DDC in DSpace: Integration of Multi-lingual Subject Access System in Institutional Digital Repositories

Bijan Kumar Roy*, Subal Chandra Biswas**, Parthasarathi Mukhopadhyay***

ARTICLE INFO

Article history:
Received 15 June 2017
Revised 19 July 2017
Accepted 03 August 2017

Key words:
Institutional Digital Repository, Knowledge Organization System, DDC, DSpace, Classification, Resource Discovery

ABSTRACT

The paper discusses the nature of Knowledge Organization Systems (KOSs) and shows how these can support digital library users. It demonstrates processes related to integration of KOS like the Dewey Decimal Classification, 22nd edition (DDC22) in DSpace software (http://www.dspace.org/) for organizing and retrieving (browsing and searching) scholarly objects. An attempt has been made to use the DDC22 available in Bengali language and highlights the required mechanisms for system-level integration. It may help a repository administrator to build an IDR (Institutional Digital Repository) integrated with SKOS-enabled multilingual subject access systems for supporting subject descriptors based indexing (DC.Subject metadata element), structured navigation (browsing) and efficient searching.

1. Introduction

The enormous growth of scientific and research outputs require comprehensive indexing, application of classification in order to facilitate structured navigation, and intuitive user interface to support seamless resource discovery. In digital library (DL) environment, the importance of developing different knowledge organization systems (KOSs) has been recognized quite early to support effective browsing by end users, proper organization of collections in helpful order and more importantly to categorise items during ingesting in IDR on the basis of standard knowledge modeling tools like library classification system. Almost all DLs started using KOSs in different forms to satisfy their users’ needs in late nineties (Brummer et al., 1997; Koch et al., 1997) and ranges from authority files to classification schemes, thesauri, and ontologies. But the real-time integration of Indic-script based KOS in an IDR (both ways linking to support selection of the term from integrated KOS at the time of metadata encoding as well as by searchers at the time of retrieval) reported by the authors (Mukhopadhyay, 2010, 2015).

* Assistant Professor, Dept. of LIS, The University of Burdwan, WB (bijankumarroy@yahoo.co.in)
** Professor, Dept. of LIS, The University of Burdwan, WB (scbiswas_56@yahoo.co.in)
*** Associate Professor, Dept. of LIS, The University of Kalyani, WB (psmukhopadhyay@gmail.com)

http://dx.doi.org/10.5865/IJKCT.2017.7.4.071
Our discipline (Library and Information Science) is devoted to developing SKOS (Simple Knowledge Organization System) based vocabulary control tools for organizing, disseminating and preserving large collections of objects both available in print and non-print formats.

In view of the foregoing, this paper focuses on a very specific problem at hand in the context of a multilingual country like India (India has 418 languages of which 407 are living and 11 are extinct (Maitra, 2002) - is it possible to convert an existing KOS into SKOS-enabled and Unicode-compliant format for its easy integration with an existing IDR system. This paper demonstrates an integration of the Dewey Decimal Classification, 22nd edition (DDC22) with a prototype institutional digital repository (IDR) termed as BURA (Burdwan University Research Archive). BURA is based on the open source information retrieval software namely DSpace (version 4.2). DDC22 as KOS has been incorporated into BURA framework and has been integrated with the indexer interface as well as the user interface so that it (DDC) can be used both at the time of indexing and searching. In view of the needs of multilingual resource discovery, the DDC22 (up to the 3rd summary level) is translated into Bengali language and made available along with English version for users and indexers of the information representation and retrieval (IRR) system. The advantages of using integrated SKOS-enabled controlled vocabulary includes standardized subject indexing, subject category based hierarchical display of scholarly resources and structured navigation by end users.

2. Literature Review

The value of classification has already been recognized as significant in the digital and networked environment (Soergel, 1999; Hodge, 2000; Currier & Wake, 2001). Gunjal, Urs, & Shi (2008) analyzed the various aspects of KOS in DLs. An attempt has been made to integrate KOS with the digital library architectures using emergent semantic technologies and data by Babu, Sarangi & Madalli (2012). A JISC (Joint Information System Committee) report reviewed the state of the art in this area, in particular, with regard to vocabulary types, indicative use cases, best practice guidelines and current research (Tudhope, Koch, & Heery, 2006). McKiernan (1996) listed 14 sites that use, or claim to use, DDC for the organization of resources. In another study, Shiri & Molberg (2005) reported that 33 digital collections in Canada have already employed some type of KOS in their search interfaces. Whereas, Manaf, Bechhofer & Stevens (2012) surveyed the current state of SKOS vocabularies on the Web and identified 478 SKOS vocabularies, which were gathered through collections and Web crawling. Smith (2013) reported on the current status of the LCMPT (Library of Congress Medium of Performance Thesaurus) to describe music resources and outlined the future steps that will allow the thesaurus to operate in a linked data environment.
Digital repositories are now using some kind of KOS for organizing resources in order to assist users to better find and locate resources, providing valuable additional information (Joo & Lee, 2011). A number of studies (Roy, Biswas, & Mukhopadhyay, 2016a, 2016b, 2016c, 2017) have advocated for adopting ontology driven SKOS in organizing and managing IDR collections to fulfill the subject approach of the users. Lin et al. (2016) presented a unified visual interface based on metadata aggregation and automatic classification mapping that can aggregate metadata records from multiple unrelated repositories. Another paper (Eric Si, O’Brien, & Probets, 2010) reported the development of a prototype to improve the interoperability between different terminology resources in order to provide a better subject cross-browsing service for metadata repositories.

There have been several projects aimed at implementing controlled vocabularies in organizations and in specific contexts. The Nordic WAIS/WWW Project (from Summer 1993 to Summer 1994) at Lund University Library (http://www.ub2.lu.se/W4.html) was the first project which tried to apply simple methods of automatic classification in order to improve the discovery and retrieval of Internet resources. Hill et al. (2002) discussed the importance of integrating KOS into digital library architectures and the approaches used at the Alexandria Digital Library (ADL project). Koch et al. (1997) in DESIRE project reported the development and implementation of automatic classification in networked environment. Another project (RUBRIC, 2007) presented a controlled vocabulary of ASRC (Australian Standard Research Classification) subject terms for use with the DSpace digital library suite. Zeng & Chan (2004) provided an extensive review of problems of mapping between different KOSs which were being applied in the European ALVIS project. Ferreira & Baptista (2005) reported that the University of Minho was the first institution in the Portuguese speaking world to use a translated version of DSpace in the context of institutional repositories. The first thesaurus they have imported into their repository system was the publicly-available Association for Computing Machinery (ACM) Computing Classification System (CCS). In the same fashion, Solomou & Koutsomitropoulos (2009) reported the success story of the University of Patras in incorporating SKOS into their DSpace system. They have incorporated a real SKOS vocabulary: the thesaurus of Greek Terms into their repository system. Witten (2003) offered a unique way of knowledge organization by providing for hierarchical phrase browsing in Greenstone Digital library system.

3. Why KOS in IDR Environment?

In Institutional Digital Repository (IDR) environment, resources are organized in many ways (OpenDOAR, 2017; ROAR, 2017). Even subject repositories (SRs) are no exception. No consensus has been developed and no unique system has been reported any initiatives for organizing and representing open knowledge resources. But there are flexibilities in organizations of collections in a proper perspective depending on the institutional preferences. After analyzing ROARMAP database, Roy (2014) reported that resources may be arranged by ‘subject’, by ‘collection type’ or by ‘departments and schools’ (a form of crude classification). Almost all IDRs use and/or have their own controlled vocabularies (in built) but in most of the cases such tools are insufficient
in representing taxonomies of subject categories in regional or local languages. So, resources need to be organized under suitable categories or sub-categories and are required to be displayed in local languages in such a way that it should reflect and cover the thrust areas or areas of information demand of the community members (Roy, 2015).

4. Methodology: Development of Subject Access System

In DSpace software, controlled vocabulary file is represented in a simple XML (extensible markup language) format. Users search resources using set of keywords that are organized in a tree (taxonomy), which appears during the search and submission process. In the present work, DDC22 (up to 3rd summary in English) has been incorporated in DSpace (version 4.2) as KOS along with translations in Bengali. All subject terms are enclosed in a <node> element and expression of a hierarchical (narrower in meaning) relationship is allowed through the use of the <isComposedBy> sub-element. Furthermore, it has been made possible to incorporate a simple annotation mechanism by using <hasNote>. The steps of transforming DDC in Bengali language and making it available at the time of indexing as well as searching (in both languages e.g. Bengali and English) have been demonstrated through the following steps and different screen snapshots (Fig. 1 to Fig. 4). The summary of the process may be illustrated as below.
Step 1

The first step of integrating DDC22 in DSpace is by making necessary changes in DSpace.cfg file (Fig. 1). This step provides a link between repository/archive and available controlled vocabularies through the parameter `webui.controlledvocabulary = true`.

![Fig. 1. Link to Control Vocabulary Device](image1.png)

Step 2

The next logical step is to link the target controlled vocabulary (here DDC22) with submitter and editor interfaces through necessary modifications in the `input.xml` file (Fig. 2). It includes modifications in two `DC.Subject` blocks of `input.xml` file in such a way that indexers/submitters during metadata encoding must have to pick up at least one subject category from DDC22 (up to the 3rd summary level) in English and one subject category from its Bengali translation.

![Fig. 2. Modified Input.xml File](image2.png)
Step 3

English language subject divisions and subdivisions are based on DDC22 and Bengali language equivalents are based on the available translation work (Saha, 2008). The structure of the SKOS-enabled XML-formatted file displaying the hierarchy of Social Science >> Education >> Subject categories under ‘Education’ is given here in English (Fig. 3).

Similarly, the hierarchy Social Science >> Education >> Subject categories under ‘Education’ is also displayed for Bengali script (Fig. 4).
5. Selection of Subject Categories

The DDC22 is displayed both at the time of submission and searching, and the submitter can add standard subject term(s) from DDC for both in English (Fig. 5) and Bengali (Fig. 6).

**Fig. 5.** Selection of Subject Categories from DDC (Language 1 - English)

**Fig. 6.** Selection of Subject Categories from DDC (Language 2 - Bengali)
6. Browsing and Searching of Subject Access System

The integrated subject access system allows users browsing and searching specific subject categories which may not be organized under proposed Communities and Sub-communities of the software framework or has not been categorized in the proposed IRR system. It (DDC) can be displayed by simply clicking on the link on the left side of the navigation panel and user can navigate throughout the list and can select appropriate subject category(s) (Fig. 7). This window displays the main divisions and sub-divisions of subject categories of DDC22 in English.

Another novelty, as stated earlier, of this SKOS-enabled subject access system is that, it supports browsing and searching resources through specific subject categories in Bengali. The subsequent window (Fig. 8) displays all subject categories in Bengali. Each broad subject has a hierarchical listing of subject categories/sub-categories and is grouped under those subjects. These keywords are organized in a subject tree (or subject taxonomy) which appears during searching as well as indexing process. Top level terms are displayed and user can navigate any of the top terms by simply clicking on it. The plus sign (+) indicates that the category concerned has sub-categories and/or links to resources under it.
7. Filtering of Subject Categories

This system allows users to filter documents against a standard subject division/sub-division (taken from DDC22). After putting a term in search box (e.g. America - here অমেরিকা), it displays all the subject categories that matched with the term. It shows all the links/or fields related with that particular term (Fig. 9).

The filtering process is also available at the time of indexing. Indexer can either put desired term(s) directly in appropriate box at the time of searching or can pick up category(s) from integrated vocabulary control device. It should be pointed out that indexer may opt for any number of subject categories or subject divisions for populating subject access field in a given metadata schema (here DC.Subject). Fig. 10 gives the result against a search term (e.g. education) in English.
Fig. 9. Filtering of Subject Searching

Fig. 10. Indexing of Subject Term: Language 1 (English)
In the same fashion, indexer can select term in Bengali. The moment indexer clicks against a search query (e.g. library - here গুরুনিধার), it will expand the subject categories and the system displays all the sub-divisions of the term matched (Fig. 11).

8. Conclusions

The mechanisms as demonstrated above through different snapshots in methodology section may be considered as an add-on for DSpace that enables repository administrators as well as community members to use a controlled set of subject categories to describe self-archived items and to display resources in a structured way. At present, cross-browsing features including controlled-vocabulary based searching and authority control are lacking in most IDR systems. This prototype may be viewed as an alternative solution to fill this gap. This Integrated subject access system in SKOS-enabled format has possibilities in improving retrieval effectiveness and will provide unified access to materials in different media and in different languages. Like other SKOS-enabled subject access systems in an Internet retrieval, this model supports browsing, searching resources in a number of ways including multilingual features (here Bengali language) in an IDR. The advantage of using a library classification system like DDC22 (up to the third summary) is to ensure that the concepts are represented in the form of standard subject categories and subject descriptors are arranged in the form of a tree (taxonomy) to show each category or sub-category is placed under its broader concept. Users can easily find the categories (and resources associated with the categories) they are looking for by expanding just a few branches of the taxonomy. So, it may be concluded that this Web-enabled
Bengali-script based DDC-driven SKOS-enabled subject access system has the potential to populating DC.Subject metadata element in a standardized manner to support both library professionals as well as end users of an IDR.

References


http://www.iastate.edu/~CYBERSTACKS/CTW.htm.


[ About the authors ]

Bijan Kumar Roy, M.Com, MLIS, PhD is Assistant Professor in Library and Information Science, The University of Burdwan, West Bengal, India. He started his career as full time JRF and later joined as Librarian in Government-aided College in 2009. His research interest includes open access, open source software, digital repository.

Subal Chandra Biswas, b. 1955, M.A. (Economics), MLIS, Ph.D. (Loughborough) has recently retired as Professor of LIS, The University of Burdwan, West Bengal, India. Recipient of Commonwealth Scholarship (UK), 1985-1989. He has an experience in teaching and research of more than three decades both at home and abroad. Research interests include information seeking, information retrieval, and public libraries. Has supervised more than a dozen doctoral theses.

Parthasarathi Mukhopadhy, MLIS, PhD is Associate Professor in Library and Information Science, Kalyani University, Kalyani, West Bengal, India. His research interest includes Open Access resource organization, Open source applications in library organization and multilingual information retrieval. He is presently associated with two mega digital library projects in India namely National Digital Library Initiative and National Virtual Library Initiative.