

Reference number of working document: **ISO/IEC JTC 1/SC 32 WG2 N 1473**

Date: 2010-11-16

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

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

Secretariat: **US**

Information technology — Procedures for achieving metadata registry content consistency — Part 5: 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.

Document type: **Technical Report**
Document subtype:
Document stage: **(20) Preparatory stage**
Document language: **E**

Copyright notice

This ISO document is a working draft or committee draft and is copyright-protected by ISO. While the reproduction of working drafts or committee drafts in any form for use by participants in the ISO standards development process is permitted without prior permission from ISO, neither this document nor any extract from it may be reproduced, stored or transmitted in any form for any other purpose without prior written permission from ISO.

Requests for permission to reproduce this document for the purpose of selling it should be addressed as shown below or to ISO's member body in the country of the requester:

ISO copyright office

Case postale 56 • CH-1211 Geneva 20

Tel. + 41 22 749 01 11

Fax + 41 22 749 09 47

E-mail copyright@iso.ch

Web www.iso.ch

Reproduction for sales purposes may be subject to royalty payments or a licensing agreement.

Violators may be prosecuted.

Contents

Page

Foreword	iv
Introduction.....	v
1 Scope	1
2 Normative references	1
3 Terms and definitions.....	2
4 Semantic metadata mapping procedure	2
4.1 General	2
4.2 Identifying metadata element sets (First process)	3
4.2.1 Method	3
4.2.2 Examples	3
4.3 Grouping data elements (Second process)	4
4.3.1 Method	4
4.3.2 Examples	4
4.4 Semantic mapping (Third process)	5
4.4.1 Method	5
4.4.2 Examples	6
4.5 Value mapping	6
4.5.1 Conversion	7
4.5.2 Structural rearrangement	7
4.5.3 Examples	7
Annex A (normative) Types of semantic heterogeneity.....	8
Annex B (Informative) Semantic metadata mapping and metadata registry.....	10

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 element sets applicable to an information object. For example, **metadata such as** DC (Dublin Core), MARC (MACHine Readable Cataloguing), and MODS (Metadata Object Description Schema) 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 database developers. Consequently, data exchange among databases becomes difficult or almost impossible.

ISO/IEC 11179 provides a good provision for improving the semantic interoperability of metadata. A metadata registry based on ISO/IEC 11179 offers a good way to secure interoperability among databases. However, it is just a prearranged measure. In order to mediate among plural metadata element sets already developed or used, other measures are necessary. For example, metadata crosswalk is the most commonly used way to map a metadata element set to another metadata element set. However, the metadata crosswalk has poor semantics; it provides a simple one-to-one mapping table among data elements without any explanation about the semantic relationship. Therefore, the metadata crosswalk needs to be elaborated in order to give semantics and to cover cases other than one-to-one mapping. The basic concept of ISO/IEC 11179 can still be applicable to the improvement of semantic metadata crosswalk because it addresses the semantics of metadata and naming principles for data elements.

This standard describes a semantic metadata mapping procedure (SMMP), which can maximize the interoperability among metadata element sets. The procedure consists of three main processes: identifying metadata element sets, grouping data elements, and semantic mapping. This standard also includes **types of value mapping**.

Information technology — Procedures for achieving metadata registry content consistency — Part 5: Semantic metadata mapping procedure (SMMP)

1 Scope

1.1 Background

Data elements having different names even though they have the same meanings may cause a data discrepancy problem when such data are shared or **interchanged**. Thus, semantic metadata mapping is required to mediate among these data elements to allow sharing or interoperable use. A metadata crosswalk is the most commonly used method to map a metadata element set to another metadata element set. However, it has poor semantics because it is meaningful only for simple one-to-one mapping. Therefore, the metadata crosswalk needs 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 the ISO/IEC 11179 standard, subsequently improving the semantic harmonization of metadata.

1.3 Scope

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

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

This standard **addresses** interoperability of values **as well**.

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 11179-5:2005, 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 that cannot be harmonized

3.3

Domain difference

a type of semantic heterogeneities arising from the different kinds of contexts or cultures

3.4

Hierarchical difference

a type of semantic heterogeneities arising from the different levels of details

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 arising from different appearances

3.7

Primary metadata element set

a metadata element set to which other metadata element sets are mapped

3.8

Syntactic difference

a type of semantic heterogeneities arising from varying arrangement of parts

4 Semantic metadata mapping procedure

4.1 General

In this standard, metadata mapping covers data element mapping and value mapping. Data element mapping will be performed before value mapping.

The procedure for **data element** mapping consists of three main processes as shown in Figure 1.

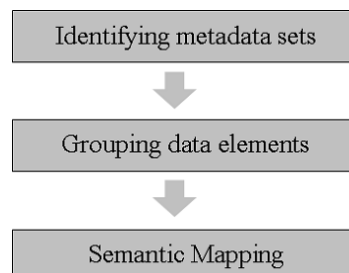


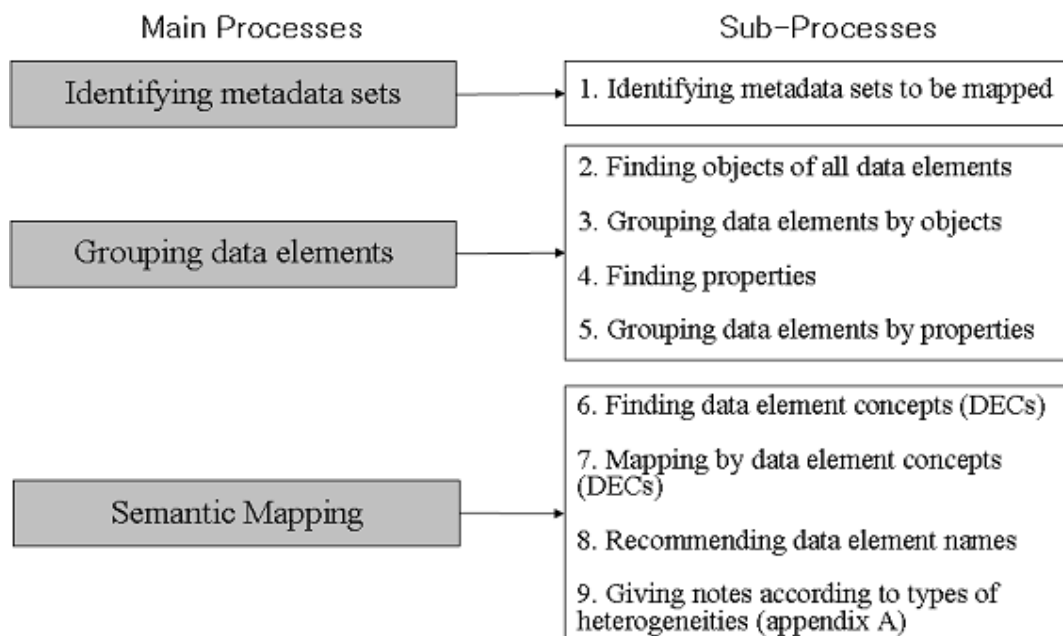
Figure 1 — Procedure for data element mapping

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

The second process is to group data elements obtained from the identified metadata element sets, including four consecutive sub-processes namely, finding objects, grouping all data elements by object, finding their properties, and grouping all data elements by property.

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

Figure 2 shows all sub-processes related to corresponding main processes.

**Figure 2 — Main and sub-processes for data element mapping**

4.2 Identifying metadata element sets (First process)

4.2.1 Method

First, it is necessary to collect available metadata element sets and to identify candidate metadata element sets to be mapped. Then, what the domain or service DB is should be checked, how many numbers of fields should be counted, and whether sample data exists or not should be checked. Who or which organization has the authority over each metadata element set should also be checked.

If the metadata element set or target object is not suitable for mapping, it may not be chosen.

4.2.2 Examples

- Domain: e-book cataloging

- Available metadata element sets: OpenEBPS, MODS and TEI

Table 1 — Analysing available metadata element sets

Metadata element set name	OpenEBPS	MODS	TEI header
Domain service DB or	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 Method

The next process is to group data elements including to find objects and to group data elements by object, and then to find properties involved in the objects and sub-grouped data elements by properties.

For convenience, it is helpful to select a primary metadata element set among the collected metadata element sets and aggregate data elements according to the primary metadata element set. The simplest or the highest-level metadata element set is recommended to be the primary metadata element set.

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

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

4.3.2 Examples

In the sample object class, the e-Book has plural properties as shown below.

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

Table 2 — Example of grouping data elements by property

OpenEBPS*	MODS	TEI
Title	Title subTitle <i>partNumber</i> <i>partName</i> <i>nonSort</i>	title seriesStmt:title <i>seriesStmt:idno</i>

Creator(role)	name:role	
Creator(file-as)	name:namePart <i>name:displayForm</i> <i>name:affiliation</i> <i>name:discription</i>	author
Subject	Topic classification <i>catographics</i> <i>occupation</i>	keyword classCode <i>catRef</i>
...
(no element) data	Edition	<i>fileDesc_editionStmt_date</i> <i>fileDesc_editionStmt_edition</i> <i>fileDesc_editionStmt_respStmt</i> <i>fileDesc_editionStmt_respStmt_name</i> <i>fileDesc_editionStmt_respStmt_resp</i>

* primary metadata element set

Similar properties of MOD and TEI are grouped according to those of the primary metadata element set, OpenEBPS. In the table, the italicized parts refer to properties considered less important in the target application domain.

4.4 Semantic mapping (Third process)

4.4.1 Method

After identifying object classes and properties hidden in and related to all data elements of the primary metadata element 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 is known, it will be very helpful to construct common DECs.

Finally, all candidate data elements are arranged into a table by the common DECs. Types of heterogeneity can be described near the data elements in the table. The types consist 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/abb, L/acr, L/cas, L/lan, L/var
- **Syntactic difference:** S/ord, S/del, S/mis
- **Complicated difference:** C

A recommended set of metadata can be provided to guide future standardization.

4.4.2 Examples

The data element concepts found can be shown as follows:

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

Finally, we can create DEC's according to ISO/IEC 11179-1. New DEC's should be also created for data elements set aside during the second process.

Table 3 shows the final result obtained through the procedure. The common DEC's are described in the first column while the recommended data elements are done in the right end column. Between them are data elements from the candidate metadata element sets.

Table 3 — Semantic mapping of metadata

Common DEC	OpenEBPS		MODS		TEI		Recommended DE
ebookTitle	Title		Title	H/dec	Title	H/dec	ebookTitle
			Subtitle	H/dec	seriesStmt:title	H/dec S/del	
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/syn	ebookSubject
			Classification	L:syn	Class	L/syn	
ebookEdition			Edition	L/cas S/mis	Edition	L/cas S/mis	ebookEditionNumber

4.5 Value mapping

Even though data elements are successfully mapped, it is not enough when the data type is different to each other. Value domains need to be checked and harmonized in order to interchange real data of databases. The value of each data element may be converted or rearranged to solve the inconsistency of data type, when the data types are not equal.

NOTE The conversion or rearrangement of the value is applicable only when the conceptual domains which govern the value domains to be mapped are same. If the conceptual domains are different to each other, value mapping may be performed partially or impossible.

4.5.1 Conversion

Conversion can only apply to determinate values.

Code conversion: In case of using different code sets sharing same conceptual domain for the value domain, a code set is converted to the other code set.

Letter conversion: In case of using characters such as abbreviation, acronyms, different case, lexical variation, or different delimiters for the value domain, a part of the value is converted to a preferred representation.

Unit conversion: In case of using numeric value for the value domain, the numeric value is converted using an appropriate conversion rule.

4.5.2 Structural rearrangement

Composition: If a value is separated into two or more data elements, those values are composed into one value of an appropriate data element.

Decomposition: If a value should be separated into two or more data elements, the value is decomposed into two or more values of appropriate data elements.

Order change: If a value is composed of two or more words which are ordered differently, the order of the words are changed in an appropriate order.

4.5.3 Examples

Table 4 shows the examples of value mapping.

Table 4 — Examples of value mapping

Mapping types		Examples		
Conversion	Code conversion	KOR to KR		
	Letter conversion	KATS to Korea Agency for Technology and Standards		
	Unit conversion	1 in to 2.54 cm		
Structural rearrangement	Composition	First Name: John Given Name: Kennedy	to	Full Name: John Kennedy
	Decomposition	Full Name: John Kennedy	to	First Name: John Given Name: Kennedy
	Order change	John Kennedy to Kennedy, John		

Annex A (normative)

Types of semantic heterogeneity

The 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.

Identical data elements can be mapped using one-to-one mapping.

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

Domain difference is caused by different contexts or cultures, e.g. summary vs. synopsis. In this case, one-to-one mapping is possible.

Lexical difference refers to the different appearance of a data element related to synonyms, abbreviations, acronyms, case sensitivity, languages, or variation. This kind of heterogeneity is certainly solved by one-to-one mapping.

Syntactic difference arises from the varying 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). This type of heterogeneity can likewise be solved by one-to-one mapping.

Complicated difference is the type of heterogeneity that cannot be solved without human intervention.

Table 4 — Types of semantic heterogeneity

Type	Sub-Type	Mark	Examples
Ways of harmonization (<i>Types of mapping</i>)			
Same	Same		
	One-to-one mapping		
Hierarchical	Generalization	H/gen	Price
	Specialization	H/spe	Retail price, Wholesale price, ...
	One-to-one mapping (<i>dumb down</i>)		
	Composition	H/com	Name
	Decomposition	H/dec	Family name, given name, ...
One-to-many or many-to-one mapping (if required)			
Domain	Domain	D	Summary : Synopsis

	One-to-one mapping (if required)		
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
	Variation	L/var	Color : Colour
	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	Complicated	C	
	Mapping is impossible.		

Annex B (Informative)

Semantic metadata mapping and metadata registry

When semantic metadata mapping is performed among metadata registries based on ISO/IEC 11179, the mapping agent can act as a read-only user and a submitter of the metadata registries. During the first and second processes, it uses contexts, object classes, properties, data element concepts, and data elements out of the metadata registries. It can also submit new object classes, properties, data element concepts, and data elements newly obtained in the last process.

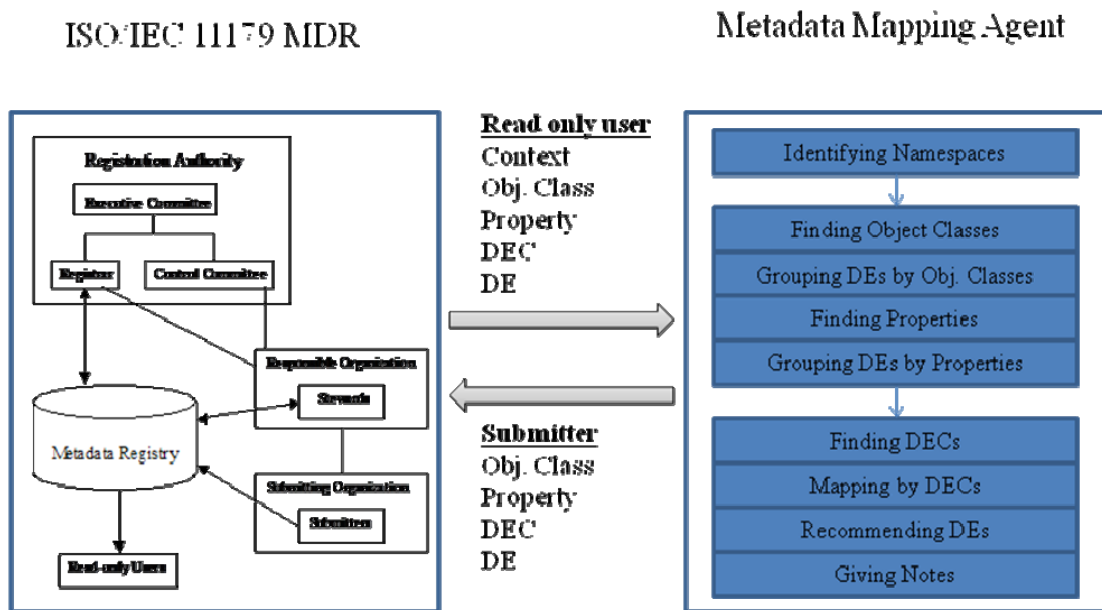


Figure 3 — Relationship between metadata mapping and registry

Bibliography

- [1] A Model for Semantic Equivalence Discovery for Harmonizing Master Data, *Baba Piprani*, OTM 2009 Workshops, LNCS 5872, pp. 649–658, 2009