

Introduction to Thesauri and the UK Archival Thesaurus

Peter Garrod
School of Oriental and African Studies
(pg7@soas.ac.uk)

Overview

- What is a thesaurus
- Why use a thesaurus? Thesauri vs natural language searching
- Thesaurus standards
- Thesaurus concepts
- Background to UKAT
- How UKAT works

What is a thesaurus?

- A type of *controlled vocabulary*
- A standard for development of indexes, card catalogues and other subject-based retrieval tools
- Improves effectiveness of retrieval tools by providing:
 - A consistent, restricted terminology
 - Rules for establishing relationships between terms

Why use a thesaurus?

Strengths of thesaurus-based searching (vs natural language searching):

1. Higher precision through avoidance of false hits
2. Eases burden of searching by guiding user to:
 - Equivalent terms (terms expressing the same or a similar concept): e.g. students ≠ pupils
 - Related terms: more general, more specific or associated terms – e.g. schools → education

Why use a thesaurus?

Strengths of thesaurus-based searching:

3. Less ambiguity - if thesaurus uses scope notes to explain the meaning of terms
4. Brings out concepts that may not be explicitly expressed in the text (e.g. blacksmith's apprenticeship indenture → iron working)

Why use a thesaurus?

Strengths of natural language searching:

1. Higher precision when searching for specific concepts or concepts with little variation in expression (e.g. names)
2. Exhaustivity: potentially more hits (but also more false hits)
3. No maintenance overhead - use of existing textual data

Why use a thesaurus?

My view ...

Ideally, you should provide both natural language and controlled vocabulary searching of your catalogue data ... in a way that makes the differences between them largely transparent to the user

Thesaurus standards

- BS 5723:1987 *British Standard Guide to Establishment and Development of Monolingual Thesauri* (1987) – undergoing revision (to be reissued as BS 8723)
- BS 6723:1985 *Guidelines for the Establishment and Development of Multilingual Thesauri* (1985)
- ANSI/NISO Z39.19 (available on web)
- Jean Aitchison, Alan Gilchrist and David Bawden, *Thesaurus Construction and Use: A Practical Manual*, 4th ed (2000)

Other sources

- UKAT website: 'Thesaurus basics' page: www.ukat.org.uk/help/
- Willpower Information (Leonard Will): www.willpowerinfo.co.uk

Thesaurus standards

Standards provide:

- Guidance for selecting and formulating terms
- Rules for establishing/expressing relationships
- Guidance on display/presentation of thesauri

Thesaurus concepts - Selection of terms

- Decide scope of thesaurus – general or specialized?
- If specialized – which areas are core subjects and which are peripheral?
- Level of detail – e.g. whether to include *classes of one* (specific instances of general concepts)?
 - UKAT excludes personal/place/corporate names and terms relevant to only one repository – but includes other ‘classes of one’

Thesaurus concepts - Formulation of terms

- Use plural for concrete entities which are *count nouns* ('how many?'): e.g. books, archives
- Use singular for:
 1. Concrete entities which are *non-count nouns* ('how much?'): e.g. Beer, Soil
 2. Abstract concepts: e.g. Love, Marxism
 3. Unique entities (classes of one): e.g. Ibuprofen, Microsoft Windows

Formulation of terms

- Use qualifiers to distinguish *homographs*: e.g. Cells (biology) and Cells (electric)
- *Scope notes* should *always* be used to provide a definition if the meaning of a term is unclear or restricted
 - E.g. *Teaching materials* SN “Any printed and/or non-printed material designed specifically for or used in instruction. Use more specific descriptor where appropriate” (*UNESCO Thesaurus*)

Formulation of terms

Compound terms can be included in thesauri if:

1. The term is in common usage (e.g. Women's rights)
2. The term is a proper noun: e.g. Salvation Army
3. The term expresses a concept which is more than the sum of its parts, or which cannot be factored into its components without loss of meaning
 - e.g. Group therapy, Electrical engineering, White elephants

Formulation of terms

- Otherwise, compound terms should be factored into their components:
 - E.g. ‘Women teachers’: factor into ‘Women’ + ‘Teachers’
 - ‘Boiler explosions’: factor into ‘Boilers’ + ‘Explosions’
- Aim is to avoid *pre-coordination*: wherever possible, thesaurus terms should represent discrete concepts

Thesaurus relationships

Relationships between terms - three basic types:

- Equivalence relationship
- Hierarchical relationship
- Associative relationship

Equivalence relationship

- One term chosen to represent a concept - the *descriptor/preferred term*
- Other terms that express the same concept become *non-descriptors/non-preferred terms*
- Users are referred from the non-descriptors to the descriptors
- When indexing - descriptors become indexing terms, non-descriptors added as guide terms

Equivalence relationship

Thesaurus conventions USE and UF:

- Boats USE Ships
- Non-descriptor → Descriptor
- Ships UF Boats
- Descriptor → Non-descriptor
- Descriptors are not inherently ‘better’ than non-descriptors

Equivalence relationship

Non-descriptors:

- *Synonyms*: e.g. Railways/Railroads
- *Quasi-synonyms*
 - Not true synonyms but treated as synonymous for purposes of thesaurus
 - E.g. Car parks/Parking spaces, Cities/Urban areas

Equivalence relationship

Upward posting:

- Technique for reducing size of thesaurus
- Narrower terms treated as equivalent to their broader term
- E.g. Incunabula, Antiquarian books USE Rare books
- Should only be used in peripheral subject areas

Hierarchical relationship

Relationship between *Broader terms* (BT) and *Narrower terms* (NT)

- E.g. Museums – broader term; National museums, Local museums – narrower terms
- Recursive: if A is NT of B, and B is NT of C, then A is NT of C (child-parent-grandparent etc)
- Levels can be indicated in thesaurus notation (BT1, BT2, NT1, NT2 etc) – e.g. UNESCO

Hierarchical relationship

Rules for hierarchical relationships:

- BT and NT must be of the same fundamental category (e.g. entities, activities, agents or properties) *and*
- *Generic relationship* holds: NT is a subclass/subtype of BT – e.g. Rodents (BT), Squirrels (NT), *or*
- *Instance relationship* holds: NT is a specific instance of BT: e.g. Seas (BT), Baltic Sea (NT)
- BT-NT relationship can also be thought of as an *IS A* relationship: e.g. squirrel *is a* rodent, Baltic Sea *is a* sea

Hierarchical relationship

Invalid hierarchical relationships:

- *Whole-part (or has-a) relationship*: where one concept is a constituent part of another: e.g. Cars, Car engines
- *Exceptions: systems/organs of bodies, geographical locations, disciplines/fields of study, social structures*
 - For these, *Hierarchical whole-part relationship* is valid: e.g. Ear BT, Inner ear NT

Hierarchical relationship

Monohierarchies vs polyhierarchies

- Monohierarchical thesaurus: where NT can have only one immediate BT (e.g. UNESCO)
- Polyhierarchical thesaurus: NT can have one or more immediate BTs (e.g. UKAT)
 - E.g. in UKAT ‘Slave emancipation’ has BTs ‘Social reform’ and ‘Civil and political rights’

Hierarchical relationship

Microthesauri and top terms

- *Top terms*: descriptors which are at the uppermost level of thesaurus (no BTs)
- *Microthesauri* (MT): used in UNESCO & UKAT to group descriptors under general themes, e.g. 'Legal systems'
 - Not descriptors (BUT can be identical to descriptors!)
 - Organized into *Areas of knowledge* (7 in UNESCO)

Associative relationship

- Related term (RT) relationship
- Established between conceptually related terms which are not equivalent and not related hierarchically
- E.g. Cars RT Automobile engines
- Reciprocal: if A is RT of B, B is RT of A
- Dependent on scope and context of thesaurus
- Danger of overloading

Presentation of thesauri

Most thesauri provide:

- Alphabetical listings of descriptors (with relationships) and non-descriptors
- Hierarchical listings by top terms, microthesauri, themes etc
- Searching facility
- Should indexes based on thesauri mimic how thesauri are displayed?

Thesauri vs other controlled vocabularies

Narrowing gap:

- Faceted classification schemes (e.g. Bliss, Colon classification) can be used to build thesaurus hierarchies
- Library of Congress Subject Headings: adopting thesaurus-like relationships

Thesauri vs other controlled vocabularies

Differences:

- LCSH: subject headings combined by indexer to produce *pre-coordinated* strings
 - e.g. Baptists – Wales – Glamorganshire - Records and correspondence
- Thesauri are *post-coordinated*: index using discrete concepts which are combined by the user at the point of searching (e.g. using Boolean and/or/not)

What is UKAT?

- Subject thesaurus created for the archive sector, to
 - Promote more consistent subject indexing
 - Promote more effective subject searching
 - Support the discovery of resources relating to under-represented groups
- Created June 2003 - August 2004
- Lead partners: ULCC and TNA, HLF funding

What is UKAT?

- Website: www.ukat.org.uk
- Searching
- A-Z browsing
- Hierarchical browsing
- Download files in Calm, SKOS-Core

UKAT's approach

- *UNESCO Thesaurus* used as basis - excluding countries
- Integrated terms submitted by repositories and projects
- Added polyhierarchies
- Sought terms relevant to inclusion objectives

UKAT content

- 19,700 terms (descriptors and non-descriptors)
- 6360 terms inherited from UNESCO
- 13,340 contributed by archives and projects
- 700 terms rejected

UKAT content

Sources of terms:

- Projects: A2A, Archives Hub, Archives Network Wales, AIM25, Baillie, CASBASH, GASHE, Mundus, Rebuilding the City, West Sussex Picture the Past
- Nationals: BBC, UK National Archives, NDAD

UKAT content

Sources of terms:

- Local/regional archives: Cumbria, Essex, Gloucestershire, Warwickshire, Corporation of London Guildhall Library, South East Film and Video Archive
- Others: Britten-Pears Library (Cecilia), Hallé Orchestra, Modern Records Centre (University of Warwick), Salidaa, HEREIN, Government Category List

UKAT content

Contributions by sector:

- HE institutions and projects: 39%
- Local authority sector: 23%
- National institutions and projects: 27%
- Others: 11%
- Web submissions: 1.5%

What we learnt

- Had to be proactive in seeking out terms
- Question of UKAT's scope – e.g. whether to include names, document types

What we learnt

- Incorporation of existing terms vs building from scratch - what if the terms don't exist?
- How to reach out to non-archivists?
- Issue of sustainability

UKAT

UK Archival Thesaurus

Future of UKAT

- Website still hosted by ULCC
- Limited editing
- Can email support@ukat.org.uk
- Incorporation into Linking Arms