2020

Developing and integration of an online thesaurus: with special reference to LGBTQIA related terms

SAYANI MUKHERJEE
sayani.mukherjee22@gmail.com

MUKHERJEE, SAYANI, "Developing and integration of an online thesaurus: with special reference to LGBTQIA related terms" (2020). *Library Philosophy and Practice (e-journal)*. 4078.
https://digitalcommons.unl.edu/libphilprac/4078

Developing and integration of an online thesaurus:
with special reference to LGBTQIA related terms

Abstract: The semantic web has opened a new arena in thesauri, as supporting tool for semantic searching and browsing. Vocabulary managing tools are permit to create, edit and also import and integrate thesauri. Integrality, interoperability and information retrieval are key role players with the advent of SKOS standards and ISO-25964, as a W3C recommendation. In this paper four open access (OA) vocabulary managing softwares are evaluated and compared and a domain specific thesaurus on LGBTQIA developed by TemaTres. Collect the LGBTQIA related terms from various universities of West Bengal and developed the thesaurus based on these terms. Finally integration has been done of this online thesaurus to OPAC and retrieved the documents by selecting vocabulary terms.

Key words: LGBTQIA, Thesaurus, SKOS, RDF, ISO-25964

0. Introduction:

Thesaurus is a semantic network of terms. It is a terminological control device used in translating from natural to systematic languages and a controlled and dynamic vocabularies of semantically and generically related terms which covers a specific domain. Thesaurus is a conceptual tool which permits to map different concepts and ideas of a given field of knowledge and ensure a consistent information storage and retrieval. The W3C standard for SKOS (Standard Knowledge Organization System) standard (Miles & Bachhofer, 2009), RDF/SKOS (Lacesta et.al., 2007; Faro et.al., 2008, Polo et.al., 2008) and ISO-25964 (ANSI/NISO) are the semantic web standards that help to achieve the goal of interoperability across today's immensely expanded and infinitely extensible network endeavor. New thesauri tools not only permit to create, edit, manage but also visually represent the thesauri. In this paper four vocabulary managing OA software are evaluated on the basis of compatibility with semantic web. A set of criteria have defined for comparison which include the purpose, functionality, web compatibility as components of information system.

1. Objectives:
The main objectives of this article are as follows –

➔ To collect the LGBTQIA related terms from web-OPAC of various universities situated in West Bengal. Then list up the top terms along with their Bengali synonymous terms.

➔ To create an online thesaurus with these terms by using an open source vocabulary manager tool. Then try to integrate this online thesaurus in local OPAC for searching and retrieving documents.

2. Literature review:

David Philip Norris (2015) created a document on LGBTQ-related Library of congress subject headings. Included is a brief history of LGBTQ cataloging. A review of existing LC subject guide, terminology used in LC for cataloging to describe LGBTQ individual and communities. Yale University Library (1701) derived a LGBTQIA subject heading list and divided into 15 parts. Such as, general, art and architecture, business, culture and identity, film and theater, health and medical, history, law and legal, literature, mass media, psychology, religion and spirituality, sociology, sports and fitness, youth and education. American Library Association (ALA) (1876) derived a GLBT Controlled Vocabularies and Classification Schemes. The provision of unambiguous and non-pejorative physical and intellectual access to library materials on gay, lesbian, bisexual and transgender people has been a goal of the GLBTRT since its inception. One of the first sponsored sessions of the Round Table (then the Task Force on Gay Liberation) was a panel discussion featuring Joan Marshall and Steven Wolf at the 1971 ALA annual meeting in Dallas. They tartly criticized the sexist and homophobic labeling which then prevailed in the Library of Congress Subject Headings and the Dewey Decimal Classification. A subject thesaurus of terms describing the Lesbian/Gay experience was established as an arm of the Gay and Lesbian task force during the annual conference in July of 1985. International

Association of Lesbian and Gay Archives and Libraries (IALGAL), also formed in 1985 by Robert B. Marks Ridinger, librarian at Northern Illinois University. Phil Parkinson, serials librarian at the Alexander Turnbull Library at the National Library of New Zealand and librarian of Lesbian and Gay Rights Resource Centre New Zealand,developed a Gay Decimal Classification (GDC). The Canadian Gay Archives (CGA) also worked upon Dewey decimal classification. The Task Force compiled The Pink Thesaurus published by the Hall-Carpenter Archives in London, Ridinger's work with the index of the Advocate, the Canadian Gay Archives' Alternative Press Index, The Lesbian Periodical Index from the Lesbian Herstory Archive and Parkinson's work with the Gay Decimal Classification(Gregg, 1986). The Gay and Lesbian task force published The International Thesaurus of Gay and Lesbian Index Terms in 1988, based on Gregg and Ridinger's years of work with the Thesaurus Working Group and input from gay and lesbian libraries and archives from across the globe. Gregg's impact on the Gay and Lesbian archival community is significant. Indiana University Bloomington derived a list which is based on subject headings compiled by John De Santis in 2000 as well as the "Library of Congress Subject Headings for Lesbian, Gay, Bisexual, Transgender, Queer (LGBTQ) Topics" compiled by Thomas Bolze at Yale University Library (1701), and updated by Deborah LaFond and Suny Albany in 2010. Again it was updated in the month of February 2016. The secular librarian derived a compendium of Lesbian, Gay, Bisexual, Transgender, and Queer-

related authorized Library of Congress subject headings from the LC Authorities. This list is important to me professionally as a cataloger and librarian, as well as personally as a gay man. The language we use to catalog and classify our literature and media matters, and incorrect or outdated language can create access barriers to those resources. So it is critical that we study and are familiar with these headings. This compendium is hopefully a step in that direction. Lacasta et al. (2007) demonstrated a SKOS model with the help of an open source tool ThManager. They also map GEMET thesaurus with Dublin Core metadata .

3. Methodology:

To carry out this study the methodology is as followed – i) Few controlled vocabulary managing software are available in open source domain have been selected, ii) Selected software are evaluated and compared based on some essential parameters, iii) From this study choose the most useful and comprehensive controlled vocabulary managing tools and a domain specific thesaurus developed by using that tool. General common top terms of the specific domains are selected from by consulting and evaluating both subject heading lists. As the part of study the web OPAC Of different universities of West Bengal have been searched so the bengali synonymous terms are also studied. The top English terms along with Bengali synonyms are listed below:

Table 1: Top terms and their Bengali synonymous terms

BENGALI TERMS	ENGLISH TERMS	BENGALI TERMS	ENGLISH TERMS
মানবিক প্রবৃত্তি	Human behavior	যৌন পরিচয়	Sexual identity
যৌন আচরণ	Sex role	লিঙ্গ সাদৃশ্য	Gender conformity
সমকাম	Homosexuality	লিঙ্গ পরিচয়	Gender identity
বিষমকাম	Heterosexual	লিঙ্গ বৈসাদৃশ্য	Gender nonconformity
নিষ্কাম	Asexual	ট্রান্স সেক্সুয়্যাল	Transexuals
উভকাম	Intersex	যৌন সংখ্যালঘু	Sexual minorities
সর্বকাম	Pangender	সমকামী-পুরুষ	Gays
রূপান্তরকাম	Transexual	সমকামী-স্ত্রী	Lesbians
দ্বৈতকাম	Bisexual	যৌন অভিমুখীতা	Sexual orientation
সিস্ জেন্ডার	Cisgender	যৌনতা	Sexuality
প্রান্তিক যৌনতা	Gender minorities		

4. Criteria for evaluation:

Primarily, purpose/description for which the software created, development stage – whether it is ready for use or not and general impression are considered. Then technical support like supported programming language, software type – backend, middleware, framework, complete solution, batch import, batch editing of controlled vocabulary term, linked open data publishing, browse compatibility, advance search mechanism, data visualization, SPARQL endpoint and finally ease of installation for digital infrastructure development are considered.

4.1. Evaluation of vocabulary managing tools:

❖ **Apache Jena[1]:** A free, open source software for building semantic web and linked open data application. Interact with the core API to create and read Resource Description Framework (RDF) graphs. Serialize your triples using popular formats such as RDF/XML or Turtle. The features are –

- Ready for testing: No
- General impression: Not investigated
- Technical support requirement : Java
- Software types: Framework
- Batch import support: Not supported
- Batch editing of controlled vocabulary term: Not supported
- LOD publishing: Not supported
- Browse compatibilities : Not supported
- Advance search capabilities: Not supported
- Data visualization: Not supported
- SPARQL endpoint: ARQ for SPARQL 1.1 (Query your RDF data using ARQ, a SPARQL 1.1 compliant engine. ARQ supports remote federated queries and free text search.)

❖ **TemaTres[2]:** An open source vocabulary server, web application to manage documentation language and exploit vocabularies, thesauri, taxonomies and format representation to knowledge.

- Ready for testing: Yes
- General impression/ testing notes: Good for simple vocabularies. Additional granularity could be accomplished by putting more information in notes and meta-terms, and/or adopting custom URL scheme to contain additional elements of descriptive information about our vocabulary.
- Technical support requirements: Apache, PHP, MySQL

- Software types: Complete solutions.
- Batch import support: Simple method for batch import for names only. Would then need to edit names individually to have additional information that we've already developed
- Batch editing of controlled vocabulary term: Only available through find and replace operations
- LOD publishing: Supports serialization in multiple formats such as DC, MADS, SKOS-Core, VDEX, XTM, Zthes, JSON, JSON-LD.
- Browse compatibilities: Yes.
- Advance search capabilities: simple advanced search form.
- Data visualization: here might be some functionality with additional plugins such as https://code.google.com/archive/p/tematresvisualvocabulary.
- SPARQL endpoint: Yes.
- Ease of Installation for Digital Infrastructure Development.
- Simple PHP-based app.
- Minor tweaks and updates probably relatively easy.
- Most comments in the code aren't in English.
- A bug in the code caused installation instructions.
- Required special configuration options to be turned on to create its DB in MySQL.

❖ **SKOS-mos[3]:** Open source web-based SKOS browser and publishing tool. Access to controlled vocabularies for indexing, information retrieval and vocabulary development. The features are –

- Search and browse vocabularies, alphabetical index, thematic index, structured concept display, visualized concept hierarchy, multilingual user interface.
- Ready for testing: No
- General impression / testing notes: Not tested in sandbox environment after reviewing specifications. Would be useful if our vocabulary was more like a thesaurus.
- Technical support requirement: From website: one or more SKOS vocabularies PHP capable web server SPARQL triple store (Apache Jena Fuseki with jena-text) HTTP cache (e.g. Varnish) – optional but recommended.
- Software types: complete solution.
- Batch import support.
- Batch editing of controlled vocabulary term.
- LOD publishing: Yes.
- Browse compatibilities: yes – alphabetical and hierarchical.
- Advance search capabilities: No.

- Data visualization: Not supported
- SPARQL endpoint: YES

❖ **ThManager[4]:** ThManager is an open source tool for creating and visualizing SKOS RDF vocabularies, a W3C initiative for the representation of knowledge organization systems such as thesauri, classification schemes, subject heading lists, taxonomies, and other types of controlled vocabulary. ThManager facilitates the management of thesauri and other types of controlled vocabularies, such as taxonomies or classification schemes. Multi-platform (Windows, Unix). As it has been developed in Java and the storage of metadata records is managed directly through the file system, the application can be deployed in any platform with the minimum requirement of having installed a Java virtual machine. Multilingual, selection and filtering of the thesauri stored in the local repository, metadata in compliance with a Dublin Core based application profile for thesaurus. These metadata can be either visualized in HTML or edited through a form, Edition of thesaurus content. The tool provides an edition interface to modify the content of a thesaurus: creation, deletion, updating of concept, exchange of thesauri according to SKOS format. The export operation includes the export of thesaurus metadata.

- Ready for testing: No
- General impression / testing notes: Combination of Dublin core application profile for thesauri and SKOS. Might be best suited for a true thesaurus with broader and narrower terms, than the type of data associated with personal and corporate names.
- Technical support requirement: Java
- Software types: Complete solution for SKOS vocabularies
- Batch import support: Yes - can batch import SKOS records.
- LOD publishing: Thesaurus data can be exported as RDF/XML
- Browse compatibilities : yes, a number of viewers are provided
- Alphabetic viewer: It provides the list of thesaurus concepts alphabetically ordered in the selected language.
- Hierarchical viewer: It provides a tree showing the hierarchical structure of thesaurus concepts.
- Concept viewer: For a selected concept it shows all the properties allowing additionally the navigation to the related concepts by means of hyperlinks.
- Search tool: It facilitates search of concepts. The searching process is based on preferred labels allowing the following criteria: 'equals', 'starts with' and 'contains'.

4.2. Findings: Comparison of above four vocabulary managing tools is following table –

Software	Apache-Jena	TemaTres	SKOSmos	ThManager

Technical support requirements	0.5	1	1	0.5
Batch import support	0	1	0	1
Batch editing of controlled vocabulary terms	0	1	0	1
LOD publishing	0	1	1	1
Browse compatibilities	0	1	1	1
Advance search capabilities	0	1	0	0
Data visualisation	0	1	0	1
SPARQL endpoint:	0.5	1	1	0
Total	**1.5**	**8**	**4**	**5.5**

[Fully support – 1, Partial support – 0.5, Not supported – 0]

- Firstly, there are not so many open source vocabulary managing tools available. In this paper only four could be evaluated. Among four only TemaTres is reday for testing where rest are in developing stages.

- TemaTres is supporting all the required features for semantic web networks where SKOSmos, ThManager are also in a considerable stage rather Apache Jena.

- Interoperability has received the main attention for information exchange. SKOS/RDF semantic web standard are not yet supported in all tools.

- Thus, TemaTres is highly recommended on the basis of above parameters. TemaTres is oriented to the development of hierarchical thesauri. TemaTres presents both semantic and hierarchical relationship. Terms can be annoted with scope notes. Though there is no option for import data but data can be exported to text format, Dublin Core, SKOS-core, Zthes, MADS, XTM, etc.
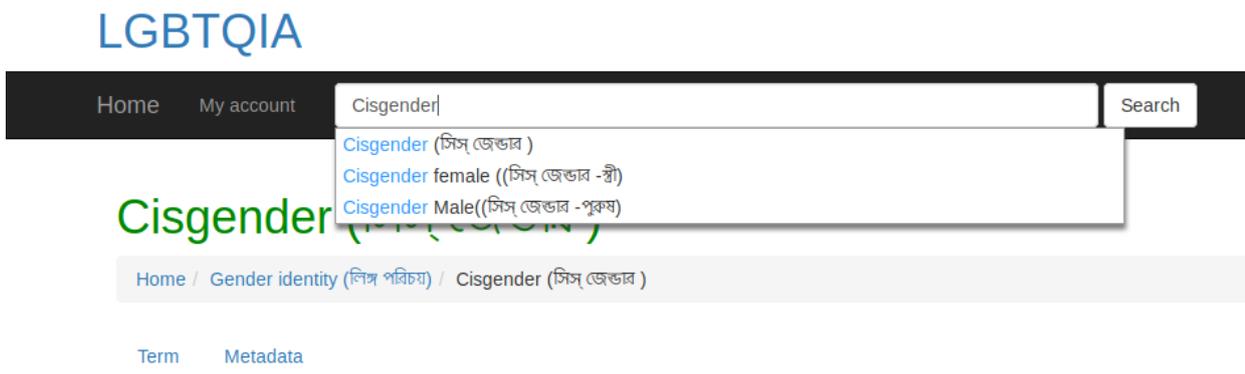
5. Development of a domain specific thesaurus through TemaTres:

A domain specific thesaurus has been developed by TemaTres. The thesaurus is on LGBTQIA – Lesbians, Gays, Bisexuals, Transgender, Queer, Intersex, and Asexual. Terms are mainly derived from Yale University LGBT subject heading list and The Secular Librarian subject heading both are based on Library of Congress subject headings lists (LCSH). To develop this thesaurus TemaTres has been downloaded, terms are added as per hierarchical structure, linked with Koha. After linking terms can be searched through Koha – OPAC.

5.1. Software related steps:

- Download TemaTres software.(https://sourceforge.net/projects/tematres/);

- Build collection according to hierarchy with top terms along with bengali synonymous terms within parenthesis.

- Download visual thesaurus and link with TemaTres (www.vocabularyserver.com)

- Path to download:Get TemaTres > documentation in English In search box search with 'visual vocabulary' > open a new page > instructions of downloading will be there > click on latest version;

- After downloading the software,zip file of the same will be stored in downloads folder of other location> extract the file and rename as 'vis' and copy and paste into other location > thes;
- vis > common > config.php;
- Open TemaTres edit the url;
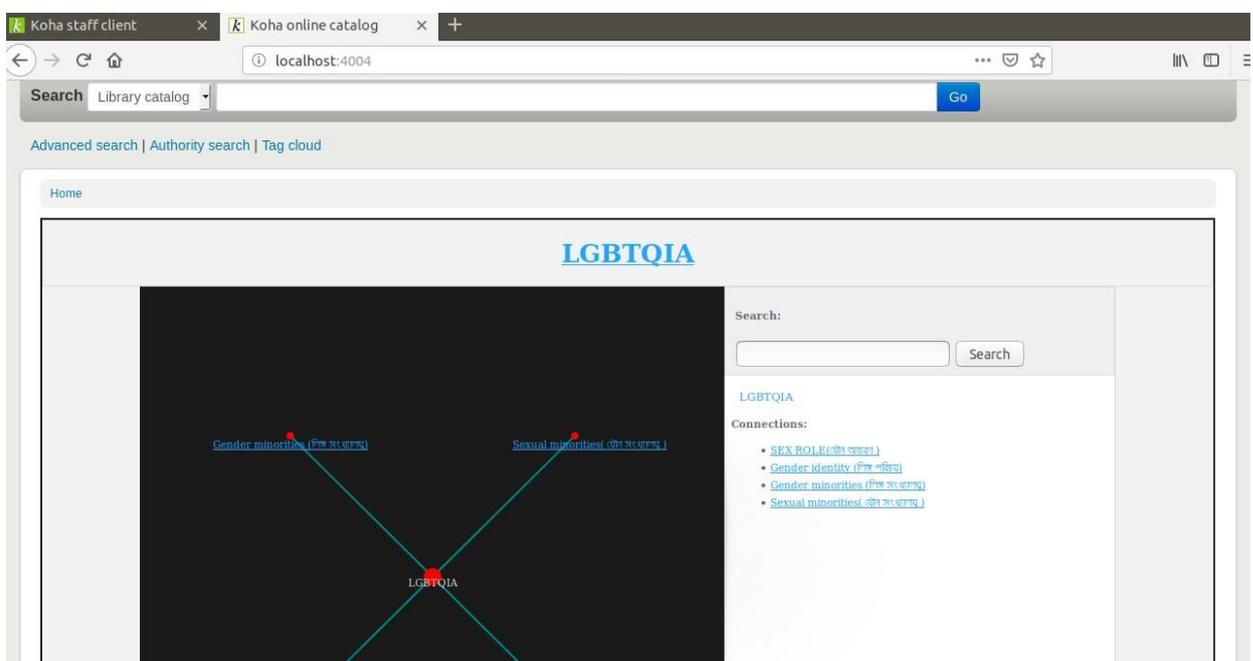- Mapping with koha – OPAC by editing Global system preference of Koha administration.

Picture 1: Thesaurus on LGBTQIA

The above picture shows the search process of specific terms and how it appears in TemaTres in a hierarchical manner according to Broader term (BT), Narrower terms (NT), Related terms (RT).

: Thesaurus in Koha OPAC (Visual representation of the top terms)

6. Conclusion:

In today's digital domain search terms are key role players for retrieval of appropriate and relevant document. Subject headings help to collocate document under specific subject terms and Thesauras provides the specific subject domain under one umbrella. Vocabulary managing softwares incorporates the systematic knowledge organisation and support state of the earth technologies to retrieve precise information. TemaTres is the only vocabulary managing software which is in matured state and open access. The subject of the developed thesauras is a very prominent,buzzing and trending too. As we know about the struggle of the respective community and how they had achieved their liberty and equality in society. But being a part of the society specially academic libraries of Indian leading institutes of social sciences are very much reluctant. Most of the universities donot have web OPAC.Those who are having the search results are not robust and number of retrieve documents are either very few in number or not prominent. The purpose of this paper is to compare the features of four above mentioned softwares and development of a thesaurus by using the most appropriate one. To conclude, more vocabulary managing softwares need to be developed or development or further research is required for the exsisting one.In the age of semantic web, open source, well equiped vocabulary managing tools are the only solution for appropriate retrieval.

Reference

[1]https://jena.apache.org/
[2]https://www.vocabularyserver.com/
[3]http://skosmos.org/
[4]http://thmanager.sourceforge.net/
https://libraries.indiana.edu/
https://secularlibrarian.com/lgbtq-lcsh
https://guides.library.yale.edu
https://itspronouncedmetrosexual.com/2013/01/a-comprehensive-list-of-lgbtq-term-definitions/

Lacasta, J., Nogueras-Iso, J., López-Pellicer, F. J., Muro-Medrano, P. R., & Zarazaga-Soria, F. J. (2007). ThManager: An Open Source Tool for Creating and Visualizing SKOS. *Information Technology and Libraries, 26*(3), 39-51. doi:10.6017/ital.v26i3.3274

Norris, D. (2016). *Queerying the catalog: LGBTQ subject headings*. Retrieved from https://www.slideshare.net/DavidPhilipNorris/lgbtq-subject-headings-presentation on 10 March 2020

Miles, A. & Bachhofer, S. (2009). *SKOS Simple Knowledge Organization System RDF Schema.* Retrieved from https://www.w3.org/TR/2009/CR-skos-reference-20090317/skos.html on 10 March 2020