Audience-centric taxonomy: Using taxonomies to support heterogeneous user communities

Pei Jiun Tan
National Library Board,
Singapore
Pei_Jiun_TAN@nlb.gov.sg

Dave Clarke
Dow Jones, USA
dave.clarke@dowjones.com

Abstract

Controlled vocabularies enhance precision and recall but sometimes they achieve this at the expense of imposing a prescribed terminology and a homogeneous worldview upon a heterogeneous user community. Folksonomies allow end-users the freedom to describe content any way they want, but in doing so they create meta noise which diminishes precision and recall. This paper presents an alternative model called audience-centric taxonomy, which blends the best practices of top-down controlled vocabularies with the bottom-up approach of folksonomy. The result is a semantically rich and well structured vocabulary that can adapt how it presents itself to different end-user communities ensuring each audience sees the language and worldview that it prefers. The paper describes how the National Library Board Singapore intends to utilize audience-centric taxonomy to provide enhanced information access to its multilingual, multi-cultural user community.

Keywords: audience-centric taxonomy; meta noise; taxonomy; ontology; controlled vocabulary; folksonomy; information access.

1. Problems of Using Controlled Vocabularies for Heterogeneous User Communities

Controlled vocabularies, defined by standards like ANSI/NISO Z39.19-2005, enhance precision and recall but sometimes achieve this at the expense of imposing a prescribed terminology and single worldview upon a less cohesive user community. Folksonomies allow end-users the freedom to describe content any way they want, but they do this at the expense of precision and recall.

Blending top-down controlled vocabulary methods with bottom-up folksonomy and social tagging methods has recently gained significant interest. This project report describes an approach to this which attempts to segment user communities into defined audience segments and to serve these segments with terminology and conceptual structures that reflect their differing language and differing world-views.

2. Project Outline

2.1. Library 2010 – The Vision of The National Library Board Singapore

The Library 2010 project (National Library Board of Singapore, 2005) addresses the key challenges for the Singapore society and economy in the coming years and provides a strategic response. As Singapore moves into the era of a knowledge-based economy, the role of libraries in such an environment will become much more significant.

Libraries help in nurturing a society of life-long learners who can accelerate the creation of intellectual capital and create a new cycle of national innovation. This is an important factor of competition, much needed for success in a competitive knowledge-based economy. In the coming years, knowledge will take on an even more critical role as a social differentiator between communities and between nations. The National Library Board Singapore (NLB) will therefore

have to ensure that it continues to deliver learning and knowledge easily and affordably to all user communities to help maintain social cohesion. At the same time NLB will provide individuals, companies and government agencies with the real-time knowledge access that is necessary to succeed in a globally competitive environment.

To support this vision NLB identified the need to create three taxonomy/metadata components: seamless access to physical and digital content, multilingual search, and audience-centric terminology and taxonomic structures. Each of these three components is discussed in the remaining sections of this project report.

2.2. Unifying Access to Physical and Digital Content Collections

To provide access to knowledge wherever it resides, the NLB is developing sophisticated taxonomies that will support cross-searching of both the Library's physical book and media collections and also its expanding digital content collection.

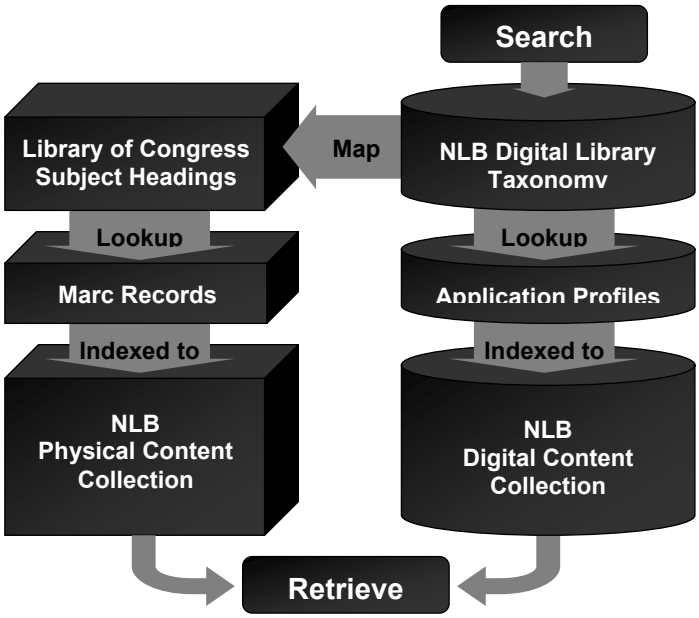

FIG. 1. Using taxonomy to unify physical and digital content.

Nearly all physical content is already catalogued in MARC format which includes subject metadata from the Library of Congress Subject Headings. A crosswalk between the LC Subject Headings and a custom taxonomy created to classify and subject index NLB's rapidly expanding Digital Library collection enables cross-searchability of content from both collections.

2.3. Supporting a Multilingual User Community

NLB is using a multilingual thesaurus and taxonomy management tool (Synaptica® from Dow Jones) to create, organize and map together multi-domain, multilingual search thesauri and navigational taxonomies.

A master taxonomy in one language (English) serves as the hub to a collection of three satellite taxonomies representing Singapore's other three official languages: Chinese, Tamil and Malay. These taxonomies are mapped to the hub taxonomy using the language equivalency methods recommended by the ISO standard for multilingual thesauri – ISO 5964:1985 (International Organization for Standardization, 1985).

Language equivalency mapping is not as simple as one-to-one translations. Cultural differences inherent in natural languages mean that sometimes concepts can exist in one language

that have no direct equivalent in another language. ISO 5964:1985 supports this condition by allowing independence in the semantic structures within each language and equivalence maps between them.

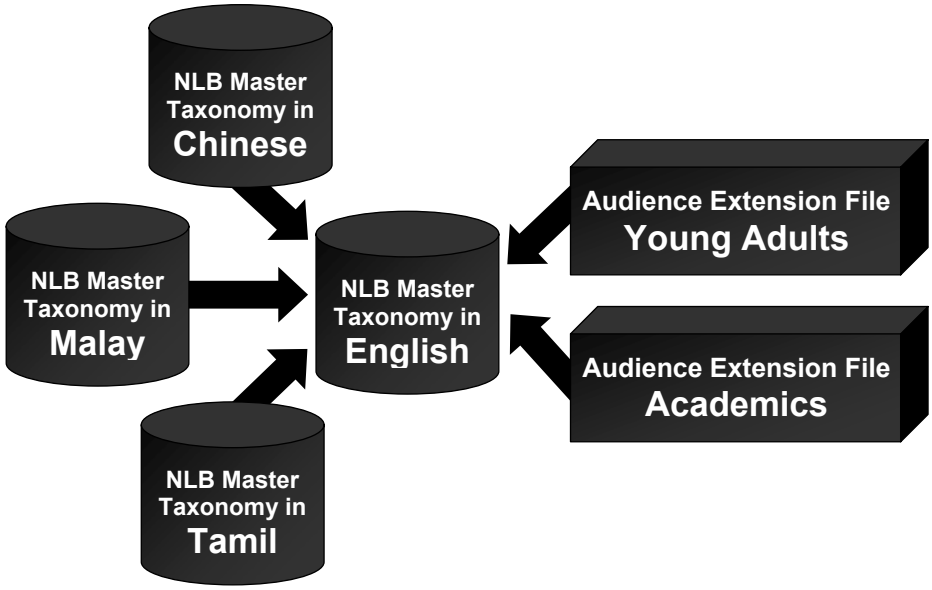

FIG. 2. Multilingual and multi-audience taxonomy.

This multilingual mapping means that a search performed in any of the four official languages will always generate a consistent set of results.

2.4. Supporting Audience-Centric Views

Besides supporting access in all four of Singapore's official languages, the NLB also recognizes that there can be significant heterogeneity within the community using a single language. For example, the NLB provides services to at least three different communities: (i) it has a function as the national book repository where its users are professional librarians; (ii) it functions as a lending and reference library serving children and adults from the general public; and (iii) it functions as a research resource serving academics and business professionals.

Each of these audience segments has different needs with regard to information access. Children may use different terminology than adults or academics. Research professionals need a more granular and technical level of terminology than the general public.

Standard controlled vocabulary methodologies require that where a concept can be expressed by variant forms one of the terms should be designated as the preferred term and all others should become signposts pointing to it. This method can work perfectly well as a behind-the-scenes lookup in search but it fails for browsable/navigable interfaces where a traditional system would revert to display only the standard preferred term.

An innovative response to this challenge has been developed by Dow Jones for its taxonomy management software application, Synaptica®. The solution is called Audience-Centric Taxonomy (ACT). It builds upon the foundation of a master controlled vocabulary by allowing multiple audience segments to be defined in which alternative preferred terminology may be expressed; concepts that are not relevant to a particular audience can be suppressed and extra granularity that is needed by a particular audience can be developed. The system automatically manages crosswalks between all the alternative audience-centric extensions, thereby enabling consistent searching using diverse taxonomic views.

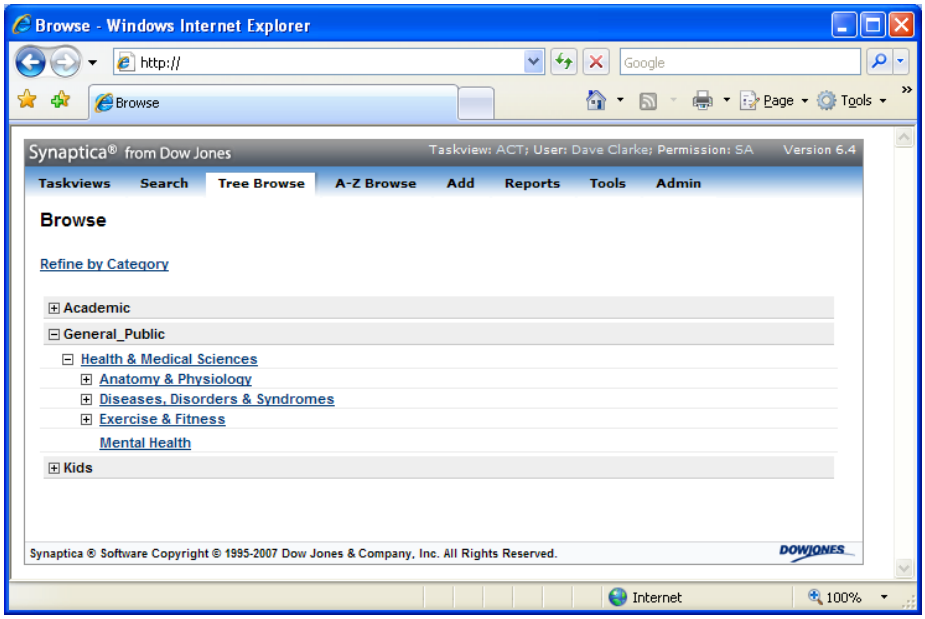

FIG. 3. Screenshot illustrating three top level audience-centric views.

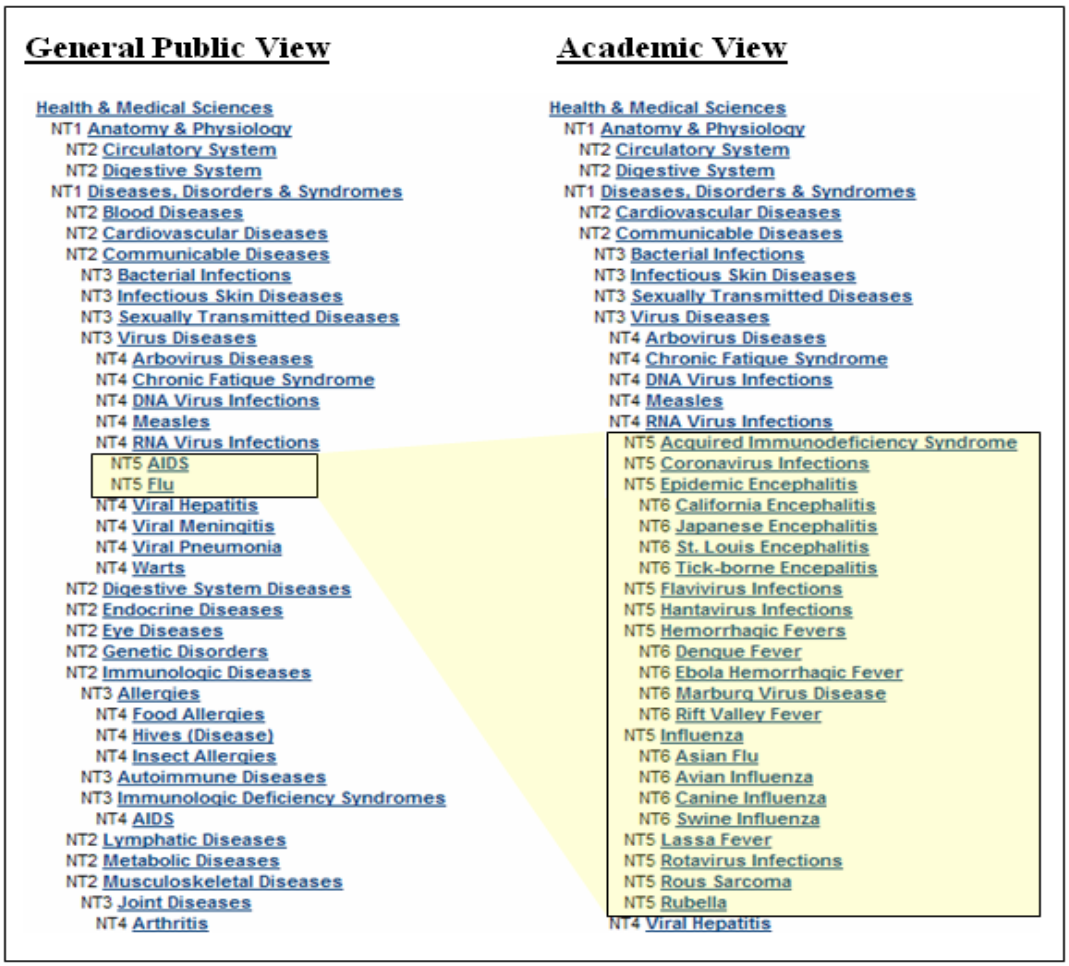

FIG. 4. Composite screenshot illustrating terminology and granularity differences.

Figure 3 illustrates the management of three alternative audience-centric views: Academic, General Public, and Kids. The "Diseases, disorders and syndromes" term is expanded in Figure 4 and shown side-by-side in both its "General Public" view and its "Academic" view.

In the General Public view the term "RNA Virus Infections" has just two narrower terms: AIDS and Flu. In the Academic view the granularity is greatly expanded with 11 narrower terms under "RNA Virus Infections" and an additional 12 terms under those.

Besides expanding the granularity the system is able to change the terminology. For the General Public audience "AIDS" and "Flu" are preferred terms but in the Academic view "Acquired Immunodeficiency Syndrome" and "Influenza" are the preferred terms.

To completely rebuild taxonomies many times over to suit the variances required by different audiences can be prohibitively expensive both to build and to keep synchronized. This obstacle has been overcome by a methodology that builds the views dynamically by blending a core master taxonomy with small audience- centric extensions. In the branch of terms used in Figure 4 the vast majority of the view is generated by borrowing terms and structure from the General Public view and the only additional work needed is to create the substitute preferred terms and the additional terms. These additional terms are directly connected to the core taxonomy thus obviating the need for manual synchronization.

3. Project Next Steps

The development of NLB's taxonomies for the digital library is still at an early stage. All legacy taxonomies have been imported into the central taxonomy management system. The new taxonomies and taxonomy crosswalks have been designed and prototyped but the majority of the development work is planned for the remainder of 2007 and 2008.

References

ANSI/NISO. (2005). *Guidelines for the construction, format, and management of monolingual controlled vocabularies* (ANSI/NISO Z39.19-2005). Approved July 25, 2005, by the American National Standards Institute. Retrieved July 2, 2007, from http://www.niso.org/standards.

National Library Board of Singapore. (2005). *Library 2010: Libraries for life, knowledge for success*. Retrieved July 2, 2007, from http://www.nlb.gov.sg/L2010/L2010.pdf.

International Organization for Standardization. (1985). *Guidelines for the establishment and development of multilingual thesauri* (ISO 5964:1985).