



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

### 2 **Subject:**

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

### 4 **Document identifier:**

5 <http://dx.doi.org/10.3886/DDIBestPractices03>

### 9 **Authors:**

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

### 13 **Target Audience:**

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

### 16 **Editors:**

17 Mary Vardigan

### 18 **Abstract:**

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

### 22 **Status: Draft**

23 This document is updated periodically on no particular schedule. Send comments to  
24 editor -- [ddi-bp-editors@icpsr.umich.edu](mailto:ddi-bp-editors@icpsr.umich.edu)



25

26	<b>1 INTRODUCTION.....</b>	<b>3</b>
27	1.1 Problem statement .....	3
28	1.2 Terminology .....	3
29	<b>2 BEST PRACTICE SOLUTION .....</b>	<b>4</b>
30	2.1 Definitions .....	4
31	2.2 Best Practice behavior .....	4
32	2.3 Discussion .....	8
33	2.4 Example .....	8
34	<b>3 REFERENCES .....</b>	<b>9</b>
35	3.1 Normative .....	9
36	<b>APPENDIX A. ACKNOWLEDGMENTS .....</b>	<b>10</b>
37	<b>APPENDIX B. REVISION HISTORY .....</b>	<b>12</b>
38	<b>APPENDIX C. LEGAL NOTICES .....</b>	<b>13</b>
39		



40 **1 Introduction**

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

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

53 **1.1 Problem statement**

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

59 **1.2 Terminology**

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



64

## 65 **2 Best Practice Solution**

### 66 **2.1 Definitions**

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

69 SIP: Submission Information Package

70 AIP: Archival Information Package

71 DIP: Dissemination Information Package

72 METS: Metadata Encoding and Transmission Standard

73 PREMIS: Preservation Metadata Implementation Strategies

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

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

81 DTD: Document Type Definition is one of several [SGML](#) and [XML schema](#) languages, and  
82 is also the term used to describe a document or portion thereof that is authored in the DTD  
83 language.

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

### 86 **2.2 Best Practice behavior**

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

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



## Data Documentation Initiative

- 95 terminology as a Submission Information Package (SIP). A Submission Information  
96 Package (SIP) to be ingested into the archive might typically include:
- 97 • Codebook (ASCII, Word, DDI 1, 2, or 3), ideally with full variable names, variable  
98 labels, and value labels
  - 99 • SPSS portable, SAS transport, Stata data file, or ASCII raw data file with setup  
100 (command/syntax files) files
  - 101 • Questionnaire and show cards
  - 102 • Methodological documentation
  - 103 • Organizational and other bibliographic information
  - 104 • Formulas for calculating variables and weighting instructions
  - 105 • Frequency counts and other univariate statistics
  - 106 • Citations to publications related to the data
  - 107 • Other supporting material
- 108 The SIP may be delivered to the archive in a METS wrapper.
- 109 Depending on the content and format of the digital assets ingested into an archive,  
110 workflows that an archive undertakes to add value to and preserve information will vary.
- 111 The workflows described in this document may differ depending on whether the metadata  
112 are stored in native XML format, in XML-based database platforms such as eXist, or other  
113 database platforms such as Oracle or PostgreSQL. A discussion of metadata storage is an  
114 important consideration but outside the scope of this document.
- 115 Below we consider three different cases related to documentation format:
- 116 (1) The case when an archive receives DDI 3 documentation
  - 117 (2) The case when DDI 1 or 2 documentation is deposited
  - 118 (3) The case when the archive receives non-DDI documentation. In general, we  
119 recommend that conversion to DDI 3 occur at the earliest stage possible in order to  
120 maximize the potential to realize specific DDI benefits for archival processing.



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

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

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

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



## Data Documentation Initiative

- 151           d. Referencing of master questions to country/language versions
- 152           e. Translation of metadata
- 153        7. To ensure compliance with the OAIS standard, add required content related to  
154        preservation (PREMIS)
- 155        8. Create archival metadata record (a subset of DDI 3)
- 156        9. Define any access restrictions for all/part of data
- 157        10. Make decision about what goes into the Archival Information Package (AIP) for long-  
158        term preservation
- 159        11. Create dissemination files from DDI 3 or from DDI-compliant repository or database
- 160        12. Version and publish the Dissemination Information Package (DIP) (including DDI  
161        XML along with the style sheet to render it for presentation). See for instance:  
162        <http://www.ddialliance.org/DDI/related/xml-xslt.html>

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

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

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

173        Repeat steps 1 to 12.

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

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



## Data Documentation Initiative

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

183 Repeat steps 1 to 12.

### 184 **2.3 Discussion**

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

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

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

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

### 194 **2.4 Example**

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





197

198 **3 References**

199

200 DDI Best Practice: Workflows for Metadata Creation Regarding Recoding, Aggregation and  
201 Other Data Processing Activities: <http://dx.doi.org/10.3886/DDIBestPractices04>  
202

203 DDI Best Practice: Workflows - Data Discovery and Dissemination: User Perspective:  
204 <http://dx.doi.org/10.3886/DDIBestPractices02>  
205

206 DDI tools Web page: <http://tools.ddialliance.org/>

207 PREMIS: <http://www.loc.gov/standards/premis/>

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

209 METS: <http://www.loc.gov/standards/mets/>

210 **3.1 Normative**

211

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

214

215 OASIS, Best Practice, <http://www.oasis-open.org/committees/uddi-spec/doc/bp/uddi->  
216 [spec-tc-bp-template.doc](http://www.oasis-open.org/committees/uddi-spec/doc/bp/uddi-spec-tc-bp-template.doc), 2003

217

218 **Appendix A. Acknowledgments**

219 The following individuals were members of the DDI Expert Workshop held 10-14 November  
220 2008 at Schloss Dagstuhl, Leibniz Center for Informatics, in Wadern, Germany.

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

222 Karl Dinkelmann, University of Michigan

223 Michelle Edwards, University of Guelph

224 Janet Eisenhauer, University of Wisconsin

225 Jane Fry, Carleton University

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

227 Arofan Gregory, Open Data Foundation

228 Rob Grim, Tilburg University

229 Pascal Heus, Open Data Foundation

230 Maarten Hoogerwerf, Data Archiving and Networked Services (DANS)

231 Chuck Humphrey, University of Alberta

232 Jeremy Iverson, Algenta Technology

233 Jannik Vestergaard Jensen, Danish Data Archive (DDA)

234 Kirstine Kolsrud, Norwegian Social Science Data Services (NSD)

235 Stefan Kramer, Yale University

236 Jenny Linnerud, Statistics Norway

237 Hans Jørgen Marker, Danish Data Archive (DDA)

238 Ken Miller, United Kingdom Data Archive (UKDA)

239 Meinhard Moschner, GESIS - Leibniz Institute for the Social Sciences

240 Ron Nakao, Stanford University



Data Documentation Initiative

- 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



250

## 251 **Appendix C. Legal Notices**

252 Copyright © DDI Alliance 2009, *All Rights Reserved*

253

254 <http://www.ddialliance.org/>

255

256 Content of this document is licensed under a Creative Commons License:

257 Attribution-Noncommercial-Share Alike 3.0 United States

258

259 This is a human-readable summary of the Legal Code (the full license).

260 <http://creativecommons.org/licenses/by-nc-sa/3.0/us/>

261

262 You are free:

- 263 • to Share - to copy, distribute, display, and perform the work
- 264 • to Remix - to make derivative works

265

266 Under the following conditions:

- 267 • Attribution. You must attribute the work in the manner specified by the author or  
268 licensor (but not in any way that suggests that they endorse you or your use of  
269 the work).
- 270 • Noncommercial. You may not use this work for commercial purposes.
- 271 • Share Alike. If you alter, transform, or build upon this work, you may distribute  
272 the resulting work only under the same or similar license to this one. For any  
273 reuse or distribution, you must make clear to others the license terms of this  
274 work. The best way to do this is with a link to this Web page.
- 275 • Any of the above conditions can be waived if you get permission from the  
276 copyright holder.
- 277 • Apart from the remix rights granted under this license, nothing in this license  
278 impairs or restricts the author's moral rights.

279

### 280 **Disclaimer**

281

282 The Commons Deed is not a license. It is simply a handy reference for understanding the Legal  
283 Code (the full license) — it is a human-readable expression of some of its key terms. Think of it as  
284 the user-friendly interface to the Legal Code beneath. This Deed itself has no legal value, and its  
285 contents do not appear in the actual license.

286

287 Creative Commons is not a law firm and does not provide legal services. Distributing of, displaying  
288 of, or linking to this Commons Deed does not create an attorney-client relationship.

289 Your fair use and other rights are in no way affected by the above.

290

291 **Legal Code:**

292 <http://creativecommons.org/licenses/by-nc-sa/3.0/us/legalcode>