DESIDOC Journal of Library & Information Technology, Vol. 33, No. 4, July 2013, pp. 289-294

Multilingual Thesaurus and Interoperability

A. Neelameghan and S.K. Lalitha

Sarada Ranganathan Endowment for Library Science
702, Upstairs, III Block, Rajajinagar, Bangalore-560 010, Karnataka
E-mail: anm2002@vsnl.net

ABSTRACT

This paper briefly describes a pilot project to design and develop a multilingual- English and eight other languages of India – thesaurus for 'Education'. The thesaurus is derived from the bilingual (English-Kannada) Colon Classification (CC), (Ed. 7 revised). The CC number and Dewey Decimal Classification (DDC) number are metadata in the thesaurus strtucture. The thesaurus for each language is separately built using Greenstone Digital Library software. The descriptors are hyperlinked. Descriptor terms, terms in the Scope Note (text), CC and DDC numbers are indexed. Two or more of the thesauri can be searched simultaneously. The advantages of using class number in the search and retrieval are mentioned. Other issues discussed include interoperability as key to effective interlinking among knowledge organising tools (KOTs) and in integrating two or more KOTs, and issues and problems in building multilingual thesaurus (thesauri) with particular reference to the cultural frames and languages of India.

Keywords: Multilingual thesauri, class number, interoperability

1. INTRODUCTION

Information resources are of various kinds– books, chapters in books, papers in periodicals and conference volumes, newspapers, case records, data tables, graphs, images, maps, music sheets, etc. The contents may be in different languages. One or more of such resources may be required to satisfy the needs of information seekers. These may be available in their conventional physical document forms and/or in digital form, and as audio, video records, CDs, transparencies, power point presentations, etc. We may also include e-mail messages, websites, virtual dialogues/conferences, social networks (by blogs, twitter, facebook, You-Tube, etc.) together with their printed versions, if any. To know about their respective contents and to facilitate retrieval of desired information directories, indexes, lists, catalogues and such other tools are used. These may be on conventional paper medium and/or in digital form.

With a view to managing the vocabulary/terminology of directories, indexes, lists, catalogues, etc., and for enhancing and supplementing their capabilities KOTs are useful. The KOTs include ontologies, taxonomies, lexicons, dictionaries, schemes for subject classifications, thesauri, wordnets, semantic nets, self-organising systems, etc. These help in standardising and/or managing vocabulary in indexes, formulating search expressions and retrieval from databases (including online web-enabled databases) and also in organising information resources mentioned above and their surrogates (e.g. entries in catalogues and bibliographies) in a helpful sequence. Some of the KOTs, e.g. classification schemes and thesauri for specific domains can also provide a map of the content of the related domain and to learn about inter-relationships among the components and concepts of the domain.

2. OBJECTIVES

The objectives of this paper are to:

(a) Show with a pilot project that 'Interoperability' is key to effective interlinking among knowledge organising tools (KOTs) and in integrating two or more KOTs;

(b) Affirm that use of a classification code (e.g. class number) has advantages specially in a multilingual thesaurus system.

(c) Discuss selected issues and problems in building multilingual thesaurus (thesauri) with particular reference to the cultural frames and languages of India.

Received 20 May 2013, online published 22 July 2013

3. MULTILINGUAL THESAURUS: PILOT PROJECT

A collaborative pilot project for developing thesauri in different languages for a subject domain is in progress. In this pilot thesaurus a part of the 'Education' domain is covered. Currently thesauri in English, Tamil, Kannada, Tulu, Malayalam have been completed; Thesauri in Punjabi (gurumuki), Hindi (devnagari), Sanskrit (devnagari),

Urdu and Arabic will be added. The initial thesaurus is built using the English CC Personality Facet schedule.

The starting point, that is, collection of relevant concepts/terms to create a corpus of terms for the subject domain was facilitated with the availability of the bilingual (English-Kannada) version of S.R. Ranganathan's Colon Classification (CC) scheme (English edition 7, partly revised) in print and machine-readable form[1]. The schedule for 'Education' was enriched with terms selected from ERIC thesaurus and the UNESCO/IBE thesaurus.

Each descriptor is given a CC number and corresponding Dewey Decimal Classification (DDC) number, and Scope Note (SN), Broader Term (BT), Narrower Term (NT) as applicable. The pilot thesaurus for each language is separately built. In the pilot version the SN is given only in the English language thesaurus. The terms in the schedules in each language are hyperlinked among themselves and with corresponding terms in the thesauri for the other languages.

The descriptor, class number, and terms in the SN (in each language) are indexed. The terms in the schedule of each thesaurus can be browsed; and searching by any of the language term(s) and by class number (CC or DDC) is provided for. By clicking on a selected term or class number the corresponding terms in the other languages may be displayed and used in a search. Display of schedule of terms in one or more than one language is provided for. A thesaurus descriptor can be hyperlinked to online/web-based databases for Education and allied domains; term(s) from the thesaurus can be cut and pasted or dragged and dropped in the search box of the database(s) for search and retrieval of hit records from the database(s). Selected screen shots are given in the Annex.

4. SOFTWARE

The Greenstone Digital Library Software (GSDL) is used for building the different language thesauri. The GSDL is a suite of software for building and distributing digital library collections. It provides for organising information records and publishing it on the iInternet or on CD-ROM. It is produced by the New Zealand Digital Library Project at the University of Waikato, New Zealand and developed and distributed in cooperation with UNESCO and the Human Info NGO. It is an open source, multilingual software, issued under the terms of the GNU General Public License. GSDL is Unicode-compliant. User interfaces have been prepared in several languages, including languages of India through collaboration with institutions and individuals in different countries.

Several institutions and individuals have input terms to the Education thesauri in the different languages and also have assisted in editing the inputs.

5. DISCUSSIONS

5.1 Need for Knowledge Organising Tools

Ontologies "firstly, help sustain a communications framework around the domains of interest between people, organisations, and systems by providing a shared and common understanding of the domain. Secondly, they enable knowledge re-use and sharing since other researchers can adopt or integrate an ontology for their own purposes..... They facilitate inter-operability among systems by specifying and translating different concepts and languages in a domain (or across several domains). A well-developed ontology produces cost-time benefits by eliminating or reducing the cost of re-inventing a knowledge-base system for each use. Furthermore, ontologies help users learn domain knowledge, and, in addition, browse and search for information by providing structured knowledge representation"[2]

A knowledge organising system (KOS) consists of two or more KOTs (or other information sources–e.g. full-texts, abstracts, images, audio, video material - that are interlinked or integrated. These assist vocabulary management functions in information/ subject indexing, searching, retrieval, and presentation, to provide additional information and to support knowledge discovery, enhance the scope for research in the subject domain and in the field of KOS theory and practice.

5.2 Interoperability

Interoperability is the key to simultaneous search and retrieval from two or more databases containing a single or multi-genre resource, in one or more languages. Interoperability capability is required to merge or integrate databases holding multi-genre and/or multi-lingual records.

Interoperability is the ability of diverse systems and organisations to work together. The term is usually used in a technical systems engineering sense, or alternatively in a broad sense, taking into account social, political, and organisational factors that impact system performance. Initially, interoperability was defined for IT systems and services restricted for information exchange between systems.

5.2.1 Interoperability: General

- A property of a product or system, whose interfaces are completely understood/transparent,

to work with other products or systems, present or future, without any restricted access or implementation.

- The ability of two or more systems or components to exchange information and to use that information[3] (IEEE Glossary).
- Enable end-user applications using different types of computer systems, operating systems, and application software, interconnected by different types of local and wide area networks[4].

This generalised definition can be applied to any system. It defines several criteria that can be used to discriminate between systems that are 'really' inter-operable and systems available as such but are not so as they may not comply with one or more criteria, namely: non-disclosure of one or several interfaces, and/or implementation or access restriction built in the product/system/service. Two aspects of interoperability are: Syntactic interoperability and semantic interoperability.

5.2.2 Syntactic Interoperability

Two or more systems should be capable of communicating and exchanging data, using specified data formats communication protocols, etc. XML or SQL standards are tools of syntactic interoperability. This applies to ASCII or a Unicode format (for English or text in other languages) in all communicating systems.

5.2.3 Semantic Interoperability

The ability of systems to automatically interpret the information exchanged meaningfully and accurately to produce results as defined by the end users of the systems. To achieve this, both sides must refer to a common information exchange reference model: what is sent is the same as what is understood (e.g. ISO 2709).

Interoperability between KOSs has become crucial in facilitating simultaneous searches in several databases or to merge different databases into one. The standards for KOS design and development, the American Z39.19:2005 and the British 8723-4:2007, provide detailed recommendations for interoperability. A new ISO standard about thesauri and interoperability is ISO 25964-1.

The available technology provides tools for interoperability, e.g. formats and functional requirements for subject authority, as well as those for Semantic Web RDF/OWL, SKOS Core and XML. Given the high cost of designing and developing new KOSs, interoperability techniques makes it possible to take advantage of the existing ones. Martinez Tamayo[5], et al. review basic concepts, models and methods recommended by the standards as well as several experiences on interoperability between KOS - monolingual and multilingual.

- Interoperability – Terminology management; terminology mapping; terminology authoring; and terminology browsing.
- Compatibility/convertibility between databases usually require these capabilities in respect of following:
  - Field: Tag; Metadata; Name
  - Field: Structure
  - Field: Demarcation/Delimiters
  - Language and presentation of text of different genres
- Use of interface (e.g. BSO, GSDL language interfaces)
- Programming (e.g. Pascal programmes for CDS/ISIS) for facilitating exchanges of records/information among databases
- Unicode compliance to handle different languages
- Codes/standards for rendering of names, etc.
- Input and retrieval: Use of standards, dictionaries, glossaries, lexicons, etc. as vocabulary management tools.

6. BILINGUAL/MULTILINGUAL KOTs: ISSUES

- Identifying equivalent/near-equivalent concept terms in all the languages of the KOT. If such bi-lingual (online) dictionaries are available or prepared, the system may be enabled to pick up all the equivalents (and near-equivalents) from the dictionary for matching with the search term(s)[6].
- Hyper-linking of terms in a KOT, e.g., glossary, thesaurus and map tree structure of the subject field to other KOTs, data collections, images, etc., can be provided to enable comprehensive search and knowledge discovery.

6.1 Multilingual KOTs

Absence of equivalent concepts; meaning of concept terms not exact equivalents in different languages; cultural constraints and frames are issues to be taken into account. So also with homonyms, homographs and synonyms.

6.2 Semantic Relationships in Culture-specific Domains

Conceptual relationships between terms especially in the subjects in the Humanities (religion, philosophy, socio-cultural topics, etc.) are not as precisely defined as in the physical or life sciences. In addition to this general issue, building a multilingual thesaurus for a culture-specific domain raises certain issues that are related to the nature of such domains. Firstly, concepts encountered in and associated with

culture-specific domains in particular are abstract in nature and rarely can they be related to concrete referents. Secondly, a large number of concepts of a culture-specific domain are those that have some meaning in the life of the members of the community belonging to the culture. These have implications for a multilingual thesaurus. A language is a product of, and reflects the culture of the particular community. In other words, it is the culture and lifestyle prevalent among the members of a particular community that necessitates and results in the formation of lexemes/ expressions (words/terms) for concepts associated with that culture and lifestyle. It is therefore very likely that, unless the communities that speak two different languages share the same culture, certain concepts in culture-specific domains may have verbal expressions only in a particular language. Therefore, in building multilingual thesauri, it was indeed difficult to find exact equivalent concepts in the the different languages language. The notion of 'equivalence relationship' widely used in information retrieval thesauri had to be extended to include both equivalent terms in the source language (Synonyms in Tamil for Tamil terms) and terms in English denoting near-equivalent/similar concepts. The second major issue, also related to the notion of 'equivalence relationship', was the occurrence of several different terms, for instance, in Tamil with closely similar meanings/connotations. Neelameghan and Raghavan have discussed such issues more elaborately in a recent paper[7].

6.3 Advantages of using Class number

The main advantages from a scheme for subject classification such as DDC, CC, etc., include:

(a) If in a language a single term or phrase is not available as equivalent or near-equivalent to, say, an term in English but a description or meaning is given, say, in the Scope Note the class code can be given in this field (metadata); the system will pick up this class code in the sear and display the field content. This can be done irrespective of the language of the KOT schedules.

(b) The hierarchical division(s) of a concept can be displayed.

(c) The interlinks among a concept/term and the narrower, broader and other associatively related concepts/terms can be mapped for a visual display.

(d) The class numbers from different classification schemes asssigned to the descriptors facilitates comparative study of the strength and weaknesses of the schemes, and

(e) These enables corrections and improvements in the classification schemes used.

ACKNOWLEDGEMENTS

The authors thank the following for their respective inputs to the preparation of the multilingual Education thesasurus: Dr B.A. Sharada – the Kannada-English version of Colon Classification (CC), and the initial Kannada Education thesaurus; Dr K.S. Raghavan and Dr K.N. Prasad for their inputs to the Kannada-English version of CC; Dr I.K. Ravichandra Rao for arranging to get the Tulu version of the thesaurus (in progress); Dr M.P. Satija, *et al*. for preparing the Punjabi and Hindi versions of the thesaurus (in progress); Mr Vinod Kumar for preparing the Sanskrit version of the thesaurus (in progress); Mr and Mrs Abdul Rahman Kamaruddin for reviewing and editing the Urdu and Arabic versions; and Dr A.Y. Asundi for assigning DDC numbers to the descriptors.

REFERENCES

1. Sharada, B.A. Ranganathan's Colon Classification: Kannada-english version 'dwibindu vargikaraNa'. *In* Categories, Contexts and Relations. Proceedings of the 12th International ISKO Conference, 6-9 August 2012, Mysore, edited by A. Neelameghan and K.S. Raghavan, Ergon Verlag GMBH, Wurzburg 2012, pp. 63-68.
2. Ok nam Park. Opening ontology design: A study of the implications of knowledge organisation for ontology design. *Knowledge Organisation*, 2008, **35**, 209-21.
3. Institute of Electrical and Electronics Engineers. IEEE Standard Computer Dictionary: A compilation of IEEE standard computer glossaries. New York, NY, 1990.
4. O'Brien, James & Marakas, George. Introduction to information systems. Ed. 13, McGraw-Hill/ Irwin; ISBN 0-07-304355-9.
5. Martinez Tamayo, Ana M; *et al.* Interoperability of knowledge organization systems: The state of the art. *Información, cultura y sociedad* [*Information Culture and Society*], 2011, **24**, 15-37.
6. Neelameghan, A. & Lalitha, S.K. Searching in and retrieval from GSDL multilingual multimedia databases simultaneously using a multilingual dictionary. *Information Studies*, 2011, **17**, 47-64.
7. Neelameghan, A. & Raghavan, K.S. Frames of knowledge: A perspectivde of vedic hinduism and dravidian culture. *In* Cultural frames and knowledge, edited by Richard P. Smiraglia & Hur-Li Lee. Ergon Verlag GMBH, Wurzburg, 2012, 19-62.

Annexure

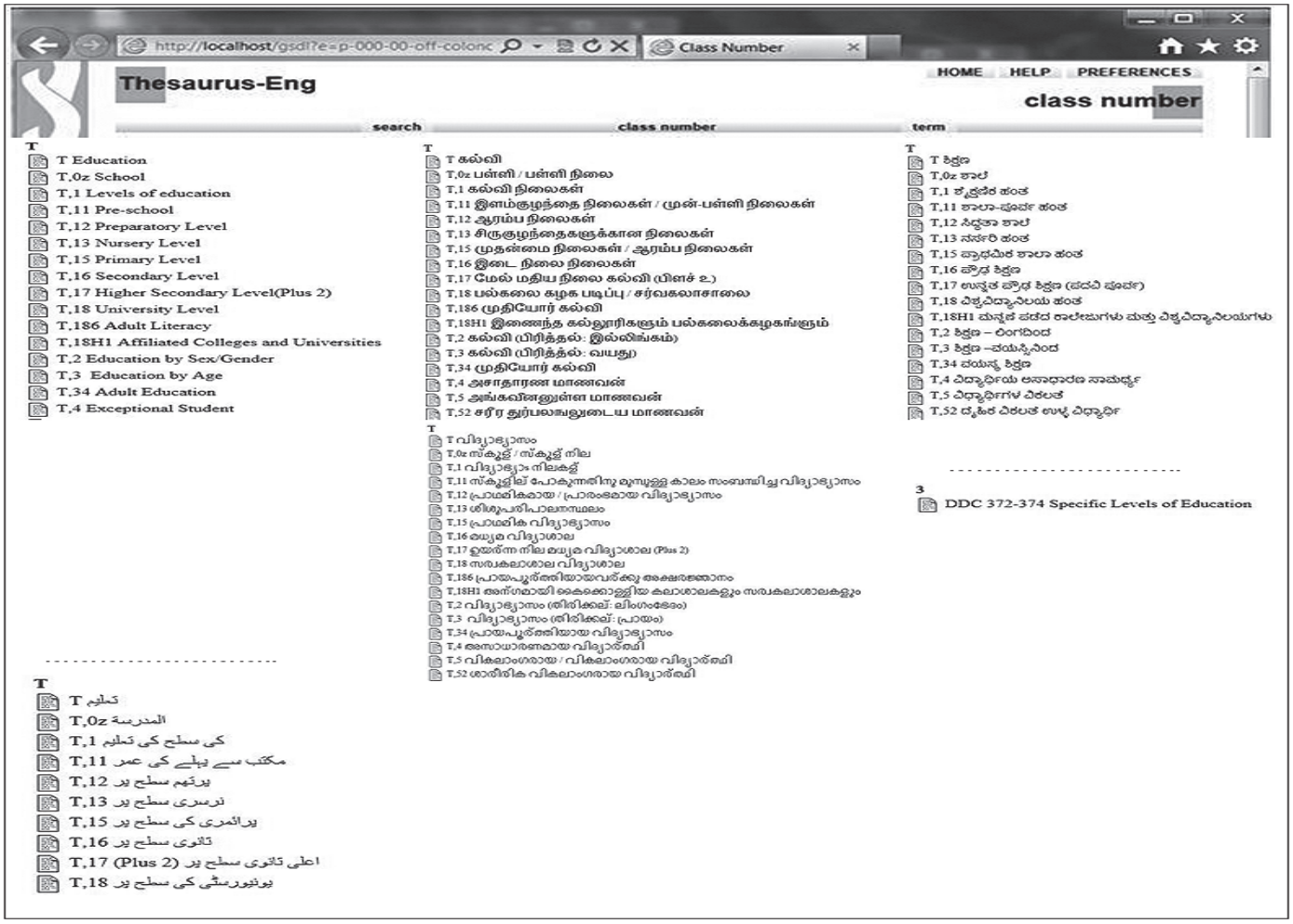

Figure 1. Thesaurus–Five languages and demo DDC line clicking on T, 15 primary level will retrieve as in Fig. 2

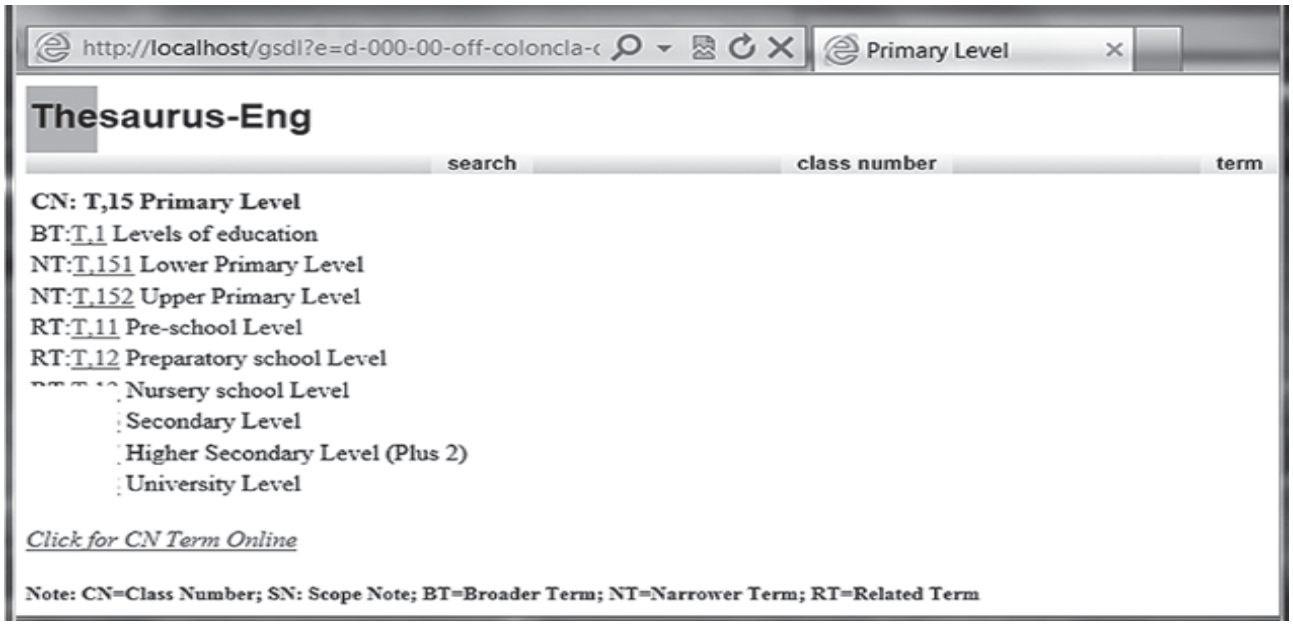

Figure 2. Clicking on T,12 Prepartory School Level in Fig. 1 will retrieve as in Fig. 2.

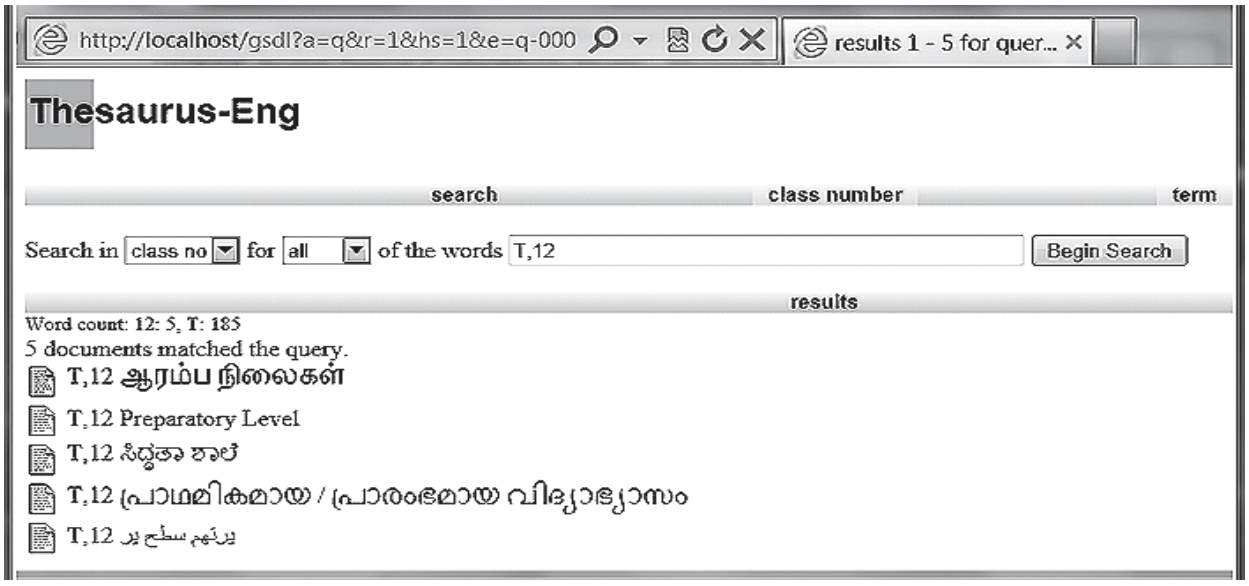

Figure 3. Clicking on T,12 which is hyperlinked internally will retrieve as in Fig. 3.

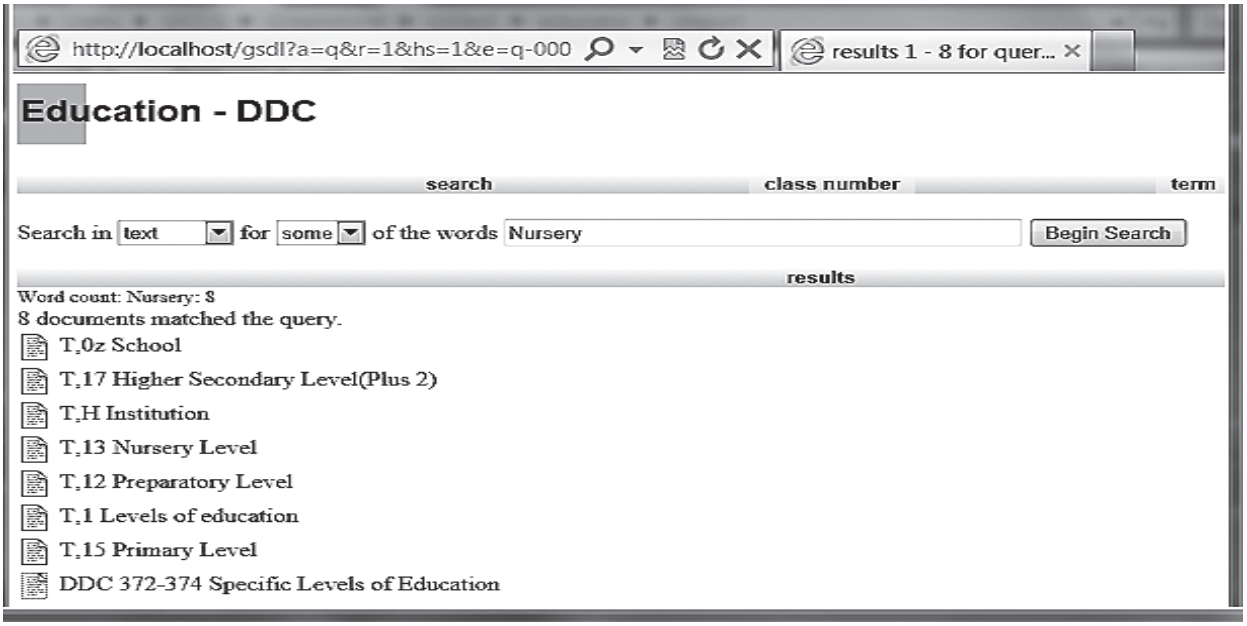

Figure 4. Clicking on the icon against T,12 ஆரம்ப நிலைகள் ı Fig. 3 will retrieve as in Fig. 4.

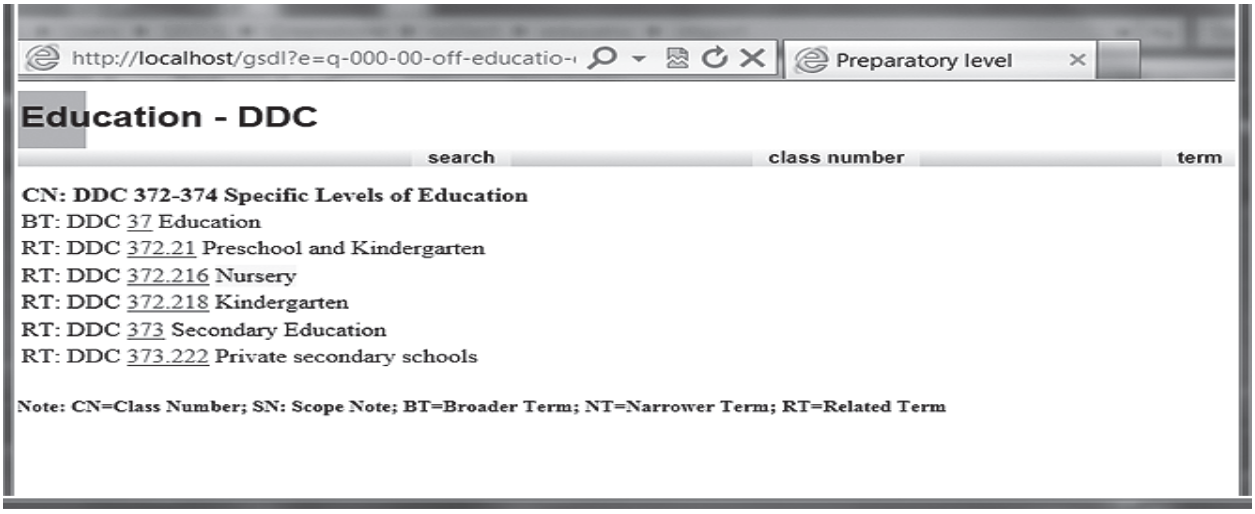

Figure 5. Showing search by DDC number.