ISSN (Print): 0972-2467
ISSN (Online): 0976-2477
SRELS Journal of Information Management, Vol 55(1), February 2018, p. 8-12 DOI: 10.17821/srels/2018/v55i1/117931

Integrating HASSET Thesaurus with ICSSR Data Service

Miteshkumar Pandya*

INFLIBNET Centre, Gandhinagar, Gujarat - 382007, India; miteshpandya21@gmail.com

Abstract

This article studies effective use of thesaurus which provides access to wide range of subject keywords and pick right one which is best suited to describe the dataset being indexed. The article provides brief information on ICSSR Data Service followed by aims and objectives. It also elaborates need for the integrating thesaurus with a data input worksheet of the online databases. It also enlists steps of assigning subject keywords to a dataset being indexed in the repository.

Keywords: Humanities and Social Science Electronic Thesaurus (HASSET), ICSSR Data Service, Indexing Languages, Thesaurus

1. Introduction

Indian Council of Social Science Research ICSSR Data Service (2017)[2] hosts a comprehensive set of statistical datasets in social sciences generated and contributed by the Ministry of Statistics and Programme Implementation (MoSPI), New Delhi, social science institutes under direct purview of ICSSR and other Government agencies. The project on setting-up of ICSSR Data Service is executed by the INFLIBNET Centre with funding from ICSSR to support researchers, teachers and policymakers who heavily rely on high-quality socio-economic data for their research. ICSSR has signed a MoU with MoSPI to provide datasets for hosting on this platform to provide a single point access to a wide range of primary datasets including datasets generated by large-scale government surveys i.e. ASI and NSS that provide unit level data as well as qualitative studies. The ICSSR Data Service, as a policy, promotes data sharing to encourage the reuse of data and provide information on developing and generating social science research data and its management.

2. Aims and Objectives of ICSSR Data Service

The ICSSR Data Service[2] is the first of its kind data repository in India which hosts a large amount of socio-economic and industrial statistical data in raw format as well as processed form. The first and foremost objective of this portal is to provide seamless and integrated access to a wide range of datasets generated by the MoSPI, New Delhi, social science institutions under direct purview of ICSSR and other Government organizations, researchers who are looking for high quality social and economic research datasets, etc. It aims to serve as a national research data sharing platform to facilitate sharing, use and reuse the socio-economic data by the social science research community in India. Following are the major aims and objectives:

- To serve as a national data service for promoting powerful research environment through sharing and reuse of data among social science community in India;
- To acquire, process, organize, preserve and host research data and its metadata along with Extract, Transform and Load (ETL) facilities of raw data in social sciences and related domains collected from diverse sources for easy sharing and access;
- To facilitate online submission, access, search, browse, discovery, conversion, analysis and visualization of data through intuitive interfaces;
- To impart training and spread awareness about benefits of data sharing and reuse amongst social science research community in India; and
- Interact, cooperate and collaborate with other national and international data services and repositories for data and resource sharing and

*Author for correspondence

improved management of data services.

3. Statement of Problem

The datasets hosted into the repository would be more meaningful if it is made discoverable using a variety of approaches. While on one hand, it is essential to organize datasets into appropriate categories and collections, on the other hand, datasets should be searchable using controlled vocabulary as well as free-text search. There are two methods of indexing datasets, i.e. 1. using controlled vocabulary/thesaurus; 2. using uncontrolled vocabulary or free-text search. In the absence of the standard thesaurus for describing Indian social science datasets, it was decided to use HASSET thesaurus for indexing all the datasets hosted on ICSSR Data Service.

4. Need for the Integrating HASSET with ICSSR Data Service

With the advent of Information and Communication Technology (ICT), the researchers approach has also changed. The socio-economic, political and census research data generated by various agencies are available in public domain for use and reuse by social science researchers. There are a number of datasets hosted on various repositories and platforms and it is a challenging task for a researcher to get the appropriate and authentic dataset for his/her research work from the repositories. The datasets hosted on various platforms will be meaningful if it is retrieved efficiently and as a ranked output. The repository provides facility to upload datasets along with metadata from a remote machine connected with the network. Many times, while indexing online resources in a database, it is difficult to provide hard copy of thesaurus to every indexer and therefore, integrating entire thesaurus with indexing database will ease the job of the indexer. The thesaurus is one of the tools for indexing which helps indexer as well as researcher to retrieve appropriate dataset within time. Integration of HASSET thesaurus with ICSSR Data Service will enable researchers to find their desired data using pre-defined keywords. Integration of searchable thesaurus interface will result into providing seamless access to the terms stored with all its semantic relations in a database.

5. Review of Literature

There are a few attempts at integrating thesaurus with data input sheet of various databases. Manjunath and Sangam observed that online indexing databases should have facility to browse standard subject headings while indexing literature[3]. Indexing is a crucial process and should be done with utmost care. Moreover, IGIDR, Mumbai have entered into an agreement with German Social Science Infrastructure Services (GESIS), Bonn and entire thesaurus was shared / provided by the institute in tabbed separated text file for integrating with their database. Entire thesaurus along with all its semantic relations have been converted into MySQL database and integrated with Open Index Initiative (OII) data input worksheet. Oza and Pandya (2016)[4] have successfully integrated Library of Congress Subject Heading with data input sheet of IndCat's Theses database. They have also discussed and elaborated all the steps for assigning subject descriptor to every thesis being indexed in the database.

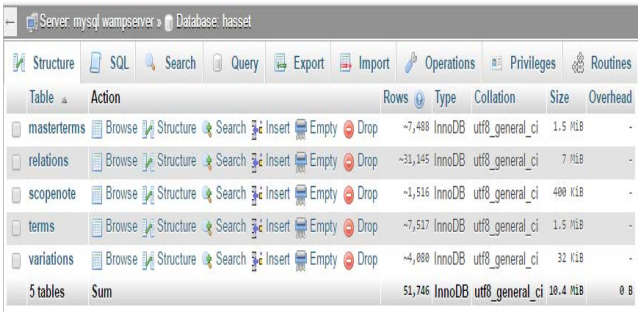

Figure 1. Tables in HASSET database.

6. Humanities and Social Science Electronic Thesaurus

The HASSET is one of the well-known thesauri in the area of social sciences devised by UK Data Service. HASSET Thesaurus (2017)[1] has also been used as base thesaurus for compiling European Language Social Science Thesaurus (ELSST). The entire thesaurus with all its semantic relations has been prepared according to conform to international standards and it is interoperable across different systems. ICSSR Data Service uses HASSET Thesaurus for indexing social science datasets. INFLIBNET Centre and ICSSR have signed License Agreement with UK Data Service for integrating thesaurus with its metadata input sheet.

With signing of license agreement, entire thesaurus with total number of 11,678 keywords, including 7,605 descriptors and 4,073 non-descriptors were provided to INFLIBNET Centre. Entire thesaurus was provided as a single ***.nt*** file which is a plain text format for encoding RDF graph. The HASSET thesaurus was converted into SKOS

based RDF/XML format using online converter available at http://www.easyrdf.org/converter. A PHP script was written to decode and store all the keywords with all its semantic relations into ***MySQL*** database. The MySQL database containing HASSET Thesaurus was stored in the server hosting ICSSR Data Service. As shown in Figure 1 below, the database contains all the keywords along with 1517 scope notes with respective keywords in five different MySQL tables, namely *masterterms, relations, scopenotes, terms, variations*.

7. Assigning Subject Keywords using HASSET

Steps taken for choosing/assigning appropriate keywords in ICSSR Data Repository using HASSET Thesaurus are as follows:

1. Log into the ICSSR Data Service system module with administrative privileges using appropriate credentials. As soon as a user is logged-in with administrative rights, he / she is allowed to create or update metadata. As shown in the Figure 2 below, the selected dataset belongs to "Domestic Tourism". The "edit metadata" option is available on the right side navigation bar that redirects a user to metadata input sheet.

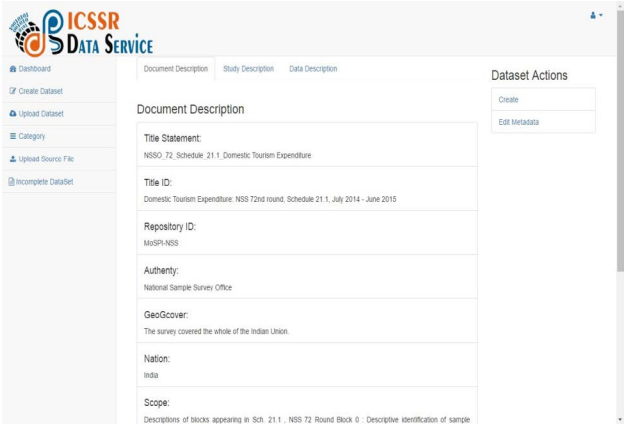

Figure 2. Administrative user interface for ICSSR data service.

2. As per the system structure of NADA software, the subject-related information is being stored in study description tab as shown in the Figure 3. User needs to expand the study description field by clicking on + symbol to insert/select subject field.

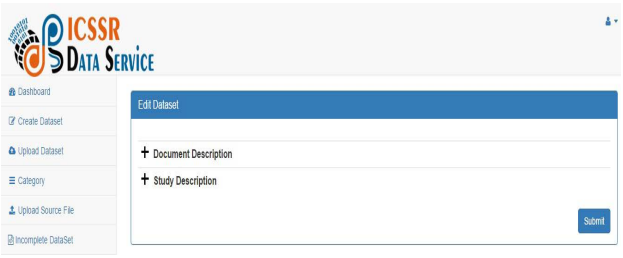

Figure 3. Insert/Update metadata elements.

3. As soon as the field is expanded, users can see a link named "Click here for HASSET" as shown in Figure 4. On clicking at the link; a new popup window opens as shown in Figure 5.

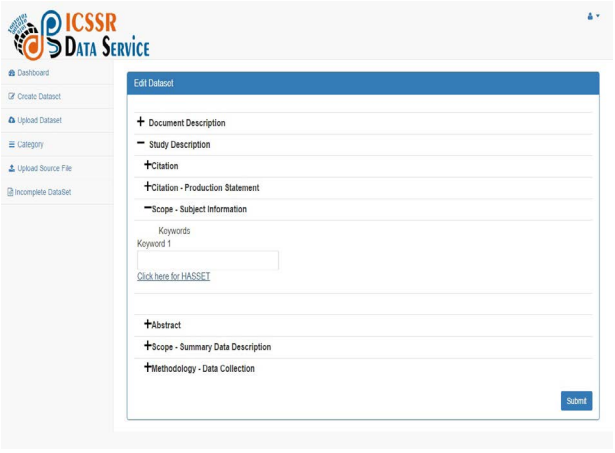

Figure 4. Data input sheet for keywords.

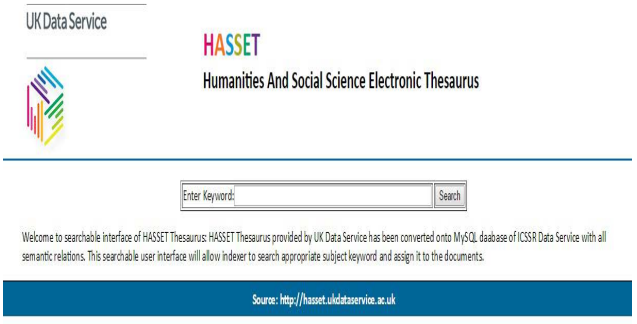

Figure 5. Integrated search interface of HASSET.

4. The Figure 5 is an integrated search interface for the HASSET thesaurus stored on ICSSR Data Service. This interface facilitates users to search desired keyword from the database. The system displays the

results related to the keyword if it is available in the thesaurus.

5. The selected dataset is related to "Tourism Expenditure". On searching "tourism", the system retrieved all the terms which are related to tourism as displayed in the Figure 6.

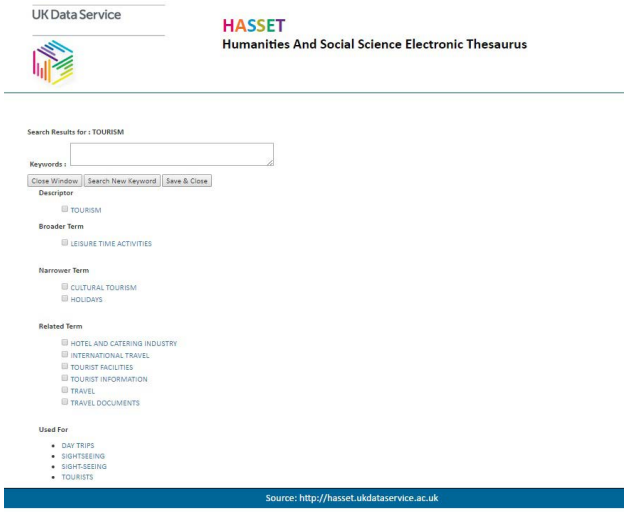

Figure 6. Search results.

6. Users can select appropriate keyword(s) suitable to describe content of dataset. As soon as user selects keyword by clicking on check box displayed on the left hand side, the selected keyword(s) get copied in the text box with # as separator as shown in the Figure 7.

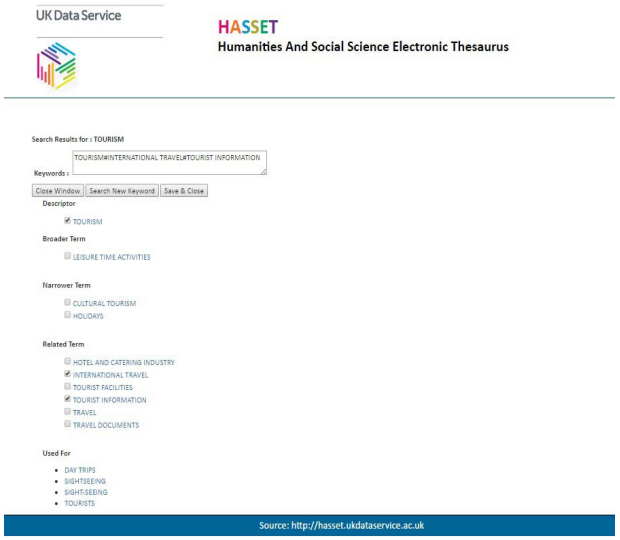

Figure 7. Selection of appropriate keywords.

7. The interface also has a feature to search with alternate keyword(s) if relevant results are not retrieved. As soon as user clicks at "Save & Close" button, the system creates text boxes for each of the selected keywords and automatically all the terms get copied into respective fields as shown in the Figure 8.

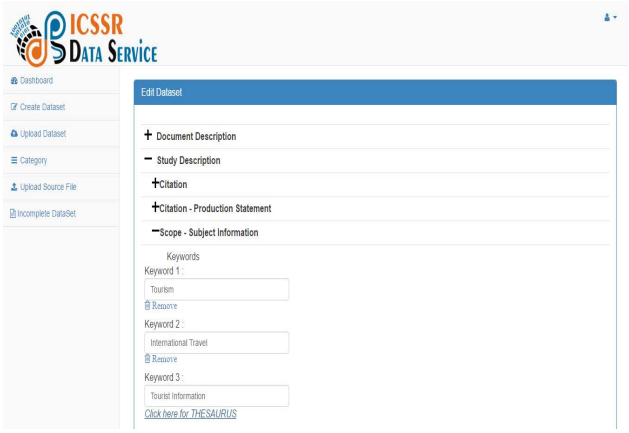

Figure 8. Automatic insertion of subject headings into keyword field.

8. Conclusion

This article describes need for the integrating thesaurus with online databases and process of assigning subject keywords to datasets hosted on ICSSR Data Service using HASSET thesaurus. While searching desired datasets from the repository, if standard terms are used, the result will be refined and displayed on the basis of relevancy rank. The main aspect of integrating HASSET thesaurus with ICSSR Data Sercice is to provide standard terminology for describing the datasets hosted on the repository. From the perspective of library professionals, they will be able to browse entire thesaurus with all its semantic relations and from the end users perspective, they will be able to search/browse their desired dataset by keying-in standard keywords. This will result in effictive and efficient retrieval of social science datasets hosted on the platform.

9. Bibliography

1. HASSET Thesaurus. (2017). Retrieved August 29, 2017. Available at: https://hasset.ukdataservice.ac.uk/.
2. ICSSR Data Service. (2017). Retrieved August 29, 2017. Available at: http://icssrdataservice.in
3. Manjunath GK and Sangam SL. (2007). Integrating an online thesaurus with Open Index Initiative: A case study. In:

Libraries without Boundaries: Reaching the Unreachable in Knowledge Era, 20-23, November 2007, New Delhi, India, Ed. by Kaul, S and Kaul S.K, 2007, p. 246–55.
4. Oza ND and Pandya MY. (2016). Integrating library of congress subject heading to IndCat's theses database: A case study, Journal of Library and Information Science. 10(1):68–77. Crossref
5. Pandya MY. (2016). Thesaurus development for Indian social science literature on relational database management system and its integration with OII. Sardar Patel University, Vallabh Vidyanagar.