Kekal Abadi 29(1) 2011

Controlled Vocabulary in the Digital Age

Muhamad Faizal Abd Aziz
University of Malaya Library
mfaizal@um.edu.my

Abstract

Libraries are undergoing changes from managing traditional to hydrid information. With the rapid development in IT, especially the Internet, libraries need to establish a standard to manage the information such as having a controlled vocabulary. This article describes the underlying factors in the digital age for the establishment of controlled vocabulary.

Abstrak

Perpustakaan kini sedang mengalami perubahan dari segi pengurusan maklumat, iaitu secara tradisional kepada hibrid. Dengan perkembangan pantas teknologi maklumat (IT), terutama internet, perpustakaan perlu membangunkan satu piawaian untuk perbendaharaan kata terkawal (controlled vocabulary). Artikel ini menerangkan faktor-faktor asas untuk membangunkan perbendaharaan kata terkawal dalam era digital.

Introduction

Libraries are undergoing change from traditional to hybrid and currently to the electronic or virtual library. Challenges faced by librarians are tremendous in maintaining the collection and services. As we witness the development of the physical library, we should not forget the development of its content or collection. Today, libraries are not only holding just books, but also different kind of information and formats whether in printed, audio, digital, and electronic resources available on the web which have substantially increased in the recent years. The resurgence of interest in controlled vocabularies in the recent decade is related to the development of contents and format of library materials. Basically, these are caused by three main factors, which are: Development of Internet Technology, Development in Integrated Library System and Variations of Information Content and Format. All these factors are seen as reasons why researchers and librarians are paying more attention in the vocabulary control activities. In this article, the definitions from various sources, history and development will be discussed.

Definition of Controlled Vocabularies

Before looking at the definition of the term 'controlled vocabularies' (CVs), we need to know why controlled vocabularies are important and what are the effects of not having it. Vocabulary control is used to improve the effectiveness of information storage and retrieval system, web navigation systems, and other environments that seek to both identify and locate desired content via some sort of descriptions using language. The primary purpose of vocabulary control is to achieve consistency in the description of content objects and to facilitate retrieval. Basically, the need for vocabulary control arises from two basic features of natural language which are: (i) two or more words or terms can be used to represent a single concept and (ii) two or more words that have the same spelling can represent different concepts.

Controlled vocabulary can be simply defined as a list or collection of terms or words available for use. In library and information science, controlled vocabulary is a carefully selected list of words and phrases, which are used to tag units of information (document or work) so that they may be more easily retrieved by a search. To get a better understanding of controlled vocabulary let us look into the definitions below.

Wikipedia.com gives a definition of controlled vocabularies as providing a way to organize knowledge for subsequent retrieval which are used in subject indexing schemes, subject headings, thesauri and taxonomies.

According to Larson (1998), controlled vocabulary is the attempt to provide a *standardized* and *consistent* set of terms (such as subject headings, names, classifications, etc.) with the intent of aiding the searcher in finding information.

History And Development of Controlled Vocabularies

Controlled Vocabularies have been in place since early 1900, when the first printed subject headings were published in 1909. In any case, Library of Congress Subject Headings was published earlier than that in 1898, after it was converted from an author-plus a classed-catalogue to a Dictionary Catalogue. The first Library of Congress Subject Headings used the American Library Associations List of Subject Headings for use in dictionary catalogue. The actual printed subject headings used in the Dictionary Catalogues at the Library of Congress (later been titled as Library of Congress Subject Headings) began in summer 1909. Dewey Decimal Classification (DDC) was originated in 1873 and had published and patented in 1876. In 1950s, government agencies began to develop controlled vocabularies for the burgeoning journal literature in specialized fields; for example Medical Subject Headings (MeSH) was developed by United States National Library of Medicine. Sears List of Subject Headings, which first appeared in 1923. The development of controlled vocabularies did not stop at that point; it continues to develop in more modern ways. Now the subject headings are available in online format, for example; Classification Web for Library of Congress and MeSH online.

Types of Controlled Vocabularies

There are many types of controlled vocabularies. Listed below are some common ones:

i) List or "pick list"

A list or pick list is a limited set of terms arranged in a simple alphabetical list or in some other logically evident ways. Lists are used to describe aspects of content object or entities that have limited number of possibilities. Examples of lists would be that for Geography which list country, state and city; for Language (English, France and Germany).

ii) Synonym ring

Synonym ring is a set of terms considered equivalent for the purposes of retrieval. Synonym rings usually occur as flat lists. Use of synonym ring ensures that a concept can be described by multiple synonyms or quasi-synonym terms and retrieved if any one of the terms is searched.

iii) Taxonomy

Taxonomy is controlled vocabulary consisting of preferred terms, all of which are connected in a hierarchy or poly-hierarchy way.

iv) Thesaurus

A Thesaurus is a structured controlled vocabulary arranged in a known order so that the various relationship among terms are displayed clearly and identified by standardized relationship indicators. Relationship indicators are usually employed reciprocally.

Factors Contributing to the Resurgence of Interest

There has been an increased interest in the development of technology that affects library and its collections. The first reason would be the development of Internet Technology which drives a drastic change in the way we use and access library collection

a) Development of Internet Technology

It was reported that librarians were among the earliest professionals to use the Internet. Internet technology underwent rapid development in the early stages, which started with four computers, telnet, dial-up and the latest is wireless connection. These developments have played an important role in changing the library and librarianship fields. Internet enables access to various information resources in many formats. Traditionally, library is only accessible to a group of people or community staying nearby and the collection will only be available within the library building. With internet technology the library has become borderless and its contents virtually accessible from anywhere. To accommodate these changes, librarians and researchers have to find ways to make information retrieval possible through Internet. The searching criteria or access point need to be refined to get an accurate search result. Most websites, search engines and web portals use natural language or free-text language as their controlled vocabularies that results in wider and broader search results, increasing the hit list but decreasing the precisions. The use of natural language or free-text language is to accommodate the layman searching capabilities. Arguments arose among professionals on the advantages and disadvantages of these two options as accurate and reliable retrieval tools. Thomas (2000) commented: "with the Web estimated to be increased by 10 million pages weekly, the task of indexing the internet resources is clearly argentums, and not something that can be done overnight by the cataloguer. Instead of relying on the catalogue to identify and retrieved web pages, users have to turn to web portals which use metadata".

Research prove that controlled vocabulary has more advantages over natural language and free-text language. Gerhan (1991) found that catalogue users retrieved more records in fewer attempts making use of the Library of Congress Subject Headings. Arellano

(1991) discovered that a great deal of material was missing. Referring to the importance of a controlled vocabulary, Tillet (2000) pointed out, "Authority control enables "precision and recall" which are lacking from today's web searches. The above findings show the importance of controlled vocabulary for subject retrieval in a network environment.

b) Development of Integrated Library Systems

Libraries began to automate and network their catalogue in the late 1960s. Frederick G. Kilgour at the Ohio College Library Center (know OCLC, Inc) led the networking at Ohio libraries during the '60s and '70s. The automated catalogue became available to the world, first through telnet or TN3270 via IBM and only became web-based on 1997 with the introduction of HyWebCat. In the conventional way of searching, it will be done through a catalogue card which is made available in the library. Today, with the enrichment of Internet technology, the library integrated system has replaced the catalogue card and the Online Public Access Catalogue (OPAC) to provide more precise searching options. With this Web-based Catalogue or OPAC, users can retrieve information not only about holdings in the individual library, but also can examine holdings from other libraries. Card catalogues have given way to online catalogues to incorporate new search options, particularly subject searches. In card catalogues, the options for retrieving information about the holdings of a library are by author, title and subject. In comparison, online catalogues enable searches by title words or words included in any other field as surrogates possible. In this way, the possibilities of subject access in online and web-based catalogue are not limited to subject headings and a controlled language, but they are extended to key words, mainly those from titles which are the basic constituent of free-text. Realizing the new needs to accommodate the current trend, the existing controlled vocabularies need to be improvised for new roles in the electronic environment, with the aspects of improvement in areas such as:

1. Improved currency, hospitality for new topic, and capability for accommodating new terminology
2. Flexibility and expandability – including possibilities for decomposing faceted notation for retrieval purposes
3. Intelligibility, intuitiveness, and transparency – it should be easy for users to use, responsive to individual learning style, able to adjust to the interest of users, and allow for custom views
4. Universality – the scheme should be applicable for different types of collections and communities and should be able to be integrated with other subject languages, and
5. Authoritativeness – there should be a method of reaching consensus on terminology, structure, revision that includes user communities.

Some of the controlled vocabularies have already adjusted to the electronic environment such as AGROVOC the agricultural thesaurus, WebDewey, which is Dewey Decimal Classifications adapted to electronic environment and California Environmental Resources (CERES) thesaurus.

c) Variations of Information Content and Format

Nowadays library holdings are not just limited to books, but also different formats of information such as visual images, audio recordings, electronic resources. Many organisations and individuals are using the internet for generating and delivering electronic information. The amount of electronic resources that are available on the web have substantially increased in recent years and there is an urgent need to include them into the library collection and consequently, to include their surrogates in the library catalogue. New terminology specifically in the Internet and Information technology fields have been forced to burst-up in this recent decade. The new contents that entered the library collections among others are websites and web portals. Websites are very unique format of materials and new controlled vocabulary have to be developed. Web pages have specific characteristics such as hyperlinks, anchors and metadata. Web portals use free-text and natural language types of controlled vocabularies which are not really reliable when searching. The World Wide Web has transpired a new type of controlled vocabulary which is ontology and directory-style subject browsing that is very popular in commercial search engines (directories and web pages).

Conclusion

Although the use of free-text language and natural language is an easy and cheap option for indexing activities, there is still a need to use controlled vocabularies for the storage and retrieval of the precise information that matches user needs. Any search engine or directory and other home grown scheme in the web, even those with well-developed terminological policies such as Yahoo and Google still suffer from a lack of understanding of principles of classification design and development. In this way controlled vocabulary will continue to play an important role in the organization of knowledge and librarians will have to be more adequately prepared to

face the challenges that technology and the new types of information resources impel in future.

References

American National Standards Institute. (2007). *Guideline for the construction, format, and management of monolingual controlled vocabularies*. Retrieved 1 January 2009, from http://www.slis.kent.edu

Bates, M.J. (1988). How to use controlled vocabularies more effectively in online searching. *Online*, *12*(6), 45-56. Retrieved from http://proquest.umi.com/pqdweb

Chan, L.M. (2001). *Subject vocabulary for web resources*. Retrieved 1 January 2009, from http://klement.nkp.cz/Csalin/caslin01/sbornik/subjectvoc.html

Controlled Vocabulary. Retrieved 1 January 2009, from http://en.wikipedia.org

Golub, K. (2006). Automated subject classification of textual web pages, based on a controlled vocabulary: Challenges and recommendations. *New Review of Hypermedia and Multimedia*, *12*(1), 11-27. Retrieved from http://www.informaworld.com

Golub, K. (2006). Using controlled vocabularies in automated subject classification of textual web pages, in the context of browsing. *TCDL Bulletin*, *2* (2), 1-10. Retrieved from http://www.ieee.tcdl.org

Hornby, A. S. (1953). Vocabulary control: History and principles. *ELT Journal*, *VIII*(I), 15-21. Retrieved from http://eltj.oxfordjournals.org

Lancaster, F. W. (1986). *Vocabulary control for information retrieval*. (2nd Ed). Virginia: Information Resources Press.

Lima, C., et al., [n.d.]. *A historical perspective on the evolution of controlled vocabularies in Europe*. Retrieved 1 January 2009, from http://www.irbdirekt.de/daten/iconda/CIB7425.pdf

Marshall, J. (2006). Control vocabularies: Implementation and evaluation. *Key Words*, *14*(2), 53-59. Retrieved from http://www.informaworld.com

Martinez-Arellano, F.F. (2001). Teaching of subject access and retrieval at Mexican LIS schools. Paper presented at the 67th IFLA Council and General Conference , Boston. Retrieved from http://www.ifla.org/IV/ifla67/papers/026-142e.pdf

Should We Control Vocabulary?. Retrieved 1 January 2009, from http://www.nelinet.net/edserv/conf/cataloging/2007/ohnmitchell.pdf

Stone, A. (2000). The LCSH century: A brief history of the library of congress subject headings, and introduction to the centennial essays. *Cataloging & Classification Quarterly*, *29*(1&2), 1-15. Retrieved from http://cataloging andclassificationquarterly.com/ccq29nr1-2ed.htm

Tenopir, C. (1987). Searching by controlled vocabulary or free text?. *Library Journal*, *112*(19), 58-59. Retrieved from http://web.ebscohost.com

Windsor, R. (1995). *Designing a controlled vocabulary for use with Digital Asset Libraries.* Retrieved 1 January 2009, from http://www.daydream.co.uk/controlled_vocabulary.asp