DESIDOC Journal of Library & Information Technology, Vol. 43, No. 4, July 2023, pp. 219-225, DOI : 10.14429/djlit.43.4.19227

Machine Learning Applications in Digital Humanities: Designing a Semi-automated Subject Indexing System for a Low-resource Domain

Roshni Mitra* and Parthasarathi Mukhopadhyay

Department of Library and Information Science, University of Kalyani, Kalyani, Nadia, West Bengal- 741 235, India
**E- mail: roshnimitrakly@gmail.com*

ABSTRACT

This research study explores the potential of machine learning tools and techniques to organize knowledge objects pertaining to various aspects of the gender spectrum (LGBTQIA+) in order to address the low-resource features of the LGBTQIA+ knowledge domain in Indian libraries. It aims to develop a semi-automated subject indexing system using an open source machine learning framework (Annif) and deploying the Homosaurus, a domain-specific vocabulary system. It develops programmatically a comprehensive training dataset from open-access bibliographic data sources with the help of data carpentry tools and NLP services from OpenAI. The study also measures the efficiencies of the automated indexing framework and investigates the potential for widespread adoption of a REST/ API call-based approach for rapid indexing of a substantial number of records related to the LGBTQIA+ domain.

Keywords: Annif; Homosaurus; Inclusive librarianship; Large language model (OpenAI); LGBTQIA+; Machine learning; Retrieval metrics

1 INTRODUCTION

Digital humanities is a broadly defined area of study, but most definitions emphasise the utilisation of digital technologies to create, connect, interpret, collaborate, investigate, and acknowledge the diversity of human culture.[1-4] With the aim of investigating the intersection of inclusive librarianship and Digital Humanities, this research study endeavors to explore the potential of machine learning tools and techniques in organising knowledge objects pertaining to various aspects of the LGBTQIA+ domain. In academic libraries of India, LGBTQIA+ resources are low in number, for example, the union catalogue IndCat (indcat.inflibnet.ac.in) has produced only 1,349 book records (out of 2,08,33,966 book records as on 28th February 2023 – 0.0065 % only) against a broad free search in the subject field (Lesbian OR Gay OR Bisexual OR Transgender OR Queer OR Intersex OR Asexual OR LGBT*), whereas a similar broad subject search in the WorldCat (worldcat.org) has provided 1,27,794 book records (print & e-books) out of 405 million book records (0.032 %). Similarly, the OPAC of the National Library, India (nationallibraryopac.nvli.in) has listed a mere 668 book records against the same search (out of 15,54,997 book records as on 28th February 2023). Another issue in processing the LGBTQIA+ resources in Indian libraries is the use of generic knowledge organisation systems like Library of Congress Subject Headings List (LCSH), Dewey Decimal Classification and so on. The widely used Knowledge Organisation Systems (KOSs) in libraries around the world, such as LCSH and DDC, perpetuate sexist and homophobic attitudes. Moreover, the traditional approach to classification and cataloging continues to be Eurocentric, Christocentric, patriarchal, and heteronormative.[5-8] Sexual prejudice, akin to other forms of prejudice, is a mindset that leads to negative assessment of certain individuals or social groups, resulting in animosity and aversion.[9] This leads to a flawed retrieval system, making knowledge resources inaccessible to marginalised communities.[10]

This research study, keeping in view the low-resource features of the LGBTQIA+ domain, attempts to develop a semi-automated subject indexing system by deploying Homosaurus as a domain-specific vocabulary system and by applying a comprehensive set of training dataset developed programmatically by using data wrangling tools and ODbL- based bibliographic data sources. The

Received : 02 February 2023, Revised : 21 April 2023
Accepted : 30 June 2023, Online published : 04 September 2023

Homosaurus is a globally-known vocabulary standard for the LGBTQIA+ area of study[11], available in the public domain, supports different RDF serialising export formats (like N-Triples, JSON-LD & TTL), and much comprehensive in comparison to generic vocabularies like LCSH (see Table 1).

Table 1. Homosaurus vs LCSH

LCSH Terms	Homosaurus terms
Older sexual minorities	Older lesbians Older queer people Older bisexual people Older gay men Older (LGBTQ)
Gender-nonconforming people	Agender people Non-binary people Pangender people
Closeted gays	Closeted bisexual people Closeted gay men Closeted lesbians

(source: Mukhopadhyay& Mitra)

2 OBJECTIVES

This study is based on three components – a) Homosaurus as a vocabulary device[12]; b) short text corpus for the purpose of training of the selected machine learning backend by fetching bibliographic data elements like title and abstract/summary over REST/API call from ODbL-based bibliographic data sources[13]; and c) an open source machine learning framework named Annif developed by the National Library of Finland.[14] In this context, the fivefold objectives of this research study are:

- To select and deploy a comprehensive domain-specific vocabulary (here Homosaurus) in an open source machine learning framework (here Annif) by converting the vocabulary standard into a SKOS-compliant format as prescribed by Annif;
- To develop a comprehensive training dataset (at least 100K bibliographic records with short text corpus – title, abstract or summary, corresponding keyword(s) from Homosaurus, and URI(s) of the keyword(s);
- To utilise free API based NLP services from OpenAI to define scope of undefined concepts in the Homosaurus (around 100 such undefined concepts in a set of 2306 descriptors of the Homosaurus version 3.3 released in December 2022), to extract keywords to finalise inclusion of the records in the final training dataset, and to create concise abstract from an extremely lengthy abstract;
- To measure efficiencies of the automated indexing framework on the basis of a set of retrieval metrics like Recall, Precision, F1@5, NDCG (Normalised Discounted Cumulative Gain) etc; and
- To investigate the potential for widespread adoption of an automated indexing framework, which utilises a REST/API call-based approach for rapid indexing of a substantial number of records related to the LGBTQIA+ domain.

3 METHODOLOGY

This study, apart from Annif, deploys data carpentry tool (OpenRefine-an open source data wrangling tool), data wrangling activities to gather content from ODbL-based bibliographic data sources through REST/API call and the GPT-3 based language model for fine tuning the datasets. The entire methodology from vocabulary integration to configuration and training is discussed here under four major heads.

3.1 Project Configuration

The installation of Annif is a straight forward process. It needs a Linux-based OS (Ubuntu version 22.04 LTS is the platform for this research study), Python programming environment (version 3.8+), PIP (version 21.1+), Python virtual environment and NLTK Punkt (Sentence tokeniser). All necessary tools like backend algorithms, analysers are installed automatically during the installation of Annif. The post installation tasks are - a) adding a vocabulary; b) selection of an appropriate backend algorithm; and c) assortment of a suitable analyser. This research applied Simplemma analyser for all machine learning backends. It is a machine learning tool that performs rule-based lemmatisation for various languages.

3.2 Vocabulary Integration

As previously stated, a domain-specific vocabulary standard - the Homosaurus - is chosen and implemented in the Annif framework. The process is not an easy one, as the RDF formats of the Homosaurus available for downloading are SKOS-compliant, but the Annif presently does not understand the syntax followed there. The problem has been solved by using the CSV format of the Homosaurus and then converting the CSV file into a Skos-compliant TTL file. The TTL file as developed accepted by the Annif framework and the load-vocab command created the vocabulary support system for this project.

3.3 Preparation of Training Datasets

The most challenging task for this project was to develop a comprehensive training dataset that uses the Homosaurus vocabulary standard for indexing the subject content of the resources related to the LGBTQIA+ domain. In fact, no such bibliographic dataset exists presently. It has been decided after much deliberation and exploring different possibilities that this study will develop training datasets in two levels: a basic dataset containing terms and term definitions alongside term URIs (term URI and URIs of the related terms, narrower terms and broader terms); and a bibliographic dataset by gathering 25 records for each term or descriptor included in the Homosaurus vocabulary by fetching data from four ODbL-based bibliographic data sources, namely CoRE, CrossRef, OpenAlex, and Semantic Scholar.

Table 2. Structure of the basic dataset

Term & definition/scope	Term URI & URIs of RT, NT, BT
Xenogender people # A person who identifies their gender in relation to non-human understandings of gender and relates to animals, plants, or other things.	<https://homosaurus.org/v3/homoit0001671><https://homosaurus.org/v3/homoit0001670><https://homosaurus.org/v3/homoit0000571><https://homosaurus.org/v3/homoit0001048>

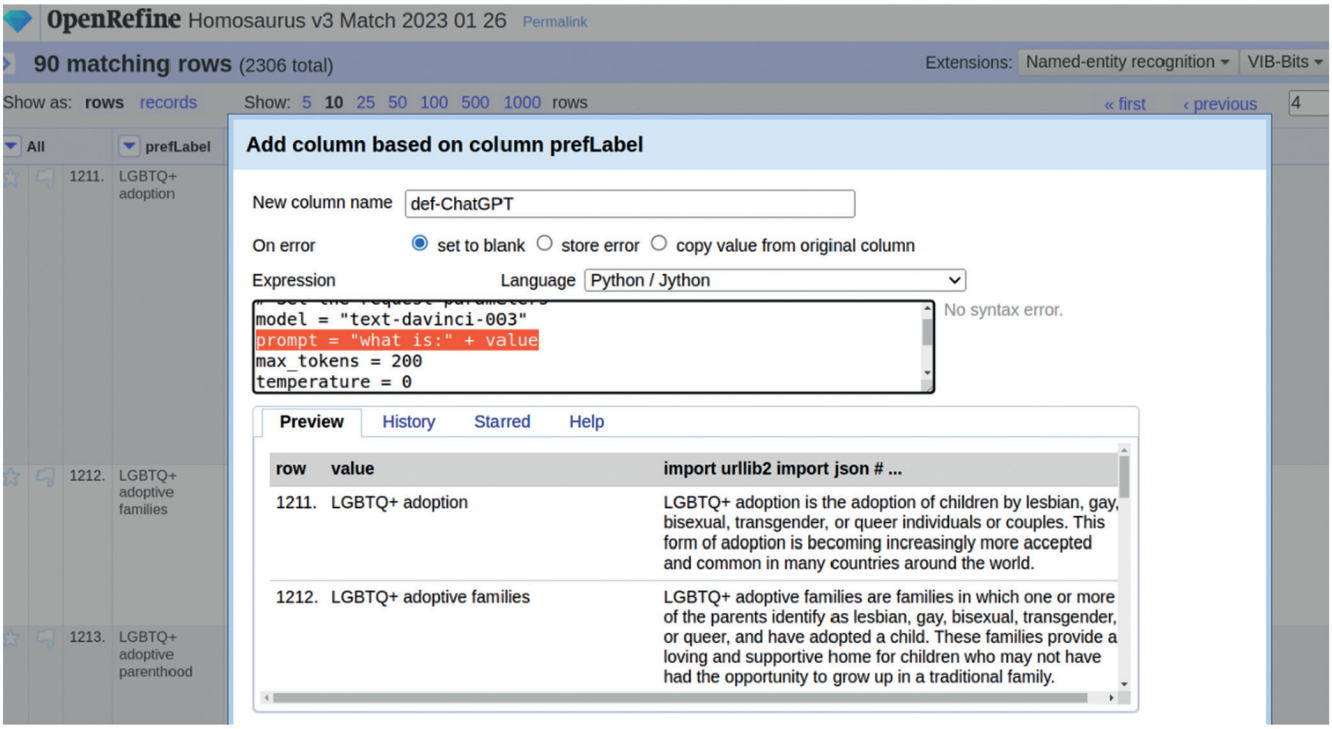

Figure 1. Automatic preparation of term definition from OpenAI (GPT-3 language model).

3.3.1 Basic Dataset

The deplyod Homosaurus version (version 3.3 released in December 2022) includes 2306 preferred terms (skos: prefLabel), with the majority of the preferred terms having a short definition or scope. Around 100 preferred terms do not presently have any definition or scope statement. This research study has developed a generic Python/Jython program in OpenRefine software that can negotiate with OpenAI to automatically collect definition or scope of a term (for those terms not having definitions or scope statements). Figure 1 illustrates how OpenAI can be utilised to fine tune the dataset and Table 2 shows the final structure of the basic dataset.

3.3.2 Bibliographic Dataset

The bibliographic dataset has been developed by using four ODbL-based data sources, as mentioned in Section 3.3. The content negotiation processes were executed for all four selected sources through REST/API-based data fetching in JSON format. It has been decided to select only 25 top search results (in all these sources, results are ranked by relevancy scores) against each preferred term or subject descriptor in the Homosaurus. The processes of the data fetching and data curation are illustrated in Table 3 with one example from the CoRE database.

The same processes are followed for other three databases (namely CrossRef, OpenAlex, Semantic Scholar) and finally the results (obtained after curation) are merged to form a set of 1,38,100 curated bibliographic records. A set of 551 records from this final dataset kept aside as a test dataset and the rest 1,37,549 records are meant for training the Annif framework (see Fig. 2).

3.4 Machine Learning Backends

The machine learning backends are responsible for performing the heavy lifting such as data preprocessing, feature engineering, model training, and prediction. The choice of a backend algorithm depends on the type of problem being solved, the available data, and the desired outcome. The Annif framework includes an array of backend algorithms in a default installation instance. These backends in Annif can be categorised into two basic groups - Lexical models and Associative models. In a bibliographic data environment, the use of standardised vocabularies and labeled datasets, such as subject indexes, class numbers, and metadata elements, provides a solid foundation for deploying supervised learning techniques.[15-19]

This research study adopted associative models of machine learning backends as included in the Annif framework. These associated models in Annif may be classified into two broad groups - Regular backends (like

Table 3. Data fetching and data curation

About CoRE: CoRE is one of the world's largest aggregator of open access research papers from repositories and journals and is managed by by the Open University and Jisc, UK						
API call syntax in OpenRefine: "https://core.ac.uk:443/api-v2/articles/search/" + value.escape('url') + "?page=1&pageSize=25&metadata=true&fulltext=false&citations=false&similar=false&duplicate=false&urls=false&faithfulMetadata=false&apiKey=<API-Key-Goes-Here>" **value**=term/descriptor in Homosaurus						
Queries sent	**No results**	**Total records**	**After curation**			
2, 306 terms	676 terms	27,678	Length of corpus (<350 words)	Language (Non-English)	Content (Non-relevant)	Final Records
			1,443 records	1,925 records	2,500 records	**21,810**

Figure 2. Final form of the training dataset for the Annif framework.

```
(annif-venv) roshni@roshni-HP-Pavilion-Laptop-14-dv0xxx:~/annif$ echo "Using Social Media to Advocate LGBT Rights in Black
 Africa :: The prevalence of draconian homophobic laws in Cameroon and Nigeria has systematically stultified sympathy for
the LGBT communities and made pro-gay street activism a risky venture in these two countries. In view of this, a good numb
er of gay rights activists have resorted to the social media as a suitable platform for a less risky advocacy. Using the s
ocial media has afforded them the opportunity to explore interactive, post-modern, and personified approaches to sensitizi
ng and mobilizing their readership in favour of gay proselytism in Cameroon, Nigeria, and some other parts of Africa. Base
d on a content analysis of 200 blog posts and web/facebook pages generated by Cameroonian and Nigerian gay activists, this
 chapter measures the extent to which gay activists adopt a national/local perspective versus the level to which they adop
t an international perspective in their online advocacy. The chapter equally examines the degree to which these citizen jo
urnalist/activists construct their advocacy discourse from the prism of a cultural war between the West and Africa." | ann
if suggest homosaurus-nn
To enable them in other operations, rebuild TensorFlow with the appropriate compiler flags.
<https://homosaurus.org/v3/homoit0000946>       LGBTQ+ social media     0.4629320502281189
<https://homosaurus.org/v3/homoit0001321>       Social media    0.3833159804344177
<https://homosaurus.org/v3/homoit0000823>       LGBTQ+ blogs    0.34826794266700745
<https://homosaurus.org/v3/homoit0000846>       LGBTQ+ dating applications      0.32926449179649353
<https://homosaurus.org/v3/homoit0001983>       LGBTQ+ vlogs    0.3106234669685364
<https://homosaurus.org/v3/homoit0000888>       LGBTQ+ Internet forums  0.29822877049446106
<https://homosaurus.org/v3/homoit0000832>       LGBTQ+ chatrooms        0.29165834188461304
<https://homosaurus.org/v3/homoit0000962>       LGBTQ+ websites 0.2627161741256714
<https://homosaurus.org/v3/homoit0000900>       LGBTQ+ meeting places   0.23306217789649963
<https://homosaurus.org/v3/homoit0000007>       LGBTQ+ activists        0.17895308136940002
(annif-venv) roshni@roshni-HP-Pavilion-Laptop-14-dv0xxx:~/annif$
```

Figure 3. Subject descriptor suggestions from Homosaurus in neural network backend.

TF-IDF, Omikuji Bonsai, Omikuji Parabel) and Fusion backends that combine results from many backends (e.g. Ensemble Simple, Ensemble PAV and Ensemble Neural Network). Each of these machine learning backend algorithms has its own strengths and weaknesses. This research study shows that both Omikuji as a regular backend and the neural network as a fusion (ensemble) backend performed more efficiently than the TF-IDF backend (see Section 4). What really distinguishes the neural network backend (Fig. 3) is its ability to support successive learning, enabling further training while in use through the learn command in the Annif framework.

4 RESULTS

Measuring the efficacy of machine learning backends is a crucial step in building efficient and accurate information retrieval systems e.g. an automated subject indexing system. One common approach to evaluating machine learning backends is by using retrieval metrics, which provide a quantitative measure of the system's performance. These metrics include precision, recall, F1-score, and Normalised Discounted Cumulative Gain (NDCG), among others. The Annif framework extends support to measure efficacy of a machine learning backend with the help of an array retrieval matrices of which F1@5 and NDCG are considered as the most important indicators. F1@5 (F1 at 5) measures the harmonic mean of precision and recall at a cut-off point of 5 results. It provides a way to evaluate the system's performance in terms of both precision and recall for the top 5 predicted subject descriptors. The F1@5 is an order-unaware metric. NDCG, on the other hand, is an order-aware retrieval metric that takes into account both the relevance of the retrieved documents and their position in the ranked list. F1@5 only considers the precision of the top 5 retrieved documents, whereas NDCG considers the precision of all retrieved documents in the ranked list. NDCG is also more sensitive to the position of relevant documents in the ranked list, whereas F1@5 treats all relevant documents equally regardless of their position in the list. The comparison of the efficacy of the selected machine learning backends in terms of retrieval metrics (based on 551 records kept aside as a test dataset) is given in Table 4.

An analysis of the comparative performances of the adopted machine learning backends for this research study (as tabulated in Table 4) leads to the following significant observations:

Table 4. Efficacy of different machine learning backends through retrieval metrics

Retrieval metrics	TF-IDF	Omikuji (Bonsai)	Omikuji (Parabel)	Ensemble (NN-cycle1)	Ensemble (NN-cycle2)
Precision (doc avg):	0.101634	0.155354	0.154628	0.155659	0.160126
Recall (doc avg):	0.237625	0.355162	0.352026	0.335414	0.334655
F1 score (doc avg):	0.130473	0.20041	0.199552	0.196346	0.199393
Precision (subjavg):	0.106567	0.107767	0.10664	0.115987	0.112513
Recall (subjavg):	0.127574	0.158171	0.156506	0.152875	0.151994
F1 score (subjavg):	0.100542	0.11406	0.113311	0.116588	0.114847
Precision (weighted subjavg):	0.231017	0.22924	0.229441	0.239383	0.236226
Recall (weighted subjavg):	0.200072	0.305824	0.304395	0.28939	0.290104
F1 score (weighted subjavg):	0.180898	0.23109	0.23205	0.229889	0.230164
Precision (microavg):	0.101634	0.155354	0.154628	0.153468	0.158656
Recall (microavg):	0.200072	0.305824	0.304395	0.28939	0.290104
F1 score (microavg):	0.134794	0.206042	0.205079	0.20057	0.205129
F1@5:	**0.139707**	**0.223012**	**0.219605**	**0.21098**	**0.211939**
NDCG:	**0.199271**	**0.33887**	**0.335462**	**0.322758**	**0.322949**
NDCG@5:	0.183383	0.318747	0.313858	0.302617	0.303853
NDCG@10:	0.202871	0.34295	0.339468	0.326712	0.326902
Precision@1:	0.176044	0.406534	0.401089	0.3902	0.39383
Precision@3:	0.162735	0.275258	0.265578	0.259831	0.263158
Precision@5:	0.139746	0.221053	0.216334	0.209166	0.210708
LRAP:	0.151515	0.271484	0.267398	0.257715	0.258895
True positives:	560	856	852	**810**	**812**
False positives:	4950	4654	4658	**4468**	**4306**
False negatives:	2239	1943	1947	**1989**	**1987**
Documents evaluated:	551	551	551	551	551

1. Significant retrieval metrics like F1@5, NDCG, Recall, and Precision are all evaluation metrics that have a range from 0 to 1. A value of 0 indicates that the metric has performed poorly, while a value of 1 indicates perfect performance.
2. Both Omikuji (two models namely Bonsai and Parabel) and Ensemble-NN machine learning backends outperformed the TF-IDF backend in terms of all significant metrics;
3. Omikuji Bonsai is ahead of Omikuji Parabel in almost all significant retrieval metrics;
4. The Neural network (Ensemble-NN) backend performed almost at par with the Omikuji Bonsai (trained with 1,37,549 records) after training with only 10 % of training dataset (cycle 1 - 13,755 records);
5. The Neural network (Ensemble-NN) backend at cycle2 i.e. after successive learning with another 10 % of training dataset (total 27,510 records) shows visible improvements in terms of all significant retrieval metrics (F1@5, NDCG, NDCG@5, NDCG@10 and so on);
6. The Neural network (Ensemble-NN) backend shows promises that with successive learning (by using learn command in Annif) it may achieve near 50 % scores for F1@5 and NDCG.

Automated subject indexing is a challenging task and is considered a hard problem. Research studies show that indexers agreed with only ⅓ of the descriptors when they indexed same set of documents by using the same vocabulary control device.[18, 20-21] Therefore, accomplishing a 50 % mark for important retrieval metrics like F1@5 and NDCG is considered a significant achievement in automated indexing. This research study shows that the machine learning backend algorithms (like Omikuji and Ensemble-NN) are identifying and classifying relevant topics or concepts from the input text with considerable accuracy. Nevertheless, the degree of significance of achieving a 50 % score also depends on the specific task, the dataset, and the context in which the automated indexing system is being used. While designing and developing a prototype, the focus is usually on creating a system that can perform a specific task with high accuracy on a limited dataset. However, deploying such a system for large-scale application requires careful consideration of various factors like data distribution, computational resources, and scalability. It also involves addressing issues like model explainability and data accessibility. The recommended method for programmatically obtaining descriptor suggestions against a text corpus is through REST/API call-based access. Annif presently supports the following REST/API endpoints (with the base URL being http://<IP or DNS>:5000/v1/): a) /projects (which returns a list of projects); b) /projects/{project_id} (which provides information for a specific project); and c) /projects/{project_id}/suggest (which suggests subject descriptors from the KOS for a given text).

6 CONCLUSIONS

Historically, AI and ML tools have been reserved for commercial enterprises or large-scale organisational initiatives. However, open source software solutions and open datasets have now made it possible for LIS professionals to experiment with these cutting-edge tools. The present research study serves as a preliminary account of such experimentation with an open source AI/ML tool named Annif, achieving almost 33 % accuracy in subject prediction (see Table 4) with an Ensemble neural network backend. This result of the reported study is important in view of a low-resource domain like LGBTQIA+, which has almost no human-indexed bibliographic datasets based on a domain-specific vocabulary standard like Homosaurus. This research also demonstrates how it is possible to use recently available large language models (LLMs) to gather the definition and scope of the undefined technical terms in a given vocabulary standard (see Figure 1), to extract keywords from a text corpus to identify important concepts therein, and to summarize a long corpus into a concise one. In conclusion, the convergence of data carpentry techniques, large language models, and open source AI/ML frameworks has the potential to fundamentally transform knowledge organisation in libraries of any type or size.

REFERENCES

1. Burdick, A.; Drucker, J.; Lunenfeld, P.; Presner, T. & Schnapp, J. Digital humanities. MIT Press, Cambridge, MA, 2016.
2. Jeffrey L. Meikle. Digital humanities, by Anne Burdick; Johanna Drucker; Peter Lunenfeld; Todd Presner & Jeffrey Schnapp, Design and Culture, 2014, **6**(3), 431-433. Doi: 10.2752/175470814X14105156869746.
3. Liu, A. The meaning of the digital humanities. PMLA, 2013, **128**(2), 409-423. Doi: 10.1632/pmla.2013.128.2.409.
4. Schnapp J. & Presner P. Digital Humanities Manifesto 2.0, 2009 Available at: https://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf (Accessed on 5 January 2023).
5. Berman, S. Prejudices & antipathies: A tract on the LC subject heads concerning people.ed. Jefferson, N.C. McFarland & Co., London (2nd edition, first published in 1971), 1993.
6. Drabinski, E. Queering the catalog: Queer theory and the politics of correction. *Library Quarterly*, 2013, **83**(2), 94-111. Doi: 10.1086/669547.
7. Gibson, K.; Ladd, M. & Presnell, J. Traversing the gap: subject specialists connecting humanities researchers and digital scholarship centers. In Digital humanities in the library: Challenges and opportunities for subject specialists, edited by A. Hartsell-Gundy, L. Braunstein & L. Golomb. American Library Association, US, 2015, 3-18.
8. Knowlton, S.A. Three decades since prejudices and antipathies: A study of changes in the library of congress subject headings. *Cataloging & Classification Quarterly*, 2005, **40**(2), 123-145.

Doi: 10.1300/J104v40n02_08.
9. Olson, H.A. The power to name: locating the limits of subject representation in libraries (soft cover reprint of the original 1st ed. 2002 edition). Springer, 2011.
10. Watson, B.M. There was sex but no sexuality: Critical cataloging and the classification of asexuality in LCSH. *Cataloging & Classification Quarterly*, 2020, **58**(6), 547-565.
11. Zwaaf, K. The Homosaurus - http://homosaurus.org. *Technical Services Quarterly*, 2020, **37**(2), 207-208.
12. Mukhopadhyay, P. & Mitra, R. Digital humanities and inclusive librarianship: designing a collaborative, multi-lingual, Skos-compliantl linked open vocabulary for LGBTQIA+. *Indian J. Information, Library & Society,* 2022, **35**(1-2), 16–33.
Doi: 10.5281/zenodo.6814869.
13. Mukhopadhyay, P.; Mitra, R. & Mukhopadhyay, M. Library carpentry: Towards a new professional dimension (part i – concepts and case studies). *SRELS J. Infor. Management*, 2021, **58**(2), 67–80.
Doi: 10.17821/srels/2021/v58i2/159969.
14. Suominen, O. Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly: The J. Association of European Research Libraries,* 2019, **29**(1), 1–25.
Doi: 10.18352/lq.10285.
15. Kasprzik, A. Putting research-based machine learning solutions for subject indexing into practice. *In* Proceedings of the Conference on Digital Curation Technologies (Qurator 2020) Berlin, Germany, January 20–21, 2020, ed. Adrian Paschke, et. al. CEUR Workshop Proceedings 2535, CEUR-WS.org 2020, Available at: https://ceur-ws.org/Vol-2535/paper_1.pdf (Accessed 15 February 2023).
16. Golub, K. Automated subject indexing: An overview. *Cataloging & Classification Quarterly*, 2021, **59**(8), 702–719.
Doi: 10.1080/01639374.2021.2012311.
17. Sfakakis, M.; Papachristopoulos, L.; Zoutsou, K.; Papatheodorou, C. & Tsakonas, G. Automated subject indexing using word embeddings and controlled vocabularies: A comparative study. *International Journal of Metadata, Semantics and Ontologies,* 2021, **15**(4), 233-243.
Doi: 10.1504/IJMSO. 2021. 125884.
18. Suominen, O.; Inkinen, J. & Lehtinen, M. Annif and Finto AI: Developing and implementing automated subject indexing. *JLIS. It*, 2022, **13**(1), 265–282.
Doi: 10.4403/jlis.it-12740.
19. Lehtinen, M. Developing and implementing automated subject indexing. *In* Bibliographic Control in the Digital Ecosystem, edited by G. Bergamin & M. Guerrini. Associazione italiana biblioteche, Italy, 2022.
20. Wu, M.; Brandhorst, H.; Marinescu, M-C.; Lopez, J.; Hlava, M. & Busch, J. Automated metadata annotation: What is and is not possible with machine learning. *Data Intelligence,* 2023, **5**(1), 122–138.
Doi: 10.1162/dint_a_00162.
21. Hahn, J. Semi-automated methods for Bibframe work entity description. *Cataloging & Classification Quarterly,* 2021, **59**(8), 853–867.
Doi: 10.1080/01639374.2021.2014011.

CONTRIBUTORS

Ms Roshni Mitra is a UGC-SRF scholar working in the Department of Library and Information Science at the University of Kalyani. Her research interests include knowledge modelling through vocabulary control and the development of multilingual, collaborative, and semantic web-based knowledge organisation systems in low-resource domains like LGBTQIA+.
Her role in this research study includes constructing the research problems, conducting a literature review, data wrangling and curation, and measuring the efficacy of the semi-automated subject indexing system as developed using the Annif framework.

Prof Parthasarathi Mukhopadhyay, Department of Library and Information Science, University of Kalyani, is a noted contributor in the research areas of application of open source and open standards in LIS, data carpentry, library discovery systems, and AI/ML-based applications.
His role in this research study includes configuring the software framework, constructing the methodologies, optimising results through an array of retrieval metrics, and finalising the manuscript.