Annals of Library and Information Studies
Vol. 71, December 2024, pp. 370-383
DOI: 10.56042/alis.v71i4.13392

Library Carpentry and the Bibliographic Data Universe: What Librarians Can Do with Open Data

Parthasarathi Mukhopadhyay[a]* and Mondrita Mukhopadhyay[b]

[a]Professor, Department of Library and Information Science, University of Kalyani, Kalyani, WB, India
E-mail: psm@klyuniv.ac.in

[b]University Research Scholar (Senior), Department of Library and Information Science, University of Kalyani, Kalyani, WB, India
E-mail:mondrita24c@gmail.com

Received: 09 August 2024; Accepted: 20 August 2024

This study emphasizes the significance of open data in libraries and the necessity for library professionals to acquire expertise in managing and analyzing data. It introduces the concept of library carpentry, a specialized form of data carpentry tailored for library professionals, focusing on skills like data curation, textual data management, and bibliographic content negotiation. Library carpentry aims to equip library staff with these skills to improve current services and introduce new, data-driven information services. The paper discusses case studies illustrating the potential of library carpentry techniques, particularly in understanding bibliographic datasets under the ODbL license, emphasizing how embracing library carpentry techniques like bibliographic content negotiation, regular expressions (regex), named entity recognition, machine translation and data reconciliation, among others can help libraries evolve in response to the changing information landscape and better serve their communities.

Keywords: Data carpentry, Data reconciliation, Data wrangling, GREL, Library carpentry, Machine translation, Named entity recognition, Open data, OpenRefine, REST/API

Key points:

This research study discusses deployment of data carpentry techniques and open data sources in libraries to:

- measure popularity of books and e-books
- assess open access friendliness of an institute
- initiate large-scale bibliometric and altmetric studies
- gendrize authorship of scholarly publications
- support machine translations in Indian languages

1 Introduction

Library and Information Science (LIS) professionals play a pivotal role as architects and custodians of diverse databases, encompassing bibliographic records, financial data, member information, and more. In this context, it is essential for LIS professionals to arm themselves with the necessary tools and expertise to navigate through extensive datasets, enabling them to make informed decisions and enhance the quality and currency of their services. Data is the new oil in the 21st century, and much like oil, the true value of data lies in its refinement, enhancement, and extraction. Therefore, it's necessary for library professionals to transition from being merely tech-savvy to becoming data-savvy professionals. This shift is essential to managing the complex information landscape in libraries, regardless of their size or type (Dennis, 2019; Wouter, 2019). The concept of 'library carpentry' stems from 'data carpentry,' which refers to the fundamental data skills necessary for various research activities, ranging from data gathering to dissemination. These skills are highly specific to different domains of research. Library carpentry, a subset of data carpentry, focuses on domain-specific data skills such as data curation, textual data management, bibliographic content negotiation, including citations, regular expressions (regex), named entity recognition (NER), and data reconciliation, among others, as outlined by researchers (Baker *et al.*, 2016; Burton and Lyon, 2017). The goal of library carpentry is to empower library personnel to apply these data skills, enhance existing services, and introduce new, data-intensive information services. This case study-based approach

*Corresponding author

paper explores the fundamental concepts, general methods, data wrangling tools, and associated techniques. It also presents concepts, tools, and steps through an array of case studies employing specific methodologies. These case studies provide readers with insights into the possibilities offered by library carpentry methods and their applications, especially in understanding bibliographic datasets available in the public domain, more specifically under the ODbL (Open Data Commons Open Database License). The ODbL is a legal framework designed for the licensing of open data. It allows individuals and organizations to freely use, modify, and distribute databases while providing proper attribution to the original creators. ODbL sets guidelines for sharing databases and ensures that they remain open and accessible to the public. It encourages collaboration and innovation by allowing users to build upon existing datasets, promoting the free exchange of information and knowledge. In essence, a case-study-based approach equips library professionals with the practical skills and problem-solving abilities necessary to handle the complexities of managing and analyzing data in modern libraries. It not only enhances their technical expertise but also empowers them to make informed decisions, ultimately improving the quality of library services and user experiences.

2 Related Literature

Data carpentry operates on the fundamental principle that 'good data triumphs over opinion,' aligning with the FAIR principles of data management—findability, accessibility, interoperability, and reuse (Mani *et al*., 2021; Virkus and Garoufallou, 2020; Wilkinson *et al*., 2016). Library carpentry, on the other hand, is designed to cultivate software and data proficiency within the professional community, equipping them to meet the challenges of 21st-century jobs in library and information science (LIS). This initiative focuses on empowering LIS professionals to adeptly employ software and data techniques, particularly within bibliographic datasets. It emphasizes mastering reproducible data and software practices. Library carpentry, as outlined on *librarycarpentry.org* and as reported by a research group (Mukhopadhyay *et al*., 2021), encompasses several key areas: a) applying data science concepts in library-related tasks; b) recognizing and implementing best practices in data structuring; c) earning the skills to programmatically transform and map data from one format to another; d) familiarity with data visualization tools and techniques; e) developing effective collaboration with researchers, ICT staff, and systems colleagues; and f) automating repetitive and error-prone tasks. By honing these skills, library professionals are better prepared to harness the power of data and software, enhancing their efficiency and effectiveness in the dynamic landscape of LIS (Atwood *et al*., 2019; Cope *et al*., 2020; Dennis *et al*., 2017).

A plethora of specific instances exist wherein library and information science (LIS) professionals have effectively implemented data carpentry principles and techniques. Examples of these implementations include the continual updating of authority datasets in response to the evolving landscape of the bibliographic universe, as discussed in works such as Vellucci (Vellucci, 2004) and Zhu & Seggern (Zhu and Seggern, 2005). Furthermore, scholars have explored the potential of 'context control' as an alternative or complementary approach to traditional authority control, as evidenced in Dryden's work (2008). Some other data-intensive activities include the application of data mining for author name extraction, followed by hierarchical clustering to generate name authority files automatically (Diaz-Valenzuela *et al*., 2010); the development of an open-source authority control tool based on Mosul technology (Manghi and Mikulicic, 2011); the establishment of a process for automatic name authority control in the context of electronic theses and dissertations (McCutcheon, 2011); the creation of a technical framework for automated name authority control in digital libraries (Diaz-Valenzuela *et al*., 2013); the development of a tool for authority control utilizing semantic web technologies (Leiva-Mederos *et al*., 2013); the scripting for the reconciliation of authority datasets from Linked Open Data sources (Harlow, 2015); the creation of Linked Open Data for geographic name authority files for specific countries (Ryan *et al*., 2015); the merging of authority data into Linked Open Data formatted datasets for nationwide library networks (Bensmann *et al*., 2017); the validation of local authority datasets through the utilization of OpenRefine (Carlson and Seely, 2017); the enrichment of authority datasets in libraries using the Wikidata platform (Allison-Cassin and Scott, 2018); measuring open access friendliness of a given set of institutions (Mukhopadhyay, 2022; Roy and Mukhopadhyay, 2022a, 2022b, 2022c);

analyzing patterns in retraction of journal articles (Mukhopadhyay et al., 2023) and the creation of geographic name authority files in MARC format for Indian place names by applying data carpentry method (Mukhopadhyay and Mukhopadhyay, 2022), among others.

The other important concepts related to data carpentry from which LIS professionals can benefit are name-to-gender inferences, sentiment analysis, machine translations, named-entity recognition (NER), and data reconciliation. Researchers reported applications of different REST/API-enabled name-to-gender inference services for determining the male-female ratio in authorship (Mukhopadhyay *et al*., 2021; Santamaría and Mihaljević, 2018). A translation platform facilitates the automatic conversion of specified phrases or terms from one language into numerous other languages. It incorporates a sophisticated amalgamation of tools for managing terminologies, human translation, and localization, catering to both ASCII and Unicode environments (Bowker, 2019). When employed in the context of multilingual bibliographic databases, especially in linguistically diverse countries like India, utilizing translation APIs through data wrangling processes (such as MARC or DCMES formatted records) opens up vast opportunities for libraries in those nation (Bowker and Buitrago Ciro, 2015). Machine translation initiatives have a long history rooted in computational linguistics. However, recent advancements have introduced API-based automatic translation processes, marking a new era in data carpentry for knowledge fusion (Knight and Koehn, 2003). These processes are instrumental in creating multilingual metadata, showcasing their significance in enhancing language-related tasks in libraries. The evolution from traditional machine translation methods to modern API-based approaches reflects the continuous efforts to improve translation quality and address the challenges posed by linguistic variations and structural differences between languages (Wani *et al*., 2017).

In a comprehensive study (Purkayastha, 2019) exploring the applications of translation APIs within the data wrangling tool OpenRefine, a range of REST/API-enabled machine translation platforms were identified. Notably, two research studies have documented the utilization of such machine translation platforms within OpenRefine. These studies demonstrate the effective generation of translated strings in 12 different Indian languages, thereby facilitating the development of multilingual bibliographic and authority datasets (Mukhopadhyay and Mitra, 2021; Mukhopadhyay and Mukhopadhyay, 2021).

Studies on Named Entity Recognition (NER) employing Conditional Random Fields (CRF) have been conducted since 1996, as evidenced by McCallum and Li's work in 2003 (McCallum and Li, 2003). Additionally, research has explored into the reconciliation process between local metadata and established controlled vocabularies in cultural heritage collections, as highlighted in publications by van Hooland et al. (2015) and Park and Kim (2014). These scholarly works underscore the potential utility of linked data in the field of library and information science (LIS), particularly in enhancing bibliographic and authority data and offering valuable insights for LIS professionals. The importance of tools such as OpenRefine in managing large datasets has been a subject of extensive exploration in various research papers and articles, as demonstrated in a series of research reports (Carlson and Seely, 2017; Hill, 2016; Mukhopadhyay and Mitra, 2021; Tillman, 2016; Verborgh and Wilde, 2013). These studies underscore the indispensable nature of OpenRefine for handling extensive datasets effectively. Furthermore, efforts have been made to establish a unified framework capable of handling both flat and nested NER tasks, as discussed by Li et al. in 2020 (Li et al., 2020). Additionally, NER techniques have been applied innovatively, including measuring gender bias. For instance, Mehrabi et al. (Mehrabi et al., 2020) conducted a study employing NER techniques on a dataset spanning 139 years of US Census Reports on baby names, revealing a disparity in the recognition of female names as person entities. These scholarly investigations contribute significantly to advancing the understanding and application of NER methodologies in various contexts, providing valuable knowledge for researchers and professionals in the field.

In recent years, the field of LIS has witnessed a surge in research concerning data reconciliation, a concept intricately linked with Named Entity Recognition (NER) but significantly broadening its scope to encompass the retrieval of suggested access points from linked open authority datasets within the LIS domain. A notable study conducted by Delpeuch (Delpeuch, 2019) shed light on the extensive

availability of linked open data (LOD) datasets, emphasizing their seamless integration with data wrangling tools like OpenRefine. The study meticulously compiled a comprehensive list of API-based data services, which were instrumental in fetching suggested access points and streamlining the organization of information resources. Furthermore, the years 2016 and 2018 witnessed pivotal international surveys conducted by Smith-Yoshimura, exploring the landscape of LOD-based data sources in the domain of LIS (Smith-Yoshimura, 2016, 2018). These surveys aimed to identify readily deployable services for data reconciliation, particularly within the domain of bibliographic data. Encouragingly, the research community has increasingly recognized the potential of Wikidata as a robust source for data reconciliation. Several studies have underscored the efficacy of integrating Wikidata with bibliographic datasets (Allison-Cassin and Scott, 2018; Bianchini and Bargioni, 2021; Lemus-Rojas and Pintscher, 2017; Ruttenberg, 2019). These efforts have not only advocated for its application but have also explored the mechanisms of multilingual data reconciliation, highlighting the platform's versatility. In addition to exploring generic datasets such as Wikidata and DBpedia, researchers have embarked on innovative endeavors within specific contexts. Brando et al. (2016) explored the linkage of named entities within digital library setups, while Downey (2019) focused on name authority control within institutional repositories. Furthermore, research studies explored the application of LOD datasets for creating geographic name authority files, emphasizing the vital role of data reconciliation in geographic contexts (Mukhopadhyay and Mukhopadhyay, 2022; Ryan et al., 2015). Moreover, Parker and Gray (Parker and Gray, 2019) made significant strides in the development of a digital retrieval system, employing local authority datasets effectively through the reconciliation process, thereby enhancing the efficiency of information retrieval methodologies. These studies collectively contribute to the growing body of knowledge, enriching our understanding of data reconciliation practices in the dynamic landscape of LIS.

3 Objectives

The primary objective of this case study-based research work is to communicate to LIS professionals the possibilities of data carpentry tools, techniques, and the use of open datasets in introducing new information services and improving existing library services. The six specific objectives are – a) to rank books/e-books by their popularity to help effective book selection process (case study 1); b) to assess open access (OA) friendliness of a given set of institutes (case study 2); c) to develop a framework for data intensive bibliometric studies by utilizing ODbL-based citation and altmetric data sources (case study 3); d) to apply name-to-gender inference services in genderizing authorship of research papers in different disciplines (case study 4); e) to apply named entity recognition (NER) and data reconciliation services to improve the quality of name authority and subject authority datasets in libraries (case study 5); and f) to deploy machine translation platforms for automatic translation of English text to Indian languages (case study 6).

4 Generic Methodology

The generic methodology of data wrangling through an open-source tool, namely OpenRefine (openrefine.org), requires sequential steps for various data wrangling activities. These essential steps, generically applicable across diverse contexts, provide a systematic approach to the complex process of data preparation and manipulation. First and foremost, we need to establish the purpose and scope of the data wrangling endeavour, setting the foundation for subsequent actions. The second phase involves the development of primary datasets and the identification of key parameters vital for effective data wrangling (e.g. DOI for an article or ISBN for a book or ORCID ID for a researcher). This meticulous process ensures that the datasets are structured and organized in a manner conducive to the wrangling process. Identifying the primary key, in particular, plays a pivotal role, as it serves as a fundamental identifier, facilitating seamless data management. Subsequently, the selection of appropriate data-wrangling sources becomes important. Careful consideration needs to be given to source selection, aligning data sources with the research objectives, and the type of data to be processed. This strategic selection enhances the reliability and relevance of the data being wrangled. The fourth step entails the design of queries through the General Refine Expression Language (GREL), a vital component of the data wrangling process in OpenRefine. GREL empowers researchers to formulate complex API-based queries, enabling precise data extraction and transformation. It acts as the bridge between raw data

and its refined, usable form. The final phase involves a meticulous analysis of the JSON responses, extracting pertinent information while adhering to the established queries. The iterative application of GREL ensures the extraction of relevant data, paving the way for further analysis and interpretation. In essence, this generic methodology of data wrangling encapsulates a systematic series of steps, from defining research objectives to extracting refined data, providing researchers with a structured approach to navigate the complexities of data wrangling. By adhering to these methodological principles, scholars can enhance the efficiency and accuracy of their datasets, laying a robust foundation for insightful research outcomes.

5 Case Studies

In the subsections, readers will find a comprehensive exploration of six distinct case studies carefully curated to cover six prominent areas in data-intensive library services, keeping in view the practical utilities of these kinds of services in libraries of any type or size. These case studies serve as a gateway, showcasing the myriad ways in which the profound theories and advanced techniques of data science can be harnessed to unlock uncharted areas within the field of data librarianship. The first case study was meticulously dissected and laid out in minute detail. However, owing to space constraints, the exposition of the remaining five case studies is streamlined to encompass only the essential facets. Despite this brevity, every effort has been made to distill the essence of these studies, ensuring that the essence of their methodologies remains intact. Moreover, it is important to note that each case study is supported by examples to help readers acquire the necessary skills.

5.1 Case Study 1: Measuring popularity of books/e-books

The purpose of this case study is to enable LIS professionals to gauge the popularity of a set of books or e-books that have an ISBN. This assessment may be conducted by analyzing global user ratings, engaging with discussions within socio-academic networks, and evaluating the book's altmetrics score. These evaluations are facilitated through the application of data wrangling techniques as described in the steps given below:

A: Obtain MARC records for books/e-books from publisher/catalogue datasets

In the majority of cases, publishers provide MARC datasets for e-books intended for procurement without any additional charges. These datasets encompass both content and metadata, constituting a comprehensive e-book package. Several publishers have established platforms where e-book metadata can be easily downloaded. For instance, Springer-Nature allows users to freely access MARC data for e-books through their metadata downloader portal, which is accessible at http://metadata.springernature.com/metadata/books. Alternatively, one can export MARC records from his or her own library or from other libraries. The popularity of books with ISBNs can then be measured on the basis of indicators available from global services like Goodreads.com and Altmetrics.com through the data wrangling process as discussed in the following paragraphs.

B: Convert MARC records in OpenRefine format

The chosen tool for data wrangling, OpenRefine, doesn't directly handle MARC data. So, the initial step involves converting MARC records into a format compatible with OpenRefine. MARCEdit, a trusted tool among librarians, proves invaluable here. The latest versions (starting from 6.x) of MARCEdit are OpenRefine compatible, allowing MARC records to be converted into OpenRefine-supported formats such as TSV and JSON. Converting MARC records to TSV format in MARCEdit offers additional benefits, simplifying the management of indicator positions. The process in MARCEdit is straightforward: access MARC records, export them in the desired format (TSV), and seamlessly import them into OpenRefine for further processing.

C: Create a project in OpenRefine

The ‘create project’ option in OpenRefine can handle converted MARC records efficiently, including leader and indicator position values. Each record is separated from the other through a blank row. We need an ISBN for each book (tag 020) without hyphenation (raw ISBN), and it is quite easy to add a column in OpenRefine by using a suitable GREL *value.split("$9")[0]. split("$a")[0]* (see Fig. 1). Now, we can start content negotiation with selected services (Goodreads and Altmetrics) for fetching popularity indicators in JSON format.

D: Data wrangling from Goodreads and Altmetrics

This plain ISBN can now be used as an input value to develop the API call syntax in the prescribed format. The GREL expression to create an API call, the response from Goodreads in JSON format, and data point extraction (here average_rating) are given

here (Table 1, Row 1) for a better understanding of the automatic data fetching process. The same steps and method can be applied to obtain an altmetrics score for the books from Altmetrics.com by only changing the REST/API call syntax and data extraction format (Table 1, Row 2).

The scope for application of the method and tools as explained is quite obvious as an effective means in the book selection process, particularly in the pick-and-choose model for procuring e-books.

5.2 Case Study 2: Assessing open access friendliness

It is quite possible to develop a data-intensive research framework to measure open access (OA) support in a given set of institutes. A data-carpentry-based research study has proposed a distributed weight

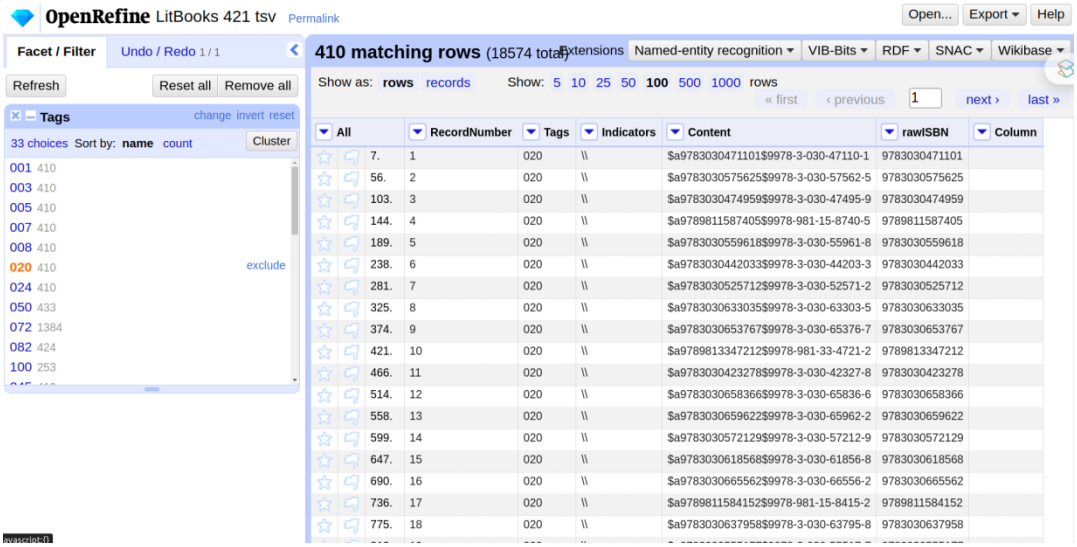


Fig. 1 — Extracting ISBN from MARC records in Open Refine

Table 1 — API based data fetching and data extraction in Openrefine from Goodreads,com and Altmetrics,com

GREL expression	Resultant API call	Response in JSON format	Data extraction
"https://www.goodreads.com/book/review_counts.json?isbns=" + **value** Note: value is the plain ISBN of the books	https://www.goodreads.com/book/review_counts.json?isbns=9783030471101	{ "books": [{ "id": 73954637, "isbn": "3030471101", "isbn13": "9783030471101", "ratings_count": 0, "reviews_count": 0, "text_reviews_count": 0, "work_ratings_count": 1, "work_reviews_count": 3, "work_text_reviews_count": 0, "**average_rating": "4.00**" }] }	GREL value.parseJson().books[0].average_rating Result 4.00 (average_rating in Goodreads varies from 1 to 5 point scale)
"https://api.altmetric.com/v1/isbn/" + **value** Note: value is the plain ISBN of the books	https://api.altmetric.com/v1/isbn/9783030471101	{"title":"Beckett and Politics","isbns":["9783030471095","9783030471101"],"cohorts":{..."altmetric_id":93419258,"schema":"1.5.4","is_oa":false,"cited_by_**posts**_count":9,"cited_by_**tweeters**_count":7,"cited_by_msm_count":2,"cited_by_accounts_count":9,"last_updated":1614112764,"**score":23.7**,"history":{.... "readers":{"citeulike":0,"**mendeley**":5,"connotea":0},"readers_count":5,"images":{..... }	GREL value.parseJson().score Result 23.7

age-based ranking framework to measure the OA friendliness of a set of given Indian institutions (publications of 16 top IITs as sample datasets in the year range from 2010 to 2019) by taking into consideration four primary areas, namely: OA publications share, OA licenses share, OA citations share, and OA altmetric scores share (Mukhopadhyay, 2022). This data collection method was later adopted by researchers to measure OA friendliness for other groups of institutions like NITs, central universities, and state universities in India. The methodology may be represented as:

A: Development of a primary dataset

The primary dataset of publications (DOI or PMID of a journal paper or book chapter as a data element) for a given institute or a set of given institutions may be collected from different sources like Scopus (commercial), Lens (OA), OpenAlex (OA) or CrossRef (OA). OpenAlex is the largest bibliographic data repository with 250+ million records as of May 31, 2024, available under ODbL. CrossRef and OpenAlex allow REST/API-based query syntax by ROR ID of an institution (the Research Organization Registry (ROR) is an internationally recognized registry led by the community, providing open persistent identifiers for research organizations worldwide). The responses from these two data sources are available in JSON format and can readily be converted to a CSV file by using any online JSON to CSV converter (e.g., konklone.io/json/). The example given in Table 2 shows how we can collect the DOI and title of all the publications of the Council of Scientific and Industrial Research, India (ROR ID: https://ror.org/021wm7p51) in the year range from January 1, 2019 to December 31, 2023 (5 years) from OpenAlex and by state name (Kerala) -based call from CrossRef. All the IDs and data elements can be changed suitably for obtaining data for other institutes from these two large and open data repositories.

B: Data wrangling for open access status

Currently, Unpaywall stands as the largest bibliographic repository for open content, boasting over 50,129,992 records as of May 31, 2024, and continuously growing. Unpaywall collects open content from over 50,000+ publishers and provides free access to its dataset through REST/API calls, specifically through version 2 with the DOI endpoint

Table 2 — REST/API call based data gathering from OdbL based bibliographic data sources

Data source	REST/API call	Response in JSON format
OpenAlex	https://api.openalex.org/works?filter=authorships.institutions.ror:https://ror.org/021wm7p51,from_publication_date:2019-01-01,to_publication_date:2023-12-31&select=doi,title&page=1&per-page=200	{ "meta": { "count": **4044**, "db_response_time_ms": 79, "page": 1, "per_page": 200, "groups_count": null }, "results": [{ "doi": "https://doi.org/10.1039/c9cs00648f", "title": "Multifunctional sonosensitizers in sonodynamic cancer therapy" }, ...
CrossRef	https://api.crossref.org/works?query.affiliation=Kerala&filter=from-pub-date:2022-01-01,until-pub-date:2022-12-31,type:journal-article&select=DOI,title&rows=1000&offset=0	{ "status": "ok", "message-type": "work-list", "message-version": "1.0.0", "message": { "facets": {}, "total-results": **2060**, "items": [{ "DOI": "10.1117/1.oe.61.4.044106", "title": ["Fractal and inertia moment analyses for thin-film quality monitoring"], ...

(GET /v2/:doi). Users benefit from a generous call limit of 100,000 calls per day. The API call structure and corresponding valid responses in OpenRefine are detailed in Table-3 for all 2,060 records published by institutions with 'Kerala' in affiliation for the year 2022 obtained from the CrossRef (Table 2, Row 2).

This methodology shows how easy it is through data carpentry to determine the access status for publications of a given institute, institutes in a given state, or even for the entire country.

5.3 Case Study 3: Citation intensive research studies

In the domain of LIS research, particularly within doctoral dissertations, bibliometric studies stand as a predominant domain. Regrettably, a significant limitation observed in these studies lies in the restricted scope of their sample datasets, often encompassing only a few thousand records. Consequently, such studies frequently fail to discern the genuine inherent patterns within the data. This limitation can be attributed to the prevalent use of manually managed citation datasets, which inherently restrict the scale of data collection. A potential remedy to this issue could be the adoption of data-carpentry-based approaches, which have the capacity to alleviate this data inadequacy challenge in bibliometric research. The following four ODbL-based data sources can be utilized to gather citations (Scite, Dimensions, and OCC) and altmetrics (altmetric.com) against the DOI of a knowledge object. All of these four REST/API services provide responses in JSON format, which can thereby be easily extracted, as explained in Table 4.

The utilization of the data carpentry method in conjunction with accessible open-citation data sources holds significant promise for transforming contemporary bibliometric and altmetric studies characterized by limited data availability into robust informetric

Table 3 — REST/API call based data wrangling from Unpaywall

API call structure for Unpaywall	No. of queries sent	Responses received
"https://api.unpaywall.org/v2/" + **valu**e + "?email=<your-mail-id-goes-here>"	2,060	2,060

value is DOI	publications with DOI	100% for publications with DOI
Response in JSON (truncated - **.....**)	GREL syntax to extract important data points	Extracted value
{"doi": "10.1080/07391102.2022.2126889", **...** , "journal_name": "Journal of Biomolecular Structure and Dynamics", "journal_issns": "0739-1102,1538-0254", "journal_issn_l": "0739-1102", "**journal_is_oa**": false, "**journal_is_in_doaj**": false, "publisher": "Informa UK Limited", "**is_oa**": true, "**oa_status**": "green", "has_repository_copy": true, "**best_oa_location**": {.... "oa repository (via OAI-PMH title and first author match)", "**license**": "cc-by", "version": "submittedVersion", "**host_type**": "repository", "is_best": true, "pmh_id": "oai:figshare.com:article/21201945", **.....** **.....** }]}]}	value.parseJson().journal_is_oa	false
	value.parseJson(). journal_is_in_doaj	false
	value.parseJson().is_oa	true
	value.parseJson().oa_status	green
	value.parseJson().best_oa_location.license	cc-by
	value.parseJson().best_oa_location.host_type	repository
		Result summary Closed: 1167 Open: 893 Gold: 580 Green: 113 Bronze: 99 Hybrid: 101

Table 4 — Sources for data wrangling for citations/altmetrics study

SL	Data source	REST/API syntax in OpenRefine	Purpose	Scope of data elements	No. of responses (2060 records as reported in Table-2)
1	Scite	"https://api.scite.ai/tallies/" + <value>	Citation data	1. Total citations 2. Supporting citations 3. Contradicting citations 4. Mentioning citations	2057 (against 2060 queries)
2	Dimensions	"https://metrics-api.dimensions.ai/doi/" + <value>	Citation data	1. Total citations 2. Recent citations	2059 (against 2060 queries)
3	Open Citation Corpus (OCC)	"https://opencitations.net/index/api/v1/citation-count/" + <value>	Citation data	1. Total citations	401 (against 2060 queries)
4	Altmetric	"https://api.altmetric.com/v1/doi/" + <value>	Altmetric data	1. Altmetric attention score 2. Subjects of the items	592 (against 2060 queries)

Note: For SL no. 1 to 4 <value> is DOI.

investigations characterized by extensive datasets. This synergy presents an opportunity to achieve a comprehensive understanding of scholarly publications and citation networks, thereby facilitating a broader and more nuanced analysis within the academic domain.

5.4 Case Study 4: Name-to-gender inference

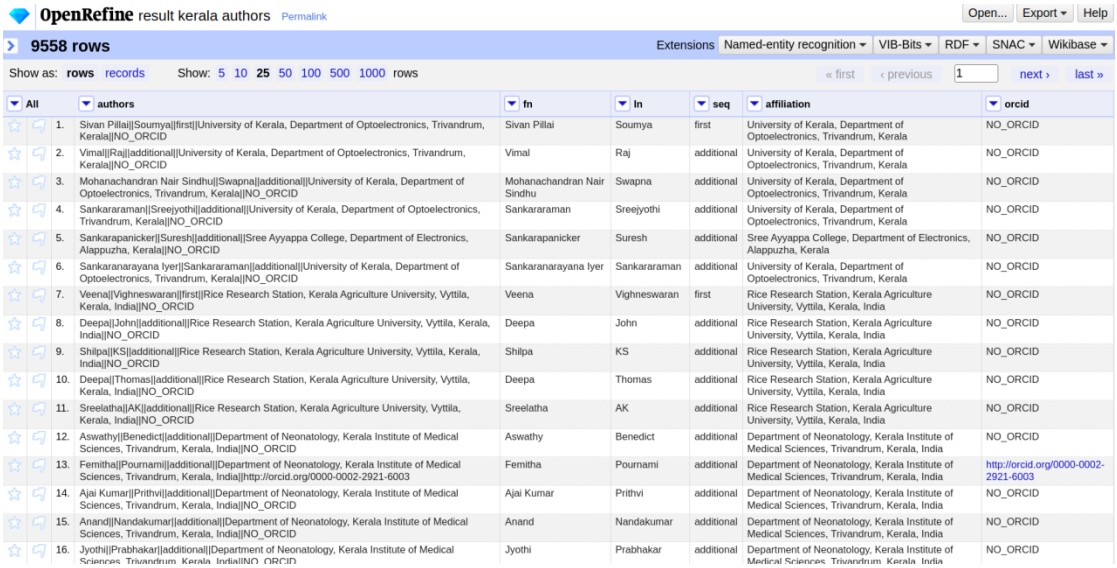
This case study investigates the practical applications of data carpentry techniques in determining the gender of authors. The study employs name-to-gender inference services as data sources, with the primary goal of inferring the gender of authors from the dataset in case study 2 (Table 2, Row 2). To achieve this, the study utilizes name-to-gender inference services in combination with data wrangling techniques in OpenRefine. Several prominent name-to-gender inference services are available for API-based data wrangling, including Gender API (https://gender-api.com/), genderize.io (https://genderize.io/), Gender-guesser (https://pypi.python.org/pypi/gender-guesser/), NameAPI (https://www.nameapi.org/), and NamSor (http://www.namsor.com/). For this study, genderize.io was chosen for its specific advantages: it provides free API calls at a rate of 1000 per day, offers customization options to focus on specific countries like India, and provides a probability score indicating the accuracy of gender identification. The study utilized a dataset of 2060 papers obtained from CrossRef (as shown in Table 2, Row 2) as the foundation. The Open Access (OA) status for these papers was obtained from Unpaywall (refer to Table 3), which also supplied author data for all these papers. The author array for each paper was extracted from Unpaywall responses using suitable General Refine Expression Language (GREL) techniques. This information was then used to create a separate project in OpenRefine for name-to-gender inference. In cases where multiple authors contributed to a paper, their names were joined by the '##' symbol. Each element of the author data, including name, author sequence, affiliation, and ORCID ID, was isolated using the '||' symbol in the new project (as depicted in Fig. 2).

The dataset encompassed a total of 9558 authors involved in these 2060 papers. Among these, 1909 given names for the first authors (2060 first authors) were available clearly, after excluding ambiguous author names like J.S. Reddy or R.V. Chinchilu, where given names were not clearly identifiable. The outcomes of name-to-gender inference are presented in Table-5 at various probability points for reference and analysis.

The application of name-to-gender inference techniques serves as a valuable tool in assessing the gender distribution within scholarly publications. By leveraging computational algorithms to infer gender from author names, researchers can gain insights into gender representation trends within academic literature, thereby contributing to the understanding of gender dynamics and disparities within various fields of study.

5.5 Case Study 5: NER and Data reconciliation

The central focus of this case study encompasses two key areas. The first, Named Entity Recognition (NER), showcases the automated extraction of concepts or entities from unstructured text, such as abstracts and notes. This process is highly valuable for indexing activities as it aids in the identification of pertinent information. The second area explores data



OpenRefine result kerala authors Permalink — Open... Export Help

9558 rows — Extensions: Named-entity recognition | VIB-Bits | RDF | SNAC | Wikibase

Show as: rows records — Show: 5 10 25 50 100 500 1000 rows — « first ‹ previous 1 next › last »

All	authors	fn	ln	seq	affiliation	orcid
1.	Sivan Pillai||Soumya||first||University of Kerala, Department of Optoelectronics, Trivandrum, Kerala||NO_ORCID	Sivan Pillai	Soumya	first	University of Kerala, Department of Optoelectronics, Trivandrum, Kerala	NO_ORCID
2.	Vimal||Raj||additional||University of Kerala, Department of Optoelectronics, Trivandrum, Kerala||NO_ORCID	Vimal	Raj	additional	University of Kerala, Department of Optoelectronics, Trivandrum, Kerala	NO_ORCID
3.	Mohanachandran Nair Sindhu||Swapna||additional||University of Kerala, Department of Optoelectronics, Trivandrum, Kerala||NO_ORCID	Mohanachandran Nair Sindhu	Swapna	additional	University of Kerala, Department of Optoelectronics, Trivandrum, Kerala	NO_ORCID
4.	Sankararaman||Sreejyothi||additional||University of Kerala, Department of Optoelectronics, Trivandrum, Kerala||NO_ORCID	Sankararaman	Sreejyothi	additional	University of Kerala, Department of Optoelectronics, Trivandrum, Kerala	NO_ORCID
5.	Sankarapanicker||Suresh||additional||Sree Ayyappa College, Department of Electronics, Alappuzha, Kerala||NO_ORCID	Sankarapanicker	Suresh	additional	Sree Ayyappa College, Department of Electronics, Alappuzha, Kerala	NO_ORCID
6.	Sankaranarayana Iyer||Sankararaman||additional||University of Kerala, Department of Optoelectronics, Trivandrum, Kerala||NO_ORCID	Sankaranarayana Iyer	Sankararaman	additional	University of Kerala, Department of Optoelectronics, Trivandrum, Kerala	NO_ORCID
7.	Veena||Vighneswaran||first||Rice Research Station, Kerala Agriculture University, Vyttila, Kerala, India||NO_ORCID	Veena	Vighneswaran	first	Rice Research Station, Kerala Agriculture University, Vyttila, Kerala, India	NO_ORCID
8.	Deepa||John||additional||Rice Research Station, Kerala Agriculture University, Vyttila, Kerala, India||NO_ORCID	Deepa	John	additional	Rice Research Station, Kerala Agriculture University, Vyttila, Kerala, India	NO_ORCID
9.	Shilpa||KS||additional||Rice Research Station, Kerala Agriculture University, Vyttila, Kerala, India||NO_ORCID	Shilpa	KS	additional	Rice Research Station, Kerala Agriculture University, Vyttila, Kerala, India	NO_ORCID
10.	Deepa||Thomas||additional||Rice Research Station, Kerala Agriculture University, Vyttila, Kerala, India||NO_ORCID	Deepa	Thomas	additional	Rice Research Station, Kerala Agriculture University, Vyttila, Kerala, India	NO_ORCID
11.	Sreelatha||AK||additional||Rice Research Station, Kerala Agriculture University, Vyttila, Kerala, India||NO_ORCID	Sreelatha	AK	additional	Rice Research Station, Kerala Agriculture University, Vyttila, Kerala, India	NO_ORCID
12.	Aswathy||Benedict||additional||Department of Neonatology, Kerala Institute of Medical Sciences, Trivandrum, Kerala, India||NO_ORCID	Aswathy	Benedict	additional	Department of Neonatology, Kerala Institute of Medical Sciences, Trivandrum, Kerala, India	NO_ORCID
13.	Femitha||Pournami||additional||Department of Neonatology, Kerala Institute of Medical Sciences, Trivandrum, Kerala, India||http://orcid.org/0000-0002-2921-6003	Femitha	Pournami	additional	Department of Neonatology, Kerala Institute of Medical Sciences, Trivandrum, Kerala, India	http://orcid.org/0000-0002-2921-6003
14.	Ajai Kumar||Prithvi||additional||Department of Neonatology, Kerala Institute of Medical Sciences, Trivandrum, Kerala, India||NO_ORCID	Ajai Kumar	Prithvi	additional	Department of Neonatology, Kerala Institute of Medical Sciences, Trivandrum, Kerala, India	NO_ORCID
15.	Anand||Nandakumar||additional||Department of Neonatology, Kerala Institute of Medical Sciences, Trivandrum, Kerala, India||NO_ORCID	Anand	Nandakumar	additional	Department of Neonatology, Kerala Institute of Medical Sciences, Trivandrum, Kerala, India	NO_ORCID
16.	Jyothi||Prabhakar||additional||Department of Neonatology, Kerala Institute of Medical Sciences, Trivandrum, Kerala, India||NO_ORCID	Jyothi	Prabhakar	additional	Department of Neonatology, Kerala Institute of Medical Sciences, Trivandrum, Kerala, India	NO_ORCID

Fig. 2 — Final dataset for gender inference (9558 authors in 2060 papers; 8192 given names are extracted)

reconciliation, which involves the automatic matching of concepts or entities, whether extracted or pre-existing, with domain-specific subject authority or name authority datasets. These datasets are accessible as linked open data and provide the flexibility to adapt the suggested values, enhancing the accuracy of the matching process. The concept of Named Entity Recognition (NER) has evolved progressively alongside the technological advancements in Natural Language Processing (NLP) tools and techniques. Notably, significant developments have occurred in the domain of NER over the past five years, as well as in the areas of mapping unstructured raw text to structured databases and text mining. In contrast, data reconciliation, although related to NER, expands its scope by retrieving suggested access points from linked open authority datasets within the Library and Information Science (LIS) domain. A recent study has highlighted the availability of numerous Linked Open Data (LOD) datasets that can be readily employed in data wrangling tools like OpenRefine. This study also enumerates various API-based data services capable of fetching suggested access points to facilitate the organization of information resources effectively (Delpeuch, 2019). Let us take up the MARC dataset of case study 1(tag 520 – summary note in particular) to extract key concepts in each of these books automatically by using NER services. The available tools for named entity recognition in OpenRefine encompass DBpedia Spotlight (accessible at https://www.dbpedia-spotlight.org/ without authentic-cation), Dandelion Entity Extraction (accessible at https://dandelion.eu/docs/api/datatxt/nex/v1/ with an authentication key requirement), and Stanford NER, which is built on Stanford University's NLP toolkit (StanfordNLP – accessible at https://stanfordnlp.github.io/CoreNLP/download.html, requiring local installation on machines pre-loaded with Java 8). Fig. 3 shows results of NER based concept(s) extraction by

Table 5 — Data wrangling for name-to-gender inference

Conditions	Result (N=1909 – first authors only)			
	Female	Male	Unsure	Female:Male
Extremely sure: if(probability == **1**, cells.genderData.gender, "Unsure")	334	1412	3640	1:4.22
Fairly sure: if(probability >= **0.90**, cells.genderData.gender, "Unsure")	418	2658	2310	1:6.35
Moderately sure: if(probability >= **0.80**, cells.genderData.gender, "Unsure")	777	3169	1440	1:4.07
Somehow sure: if(probability >= **0.60**, cells.genderData.gender, "Unsure")	872	3330	1184	1:3.81
API call: "https://api.genderize.io?name=" + value + "&country_id=IN" (value is here given name /first name of an author)				
Response: (JSON)	{"count":11152,"name":"Veena","country_id":"IN","gender":"female","probability":1.0} {"count":72,"name":"Ajai","country_id":"IN","gender":"male","probability":0.97}			

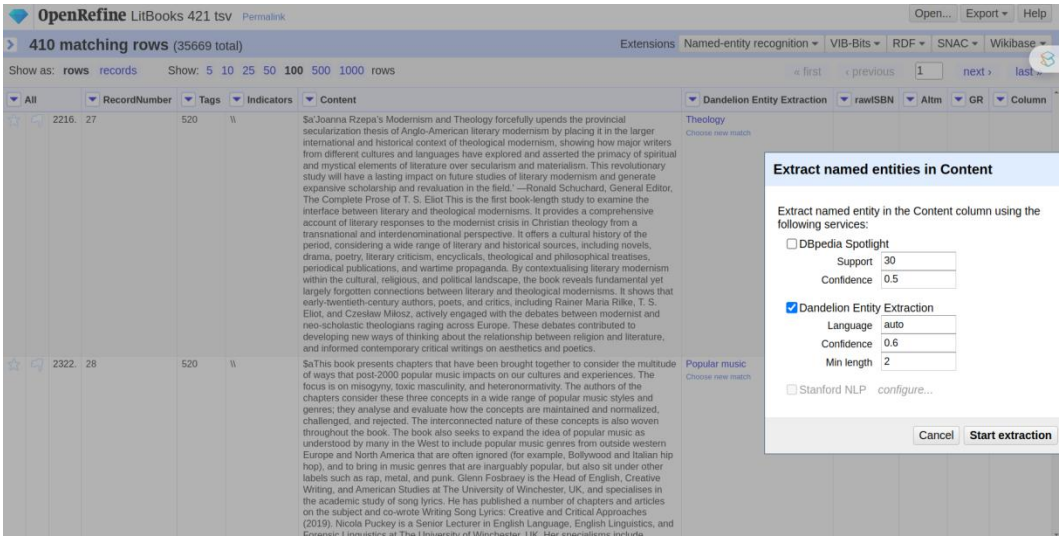


Fig. 3 — Dandelion based NER in OpenRefine

using Dandelion NER services by using an extension of Open Refine (https://github.com/stkenny/Refine-NER-Extension). It takes text content, throws it to Dandelion and produce results of concept analysis and extraction in a new column in the project.

Data reconciliation, on the other hand, is a process involving the validation of a data item's accuracy against a reliable data source, rooted in the concept of data validation. Within the framework of a library environment, data reconciliation can be perceived as a data wrangling activity. In this process, a textual representation of a given entity (such as a person, place, or subject) is matched with a standardized data source, either local or remote, such as subject headings lists, name authority files, or thesauri. This process is essentially semi-automatic. In cases where there is no direct match, the system provides suggestive name headings. It then requires human judgment to either reject the suggested name option or select an appropriate name from a list of suggestions, which is essentially a ranked array of potential entities. OpenRefine, functioning as a robust data wrangling tool, facilitates the reconciliation of names against various name registries within the Library and Information Science (LIS) domain. OpenRefine achieves this reconciliation by employing either API-based reconciliation services or by allowing the download of RDF datasets for local reconciliation. Additionally, it can interact with SPARQL endpoints. For instance, when presented with a column containing author names, OpenRefine can reconcile these names against the Virtual International Authority File (VIAF) name authority (Fig. 4). Similarly, when dealing with a column containing subject descriptors, the tool can reconcile these names against various knowledge organization tools like Library of Congress Subject Headings (LCSH), FAST, MeSH, or UNESCO thesaurus (depending on the user's selection made during the process).

5.6 Case Study 6: Automatic Translation

A translation platform facilitates the automatic conversion of phrases or terms from one language to multiple languages. It incorporates a range of tools for managing terminology, human translation, and localization in both ASCII and Unicode environments. These platforms typically offer translation APIs that enable automatic machine translation for various types of text inputs such as web pages, paragraphs, phrases, notes, and terminology. The key features of a translation API include: a) supporting automatic text translation between selected language pairs; b) automatically detecting input languages; c) combining machine translation with human-generated text; d) providing translation results in various formats like xml, json, tmx, etc; and e) improving accuracy through machine learning advancements. Utilizing a translation API in creating multilingual bibliographic content in a diverse country like India presents significant opportunities for libraries. By deploying data wrangling processes to handle MARC or DCMES formatted records, libraries can enhance accessibility and usability of information across different languages, thereby expanding their reach and impact within the

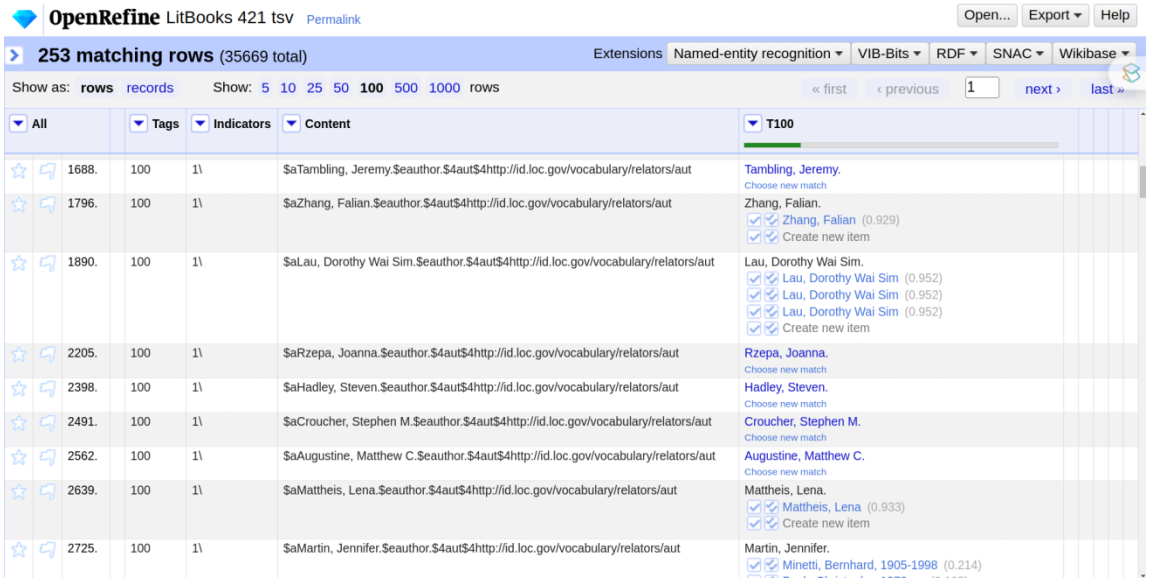


Fig. 4 — Data reconciliation for name authority from VIAF

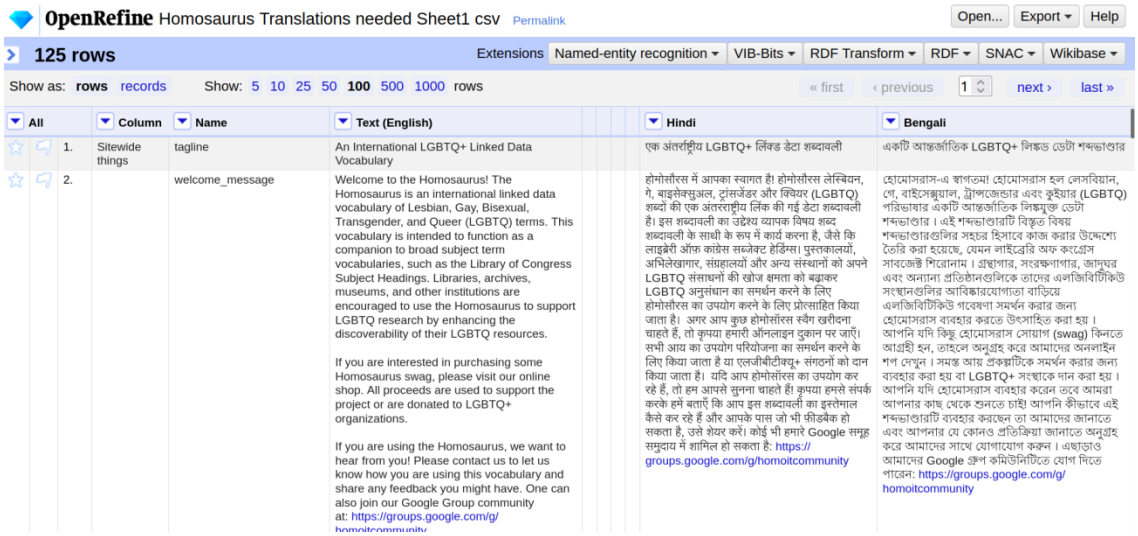


Fig. 5 — Machine translations of Homosaurus vocabulary through My Memory

multilingual community (e.g. translations of user interface of the library software popular in India like Koha, DSpace, VuFind etc).

The extensive study by Purkayastha (2019) explored the applications of translation APIs in the data wrangling tool OpenRefine. This research identified several entities that provide translation APIs, including: 1. Google Translate API; 2. Microsoft Translation API; 3. Translate API; 4. Text Translation API; 5. SYSTRAN.io Translation API; 6. My Memory Translation API; 7. My Translator Pro API; 8. Linguatools Translate API; 9. Yandex Translate API; and 10. IBM Watson Language Translator API. After experimenting with each API and considering factors relevant to Indian libraries, such as usage limits in the free tier, support for Indian languages, ability to support the GET method for data fetching, accuracy of text translations in Indian languages, and the ability to translate between Indian languages (e.g., Hindi to Bengali), this study selected the MyMemory Translation API (https://mymemory.translated.net/doc/spec.php). The MyMemory Translation API was chosen because it meets all the required criteria and offers the highest usage limit in the free tier (50,000 words/day with email-based authentication). To use the API, the syntax should include:

a) input text (UTF-8) in a given language (max 500 bytes); b) language pair (input/output) in ISO codes (ISO-639 – https://www.iso.org/iso-639-language-codes.html); c) output format (json (default) or tmx); d) valid email (to ensure full limit of the free tier); and e) option for machine translation on/off (1 (default), 0). This study also conducted a detailed analysis of the coverage of Indian languages (22 constitutionally recognized languages in the 8th schedule) in the MyMemory Translation Platform (MTP). The analysis revealed that the MTP can programmatically recognize 19 Indian languages at the input level (except the newly added languages Dogri, Meiti, and Santali), but it can provide content-level translation services for only 12 out of the 22 recognized languages.

Fig. 5 depicts how this study translated Homosaurus (a comprehensive LGBTQ+ vocabulary) control system) in Indian languages as a voluntary contribution to the Homosaurus community by using a machine translation tool

6 Conclusion

Library carpentry, a growing field in Library and Information Science, shows great promise for future progress. It focuses on organizing bibliographic data and solving real challenges for LIS professionals. This research looks at how library carpentry can be used, including tasks like data reconciliation and advanced techniques like Named Entity Recognition. Named Entity Recognition helps extract information from unstructured text. Sentiment analysis is also useful for sorting comments and reviews. Library carpentry can also create authority files like personal name authority, geographic name authority and convert them into a specific format like MARC 21 authority format. These opportunities highlight the diverse possibilities in library carpentry. As this field advances, it will improve how LIS professionals manage data, making library services more efficient.

7 References

1 Allison-Cassin, S. and Scott, D. (2018). Wikidata: a platform for your library's linked open data. The Code4Lib Journal, 40. https://journal.code4lib.org/articles/13424

2 Atwood, T. *et al*. (2019). Joining together to build more: the New England software carpentry library consortium. *Journal* of eScience Librarianship, 8(1). https://doi.org/10.7191/jeslib.2019.1161

3 Baker, J. *et al*. (2016). Library carpentry: software skills training for library professionals. *Liber Quarterly*, 26(3), 141–162. https://doi.org/10.18352/lq.10176

4 Bensmann, F., Zapilko, B. and Mayr, P. (2017). Interlinking large-scale library data with authority records. Frontiers in Digital Humanities, 4. https://doi.org/10.3389/fdigh.2017.00005

5 Bianchini, C. and Bargioni, S. (2021). Automated classification using linked open data. a case study on faceted classification and wikidata. Cataloging and Classification Quarterly, 59(8), 835–852. https://doi.org/10.1080/01639374.2021.1977447

6 Bowker, L. (2019). Machine translation literacy: Academic libraries' role. Proceedings of the Association for Information Science and Technology, 56(1), 618–619. https://doi.org/10.1002/pra2.108

7 Bowker, L. and Buitrago C. J. (2015). Investigating the usefulness of machine translation for newcomers at the public library. Translation and Interpreting Studies, 10(2), 165–186. https://doi.org/10.1075/tis.10.2.01bow

8 Brando, C., Frontini, F. and Ganascia, J. (2016). REDEN: named entity linking in digital literary editions using linked data sets. Complex Systems Informatics and Modeling Quarterly, 7, 60–80. https://doi.org/10.7250/csimq.2016-7.04

9 Burton, M. and Lyon, L. (2017). Data science in libraries. Bulletin of the Association for Information Science and Technology, 43, 33–35. https://doi.org/10.1002/bul2.2017.1720430409

10 Carlson, S. and Seely, A. (2017). Using Openrefine's reconciliation to validate local authority headings. Cataloging and Classification Quarterly, 55(1), 1–11. https://doi.org/10.1080/01639374.2016.1245693

11 Cope, J., Reed, P. and Ganeshwaran, N. (2020). Building a library carpentry community in the UK. UKSG eNews, 478. https://www.uksg.org/newsletter/uksg-enews-478/building-library-carpentry-community-uk

12 Delpeuch, A. (2019). A survey of OpenRefine reconciliation services. arXiv:1906.08092 [Cs]. http://arxiv.org/abs/1906.08092

13 Dennis, T. (2019). The Carpentries: Building local and global communities of practice to improve data skills. https://doi.org/10.5281/zenodo.3576151

14 Dennis, T., Chodacki, J. and Schneider, J. (2017). Taking the carpentry model to librarians. https://doi.org/10.5281/zenodo.1209481

15 Diaz-Valenzuela, I., Martín-Bautista, M. J. and Vila, M. A. (2010). An automatic data mining authority control system: a first approach. 2010 10th International Conference on Intelligent Systems Design and Applications, 569–574. https://doi.org/10.1109/ISDA.2010.5687205

16 Diaz-Valenzuela, I., Martin-Bautista, M. J., Vila, M. A. and Campaña, J. R. (2013). An automatic system for identifying authorities in digital libraries. Expert Systems with Applications, 40(10), 3994–4002. https://doi.org/10.1016/j.eswa.2013.01.010

17 Downey, M. (2019). Assessing author identifiers: preparing for a linked data approach to name authority control in an institutional repository context. Journal of Library Metadata, 19(1–2), 117–136. https://doi.org/10.1080/19386389.2019.1590936

18 Dryden, J. (2008). From authority control to context control. Journal of Archival Organization, 5(1–2), 1–13. https://doi.org/10.1300/J201v05n01_01

19 Harlow, C. (2015). Data munging tools in preparation for RDF: Catmandu and LODrefine. The Code4Lib Journal, 30. https://journal.code4lib.org/articles/11013

20 Hill, K. M. (2016). In search of useful collection metadata: using OpenRefine to create accurate, complete, and clean title-level collection information. Serials Review, 42(3), 222–228. https://doi.org/10.1080/00987913.2016.1214529

21 Knight, K. and Koehn, P. (2003). What's new in statistical machine translation. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Tutorials - NAACL '03, vol. 5, 5–6. https://doi.org/10.3115/1075168.1075173

22 Leiva-Mederos, A., Senso, J., Domínguez-Velasco, S. and Hípola, P. (2013). AUTHORIS: a tool for authority control in the semantic web. Library Hi Tech, 31(3), 536–553. https://doi.org/10.1108/LHT-12-20112-0135

23 Lemus-Rojas, M. and Pintscher, L. (2017). Wikidata and libraries: facilitating open knowledge. In Leveraging Wikipedia: Connecting Communities of Knowledge (ALA Editions, pp. 143–158). ALA. https://scholarworks.iupui.edu/handle/1805/16690

24 Li, X., Feng, J., Meng, Y., Han, Q., Wu, F. and Li, J. (2020). A unified MRC framework for named entity recognition. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5849–5859. https://doi.org/10.18653/v1/2020.acl-main.519

25 Manghi, P. and Mikulicic, M. (2011). PACE: A general-purpose tool for authority control. In E. García-Barriocanal, Z. Cebeci, M. C. Okur, and A. Öztürk (Eds.), Metadata and Semantic Research (pp. 80–92). Springer. https://doi.org/10.1007/978-3-642-24731-6_8

26 Mani, N. S., Cawley, M., Henley, A., Triumph, T. and Williams, J. M. (2021). Creating a data science framework: a model for academic research libraries. Journal of Library Administration, 61(3), 281–300. https://doi.org/10.1080/01930826.2021.1883366

27 McCallum, A. and Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 4, 188–191. https://doi.org/10.3115/1119176.1119206

28 McCutcheon, S. (2011). Basic, fuller, fullest: Treatment options for electronic theses and dissertations. Library Collections, Acquisitions, and Technical Services, 35(2–3), 64–68. https://doi.org/10.1080/14649055.2011.10766300

29 Mehrabi, N., Gowda, T., Morstatter, F., Peng, N. and Galstyan, A. (2020). Man is to person as woman is to location: Measuring gender bias in named entity recognition. Proceedings of the 31st ACM Conference on Hypertext and Social Media, 231–232. https://doi.org/10.1145/3372923.3404804

30 Mukhopadhyay, P. (2022). How green is my valley? Measuring open access friendliness of Indian Institutes of Technology (IITs) through data carpentry. In Biswas, A. & Das Biswas, M. (eds.) Panorama of Open Access: Progress, Practices and Prospects. pp. 67–89. New Delhi: Ess Ess. https://doi.org/10.5281/zenodo.6511080

31 Mukhopadhyay, P., Mitra, R. and Mukhopadhyay, M. (2021). Library carpentry: towards a new professional dimension (part i – concepts and case studies). Journal of Information and Knowledge, 58(2), 67–80. https://doi.org/10.17821/srels/2021/v58i2/159969

32 Mukhopadhyay, P. and Mukhopadhyay, M. (2021). Library carpentry: towards a new professional dimension (part ii – automatic authority control to enhance retrieval). Journal of Information and Knowledge, 58(3), 135–155. https://doi.org/10.17821/srels/2021/v58i3/163890

33 Mukhopadhyay, P. and Mitra, R. (2021). Library carpentry: Towards a new professional dimension (part iii – data reconciliation, named entity recognition and advanced utilities). Journal of Information and Knowledge, 58(5), 287–303. https://doi.org/10.17821/srels/2021/v58i5/166770

34 Mukhopadhyay, P. and Mukhopadhyay, M. (2022). Developing geodetic search interface through auto-generation of geographic name authority datasets: In Lamba, M. (ed.) Advances in Library and Information Science. pp. 59–81. Canada: IGI Global. https://doi.org/10.4018/978-1-7998-8942-7.ch004

35 Mukhopadhyay, P., Mukhopadhyay, M. and Ahmed, M. (2022). Retractions in India since independence: A multifaceted analysis for 75 years through data carpentry. Annals of Library and Information Studies (ALIS), 69(4), 304-322. https://doi.org/10.56042/alis.v69i4.67478

36 Park, Z. and Kim, H. (2014). Organizing and sharing information using linked data. In Park, J.R. & Howarth, L. C. (eds.) New Directions in Information Organization. pp. 61–87. Bingley: Emerald Group Publishing. https://doi.org/10.1108/S1876-0562(2013)0000007008

37 Parker, B., and Gray, A. (2019). Rethinking the University of Maryland authority file for the linked data environment. Journal of Library Metadata, 19(1–2), 69–81. https://doi.org/10.1080/19386389.2019.1589699

38 Purkayastha, S. (2019). Top 10 Best Translation APIs [2021] for Developers 20+ API reviewed. In Rakuten Rapid API Blog. https://blog.api.rakuten.net/top-10-best-translation-apis-google-translate-microsoft-translator-and-others/

39 Roy, A. and Mukhopadhyay, P. (2022a). Assessing open access friendliness of National Institutes of Technology (NITs): a data carpentry approach. DESIDOC Journal of Library and Information Technology, 42(5), 331–338. https://doi.org/10.14429/djlit.42.5.18263

40 Roy, A. and Mukhopadhyay, P. (2022b). Measuring open access friendliness of Indian central universities through data carpentry. Journal of Information and Knowledge, 59(3), 131–139. https://doi.org/10.17821/srels/2022/v59i3/170100

41 Roy, A. and Mukhopadhyay, P. (2022c). Measuring the open access friendliness of the state universities in India through data carpentry. Annals of Library and Information Studies, 69(3). https://doi.org/10.56042/alis.v69i3.63837

42 Ruttenberg, J. (2019). ARL white paper on Wikidata: opportunities and recommendations (p. 60). https://www.arl.org/wp-content/uploads/2019/04/2019.04.18-ARL-white-paper-on-Wikidata.pdf

43 Ryan, C. *et al*. (2015). Linked data authority records for Irish place names. International Journal on Digital Libraries, 15(2), 73–85. https://doi.org/10.1007/s00799-014-0129-8

44 Santamaría, L. and Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. PeerJ Computer Science, 4, e156. https://doi.org/10.7717/peerj-cs.156

45 Smith-Yoshimura, K. (2016). Analysis of international linked data survey for implementers. D-Lib Magazine, 22(7/8). https://doi.org/10.1045/july2016-smith-yoshimura

46 Smith-Yoshimura, K. (2018). Analysis of 2018 international linked data survey for implementers. The Code4 Lib Journal, 42. https://journal.code4lib.org/articles/13867

47 Tillman, R. K. (2016). Extracting, augmenting, and updating metadata in Fedora 3 and 4 using a local Openrefine reconciliation service. The Code4Lib Journal, 31. https://journal.code4lib.org/articles/11179

48 van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., and Van de Walle, R. (2015). Exploring entity recognition and disambiguation for cultural heritage collections. Digital Scholarship in the Humanities, 30(2), 262–279. https://doi.org/10.1093/llc/fqt067

49 Vellucci, S. L. (2004). Commercial services for providing authority control: outsourcing the process. Cataloging and Classification Quarterly, 39(1–2), 443–456. https://doi.org/10.1300/J104v39n01_12

50 Verborgh, R. and Wilde, M. D. (2013). Using OpenRefine. (Revised edn.). Burmingham, UK: Packt Publishing.

51 Virkus, S. and Garoufallou, E. (2020). Data science and its relationship to library and information science: a content analysis. Data Technologies and Applications, 54(5), 643–663. https://doi.org/10.1108/DTA-07-2020-0167

52 Wani, N. J., Mohanty, S. P., Purini, S. and Sharma, D. M. (2017). Anuvaad pranaali: a restful API for machine translation. In Drira, K. et al (eds.) Service-Oriented Computing – ICSOC 2016 Workshops (Vol. 10380). pp 179–183. New York: Springer International Publishing. https://doi.org/10.1007/978-3-319-68136-8_20

53 Wilkinson, M. D. et al (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

54 Wouter, K. (2019). IFLA – a concept framework for data science in libraries. In A concept framework for data science in libraries. https://www.ifla.org/publications/node/92282

55 Zhu, L. and Seggern, M. V. (2005). Vendor-supplied authority control: some realistic expectations. Technical Services Quarterly, 23(2), 49–65. https://doi.org/10.1300/J124v23n02_04