Chapter 15
Merging Controlled Vocabularies for More Efficient Subject–Based IR Systems

Ioannis Papadakis
University of Ionio, Greece

Konstantinos Kyprianos
University of Ionio, Greece

ABSTRACT

One of the most important tasks of a librarian is the assignment of appropriate subject(s) to a resource within a library’s collection. The subjects usually belong to a controlled vocabulary that is specifically designed for such a task. The most widely adopted controlled vocabulary across libraries around the world is the Library of Congress Subject Headings (LCSH). However, there seems to be a shifting from traditional LCSH to modern thesauri. In this paper, a methodology is proposed, capable of incorporating thesauri into existing LCSH-based Information Retrieval–IR systems. In order to achieve this, a mapping methodology is proposed capable of providing a common structure consisting of terms belonging to LCSH and/or a thesaurus. The structure is modeled as a Simple Knowledge Organization System (SKOS) ontology, which can be employed by appropriate subject-based IR systems. As a proof of concept, the proposed methodology is applied to the DSpace-based University of Piraeus digital library.

1. INTRODUCTION

One of the most common and time-consuming responsibilities of librarians and information scientists in their line of work is the assignment of appropriate values to the metadata element “subject”. Subjects are employed in order to describe resources within a repository. Quite often, subjects are assigned values that belong to a controlled vocabulary. Controlled vocabularies are described as carefully selected lists of words and phrases that are used to describe digital resources and other things (Williamson, 1996).

Currently, Library of Congress Subject Headings (LCSH) (Library of Congress, 2007, 2011) and thesauri are the most popular controlled vo-
Merging Controlled Vocabularies for More Efficient Subject-Based IR Systems

cabularies employed in order to assign values to the subjects of their corresponding resources. Subjects that belong to LCSH have been interpreted into many languages and accordingly employed by many libraries around the world. However, the shifting of information exchange to the web and the consequent requirement to comply with its dominating technologies brought up the need for a more suitable type of controlled vocabulary than LCSH. Along these lines, thesauri emerged as subject-based controlled vocabularies that are particularly compatible with the web. Thus, there seems to be a need to incorporate thesauri to modern Information Retrieval - IR systems within libraries (e.g. search modules of Online Public Access Catalog – OPAC) that nevertheless employ LCSH for the subject indexing of their collections.

In this paper, in order to address this seemingly contradictory requirement, a methodology is proposed capable of integrating existing LCSH-based indices with thesauri, in the context of interactive, subject-based IR systems. This way, the inherent advantages of thesauri are exploited without the need of re-indexing the underlying library collection. In order to achieve this, a mapping methodology is presented capable of providing a common structure for both the subjects of LCSH, namely LCSH terms and the subjects of a thesaurus, namely thesaurus descriptors. The structure is expressed as a Simple Knowledge Organization System - SKOS (W3C, 2004; Miles & Perez-Aguera, 2007) ontology, which can be employed by appropriate subject-based IR systems. As a proof of concept, the proposed mapping methodology is applied to the University of Piraeus digital library IR system (http://www.lib.unipi.gr/multi.php?lng=en).

The rest of the paper is structured as follows. The first section refers to LCSH and thesauri in general. Then, a detailed analysis about the advantages and drawbacks deriving from the employment of LCSH and thesauri in subject indexing is presented. The next section refers to similar efforts regarding the integration of thesauri and LCSH. In section 4, the proposed approach is described and accordingly evaluated. Finally, in section 5, conclusions concerning the proposed approach are presented.

2. LCSH AND THESAURI AS CONTROLLED VOCABULARIES FOR PRE- AND POST-COOORDINATED SUBJECT-BASED IR SYSTEMS

As stated in Harper and Tillet (2007), there are various types of controlled vocabularies concerning the assignment of values to the "subject" metadata element. LCSH and thesauri seem to be the predominant options for such a task.

According to the LCSH guidelines (Library of Congress, 2007), subjects employ verbal strings for the description of the resources within a repository. They, also, define relations between these strings in order to demonstrate the synonyms and the associations between the subjects. A subject consists of the main heading, which corresponds to the prevalent concept of the resource. In case of resources adhering to more specific concepts, the main heading can be further specialized with subdivisions. Such subdivisions represent various aspects of the main heading. There are four types of subdivisions according to the LCSH guidelines: topical, form, geographical and chronological subdivisions. The LCSH guidelines impose certain restrictions concerning the order of subdivisions within a subject. More specifically, such order normally conforms to the pattern ‘topic-place-time-form’:


The above pattern can be altered, particularly with regard to the position of the geographical subdivision, which may appear in any position after the main heading.