Revision and extension of thesaurus standards

Stella G Dextre Clarke

Convenor, IDT/2/2 Working Group

Existing thesaurus standards

- ISO 2788-1986 Guidelines for the establishment and development of monolingual thesauri

  = BS 5723:1987

- ISO 5964-1985 Guidelines for the establishment and development of multilingual thesauri

  = BS 6723:1985

- ANSI/NISO Z39.19-1993 Guidelines for the construction, format and management of monolingual thesauri [Not discussed further in this presentation, which is limited to BS/ISO]

Procedural aspects

- Constituting a working group for an international standard takes a while
- It seemed easier to work first on the British Standards; then offer these up to the international community
- The BS committee must be UK-based, but invites international input wherever possible without offending BSI/ISO

Procedure in detail

- Small UK Working Group of unpaid volunteers submits drafts to BSI committee IDT/2/2
- Comments are invited informally from everyone interested
- Progress is very slow... but worthwhile, we hope

Outline of new standard

- BS 8723: Structured vocabularies for information retrieval – Guide
- Part 1: General
- Part 2: Thesauri
- Part 3: Vocabularies other than thesauri
- Part 4: Interoperability between vocabularies
- Part 5: Interoperability with applications

Progress to date

- Parts 1 and 2 soon to be published
- Draft chapters of Part 3 exist, including Classification schemes, Taxonomies, Subject heading lists, Ontologies, Search thesauri
- Draft of Part 4 well advanced
- Part 5 to follow the rest

New features in Part 2 (monolingual thesauri)

- Clearer guidance on applying facet analysis to thesauri
- Some changes to the ‘rules’ for compound terms
- More guidance on managing thesaurus development and maintenance
- Functional specification for software to manage thesauri
- General overhaul in all areas

Part 4: Interoperability between vocabularies

- Huge demand for accessing information that has been indexed with another language and/or vocabulary. The buzzword is ‘Mapping’. The Semantic Web is just one application.
- Part 4 to include multilingual thesauri as a special case of mapping between vocabularies
- But how do you map between precoordinate and postcoordinate schemes?
- How do you provide for mapping when notations are not all enumerated, but built up by rules for synthesis?

Part 3 chapters

- Classification schemes
- Subject heading lists
- Taxonomies
- Ontologies
- Search thesauri

Part 5: Interoperability with applications

- Vocabularies must work with
  - Search engines
  - Content Management Systems
  - Web publishing software, etc.
- Build on existing formats and protocols for data exchange
- e.g. Z39.50 and Zthes, XML schema? DTD? MARC? SKOS Core Schema? Topic Map? ADL gazetteer protocol? Anything else?

The Modelling Challenge

- For Part 5, we have been assuming that the formats and protocols will ‘fall out’ straightforwardly from the models established in Parts 2, 3 and 4
- But there may be difficulties accommodating the flexible options described in Parts 2-4

The model derives from the definition:

A thesaurus is a controlled vocabulary in which concepts are represented by preferred terms, formally organized so that paradigmatic relationships between the concepts are made explicit, and the preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms. The purpose of a thesaurus is to guide both the indexer and the searcher to select the same preferred term or combination of preferred terms to represent a given subject.

The model derives from the definition:

A thesaurus is a controlled vocabulary in which concepts are represented by preferred terms, formally organized so that paradigmatic relationships between the concepts are made explicit, and the preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms. The purpose of a thesaurus is to guide both the indexer and the searcher to select the same preferred term or combination of preferred terms to represent a given subject.

Modelling challenge from a personal perspective

- Model must allow for 3 ‘levels’ of complexity
  - Term level
  - Concept level
  - Display level
- The description must be understandable to XML, UML and AI communities, data modellers as well as thesaurus editors, who often use words like ‘element’, ‘entity’, ‘level’, ‘object’, etc. in different ways

Want to get involved?

- If you would like to contribute to the next stages, please get in touch!
- Contact SDClarke@LukeHouse.demon.co.uk