Reference number of working document: **ISO/IEC JTC 1/SC 32 WG2 N 1217**

Date: 2008-09-18

Reference number of document: **ISO/IEC WD 20943-5**

Committee identification: ISO/IEC JTC 1/SC 32/WG 2

Secretariat: US

Information technology — Semantic metadata mapping procedure (SMMP)

Warning

This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO nnn-n was prepared by Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 32, *Data management and interchange*.

Introduction

There may be two or more metadata sets applicable to a single information object. For example, there are several metadata sets including DC (Dublin Core), MARC (MAchine Readable Cataloguing), and MODS (Metadata Object Description Schema), which can be used to describe a book. Thus, a data element for an information object may be differently named due to the preferences of individual developers of databases. Resultantly, data exchange between databases becomes difficult.

ISO/IEC 11179 provides a good provision for improving semantic interoperability of data. A metadata registry based on ISO/IEC 11179 is a good way to secure interoperability among databases. However, it is just a beforehand measure. A measure is needed to mediate between metadata sets already developed or used. Metadata crosswalk is the most commonly used way to map a metadata set to another metadata set. However, a metadata crosswalk has poor semantics because it is usually based on a simple one-to-one mapping between data elements. Therefore, a metadata crosswalk is required to be elaborated in order to give semantics and to cover cases other than one-to-one mapping. The idea of ISO/IEC 11179 can be still helpful to improve metadata crosswalk semantically because it addresses the semantics of data and naming principles for data elements.

This standard describes a semantic metadata mapping procedure (SMMP), which is able to maximize the interoperability among metadata sets. The procedure consists of thee main processes such as identifying metadata sets, grouping data elements, and semantic mapping. This standard includes a simple example to explain each process.

Information technology — Semantic metadata mapping procedure (SMMP)

1   Scope

1.1 Background

Differently named data elements may cause a data discrepancy problem. Semantic metadata mapping is required to mediate among those data elements named differently. Metadata crosswalk is the most commonly used way to map a metadata set to another metadata set. However, it has poor semantics because it is meaningful for simple one-to-one mapping. Therefore, a metadata crosswalk is required to be elaborated in order to have semantics and to cover cases other than one-to-one mapping.

1.2 Purpose

The purpose of this standard is to set up a procedure for making metadata crosswalks that conform to ISO/IEC 11179 standard, and thus, to improve semantic harmonization of metadata.

1.3 Scope

This standard describes a semantic metadata mapping procedure (SMMP), which is able to maximize the interoperability among metadata. The procedure consists of three main processes and nine sub-processes. The main processes are identifying metadata sets, grouping data elements, and semantic mapping. This standard includes a simple example to explain each process.

This standard is recommended to be used in a specific subject domain because the procedure can be more meaningful when a specific information object is concerned.

This standard does not consider the interoperability of values.

2   Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 11179-1:2004, Information technology — Metadata registries (MDR) — Part 1: Framework for the specification and standardization of data elements

ISO/IEC 11179-5:2003, Information technology — Metadata registries (MDR) — Part 5: Naming and identification principles for data elements

3   Terms and definitions

For the purposes of this document, the terms and definitions given elsewhere in ISO/IEC 11179 and the following apply.

3.1
Crosswalk
a mapping of the elements, semantics, and syntax from one metadata scheme to those of another [NISO, 2004]

3.2
Complicated difference
a type of semantic heterogeneities which isn't able to be harmonized

3.3
Domain difference
a type of semantic heterogeneities due to different context or culture

3.4
Hierarchical difference
a type of semantic heterogeneities due to different level of detail

3.5
Interoperability
the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality [NISO, 2004]

3.6
Lexical difference
a type of semantic heterogeneities due to different appearance

3.7
Syntactic difference
a type of semantic heterogeneities due to different arrangement of parts

4   Semantic metadata mapping procedure

4.1   General

The procedure for semantic metadata mapping consists of three main processes as Figure 1.
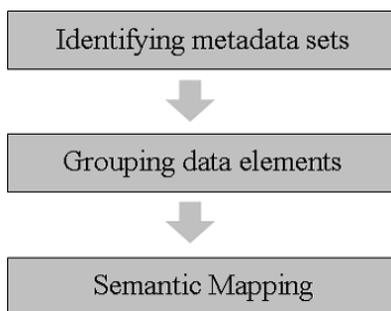
Figure 1 — The procedure for semantic metadata mapping

The first process is to identify metadata sets required to be mapped. We need to survey available metadata sets (in a specific domain).

The second process includes four consecutive sub-processes such as finding objects, grouping all data elements by object, finding properties, and grouping all data elements by property.

The last process is mapping data elements semantically. In this process, we need to arrange all data elements into a table. Notes on the accuracy of matching are included in every slot. A recommended set of metadata can be also provided in the process for guiding future standardization.

Figure 2 shows all sub-processes having relation with main processes.
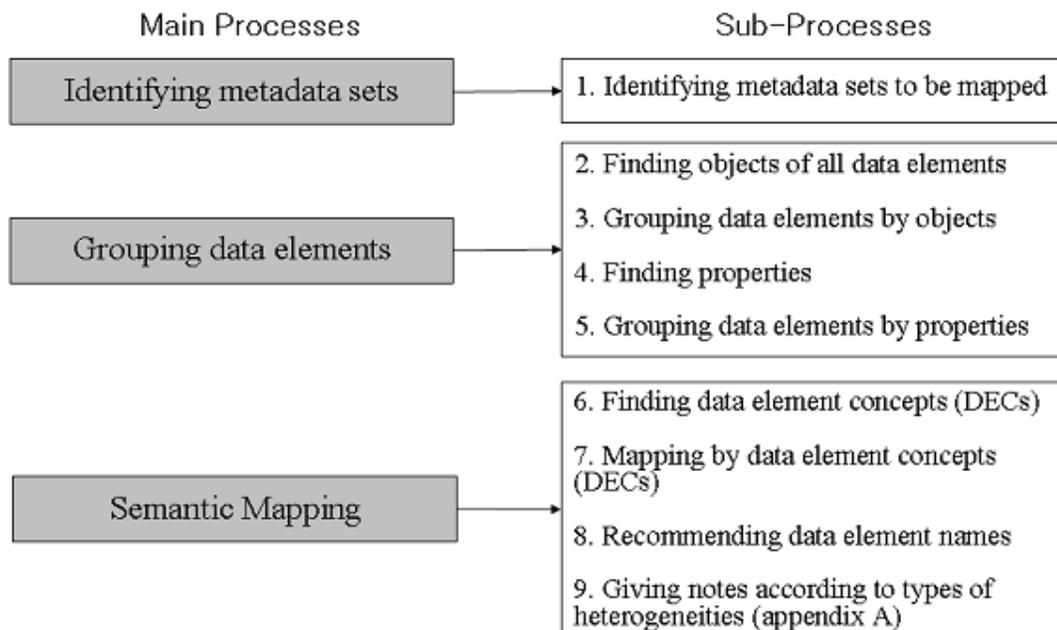
Figure 2 — The main and sub-processes for semantic metadata mapping

4.2   Identifying metadata sets (first process)

4.2.1   Explanation

At first, it is needed to collect available metadata sets, and to identify candidate metadata sets to be mapped. Then, we should check what the domain or service DB, numbers of fields, and sample data are. The authority of each metadata set should be also checked.

If metadata set or target object is not fit for mapping purpose, it may not be chosen.

4.2.2   Example

· Domain: e-book cataloging
· Available metadata sets: OpenEBPS, MODS and TEI

Table 1 — Analysing available metadata sets

Metadata set name	OpenEBPS	MODS	TEI header
Domain or service DB	Description of Electronic Book	Description of Library resources	Encoding methods for machine-readable texts
Number of fields	15	About 60 (top level: 20)	Over 20

Sample data	yes	No	yes
Authority	Open eBook Forum	LC	TEI Consortium

4.3 Grouping data elements (second process)

4.3.1 Explanation

The next process is grouping data elements including finding objects and grouping data elements by object, and then finding properties involved in the objects and sub-grouping data elements by properties.

For convenience, it is helpful to select a standard metadata set among the collected metadata set and aggregate data elements according to the standard metadata set. The simplest or the highest-level metadata set is desirable to become the primary metadata set.

All data elements included in the candidate metadata sets should be aggregated by property. Data elements less important may be eliminated. Some data elements, which can't be grouped, are supposed to be set aside.

In this process, metadata experts should perform the work along with domain experts.

4.4.2 Example

In the example there is one object class, e-Book, while properties are many as shown below.

- **Object class**: e-Book
- **Properties**: title, author, subject, …, edition

Table 2 — An example of grouping data elements by property

OpenEBPS*	MODS	TEI
Title	title subTitle *partNumber* *partName* *nonSort*	title seriesStmt:title *seriesStmt:idno*
Creator(role) Creator(file-as)	name:role name:namePart *name:displayForm* *name:affiliation* *name:discription*	author

Subject	topic	keyword
	classification	classCode
	catographics	*catRef*
	occupation	
…	…	…
(no data element)	edition	*fileDesc_editionStmt_date*
		fileDesc_editionStmt_edition
		fileDesc_editionStmt_respStmt
		fileDesc_editionStmt_respStmt_name
		fileDesc_editionStmt_respStmt_resp

* standard metadata set

Similar data elements are grouped according to the standard metadata set, OpenEBPS. In the table, the italicized parts mean data elements considered less important in the target application domain.

4.4   Semantic mapping (third process)

4.4.1   Explanation

Finding out object classes and then properties hidden in and related to all data elements of the standard metadata set, we can create common DECs according to ISO/IEC 11179-1.

The third process starts from finding common data element concepts in each group of data elements based on objects and properties found in the second process. If the domain ontology or taxonomy was known, it will be very helpful to construct DECs.

Finally, all candidate data elements are arranged into a table by the common DECs. Types of heterogeneity can be described near by the data elements in the table. The types are composed of six categories. (See detail in Appendix A)

- **Same, no difference**: no description
- **Hierarchical difference**: H/gen, H/spe, H/com, H/dec
- **Domain difference**: D
- **Lexical difference**: L/syn, L/pre, L/abb, L/sim, L/acr, L/cas, L/lan
- **Syntacticdifference**: S/ord, S/del
- **Complicated difference**: C

A recommended set of metadata can be provided for guiding future standardization

4.4.2   Example

The data element concepts can be found as below:

5

- **DECs**: ebookTitle, ebookAuthor, ebookSubject, …, ebookEdition

Finally, we can create DECs according to ISO/IEC 11179-1. New DECs should be also created for data elements set aside at the process two.

Table 3 shows the crosswalks finally obtained through the procedure. The first column is standard DECs while the right end column is recommended data elements. Between them are data elements from the candidate metadata sets.

Table 3 — Semantic mapping of metadata

DEC	OpenEBPS		MODS		TEI		Recommaned DE
ebookTitle	Title		title	H/dec	title		**ebookTitle**
			subTitle	H/dec	seriesStmt:title	T:pre	
ebookAuthor	Creator(role)	C	name:role	C	author	S/mis	**ebookAuthorName**
	Creator(file-as)	D	name:namePart	D			
ebookSubject	Subject		topic	L/syn	keyword	L/sim	**ebookSubject**
			classification	L:pre	class	L/pre	
ebookEdition			edition	L/cas S/mis	edition	L/cas S/mis	**ebookEditionNumber**

Annex A
(normative)

Types of semantic heterogeneity

Types of semantic heterogeneity of metadata can be classified into six categories: no difference, hierarchical difference, domain difference, lexical difference, syntactic difference, and complicated difference.

Exactly same data elements are able to be mapped by one-to-one mapping.

Hierarchical difference is caused by different level of detail such as generalization/specialization and composition/decomposition. For example, a general term 'price' can be specialized to 'retail price', or 'whole sail price'; and a person's 'name' is composed of 'first name', 'middle name', and 'family name'. Data elements having hierarchical difference should be mapped using the higher level or composed terms if required.(See more detail in ISO/IEC 20943-1:2003)

Domain difference is caused by different context or culture, e.g. summary vs. synopsis, and color vs. colour. In this case one-to-one mapping is possible.

Lexical difference means different appearance of a data element related to synonyms, preferred terms, similar terms, abbreviations, acronyms, case sensitivity, or languages. This kind of heterogeneities is certainly solved by one-to-one mapping.

Syntactic difference is due to different arrangement of parts within a data element name. The order of words, the type of delimiter, and missing words can cause heterogeneity between semantically same data elements, e.g. ordering (family name : name (family)), delimiters (Family-Name : FamilyName), and missing (classification code : classification). Of course this type of heterogeneity solved by one-to-one mapping as well.

Complicated difference is the type of heterogeneity which isn't able to be solved without human intervention.

Table 4 — Types of semantic heterogeneity

Type	Sub-Type	Mark	Examples
	Sub-Type	Mark	Examples
	Ways of harmonization		
Same		none	
	One-to-one mapping		
Hierarchical	Generalization	H/gen	Price
	Specialization	H/spe	Retal price, Wholesale price, …
	Composition	H/com	Name
	Decompositon	H/dec	Family name, given name, …
	One-to-many or many-to-one mapping (if required)		
Domain		D	Summary : Synopsis
			Color : Colour

7

	One-to-one mapping		
Lexical	Synonyms	L/syn	First name : Given name
	Abbreviation	L/abb	Address : Addr.
	Acronyms	L/acr	Serial Number :SN
	Case sensitivity	L/cas	Address : ADDRESS
	Language	L/lan	Name : 이름
	One-to-one mapping		
Syntactic	Ordering	S/ord	Family name : Name (family)
	Delimiters	S/del	Family-name : Family_name
	Missing	S/mis	Author name : Author
	One-to-one mapping		
Complicated		C	
	Mapping is impossible.		