Standards for KOS
Present status
and future requirements

NKOS ECDL 2003

Dagobert Soergel
College of Information Studies
University of Maryland

Outline

- Functions of standards
- The many forms of Knowledge Organization Systems (KOS) and their standards
  - A critique of ANSI Z39.19
- Future requirements:
  An integrated comprehensive standard,
  or family of standards,
  for all types of KOS

Functions of standards

Facilitate the following functions across any type of KOS

1 **Input of KOS data** into programs / **Transfer of data** from one program to another

2 **Accessing KOS for applications**. Includes querying KOS and viewing results (for example, using Z39.50)

3 Identifying specific terms/concepts in specific KOS

4 Prescribing or giving guidance on **good practices**

Functions of standards 1

1 **Input of KOS data** into programs / **Transfer of data** from one program to another (across any type of KOS)

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 program that uses a KOS **for display, browse, search, etc.**

Functions of standards 2

2 **Accessing any type of 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.

Functions of standards 3

3 **Identifying specific terms/concepts in specific KOS of any type**

This requires rules for URIs that uniquely identify specific term/concept records in specific KOS. Needs a name resolution service (such as a KOS registry)

3.1 **Links from one KOS to another**

3.2 **Links from index terms/concepts** in the metadata for an object, or any other reference to a term/concept in a text/object

Functions of standards 4

4 Prescribing or giving guidance on **good practices** in constructing a KOS

KOS construction practices should be given in a standard **only if this is the only way to guarantee properties to be standardized**

The many forms of (KOS) and their standards

- Dictionaries, glossaries
- Thesauri
- Topic maps
- Concept maps
- Classification schemes
- Ontologies
- Generic standards for knowledge structures, entity-relationship models
- Many ISO terminology-related standards

The many forms of (KOS) and their standards 1

- **Dictionaries, glossaries**

  ISO 12200:1999, Computer applications in terminology--Machine Readable Terminology Interchange Format (MARTIF)--Negotiated Interchange

  ISO 12620:1999, Computer applications in terminology--Data Categories.

- **Thesauri**

  ISO 2788-1986(E) / ANSI/NISO Z39.19-1993(R1998) (www.niso.org)

  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)

The many forms of (KOS) and their standards 2

- **Topic maps** (reference works, encyclopedias) (http://www.topicmaps.org/about.html)

  ISO/IEC 13250:2000 Topic Maps

  XML Topic Maps (XTM) 1.0 (http://www.topicmaps.org/xtm/1.0/)

- **Concept maps**

- **Classification schemes**

- USMARCformat for classification data http://lcweb.loc.gov/marc/classification/eccdhome.html

The many forms of (KOS) and their standards 3

- **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)

  DARPA Agent Markup Language (DAML) / 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/)

  The Topic map standard belongs here as well

- **Many ISO terminology standards**

The many functions of KOS

- Provide a semantic road map
- Improve communication generally. Support learning and assimilating information.
  - Support learning and the development of instructional materials through conceptual frameworks. .
  - Assist readers in understanding text Assist writers
  - Support foreign language learning.
- Provide the conceptual basis for the design of good research and implementation.
  - Assist researchers and practitioners with problem clarification
  - Consistent data collection, compilation of statistics
- Provide classification for action. Classification for social and political purposes
  - a classification of diseases for diagnosis,
  - of medical procedures for insurance billing,
  - of commodities for customs.

The many functions of KOS 2

- Support information retrieval and analysis. Organizing and keeping track of goods and services for commerce (esp. ecommerce) and inventory
- Support meaningful, well-structured display of information
- Ontology for data element definition. Data element dictionary
- Conceptual basis for knowledge-based systems.
- Do all this across multiple languages
- Mono-, bi-, or multilingual dictionary for human use
- Dictionary/knowledge base for automated language processing

A critique of ANSI Z39.19

1. Is rooted in the sixties and in the print world
2. Takes a very limited, if not myopic, view of thesauri, never mind dealing with other types of KOS
3. Operates on the level of terms and not on the level of concepts, treating conceptual problems from a mostly linguistic perspective, creating a muddle
4. Contradicts itself
5. Has rules that are based on formalistic considerations rather than on usefulness for retrieval and other functions

Critique of ANSI Z39.19 cont.

6 Does not promote faceted structure or meaningful hierarchical structure that provide an overview of a domain

7 gives priority to a very small set of relationship types

8 Contains many rules that may have made sense at one time but do not make sense now

9 Contains many misconceptions stemming from insufficient understanding of the problem

10 Acts as a textbook (and a poor textbook at that) by describing thesaurus construction procedures when this is not necessary to define a product

11 Does not specify data formats

Limited view of thesauri

- “A thesaurus, for purposes of this standard, is a controlled vocabulary of terms in natural language that is designed for postcoordination. ... generally employed in indexing.”
- Document-oriented indexing through “human intellectual decisions” ...“assignment indexing by humans (as opposed to derivative indexing by machines)” [1]

Operates on the level of terms and not on the level of concepts, creating a muddle

- Descriptors should represent single concepts, expressed by a single word or a by a multiword term [4.1.2]

  yet

- “The factors enumerated below may be considered in deciding which multiword terms should be split into separate descriptors and which should be retained in compound form [4.1.3] and
- 4.3 Criteria when compound terms should be split
- Many more examples could be given,
  see also next slide

Contradicts itself

- Thesaurus is by the standard's definition intended for postcoordination (better called postcombination)
- "Descriptors should represent single concepts" [4.1.2]
- Yet admits precoordinated descriptors (better called precombined descriptors):
  "Compound terms increase the number of descriptors in the thesaurus" (4.1.3b)

rules based on formalistic considerations rather than on usefulness for retrieval

"To be acceptable as a descriptor, a compound term should express a single concept or unit of thought, capable of being arranged in a genus-species relationship within a hierarchy or tree structure." [4.1]

Allows

adopted children

educational television

But not

*children and television (*does not fit in isa hierarchy)

Unnecessarily rules out verbs and adverbs and discourages adjectives (imagine a thesaurus of business functions without verbs)

4.3.2.1 Focus and difference

a) A Compound term should be split when its focus refers to a property or part and its difference represents the whole or possessor of that property or part

Examples

hospital personnel = hospital + personnel

soil acidity = soil + acidity

Conversely, a Compound term should not be split when the focus term refers to a whole and the difference is a term for its part or property

Examples

acid soils

skilled personnel

Note: This contradicts 3.4.2.1, which explicitly allows to use *mobile* as a descriptor which in indexing can be used in conjunction with the descriptor *homes* to index a document on *mobile homes*

Contains many rules that may have made sense at one time but do not make sense now

For example, complex rules on when to use singular and when to use plural

Just use dictionary conventions (use singular) unless meaning or general usage require plural

In any event, with more forgiving query input, the significance of such rules is much diminished

Future requirements:

An integrated comprehensive standard,
or family of standards,
for all types of KOS

Future requirements

- A new standard should support **access to and data exchange between all kinds of KOS**
- To this end, it should specify the **representation of all kinds of information about concepts and terms and the relationships among and between them**, drawing on and unifying the various standards that exist
  - It should specify a general framework for data formats so that systems can exchange data on a formal level
  - It should codify a common understanding on the semantic level (content and interpretation of the data) to the extent possible and specify a method for each system to describe its semantics when it deviates from or extends the common standard
- It should give **guidelines for various forms of external representation in print and online** for the benefit of users

Future requirements

- n
  - H

Example
Treatment of relationships

- Specify a common set of relationships
- Specify a method for
  - describing deviations from these definitions;
  - defining new relationships (including a description of how they fit into the common set).
    The standard would specify minimal elements a relationship definition contain, including rules a system could use to check for constraints the relationship must fulfill.
    Where possible, definition of semantics should be interpretable by computer programs, but definitions that require human interpretations are preferable to no definitions.
- Set up relationship inventory where relationship definitions could be entered easily.

Take-home message

Make a new standard that

- moves the field forward
  rather than holding it back.

- brings separate communities together for mutual benefit
  rather then solidifying existing divisions.

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H

S

- n
  - H