Integrating Blaise 5 and DDI Lifecycle 3.3

Dan Smith, Jeremy Iverson, Colectica

1. Abstract

This paper is pleased to introduce a new Blaise 5 to Data Documentation Initiative (DDI) Lifecycle conversion tool. DDI Lifecycle is an open metadata standard that is used for describing survey specifications and the resulting datasets, along with study, process, and linage information. Colectica Designer has supported a Blaise 4 and Blaise 5 to DDI Lifecycle converter for many years, utilizing a custom-built grammar for the Blaise language. While this has worked well for converting the major structures of a Blaise survey into DDI Lifecycle, not all aspects of the Blaise language were supported. External language definitions within Blaise .bitt files were not processed, as only the Blaise source code was processed and the translations are added later in the .bmix build process. Another issue was keeping the grammar updated with the new language features that are being added into each new Blaise 5 release.

To improve the story for Blaise and DDI Lifecycle integration, Statistics Netherlands and Colectica partnered to jointly develop several DDI tools for the Blaise ecosystem. The first tool developed was the Blaise Colectica Questionnaires tool, which is used to specify a survey specification in DDI Lifecycle and generate corresponding Blaise code. This paper introduces a second tool, a Blaise 5 to DDI Lifecycle 3.3 converter. The Blaise Colectica DDI Connector takes the approach of using the Blaise API directly to process compiled Blaise .bmix survey specifications and .bdix data definitions to create DDI Lifecycle 3.3.

The DDI items are uniquely identified and versioned, and the converter additionally computes hashes that can be used to help locate identical question or answer choices previously documented within question banks. The resultant DDI can then be exchanged or imported into any other tools that support the DDI standard.

2. Introduction

The DDI Lifecycle metadata standard is a comprehensive metadata ontology, primarily designed for documenting and managing survey specifications and statistical datasets. It provides a structured framework for describing the entire data lifecycle, from data collection and processing to preservation and dissemination. The standardization provides a common language and structure for describing data, making it easier for researchers, data producers, and data archives to communicate and understand the content and context of survey data.

2.1 History of DDI

The DDI initiative began in the late 1990s as a collaborative effort among data archives, libraries, and research organizations to develop a standardized way of documenting and managing social science and survey data. The original version of the standard was called DDI Codebook (DDI 2) and focused on single data files. DDI Codebook made no distinction between a variable description and a question description. This version of DDI also lacked unique identifiers for the various metadata being documented, which hampered efforts to reuse, link, and share metadata information across multiple datasets, studies, and organizations.

Colectica and Statistics Netherlands both had representatives participate in the drafting of a new version of DDI, which became DDI Lifecycle (DDI 3). This newer version had several new capabilities including a particular emphasis on questionnaire specification and metadata reuse:

- **Questionnaire Design and Documentation**: DDI Lifecycle allows researchers and survey designers to document survey questionnaires comprehensively. This includes specifying question wording, response options, skip patterns and routing instructions, multiple languages, and track question and block reuse. This detailed documentation helps maintain the integrity and consistency of surveys.
- **Data Collection**: Information can be tracked using DDI that describes fielding periods, populations of the respondents, and specific versions or revisions of the fielded survey instrument.
- **Data Analysis**: Researchers can use DDI to understand the structure and content of the dataset produced by a survey instrument. This information is crucial for accurately analyzing the data, including identifying the meaning of variables, handling missing data, and understanding the survey's design.
- **Data Sharing and Data Discovery**: DDI-compliant metadata enhance the discoverability of survey data. Researchers and data archives can use DDI metadata to search for relevant datasets and evaluate their fitness for research purposes. Questions, response options, and survey blocks can be reused and linked in documentation.
- **Long-Term Data Preservation**: DDI helps in preserving the context and documentation of survey data over time. This is essential for ensuring the long-term usability and integrity of social science datasets.

2.2 DDI and Blaise Terminologies

The representatives participated in several years of design discussions to ensure that DDI Lifecycle could document the general structure of Blaise surveys. The Blaise rules, blocks, fields, groups, rosters, and computations all correspond to elements of the DDI Lifecycle ontology.

- **Rules**: Rules in the Blaise survey system are akin to DDI's Question Flow and Logic elements. They specify conditional actions or skip patterns that control the flow of questions or the survey path, based on respondent inputs. These correspond to DDI's control constructs, which are nested components that describe when questions are asked and how the survey progresses.
- **Blocks**: In Blaise, blocks are groups of related questions and rules within a survey. These correspond to DDI's sequence control constructs. Sequences captures the structure and organization of questions and content within a survey specification, which is similar to how blocks function in Blaise.
- **Fields**: Fields in the Blaise survey system are equivalent to DDI's Question and Measurement items. They represent individual data elements or questions in the survey. DDI Questions is concerned with documenting variables in terms of their type, labels, question text, response options, and other metadata, which maps to how fields, their types, and role texts are defined and documented in Blaise.
- **Groups**: Blaise groups allow you to organize related questions or fields within a survey into special composite displays. This corresponds to DDI's concept of a Question Grid or Question Block.
- **Rosters**: Rosters in Blaise are used to create dynamic repeating sets of questions, such as household members or survey responses, to a list of items. DDI's Question Grids contain roster dimensions, and the DDI Looping concepts of Loop, RepeatWhile, and RepeatUntil are similar in

that they capture how certain questions or data elements are repeated for multiple items or cases within a survey.

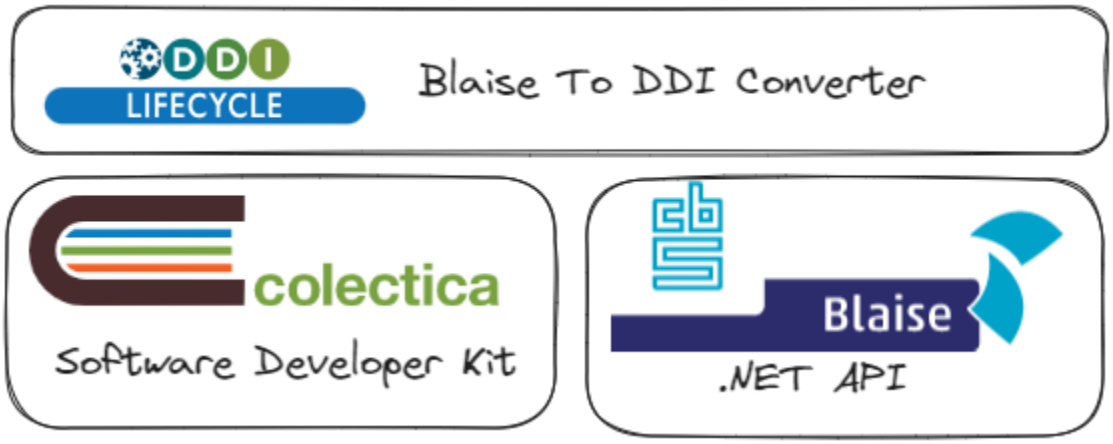
- **Computations**: Blaise allows you to define in the rules section computations or derived variables based on responses to other questions or data and store this information into fields. DDI includes a concept of computations that documents how new data values are computed from existing ones during a survey. This corresponds to how computations are used in Blaise to generate new data based on respondent inputs.

A complete detailed mapping of the Blaise metadata fields to DDI items and properties can be found within the tool distribution.

3. Blaise to DDI Converter

3.1 Components

The Blaise to DDI Converter is a tool that reads compiled Blaise .bmix and .bdix files to generate DDI Lifecycle descriptions of the survey specification and resulting raw dataset. The Blaise .NET API and the Colectica SDK are used together to convert the compiled Blaise metadata structures into DDI Lifecycle metadata.



- **Blaise .NET API**: The Blaise .NET API is a set of libraries distributed as a NuGet package to allow developers to interact with Blaise survey projects programmatically. It provides methods for reading and manipulating Blaise survey instruments and their associated metadata.
- **Colectica SDK**: The Colectica SDK is a software development kit for working with the DDI Lifecycle model. It includes libraries and tools for creating, reading, and exporting metadata in various DDI formats, as well as creating integration Addins for other Colectica tools.

Users of the Blaise to DDI Converter can use either the command line interface to create batch processes or use the GUI user interface. A .bmix file will be chosen to load, and the resulting DDI Lifecycle metadata can be written to a .xml file or saved directly to a Colectica repository. All described metadata content that is generated, including questions and type definitions, will be given globally unique identifiers to allow storage in question banks or type libraries.

3.2 Reuse with Question Banks and Type Libraries

It is useful to find all instances of identical Blaise fields and types, DDI questions, and codelists that appear across many different Blaise survey instruments during the conversion process. This allows reusing the unique identifiers assigned to metadata items and producing documentation and web portals showing question reuse.

During each conversion of a .bmix to a DDI Lifecycle description of a survey, new identifiers are generated for each metadata item. To allow finding reused questions and types, the Blaise to DDI Converter creates unique hashes of the content of the items. Creating a unique hash of the contents of question text and response options is a useful technique for identifying and finding duplicates that may already be stored in a repository. These hashes serve as fingerprints for the field or type and can be used to compare and match questions efficiently.

DDI Lifecycle and Blaise are multilingual; the normalization process for each piece of text involves sorting the metadata by language tag and creating a string in the format of {language-tag}:{text}:{languagetag}:{text}: and so forth. A similar normalization is done for type codes and multilingual value labels. Text fields of the questions and codelists are concatenated to create a normalized concatenated string. A SHA-256 hash is created from the normalized string and stored as an additional identifier within each metadata item. This hash can now be used as a reference for finding and identifying duplicate questions efficiently.

During the conversion process, the Blaise to DDI Converter can optionally cross-reference the computed SHA-256 hashes in a Colectica repository to see if the question or codelist with identical content has already been created. The converter employs the computed hash and does a search of the repository. If a match is found, the converter can utilize the pre-existing metadata item and its corresponding identifier instead of generating a new duplicate and redundant metadata item. This allows for deduplication during the conversion process instead of as a post-processing step. Using this approach, you can quickly identify duplicate questions and response types without the need for a more resource-intensive textual or semantic comparison.

While this process of using hashes of metadata fields finds exact matches, DDI is also capable of relating similar questions. A question can be associated with a concept. Similar questions can be related by linking to the same concept. Conceptual linking of questions can be addressed as a step after the Blaise to DDI conversion, done by this tool.

4. Summary/Reflections

The introduction of the new Blaise to DDI Converter represents a significant advancement in terms of sustainability and efficiency compared to the previous Colectica Designer source code parser. The new converter leverages the official Blaise API, marking a fundamental shift from the previous approach. This integration with the official API allows the tool to directly access and interact with Blaise .bmix survey projects and their metadata without having to parse Blaise source files. This is a pivotal enhancement because it means the converter is now closely aligned with the Blaise survey platform itself, allowing the utility to easily keep pace with any changes introduced to the Blaise language.

One of the primary advantages of utilizing the official Blaise API is the new converter's ability to adapt to changes in the Blaise language. As Blaise evolves and introduces new features or modifications, the converter can readily accommodate these changes. This adaptability ensures that the tool remains

compatible with the latest versions of Blaise 5, minimizing the risk of compatibility issues or data parsing errors that may have occurred with the previous Colectica source code parser.

The development of the new converter is a collaborative initiative between CBS (Central Bureau of Statistics) and Colectica. This partnership not only demonstrates a commitment to the ongoing improvement of the tool, but also enhances its stability. The combined expertise and resources of both organizations contribute to the tool's robustness and reliability, and will lead to continuous updates, maintenance, and support for the converter.

DDI Lifecycle places a strong emphasis on reusable metadata descriptions, change tracking, documenting lineage, and unique identification. Whether comparing different draft versions of a survey during the developmental stage or linking different questions across a multitude of surveys and projects, DDI offers a production-ready framework for storing and establishing relationships among this information. When applying this to Blaise fields and response types, many new opportunities for reporting and visualizations can be imagined. While starting with exact matching for linking questions across surveys is a solid foundation, it will be interesting to see what other types of question comparison the community could find useful to be included within the tool during the DDI conversion process.

In summary, the new Blaise to DDI Converter is a sustainable solution that benefits from its integration with the official Blaise API, adaptability to changes in the Blaise language, and the collaborative effort between CBS and Colectica. These factors combine to ensure that the converter will be a dependable and up-to-date tool for converting Blaise survey projects into DDI Lifecycle-compliant metadata.