Improving information retrieval through metadata tagging

Jonathan Engel, Information Architect

Date 25/09/15
Introduction

• Organising and tagging content using standard structures will always enhance navigation, search and information delivery

• This consistent structuring -- or Information Architecture -- can be simple or comprehensive

• Either way, it should seek to unite the often separate realms of structured and unstructured information

“Anybody seen a needle?”

Copyright TFPL 2015
Introduction - Aims and objectives

• Review components and benefits of faceted taxonomies
• See how to extend taxonomy structure with synonyms, related terms and keywords (using thesaurus, ontology and folksonomy)
• Review standards supporting an extended taxonomy and alignment with data models
• Consider combining components of a controlled vocabulary to improve classification -- and thus information navigation, searching and delivery
• Look at three case studies
Introduction -- Background

- Reuters Manager for Multimedia News Production (15 years at company)
- Running own information management consultancy for last 12 years – InfoArk
- Recently Senior Taxonomy Manager at EY and Information Architect at O2
- Designer of customised taxonomies and related metadata for classifying content
- Specialist in linking classification schemes with automated tagging and search software, plus content filters, for targeted client delivery
- Associate at consultants such as TFPL
Introduction – Selected clients

- Cambridge University
- Clifford Chance law firm
- Unilever
- Institute of Chartered Accountants
- Dow Jones newswires online
- Times and Sunday Times online
- Shop Direct Group (Littlewoods, Very)
- UK Care Quality Commission
- NHS Education for Scotland
- UK Department for International Development
- Oxfam International
- Barnardo’s
Phase one – Reviewing benefits of multi-faceted taxonomies
What is Information Architecture?

- Information Architecture is the accurate, logical and consistent description of enterprise content and data to improve information retrieval and delivery
- It helps get the right information to the right people at the right time
- The key structural component of this architecture is metadata, or information about information...a concise summary of the knowledge assets
Problems IA can address

- Ineffective search
- Poor website navigation
- Inability to retrieve useful, related information
- Content “silos” and “Wild West” folder structures
- Chaos of personal classification systems
- Obstacles to personalised information delivery
- Plague of duplicate and outdated content
- Inconsistent, incomparable data fields and values
- Burden of creating and maintaining custom web page content
- Limitations of purely technical solutions
What’s coming -- key points

• Types and uses of metadata
• How a key metadata element -- a multi-faceted taxonomy -- lets you classify content in various ways, providing multiple paths into the “information pool”
• Combining a “top down,” “bottom up” strategy helps align detailed topics under relevant higher-level headings
• A “building block” approach for combining terms from each facet keeps the taxonomy lean and flexible
What has metadata ever done for us?

- Metadata is a concise summary of the full contents
- A restaurant menu is a good example – diners can see what the kitchen produces without going in themselves!

- Metadata tags improve information retrieval and delivery by matching appropriate users to relevant content
- Users navigate hierarchies, generate search queries or select delivery filters reflecting these metadata values – and thereby find information
Metadata types and management

Metadata types

- Administrative metadata -- often system-generated and includes Creator, Date, Publisher. The Dublin Core metadata designed for online use is mostly administrative.

- Descriptive metadata -- tells users what content is about or who needs it.

- Structural metadata – defines chapters, series or media packages.

Metadata management

- Controlled vocabularies – someone is in charge, defining preferred terms, synonyms and related terms.

- Folksonomies – no one is in charge! Users are free to tag content with personally meaningful key words.
Metadata structures and their key attributes

- **Taxonomy**: Hierarchical – logically nested topics for navigating
- **Thesaurus**: Equivalence – adds synonyms for improved searching
- **Ontology**: Associative -- with defined relationships for navigating, searching and filtering
Taxonomy is described sometimes as a science and sometimes as an art, but really it's a battleground.

*Bill Bryson*

*from "A Short History of Nearly Everything"*
Taxonomy -- at the heart of all IT systems
Questions that define high-level taxonomy facets – “top down”

• Who and where are we?
• Who do we work with?
• What do we do?
• How do we do it?
• Who do we do it for?
Multi-faceted taxonomy goes beyond “subjects”

- Geography for location, jurisdiction
- Organisation’s business units
- External organisations by type, size
- List of statutes, products, etc.

- Business activities and issues
- Business sectors, e.g. financial services

- Events, projects and initiatives
- Information types, access and focus
- Language
- Roles and stakeholders

Entities

Who? Where?

Subject matter

What?

Focused filters

How? Who for?
Where to find content topics – earlier attempts help with “bottom up”

- Business goals and processes
- Database fields, e.g. customers and contacts (CRM)
- Paper filing system/library
- Electronic file directory
- Library catalogue
- Intranet display categories
- Project plans, e.g. document types
- Departmental locations, staff roles
Building block approach

When combining the “top down,” “bottom up” strategy, topics can be viewed as building blocks, like Lego bricks. They can be stacked, mixed, matched and joined together to describe multiple facets of entities, subjects or focused filters. Take the following example:

Western Europe

Europe

U.K.

Taxation

Personal taxation

Inheritance tax
Building block approach -- inheritance

- The dual description from the combined blocks includes information “inherited” from the blocks above.
- This inheritance between “parent” topics and the lower “child” terms means any content about “Inheritance tax” would also be about the parent term of “Personal taxation”.
Building block approach – 2 issues

• Mono-hierarchical v Poly-hierarchical structure – one taxonomy location or many?

  Suggestion – be decisive and avoid tangential metadata paths that dilute meaning

• Dual v Compound descriptors – does “France” + “Sport” produce similar results to “French sports?”

  Suggestion -- Reserve compound terms where precision required on large document volumes
Phase two – Extending taxonomies with synonyms, related terms and keywords (and supporting standards)
What’s coming -- key points

• How to extend taxonomy “horizontally” with elements that enrich description of preferred term

• The roles played by these additional elements -- synonyms, related terms and contextual keywords

• The structures that deliver this added meaning – thesaurus and ontology

• How unstructured metadata helps out – folksonomy

• Useful standards for structured vocabularies (and alignment with data models)
Combine structures for “extended” taxonomy

- **Taxonomy**: Hierarchical – logically nested topics for navigating
- **Thesaurus**: Equivalence – adds synonyms for improved searching
- **Ontology**: Associative -- with defined relationships for navigating, searching and filtering
What’s an extended taxonomy look like?
Components of an extended taxonomy

- Preferred term
- Synonyms
- Hierarchical parent
- Related term
- Contextual keywords
Extended taxonomy term – examples

- **Preferred term**
  - Law
  - Law and crime

- **Synonyms**
  - Legal system
  - Legislation
  - Statutes

- **Hierarchical parent**
  - Law and crime

- **Related term**
  - Police
  - Legislature
  - Politics

- **Contextual keywords**
  - Lawmaker, MP,
  - White paper,
  - Vote

*tfpl* a Progility company
Folksonomies (or contextual keywords)

- Supplements, not replaces, a controlled vocabulary
- Helps users see immediate benefits of metadata, e.g. with better search results on terms they contributed
- Improved search also provides incentives for contributing up-to-date terms for inclusion in controlled list
- Administrators can modulate search for folksonomy key words, so primacy and performance of controlled search terms are not threatened
Useful standards for structured vocabularies

- British Standard 8723 (ISO 25964) – models 3 types of relationships – hierarchical, equivalent, associative
- Simple Knowledge Organisation Systems (SKOS) standard -- the Resource Description Framework (RDF) -- developed by the World Wide Web consortium (W3C)
- W3C’s Web Ontology Language (OWL) for describing relationships for the semantic web
Resource Description Framework supports Linked and Open Data

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity</td>
<td>Property / Attribute</td>
<td>Value</td>
</tr>
<tr>
<td>Book</td>
<td>Title Author</td>
<td>30 Minute Meals, Jamie Oliver</td>
</tr>
</tbody>
</table>

Official RDF terms
Meaningful equivalents
Examples
Data models use similar structure – whether conceptual, logical or physical

- **Common elements** – Entities, Attributes and Values, e.g. Person and Date of Birth or Product and Serial number
- **Conceptual** -- for high-level business requirements; model connects Entities and defines their Attributes
- **Logical** – Describes relationships among Entities by type and number, e.g. between Person and Organisation relationship is “member of,” while number (aka cardinality) is “one to many”
- **Physical** – defines data types and formats within specific applications and systems; information is held in a “data dictionary”
Phase three – uses for extended taxonomies, plus three case studies
What’s coming -- key points

• The extended structure supports automated content classification, improved navigation and search

• It also allows persona-based information delivery and document lifecycle management

• An extended taxonomy can also be aligned with high-level, conceptual data models
Problem – search techniques “outperform” most users

Preferred search terms, synonyms, related topics

Proper Boolean logic (AND, OR, (brackets))

Correct search syntax ([QLRegion Title], etc.)
Solution -- standard combination rule

• **Title test:**
  Preferred term and Synonyms *in* Title

  **AND**

• **Proximity test:**
  Preferred term and synonyms *within 7 words* of Hierarchical parent, Related terms and Contextual keywords

Use OR for more Recall; AND for more Precision
Same taxonomy can drive multiple rules

Taxonomy elements + Boolean logic = IDOL search syntax
Google search syntax
OpenText search syntax
FAST search syntax
SmartLogic tagging rules
Concept Searching tagging rules
Sample rule in IDOL search syntax

(Law, “Legal system”, Legislat*, Statut*):Title
OR (Law, “Legal system”, Legislat*, Statut*)
WNEAR7 (“Law and crime”, Police, Legislature, Politics, Lawmaker, MP,
“White paper”, Vote)
Extended taxonomy term – weightings

- Preferred term (30%)
  - Law
- Synonyms (30%)
  - Legal system
  - Legislation
  - Statutes
- Hierarchical parent (15%)
  - Law and crime
- Related term (15%)
  - Police
  - Legislature
  - Politics
- Contextual keywords (10%)
  - Lawmaker, MP, White paper, Vote
Example: Hitting classification threshold of 85%

- Preferred term (30) + Synonym (30) + Related Topic (15) + Keyword (10) = 85
- Preferred term (30) + 2 Synonyms (60) = 90
- Two Synonyms (60) + 1 Related Topic (15) and 1 Hierarchical Parent (15) = 90
- Two Related Topics (30) + 1 Hierarchical Parent (15) + 4 Keywords (40) = 85

Note – adding Related Topics and Keywords can boost Recall but reduce Precision
Case studies

• Reuters – better taxonomy improved search
• Shop Direct – multiple entry points enhanced navigation
• Barnardo’s – data-driven user profiles led to new opportunities to help children
Reuters – fuller, automated tagging produced “extra” news to protect market share

- Editors added full metadata hierarchy under new umbrella headings like Business Performance, or Macroeconomics, with added synonyms and related terms
- Each article thus had more “hooks” for potential search queries
- Move boosted relevant headlines from single search terms by 18%
- Results increased perceived value against lower-cost competitors like Knight-Ridder
- Cost of producing that much extra news would have been £40 mln per year
- Automated tagging provided consistent application of consistent taxonomy
Shop Direct – user-based approach to find relevant products on web

• Traditional – and cluttered -- web navigation focused on product types
• Review and usability tests proved need for new, user-centred approach
• Web team developed other paths to find products:
  - Personas, such as New parent
  - Priorities, e.g. Value shopper
  - Occasions, such as Weddings
  - Seasons, e.g. Winter, Spring
  - Price “sliders” to refine results
Barnardo’s – data-driven user profiles created new business opportunities

- Service users were identified by an attribute called “referral reason,” e.g. Sexual exploitation or Abuse
- These descriptions were supplemented by “user issues,” e.g. Homelessness, or Substance Abuse, or Low self-esteem
- Oracle Business Intelligence reports revealed connections, showing high correlation between Sexual exploitation and Low self-esteem or Substance abuse
- Barnardo’s service managers could approach Local Authorities for additional funding to address these related issues

Copyright TFPL 2015
Summary – what we’ve learned

• Reviewed faceted taxonomies – Vertical “building blocks” grouped by entities, subject matter and focused filters

• How to extend taxonomies “horizontally” with synonyms, related terms and contextual keywords

• Helpful international standards to build and extend the taxonomy and ensure alignment with data models

• How to build effective classification and search queries using the components of an extended taxonomy to improve information retrieval and delivery
Questions and wrap-up
Thank you!

Contact details:

Jonathan Engel

07966 754614 | jonathan.engel@infoark.co.uk | www.infoark.co.uk