The representation of Knowledge Organization Structure (KOS) data.

Dagobert Soergel
College of Information Studies, University of Maryland
College Park, MD  20742-4345
Office:(301) 405-2037  Fax (301) 314-9145
ds52@umail.umd.edu  www.clis.umd.edu/faculty/soergel/

JCDL NKOS Workshop Roanoke, VA 2001-6-28

The purpose of KOS standards

1  Input of KOS data into programs / Transfer of data from one program to another
   1.1 Format for original input files (XML difficult for that, need a user-friendly format)
   1.2 Transfer from one KOS management program to another
   1.3 Transfer from a KOS management program to an information system that uses a KOS for authority control, query expansion (synonym and /or hierarchic), display/browse/search, or other purposes
   1.4 Transfer from a KOS management program to a KOS use (display / browse / search / etc.) program

2  Accessing KOS for applications. Includes querying KOS and viewing results (for example, using Z39.50)
   2.1 By people. Standardized displays would be helpful here (but have the same problems as standardizing the interfaces to search engines).
   2.2 By systems to use data from internal or external KOS for many types of processing, such as inference, natural language processing, knowledge-based clustering, index construction, query term expansion etc.

3  Identifying specific terms/concepts in specific KOS
   This requires rules for URIs that uniquely identify specific term/concept records in specific thesauri. Needs a name resolution service (such a KOS registry)
   3.1 Links from one KOS to another
   3.2 Indexing terms/concepts in the metadata for an object, or any other reference to a term/concept in a text/object

4  Prescribing or giving guidance on good practices
   For some products, proper practices guarantee properties to be standardized
Two levels of standardization

Standards that give a general format, leaving the user(s) or user communities to develop specifics (e.g., relationship types)

Standards that give specifics, either hard-coded in the format or given separately as a name space supplementing a general standard.

(In the KOS domain there is a third level of standardization, standardizing concepts and terms and their relationships, but that is not the subject of this note.)

Evaluation of standards

1 Expressivity

What kinds of statements can be made about the domain. What kind of operations and inferences do these statements support. This depends on the underlying data model. This must be judged with respect to the requirements of the expected application.

1.1 How extensible

1.2 Expressing processing rules (e.g., for relationship types)

2 Ease of application

2.1 Ease of writing software

2.2 Compatibility with related standards

2.3 Ease of understanding the standard and of writing and reading specifications

2.4 Ease of writing and reading data files

2.5 Parsimony of expression

2.6 Size of data files

3 Depth of support (in place or anticipated)

3.1 Recognition of the body issuing the standard

3.2 Technical support available

3.3 Availability of software

3.4 Breadth of adoption
The many forms of Knowledge Organization Systems (KOS) and their standards

Dictionaries, glossaries

Thesauri
ZThes (using Z39.50, strictly ANSI Z39.19)
http://www.loc.gov/z3950/agency/profiles/zthes-04.html
Browser at http://muffin.indexdata.dk/zthes/tbrowse.zap
Vocabulary Markup Language (VocML) (under discussion at NKOS)
See also http://ceres.ca.gov/KOS/
ISO 5964-1985(E) (multilingual)
USMARC format for authority data
(http://lcweb.loc.gov/marc/authority/ecadhome.html)

Topic maps (reference works, encyclopedias) (http://www.topicmaps.org/about.html)
XML Topic Maps (XTM) 1.0 (http://www.topicmaps.org/xtm/1.0/)

Concept maps

Classification schemes
USMARC format for classification data
http://lcweb.loc.gov/marc/classification/eccdhome.html

Ontologies
Knowledge Interchange Format (KIF) NCITS.T2/98-004
(http://meta2.stanford.edu/kif/dpans.html)
Ontology Markup Language (OML) /
Conceptual Knowledge Markup Language (CKML)
(http://www.ontologos.org/OML/CKML-Grammar.html)
Ontology Interface Layer (OIL) (http://www.ontoknowledge.org/oil/)

Generic standards for knowledge structures, entity-relationship models
Resource Description Framework (RDF) (http://www.w3.org/RDF/)
Metadata Coalition. Open Information Model (OIM). Knowledge Management Model
(http://www.mdcinfo.com/OIM/)
XTM might also fit here
ISO terminology-related standards (two repeated)

ISO 639:1988 Code for the representation of names of languages
ISO 704:2000 Principles and methods of terminology
ISO 860:1996 Terminology work - Harmonization of concepts and terms
ISO 1087-1:2000 Terminology - Vocabulary
ISO 1951:1997 Lexicographical symbols particularly for use in classified defining vocabularies
ISO 6156:1987 Magnetic tape exchange format for terminological/lexicographical records (MATER)
ISO 10241:1992 Preparation and layout of international terminology standards
ISO 12199:2000(E) Alphabatical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet
ISO 12200:1999 Computer applications in terminology - Machine-readable terminology interchange format (MARTIF) - Negotiated interchange
ISO/TR 12618:1994 Computer aids in terminology - Creation and use of terminological databases and text corpora
ISO 12620:1999 Computer applications in terminology - Data categories

Standards in preparation:

ISO/PWI 12200-Amd 1 Computer applications in terminology - Machine-readable terminology interchange format (MARTIF) - Amendment 1: Extended MARTIF (including a normative Annex H to ISO 12200)
ISO/CD 12615 Bibliographic references for terminology work
ISO/CD 12616.2 Translation-oriented terminography
ISO/AWI 12618 Computer applications in terminology - Design, implementation and use of terminology management systems (Rev. of ISO/TR 12618)
ISO/FDIS 15188 Project management guidelines for terminology standardization
ISO/CD 16503 Computer applications in terminology - Representation format for terminological data collections - MARTIF-compatible with specified constraints (MSC)
ISO/CD 16642 Computer applications in terminology - Meta model for representing terminological data collections
ISO/CD 17241 Computer applications in terminology - Generic model (GENETER) for SGML-based representations of terminological data
Domains in and around information studies (bold = faculty strength)

**Disciplinary domains**

- Social and political process
- Information policy
- Administration and management
- Economics
- Communication and information transfer
- Information structure and systems
- Computer science
  - Artificial intelligence and knowledge-based systems
  - Natural language processing (NLP).
- Human-computer interaction, information system interfaces
  - Cognitive science
  - Psychology (cognitive psychology, decision-making)
  - Linguistics
  - Semiotics
  - Epistemology
- Mass communication and journalism
- Literature, literature and society
- Overarching domains (connected to everything)
  - History
  - Professional issues
  - Philosophy of knowledge, philosophy of science

**Context domains**

- Librarianship;
- Digital libraries;
- Archives and records management;
- Information and knowledge management;
- Information and learning, children's and young adult’s information, children's and young adult’s literature;
- School library media;
- Health information, medical informatics.

A concept map example
Specific KOS data that must be represented

Overall data structure

Consider three levels: Term variants (strings), terms, concepts

<table>
<thead>
<tr>
<th>Term variants (strings)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AST, aspartate aminotransferase, GOT, glutamate oxaloacetate transaminase</td>
</tr>
<tr>
<td>Give information about term variants</td>
</tr>
<tr>
<td>Term variants can be related to each other, e.g., a term variant may have a sort form.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term = a group of variants of the same term, represented by a preferred member of the group</td>
</tr>
<tr>
<td>aspartate aminotransferase, glutamate oxaloacetate transaminase</td>
</tr>
<tr>
<td>Give information about the term (some of this inherits to all term variants)</td>
</tr>
<tr>
<td>Link term variants to terms (AC = ACronym)</td>
</tr>
<tr>
<td>aspartate aminotransferase AC AST</td>
</tr>
<tr>
<td>glutamate oxaloacetate transaminase AC GOT</td>
</tr>
<tr>
<td>Other frequent relationship SP Spelling Variant</td>
</tr>
<tr>
<td>May give meaning-based relationships among terms; alternatively, see concepts.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concepts can be established independently from any terms. Can also establish concepts as groups of terms with the same meaning and represent them by a preferred term; all other terms are considered non-preferred synonyms:</td>
</tr>
<tr>
<td>aspartate aminotransferase</td>
</tr>
<tr>
<td>glutamate oxaloacetate transaminase</td>
</tr>
<tr>
<td>Give information about concepts and relate concepts to concepts</td>
</tr>
<tr>
<td>aspartate aminotransferase</td>
</tr>
<tr>
<td>BT aminotransferases</td>
</tr>
</tbody>
</table>
Often a much simpler structure is used, linking term variants directly to concepts through ST. Even simpler systems employ USE instead, restricting its range to concepts that have been declared descriptors (jumping over any narrower concepts that might exist).

Groups of terms or concepts might also be unit, including fixed graphs of concept relationships

**Data elements**

At each level, there are many, many pieces of information that are required by one application or another. Standards need to accommodate these pieces of information. Any of these pieces of information can be represented as relationships (as in RDF) or as data fields.

Administrative data: Source tracking for each piece of information, history of changes / audit trail.

Schema data, for example, hierarchy of relationship types.

**Information on arrangement / display**

KOS to be used by people should convey meaning. Thus information about the sequence of concepts in a linear display or the placement of concepts in a concept map becomes important.
Elements of an XML KOS data specification

This scheme illustrates some types of data needed in a sophisticated KOS management system but is by no means complete. It is parsimonious yet allows the recording of many types of data. It gives enough information to derive a full XML specification.

This spec assumes that data from each source are grouped, so that source attribution is not needed for each element; otherwise the structure would be much more complex. This works for a communications format but not for an internal database format.

The term itself is indicated in a relationship of type TERM. This allows for terms in multiple languages for the same concept and simplifies the schema since elements in term would be the same as in relationship target.

The scope element was inspired by the Topic Map Standard (http://www.topicmaps.org/xtm/1.0/)

The scheme needs a method for indicating a relationship set defined elsewhere and used within the source or for defining a relationship set for the source.

Default is minOccurs="1" maxOccurs="1"

Source (minOccurs="0" maxOccurs=""unbounded"")
  Pointer to or definition of relationship set used
  Unit: Concept or term or group of terms (minOccurs="0" maxOccurs=""unbounded"")
    Unique identifier
    Hierarchy position (minOccurs="0" maxOccurs=""unbounded"")
      Hierarchical level
      Class number / notation
    Scope for which this concept/term holds (minOccurs="0" maxOccurs=""unbounded"")
  Relationship (minOccurs="0" maxOccurs=""unbounded"")
    Relationship type
    Relationship target
      /* See below for structure. */
    Relationship strength (minOccurs="0" maxOccurs="1")
  Audience level /* Of this relationship */ (minOccurs="0" maxOccurs=""unbounded"")
  Perspective /* Of this relationship */ (minOccurs="0" maxOccurs=""unbounded"")
  Scope for which this relationship holds (minOccurs="0" maxOccurs=""unbounded"")
  Relationship, added information (minOccurs="0" maxOccurs=""unbounded"")
    /* This could be a scope note explaining the relationship, an image illustrating the relationship, another term, etc. */
    Type of added information /* Relationship types might be reused here. */
    Relationship target
    Audience level /* Of this piece of info. */ (minOccurs="0" maxOccurs=""unbounded"")
    Perspective /* Of this piece of information */ (minOccurs="0" maxOccurs=""unbounded""
Where relationship target has this structure (unifying term, text, images, multimedia document)

<table>
<thead>
<tr>
<th>Relationship target</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
</tr>
<tr>
<td>/* Includes types of terms (descriptor, other preferred term, non-preferred term and types of texts and other documents, may be an elaborate hierarchy. */</td>
</tr>
<tr>
<td><strong>Target value (a term or a document)</strong></td>
</tr>
<tr>
<td><strong>Term</strong></td>
</tr>
<tr>
<td>Term variant (minOccurs=&quot;0&quot; maxOccurs=&quot;unbounded&quot;)</td>
</tr>
<tr>
<td>Type of variant</td>
</tr>
<tr>
<td>/* Such as Preferred Spelling, other SPelling, ABbreviation, Full Term. */</td>
</tr>
<tr>
<td>Term form (complete term or Stem plus suffix)</td>
</tr>
<tr>
<td>Complete term</td>
</tr>
<tr>
<td>Stem plus suffix</td>
</tr>
<tr>
<td>Stem</td>
</tr>
<tr>
<td>Suffix</td>
</tr>
<tr>
<td><strong>Document</strong></td>
</tr>
<tr>
<td>Language (zero to many, exactly one for terms)</td>
</tr>
<tr>
<td>Audience level /* Of this rel.ship target */ (minOccurs=&quot;0&quot; maxOccurs=&quot;unbounded&quot;)</td>
</tr>
<tr>
<td>Perspective /* Of this relationship target */ (minOccurs=&quot;0&quot; maxOccurs=&quot;unbounded&quot;)</td>
</tr>
<tr>
<td>Scope for which this/term holds (minOccurs=&quot;0&quot; maxOccurs=&quot;unbounded&quot;)</td>
</tr>
</tbody>
</table>
**A taste of ISO 12620**

From Translation Research Group [www.ttt.org](http://www.ttt.org), Last updated: January 27, 2001

**CLS Framework: ISO 12620 data categories section 05**

**Menu of data classes**
| (1) Terms | (2) Term-Related Data Categories | (3) Equivalence | (4) Subject Field |
| (5) Concept Related Description | (6) Concept Relation | (7) Conceptual Structures | (8) Note |
| (9) Documentary Language | (10) Terminology Management |

**Section 5: Concept related description**

Description: Any kind of explanatory material used to elucidate a concept.

Blind MARTIF Representation: `<descrip type=X>...</descrip> or`  
Blind MARTIF Representation: `<ptr type=X target=Y/>`, where Y is the refid of a `<refObject type=binaryData>`, which as noted above, can contain embedded binary data (in hex) or a URL, and X is one of the following:

### 5.1 definition

Description: A statement that describes a concept and permits its differentiation from other concepts within a system of concepts.

Blind MARTIF Representation: `<descrip type=definition>content unrestricted</descrip>`

### 5.2 explanation

Description: A statement that describes and clarifies a concept and makes it understandable, but does not necessarily differentiate it from other concepts.

Example: Explanation of catalyst: `<chemistry> material that triggers or accelerates a chemical reaction

### 5.3 context

Description: A text or part of a text in which a term occurs.

Note 2: Contexts are documented very frequently in descriptive and translation-oriented terminology work. Some databases use "example" for contextual references. Such data categories should be converted to the data category "context" for interchange purposes.

In addition to providing information about concepts, contexts provide text-typological information valuable for determining term usage and collocational references. Consequently some databases classify context as a term-related data category.

### 5.4 example

Description: Descriptive material that provides a sample of the entity defined in the entry.
5.5 nontextual illustrations

5.5.1 figure
Description: A diagram, picture, or other graphic material used to illustrate a concept or group of concepts.

5.5.2 audio
Description: Sound, spoken words, music, or other audible representation used to illustrate or explain terms or concepts. Example: A recording of the pronunciation of a term

5.5.3 video
Description: Recorded visual images used to represent or illustrate terminological information. Example: Video images can be used to illustrate a concept, a process, a test method, etc.

5.5.4 table
Description: An array of data arranged in columns and rows used in documenting, explaining, or describing a concept within a terminology collection.

5.5.5 other binary data
Description: Any foreign data not covered by the previous categories. Example: Spreadsheets, virtual reality files, flight simulations, and the like.

5.6 unit
Description: A relationship between a reference value as defined by an authoritative body; a quantity measured. Example: force is measured in newtons; length in millimetres; weight grams
Note: There is only one unit for each quantity in SI. The unit used to measure a quantity can be extraordinarily valuable in a terminology collection. In some cases, it can represent a major characteristic for determining the precise identity of a referenced concept, especially if polysemy or lack of precision creates ambiguity in a text.

5.7 range
Description: The relationship between a set of limits within which a quantity is measured, as expressed by stating the lower and upper range values.
example: 0 - 100 °C = liquid state of water
Note: Range, like unit, can be a critical delimiting characteristic in defining a concept, particularly in materials databases.

5.8 characteristic
Description: A mental representation of a property of an object serving to form and delimit its concept. Example: compressibility (gas); flammability (fuel); liquidity (financial assets)
Appendix B. The Zthes Abstract Model in XML
(from http://www.loc.gov/z3950/agency/profiles/zthes-04.html)

Appendix B.1. The Zthes DTD for XML

This DTD was supplied by Thomas Place. It is put forward not as a ``good'' XML representation of Thesaurus information (whatever that might be construed to mean) but as a pragmatically valuable alternative encoding of the Zthes abstract record. Real Zthes data sets have been exchanged in the form of XML documents conforming to this DTD.

<!-- Zthes DTD
   Based on Z39.50 Profile for Thesaurus Navigation, version 0.1 (20 Feb 1999)
   Version of DTD: 25 Feb 1999 -->

<!-- #PCDATA: parseable character data = text
   occurrence indicators (default: required, not repeatable):
   ?: zero or one occurrence (optional)
   *: zero or more occurrences (optional, repeatable)
   +: one or more occurrences (required, repeatable)
   |: choice, one or the other, but not both
-->

<!ENTITY % term "termId, termName, termQualifier?, termType?, termLanguage?”>
<!ENTITY % admin "termCreatedDate?, termCreatedBy?, termModifiedDate?, termModifiedBy?”>

<!ELEMENT Zthes (%term;, termNote?, %admin;,relation*)>

<!ELEMENT relation (relationType, sourceDb?, %term;)

<!-- This appendix should include a crosswalk with any pre-existing Thesaurus DTDs if appropriate. -->
Appendix B.2. Sample Zthes-in-XML Document
This document was supplied by Thomas Place.

```xml
<?XML version="1.0" ?>
<!DOCTYPE Zthes SYSTEM "zthes.dtd">
<Zthes>
  <termId>102067</termId>
  <termName>video art</termName>
  <termType>PT</termType>
  <termNote>
    Use for works of art that employ video technology, especially videotapes. For the study and practice of the art of producing such works, use "video."
  </termNote>
  <relation>
    <relationType>UF</relationType>
    <termId>102067/001</termId>
    <termName>art, video</termName>
    <termType>ND</termType>
  </relation>
  <relation>
    <relationType>BT</relationType>
    <termId>185191</termId>
    <termName>[time-based works]</termName>
    <termType>NL</termType>
  </relation>
  <relation>
    <relationType>RT</relationType>
    <termId>54153</termId>
    <termName>video</termName>
    <termType>PT</termType>
  </relation>
  <relation>
    <relationType>RT</relationType>
    <termId>253827</termId>
    <termName>video artists</termName>
    <termType>PT</termType>
  </relation>
</Zthes>
```