Data Documentation Initiative

DDI Working Paper Series -- Best Practices, No. 5

Subject

Controlled Vocabularies (2009-02-22)

Document identifier:

http://dx.doi.org/10.3886/DDIBestPractices05

Authors:

Peter Granda, Stefan Kramer, Jenny Linnerud, Hans Jørgen Marker, Ken Miller, Mary Vardigan

Editors:

Ken Miller

Target audience:

Communities that want to select, develop, or extend controlled vocabularies for use with DDI.

Abstract:

The benefits of using controlled vocabularies within DDI metadata creation are the primary underlying principles behind this Best Practice document.

Use of a controlled vocabulary enhances consistency and efficiency in the production of DDI metadata, gives precision in searching the DDI metadata, and allows semantic and technical interoperability between organisations creating DDI instances.

Status:

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

26

27 # Table of Contents

28

42

1 Introduction

1.1 Problem statement

Dependence on natural language alone in indexing and searching brings with it a number of problems that use of a controlled vocabulary (CV) can address. For a fuller explanation of the problems a CV addresses, see the Discussion section.

This Best Practice does not address creating CVs themselves. Within the DDI community, the DDI Alliance Controlled Vocabularies Working Group is focusing on this task.

This best practice addresses:

- **Technical aspects:** CV publication format; application requirements for handling hierarchies, if present in a CV, for indexing and searching; separation of index terms and structural description in a CV; and management of the CV's content.
- **Business practices:** Selection of a platform for publishing CVs; decision about term structure in CVs; relationship identification in CVs; versioning and cross-referencing of terms in CVs; policies governing CV maintenance; mapping between different CVs; documentation of CVs; and CVs for system use.

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

Genericode: Genericode defines a standard format for defining code lists (also known as enumerations or controlled vocabularies).

Genericode aims to provide the following:

- A standard model and XML representation for the contents of a code list
- A standard model and XML representation for data associated with items in a code list

- A standard model and XML representation for how new code lists are derived from existing code lists

Community: In this document, the term *community* is used to identify any grouping of personal or organizational entities, at different levels of formal organization, that are considering or undertaking implementation of DDI. Examples: a national statistical service, a data producer, an archive, a consortium of data archives.

Controlled vocabulary (CV): Broadly speaking, a CV can range from a short list of clearly defined, mutually exclusive, and exhaustive terms, which are the only choices for usage in a specific context (e.g., populating certain DDI elements or attributes) through a classification to something as complex as a thesaurus with thousands of terms and term relationships. A CV has also been described as "A set of subject terms, and rules for their use in assigning terms to materials for indexing and retrieval." (http://www.cs.cornell.edu/wya/diglib/MS1999/Glossary.html)

In a CV, a term consists of one or more words used to represent a concept (example: "fear"; "females"; "child care"). Terms are selected from natural language for inclusion in a controlled vocabulary.

Pre-coordinated vs. Post-coordinated systems in CVs:

- In pre-coordinated CV systems, multiple concepts are brought together in one term. An illustrative example is the Library of Congress Subject Headings (LCSH), which yield entries such as: "Insurance, Unemployment --Switzerland --Statistics." This method allows for disambiguation of the relationship of the concepts in the term that might not be possible in post-coordinated systems, such as whether a term is a qualifier of another.

- In post-coordinated or faceted systems, concepts are kept broad and separate and selected and joined in the process of searching with Boolean operators. A representation of the above LCSH in this system could be "Insurance AND Statistics AND Switzerland AND Unemployment" – note that entry order in the query has no relevance here. An example of such a system is the American Psychological Association's *Thesaurus of Psychological Index Terms*.

2.2 Best Practice behavior

Selection or Creation

- Research whether a CV already exists that maps well with the knowledge domain to be covered by your activity, which could mean taking a subset of that existing CV.

- Research whether the desired CV has already been published in a format suitable to local application, such as Genericode. (Recommend to creators of CVs to publish the machine-actionable part of the CV in Genericode.)
- In selecting an existing CV for use, choose one that is as large and complex as necessary, but as small and simple as possible, to help optimize both indexing and searching.
- Multiple languages should be considered in the initial creation or selection of a CV.
- If selecting or creating a complex CV, decide upfront whether it is or will be pre- or post-coordinated. Post-coordination will allow for more flexibility and lend itself to Boolean searching by end users.
- If you are merging two or more complex CVs, consider the implications if some of them are pre- and some are post-coordinated.
- In making a CV publicly available, consider releasing it with the least possible restrictions (e.g., Creative Commons or Public Domain).
- Multiple communities may join to create a CV derived from existing ones.
- If a new CV is created, the reason for this activity should be explained as part of its documentation (see, for instance, ANSI/NISO Z39.19-2005 section 9.5 [see References section]). If it is made publicly available, its creation should be announced to the DDI community and any other relevant user communities.
- If terms are likely to be misunderstood by the CV's users, more detailed scope notes should be provided for them.
- Handling of phrases: multi-word terms (e.g., "child care" or "traffic regulation") need to stay bound as a phrase for handling by applications.
- If your CV contains descriptions of structure, these should be clearly distinguished from terms to insure accuracy in indexing and processing. Descriptions of hierarchies (e.g., "Age Groups" or "Vehicles by size") and other text that describes the arrangement of the CV, but does not constitute CV terms, need to be excluded from being able to be assigned as CV terms.
- The storage container for the CV should allow for: persistent identifiers, version control and dating. This facilitates machine-actionability and human understanding of the CV.

- Consider use of CVs for system-to-system coordination, such as consistent access control to shared resources among different communities.

Maintenance, Use of, and Additions to CVs

- Define what requires additions or revisions to the CV, such as: new organizations contributing to or being affected by the CV; new data acquisitions; new services or user communities; changes in the reality the CV describes; changes in language usage or jargon; etc.

- If the community of CV users broadens over time, it may become necessary to translate the CV into other languages.

- Whether using one or more complex CVs, the relationships between terms (e.g., "see" or "related terms" or "narrower term") have to be defined and represented consistently as the CV evolves.

- Deprecated terms in more complex CVs need to be identified as such, and refer to their "use instead" term.

- Assign only the most specific CV term if a hierarchy of terms is exposed, so broader terms can automatically be included in a query.

- Multiple communities that have agreed to create a CV derived from existing CVs should create and maintain a mapping between the existing and derived CVs.

2.3 Discussion

Use of a controlled vocabulary in indexing, documenting, or searching for resources by use of their assigned metadata will enhance: precision; consistency; temporal, spatial and topical comparability; semantic and technical interoperability; multilingual access; efficiency; and harmonization. These benefits of a controlled vocabulary are the primary underlying principles behind this Best Practice document.

Legal considerations including applicable laws regarding copyright for using large parts of, or entire, existing controlled vocabularies, should be taken into account. Detailed discussion of legal issues is outside the scope of this Best Practice.

The need for any crosswalks between CVs that cover different domains of knowledge should be considered, but is outside of the scope of this Best Practice.

2.4 Example

There are CVs embedded in the DDI standard already. These are short lists of clearly defined, mutually exclusive, and exhaustive terms, which have vocabularies which are considered fixed. Examples of these are:

ValueTypeCodeType: Indicates value type.

- *Restricts:* xs:NMTOKEN
- (pattern is: code, colon, term, dash, definition)
    - Code: GreaterThan - Greater Than
    - Code: LessThan - Less Than
    - Code: Equal - Equal
    - Code: GreaterThanOrEqual - Greater Than or Equal
    - Code: LessThanOrEqual - Less Than or Equal
    - Code: NotEqual - Not Equal

AggregationMethodCodeType: A list for describing aggregation methods.

- *Restricts:* xs:NMTOKEN
- (pattern is: code, colon, term, dash, definition)
    - Code: Sum - Sum
    - Code: Average - Average
    - Code: Count - Count
    - Code: Mode - Mode
    - Code: Median - Median
    - Code: Maximum - Maximum
    - Code: Minimum - Minimum
    - Code: Percent - Percent (Percentages are used to express how large one quantity is relative to another quantity)
    - Code: CumulativePercent - Cumulative percent (percentage of items in its frequency distribution which are equal to or lower than the current item [maximum value is 100%])
    - Code: PercentileRank - Percentile rank (The percentile rank of a item is the percentage of items in its frequency distribution which are lower [cannot reach 100%])

The DDI Controlled Vocabulary Group is looking at other areas of the DDI in which a CV would enhance computer actionability, but the list of terms is not exhaustive, and the terminology might change over time. These would be published in Genericode.

Examples of this kind of CV are:

Suggested CV for Collection Mode (Data Collection Module)

- Interview
  - Face-to-face
  - Telephone
  - E-mail
  - CATI
  - CAPI
  - Mixed-mode
- Self-completed questionnaire:
  - Paper/pencil
  - Web-based
  - CASI
  - ACASI
- Coding
- Transcription
- Compilation
- Synthesis
- Recording
- Simulation
- Observation
- Experiments
- Focus Group
- Other

Suggested CV for attribute “Role”:

- Data Collector
- Data producer
- Depositor
- Distributor
- Copyright holder
- Research Investigator
- Other

Within the Council of European Social Science Data Archives (CESSDA) further CVs are being considered to enhance searching and additional proposed harmonization and comparability services that would be available via the CESSDA portal.

Examples here are:

(Pattern is: hierarchy level, dashes indicating level, term)

01 - - SAMPLING PROCEDURES

241 02 - - - - COMPLETE COUNT

242 02 - - - - PROBABILITY SAMPLE

243 03 - - - - - - SIMPLE RANDOM SAMPLE

244 03 - - - - - - SYSTEMATIC SAMPLE

245 03 - - - - - - STRATIFIED SAMPLE

246 03 - - - - - - CLUSTER SAMPLE

247 03 - - - - - - MULTIPHASE SAMPLE

248 03 - - - - - - TIME SAMPLE

249 02 - - - - NONPROBABILITY SAMPLE

250 03 - - - - - - PURPOSIVE SAMPLE

251 04 - - - - - - - - QUOTA SAMPLE

252 04 - - - - - - - - RANDOM WALK SAMPLE

253 03 - - - - - - ACCIDENTAL SAMPLE

254 04 - - - - - - - - VOLUNTEER SAMPLE

255 04 - - - - - - - - CONVENIENCE SAMPLE

256 01 - - DATA FORMAT

257 02 - - - - TEXTUAL DATA

258 02 - - - - NUMERIC DATA

259 02 - - - - ALPHA/NUMERIC DATA

260 02 - - - - IMAGE DATA

261

262 01 - - DATA STRUCTURE

263 02 - - - - RECTANGULAR DATA

264 02 - - - - RELATIONAL DATA

02 - - - - HIERARCHICAL DATA

Example of the development of a community-based thesaurus:

CESSDA has already constructed a CV to act as a common topical classification scheme to describe the overall coverage of the resources held at individual member archives.

The process involved consideration of existing published classification schemes, such as the Library of Congress Subject Headings. Several CESSDA member archives had created CVs themselves and through comparing these it was agreed that a new common CV would not be that difficult to create.

After the common structure had been agreed each member archive had the choice to either adopt the new CV or map their existing one to it. The terminology of about 140 terms was then translated into the languages of the different member archives.

When it came to the adoption of a common thesaurus for allocation of keywords and concepts CESSDA chose the existing UKDA HASSET thesaurus. This was reduced and restructured slightly to remove UK specific concepts and align it more to a European resource. The terms (over 4,000) have now been translated into 9 of the 21 languages of the present CESSDA membership.

3 References

Genericode http://www.genericode.org/

DDI Technical Specification, April 2008, part I, section 4.3: *Controlled Vocabularies*. Available via: http://sourceforge.net/projects/ddi-alliance/files/Data%20Documentation%20Initiative/DDI%203.0%20%282008-04-28%29/DDI_3_0_2008-04-28_Documentation_XMLSchema.zip/download

Best Practice on Versioning and Publication: http://dx.doi.org/10.3886/DDIBestPractices08

DDI Alliance Controlled Vocabularies Working Group: http://www.ddialliance.org/alliance/working-groups

Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies: http://www.niso.org/standards/resources/Z39-19-2005.pdf

ANSI/NISO Z39.19-2005: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies: http://www.techstreet.com/cgi-bin/detail?product_id=1262086

ISO 5964: Guidelines for the establishment and development of multilingual thesauri: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=12159

Creative Commons: http://creativecommons.org/

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

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

330 Wendy Thomas, University of Minnesota

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

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

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

334

335 Appendix B. Revision History

336

Rev	Date	By Whom	What
0.1	14/11/2008	Stefan Kramer	Final version from Dagstuhl
0.2	30/11/2008	Ken Miller	Abstract added and minor text and format changes.
0.3	2009-02-15	Stefan Kramer	Formatting changes (standard across all BPs) and reference link additions.

337

Appendix C. Legal Notices