Controlled Vocabularies for DDI 3: Enhancing Machine-Actionability

by
Taina Jääskeläinen, Meinhard Moschner and Joachim Wackerow[1]

Introduction

Controlled vocabularies (CVs) are organized lists of terms used for metadata and information retrieval. They may be short (flat) lists or more complex constructs, containing hierarchical relationships. Subject thesauri, such as LCSH (Library of Congress Subject Headings) or the multilingual ELSST thesaurus used by European data archives, are examples of the more complex constructs, containing broader terms, narrower terms, related terms, synonyms, and scope notes.

Ideally, the terms in a controlled vocabulary should be exhaustive (covering the whole dimension of the issue), mutually exclusive (no overlaps between terms) and clearly defined (definitions/scope notes given for the meanings of the terms). CVs are often used in specific contexts, and definitions/scope notes clarify and disambiguate the meaning of a term in a particular context as it may differ from the meaning in natural language.

Controlled vocabularies play a critical role in metadata standards, which have two basic components: 1) semantics – definition of the meaning of metadata elements, and 2) content – declaration of instructions for what and how values should be assigned to elements (Chan and Zeng, 2006). Controlled vocabularies belong to the domain of content as they specify the values allowed in an element or attribute.

An extensive set of controlled vocabularies is now being developed for the Data Documentation Initiative (DDI) metadata standard, to be used to describe specific aspects of a dataset across the data life cycle. This paper discusses the advantages of and reasons for using controlled vocabularies; the history of the effort to create controlled vocabularies for DDI; the work of the DDI Controlled Vocabularies Group (CVG) in developing new vocabularies; and a new standards-based system for managing CVs in DDI, as well as other future developments.

Advantages of Controlled Vocabularies

Using controlled vocabularies has several advantages for DDI, many of them related to overcoming difficulties caused by natural language in documentation and information retrieval.

Control of synonyms. In natural language, there are synonyms that use different terms to refer to the same entity. In a controlled vocabulary, synonyms are no longer a source of concern because the vocabulary defines the preferred term (Chu, 2007).

Control of lexical anomalies. CVs control lexical anomalies by minimizing any superfluous vocabulary or grammatical variations that could potentially create noise in the users' results set (Chamis, 1991; Garshol, 2004), e.g., removing leading articles, prepositions, conjunctions, etc., or ensuring consistency (Macgregor & McCulloch, 2006).

For example, when describing data collection methods, a DDI vocabulary will tell us whether to use 'self-completed questionnaire' or 'self-administered questionnaire'.
Even in the case of simple issues such as describing the (same) planned frequency of data collection, there may be surprisingly many variations: 'twice every year', 'biannually', 'every 6 months', 'every six months', 'two times a year'. Add to that all the possible variations in different languages, and it becomes clear how difficult it often is to produce comparability and how vocabularies can enhance semantic interoperability between organizations and systems.

Promotion of consistency and efficiency. Controlled vocabularies enhance consistency and high-quality metadata not only by providing a single form of the term to be used but also by promoting more consistent use of DDI elements themselves. A CV is a clear indication of the intended content of an element. We must also factor in improved efficiency: persons providing metadata tend to change over time and vocabularies lessen the burden of learning. Vocabularies also make metadata production quicker – less time will be spent on trying to figure out the meaning of elements and how this or that entity should be described. Not surprisingly, new staff members tend to be fond of vocabularies.

Clearly defined terminology. Definitions/scope notes for terms provided in the DDI vocabularies are another way of improving consistency, comparability and efficiency. The definitions explicate the meaning of a term in the given context, clarifying the difference between, say, a proxy and an informant as a response unit. The Controlled Vocabularies Group members have found that meanings of terms are rarely so clear they seem at first glance. Our experience was that more often than not, discussion of a particular vocabulary had to be postponed until we had time to find term definitions from methodology handbooks or other relevant sources in order to know what we were talking about. Therefore, we fully expect that the definitions provided for the terms in DDI vocabularies will be an advantage both to metadata production and information retrieval.

The Controlled Vocabularies Group also learned during the process that institution-level vocabularies currently used in their own organizations leave a lot to be desired. We found, quite often, that they contain overlapping terms, lack some necessary terms and are too geared toward describing the collections of a particular archive. This makes us confident that the vocabularies suggested for DDI 3 elements will be an improvement to many institution-level vocabularies.

Promotion of interoperability. It is becoming generally accepted in the information community that interoperability is one of the most important principles in metadata implementation. Using the same controlled vocabularies for metadata in different collections enables cross-collection searching. If different vocabularies are used, interoperability can be provided by mappings. Interoperability at the repository level - with harvested or integrated records from varying sources - can be enabled by mapping value strings associated with particular elements (Chan and Zeng, 2006).

If a data organization feels obliged to continue to use an institutional-level and context-specific vocabulary instead of the one recommended by the DDI standard, the organization should provide a terminology mapping to the DDI vocabulary. Terminology mappings are intellectually created crosswalks from the terms in one vocabulary to the terms in another, providing a network of equivalent, broader, narrower and related term relationships (Mayr and Petras, 2008).

Support for machine-actionability. The structure of CVs also facilitates the use of codes or notation which can then be associated with terms. Such notation is mnemonic, predictable, and language-independent (Broughton, 2004). In fact, one of the most important advantages of controlled vocabularies is that they facilitate the production

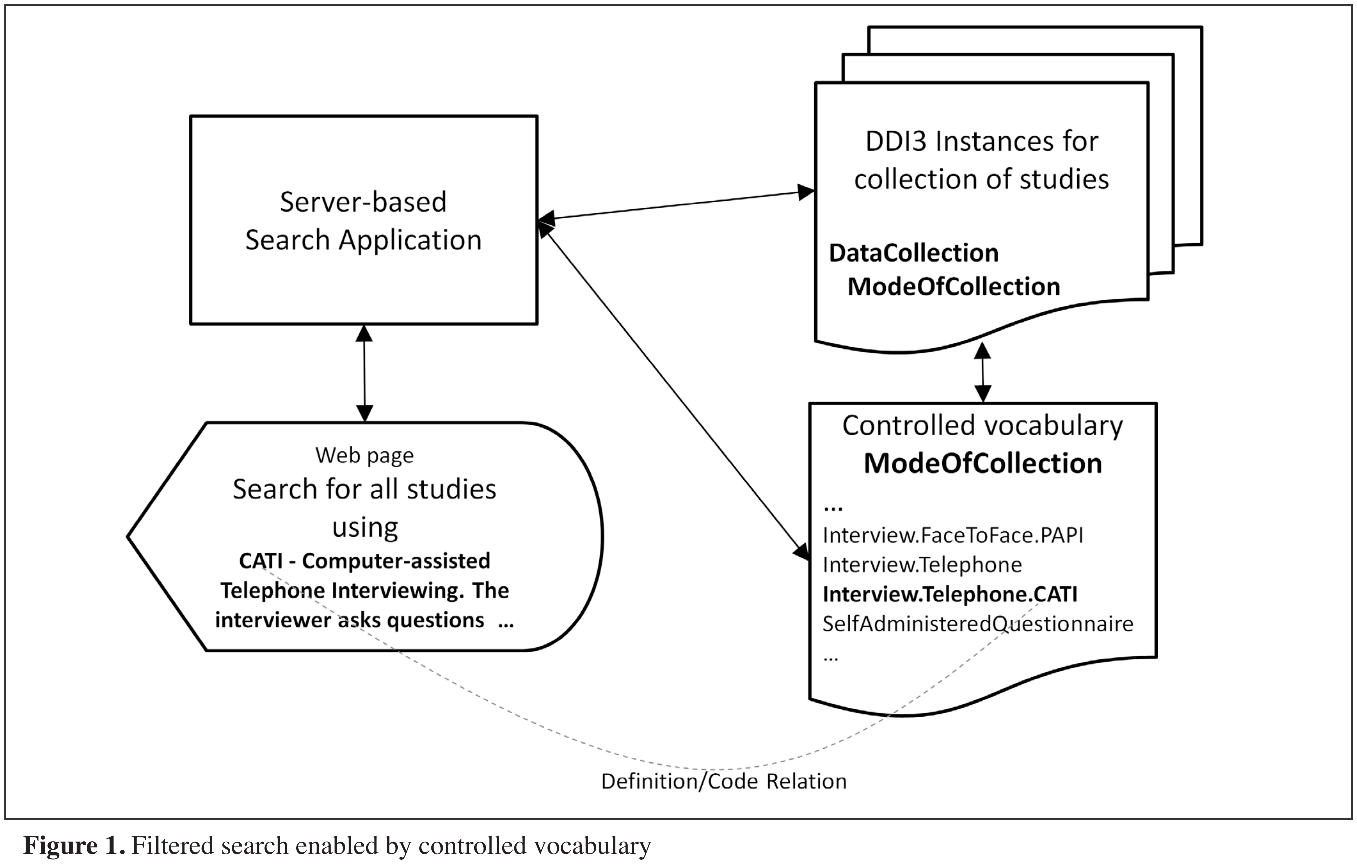
Server-based Search Application
DDI3 Instances for collection of studies
DataCollection
ModeOfCollection
Web page
Search for all studies using
CATI - Computer-assisted Telephone Interviewing. The interviewer asks questions ...
Controlled vocabulary
ModeOfCollection
...
Interview.FaceToFace.PAPI
Interview.Telephone
Interview.Telephone.CATI
SelfAdministeredQuestionnaire
...
Definition/Code Relation

Figure 1. Filtered search enabled by controlled vocabulary

of metadata that are not only machine-readable, but also machine-interpretable (DDI 2) and machine-actionable (DDI 3). CVs do not usually replace machine-readable textual descriptions used to give more in-depth information, but since DDI controlled vocabulary terms have codes, they provide precise and unambiguous means for controlling software processes, as in Figure 1 below. In this example, a search application looks for all studies conducted by CATI. The user sees the human-readable text, which is the definition of the code Interview.Telephone.CATI from the controlled vocabulary ModeOfCollection, and then retrieves a list of studies to browse.

If statistical measures, for instance, are to be extracted and compared in an automated way, resulting in a time series chart or a geographical representation, it has to be guaranteed that identical statistical measures are being compared. In addition, one needs to control for universe, analysis unit, etc. If the matching of datasets is managed by an application, the terms used to describe dataset formats or character sets have to come from a controlled vocabulary.

Another example of a CV with the potential to control software processes is the ISO 3166 alpha-2 country code provided by DDI 3 for identifying countries. The country of a data provider or user may be relevant for access conditions, and an application can use the country codes to control access. Similarly, the ISO 639-1 alpha-2 standard coding for the most common languages also facilitates machine-actionability. If questionnaires are marked up using the language CV, their selective and separate indexing by retrieval tools can be controlled by software. Software can also determine which language versions of metadata are available for which elements in a database (or DDI instances) in order to control text retrieval processes and to offer transparency to end users in multi-language applications.

Crossing language barriers in documentation and information retrieval. Controlled vocabularies form an important class of language tools. They can be used to assist both in manual and automatic translation of metadata (Svenonius, 2003). If an editing tool has implemented controlled vocabularies, the tool may be designed to produce automatic translation of some metadata elements from one language to another. This, of course, only applies to the elements or attributes that have CVs.

Controlled vocabularies can be used to support behind-the-scenes query expansion across languages. In addition to element-specified search, they can also be used as tools in free-text search. In fact, multilingual subject thesauri are useful for crossing language barriers even in cases where they have not been used for indexing data. If implemented as behind-the-scene search tools, they enable the user to discover data in different languages when the query term he/she has used is both 1) a thesaurus term and 2) appears in the text of an abstract or in question wording.

More precision and recall in information retrieval. Controlled vocabularies presuppose less previous knowledge about the content of a resource or a repository, or even a virtual collection of repositories. They may help to bridge the initial gap between the user and the resources he or she needs by focusing the request. An information seeker is more likely to achieve high recall (fraction of the relevant documents that are successfully retrieved) if all entities of the same kind are named in the same way. The seeker would achieve lower precision (fraction of the retrieved documents that are relevant) if terms have multiple meanings in different contexts. There is even more “added value” if the combination of allowed values in relevant elements provides a pre-selection mechanism for potentially comparable results, for example, data resulting from measuring the same concept at a certain point in time and space, for a comparable universe and using a certain methodology. The challenge is to build a system for “bringing like things together and differentiating among them” (Svenonius, 2000).

Search tools can display thesauri and other controlled vocabularies to allow users to improve their query formulation. Data portals may display broader, narrower and related terms of the term the user has used in his/her search. This will enable users to see which concepts/terms are related to their areas of interest and maybe give them ideas of other potentially relevant query terms to use.

DDI Controlled Vocabularies Project

The implementation of controlled vocabularies has been an ongoing topic of discussion across the history of the DDI standard. Early versions of the standard, DDI 1 and 2, were expressed in the form of a Document Type Definition (DTD) that contained some controlled vocabularies within it. This was problematic, however, because to change the embedded CVs, the entire specification had to be reissued – clearly not an ideal situation. When DDI 3 was released in 2008 as an XML Schema, it was clear that a review of controlled vocabularies was in order, both in terms of content and their relation to the XML Schema.

Accordingly, a Controlled Vocabularies Working Group (CVG) was established by the DDI Alliance in late 2007 to develop vocabularies for DDI 3 elements and attributes. The Technical Implementation Committee (TIC) provided a list of elements and attributes for which vocabularies might be considered.

The multilingual and multicultural Controlled Vocabularies Group has members from several different countries. The group was initially chaired by Ken Miller from the UK Data Archive, and after his retirement in mid-2009, by Taina Jääskeläinen from the Finnish Social Science Data Archive. The group has been meeting via videoconferences approximately every two weeks and expects to publish the vocabularies on the DDI Alliance Web site within the next few months.

Controlled Vocabularies Created for DDI

At the moment, there already are a number of controlled vocabularies embedded in the DDI 3 Schema, including ValueTypeCodeType (e.g., Greater than, Less than, Equal to, etc.). Some elements use well-established external controlled vocabularies. For example, the CountryCodeType element uses ISO country codes. The CVG has developed the following controlled vocabularies which correspond with related elements or attributes in DDI 3.1; the usage of these controlled vocabularies will be enabled with the next version in the DDI 3 development line.

- LifeCycleEvent
- Commonality
- IntendedFrequency
- TimeMethod
- ModeOfDataCollection
- ResponseUnit (for survey type data)
- AggregationMethods
- DataType
- SoftwarePackage
- CharacterSet
- CategoryStatistic
- SummaryStatistic
- DateCalendar
- AnalysisUnit
- ContributorRole
- PublisherRole
- KindOfData (referring to the kind of data disseminated)

The DDI Alliance will recommend the usage of the CVs for DDI 3, and the vocabularies will be published on the DDI Alliance Web site. Each vocabulary will have its own version number.

Each entry in a vocabulary has a code and a corresponding term and definition in English (see example in Table 1). Terms and definitions in other languages can be added as required.

It is possible to add region-specific language versions for terms and definitions. While this can make sense in some cases, in general one language version should suffice. This avoids confusion caused by multiple terms in the same language.

Some of the proposed vocabularies developed are hierarchical, containing broader and narrower terms. The narrower terms may not cover the whole dimension of the broader term, and users are advised to use the broader term if none of the narrower terms is suitable. All vocabularies have an unspecified ‘Other’ term, unless there is a clear reason for not including it. Users are advised to specify in the documentation what they mean by ‘Other’. If the element is of the CodeValueType, it has an ‘OtherValue’ attribute that can be used to enter the specific meaning. CESSDA, for example, is considering capturing the OtherValue information in order for the CVG to determine whether there are additional terms that should be added to specific vocabularies.

The work on DDI 3 controlled vocabularies is still in progress. The group has revised the original draft vocabularies after receiving comments from the CESSDA data archives and other data providers. The biggest challenge the CVG has encountered in its work is that, because DDI 3 is as yet not widely used by data organizations, there are few experts on the standard. The group has consulted TIC on several occasions. We expect that when the standard becomes more widely used, there may be suggestions for changes to some vocabularies, or requests for vocabularies for new elements/attributes. The Qualitative Data Exchange Working Group activities may eventually bring additional term suggestions. At the moment, while the CVG has done its best to take qualitative data into account, the vocabularies are somewhat more geared to quantitative data.

Code	Term	Definition
Median	Median (Mdn)	The score value below which (and above which) half of the scores in a distribution fall (50th percentile).
ValidCases	Valid Cases	Cases with observations considered to be valid, i.e., providing substantial information and to be included for calculation.
InvalidCases	Invalid Cases	Cases which are considered and defined as “missing” (e.g., not ascertained, not applicable, etc.) to be excluded from calculation.
Minimum	Minimum	The lowest valid score in a variable.

Table 1: Extract of the Controlled Vocabulary for Summary Statistics

If a vocabulary has been created for a DDI 3 element/attribute that has a corresponding element/attribute in DDI 2, the vocabulary can also be used for DDI 2. This approach is backward compatible. The documentation of the forthcoming new version in the DDI 2 development line will include information on the use of these CVs.

The Alliance has also published DDI Best Practices for controlled vocabularies .

Genericode

The DDI CVs will be published in the Genericode format, separately from the DDI XML Schemas. Genericode defines a standard format for defining code lists, also known as enumerations or controlled vocabularies. Genericode aims to provide a standard model and XML representation for the contents of a code list. This is an OASIS Genericode Committee Specification.

The Genericode format has a tabular model for code lists. The "rows" are individual entries in a code list, where an entry is a set of one or more codes, plus other metadata, that is associated with a single conceptual entry in the code list. The "columns" are individual (typed) pieces of metadata that can be applied to each entry in a code list. So columns define what kind of data can be in the code list, while rows define what actual data are in the code list (OASIS Code List Representation Requirements, 2007). An advantage of using a controlled set of semantic concepts is in localization where the associated documentation for the coded values can include descriptions in different languages, thus not requiring the coded values themselves to be translated, or where translation is desired, the semantic equivalence of values can be described (OASIS Code List Representation TC Charter).

Maintenance and Management of DDI Controlled Vocabularies

For the vocabularies to remain up-to-date and viable, they need to be maintained. The CVG will function as the management team, reviewing any suggestions for changes, monitoring the types of terms that have been used for 'Other' and their documentation, keeping track of different language versions and liaising with the bodies/persons responsible for them. Updated vocabularies will be published with new version numbers.

Flexible Approach for Specific Needs

DDI controlled vocabularies can be customized to meet local requirements or even replaced by institution-specific vocabularies, if needed. However, any extension or change of terms puts interoperability at risk, particularly if data are to be published or harvested in cross-institutional data portals. If extensions or revisions to DDI CVs are made locally, mappings from the more detailed local vocabulary version to the DDI CV are recommended.

Usage of DDI Controlled Vocabularies for Other Applications

While the DDI controlled vocabularies have been developed for usage with DDI 3, they can be used by other applications as well. The DDI controlled vocabularies are a separate product of the DDI Alliance, published independently of DDI XML Schema.

CESSDA Plans

CESSDA is planning to establish a European research infrastructure for the social sciences in 2011. All members in this coalition will use a common metadata standard for data documentation. The standard will include mandatory or recommended use of controlled vocabularies in certain DDI elements/attributes, and therefore the adopted vocabularies will be translated into the local language(s) of the member organisations. Using the controlled vocabularies will help to facilitate the eventual transformation of DDI 2 documentation to DDI 3.

References

Broughton, V. (2004). *Essential Classification*. London: Facet Publishing.

Chamis, A.Y. (1991). *Vocabulary Control and Search Strategies in Online Searching*. Westport Conn.: Greenwood Publishing Group [Ref. Macgregor & McCulloch, 2006]

Chan, L.M., and Zeng, M.L. (2006). Metadata Interoperability and Standardization – A Study of Methodology Part I, D-*Lib Magazine*, Vol. 12, No 6.

Chu, H. (2007). *Information Representation and Retrieval in the Digital Age*. *American Society for Information Science and Technology*. ASIST Monograph Series.

Garshol, L.M. (2004). Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of it All. *Journal of Information Science*, Vol. 30, No. 4, 378-391.

Mayr, P., and Petras, V. (2008). Building a Terminology Network for Search: The KoMoHe Project. Paper at the 2008 International Conference on Dublin Core and Metadata Applications. Also available online from http://dcpapers.dublincore.org/ojs/pubs/article/viewArticle/931.

Macgregor, G. and McCulloch, E. (2006). 'Collaborative Tagging as a Knowledge Organisation and Resource Discovery Tool', *Library Review*, Vol. 55, No.5, 291–300.

OASIS Code List Representation Requirements (2007). Version 1.0.1, p. 6. Retrieved 7 May 2010 from http://www.oasis-open.org/committees/download.php/23844/oasis-code-list-representation-requirements-1.0.1.pdf.

OASIS Code List Representation TC Charter. Retrieved

7 May 2010 from http://www.oasis-open.org/committees/codelist/charter.php.

Svenonius, E. (2000). *The Intellectual Foundation of Information Organization*. Cambridge, Mass.: MIT Press.

Svenonius, E. (2003). *Design of Controlled Vocabularies*. Encyclopedia of Library and Information Science, 2nd ed.

Zeng, M.L., and Chan, L.M. (2006). Metadata Interoperability and Standardization – A Study of Methodology Part II, D-*Lib Magazine*, Vol. 12, No 6.

Notes

[1] Taina Jääskeläinen, Finnish Social Science Data Archive, Finland: email taina.jaaskelainen@uta.fi. Meinhard Moschner, GESIS - Leibniz Institute for the Social Sciences, Germany: e-mail meinhard.moschner@gesis.org. Joachim Wackerow, GESIS - Leibniz Institute for the Social Sciences, Germany: e-mail joachim.wackerow@gesis.org.

[2] The alpha-2 standard can be extended to alpha-3 or supplemented by region or script subtags where necessary.

[3] DDI Best Practices documents are available at: http://www.ddialliance.org/resources/publications/working/bestpractices

[4] The Organization for the Advancement of Structured Information Standards (OASIS) is a global consortium that drives the development, convergence and adoption of e-business and Web service standards, Web site http://www.oasis-open.org/.

[5] The Council of European Social Science Data Archives, Web site http://www.cessda.org/