Data Documentation Initiative

1 DDI Working Paper Series -- Best Practices, No. 9

2 Subject:
3   DDI as Content for Registries (2009-01-19)

4 Document identifier:
5   http://dx.doi.org/10.3886/DDIBestPractices09

6
7

8 Authors:
9   Nikos Askitas, Peter Granda, Arofan Gregory, Rob Grim, Pascal Heus, Maarten
10   Hoogerwerf, Jeremy Iverson, Ken Miller, Sigbjorn Revheim, Jannik Vestergaard
11   Jensen

12 Editors:
13   Peter Granda and Arofan Gregory

14 Target Audience:
15   Research and archival communities that want to use DDI to create and share
16   indexed resources with others; decision-makers seeking to enhance research
17   infrastructure for the social sciences

18 Abstract:
19   The authors believe that it is currently premature to recommend a 'best practice' for
20   interacting effectively with registries. Instead, this document defines registries and
21   provides a justification for using them. It begins by giving relevant definitions and
22   discussing the general business case and then goes on to describe from a high-level
23   perspective the use of registries that allow researchers to access material. It
24   concludes by describing examples of researcher use cases.

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

28 **TABLE OF CONTENTS**

42

1 Introduction

44 This document defines registries and provides a justification for using them. It takes a high-
45 level perspective in describing the use of registries that allow researchers to access
46 material. While few such registries yet exist, this approach has several advantages, outlined
47 in this document.

1.1 Problem statement

49 Researchers need to discover information that they can use to answer their research
50 questions. To meet this requirement, data disseminators need to improve the
51 discovery/visibility of their data and metadata. The exposure of DDI metadata can be
52 enhanced through DDI registries that allow, for example, federated searching, as well as the
53 reuse of information from one project to another to enhance comparison and minimize
54 duplication of effort.

1.2 Terminology

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

60

2   Best Practice Solution

2.1 Definitions

63 **Registry**
64 A virtual, centralized and structured database or portal that allows you to list, do a
65 structured search, and to identify and retrieve metadata and possibly data that is distributed
66 around a network. Registries are places where various types of resources are indexed and
67 made visible and available for use throughout a community; they do not include clustered
68 servers or depend on harvesting approaches to access their contents.

69 Some survey organizations register measures (question wording and response options, for
70 example) in order to standardize the way they elicit information from respondents. The
71 implication is that there is one correct way for an organization to measure, say, income.

72 Examples of DDI registries could be question banks, concept banks, social science data
73 survey catalogs, and variable banks.

74 **Register**
75 A collection of elements and attributes that contain information on a particular subject
76 whose authors wish to share with others. Registers require support of well-defined
77 registration processes, and include provisions for dealing with provenance and auditing,
78 versioning, and– security enforcement. Registers are the basic components of registries.

79 **Catalogue**
80 Catalogues contain sets of metadata entities, identifiers, and descriptions of associated
81 items included in a registry. Registries can be thought of as smart catalogues with
82 enhanced functionality which allow for the classification of objects.

83 **Federated**
84 Federated search is the simultaneous search of multiple online databases or Web
85 resources and is an emerging feature of automated, Web-based library and information
86 retrieval systems. It is also often referred to as a portal or a federated search engine.
87 (http://en.wikipedia.org/wiki/Federated_search)

88 **Service-Level Agreement**
89 The agreement between a service provider and service consumer that describes how the
90 service will be provided. Typically a contract.

2.2 Business Case

91
92 There are important reasons for data and metadata disseminators to use DDI registries.

93 • DDI registries will help researchers, data archives, and data producers discover
94   concepts, questions, variables, and data items critical to both data collection and
95   data analyses. Registries can facilitate efficient workflow by allowing these users to
96   study what their predecessors did when investigating similar topics. Effective use of
97   registries will promote good survey design, increase comparability across time and
98   survey populations, and reduce overall costs.

99 • DDI registries will concentrate information, now stored in many different and possibly
100   inaccessible places, in a central location that will increase the visibility of these
101   resources and provide better access tools for all users.

102 • DDI registries will particularly assist new researchers by providing them with basic
103   information on questions and concepts that they can use in creating new surveys.
104   This will get surveys into the field more quickly and improve the comparability of the
105   resulting data.

106 As DDI registries develop, they will support community building among new research areas,
107 encourage interdisciplinary collaborations, and facilitate the sharing of new sources of both
108 data and metadata. To demonstrate how a researcher might interact with a DDI registry, the
109 following illustration guides us through the research process.

110 **2.3 Description of a DDI registry**
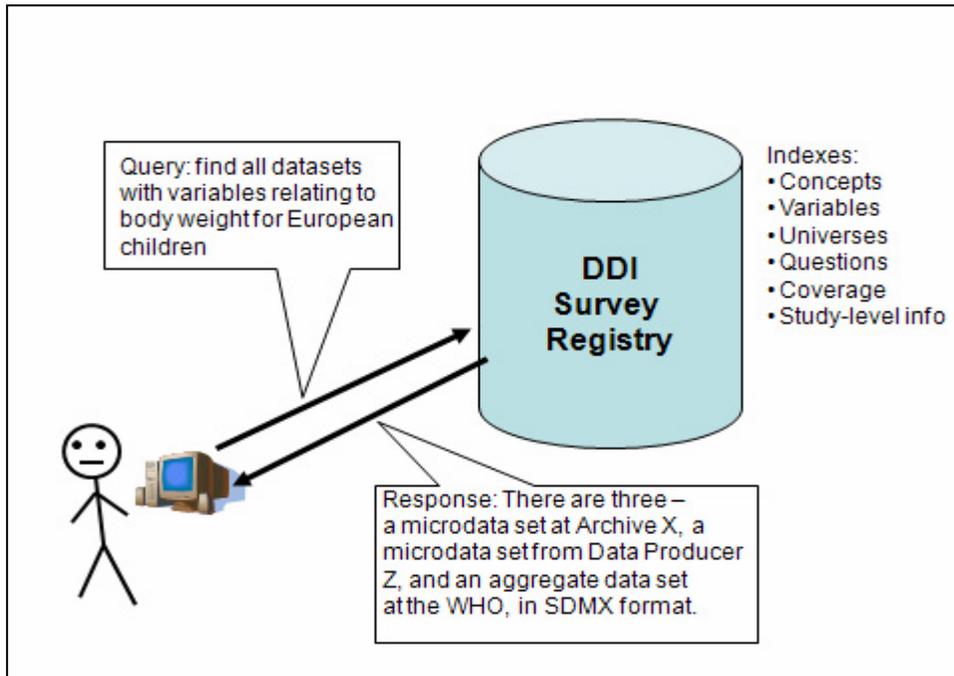
111 **1 Query**

112

113 *Figure 1*

114 A researcher wants to answer a research question about obesity in European children, and
115 needs to locate the relevant data. He uses an application that knows how to query the
116 registry to find relevant variables and their documentation.
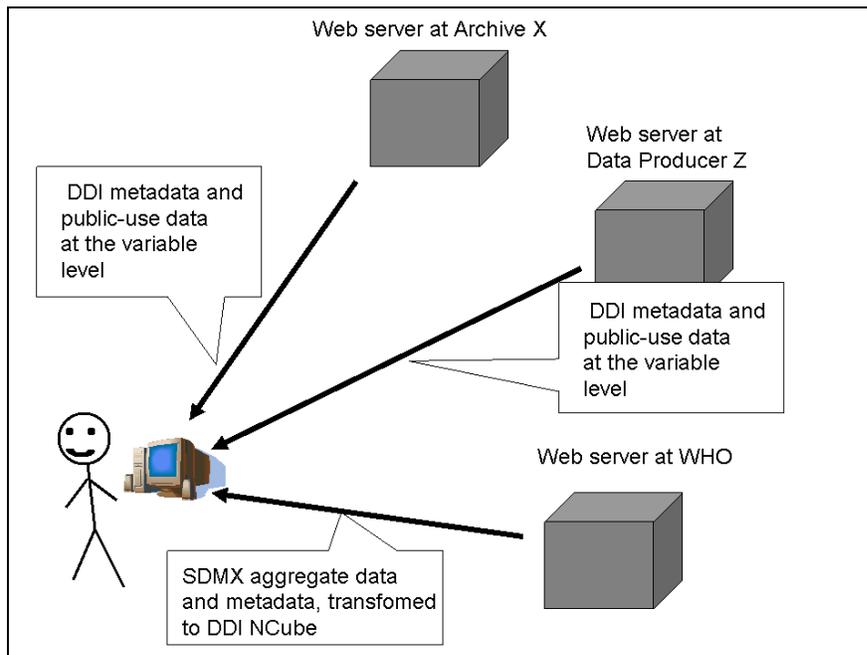
117 **2 Results**

118 *Figure 2*

119 The registry responds to the application with a list of relevant variables based on the
120 researcher's query. The registry is able to do this because it has indexed fields from all of
121 the studies submitted to it:

122 • Study   description
123 • Concept
124 • Coverage
125 • Universe
126 • Question
127 • Variable

128 **3 Retrieval**

129

130   *Figure 3*

131   The researcher browses the list of hits provided by the registry and decides to retrieve a
132   number of variables from the list. He can retrieve the metadata and look at it before
133   obtaining the data, or he can retrieve the entire data set as well. This assumes that he has
134   access to the files (e.g., they are public-use data).

135   The researcher activates the variable links he wants and is taken to the authentication
136   mechanism of the site holding the data. The site may permit automated retrieval of the data
137   and metadata. He downloads the metadata and microdata.
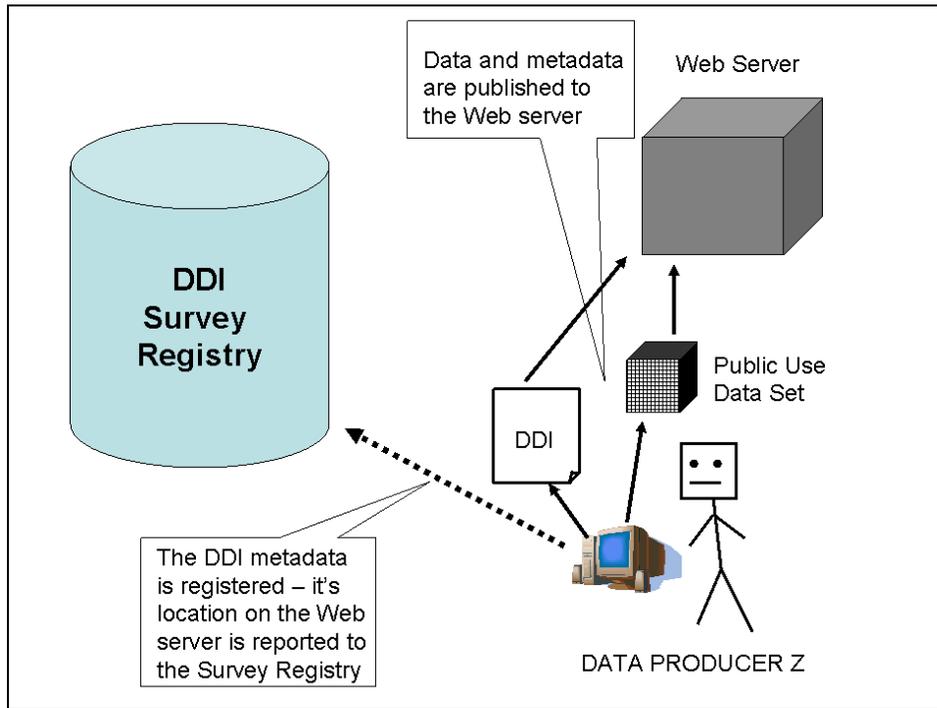
138 **4 Subscription**
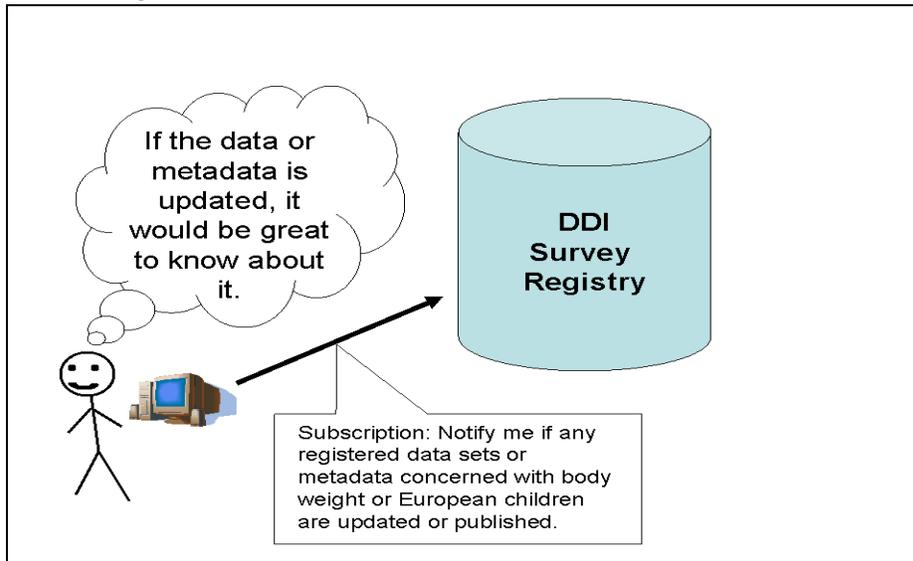
139

140 *Figure 4*

141 The researcher's application allows him to subscribe for a notification if any data or
142 metadata relevant to his research has been published or updated. The application provides
143 its location to the registry so it can receive the notification.

144 **5 Study submission (registration) and metadata indexing**

145

146   *Figure 5*

147

148   A data producer is finished with his collection processes and publishes her metadata and
149   microdata to the IT system in her own organization, which includes making it available on
150   the Web site. The IT system or the data producer also submits a message to the DDI
151   Registry containing information about the location of the metadata for the new or updated
152   material.

153
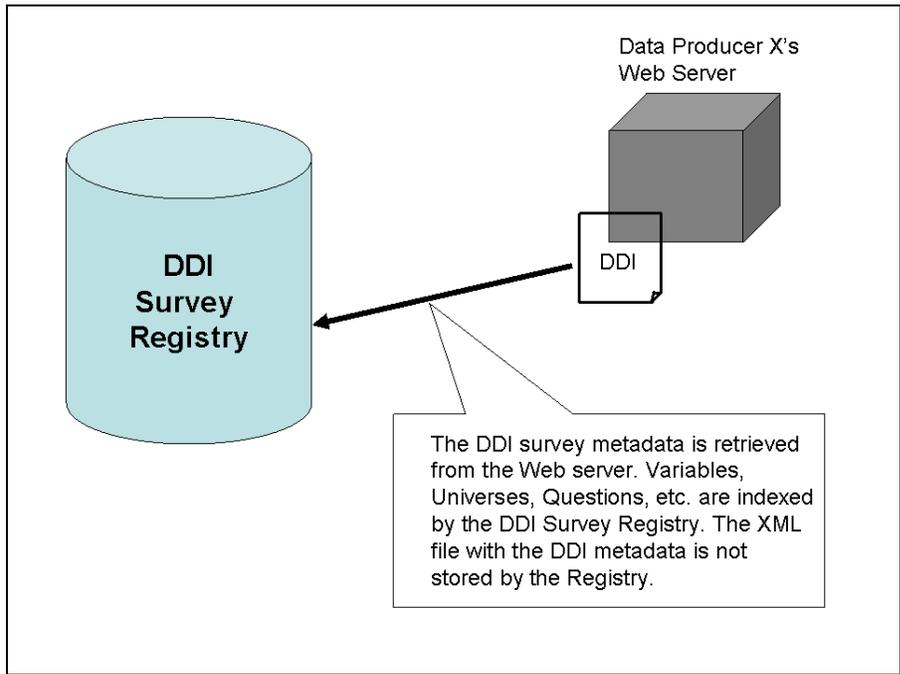
155   *Figure 5.1*

156   The DDI Registry then retrieves the new metadata and indexes it by storing only those
157   fields that will help people query for the data (variable names, labels, and descriptions,
158   universe information, etc.). After the registry has indexed the submitted metadata, it leaves
159   the file on the original Web server – the DDI XML file is not stored at the registry. The
160   registry now contains indexed information about the metadata of the submitted study and a
161   reference to where the actual metadata and its microdata reside. This means that the data
162   and its documentation remain with the organization which published them – these files are
163   always retained and managed by their publishers – there is no second copy maintained
164   separately by the registry. When the data and metadata files are retrieved by users, they
165   know its provenance.

166   **6 Notification**

167

168
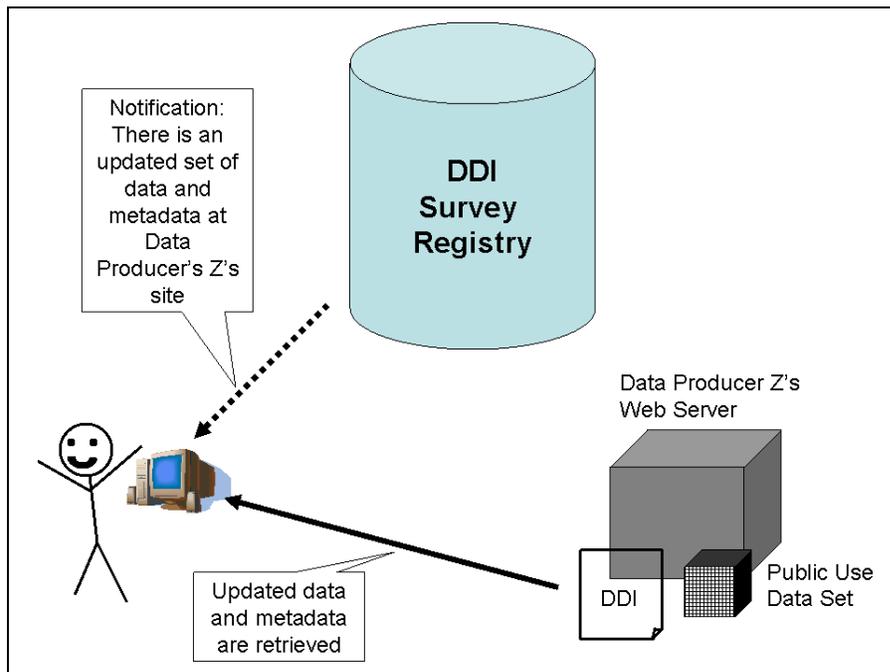
169

170   *Figure 6*

171   Because the researcher has subscribed for notifications when new data and metadata
172   became available relevant to their research, the registry will send out a message
173   automatically to the researcher's application. The application then knows to retrieve the new
174   data and metadata files.

175   Yet another way of considering the importance of registries is to describe the requirements
176   needed to create and share the information they contain.

177   **Requirements**

178   In order for researchers to take full advantage of registries, they should have the following
179   functional and technical requirements:

180   **Functional Requirements**
181   • Registration

182   A registry requires a system that can add and delete users and provides various levels
183   of authentication to interact with the registry.

184   •   Administrative management and maintenance

185   This requires the selection of individual(s) who will oversee the operation of the
186   registry and verify the process of accepting new content and revising existing

| 187 | information. A registry administrator could review and approve submissions and |
| 188 | manage interactions with users as well as provide statistical and error reports. |

| 189 | • Discovery: navigating, listing, searching, classification |

| 190 | The registry should provide an interface that permits easy access to all of its contents |
| 191 | in a manner that is logical and efficient. |

| 192 | • Presentation of results: ordering, grouping |

193	Registries must present results consistent with the items to be searched. For
194	example, the Dublin Core Metadata Initiative (MCDI) suggests: "Users must be able
195	within displays to distinguish whether terms are elements or element refinements or
196	schemes or from controlled vocabularies. Users must be able to browse and search
197	for these categories of terms separately. In addition users must be able to identify
198	relationships between terms e.g. to see that Alternative is a sub-property of Title."
199	(http://dublincore.org/groups/registry/fun_req_ph1-20011031.shtml).

| 200 | • Identification: Use DDI URNs |

| 201 | • Maintainability: add, update, versioning, workflow |

202	If the registry is responsible for maintaining certain content it must have procedures to
203	determine when there is a semantic change in a term that might require a change in
204	its URI. If the registry is merely storing such structures as externally managed
205	vocabularies, it should adopt a policy in regard to versioning since the maintenance
206	agency processeswill be independent of the Registry.

| 207 | • Content quality: quality control, coherence |

208	A number of existing data quality frameworks such as one created by Eurostat
209	emphasize the importance of well-structured content, not only for retrieval purposes,
210	but also to maximize data availability online. See:
211	http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details
212	/publication?p_product_code=2010-S-21-028868

| 213 | • Ownership: encourage users to publish DDI materials, intellectual property |

214	The registry should encourage metadata producers to share their content by providing
215	easy access to the registry and clear standards to describe how these producers can
216	deposit and update their materials.

| 217 | • Up-to-dateness: currency, subscription, timeliness, notification |

218   It is important for registries to notify users when content changes. Elements that might
219   be included in such notifications (derived from the National Science Data Library
220   [NSDL]) are:

221   o   Date/time of notification

222   o   Date/time of notified activity

223   o   Name/email of agent initiating activity

224   o   Contact for questions

225   o   Basic explanatory text about the registry

226   o   Link to help text (specific section based on notification type)

227   • Long-term   support/maintenance/sustainability

228   o   Service Level Agreements (SLAs)

229   o   Issues of trust

230

231 Technical Requirements
232   • Interface/Protocols

233   The registry should provide an interface and accompanying tools that permit
234   authorized users to upload, create, maintain, and remove resource descriptions from
235   the registry.

236   The underlying architecture of the registry should incorporate a flexible database
237   structure that precisely defines the linkages between different terms and displays
238   them to the user. The database should have equally robust storage, production, and
239   publication features. Registry architects need to consider how much content the
240   registry might contain and what demands users will place on it. Careful
241   consideration of the following features is necessary when building a registry:

242   o Identifiers

243   o Index-specif ic elements

244   o Performanc e

245   o Scalability

246   o   Relationships among registry contents

247   o Subscription -notification (e.g., RSS)

248   **Important DDI elements that need to be indexed**

249   Depending on the type of registry and its content, registry creators and contributors should
250   consider including the following DDI elements, which have particular significance for sharing
251   information between data archives:

252   • Study-level information (Title, Purpose, Abstract)

253   • Univ   erses

254   • Coverage

255   • Concepts

256   • Variables

257   • Questions

258   • Geographies

259   • Codes and categories (classifications)

260   • NCube   structures

261   • Comparison   maps

262   • Groups

263

264   ## 2.4 Examples
265   The References section provides examples of existing metadata registries. The examples
266   address the following properties:

267   • Visibility/navigation/cla   ssification

268   • Location/access   information

269   • Management of resources and submission

270   • Access   control

271 • Subscription   /notification of change

3   References

273

274 Metadata registries:

275 *http://www.datafoundations.com/solutions/data_registries.shtml*

276 Specific metadata registries:

277 http://en.wikipedia.org/wiki/Metadata_registry#Examples_of_public_metadata_registries

278 Good examples of these:

279 • Cancer Data Standards Repository --
280   http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr

281
282

283 • US Department of Defense Metadata Registry (requires sponsored registration)

284 • US Environmental Protection Agency - Environmental Data Registry --
285   https://xmdr.org/content_survey/results/edr_survey.html

286 NSDL examples (Vocabularies):

287 http://wiki.metadataregistry.org/NSDL_Registry_Use_Case_Documentation_(Vocabularies)

288 Definition from DLIB Magazine about a metadata registry for the Semantic Web:

289 http://www.dlib.org/dlib/may02/wagner/05wagner.html

290 "…where machines and humans can exchange electronic content that has clear-cut,
291 unambiguous meaning. This vision is based on the automated sharing of metadata terms
292 across Web applications. The declaration of schemas in metadata registries advances this
293 vision by providing a common approach for the discovery, understanding, and exchange of
294 semantics. However, many of the issues regarding registries are not clear, and ideas vary
295 regarding their scope and purpose. Additionally, registry issues are often difficult to describe
296 and comprehend without a working example."

297 Protocol for Object Exchange and Reuse from the Open Archives Initiative. OAI-ORE
298 facilitate access to collections of resource types.

299   http://www.openarchives.org/ore/

300   UDDI (may no longer be active)

301   http://en.wikipedia.org/wiki/Universal_Description_Discovery_and_Integration

302   ISO11179 manages terminology

303   http://en.wikipedia.org/wiki/ISO/IEC_11179

304   ISO 15000 part 3 and 4 / EBXML

305   http://en.wikipedia.org/wiki/E-Business_XML

306   GRID, e.g. GRIMOIRES, a Grid RegIstry with Metadata Oriented Interface: Robustness,
307   Efficiency, Security

308   http://twiki.grimoires.org/bin/view/Grimoires/

309   **3.1 Normative**

310

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

313

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

316

Appendix A. Acknowledgments

318 The following individuals were members of the DDI Expert Workshop held 10-14 November
319 2008 at Schloss Dagstuhl, Leibniz Center for Informatics, in Wadern, Germany.

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

321 Karl Dinkelmann, University of Michigan

322 Michelle Edwards, University of Guelph

323 Janet Eisenhauer, University of Wisconsin

324 Jane Fry, Carleton University

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

326 Arofan Gregory, Open Data Foundation

327 Rob Grim, Tilburg University

328 Pascal Heus, Open Data Foundation

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

330 Chuck Humphrey, University of Alberta

331 Jeremy Iverson, Algenta Technology

332 Jannik Vestergaard Jensen, Danish Data Archive (DDA)

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

334 Stefan Kramer, Yale University

335 Jenny Linnerud, Statistics Norway

336 Hans Jørgen Marker, Danish Data Archive (DDA)

337 Ken Miller, United Kingdom Data Archive (UKDA)

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

339 Ron Nakao, Stanford University

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

341   Wendy Thomas, University of Minnesota

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

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

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

345

Appendix B. Revision History

347

Rev	Date	By Whom	What
0.9	2009-04-05	Stefan Kramer	Began revision history tracking. Minor formatting changes.

348

349

350   **Appendix C. Legal Notices**

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

352

353   http://www.ddialliance.org/

354

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

356   Attribution-Noncommercial-Share Alike 3.0 United States

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

358   http://creativecommons.org/licenses/by-nc-sa/3.0/us/

359

360   You are free:

361   to Share - to copy, distribute, display, and perform the work

362   to Remix - to make derivative works

363

364   Under the following conditions:

365

366   •   Attribution. You must attribute the work in the manner specified by the author or

367   licensor (but not in any way that suggests that they endorse you or your use of

368   the work).

369   •   Noncommercial. You may not use this work for commercial purposes.

370   •   Share Alike. If you alter, transform, or build upon this work, you may distribute

371   the resulting work only under the same or similar license to this one.

372   •   For any reuse or distribution, you must make clear to others the license terms of

373   this work. The best way to do this is with a link to this web page.

374   •   Any of the above conditions can be waived if you get permission from the

375   copyright holder.

376   •   Apart from the remix rights granted under this license, nothing in this license

377   impairs or restricts the author's moral rights.

378

379   Disclaimer

380   The Commons Deed is not a license. It is simply a handy reference for understanding the Legal

381   Code (the full license) — it is a human-readable expression of some of its key terms. Think of it as

382   the user-friendly interface to the Legal Code beneath. This Deed itself has no legal value, and its

383   contents do not appear in the actual license.

384

385   Creative Commons is not a law firm and does not provide legal services. Distributing of, displaying

386   of, or linking to this Commons Deed does not create an attorney-client relationship.

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

388

389   Legal Code:

390   http://creativecommons.org/licenses/by-nc-sa/3.0/us/legalcode

391