, and SDTL).”<sup>34</sup>*

### Users’ community

The DDI metadata standard is used by many academic data centers including the Inter-University Consortium for Political and Social Research (ICPSR, an international consortium of more than 750 academic institutions and research organizations)<sup>35</sup>. It is also used by international organizations including the Food and Agriculture Organization (FAO), International Labour Organization (ILO), Pacific Community (SPC), United Nations High Commission for Refugees (UNHCR), UNICEF, World Health Organization (WHO), and the World Bank, and by national statistical agencies.

## **Metadata formats and tools**

---

### **The JSON and XML formats**

Metadata standards and schemas consist of structured lists of metadata elements (or “fields”). Schemas must be sufficiently intuitive and human-readable to allow data curators to generate and organize their metadata in compliance with the schema. They are also designed to be machine-readable and available in non-proprietary formats, to be exploited by software and database applications. JSON (JavaScript Object Notation) and XML (eXtended Markup Language) are the most suitable formats for these purposes. Both are plain text files (i.e. not allowing complex

---

<sup>34</sup> Source: <https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/929792030/DDI+Codebook+Development+Work>

<sup>35</sup> See <https://www.icpsr.umich.edu/web/pages/>

formulas or text formatting). Although this can be seen as a limitation, it is a guarantee of durability and portability. Metadata stored in XML and JSON can be converted into html pages, PDF files, and other user-friendly outputs. The example below shows how information on the authoring entity (primary investigator), data collection dates, and country, are stored in a DDI-compliant format, respectively in XML and JSON format.

#### ***In XML format***

```
<studyDscr>
  <citation>
    <rspStmt>
      <AuthEnty name="National Statistics Office"
        affiliation="Ministry of Planning"></AuthEnty>
    </rspStmt>
  </citation>
  <studyInfo>
    <sumDscr>
      <timePrd date="2021-01-15" event="start"></timePrd>
      <timePrd date="2021-03-30" event="end"></timePrd>
      <nation abbr="POP">Popstan</nation>
    </sumDscr>
  </studyInfo>
</studyDscr>
```

#### ***In JSON format***

```
{
  "study_desc": {
    "authoring_entity": [
      { "name": "National Statistics Office",
        "affiliation": "Ministry of planning" }],
    "study_info": {
      "coll_dates": [{"start": "2021-01-15", "end": "2021-03-30"}],
      "nation": [{"name": "Popstan", "abbreviation": "POP"}]
    }
  }
}
```

Both JSON and XML are (somewhat) human and machine-readable text files, hierarchical (they can contain values within values), which can be parsed and used by programming languages like R or Python. XML files must be parsed with an XML parser, while JSON files can be parsed by standard JavaScript functions. JSON files are easier to generate and parse than XML, and easier to read by humans.

### **Tools for the production of standard-compliant metadata**

Metadata standards will only be broadly adopted if their implementation does not represent a heavy financial or technical burden. Metadata editors can make the production of standard-compliant metadata simple and effective, and open-source cataloguing application facilitate their publishing and cataloguing. Some tools already exist. A common effort by the statistics community to further develop them would be very beneficial.

#### Multi-standard metadata editor

Data archives have for many years relied on a freeware application developed by the Norwegian Research Data Center, the *Nesstar Publisher*<sup>36</sup>, to document their microdata. This application was

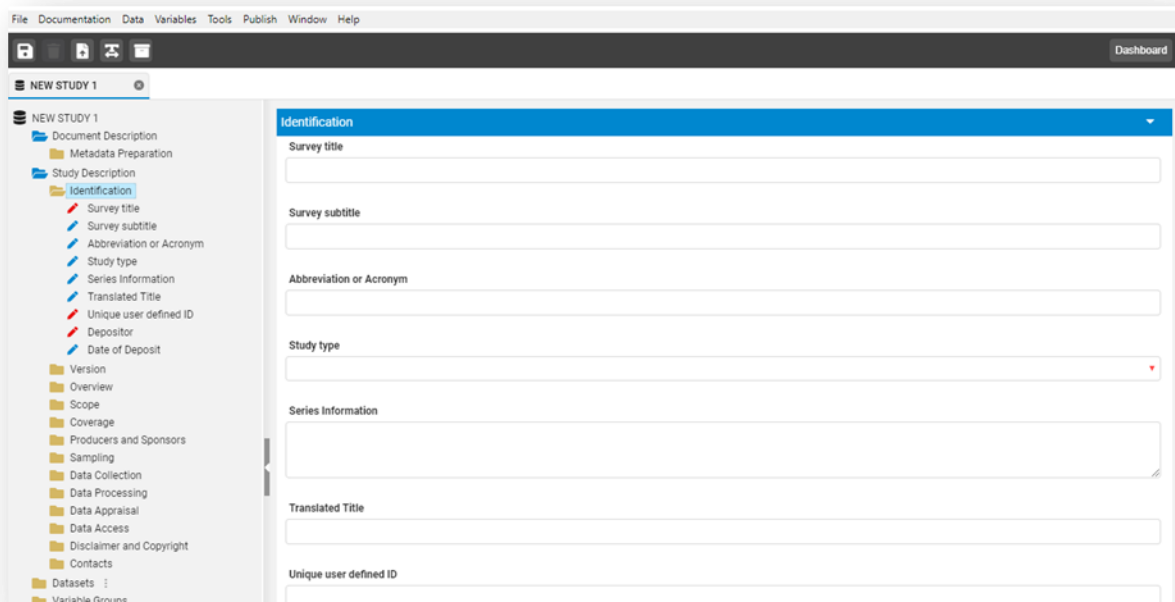
---

<sup>36</sup> See <http://www.nesstar.com/software/publisher.html>



built on a now-obsolete technology and around an outdated version of the DDI Codebook. It is not maintained anymore. The World Bank Data Group has initiated the development of a multi-standard Metadata Editor, expected to be published as open-source software. The screenshots below provide an overview of how a specialized Metadata Editor is used to generate DDI-compliant metadata. The steps will typically be:

1. Selection of a pre-designed DDI template (a subset of the DDI elements considered relevant by an organization). This will automatically generate metadata entry forms (Figure 15).
2. Import the data files, e.g., from Stata, SPSS, or another file format. This will automatically generate a list of files and variables, import the metadata stored in the original files (variable names, variable and value labels), and generate summary statistics for each variable (Figure 16).
3. Enter additional available information in the appropriate fields, at the study, file, and variable levels (Figure 17).
4. Save the metadata and export it as a DDI-compliant text file (XML, JSON), ready to be used as input to data cataloguing systems.



The screenshot displays a web-based application interface for creating metadata. The top navigation bar includes 'File', 'Documentation', 'Data', 'Variables', 'Tools', 'Publish', 'Window', and 'Help'. Below this is a 'Dashboard' button. The main content area is titled 'NEW STUDY 1' and features a left-hand sidebar with a tree view of metadata categories. The 'Identification' category is selected and expanded, showing sub-items: Survey title, Survey subtitle, Abbreviation or Acronym, Study type, Series Information, Translated Title, Unique user defined ID, Depositor, and Date of Deposit. The main panel on the right is titled 'Identification' and contains several input fields: 'Survey title' (text input), 'Survey subtitle' (text input), 'Abbreviation or Acronym' (text input), 'Study type' (dropdown menu), 'Series Information' (text area), 'Translated Title' (text input), and 'Unique user defined ID' (text input).

*Figure 15 – Metadata entry forms are generated based on a selected template*

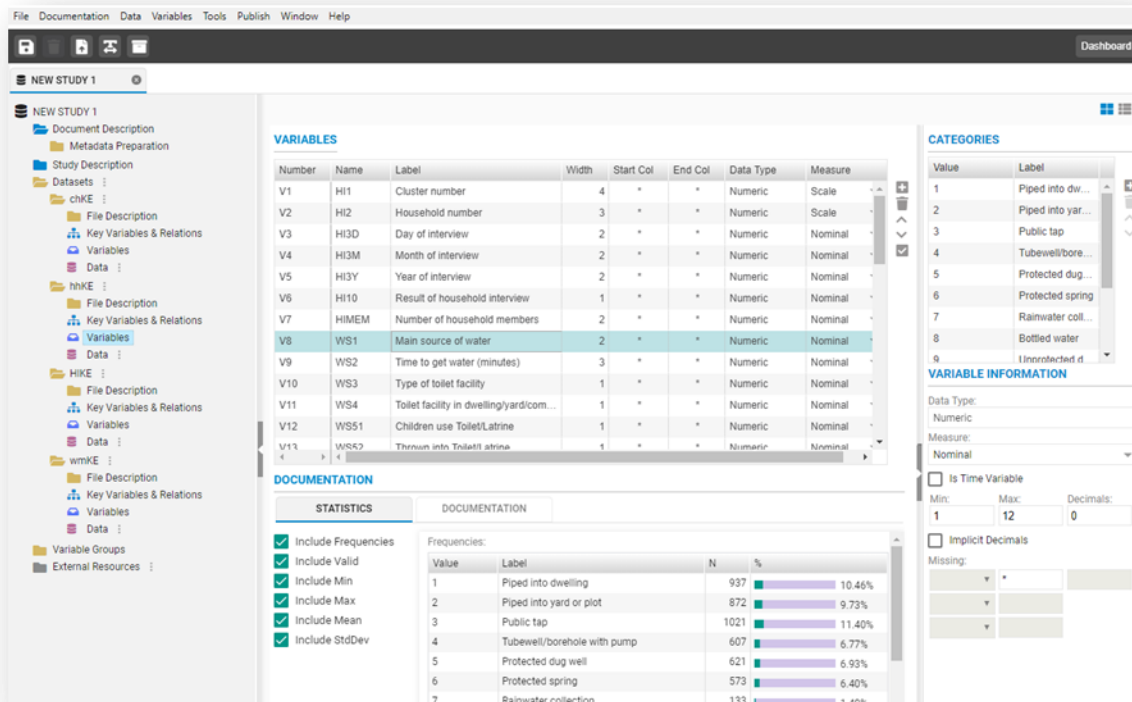


Figure 16 – After importing data files, variable-level information is automatically generated

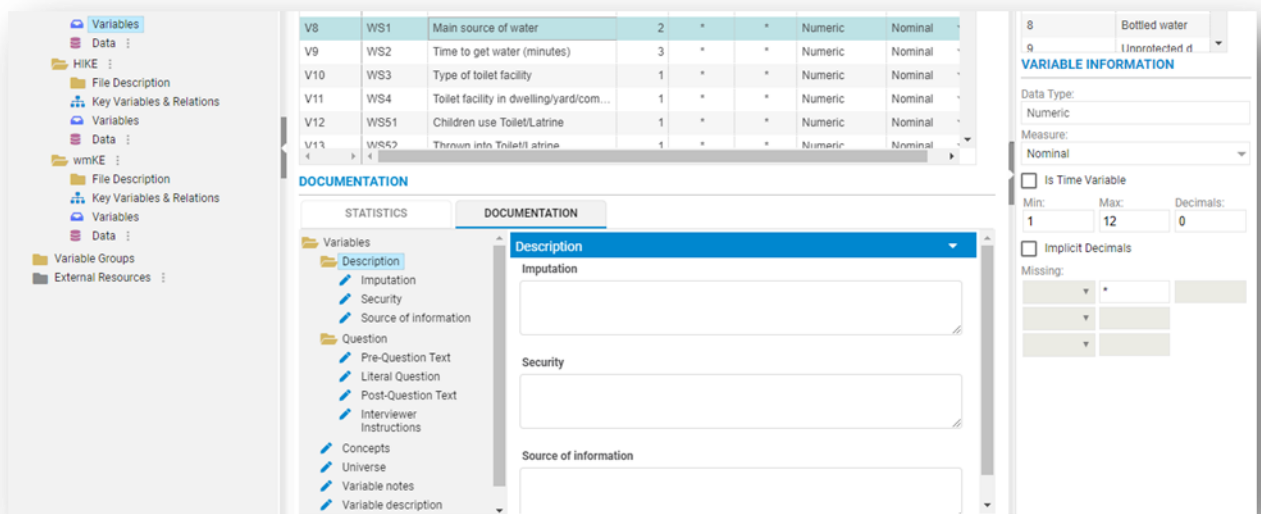


Figure 17 – Detailed metadata is added, such as literal questions or interviewer's instructions for each variable

## R packages, Python libraries, other utilities

Metadata files can also be generated using a programming language like R or Python. This option does not have the user-friendliness of a specialized metadata editor, but it provides opportunities to automate part of the process and to embed metadata augmentation in it. For simple datasets,

metadata entry forms could also be developed in applications like MS-Excel, then converted to XML or JSON files using VBA, R or Python scripts.

#### DDI Codebook in software applications

CsPro (data entry software developed by the US Bureau of Census) and Survey Solutions (CAPI application developed by the World Bank) include a feature to export DDI Codebook variable-level metadata. Once exported, these metadata that include the list of collected variables, formulation of the questions, variable categories, and interviewer's instructions can be imported in a DDI Metadata Editor to be complemented with study-level metadata. A tool that would map XLSForm files used by applications like Kobo Toolbox could also be developed, to automate the production of DDI-compliant metadata by users of data entry applications compatible with XLSForm.

#### **Tools for the exploitation of standard-compliant metadata**

DDI-compliant data cataloguing applications are the main tools used to exploit metadata. Data catalogs make the data and metadata visible, discoverable, and accessible. Most data cataloguing applications only provide full-text (keywords-based) searchability. The availability of standardized and structured metadata justifies a global effort to develop more advanced discoverability solutions, with semantic capability and able to operate as recommender systems. Such development has been initiated but should be accelerated and governed by a larger community of practice organized around common standards and open-source solutions. We summarize below some of the features of modern data cataloguing applications that the availability of standard-compliant, rich metadata can enable.

#### Features enabled by rich and structured metadata

##### *Filters / facets*

Data catalogs must provide filters (facets) to subset suitable datasets based on criteria applied to categorical metadata elements. The structure of the metadata combined with the use of controlled vocabularies provide the necessary flexibility to implement such facets.

##### *Advanced search*

A search engine can be lexical (“full-text”), i.e., based on a search for literal matches between terms entered in a query and the terms found in the indexed metadata, or semantic, i.e., seeking to find datasets whose metadata are semantically close to the semantic composition of the query. Ideally, a combination of these two options will be provided in by data catalogs. Implementing semantic searchability is complex as it relies on machine learning solutions.

When only a keyword-based search is implemented, efforts should be made to optimize the indexing using open-source tools like Solr or ElasticSearch. Out-of-the box solutions (like those provided by SQL databases) will rarely perform in a satisfactory manner. Structured metadata allows optimizing search engines (by “boosting” some metadata elements in the calculation of results relevancy scores to ensure that the most relevant results appear at the top of the list). It also allows advanced search features to be implemented (searching for keywords in specific metadata elements instead of the whole metadata).

##### *Variable-level search and comparison*

The integrated use of data from different sources requires a certain degree of comparability and consistency across datasets. Structured metadata make it possible to implement user-friendly tools

for variable comparison (see section “Question banks and harmonization of data collection” of this note).

### ***Data and metadata API***

A modern data catalog application must provide users with access to the data and metadata via an application programming interface (API). The structured metadata allows users to extract specific components of the available metadata. For example, a user may want to extract the identifier and the title of all microdata and geographic datasets conducted after year 2000 in a data catalog. This can be done easily using an API but would be tedious to do otherwise. Making data accessible via API, although not critical for microdata, allows users to acquire the datasets or subset of datasets in an automated and effective manner. This can also enable features internal to a data catalog, such as dynamic visualizations and data previews.

### ***Recommendations***

Not all users will search data catalogs knowing exactly what data they need. Some will explore more than search data catalogs. E-commerce platforms build recommender systems to recommend products to their customers (under headings like “You may also be interested in ...”, “Products related to this item”, or “Frequently bought together: ...”) Data catalogs will ideally provide a similar option, to bring relevant resources to the attention of their users. Machine learning tools (like topic models and word embedding models) make it possible to measure the topical or semantic closeness between catalog entries, which can be exploited to implement recommender systems. In the example below (Figure 18, extracted from the World Bank’s exploratory application [NLP4DEV](https://www.nlp4dev.org/)<sup>37</sup>), we show how a document (in this case a 7-page document on “Poverty and the environment/climate change”) can be submitted as a query. The application will process the document (submitting it to machine learning models via a public API) and return a list of related documents and data (Figure 19).

---

<sup>37</sup> See <https://www.nlp4dev.org/>

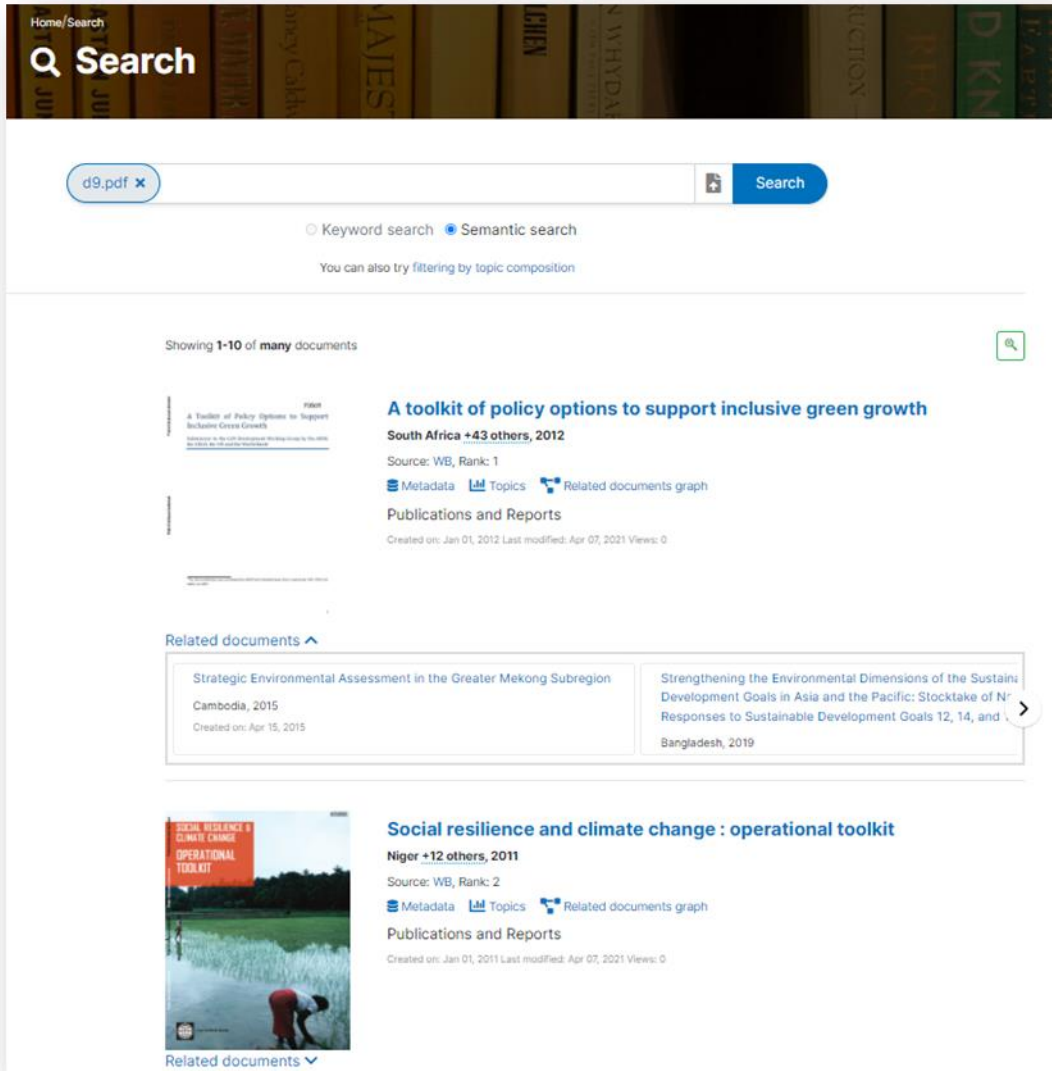
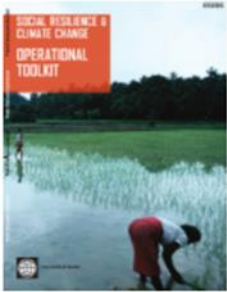


Figure 18 - Submitting a document as a query to a semantic-search tool

The accuracy of the recommendations depends on (i) the quality of the embedding machine learning model, and (ii) the quality (richness) of the metadata associated with each dataset. Further work on machine learning models and automated metadata enhancement is needed to improve the quality of recommender systems.



## Social resilience and climate change : operational toolkit

This note is written for World Bank task teams and explains how an understanding of the social dimensions of climate change can enhance the sustainability and quality of Bank supported operations while mitigating potential risks. The note reviews major challenges involved in addressing the social dimensions of climate change; outlines how social development approaches can help to solve these challenges; highlights the main social development analytical and operational tools in relation to the so...

[read more](#)

World, 2011

Category: Publications and Reports

Source: WB

Open in: [World Bank Documents and Reports](#)

Created on: Jan 01, 2011 Last modified: Apr 07, 2021 Views: 0

Metadata [📄](#) [JSON](#)

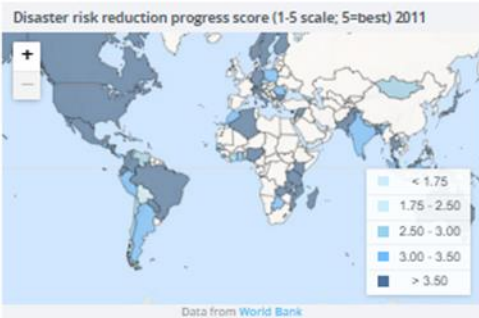
---

Metadata
View document
Related documents
Related data
Related documents graph

---

### Related World Development Indicators

Disaster risk reduction progress score (1-5 scale; 5=best)



Data from World Bank

Disaster risk reduction progress score is an average of self-assessment scores, ranging from 1 to 5, submitted by countries under Priority 1 of the Hyogo Framework National Progress Reports. The Hyogo Framework is a global blueprint for disaster risk reduction efforts that was adopted by 168 countries in 2005. Assessments of "Priority 1" include four indicators that reflect the degree to which countries have prioritized disaster risk reduction and the strengthening of relevant institutions.

[Link to data](#)

[Link to metadata](#)

### Related SDG Indicators

Number of local governments that adopt and implement local DRR strategies in line with national strategies (number)

Figure 19 - Results of the query - Semantically-close documents, indicators, and survey datasets (not shown)

### Efficiency and affordability of catalog maintenance

A data cataloguing application should be an efficient and effective tool for its administrator. Efficiency and effectiveness will be increased when:

- The application is available as an open-source application, build around open standards.
- The catalog is inter-operable with other catalogs, which requires consistency in the metadata structure (and is enhanced with the use of standard controlled vocabularies).

- The content of the catalog (both data and metadata) can be maintained and disseminated via APIs, which requires structured metadata.
- Search Engine Optimization (SEO) procedures are embedded in the application (which requires mapping of the metadata to the DCAT or schema.org standards, which itself requires structured metadata).
- The application is used and maintained by a community of users, to foster information exchange, sustainability, and partnerships.

## Capacity building and support

---

Many organizations around the world have adopted the DDI Codebook standard to document their microdata. Although expertise in programming languages and in the use of APIs opens multiple opportunities to exploit DDI-compliant metadata and to automate metadata management processes, the production, dissemination, and use of DDI-compliant metadata do not require such expertise. The standard is simple and intuitive, and the availability of applications like DDI metadata editors and cataloguing tools make it accessible to all organizations, including those who face financial and technical constraints. The maintenance of a set of free, open-source tools available in multiple languages, and the production of related guidelines and training materials, are a responsibility that the international statistical community has already largely taken. Experience has demonstrated that building capacity in the production and use of DDI-compliant metadata is not a major challenge. The rapid scaling-up of the adoption of the DDI Codebook by NSOs and other official data producers in resource-constrained environments is thus an achievable objective, which calls for a concerted effort by development partners involved in statistical capacity building. Training on data documentation should be complemented by training and support to the formulation of microdata dissemination policies and to data protection.

The broad adoption of the DDI Codebook in countries where statistical data production is supported by external funding will also be fostered by the inclusion of requirements related to metadata in the funding agreements and in the description of expected deliverables in consultant's contracts.

## A 10-tasks action plan

---

The following ten tasks are proposed to support the broad and rapid adoption of the DDI Codebook and possibly other metadata standards, and the improvement and harmonization of data documentation and dissemination practice.

No	Deliverable
1	Contribution to the development of the DDI metadata standard (by providing input to the DDI Alliance as and when relevant), to the review of other metadata schemas (for documents/time series, tables, and reproducible scripts), and to the development of controlled vocabularies. The output will include (i) updated version of the standards/schemas available on GitHub, and (ii) updated technical documentation of the schemas (including the <i>Guide on the Use of Metadata Standards and Schemas</i> ).
2	Finalization of an open-source, multi-standard <b>Metadata Editor</b> software application. This will be an improved version of the Metadata Editor developed by the World Bank (currently in beta version).

3	Production of tools and guidelines for <b>metadata augmentation</b> . This will include open-source scripts (R, Python), machine learning models accessible openly via API, and guidelines on the use of these solutions.
4	Development of an <b>assessment framework</b> to evaluate the readiness of micro-datasets for dissemination. This framework will be used by data producers and curators to conduct self-assessments of their microdata/metadata, or for external review/audit of microdata dissemination practice.
5	Tools and utilities to enable <b>search and recommender systems optimized for data discoverability</b> . This will include the training of machine learning models (embedding models and topic models) for semantic searchability (with API and technical documentation accessible openly), and recommendations for implementing advanced indexing and ranking solutions (Solr/ElasticSearch) in data catalogs. These solutions will be implemented in the <b>open-source NADA cataloguing application</b> (and available as open-source code for implementation in other cataloguing applications).
6	Development of an open-source solution for <b>publishing data and metadata via API</b> .
7	Advocacy and support for the adoption of a <b>DOI-based identification system</b> of micro-datasets by national and international statistical organizations. This will foster the inter-operability of data catalogs and facilitate the maintenance of citations catalog.
8	Production and dissemination of practical guidelines for <b>Search Engine Optimization</b> , specifically produced for improving the visibility/ranking of on-line data catalogs and statistics websites.
9	Maintenance of a central <b>Data Library</b> , as a hub (aggregator) of metadata (not data) from a network of contributing DDI-compliant microdata catalogs. The Data Library would coordinate the development, hosting, and maintenance of a <b>central metadata catalog/registry</b> of datasets. This Data Library could build on the IHSN catalog.
10	Provision of a <b>training program and technical support</b> to data disseminating organizations in low and lower-middle income countries.

These tasks would have to be implemented in partnership with multiple organizations including (but not limited to):

- National statistical agencies in low-, middle-, and high-income countries.
- International organization (mainly specialized agencies of the United Nations) that are part of the International Household Survey Network or of the Committee for the Coordination of Statistical Activities (CCSA), including FAO, ILO, OECD/PARIS21, UNHCR, UNICEF, UNSD, and WHO.
- Other regional and international organizations and multilateral development banks.
- The developers of metadata standards including the DDI Alliance.
- Academic data centers.
- Financial sponsors including foundations and bilateral donors.