Empowering Digital Libraries Users through Combining Taxonomies with Folksonomies

Daniel Gelaw Alemneh
Digital Libraries Services,
University of North Texas Libraries
Daniel.Alemneh@unt.edu

Abebe Rorissa
Department of Information Studies, University at Albany, State University of New York (SUNY)
arorissa@albany.edu

ABSTRACT
The increase in the number and heterogeneity of digital resources has led cultural heritage institutions to develop tools, workflows, and quality assurance mechanisms that allow effective digital resource management. This poster assesses the current landscape in digital libraries as well as best practices and identifies emerging trends in information indexing. It also explores the potentials of and controversies surrounding user supplied tags or keywords in terms of complementing established controlled vocabularies in a diverse and collaborative environment.

Keywords
Index, tags, digital libraries. Social networking

INTRODUCTION
Digital libraries and supporting technologies have now matured to the point where their contents are incorporating complex and dynamic resources and services. Given the proliferation of scholarly digital contents, research in almost every discipline is becoming more data-intensive and current users demand access to various formats regardless of temporal and spatial restrictions and the types of devices used.

Trends
Advancement of digital technologies is shaping creation, access, use and preservation of information resources in ways that are so profound that traditional methods and concepts of access and organization are becoming less effective.

Various emerging (Web 3.0) applications, driven by semantic web technologies such as OWL, RDF, SPARQL, SWRL, etc., indeed offer powerful data organization, combination, and query capabilities. Some of the trends that are already taking shape include:

- Huge multimedia digital libraries instead of documents
- Complex retrieval systems instead of matching queries and document representations
- Visualization of the information space instead of a ranked list of search results
- Human information behavior instead of information need
- Users as both creators and consumers of information
- Bottom-up/user assisted indexing instead of “top-down” approach to indexing by trained professionals

As more users move into the more self-structured digital environment, a new paradigm for user experience will be required.

Indexing and Representation
An index term is simply a systematic representation of an information-bearing object (text, images, audio, video, etc) which points users to specific items on topics of interest. In other words, it is an information retrieval tool. As noted by Caropreso, Matwin and Sebastiani, (2001), one of the key issues for information retrieval (IR) and all other content-based text management applications is document indexing.

The generation of an accurate indexing term, i.e. a representation of an information-bearing object, is fundamental to the discovery, use, and reuse of digital resources. Indexing enhances the accessibility and value of a resource, provided that it is based on a thorough analysis of the resource. A good index helps users find what they need, even when they are not sure of what they need or looking for.

To fully understand what a good index is, it is necessary to be both micro- and macro-minded. On the micro level, we concern ourselves with the specific mechanics of creating an index term. On the macro level, indexing could be thought of as part of a larger context of an information retrieval system. At the basic level, retrieval of information then involves the user expressing an information need in the form of an information retrieval request by using terms...
from the common vocabulary and matching requests with stored records.

Considering the growing interdisciplinarity and constant changes and development of information and knowledge management; digital resources demand a more specialized treatment and characterization that can help to better capture the semantics and relations of the underlying concepts.

Different metadata elements describe different characteristics or aspects of an object or digital resource. However, users are more interested in the contents and the subjects, rather than in what the objects are. The most useful metadata about a digital object is the subjects (or keywords), since they explicitly describe what it is about. To describe digital resources accurately, metadata creators and/or catalogers try to follow (as closely as possible) the thinking of the creator/author, and also anticipate what and how the users might want to discover and retrieve them. Otherwise, the descriptions or subject headings will be ineffective. Wichowski (2009) noted that in the rapidly growing information environment, unidentified and unorganized content, however useful it may be, is at risk of being rendered unfindable, and thus obsolete.

A number of researchers, Bates (1998), Peterson (2006) and Spiteri (2007); among others, analyzed content indexing (especially subject indexing) and described the general behavior of users' information seeking and their queries. Many agreed that the two major reasons why users experience problems with subject access are the quality and application of subject index on the one hand, and the complexity of knowledge as well as information literacy skills required for successful subject access on the other. To maintain the consistency of search results and high recall of available resources, it is critical to ensure the quality of the keywords and taxonomies used to index heterogeneous digital resources within digital libraries.

CONTROLLED TERMS
Libraries have been developing various systems for creating and managing controlled vocabularies for use in digital library initiatives. Selecting a term from a controlled vocabulary ensures indexing consistency and enhances retrieval precision across all digital resources. In this regard, controlled terms provide a broad navigational tool for browsing through digital content and digital library collections. Users can drill down through subordinate subject terms to find other content within that subject category. Such an approach promotes consistency and enhances a digital library user's ability to find and use available digital resources.

Considering the complexities and multifaceted issues involved in determining the level of indexing term quality, traditional approaches may not adequately address the diverse users’ requirements and needs. It is critical for digital libraries to assess the practice that shapes the generation of subject terms which determines the effectiveness of subject and keyword access.

This raises the question of whether user-supplied tags complement traditional indexing by professionals in ways which significantly improve information retrieval.

Folksonomy
Folksonomy is a user-generated system that allows users to tag their favorite digital resources using their natural-language words. Trant (2009) summarized both the negatives and positives of folksonomy. Most critics point to the fact that it is an uncontrolled vocabulary and leads to less effective information retrieval. On the other hand, proponents point to the fact that it is user-friendly and enables personalized information retrieval by users. As folksonomies are in a continual state of flux, they are better able to accommodate current terminology and concepts than traditional indexing tools and systems such as the Dewey Decimal Classification and the Library of Congress Subject Headings.

As some commentators noted, (e.g., Peterson, 2006; Spiteri, 2007; and TechSmith, 2008; among others) both approaches share a basic problem: the potential users of information are disconnected from the process. They believe combining both traditional indexing systems with folksonomies is the solution for delivering a richer user experience of digital libraries as well as on the Web while leveraging the benefits of composite applications, mash ups, and Service-Oriented Architectures.

SUMMARY
While the capacity to create digital content is great and the appetite for it seemingly insatiable, much work remains to be done in order to make the infrastructure robust. In the increasingly self-structured digital environment, it is clear that traditional information indexing will be of limited use to user experience.

In view of the growing interdisciplinarity and constant changes in users’ requirements, access to digital resources relies on a seamless discovery process that offers all possible options to users. Accordingly, the roles of information environments such as cultural heritage institutions evolved from that of local resource repositories to global gateways for access.

Although different metadata schemas are often mutually complementary, good subject or keyword terms help users find what they need, even when they are not aware of their needs. In their search for better discoverability for existing digital resources; cultural heritage institutions attempt to enrich the traditional catalog and metadata with additional user-supplied terms and descriptions. Both topical and natural approaches to the subject matter in a digital collection provide high-level descriptions and representations. Such an aggregation of digital items will
add values and enhance a digital library user's ability to find, access, use, and re-use available digital objects.

Considering the emerging applications, combining the strengths of the two approaches indeed offer powerful data organization, combination, and query capabilities. In the poster, we will present the current landscape, best practices, and emerging trends in indexing resources in digital libraries. We will also explore the potentials of and controversies surrounding folksonomies.

REFERENCES


