Data Documentation Initiative

DDI Working Paper Series -- Best Practices, No. 4

Subject:

Workflows for Metadata Creation Regarding Recoding, Aggregation and Other Data Processing Activities (2009-03-21)

Document identifier:

http://dx.doi.org/10.3886/DDIBestPractices04

Target audience:

Producers of DDI metadata; researchers in the sense of both data collectors and analysts; data processors whenever and wherever processing occurs (data collection, data cleaning, dataset creation, archiving process, or other)

Authors:

Hans Jørgen Marker, Wolfgang Zenk-Möltgen, Wendy Thomas, Achim Wackerow

Editors:

Hans Jørgen Marker, Wendy Thomas

Abstract:

This best practice discusses the capturing, in DDI metadata, of the processes of data aggregation, recoding and data processing.

Status: Draft

This document is updated periodically on no particular schedule. Send comments to ddi-bp-editors@icpsr.umich.edu

24 INTRODUCTION 3
25 1.1 Problem statement 3
26 1.2 Terminology 3
27 2 BEST PRACTICE SOLUTION 4
28 2.1 Definitions 4
29 2.2 Best Practice behavior 4
30 2.3 Discussion 7
31 2.4 Example 8
32 3 REFERENCES 10
33 3.1 Normative 10
34 APPENDIX A. ACKNOWLEDGMENTS 11
35 APPENDIX B. REVISION HISTORY 13
36 APPENDIX C. LEGAL NOTICES 14

Introduction

DDI 3 facilitates the creation of metadata at a variety of starting points from the hypothesis for a study through the capturing of legacy metadata. How and where one starts capturing metadata depends upon the data being described, the application within which it is used, and the organizational needs of the creators. The best practices on workflow provide guidelines for setting up metadata creation processes within different environments, identifying organizational and application features that impact the process structure, addressing salient questions/issues in setting up the process, and determining the implications of various starting points and process orders:

1. Metadata Creation Regarding Recoding, Aggregation, and Other Data Processing Activities (this document)
2. Archival Ingest and Metadata Enhancement [see References section]
3. Dissemination and Discovery: User Perspective [see References section]

1.1 Problem statement

Data transformations related to aggregation, recoding, and data processing need to be captured in the metadata to support both data processing (machine-actionable) and user understanding of the resulting data content. This information must support both reuse of metadata and data but also the review and evaluation of resulting research. Stakeholders lack information if data transformations are not documented. This kind of documentation is essential to tracking the evolution of a dataset over time, as indicated in the OAIS reference model that defines requirements for digital preservation. Stakeholders include: researchers, research councils/funding agencies, data producers, archivists, librarians, data and metadata users, registry managers, and research analysts.

1.2 Terminology

The key words *must, must not, required, shall, shall not, should, should not, recommended, may,* and *optional* in this document are to be interpreted as described in **[RFC2119].** Additional DDI standard terminology and definitions are found in http://www.ddialliance.org/bp/definitions

2 Best Practice Solution

2.1 Definitions

Open Archival Information System (OAIS): A reference model of the space community that governs general archival activities and policies. Includes:

SIP: Submission Information Package
AIP: Archival Information Package
DIP: Dissemination Information Package

Control operations: Methods to facilitate data control performed by the primary investigator or by the data archive. Specify any special programs used for such operations. The "agency" attribute maybe used to refer to the agency that performed the control operation.

Cleaning operations: Methods used to "clean" the data collection, e.g., consistency checking, wildcode checking, etc. The "agency" attribute permits specification of the agency doing the data cleaning.

Weighting: The use of sampling procedures may make it necessary to apply weights to produce accurate statistical results. Describe here the criteria for using weights in analysis of a collection. If a weighting formula or coefficient was developed, provide this formula, define its elements, and indicate how the formula is applied to data.

Logical record: A reference to a data record that is independent of its physical location. It may be physically stored in two or more locations.

NCubes: Describe the logical structure of an n-dimensional array, in which each coordinate intersects with every other dimension at a single point. The NCube has been designed for use in the markup of aggregate data.

Resource package: A resource package is a means of packaging any maintainable set of DDI metadata for referencing as part of a study unit or group. A resource package structures materials for publication that are intended to be reused by multiple studies, projects, or communities of users. A resource package uses the group module with an alternative top-level element called Resource Package that is used to describe maintainable modules or schemes that may be used by multiple study units outside of a group structure.

2.2 Best Practice behavior

Data processing is captured by a set of elements that allow both descriptive and machine-actionable content regarding control operations, cleaning operations, weighting, data

appraisal, and variable generation. While commonly occurring between data collection and production of the dataset, in fact, data processing can and does take place at many points along the data life cycle. Activities or events that can trigger the capture of metadata related to data processing may include the following (not an exhaustive list):

- Instrument development
- Confidentiality checks
- Transfer from data collection to data file
- Error correction
- Incorporating externally collected data items
- Creating indicator variables
- Harmonization
- Building classifications
- Comparison
- Creating aggregated statistics
- Preparation for analysis
- Creating new logical records
- Distribution of data
- Archival ingest and data management
- Creating subsets
- Linking datasets and/or data integration
- Complex file preparation
- Analysis

In effect, data processing is not a step in an overall process but occurs at many stages and is part of many other workflows throughout the overall process. Because of the ability of this metadata to capture specific machine-actionable commands, it can be used to process not only the data, but also automate the creation of metadata later in the process.

This best practice addresses the general process of data transformation and then identifies specific cases that may require special attention.

Processing Event

When a data processing event has occurred, new metadata should be created to reflect that event. This includes researchers and analysts who recode or restructure the content of the data in the process of analysis. The specific recoding or processing action should be captured from the statistical package or other processing tool and entered into the DDI format so that their results can be understood and evaluated by peers. For analysis this includes preparation of the data, the statistical methods applied, and the results. Processing metadata should be captured in a machine-actionable way (e.g., SAS or SPSS code, mathematical formula, or system-independent formal language) when possible, but should always include a human-readable explanation of the process. In addition to the specific processing commands, metadata should include the *agent* who executed the process, *when* it was done (specifically and in relation to the overall life cycle), the *purpose* of the process, and the *rationale* for use of the specific method. In addition to entering this information in the data processing elements, major events should also be noted in the Life cycle list in the Archive module. This makes it easier for both those managing the development of data and users to understand where and when data processing occurred throughout the workflow.

The DDI Processing Event element is a packaging mechanism that contains discrete events. Each event should contain the appropriate descriptive type, e.g., Control Operation and its associated Coding.

Control Operation and Cleaning Operation should be packaged with any related codes or data appraisal events that occurred as a single integrated process.

Weighting describes the process of determining overall or specific weights. When the weighting process results in one or more weight variables, the processing event should include both the weighting description and the generation code for each weight variable. If a study contains a standard weight, this should be entered as a numeric value in weighting. It is to be expected that future versions of the standard will have a machine-actionable home for the standard weight.

Data Appraisal Information contains separate elements for response rate, sampling error description, and other appraisal information. This should be coupled with any specific coding information related to the process used for determining sampling error or other appraisal. For surveys it is considered a best practice to include both the response rate and sampling error in all documentation.

Coding contains two structural types, General Instruction and Generation Instruction. General instruction captures processes that were used on large sections of data such as imputation processes, suppression rules, and handling of non-response to questions. General instructions express overall processing of the resulting data file. A general instruction for a subset of the data may override another general instruction for the whole file. Generation instructions provide processing information for specific variables. In addition to providing specific information to the user regarding the generation of the final dataset, these fields in conjunction with information on concepts and questions can be used to generate metadata in the logical product, physical data structure, and physical instance. When creating aggregate data expressed as NCubes, the aggregation process for creating specific cell contents is captured in aggregation within generation instruction.

Statistical summary data housed in physical instance does not require processing codes as these are standard weighted and unweighted frequencies with or without filters. Statistics created through the NCube creation process should be recorded with processing events. This would include the recoding of variables to created code representations such as age cohorts. In addition, NCubes require the creation of variables to represent specific measures such as counts and percents. These are captured in generation instruction to clarify how a count was defined or in the case of a percent exactly what is being used as the numerator and denominator in each table.

2.3 Discussion

The advantage to documenting data transformation is that the structure provides a systematic and coherent means of capturing a variety of processing events that can take place at many points along the life cycle. It facilitates access to the processing documentation and makes it available for reuse within and between studies.

The use of this best practice supports the use of data processing content to drive the production of data and metadata.

It packages the description of a process with the coding so that it can be referenced as a specific step from many points in the life cycle. It provides transparency for each step within a workflow, thereby aiding both understanding of how the data arrived at its current state at any point in the process workflow, and management of the process itself.

Capturing this type of metadata requires a change in work processes to capture the specific processing done within statistical software or other applications. Currently, much of this information is lost as it is not routinely maintained by the application along with the resulting data.

In addition, this type of workflow change provides the opportunity to capture the reasoning behind specific data processing activities.

Capturing this information can easily be perceived as an additional burden especially during the analysis process, which may have short term goals. The payoff for the researcher or group of researchers is improved quality of the result. It supports the ability to replicate the analysis for evaluation and validation. This approach supports best practices as they have been defined for scientific activities.

However, there are more easily identifiable payoffs for the data producer and data manager. Capturing the process commands in a single location allows them to be reused by reference. Processing event metadata can be used to drive applications for data and metadata creation. Aggregation processes can be captured and used to produce aggregated data from microdata on demand, as opposed to creating and storing it as a dataset. Put another way, the metadata can drive rule-based aggregate data presentation.

For production and management organizations that want to ensure consistent processing such as cleaning practices, confidentiality processes, etc., the processing event information and coding structures can be published and maintained as a resource package for reuse within the organization. See the Schemes Best Practice [see References section] for more information on preparing resource packages for reuse.

The definition of what constitutes an individual processing event that results in the transformation of data from one stage to another cannot be clearly defined. A process may result in a number of transient versions of the data between declared stages. The contents of a single Processing Event should be coherent (e.g., contain a single type of processing event and its related coding) and sufficiently discrete to support clear referencing and execution of the command codes. When does a process have a stage one and a stage two and what constitutes the processing event? Is it each individual step (e.g., code) or a combination of a number of steps? If the steps within a process must occur in a specified order, then they should be listed as separate process events.

Currently there is no machine-actionable means of providing sequential information for the processing of coding; however, the order in which steps were taken in an actual chain of events can be captured in the Life cycle list.

2.4 Example

This example shows the use of Processing Event as packaging for a single step (e.g., cleaning instructions and coding).

```
<d:ProcessingEvent isIdentifiable="true" id="PE_1" isDerived="false">
<d:CleaningOperation>
<r:Description>For all NCubes without suppression and whose contents are
additive. The contents of the discrete cells in the NCube are aggregated
and compared to the universe value. All errors have been corrected or
noted.</r:Description>
</d:CleaningOperation>
```

```
<d:CleaningOperation>
<r:Description>For all geographic levels that roll up into a parent level
(e.g., U.S. Counties within U.S. States). All cells where suppression is
not used and whose contents are additive, have been aggregated and
compared to the value in the parent geography. All errors have been
corrected or noted.</r:Description>
</d:CleaningOperation>
<d:Coding isIdentifiable="true" id="Coding_1">
<d:GenerationInstruction>
<d:SourceVariable
isReference="true"><r:ID>V1</r:ID><d:Mnemonic>BaseAge</d:Mnemonic>
<r:Description>Recodes discrete age into 3 age cohorts; under 18, 18 to
64, and 65 and over.</r:Description>
<r:Command>
<r:CommandText formalLanguage="SPSS">RECODE BaseAge (Lowest thru 17=1) (18
thru 64=2) (65 thru Highest=3) INTO AGE.EXECUTE.</r:CommandText>
</r:Command>
</d:Coding>
</d:ProcessingEvent>


Representation/CodingInstructionRefrence in Variable "AGE" will reference
"Coding_1"
```

Parallel of Processing Event content and listing in life cycle

```
<r:LifecycleInformation>
<r:LifecycleEvent isIdentifiable="true" id="le_1">
<r:EventType>Cleaning Operation</r:EventType>
<r:Date><r:SimpleDate>2008-04-01</r:SimpleDate></r:Date>
<r:AgencyOrganizationReference isReference="true">
<r:ID>mpc</r:ID>
</r:AgencyOrganizationReference>
<r:Description>Completed cleaning operations verifying cell counts within
NCubes and cell counts within geographies which roll completely into a
higher level for which they provide comprehensive and complete
coverage..</r:Description>
```

3 References

DDI Best Practice: Workflows - Archival Ingest and Metadata Enhancement:
http://dx.doi.org/10.3886/DDIBestPractices03

DDI Best Practice: Workflows - Data Discovery and Dissemination: User Perspective:
http://dx.doi.org/10.3886/DDIBestPractices02

DDI Best Practice: DDI 3.0 Schemes:
http://dx.doi.org/10.3886/DDIBestPractices07

3.1 Normative

[RFC2119] S. Bradner, Key words for use in RFCs to Indicate Requirement Levels, http://www.ietf.org/rfc/rfc2119.txt, IETF RFC 2119, March 1997.

OASIS, Best Practice, http://www.oasis-open.org/committees/uddi-spec/doc/bp/uddi-spec-tc-bp-template.doc, 2003

Appendix A. Acknowledgments

The following individuals were members of the DDI Expert Workshop held 10-14 November 2008 at Schloss Dagstuhl, Leibniz Center for Informatics, in Wadern, Germany.

Nikos Askitas, Institute for the Study of Labor (IZA)

Karl Dinkelmann, University of Michigan

Michelle Edwards, University of Guelph

Janet Eisenhauer, University of Wisconsin

Jane Fry, Carleton University

Peter Granda, Inter-university Consortium for Political and Social Research (ICPSR)

Arofan Gregory, Open Data Foundation

Rob Grim, Tilburg University

Pascal Heus, Open Data Foundation

Maarten Hoogerwerf, Data Archiving and Networked Services (DANS)

Chuck Humphrey, University of Alberta

Jeremy Iverson, Algenta Technology

Jannik Vestergaard Jensen, Danish Data Archive (DDA)

Kirstine Kolsrud, Norwegian Social Science Data Services (NSD)

Stefan Kramer, Yale University

Jenny Linnerud, Statistics Norway

Hans Jørgen Marker, Danish Data Archive (DDA)

Ken Miller, United Kingdom Data Archive (UKDA)

Meinhard Moschner, GESIS - Leibniz Institute for the Social Sciences

Ron Nakao, Stanford University

311 Sigbjørn Revheim, Norwegian Social Science Data Services (NSD)

312 Wendy Thomas, University of Minnesota

313 Mary Vardigan, Inter-university Consortium for Political and Social Research (ICPSR)

314 Joachim Wackerow, GESIS - Leibniz Institute for the Social Sciences

315 Wolfgang Zenk-Möltgen, GESIS - Leibniz Institute for the Social Sciences

316

317 Appendix B. Revision History

318

Rev	Date	By Whom	What
0.9	2008-02-08	Stefan Kramer	Removed date from filename to accommodate linking. Began revision history tracking.

319

Appendix C. Legal Notices