Data Documentation Initiative

DDI Working Paper Series -- Best Practices, No. 2

Subject

Work flows - Data Discovery and Dissemination: User Perspective (2009-02-15)

Document identifier:

http://dx.doi.org/10.3886/DDIBestPractices02

Authors:

Karl Dinkelmann, Michelle Edwards, Jane Fry, Chuck Humphrey, Ron Nakao, Wendy Thomas

Editors:

Ron Nakao and Chuck Humphrey

Target Audience:

Metadata producers, funding agencies and councils, and institutions with stewardship roles

Abstract:

Describes the best practices for metadata producers to provide end users with the resources for data discovery and dissemination.

Status:

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 [see References section]
3. Dissemination and Discovery: User Perspective (this document)

1.1 Problem statement

Each phase in the data life cycle represents a group of related processes. Within a stage, specific processes or activities, when viewed collectively, represent a significant component in conducting research. While some activities and products are intrinsic to each stage, others flow across stages. For example, the design of an experiment or survey will be integral to the Data Production stage, while data products emanating from this stage will flow throughout the model. (Stewardship of Research Data in Canada: A Gap Analysis, Draft [see References section])

The “End User” Model [see Discussion section] represents the user perspective on the data life cycle. However, metadata emerges from discrete organizations over time with workflows that often do not articulate cleanly from stage to stage across handoffs.

From the end user perspective, what are the best practices that metadata producers should follow to provide potential data analysts with the resources they need for data discovery and that archives need for dissemination?

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

71

72 2 Best Practice Solution

73 2.1 Definitions

74 Stewardship: In the context of this best practice, stewardship involves taking on custodial
75 responsibilities for a stage in the data life cycle.

76 Data life cycle: The stages of the data component of the research process, from study
77 conceptualization to data analysis and archiving, feeding back to earlier stages. This
78 process has often been depicted as linear, but the diagram embedded in this best practice
79 [see the End User Model below

Figure 1: the End User Model] offers a different perspective on it, from the user's point of view.

Citizenship: Being a contributing member of the full data life cycle and realizing that one is part of a bigger scientific picture.

End user: Anyone using any product generated in the data life cycle. Examples include research council/funding bodies, researchers, data producers, archivists, librarians, users, registry managers, research analysts/authors.

Discovery: Strategies and processes used by the end user to locate and access products (metadata, data, and other related information) of the data life cycle.

Dissemination: Data distribution with the aim of access by the end user to the products (metadata, data, and other related information) of the data life cycle.

Data and knowledge repository: Places (may be virtual) where the products (metadata, data, and other related information) of the data life cycle are located.

Knowledge transfer: The act of sharing the knowledge gained throughout the data life cycle.

2.2 Best Practice behavior

An underlying principle of metadata, data, and other information is that it can be shared and accessible to a wider user community. The often-disjointed nature of the production of metadata throughout the data life cycle can lead to fragmented metadata that can undermine the achievement of this principle. While DDI 3.0 was designed to capture metadata in a dynamic process for the purposes of informing the production process and later stages in the model, including data discovery, analysis, and re-use, best practice by metadata producers should ensure consistent and exhaustive coverage. Bear in mind that these same features in DDI 3.0 can increase fragmentation. Metadata producers are responsible for retaining the metadata integrity throughout the Data Life Cycle.

From the end user perspective, published metadata should be persistent, versioned, and accessible. The quality and completeness of the content of the metadata should be evaluated against the best practices of the wider DDI community, which should ideally be addressed through certification.

Organizations and agencies involved in the production of metadata should seek certification of their metadata creation processes. Certification should be based on best practices as defined by the DDI community. Institutions with stewardship responsibilities, such as trusted data and knowledge repositories, should establish certification criteria and processes.

Funding agencies and councils should facilitate the adoption, oversight, and use of best practices in metadata creation. Compliance and certification should be viewed as rewards in and of themselves because they promote greater return on the initial investment in metadata creation and reinforce the scientific method.

Overarching Metadata Principles

While there will always be new ways to use metadata, there are a set of important principles that data and metadata producers should always keep in mind:

- Metadata may be used in ways not anticipated at the time of generation. Producers need to look beyond their own goals and uses of their metadata, and be good citizens to the general user community to support data discovery and dissemination.
- Related to the point above, metadata should never be discarded without careful thought.
- Handoffs along the trajectory of the data life cycle (for example, the handoff from a data producer to an archive) carry risks of metadata loss. It is best to be conservative and to preserve and pass along all relevant metadata. See the best practice on archival ingest [see References section].
- Documentation of data and metadata transformations over time is key to OAIS compliance and to providing a way to track content changes and chain of custody. See the best practice on metadata creation regarding recoding, aggregation, and other data processing activities [see References section].
- Search and browse features are ubiquitous on the Web. In preparing metadata content, one should always think ahead to how content will be used with these types of finding aids.
- Controlled vocabularies used consistently can help users target the data and metadata they need.
- DDI provides rich options for the description of coverage – topical, geographic, and temporal. Coverage elements should always be populated as fully as possible.
- Concepts may be assigned in DDI starting at data conceptualization. Assigning concepts at the question level is very helpful to data analysts seeking to understand the rationale behind the question.
- All stages of data and metadata production may be relevant to the end user. For example, an end user assessing data quality may want to know that the interviewers

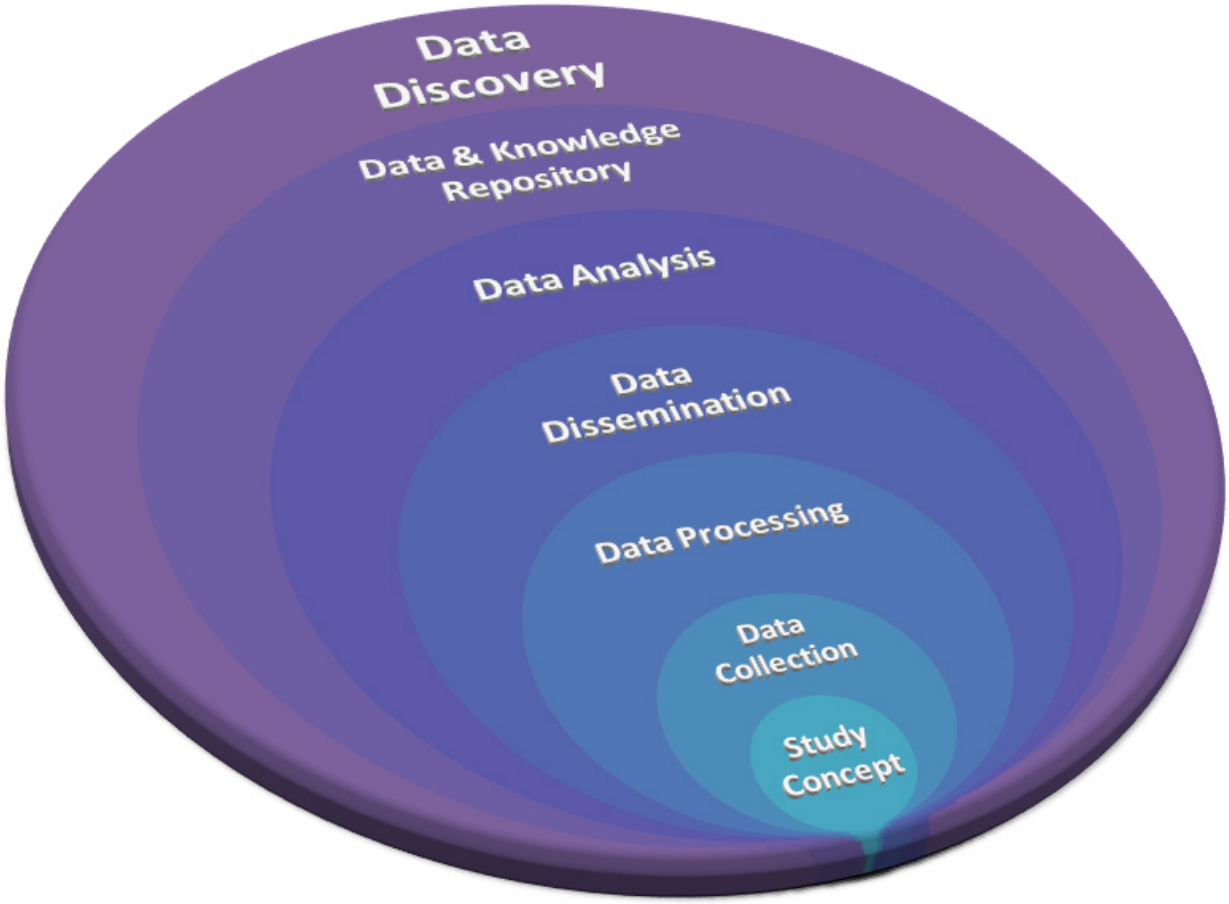
administering a questionnaire to a population whose first language is Spanish were themselves fluent in Spanish.

2.3 Discussion

The original Data Life Cycle Model used to develop DDI 3 represents a data production perspective. The challenge is to determine whether the current model reflects the metadata, data or other information needs of the end user for discovery or access. The End User Model provides a view of the original Model from a user-centered perspective. For example, in the current Model, production is often sequential, whereas end users will select relevant products throughout the entire data life cycle based upon their particular needs (sequential versus random access). Having comprehensive metadata -- in the fullness of description -- is essential for end users to locate what they require successfully.

A user-centered perspective is represented in the end user model below, which shows the nested relationship of the different stages in the data life cycle.

160 Figure 1: the End User Model



161

162

163 The end user perspective should inform the metadata production throughout the data life
164 cycle. The end users may never see actual DDI metadata markup, yet they will be totally
165 reliant upon its existence to accomplish their goals. The workflow producing metadata within
166 each stage needs to take into account that the end user may unpredictably require
167 metadata, data, and other information from any stage. Thus, the “membranes” between life
168 cycle stages must be permeable.

169 Best Practices need to be developed for the metadata produced in each phase of the Life
170 Cycle to meet the discovery and dissemination requirements of the end users. However, the
171 Data Life Cycle phases do not have a one-to-one correspondence with the audiences that
172 produce metadata. Thus, gaps in the metadata may exist.

173 2.4 Example

Two units of the University of Michigan's Institute for Social Research – Survey Research Operations (SRO) and the Inter-university Consortium for Political and Social Research (ICPSR) – worked together to create interactive documentation for the Collaborative Psychiatric Epidemiology Surveys, which were harmonized. This documentation reflects not only what the user needs to know but also how the instrument looked as the interviewers administered it. There are links to interviewer aids and to other pertinent metadata, including the universe of respondents who answered a given question.

http://www.icpsr.umich.edu/CPES/

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 - Archival Ingest and Metadata Enhancement: http://www.ddialliance/bp/DDIBestPractices_Workflows-ArchivalIngestAndMetadataEnhancement.doc.PDF

Stewardship of Research Data in Canada: A Gap Analysis, Draft: July 31, 2008, Research Data Strategy Working Group, http://data-donnees.gc.ca/docs/GapAnalysis.pdf

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

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

224 Wendy Thomas, University of Minnesota

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

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

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

228

229 Appendix B. Revision History

230

Rev Date		By Whom	What
0.9 2009	-02-08	Stefan Kramer	Removed date from filename to accommodate linking. Began revision history tracking.
0.91	2009-02-15	Stefan Kramer	Added rev. date to Subject. Changed Abstract from question to statement. Removed reference to “Intellectual Property Rights section of the DDI Alliance.” Changed sections above ToC to heading 3, rebuilt ToC for heading levels 1-2 only.

231

Appendix C. Legal Notices