Data Documentation Initiative

DDI Working Paper Series -- Best Practices, No. 3

Subject:

Work Flows - Archival Ingest and Metadata Enhancement (2009-02-22)

Document identifier:

http://dx.doi.org/10.3886/DDIBestPractices03

Authors:

Karl Dinkelmann, Michelle Edwards, Jane Fry, Chuck Humphrey, Kirstine Kolsrud, Stefan Kramer, Jenny Linnerud, Hans Jørgen Marker, Meinhard Moschner, Ron Nakao, Wendy Thomas, Achim Wackerow, Wolfgang Zenk- Möltgen

Target Audience:

Archivists, information managers, data/digital information curators, software developers, repository managers

Editors:

Mary Vardigan

Abstract:

This Best Practice discusses workflows for DDI usage in the context of archival ingest and metadata enhancement, beginning at the point of the handoff between the data provider and the archive.

Status: Draft

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

1 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 [see References section]
2. Archival Ingest and Metadata Enhancement (this document)
3. Dissemination and Discovery: User Perspective [see References section]

1.1 Problem statement

This Best Practice concerns how DDI 3 can support and enhance the intake, augmentation, and preservation functions of data archives and data libraries. Ideally, DDI 3 can drive archival activities and products and provide new benefits, including increased possibilities for lifecycle support, comparability (spatial, temporal, topical), grouping, and metadata reuse.

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

METS: Metadata Encoding and Transmission Standard

PREMIS: Preservation Metadata Implementation Strategies

Ingest: In OAIS terminology, the OAIS entity that contains the services and functions that accept Submission Information Packages from Producers, prepares Archival Information Packages for storage, and ensures that Archival Information Packages and their supporting Descriptive Information become established within the OAIS. Used in its verb form, ingest refers to the process of taking information into a repository.

Codebook: A document that provides information on the structure, contents, and layout of a data file.

DTD: Document Type Definition is one of several SGML and XML schema languages, and is also the term used to describe a document or portion thereof that is authored in the DTD language.

XML Schema: The XML Schema Definition Language is an XML language for describing and constraining the content of XML documents. XML Schema is a W3C Recommendation.

2.2 Best Practice behavior

There are many stakeholders in the research data life cycle, including research councils and funding agencies, researchers, data producers, archivists, librarians, users, registry managers, and secondary analysts. The user perspective should inform the workflows across the lifecycle, leading to data products that are high quality and in line with the needs of the end users, specifically in terms of data discovery and effective and adequate use and analysis.

This best practice begins at the point of the handoff between the data provider and the archive. The package of materials to be ingested into an archive is known in OAIS

terminology as a Submission Information Package (SIP). A Submission Information Package (SIP) to be ingested into the archive might typically include:

- Codebook (ASCII, Word, DDI 1, 2, or 3), ideally with full variable names, variable labels, and value labels
- SPSS portable, SAS transport, Stata data file, or ASCII raw data file with setup (command/syntax files) files
- Questionnaire and show cards
- Methodological documentation
- Organizational and other bibliographic information
- Formulas for calculating variables and weighting instructions
- Frequency counts and other univariate statistics
- Citations to publications related to the data
- Other supporting material

The SIP may be delivered to the archive in a METS wrapper.

Depending on the content and format of the digital assets ingested into an archive, workflows that an archive undertakes to add value to and preserve information will vary.

The workflows described in this document may differ depending on whether the metadata are stored in native XML format, in XML-based database platforms such as eXist, or other database platforms such as Oracle or PostgreSQL. A discussion of metadata storage is an important consideration but outside the scope of this document.

Below we consider three different cases related to documentation format:

(1) The case when an archive receives DDI 3 documentation

(2) The case when DDI 1 or 2 documentation is deposited

(3) The case when the archive receives non-DDI documentation. In general, we recommend that conversion to DDI 3 occur at the earliest stage possible in order to maximize the potential to realize specific DDI benefits for archival processing.

Case I: DDI 3 SIP -- When the Submission Information Package (SIP) to be ingested into the archive has full DDI 3 documentation generated by computer systems or in other ways

Because this documentation is characterized by having full descriptive content of variables (e.g., question texts are fully integrated into variable descriptions), it is possible to ultimately generate from the full document itself a codebook, instrument documentation, a metadata record, and software-specific syntax files for distribution at the end of data processing. This means that the workflow can be optimized.

After receiving the submitted DDI 3 documentation along with other files in the submission information package, best practice is to proceed in this way:

1. Create and run validation scripts
2. Conduct quality control on data and metadata
    a. Control data against submitted metadata to make sure they match
    b. Assess accuracy, consistency, and completeness
3. Process data and update metadata (see Best Practice on Data Processing). This step may include checking for confidentiality issues, data cleaning, etc., which DDI 3 can facilitate and describe. Document data cleaning steps followed in DDI.
4. Adjust metadata to reflect processing
5. Identify the features provided by DDI 3 that add benefits for end users. Be aware that in order to fully utilize the benefits of DDI 3, the structure and organization of the metadata should be optimized. Specifically at this stage, it is useful to consult the DDI 3 Schemes Best Practice and the Best Practice on Grouping (yet to be developed).
6. Enrich the metadata to ensure that the required modules/elements/attributes for desired functionality are present. At this stage it is useful to separate existing metadata into pieces that are reusable/maintainable, e.g., in question banks, and those that are not. For example, one might add to the DDI 3 instance:
    a. Comparable terms (geographic, temporal, and topical)
    b. Grouping of study units - e.g., identify whether the study is part of a series
    c. Grouping of trend variables

d. Referencing of master questions to country/language versions

e. Translation of metadata

7. To ensure compliance with the OAIS standard, add required content related to preservation (PREMIS)
8. Create archival metadata record (a subset of DDI 3)
9. Define any access restrictions for all/part of data
10. Make decision about what goes into the Archival Information Package (AIP) for long-term preservation
11. Create dissemination files from DDI 3 or from DDI-compliant repository or database
12. Version and publish the Dissemination Information Package (DIP) (including DDI XML along with the style sheet to render it for presentation). See for instance: http://www.ddialliance.org/DDI/related/xml-xslt.html

Case II: DDI 1 and/or 2 SIP -- When the documentation deposited is in DDI 2.1 or earlier

To transform DDI 2.1 (or earlier versions) to DDI 3, consult Appendix 4 of the DDI 3 Technical Specification Part I Overview. This addresses mapping of DDI 2.1 elements and attributes to 3. At this stage it is critical to ensure that elements are assigned unique ids (see DDI Identifiers Best Practice in References).

Note that because DDI 1 and 2 were expressed in XML as a Document Type Definition (DTD) and not as an XML Schema, the element definitions may not always have been consistently applied within and across organizations. For example, the names of data files may appear in different elements. The content of elements should be carefully evaluated.

Repeat steps 1 to 12.

Case III: No DDI SIP -- When documentation is not DDI conformant:

This case is the most complicated of the three because many types of archival workflows currently exist to handle incoming data and documentation. Having DDI 3 will help to harmonize workflows within organizations. Thus, the earlier an archive can transform documentation into DDI 3-compliant components, e.g., in databases, the more efficient the workflow will be.

Note that transformation into DDI 3 depends on having specific tools available. For a complete list of DDI transformation tools, refer to the DDI Alliance Tools site: http://tools.ddialliance.org/.

Repeat steps 1 to 12.

2.3 Discussion

The application of DDI 3 provides the potential for greater efficiency and effectiveness across the workflow of archival ingest and metadata enhancement.

This Best Practice has identified the need for the development of the following tools in prioritized order: DDI migration and conversion tools, an editing suite, a grouping and comparison tool, and a DDI 3 validation tool.

A Best Practice for metadata storage could discuss repository architecture recommendations.

METS has endorsed DDI as a metadata format and work is under way to determine the best practice in using these two standards together.

2.4 Example

Because the application of the recommended workflow will differ for each community or organization, we do not provide an example other than the general cases described above.

3 References

DDI Best Practice: Workflows for Metadata Creation Regarding Recoding, Aggregation and Other Data Processing Activities: http://dx.doi.org/10.3886/DDIBestPractices04

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

DDI tools Web page: http://tools.ddialliance.org/

PREMIS: http://www.loc.gov/standards/premis/

OAIS: http://public.ccsds.org/publications/archive/650x0b1.pdf

METS: http://www.loc.gov/standards/mets/

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

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

242 Wendy Thomas, University of Minnesota

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

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

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

246

247 Appendix B. Revision History

248

Rev	Date	By Whom	What
0.9	2009-02-08	Stefan Kramer	Removed date from filename to accommodate linking. Began revision history tracking.

249

Appendix C. Legal Notices