5-31-2022

Integration of Linked Open Data Authorities with OpenRefine : A Methodology for Libraries

Sukumar Mandal
Department of Library and Information Science, The University of Burdwan,
sukumar.mandal5@gmail.com

Integration of Linked Open Data Authorities with OpenRefine : A Methodology for Libraries

Dr Sukumar Mandal
Assistant Professor
Department of Library and Information Science, The University of Burdwan
Email: sukumar.mandal5@gmail.com

Abstract

The primary purpose of this paper is to explore the integration process of linked open data authority with OpenRefine for easy access of related metadata towards the creation of data cleaning and updating in a modern integrated library system. The integration process and methods are based on the API of reconciliation repositories collected from web resources. This integrated framework will be designed and developed on OpenRefine techniques and components based on RDF, CSV, SPARQL, and Turtle scripts. This integrated framework is based on JAVA and Apache Web Server for running the OpenRefine on the Ubuntu Platform. This integrated framework has been explored the data cleaning and import of bibliographic metadata from multiple linking authorities such as Open Library, ORCID, VIAF, VIAF BNF, Library of Congress Authorities data, and Wikidata. These are the essential findings of this study for creating a new interface to library professionals and advanced users. The library professionals are very much benefitted by using this system and services towards easy import and access of related linking resources from the Wikidata. Aside from these, it also explores the other facilities for data cleaning and updating the information from multiple scripts and URLs in the Web environment. It is possible to fetch related linking authorities for enhancing the advanced level services in a modern library management system. So, library carpentry and data carpentry are essential concepts for making a dynamic integrated interface for library professionals.

Keywords: OpenRefine, Reconcilation, API, Linking Authorities, Fetching URL, and Library Carpentry

Introduction

Carpentry is an essential concept in software technology. It is a part of data science for easy management of big data. Data and library carpentry are closely associated with authority linking of different repositories. The study of data using algorithms, methods, and systems is known as data science. As a result, machine learning and artificial intelligence are used to forecast and improve outcomes. To start making AI model performed in either hybrid multi-cloud scenario, gather, manage, and analyze data from any cloud. AI lifecycle and deployment should be shortened (https://www.ibm.com/in-en/analytics/data-science). OpenRefine is open source software in a linked data environment. It can be used offline to convert in columns and cells to use a web-based editor. It can discover data normalization issues by inspecting and faceting data. This page cannot cover every aspect of OpenRefine. It includes LCSH, Journal-TOCs, FAST support, and a robust API based on GRL regular expression scripting language. Many typical scriptings can be found online even without programming knowledge (Hill, 2016). Centuries of research have resulted in increased research power. Despite increased data collection, big data remains underused. As a result, libraries can educate the public about big data analytics. Libraries are becoming more data-savvy. Librarians should promote large-scale data methods and resources. Data science and large amounts of data can be used to group, recommend, and forecast outcomes. Library resources can help data scientists. Data science and big data are becoming increasingly important (https://www.discoverda tascience.org/resources/data-science-and-librarians/). Building digital libraries and research data infrastructures necessitate the use of technical experts (Borgman et al., 2015). This improves the speed of semantic inquiries. It shares data with computers via Web technologies like HTTP, RDF, and URIs on the Internet to build a global database. Tim Berners-Lee created the term in 2006 and essential components of data sets are below.:

Things should be named and identified using URIs.

Use HTTP URIs to look up, understand, and "words and actions”.

RDF and SPARQL, which are open standards, should help people find a name.

Users should be called by their HTTP URIs.

The modern Web uses hyperlinks to connect documents. To automate, find and link data collections. It is used by Google, Facebook, IBM, Oracle, and governments. JavaScript and Python are used in web-based Linked Data. HTTP URIs, RDF, and SPARQL queries are all examples. After that, there are web apps and mashups (https://www.manning.com/books/linked-data). Consider Linked Data identifiers for books, movies, actors, locations, and whatever else comes to mind. Library professionals may use it in this system and refer to it as needed. It's becoming more popular, especially since that data is readily available. Multiple linked data sets have emerged from Open Moving Data (https://cambridgesemantics.com). Ontologies are generated using the OWL language group. A taxonomy is a collection of object classes linked together by nouns and verbs. Domain models are used throughout the process. There are a lot of models that are used all the time. Use UML and MDA to work with each other. Specializing in a problem area may make it more difficult to reuse. The Web is a multi-dimensional answer—lack of use of the Web's model reuse features. It can use RDF enabled, Schema and OWL. A domain model is a way to think about how things work in a specific area (https://www.w3.org/2001/sw). The critical objectives of this paper are as (i) To explore the techniques of reconciliation using API services and software. (ii) To import the metadata from other repositories to provide library users resources. (iii) To clean and update the data for integrating the reconciliation system and services. (iv) To show the authority records and their related bibliographic metadata through the concept of data carpentry and data science.

Review of related works

It is a growing community of librarians that teach each other how to deal with data and deploy software to automate tiresome tasks. Data manipulation and organization, automating operations using a computer, and text pattern matching are only a few of the topics covered in Library Carpentry (Dennis, 2017). The "Carpentries" community created "Library Carpentry" as a lesson plan to make it easier to teach computational principles and abilities (Seidlmayer, Müller & Förstner, 2020). Every science discipline can make use of data science. A greater amount of assistance is required than can be offered by college resources. For teaching data science, academic libraries have a distinct advantage (Oliver & et.al, 2019). When scientists use computation to help them find new things in science, they are making a lot more discoveries. Scientists use a lot of scientific software in a lot of different fields. This includes geosciences, astronomy, and machine intelligence. Scientific software is hard to find, install, and compare because of inconsistencies in the documentation (manuals, quickstart guide files, websites, and code comments). Then, it is important to compare the technology that is already out there (Kelley & Garijo, 2021). Academic research software is becoming more important. Libraries at academic research schools are looking into ways to keep research software available for a long time (Chassanoff & et.al, 2018). Toolkits and principles for linked data are being adopted by the most essential points and lookup services, paving the way for the Semantic Web. It can improve the scholarly value of our content by utilising innovative technology. Open-licensed data and metadata can be used to create a web. Semantic meaning may be conveyed through a link. In addition to the URL, a link could also include the name of the author, the date the work was created, and the title of the work itself (Landis, 2014). Many librarians don't know what open data is. It looks like a lot of recent research has been focused on data that is linked. People still have to deal with things. It's written for people who are either new to LOD or need a refresher. Launch LOD and its rules and formats. In the second part, we'll talk about how LOD is being used in libraries, archives, and museums (Yoose & Perkins, 2013). It shows how to use Linked Data design and publishing tools and methods. This chapter talks about why and the way that data can be opened up. The most common methods are also shown (Konstantinou & Spanos, 2015). It's called "Scholarly Communication Brown Bag Lunch" at the UCFL Library. So, what now? Libraries and the Web (Cases, Standards and Vocabularies) (Deng, 2018). The quality of Open Data from contributing to and using them for multidimensional models, this paper illustrates how to make data from public sources generally available using semantic web standards like RDF and SPARQL (Escobar & et.al, 2020).

Reconciliation

The term API is used to refer to a user-facing software interface. The buzzword "middleware" comes from this. The word now refers to utility software and even hardware interfaces (Bloch, 2018). APIs connect computer programmes to other computer programmes. It is a piece of software that tells you how to make or use a connection like this. In this case, the API is made for the Assertion Protocol Interface. Software development can be sped up so that third-party services can be added to existing solutions or new apps can be made. These situations don't need you to look at the source code because of links and IRI. Unlike a user interface, which connects computers or software for people who aren't programmers to use, a service interface connects computers or software for people to use. Subroutines, methods, requests, and endpoints are all API calls. In this way, programmers can use APIs to meet their needs while the specifics of their own system stay the same. The Web service can be changed to work with a pair of technologies or a standard that works on many things (https://en.wikipedia.org/wiki/API). Based on some language that serves as a name or label, it generates a list of things that might be possible. The advanced level candidate terms do not have to match the official names of each category and compare OpenRefine fields with data from other sources that aren't fully automated. It can choose from a list of possible outcomes in which data are changed. Its data fields and columns need to be better and more consistent. Users could have also used this tool library carpentry to make data more significant and relevant. Comparing two comparable datasets may be found in academic, scientific, and reference libraries based on user-edited data on Wikidata reconciliation. API specifications for reconciliation services must be met by this web service for advanced library professionals. While OpenRefine attempts to match the cell values in the table CSV file to the reconciliation data, human review and approval are required.

Typographical errors, spaces, and additional characters will influence the results generated by the data cleansing and clustering processes. It can find more information on how to use OpenRefine's Wikidata reconciliation on Wikibase. Reconciliation capacity and services have been expanded, available on the Downloads page. These services use entity identifiers on VIAF and Wikidata IDs for the authority data and send these strings directly to the reconciliation service (https://docs.openrefine.org/manual/reconciling). However, the process of integration using reconciliation with different search engines and repositories are shown in Figure-1. It shows the connectivity process through proper URL based on reconciliation. It is a straightforward concept to import the data from other repositories with the help of open source tools, scripts and techniques based on reconciliation.

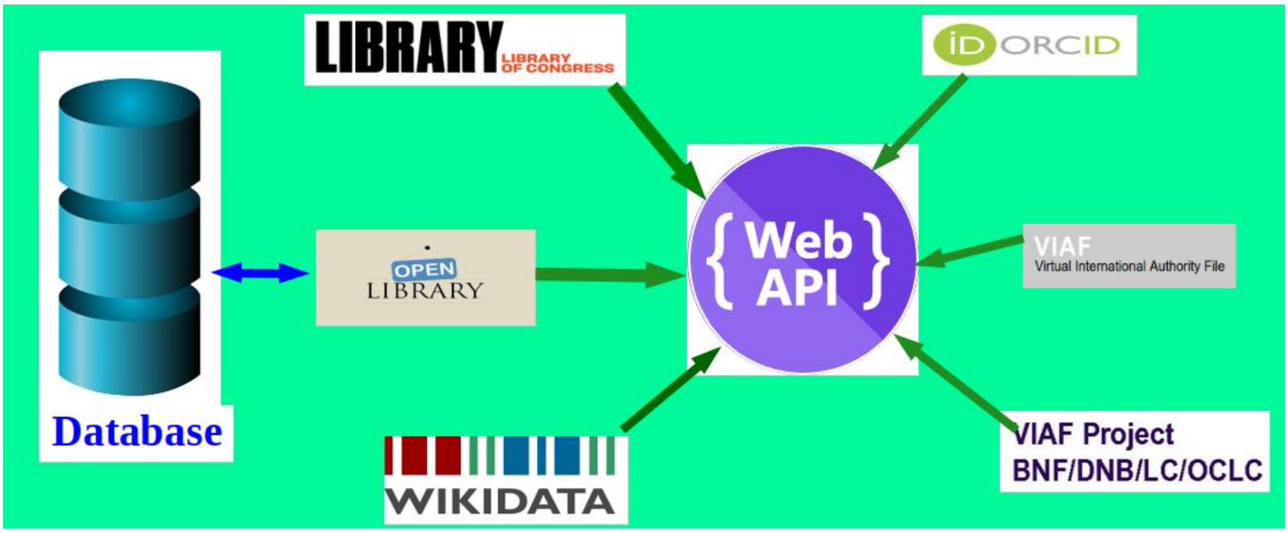


Figure-1: Reconciliation API Framework

Methodology

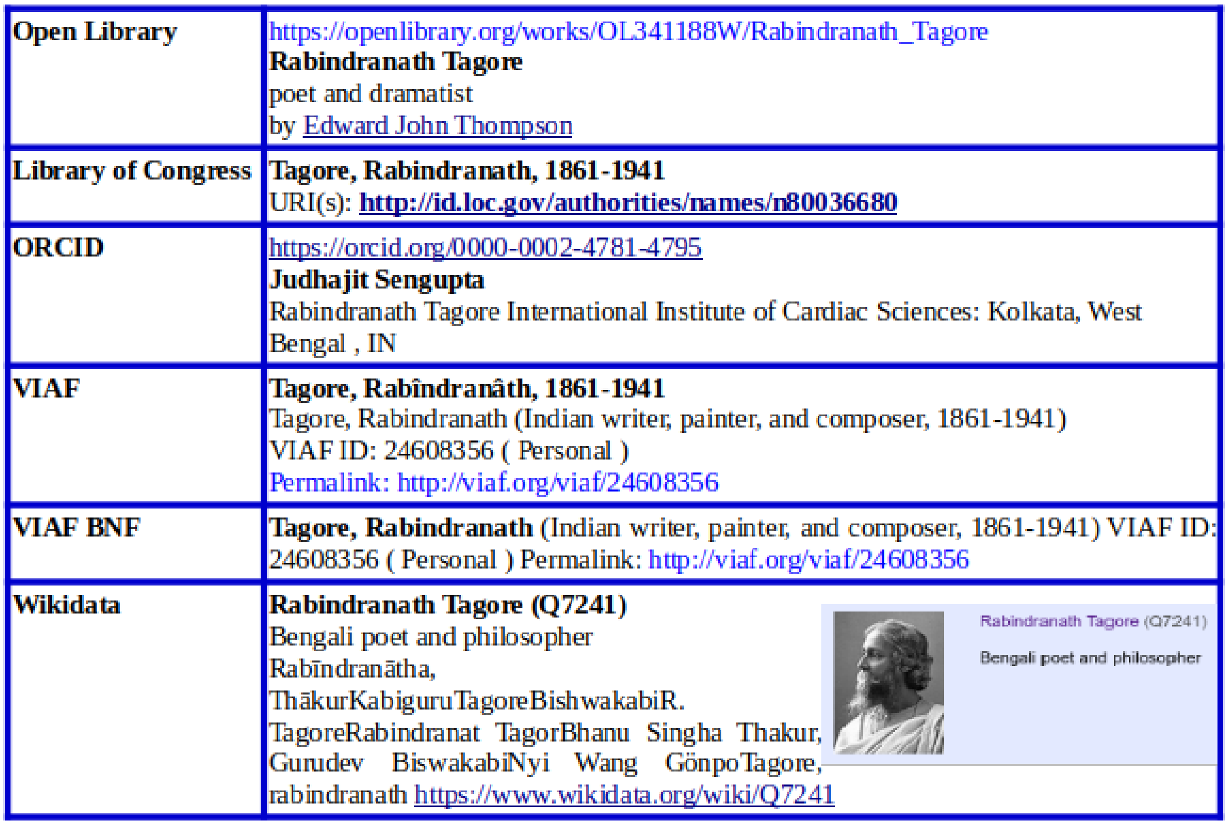
Generating the linked data access interface is very simple in this paper. It explores the methods and processes for easy integration of multiple authorities from different repositories using the OpenRefine software and technique. It is a well-known tool in modern linked open data reconciliation for increasing the innovative services in a library environment. This study is based on practical aspects consisting of three components: repository, related data authorities, and API services. The brief sketch is shown below:

Repository >> Linked Data Authorities >> API Services

e.g. Open Library >> https://openlibrary.org/works/OL341188W>> http://refine.codefork.com/reconcile/openlibrary

This paper has been selected only six linked data repositories such as the open library, library of congress, ORCID, VIAF, VIAF BNF, and Wikidata. Aside from these, it also needs to select the linked data authorities based on six repositories. Now all the tasks depend on good API services to linked data repositories. These will integrate with OpenRefine to retrieve the same information from other related authorities. This is a fundamental concept and aspect of automated and digital library services. Finally, display the results of linked data authorities (Figure-2) as below:

Rabindranath Tagore >> it will retrieved from Open Library
>> it will retrieved from Library of Congress
>> it will retrieved from ORCID
>> it will retrieved from VIAF
>> it will retrieved from VIAF BNF
>> it will retrieved from Wikidata



Open Library	https://openlibrary.org/works/OL341188W/Rabindranath_Tagore **Rabindranath Tagore** poet and dramatist by Edward John Thompson
Library of Congress	**Tagore, Rabindranath, 1861-1941** URI(s): **http://id.loc.gov/authorities/names/n80036680**
ORCID	https://orcid.org/0000-0002-4781-4795 **Judhajit Sengupta** Rabindranath Tagore International Institute of Cardiac Sciences: Kolkata, West Bengal , IN
VIAF	**Tagore, Rabîndranâth, 1861-1941** Tagore, Rabindranath (Indian writer, painter, and composer, 1861-1941) VIAF ID: 24608356 (Personal) Permalink: http://viaf.org/viaf/24608356
VIAF BNF	**Tagore, Rabindranath** (Indian writer, painter, and composer, 1861-1941) VIAF ID: 24608356 (Personal) Permalink: http://viaf.org/viaf/24608356
Wikidata	**Rabindranath Tagore (Q7241)** Bengali poet and philosopher Rabīndranātha, ThākurKabiguruTagoreBishwakabiR. TagoreRabindranat TagorBhanu Singha Thakur, Gurudev BiswakabiNyi Wang GönpoTagore, rabindranath https://www.wikidata.org/wiki/Q7241 Rabindranath Tagore (Q7241) Bengali poet and philosopher

Figure-2: Display the linked window

Linked Data Repository

Modern repositories are based on linked open data. There are a lot of related data repositories available, but this research paper has been selected only six digital repositories. All the repositories are enabled on reconciliation for easy retrieval of cross-domain information. The selected repositories are represented in Table-1. It also shows the linked data identifiers on six repositories. So, these integrated repositories provide very new and innovative services.

Table-1: Identifiers of Linked Data Repository

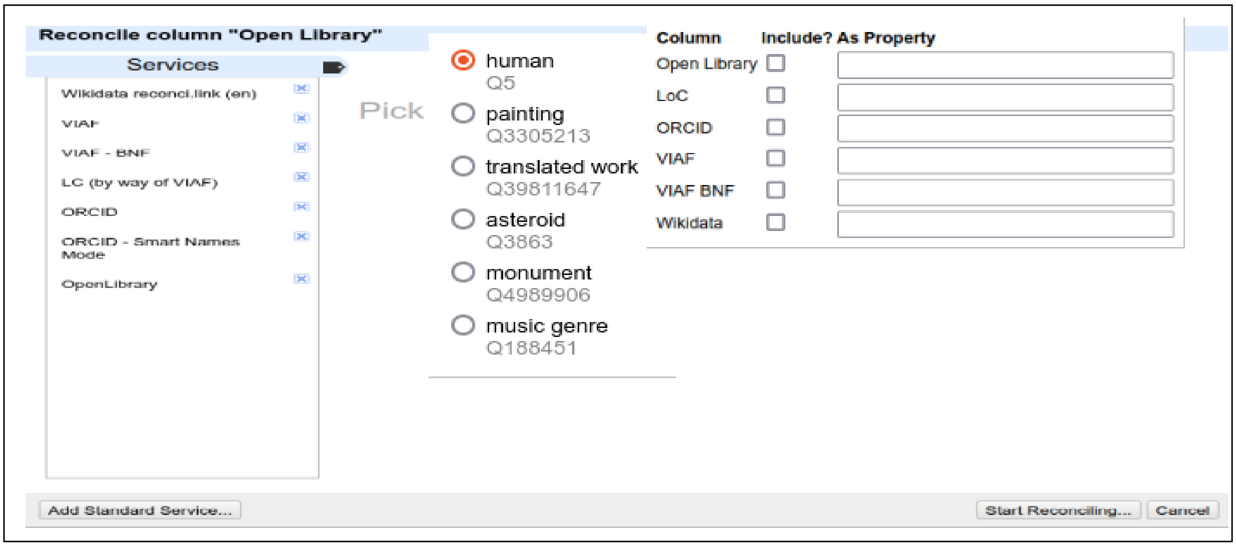
Repository	**Linked data**
Open Library	https://openlibrary.org/works/OL341188W
Library of Congress Authority Records	https://id.loc.gov/authorities/names/n80036680
ORCID	https://orcid.org/0000-0002-4781-4795
VIAF	https://viaf.org/viaf/24608356
VIAF BNF	https://viaf.org/viaf/96994048
Wikidata	https://www.wikidata.org/wiki/Q3020852

Collection and Configuration of API Services

It also needs to identify the Application Programming Interface services for integrating six repositories with OpenRefine. The API services of these repositories are explained in this section for six repositories such as Open Library, Library of Congress Authority Records, ORCID, VIAF, VIAF BNF, and Wikidata.

https://wikidata.reconci.link/en/api >> Wikidata
http://refine.codefork.com/reconcile/viaf >> VIAF
http://refine.codefork.com/reconcile/viaf/BNF >> VIAF BNF
http://refine.codefork.com/reconcile/viafproxy/LC >> Library of Congress Authorities
http://refine.codefork.com/reconcile/orcid >> ORCID
http://refine.codefork.com/reconcile/orcid/smartnames >> ORCID Smartnames
http://refine.codefork.com/reconcile/openlibrary >> Open Library

These API services are based on JSON programming scripts. However, six repositories scripts are shown in Figure-3 for easy integration of link repositories with OpenRefine. It needs to require a few steps to configure the reconcil field. First, Open the OpenRefine by Terminal and run the server as refined. Now, then required to import the RDF file into the OpenRefine for reconciliation of these repositories, click on the reconciliation option, and add these API URLs. Finally, click on the start reconciliation option towards importing the data from the other linked data repositories.



Reconcile column "Open Library"

Services

Wikidata reconci.link (en)
VIAF
VIAF - BNF
LC (by way of VIAF)
ORCID
ORCID - Smart Names Mode
OpenLibrary

Pick

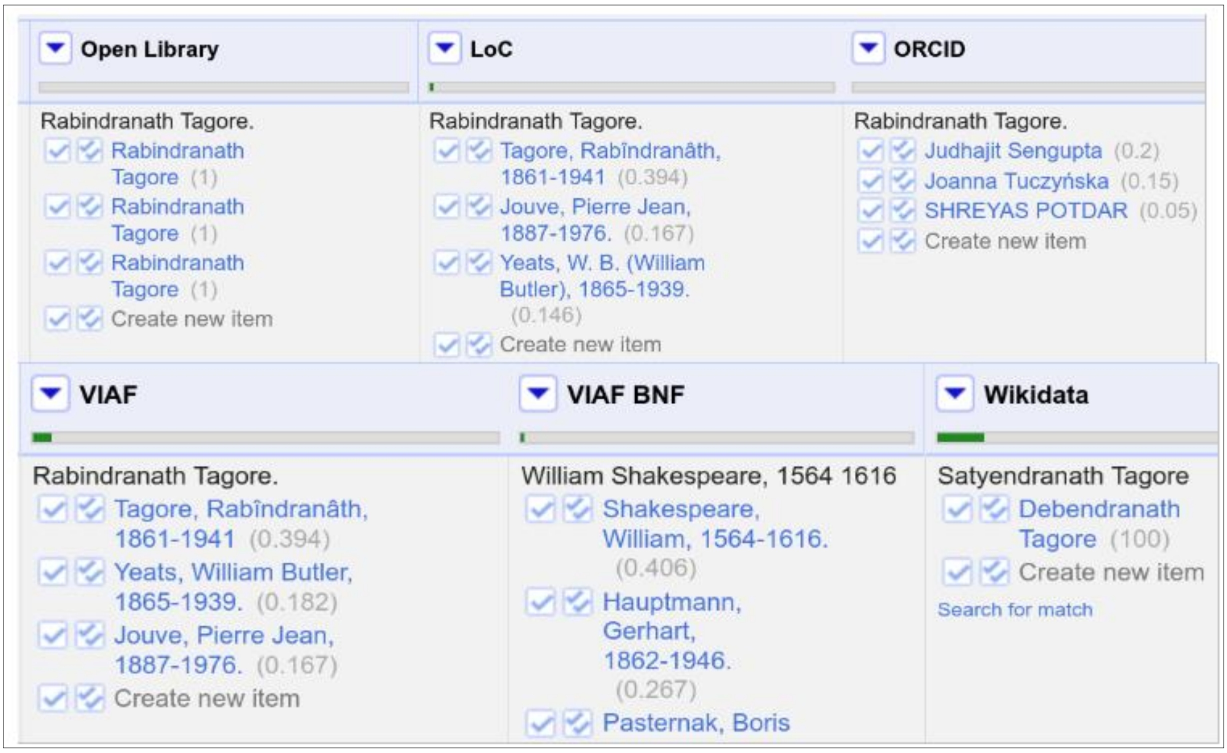
● human
Q5
○ painting
Q3305213
○ translated work
Q39811647
○ asteroid
Q3863
○ monument
Q4989906
○ music genre
Q188451

Column	Include?	As Property
Open Library	□	
LoC	□	
ORCID	□	
VIAF	□	
VIAF BNF	□	
Wikidata	□	

Figure-3: Reconcile configuration for linked authorities

Display of Linked Open Data Authorities

Linked Open Data authorities are displayed in OpenRefine search databases to enhance the web-scale, and library carpentry enabled services to develop modern digital repositories and services. This framework is based on a cross-searching platform because it can easily search the linked data authorities from other repositories. This framework explores and integrates the six popular linked repositories with OpenRefine to display the search results from multiple sources and repositories. All the data have been retrieved and accessed from six digital repositories to integrate with OpenRefine. Figure-4 shows the linked open data authorities from different electronic archives. The library professionals are benefitted from using this related data framework because it gives advanced level facilities and services.



Open Library

Rabindranath Tagore.
- Rabindranath Tagore (1)
- Rabindranath Tagore (1)
- Rabindranath Tagore (1)
- Create new item

LoC

Rabindranath Tagore.
- Tagore, Rabîndranâth, 1861-1941 (0.394)
- Jouve, Pierre Jean, 1887-1976. (0.167)
- Yeats, W. B. (William Butler), 1865-1939. (0.146)
- Create new item

ORCID

Rabindranath Tagore.
- Judhajit Sengupta (0.2)
- Joanna Tuczyńska (0.15)
- SHREYAS POTDAR (0.05)
- Create new item

VIAF

Rabindranath Tagore.
- Tagore, Rabîndranâth, 1861-1941 (0.394)
- Yeats, William Butler, 1865-1939. (0.182)
- Jouve, Pierre Jean, 1887-1976. (0.167)
- Create new item

VIAF BNF

William Shakespeare, 1564 1616
- Shakespeare, William, 1564-1616. (0.406)
- Hauptmann, Gerhart, 1862-1946. (0.267)
- Pasternak, Boris

Wikidata

Satyendranath Tagore
- Debendranath Tagore (100)
- Create new item

Search for match

Figure-4: Display the six linked data authorities

Conclusion

From the above discussions, it is clear that integration is possible with multiple repositories of authorities. Library professionals use this technique to access the related information from the other repositories. This paper has successfully shown the linked authority data in OpenRefine based on the reconciliation application programming interface. RDF and SPARQL endpoint interfaces are beneficial among the librarians towards moving metadata into linked open data. However, users quickly access the related data from six well-known repositories: Open Library, ORCID, VIAF, VIAF BNF, Library of Congress Authorities, and Wikidata. It is a dynamic process to convert and import the data from other databases using the OpenRefine based on carpentry related systems and services. So, Library carpentry and data carpentry are possible through these techniques and scripts for data cleaning in facets and text of multiple contents.

References

A semantic web primer for object-oriented software developers (n.d.). Retrieved from https://www.w3.org/2001/sw/BestPractices/SE/ODSD/ (Accessed on February 20, 2022)

Bloch, Joshua (2018). A Brief, Opinionated History of the API (Speech). *QCon. San Francisco: InfoQ*. Retrieved September 18, 2020. https://en.wikipedia.org/wiki/API (accessed 1 December 2021)

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: Data, diversity, and digital libraries. *International Journal on Digital Libraries, 16* (3–4), 207–227. https://doi.org/10.1007/s00799-015-0157-z (Accessed on February 20, 2022)

Chassanoff, A., AlNoamany, Y,. Thornton, K. & Borghi, J., (2018). Software Curation in Research Libraries: Practice and Promise, *Journal of Librarianship and Scholarly Communication*, 6 (1), 2239. doi: https://doi.org/10.7710/2162-3309.2239 (accessed 2 December 2021)

Data Science (2022). About on data science. https://www.ibm.com/in-en/analytics/data-science (Accessed on February 20, 2022)

Data Science and Library (2022). Data science for librarians. Retrieved from https://www.discoverda tascience.org/resources/data-science-and-librarians/ (Accessed on February 20, 2022)

Deng, S. (2018). Linked data in the library & OpenRefine. *Stay Savvy with Scholarly Communication Brown Bag Lunch, University of Central Florida Libraries*. https://stars.library.ucf.edu/ucfscholar/774/ (accessed 4 December 2021)

Dennis, T. (2017). Taking the Carpentry Model to Librarians. *UCLA: Library*. http://dx.doi.org/10.5281/zenodo.1112373 Retrieved from https://escholarship.org/uc/item/7hk 6f84g (accessed 9 December 2021)

Escobar, P., Candela, G., Trujillo, J., Marco-Such, M., & Peral, J. (2020). Adding value to Linked Open Data using a multidimensional model approach based on the RDF Data Cube vocabulary. *Computer Standards & Interfaces, 68*, 103378. https://doi.org/10.1016/j.csi.2019.103378 (accessed 16 December 2021)

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 (Accessed on February 20, 2022)

Introduction to linked data. (n.d.). *Cambridge Semantics*. Retrieved from https://cambridgesemantics.com/blog/semantic-university/intro-semantic-web/intro-linked-data/ (Accessed on February 20, 2022)

Kelley, Aidan & Garijo, Daniel (2021). A framework for creating knowledge graphs of scientific software metadata. *Quantitative Science Studies*. doi: https://doi.org/10.1162/qss_a_00167 (accessed 18 November 2021)

Konstantinou, N., & Spanos, D. E. (2015). Deploying linked open data: Methodologies and software tools. In *Materializing the Web of Linked Data* (pp. 51-71). Springer, Cham. DOI: 10.1007/978-3-319-16074-0_3 (accessed 21 November 2021)

Landis, Cliff (2014). A web of meaning: linked open data resources on the web. *University Library Faculty Publications*, 118. https://scholarworks.gsu.edu/univ_lib_facpub/118 (accessed 6 January 2022)

Linked data. (n.d.). Manning Publications. Retrieved from https://www.manning.com/books/linked-data (Accessed on February 20, 2022)

Oliver, J. C., Kollen, Christine, Hickson, Benjamin & Rios, Fernando (2019). Data Science Support at the Academic Library. *Journal of Library Administration*, 59 (3), 241-257, DOI: 10.1080/01930826.2019.1583015 (accessed 8 January 2022)

Seidlmayer, E., Müller, R. & Förstner, K. (2020). Data Literacy for Libraries – A Local Perspective on Library Carpentry. *Bibliothek Forschung und Praxis*, *44* (3), 485-489. https://doi.org/10.1515/bfp-2020-2038 (accessed 12 January 2022)

Tim Berners-Lee (2006). Linked Data. Design Issues. W3C. Retrieved 2010-12-18. (Accessed on February 20, 2022)

Yoose, B., & Perkins, J. (2013). The linked open data landscape in libraries and beyond. *Journal of Library Metadata*, *13* (2-3), 197-211. https://doi.org/10.1080/19386_389._2013.826075 (accessed 15 January 2022)

OpenRefine (2022). About on OpenRefine and Reconcilation. https://docs.openrefine.org/manual/reconciling (accessed 16 January 2022)