Optimization of Accessibility and Quality of Metadata for Researchers: an Example of Collaboration Between INSEE and CASD

Halima Bakia[1], Ifaliana Rakotoarisoa[1], Thomas Dubois[2]

[1]CASD, Malakoff, FRANCE

[2]INSEE, Montrouge, FRANCE

Abstract

Accessibility and quality of metadata for researchers is fundamental. Metadata is anything that can give some useful information about data sets and help the users to better understand the data produced. It enables the researchers to identify relevant data for their project, before starting the procedures for accessing data, which can be quite long. The detailed documentation outlining the content of data serves as an initial response, acting as the first filter for source selection.

The Secure Data Access Center (CASD) is an organization allowing researchers and datascientists to work remotely and securely with confidential highly detailed microdata. It develops and establishes a secure interface between the research community and the NSS. 507 data sources are made available in a secure way, mainly produced by the French NSI (INSEE) and NSOs, but also from different public institutions. In this context, the exchange of high-quality metadata is crucial.

From 2018, documentation on data content has gradually been made available online by the CASD. Researchers, through satisfaction surveys, applaud this advancement. They are still requesting more documentation to be made available prior to apply for data access.

The process for exchanging metadata between INSEE and CASD is under review with a particular emphasis on the use of a metadata standard for structural metadata documentation (list of variables, codes and their meanings). INSEE and CASD recently carried out an experimentation to exchange files in the DDI standard (in this case DDI Lifecycle). The experimentation results show the expected gains:

- on workload, due to lower metadata entry burden
- on timeliness, the time taken to put documentation online will be reduced
- on reliability, as online documentation will correspond to the documentation delivered by the producer, thus avoiding re-entry errors.

The use of standard such as DDI demonstrates how it promotes metadata interoperability between the two organizations. To achieve the complete goal of interoperability, the experimentation also shows that the use of standard must be accompanied by best practices definition.

The proper use of data by researchers plays a crucial role in enhancing quality. When researchers have access to standardized and high-quality metadata, it not only facilitates their

own analytical work but also enables them to provide constructive feedback to data producers such as INSEE. This virtuous feedback loop creates an environment where data producers can adjust and enhance their processes in response to the specific needs of researchers.

Keywords: DDI Standard, Reusability, Reliability, Timeliness, Interoperability

1. Two Organizations: a Data Producer (INSEE) and a Data Disseminator (CASD)

1.1 The Importance of Documentation

The French National Institute of Statistics and Economic Studies (INSEE) collects, produces and disseminates information on the French economy and society, so that all interested parties (public administrations, companies, researchers, media, teachers, private individuals) can use it to carry out studies, make forecasts and take decisions. To make some confidential data and metadata securely available to researchers, INSEE relies on the French Secure Data Access Center (CASD), as it manages an equipment designed to allow accredited researchers and datascientists to work remotely and securely with confidential highly detailed microdata.

A presentation of the sources (a summary with the essential information, a list of variables and modalities) is available to the researchers on the CASD website. The main purpose is to help the researchers identify relevant data for their project, before starting the procedures and submitting the application, which could be quite long and complex CASD has chosen to document using the Data Documentation Initiative (DDI) standard using Colectica Designer©. Firstly, because it is a standard in which exchanges are possible and metadata are well described, enabling interoperability between systems and the reuse of metadata by people and software. Secondly, because INSEE, CASD's main data provider, is also working on documenting its data in this standard.

Both organizations INSEE and CASD use the same metadata standard for documenting structural metadata, they still exchange these metadata from non-standard files in different formats (PDF, CSV, ODS…). Based on this observation, INSEE and CASD have started experimenting a direct exchange of DDI files.

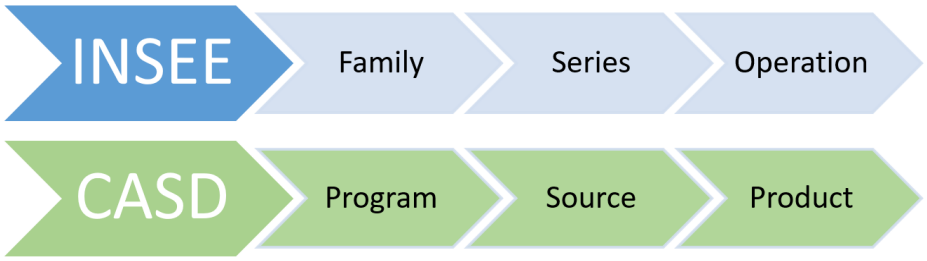
1.2 Works Between INSEE and CASD

INSEE and CASD engaged very early into a partnership on documentation, in particular with the agreement for the implementation of INSEE's Statistical Metadata Repository RMéS, which uses the DDI standard. This collaboration was carried out during ESSnet "Sharing common functionalities in ESS (European Statistical System)". It produced highly satisfactory

results, leading in particular to CASD information system being brought into line with the reference metadata disseminated by INSEE. In particular, the following results were achieved:

- Alignment of the data model with CASD data sources and INSEE metadata model with statistical operations. See the figure below for the mapping between INSEE model (Family, Series, Operation) and CASD model (Program, Source, Product)

Figure 1: Mapping Between INSEE and CASD Model



- Convergence towards a common list of series and families of statistical operations (programs and sources in CASD terminology)
- Development of a common list of themes used to categorize data sources
- Validation of a technical solution enabling CASD to directly query INSEE's metadata repository to obtain the necessary information

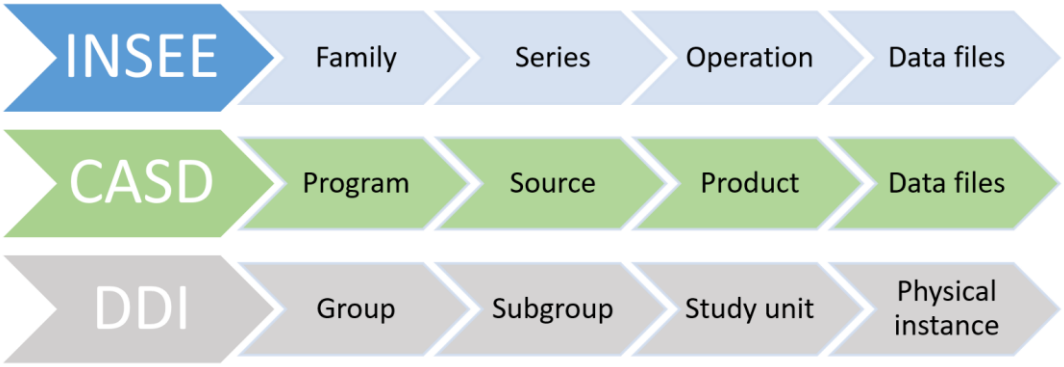
This first collaboration enabled us to measure the extent to which the CASD could benefit from the reuse of INSEE metadata, and motivated our ongoing work.

2. Metadata Exchange Methods

2.1 Experimentation: an Opportunity Offered by a Metadata Standard (DDI)

Matching the documentation's architecture of INSEE and CASD with DDI standard enables the exchange of XML files: see below INSEE model (Family, Series, Operation, Data files), CASD model (Program, Source, Product, Data files) and DDI standard (Group, Subgroup, Study unit, Physical instance).

Figure 2: Mapping Between INSEE, CASD and DDI Standard



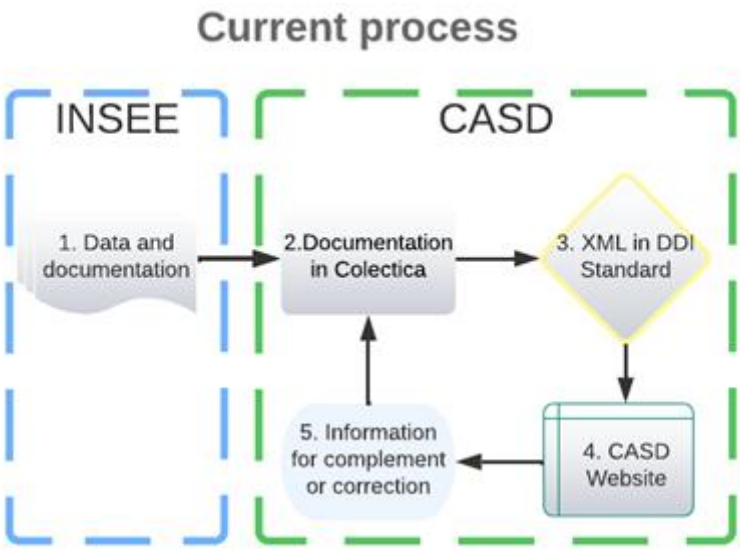
This experimentation focused on metadata at physical instance level.

2.2 Current Process for Disseminating Data and Metadata

The current workflow between INSEE and CASD is as following:

1. INSEE transfers data to CASD. CASD downloads the documentation at the same time, which is not standardized. Documentation is built upon several documents from INSEE's delivering application: file structure, codebook, methodological documentation, questionnaire, source code, and in different and various formats (PDF, CSV, ODS...).
2. All the work for transforming the documentation in DDI is done by CASD using the Colectica Designer software in a dedicated secure environment: data remain confidential and are used as input of the software.
3. DDI Metadata are extracted in XML format.
4. They are then uploaded from the secure environment in an internal application (a tool developed by CASD for supporting DDI standard) for display on the CASD website.
5. CASD asks INSEE if any information is missing. INSEE is also invited to check the results and, if needed, propose corrections. After the feedback CASD modifies the documentation if necessary.

Figure 3: Current Process

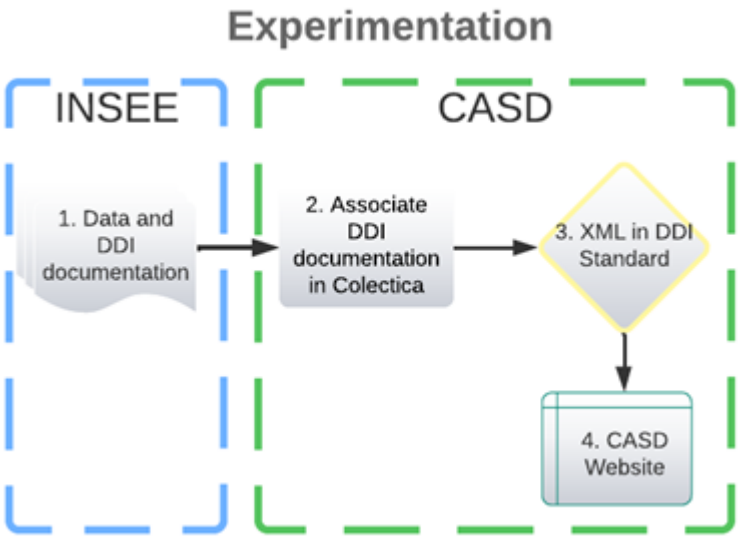


This process can be made more efficient, which is why INSEE and CASD have set up the following experimentation.

2.3 Experimented Workflow and Beyond

Experimentation focused on metadata at the lowest level, i.e. descriptions of data files, variables and associated representations (code list, numeric, date, text, etc.) according to this workflow.

Figure 4: Experimentation Process

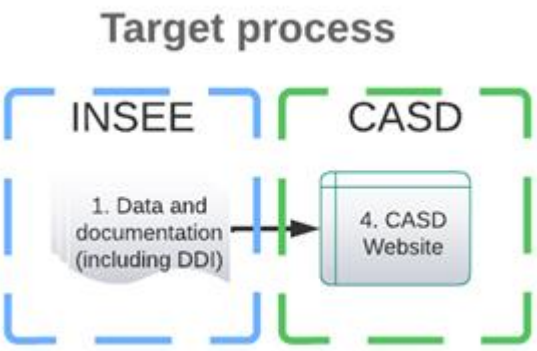


1. INSEE transfers data to CASD and CASD downloads the documentation including DDI/XML format. This new delivery in DDI/XML format represents an evolution of the current process.
2. Instead of documenting variables, this step now consists of simply attaching the physical instance (data file description) to the study unit.
3. DDI Metadata is extracted as XML (same step as above for the current process).
4. They are then uploaded in an internal application for display on the CASD website (same step as above for the current process).
5. This step is no longer necessary, as there is no need to reinterpret the delivered documentation.

Since displaying it on the CASD website requires documentation at study unit level, and the experimentation was carried out at physical instance level, it still requires manipulation via the Colectica Designer software. This one is much lighter because the heaviest tasks disappear: documenting variables (major part of step 2 in the current process) and checking metadata published or even propose corrections (step 5).

The purpose in mind is to be able to exchange DDI/XML metadata at a higher level (in particular at the level of the study unit), thus eliminating steps 2 and 3 (from figure 4), and moving on to the following very simple workflow (figure 5).

Figure 5: Target Process



3. Metadata Exchange Results

3.1 Checks Performed

In order to perform the experimentation, metadata already documented in DDI in both organisations has been chosen: in this case the data file produced by Insee "All employees databases - job position data 2019". The aim is to achieve the same result online if the DDI documentation is supplied by INSEE or if it is produced by CASD from non-standardized files as it's currently performed.

At different stages of the experimented workflow, checks have been done. Firstly, since we are using Colectica software, importing the DDI/XML file (including variable, variable labels, variable representations, and code lists) is well supported. If necessary, information has been harmonised: attribute the correct name and description of the data file.

Secondly, a new DDI/XML file is exported from Colectica for displaying the content on CASD website.

The result has been compared with the already published metadata for the same file. The content verification revealed a need for harmonization of practices: e.g. the link to the nomenclature online needs to be harmonised to make it clickable or the wording of a variable is different from that of the question.

3.2 Expected Gains

Implementing a process in which INSEE sends DDI metadata to CASD will enable gains:

- on workload, as a result the burden of metadata entry will be reduced. It will reduce internal resources dedicated to the management of the information
- on timeliness, the time taken to put documentation online will be reduced
- on reliability, as online documentation will correspond to the documentation delivered by the producer
- on relevance, improving the service quality in accordance with our public-interest mission to meet users' needs
- on accessibility and clarity of the information, the documentation for each year will be easily accessible by users in both websites, and available in the same form. This new process will facilitate harmonization between each year from the same source, and afterwards between each source from different producers
- on coherence, across datasets and within a dataset, the metadata are based on common concepts, definitions and nomenclatures (without reinterpretation of the meaning of variables)
- on interoperability, we share the best practices for the documentation, what was written or what refers to a link, how to write the label of the variable

4. What's next?

4.1 Benefits from Enrichment of INSEE Variable Description

To be able to exchange DDI variable descriptions according to the DDI standard will enable CASD to benefit from future enrichments. One idea is to track variable evolutions over time relying on the variable cascade model.

"*The variable cascade is the way the descriptions of variables is managed. The main purpose of the cascade is to increase the reuse of metadata"...*

"The cascade consists of four levels, each level corresponding to an ever-increasing descriptive detail. The levels in the cascade are:

- *Concept*
- *Conceptual variable*
- *Represented variable*
- *Instance variable*

The names of the levels indicate to the user what the main focus of the description is at each. The Concept and Conceptual Variable provide details about the concepts employed.

The Represented Variable and Instance Variable provide the details about the codes, characters, and numbers representing the concepts at the higher levels. "(DDI Alliance)

While a new instance variable is created for each dataset, the represented variable can be reused. The logical level offered by the represented variable helps to take better account for evolutions over time (by adding a relationship between them to formalize evolutions). In particular, this means tracking changes in representation: adding/removing a code, switching from a list of codes to numeric... Because the data producers have the most complete knowledge, they are better placed to specify these evolutions.

4.2 Ongoing work

INSEE is currently the largest provider of data for CASD. CASD has documented at least one year of each of all the data sources made available by INSEE through their secure access systems, including some of the older years. INSEE has retrieved 163 of these metadata files in XML format. They will re-use them to back-document their data files in the RMéS repository and thus benefit from interoperability. The benefits of implementing such a process for the provider and disseminator are more significant in this case of high-volume file exchanges.

4.3 Next Steps

CASD aims to consolidate this process and encourage its generalization within the other producers from the French Official Statistical System, so that even more quality documentation can be displayed.

At this point in time none of the other producers providing data through CASD use the DDI standard. The entry into this standard and to comply with it has a cost. It requires the mobilization of human and financial resources if we want to use Colectica. The DDI standard requires trained personnel. The use of Colectica software requires the subscription of a license. Although not free, the Colectica software has the advantage of being easy to use and well-structured, allowing you to produce DDI-standard metadata. All these constraints limit the possibilities of uses of the standard with other producers.

However, producers are showing increasing interest in this standard as a mean of providing quality metadata to researchers or administrations. Through this experiment, CASD can position itself as a reference for other producers to help produce DDI-standard metadata, to facilitate the process of sharing and displaying documentation.

5. Conclusion

To summarize, the main purpose of the CASD online documentation is to provide users with an overview of the confidential data available in the secure area, to give them an idea of the variables and modalities, and to help them in their application. Due to the large number of data sources available, documentation is time-consuming.

The experimentation shows some of the benefits of sharing the same standard. It is very promising, as it facilitates reusability and speeds up documentation display. Harmonization of documentation processes will reduce the reinterpretation of the meaning of variables. It also allows a better understanding of the metadata sharing landscape and practical insight into future needs. It requires to define and share precise best practices in order to avoid re-manipulating DDI files and so fluidizing the workflow.

References

CASD website https://www.casd.eu/en/data-used-at-casd/

Data Documentation Initiative (DDI) standard https://ddialliance.org/

Colectica Designer© https://colectica.com/

INSEE's Statistical Metadata Repository RMéS https://www.insee.fr/en/information/4195079?sommaire=4195125

All employees databases - job position data 2019 https://www.casd.eu/en/source/all-employees-databases-job-position-data/

DDI Alliance (DDI-Cross Domain Integration: Detailed Model) https://ddi-alliance.atlassian.net/wiki/download/attachments/860815393/Part_2_DDI-CDI_Detailed_Model_PR_1.pdf?version=3&modificationDate=1586887411228&cacheVersion=1&api=v2