Integrating Controlled Vocabularies into Cultural Heritage Digital Collections: the Portal to Texas History Experience

Daniel Gelaw Alemneh

University of North Texas Libraries, Digital Projects Unit, Denton, Texas, P.O. Box 305190, Denton, Texas 76203-5190 Dalemneh@library.unt.edu

Mark Edward Phillips

University of North Texas Libraries, Digital Projects Unit, Denton, Texas, P.O. Box 305190, Denton, Texas 76203-5190 mphillips@library.unt.edu

Dreanna Belden

University of North Texas Libraries, Digital Projects Unit, Denton, Texas, P.O. Box 305190, Denton, Texas 76203-5190 dbelden@library.unt.edu

Abstract

The University of North Texas Libraries developed a system for creating and managing a hierarchical controlled vocabulary for use in digital library initiatives. This short paper discusses and provides an overview of the benefits and challenges of integrating controlled vocabularies into cultural heritage digital collections. Specifically, it examines the University of North Texas Libraries' initiatives from a real-life digital resource management and search implementations viewpoint.

Introduction

The University of North Texas (UNT) Libraries created a robust application framework for integrating diverse, complex and dynamic digital information resources from a multitude of sources. The various digital projects at the UNT Libraries include: the CyberCemetery, the Congressional Research Service Reports Archive, the World War Poster collection, Federal Newsmaps and other materials drawn from collections throughout the libraries.

One UNT Libraries' digital initiative, the Portal to Texas History, is a State-wide collaborative digital program that offers students and lifelong learners a digital gateway to the rich collections held in Texas libraries, museums, archives, historical societies, and private collections. It features digital reproductions of photographs, maps, letters, documents, books, artifacts, and more. In addition, Portal Primary Source Adventures and curriculums that comply with TEKS (Texas Essential Knowledge and Skills) standards highlight relevant materials for students and classroom teachers.

The Portal to Texas History provides a metadata framework that fosters a collaborative environment for participating institutions. The Portal collaborators share a number of goals, including ensuring long-term and easy access to a wide variety of cultural heritage collections.

The UNTL Metadata

Considering the role of standardized metadata in digital resource life cycle management, the UNT Libraries actively promote metadata-based digital resource management. The existing UNT Libraries metadata system empowers participating institutions to describe digital objects in a consistent way that provides for optimum searching, discovery, and retrieval, while supporting long-term preservation of digital resources.

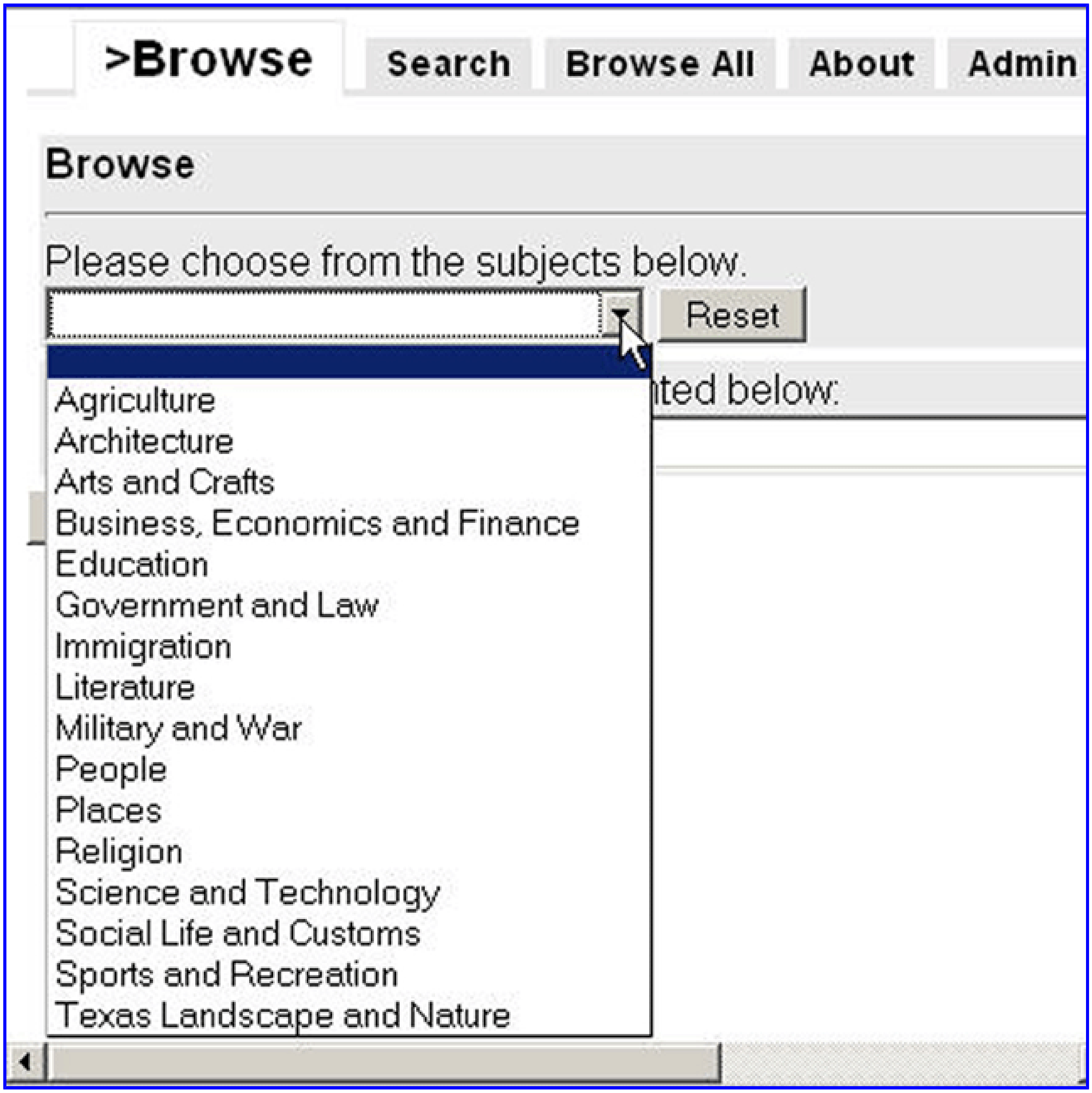
The UNT Libraries metadata element set comprises locally qualified Dublin Core based descriptive metadata along with the detailed technical and preservation metadata elements that document how digital resources are created, formatted, arranged, identified, and sustained. While promoting interoperability with widely accepted standards, the recommended UNT Libraries metadata elements allow flexibility at the local level to integrate existing and anticipated content.

The UNTL Controlled Vocabularies

As the UNT Libraries digital collections increase, opportunities for creating value-added services emerge. Among other tools, the UNT Libraries developed a system for creating and managing hierarchical controlled vocabularies for use in digital library initiatives such as the Portal to Texas History.

As a controlled vocabulary, the UNT Libraries Browse-Subjects (UNTL-BS) provide a broad navigational tool for browsing through digital content. (See Figure-1 and Figure-2). Users of the Portal can select a top level hierarchical search term, then drill down through subordinate subject terms to find other content within that subject category. Design of the browse interface and its supporting controlled vocabulary focused on creating a service that would feel similar in use and function to a commercial search engine directory page, an interface already familiar to Internet users.

Figure-1 UNT Libraries Browse Subject (UNTL-BS)



On the data entry side, The UNTL-BS enables data enterers to easily select appropriate values and place them in metadata records. Selecting a value from a controlled vocabulary ensures consistency and enhances precision across all digital resources.

Challenges and Opportunities

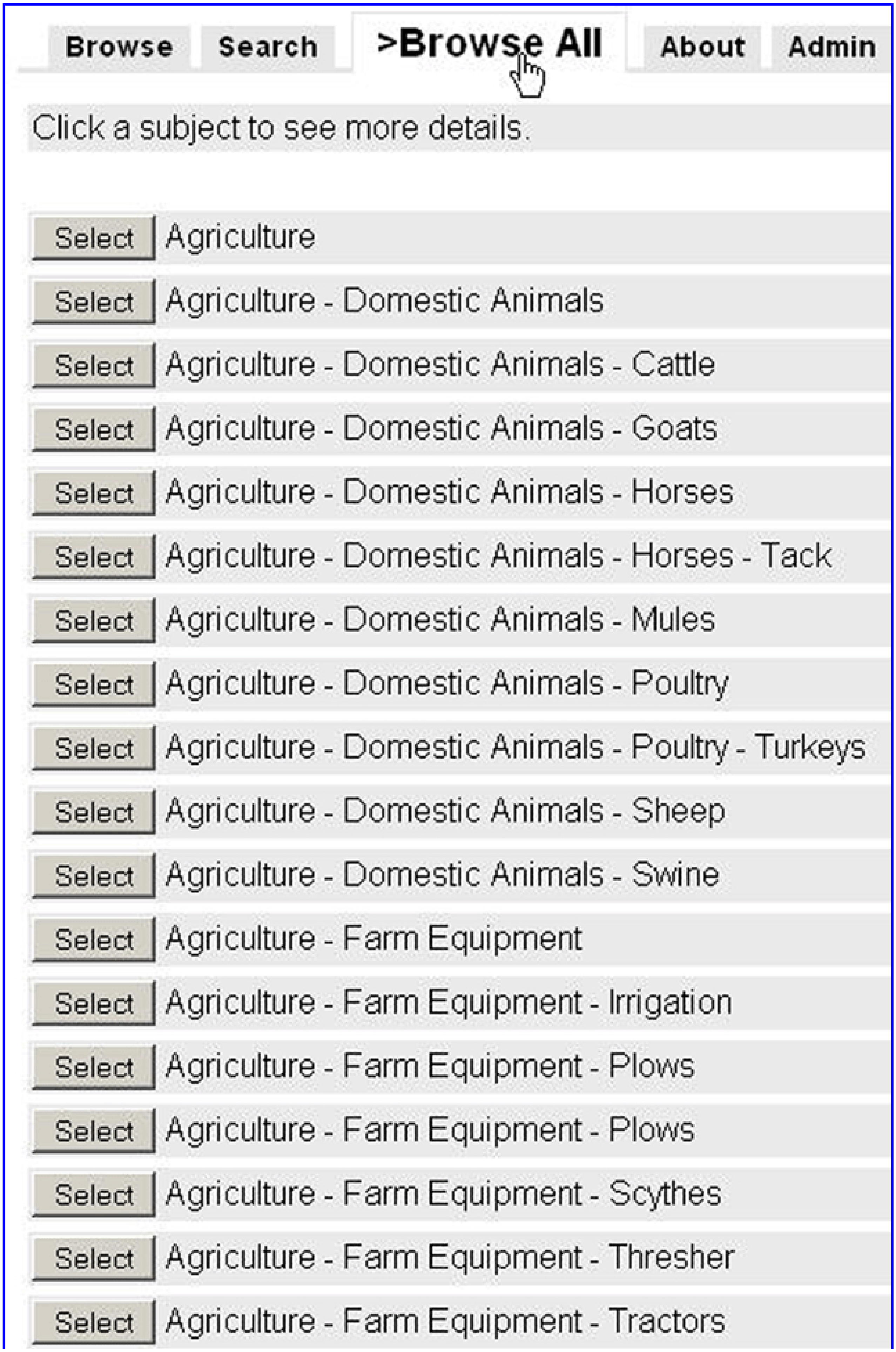
There are different ways to searching the Internet and the common approaches are field-based searches and full-text searches. With the amount of digital resources available online, providing easy and efficient search mechanism that best serve the user is more important than ever. Users prefer to use natural language searching. Natural language is cheaper and more user friendly, but controlled vocabularies provide higher precision and can be harnessed to produce a browse interface.

A number of researchers (Rowley, 1993; Budhu and Coleman, 2002; Franklin, 2003; and Baca, 2004), among others evaluate the performance of various indexing languages in retrieval and test the effectiveness of both vocabulary control and natural languages. Although no clear winner exists in this debate, if the strengths of natural language and controlled vocabulary can be combined, then stronger more effective systems will emerge.

The use of both controlled vocabulary and natural language allows flexibility to meet the changing demands of sources as well as users. One of the disadvantages of a controlled vocabulary is that new subjects and descriptors may not be added to a database quickly enough. Furthermore, considering the diversity of participating institutions and heterogeneity of their collections, all digital resources may not be described adequately using pre-determined terms. Although collaborating institutions have much in common, they may have conflicting metadata requirements. The nature of their digital objects (museum artifacts, archives, historical documents, scholarly documents, etc.) and the information-seeking behavior of their respective users (historians, genealogists, students, researchers, etc.) may require significantly different approaches to resource description.

To overcome the limitations and balance the issues and contextual interactivities, the UNT Libraries implemented a hybrid system that uses both controlled terms and free keywords in order to describe resources adequately. This flexible approach of pre-defined and custom-generated vocabularies provides maximum flexibility to capture high-quality metadata for digital resources. An integrated approach of using both structured and unstructured data provides better browsing and searching capabilities and offers users better search results and a better experience.

Figure-2 UNTL-BS Browse-All interface



Click a subject to see more details.

Agriculture

Agriculture - Domestic Animals

Agriculture - Domestic Animals - Cattle

Agriculture - Domestic Animals - Goats

Agriculture - Domestic Animals - Horses

Agriculture - Domestic Animals - Horses - Tack

Agriculture - Domestic Animals - Mules

Agriculture - Domestic Animals - Poultry

Agriculture - Domestic Animals - Poultry - Turkeys

Agriculture - Domestic Animals - Sheep

Agriculture - Domestic Animals - Swine

Agriculture - Farm Equipment

Agriculture - Farm Equipment - Irrigation

Agriculture - Farm Equipment - Plows

Agriculture - Farm Equipment - Plows

Agriculture - Farm Equipment - Scythes

Agriculture - Farm Equipment - Thresher

Agriculture - Farm Equipment - Tractors

Summary

Although the debate of Controlled vocabularies versus free-text keywords continues, there is a general recognition that controlled language and natural language should be used in conjunction.

The UNT Libraries made early steps towards synthesizing natural language and controlled vocabulary within its digital collections. The diversity of the objects, collections, institutions, and individuals creating metadata in the Portal to Texas History program provide a solid environment for demonstrating the benefits of a hybrid model of controlled vocabularies and free-text keywords searching. The UNT Libraries experiences strengthen the assertion that it is ideal to use both systems in conjunction.

REFERENCES

Alemneh, D. Hartman, C. & Phillips, M. (2005). The Issues of Compliance and Interoperability in Integrating Heterogeneous Digital Information Resources *Society for Imaging Science and Technology (IS&T) Archiving Conference* 2005 April 26 - April 29, 2005, Washington, D.C. Retrieved February 23, 2007 from http://www.imaging.org/store/epub.cfm?abstrid=32247

Baca, M. (2004). Fear of authority? Authority control and thesaurus building for art and material culture information [Electronic version]. *Cataloging & Classification Quarterly* 38(3/4), 143-151. Retrieved February 6, 2007, from Electronic Collections Online database.

Budhu, M. and Coleman, A. 2002. The Design and Evaluation of Interactivities in a Digital Library *D-Lib Magazine* 8 (11), November. Retrieved February 6, 2007, from http://www.dlib.org/dlib/november02/coleman/11coleman.html

Carrow, D., & Nugent, J. (1981). Comparison of free text and index search abilities in an operating information system. *Information Management in the 1980s: Proceedings of the American Society for Information Science 40th Annual Meeting,* Knowledge Industry Publications, 131-138.

Cousins, S. (1992). Enhancing subject access to OPACs: controlled vocabulary vs. natural language. *Journal of Documentation* 48(3), 291-309.

Dubois, C. P. R. (1987). Free text vs. controlled vocabulary: a reassessment *Online Review* 11: 243-253.

Feldman, S. (1996). natural language: comparing DIALOG, TARGET, and DR-LINK *Online,* 20, 71-74. Retrieved February 9, 2007, from Library Literature and Information

Sciences Full Text database.

Franklin, R. A. (2003). Re-inventing subject access for the semantic web *Online Information Review* 27(2), 94-102

Hartman, C. et.al. (2005). "Development of a Portal to Texas History." *Library Hi Tech* 23, no. 2; 151-163. Retrieved February 23, 2007 from http://puck.emeraldinsight.com/10.1108/07378830510605124

Henzler, R. (1978). Free or controlled vocabularies *International Classification,* 5(1), 21-26.

Marshall, C. (1998). Making Metadata: a study of metadata creation for a mixed physical-digital collection Retrieved February 20, 2007, from http://citeseer.ist.psu.edu/marshall98making.html

Rowley, J. (1994). The controlled versusd natural indexing languages debate revisited: a perspective on information retrieval practice and research *Journal of Information Science* 20(2), 108-119. The UNT Libraries Metadata Projects page. Retrieved February 20, 2007, from http://www.library.unt.edu/digitalprojects/documentation/metadata.htm

Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science and Technology* 37, 413-464.