The 'DDI-CDI Converter (Prototype): Wide Table Generation for STATA & SPS' in the context of WorldFAIR WP 6 - 'Social Surveys'

Benjamin Beuster, Hilde Orten

1. Introduction

1.1. WorldFAIR WP6 Social Surveys objectives and results

The main objectives of WP6 ‘Social Surveys’ has been to conduct a comparative study of the data management, harmonisation and integration practices between the European Social Survey European Research Infrastructure Consortium (ESS ERIC) and one of the satellite countries - Australia, through the AUSSI-ESS. Administrative procedures, data and metadata management and technical environments have been explored, and through three project deliverable reports[1].

The project has, among other, been examining how the use of the Data Documentation Initiative (DDI) metadata standard products can make multi-national data collections increasingly reusable and interoperable, first and foremost basing on the more domain specific core products of the DDI Alliance: DDI-Codebook[2] and DDI-Lifecycle[3] that are currently in use in the AUSSI-ESS and ESS ERIC data archives, but also touching upon cross-domain standard components like for example the variable cascade[4].

1.2. Looking ahead towards cross-domain interoperability

In the recent years, however, interdisciplinary research problems have become more and more prevalent, and integration of data from different disciplines is often needed to address them.

Data from diverse research domains often vary a lot in terms of structure, complexity and volume, and in order to be able to document data from different research disciplines for the purpose of integration, a metadata standard that can account for differences in data is needed. We are currently exploring how the forthcoming DDI-Cross Domain integration (CDI)[5] metadata standard can be used to increase the interoperability and reusability of disciplinary and multidisciplinary data.

---

[1] 6.1: ‘Cross-national Social Sciences survey FAIR implementation case studies’. https://zenodo.org/doi/10.5281/zenodo.7584437
6.2: ‘Cross-national Social Sciences survey FAIR implementation case studies’. https://doi.org/10.5281/zenodo.8308012
6.3: ‘Pilot Testing Harmonisation Workflows’. https://doi.org/10.5281/zenodo.10724744

[2] https://ddialliance.org/Specification/DDI-Codebook/2.5/

[3] https://ddialliance.org/Specification/DDI-Lifecycle/

[4] Variable Cascade: DDI Cross Domain Integration (DDI-CDI), version 1.0, release candidate 1 (2023)
https://ddialliance.org/Specification/ddi-cdi

[5] https://ddialliance.org/Specification/ddi-cdi

1.3. The idea of a DDI-CDI converter tool

To move forward with this work, a converter prototype application to convert common rectangular (wide) data files from proprietary programmes to DDI-CDI has been developed. The work was initiated at the CODATA/DDI Alliance Dagstuhl Workshop 2023: 'DDI-CDI: Realising interoperable data services in the metadata ecosystem'[6], and has been further developed in the context of WorldFAIR's WP6 'Social Surveys'[7].

2. The DDI - Cross-Domain Integration standard and it's role in WorldFAIR's CDIF

2.1. What is DDI-CDI - coverage and usages

DDI-CDI is a new metadata specification from the DDI Alliance, which focuses on areas often not addressed, but critical for data integration: Granular variable-level descriptions, and the full processing context which has produced the data.

It is designed to describe the integration of data coming from different sources, infrastructures, or disciplines/domains, and it describes not only wide data sets, but also long, multi-dimensional, and key-value types.

It is also designed to work with other existing domain and web standards, extending them to support the combination and reuse of data more easily. It works with other DDI specifications (DDI-Codebook and DDI-Lifecycle) as well as popular vocabularies such as W3C's DCAT[8] , SKOS[9] , PROV[10], and many others.

2.2. WorldFair's CDIF and the role of CDI

The Cross-Domain Interoperability Framework (CDIF)[11] is a deliverable under WorldFair's WP2, providing a set of guidelines, practices and metadata profiles, created in order to facilitate the use of existing and much used domain-independent metadata standards like the ones mentioned above, together and across domains. In CDIF DDI-CDI plays an important role, as it is used to cover the data description part.

The DDI-CDI metadata profile used for the converter prototype application (see the Appendix below) is a little more advanced than the CDI profile used for CDIF, in that it also covers the 'data point' component (equivalent to a 'cell' in a data file), and also contains a more elaborate use of the variable components of the standard. Still there are lots of similarities between the two profiles, and we consider the tool a useful contribution to the family of WorldFAIR results, both from within the social sciences and cross-domain.

---

[6] https://www.dagstuhl.de/en/seminars/seminar-calendar/seminar-details/23393
[7] https://worldfair-project.eu/social-surveys/
[8] https://www.w3.org/TR/vocab-dcat-3/
[9] https://www.w3.org/2004/02/skos/
[10] https://www.w3.org/TR/prov-overview/
[11] https://worldfair-project.eu/cross-domain-interoperability-framework/

3. The 'DDI-CDI Converter (Prototype): Wide Table Generation for STATA & SPSS'

3.1. What is it?

The DDI-CDI Converter Prototype is a Python-based web application developed to convert proprietary statistical files from Stata and SPSS into the open DDI-CDI format, using DDI-CDI-XML syntax for data representation. This prototype is intended to meet the growing demand for interoperability and data sharing by transforming proprietary data formats into an open, standardized, and machine-readable format.

Stata and SPSS are widely-used statistical software packages for analyzing structured, rectangular data. SPSS, in particular, presents data in a 'Data View' table, complemented by a 'Variable View' table that holds essential variable metadata such as variable types, names, labels, value labels, missing values, and measurement levels.

The tool converts both the data and its associated metadata from these statistical packages into the DDI-CDI XML, using 25 classes from the DDI-CDI model.[12]

Users are then able to view and download a comprehensive XML-file that contains both the cell data and the variable metadata, ensuring that it can be accessed and interpreted by any tool compatible with XML. This capability significantly enhances the accessibility and application of statistical data across diverse platforms and aligns with the wider objectives of data preservation and reuse within the research community.

Furthermore, the tool provides an overview of the 25 DDI-CDI model classes used to generate the XML data, complete with links to the CDI's online field-level documentation[13].

This feature is intended to fulfill a dual purpose: firstly, to facilitate the practical implementation of the DDI-CDI model, and secondly, to support training and capacity-building activities within the DDI community.

3.2. Main components

The primary components of the tool include two tables that resemble those in SPSS: the data view and the variable metadata view. Additionally, the tool generates CDI-XML from the uploaded data, which is displayed in a separate, scrollable box.

Upon accessing the web tool, users are prompted to upload a data file. The application then processes this file to generate the data view table and the variable view, as well as the XML representation of the data. Technically, the data files are imported into Pandas[14] data frames using the Pyreadstat [15]library and then converted into XML. The processing time may vary, often taking several seconds depending on the file size. Once the processing is complete,

---

[12] DDI Cross Domain Integration (DDI-CDI), version 1.0, release candidate 1 (2023) https://ddialliance.org/Specification/ddi-cdi

[13]CDI field-level documentation https://ddi-cdi-resources.bitbucket.io/2024-03-12/field-level-documentation/

[14] Pandas https://pandas.pydata.org/docs/

[15] Pyreadstat https://pyreadstat.readthedocs.io/en/latest/

the data and variable view tables are presented alongside the XML in a scrollable box. To the user, the data table appears as a conventional rectangular data table, organized into rows and columns with column names at the top.

The 'Variable View' table contains essential metadata about the variables, such as variable types, names, labels, value labels, missing values, and measurement levels. Additionally, users can select variables to tag as identifier variables, which can be used for merging across data files, even though this is not a native concept in Stata or SPSS. When one or multiple variables are selected, the XML is regenerated and updated accordingly.

3.3 CDI Metadata for Data and Variable View

DDI-CDI stands out from other DDI standards like DDI-Codebook and DDI-Lifecycle by its ability to cover both the structure and content of data at a granular level. It captures the full metadata for each cell in a dataset, ensuring a rich description of both the data and its variables within a data structure.

A dataset in DDI-CDI is simply said split into two parts: the logical and the physical. The logical part consists of a logical structure, components and variables, which are conceptual representations of the data. The physical part, known as the PhysicalDataSet, corresponds to the actual data entries and includes detailed elements like DataPoints and InstanceValues, which represent individual pieces of data within the cells.[16]

DDI-CDI's granular approach allows for flexible reorganization of data without losing metadata, enabling various structures like the WideDataStructure. This flexibility is crucial for adapting data to different analytical needs while maintaining its descriptive richness.

The Variable View in the tool reflects this structure by displaying extensive metadata for each variable, such as role, type, name, labels, and missing values, in a user-friendly table format. This metadata is part of the logical construct that defines the dataset's structure and its components within the DDI-CDI model.

3.3. Limitations of the tool

The tool is designed to process data files and generate metadata. While the tool offers valuable output in the form of CDI-XML, it does have certain limitations that should be considered:

1. Computing Resources: The tool is deployed on Azure and relies on scalable compute instances to process files and create large metadata files, including granular cell metadata. This conversion process is resource-intensive and requires significant computing power, which can be expensive. As a result, the tool is currently limited to a medium costly compute instance (CI), which may not be suitable for all users or organizations due to budget constraints.

---

[16]DDI-Cross Domain Integration: Detailed Model (2024) https://bitbucket.org/ddi-cdi-resources/ddi-cdi/src/master/build/high-level-documentation/DDI-CDI_Model_Specification.pdf

2. Prototype Capacity: Given that the tool is a prototype, its capacity to handle data is limited. It may not efficiently process data files with thousands of rows and with hundreds of variables and classifications. Users with large datasets may find the tool less effective, as it might struggle with the volume and complexity of the data.

3. Limited Cell-Metadata Generation: The tool is designed to generate metadata for a limited number of rows, by default up to 5 rows of the dataset but can be changed in the tool configuration. The current row limit is due to the CDI structure, which uses multiple elements per cell in a file and would result in huge XML-files. Consequently, the metadata representation may not reflect the entirety of larger datasets, which could be a significant limitation for users who require comprehensive metadata for their entire dataset.

4. Scalability: As the tool is currently in its prototype phase, scalability is a potential concern. Users processing large datasets regularly may encounter performance bottlenecks, leading to longer processing times or even the possibility of system overloads. A proposed solution to solve these issues is to store the converted metadata in a database. This would enable the tool to generate and retrieve metadata more efficiently, improving both performance and scalability. Such an approach would facilitate faster access to metadata and reduce the computational demands of the conversion process.

3.4 Next Steps with the Tool

The tool is now prepared for community testing. Feedback from this phase will be used to address potential bugs. More critically, for production use, we need to consider a more robust storage and retrieval system, as mentioned in the Scalability section, to manage metadata in a database more effectively.

Another idea for development is to expand the tool's capabilities by adding support for converting long data formats and facilitating conversions between long and wide formats. This would require the integration of more CDI classes from the CDI model into the existing subset of the model.

APPENDIX

DDI-CDI profile for the tool

In the DDI-CDI Subset profile, 25 classes from the DDI-CDI model (as of March 12, 2024) are selected. These classes are organized based on their roles in format description, data description, conceptual components, and representations.

For format description, the classes include:
- PhysicalDataSetStructure
- PhysicalDataSet
- PhysicalRecordSegment
- PhysicalSegmentLayout
- ValueMappingPosition

- ValueMapping
- DataPointPosition
- DataStore
- LogicalRecord

For data description, the classes are:
- WideDataStructure
- IdentifierComponent
- MeasureComponent
- PrimaryKey
- PrimaryKeyComponent
- DataPoint
- InstanceValue
- WideDataSet
- Notation

For conceptual components, the classes are:
- InstanceVariable
- Category

For representations, the classes are:
- SubstantiveValueDomain
- SentinelValueDomain
- ValueAndConceptDescription
- Codelist
- Code

The full CDI field-level documentation, including all available classes and a detailed description of each class, can be found here:
(https://ddi-cdi-resources.bitbucket.io/2024-03-12/field-level-documentation/