Towards guidelines for mapping to a classification scheme

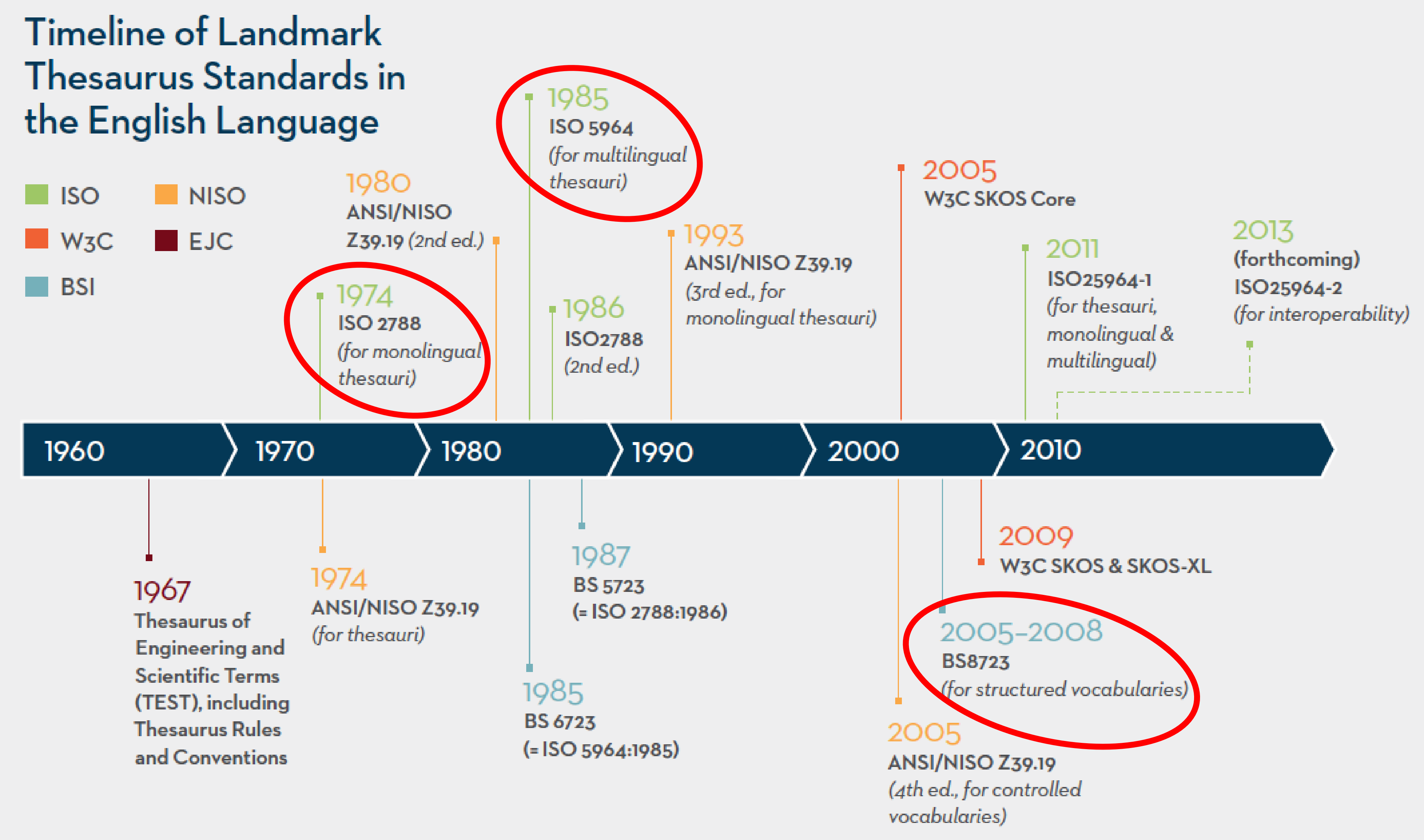
Stella G Dextre Clarke

Convenor, ISO TC46/SC9 WG8

and Vice-President, ISKO

Overview

- Adopting ISO 25964 – some health warnings
- Practical issues for mapping projects
- Some clear principles
- Some questions to explore



Timeline of Landmark Thesaurus Standards in the English Language

🟩 ISO
🟧 NISO
🟥 W3C
■ EJC
🟦 BSI

1967
Thesaurus of Engineering and Scientific Terms (TEST), including Thesaurus Rules and Conventions

1974
ISO 2788
(for monolingual thesauri)

1974
ANSI/NISO Z39.19
(for thesauri)

1980
ANSI/NISO Z39.19 *(2nd ed.)*

1985
ISO 5964
(for multilingual thesauri)

1985
BS 6723
(= ISO 5964:1985)

1986
ISO2788
(2nd ed.)

1987
BS 5723
(= ISO 2788:1986)

1993
ANSI/NISO Z39.19
(3rd ed., for monolingual thesauri)

2005
W3C SKOS Core

2005
ANSI/NISO Z39.19
(4th ed., for controlled vocabularies)

2005–2008
BS8723
(for structured vocabularies)

2009
W3C SKOS & SKOS-XL

2011
ISO25964-1
(for thesauri, monolingual & multilingual)

2013
(forthcoming)
ISO25964-2
(for interoperability)

Dextre Clarke and Zeng, 2012. http://www.niso.org/publications/isq/2012/v24no1/clarke/

ISO 25964 health warnings

- Focus is on thesauri (not other KOS types)
- Post-coordinate mind-set
- Guidelines not mandatory rules
- Coverage includes mapping between thesauri and classification schemes (not between Subject Heading Schemes and classification schemes)
- Real-world KOSs are often hybrids or variants (not conforming with the distinct KOS types delineated). Even those named “Thesaurus” often don’t comply!
- A lot of the content of Part 2 is untested (including tags/symbols)…
- …but feedback so far is positive (e.g. from MACS)

Health warnings (continued)

- Context is limited to Information retrieval (Boolean logic assumed), subdividing into just 4 scenarios:
  - Conversion of search queries
    - (1) when mapping thesaurus to other KOS
    - (2) when mapping other KOS to thesaurus
  - Conversion of metadata (index terms or codes)
    - (3) when mapping thesaurus to other KOS
    - (4) when mapping other KOS to thesaurus
  - Not much consideration of modern IR techniques e.g. statistical methods, latent semantic indexing, collaborative filtering, etc.
- Among those contexts, (3) - mapping thesaurus to classification scheme, for the purpose of converting index terms in metadata - was not recommended.
- Why not?

Examples to test conversion of index terms to class codes

- Example 1. Document is indexed with terms: **adventure trails; winter; cycles; maintenance**.

Should the class code represent **maintenance of adventure trails in winter for the use of cyclists**? Or the **maintenance of bicycles and motorbikes for rough conditions in winter**?

- Example 2. Document is indexed with terms: **antipsychotic medication, older people, care homes, dementia, residential care, nursing homes, aggression, behaviour problems, drug prescription, research**. Is the emphasis on behaviour problems? Or on medication? Or what?
- Moral: you can’t build a classmark from index terms alone. We’re not mapping like to like.
- But maybe it’s helpful to provide the components from which a classmark is built?

Finding general principles is hard!

- The big snag is moving from post-coordination to pre-coordination
- Post-coordinate index terms arise from analysis – the isolation of discrete concepts; whereas a classmark comes from synthesis – in which the concepts are combined according to how they occur in a specific context (query or document).
- Whereas most thesaurus terms are known generally enough to be accepted in normal discourse, classes in a classification scheme tend to be tailor-made for particular contexts. The coordinations within them are often “syntagmatic” rather than “paradigmatic”. Classes apply to whole documents not concepts within them.

Practical issues for mapping projects. 1: when setting up

- Pros/cons of following a standard
- Vital to agree objectives from the start
  - Definition of mapping
  - Spell out the context(s) = use case(s) e.g. which KOSs involved, when/how/where the mappings will be used, will there be human mediation, will the humans be trained, etc.
- Choose the right people for the job, and brief them thoroughly re context
- Select/refine mapping types

Practical issues for mapping projects. 2: communication formats

- SKOS should be used at the stage of publishing to the Web but cannot handle some mapping types, especially compound mappings
- Another format should be used for working with and storing the full range of mapping types, before conversion to SKOS.
- Data models for source and target vocabularies, to avoid misunderstandings with technical colleagues

Practical issues for mapping projects. 3: context issues

- A thesaurus works best in a narrow domain; the same may be true of other KOS, and is certainly true of mappings
- Never forget that some thesauri (and other KOS) are badly constructed

A few clear principles

- Exact equivalence is the ideal; can be used two-way and in fully automated situations
- Therefore use of the Exact marker (=) is worthwhile
- Intelligent mediation is advisable in the interpretation of all mappings except exact equivalence
- A caption alone is inadequate to represent or convey the scope of a class. Scope notes, superordinate/subordinate classes must be checked.
- To derive a class code for a document, mapping from assigned index terms alone is not enough
- See also guidelines in handout (ISO 25964-2 clause 13.2)

Some questions to explore

- Which mapping types?
- How, where and when to parse a class code and map to/from its components?
- Mapping to/from auxiliary tables: where/when?
- What is the role of Dewey index entries when mapping to/from classes?
- Use cases for mapping **to** DDC versus use cases for mapping **from** DDC
- Representing a class by URI rather than by notation?
- To what extent can a thesaurus concept ever be equivalent to a class in a classification scheme?

Which mapping types to use; whether/how to adapt them?

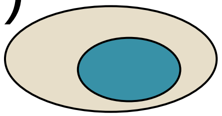
- Equivalence
  - Exact
  - Inexact
  - “unmarked”
  - Compound
- Hierarchical
  - Broader
  - Narrower
- Associative

About “equivalence”

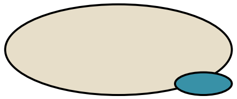
- When equivalence is Exact, does this imply that use of the class code will retrieve All and Only the items deemed relevant to the corresponding concept?
- If “Yes”, then a thesaurus concept will never find an exact equivalent in a classification scheme. Maybe a new type of equivalence is needed: “partial equivalence”?
- If “No”, =EQ could be useful.
- But are the items in auxiliary tables eligible?

The meaning of “hierarchy”?

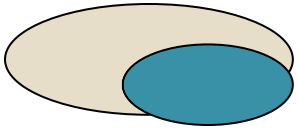
- ISO 25964 restricts BT/NT usage to logical hierarchies (generic; whole/part; instantial)



- True ontologies are even stricter – and the SW intends to use true ontologies for inferencing
- Classification schemes commonly use “display hierarchies” – organised for user navigation not logic



- Even if the overlap is considerable, it’s not a broadMatch or narrowMatch in SKOS



What about using RM (relatedMatch) for all mappings?

- May save time/effort at the stage of developing mappings
- May cost time/effort at the stage of implementing mappings
- Could be a safer option if a single vocabulary has been used with different indexing rules for different resources
- Your decision depends on context: how/where/when will your mappings be used? And what size is your budget?

And the final Guidelines?

- **Over to you!**