I J C R M
International Journal of Contemporary Research In Multidisciplinary

Research Article

Data wangling in Libraries: the power of OpenRefine software

Sheuli Hazra
PhD Research Scholar, RKDF University. Ranchi, Jharkhand, India

Corresponding Author: *Sheuli Hazra

DOI: https://doi.org/ 10.5281/zenodo.19662493

Abstract

OpenRefine is super important for our data publishing workflow with Open Context, and many of you will find it a great way to reduce the tedium in cleaning data. A few dimensions of data quality are accuracy or correctness, comparability, consistency, coherence or clarity, completeness, credibility, reliability, or usefulness, timeliness or latency, uniqueness, validity or reasonableness. OpenRefine is a software program that installs on your own computer. It uses Java to power a web server, and even through open refine runs as a web server, you interact with the application through a web browser like Safari, Chrome, or Firefox. But even though OpenRefine runs as a web server, it is running on your own computer, not on the internet. For that reason, it is important to use private or sensitive information securely, like on your own computer. Because it runs on its own device, it is not like Google Drive, Google spreadsheets, or other computing services.

Manuscript Information

- **ISSN No:** 2583-7397
- **Received:** 13-03-2026
- **Accepted:** 16-04-2026
- **Published:** 20-04-2026
- **IJCRM:**5(2); 2026: 706-710

- **Plagiarism Checked:** Yes
- **Peer Review Process:** Yes

How to Cite this Article

Hazra S. Data wangling in Libraries: the power of OpenRefine software. Int J Contemp Res Multidiscip. 2026;5(2): 706-710.

Access this Article Online

www.multiarticlesjournal.com

KEYWORDS: Open Refine, data wangling, data cleaning, data transformation.

1. INTRODUCTION

Open access relates to the liberated, open virtual access due to the results of studies, such as journal, articles, and books. (K. G. Sudhier, 2024a). OpenRefine is a robust instrument for data exploration, washing, and transformation. In this study, it will be slowly discovered how to leverage Refine to retrieve URLs and dissect network content. Data exploration, cleansing, and transformation are all made easier using OpenRefine, which is a strong tool. In this session, you will learn how to retrieve URLs by utilising the Refine algorithm and parse web content. (Williamson, 2017a). A free open-source energy tool for working with nasty data sets. A OpenRefine as a Data Exploration Tool. OpenRefine can normalise and visualise your data. OpenRefine seems to be like a spreadsheet, but it performs comparable with database. Use OpenRefine to explore, clean, and link your data. Pull with major granular details when there is a great deal of information available regarding this thing. Pinpoint very specific criteria using a combination of facets. According to Ham in 2013, said that it is quite potent. OpenRefine is a powerful tool to fetch, extract, and manage the value, volume, and variety of the datasets as gathered by following the stated methodology, and without OpenRefine, this study simply could not have been possible. (Mukhopadhyay & Roy, 2022). Users may analyse data to gain an overview, cleanse and modify data, and synchronise data with many online services.

2. LITERATURE REVIEW OF THE STUDY

Data Carpentry (Teal et al., 2015) established that organisations such as Software Carpentry (Wilson, 2013). OpenRefine (https://openrefine.org/) is a powerful, open-source software that simultaneously visualises and manipulates large quantities of data. Previously known as GoogleRefine, it was released as open-source software in 2013 and is now maintained by a diverse international community. (K. G. Sudhier, 2024b).

The software carpentry project has been around since 1998, and since it began offering its two-day workshop series in 2012, it has provided over 30,000 researchers with the opportunity to learn practical research skills. (Atwood et al., 2019). The importance of software in contemporary scientific research is comparable to that of test tubes and telescopes. More and more of the day-to-day operations of science revolve around the creation of new algorithms, the management and analysis of the large amounts of data that are generated in single research projects, the combination of disparate datasets to evaluate synthetic problems, and other computational tasks. (Crespo Garrido et al., 2025). Over the past two decades, the field of CSE has penetrated both basic and applied research in academia, industry, and laboratories to advance discovery, optimise systems, support decision-makers, and educate the scientific and engineering workforce. (Rüde et al., 2018). Computational science and engineering (CSE) applications can benefit from the adoption of some commercial software engineering practices. OpenRefine is designed for certain purposes and offers compelling reasons for its utilisation. (Bortruex, 2022). OpenRefine, an open-source program accessible on GitHub, is advantageous due to its cost-effectiveness, user-friendliness, and capability to choose a union of rows after the application of facets and filters. OpenRefine offers a platform for data management, and users, including non-programmers, do not need to know how to code to clean and transform data. (Sterner, 2019). During the course of the answers, data quality is typically associated with data preprocessing, profiling, and cleansing for subsequent tasks like data integration or data analytics. (Ehrlinger & Wöß, 2022). This web scraping exercise is intended to illustrate what "thinking in patterns" means. The patterns might not reside within the data itself but might be in how the data is presented. Offsets, fonts, colours, etc., can all be used to parse out the data if such presentation information can be recovered and fed into Google Refine. (Williamson, 2017b).

Data exploration, cleansing, and transformation are all made easier using OpenRefine, which is a strong tool. In this lesson, you will learn how to use Refine to fetch URLs and parse web content. (Williamson, 2017c).

Characteristics of OpenRefine:

Strengths

Complimentary and Open Source. It possesses far greater power than Excel. The platform is independent. The extensive historical documentation. It can export frequently utilised functions for reutilization.

Vulnerabilities

May exhibit occasional instability (some queries may execute slowly). Certain approaches need basic programming skills. Certain operations that are straightforward in Excel are more challenging or unfeasible with OpenRefine, such as adding additional rows or data and modifying specific cells.

Practical

Initiating a New Project. Records vs Rows. Rearranging Columns. Fundamental Normalisation. Monitoring Operational Histories. Faceting and Clustering. Exporting your output.

Hands-on

Creating a New Project. Records vs. Rows. Reordering Columns. Basic Normalization. Tracking Operation Histories. Faceting and Clustering. Exporting the work.

3. OBJECTIVES OF THE STUDY

OpenRefine operates with local files or data from online URLs in several file formats, including CSV, TSV, XLS, XML, and others.

- It possesses the capability to filter or search for specific items requiring modification, so limiting the display to only the pertinent cells, rows, or columns containing those elements. The user can thereafter execute the intended action just on those entries.
- It can identify duplicate entries, vacant cells, changes in entries, inconsistencies, and patterns of mistakes for mass rectification and cleansing.
- It offers a rapid examination of the data within the file; for example, the word facet tool may evaluate the words in a column and yield a tally of each unique word, with results ordered alphabetically by default, though sorting by count reveals trends at a glance.

- It offers an Undo/Redo functionality for all activities executed on the data, therefore conserving time and effort by retrieving and reapplying instructions; for instance, an address field may encompass city, state, and zip code, and an operation may be conducted to segregate the data into three distinct columns. When the identical issue arises in another project, it is a straightforward process to replicate the action from one project and implement it in the data field.
- It employs the Google Refine Expression Language (GREL) as its primary language for data transformation and creation, while also supporting Jython and other computer languages; the documentation states that GREL is structured to mimic JavaScript. A GREL expression may be utilised to identify all occurrences of a specific text string and substitute it with an alternative one.
- Users may discover that the data requires reconciliation, linkage, or augmentation with reliable sources, some of which are accessible via OpenRefine. The VIVO Scientific Collaboration Platform facilitates the reconciliation of data about VIVO entities, including faculty members, publication titles, and other data entries. The user manual references other reconciliation services for alternative applications (Ham, 2013).

Significance of the study

OpenRefine is being presented collaboratively by the Outreach and Assessment and Metadata Working Groups, alongside the Digital Projects Librarian at Colorado State University. Before her current role, she worked at the Harry Ransom Centre at the University of Texas at Austin and served as a Metadata Coordinator at the University of Colorado Boulder. She possesses extensive experience in creating metadata for diverse cultural heritage materials, including costumes, personal effects, and print. Helen, OpenRefine is accessible for anyone eager to acquire knowledge from the foundational level, without requiring much prior understanding. There are several applications. Facets are crucial to Open Refine, as indicated by its faceted gemstone emblem. Open Refine displays data and facets that quantitatively summarise distinct values across several columns of your dataset, which is highly beneficial for data evaluation and cleansing. Upon installation and initiation, Open Refine will display a screen resembling this. Open a data file describing artefacts. This data file is in CSV format (comma-separated values). CSV is a standard and non-proprietary format for the exchange of tabular data. OpenRefine is capable of reading tabular data from Excel, as well as various other formats. UTF-8 is employed here for multilingual support. Upon successfully loading a dataset into OpenRefine. It will appear as follows. By default, it will represent the significant indentation of the library domain. OpenRefine, initially developed by Google, is utilised in specific scenarios. Primarily, we employ OpenRefine for managing extensive datasets; it effectively addresses issues such as Excel freezing or prolonged operation times. Additionally, we leverage OpenRefine for several specialised functionalities. When working, it is essential to export your data, and there are various options available for exporting as a CSV. It is important to note that CSV files do not support external links. If you wish to retain this functionality, you must save your work as an OpenRefine project, which can be directly stored in your Google Drive if you are connected. If your only intention is to connect to Wikipedia to enhance your dataset, you can export as a CSV, which will be available for immediate download. Should you need to reopen your project later, you can simply click 'open' and upload it again, or if you have closed it, I can assist you further.

4. METHODOLOGY OF THE STUDY

Data cleaning, previously known as Google Refine. It is now open access and open-sourced. Open Refine can be used for record cleaning besides the standardisation system. It can be found at openrefine.org. Open Refine is an impressive instrument used for working with complicated data. Open Refine supports faceted browsing as a mechanism for seeing a big picture of your data and filtering down to just the subset of rows that you want to change in bulk. When the clustering function is used, it makes an effort to arrange the options in the text facet in such a way that the options that “look similar” contract clustered together. Resolution is a half-automatic method of equating text labels to database IDs (keys). This is utilised OpenRefine, running resolution of names with the data sets, besides any kinds of database that shows a web service accountability. Open Refine, however, is not like other tools you have used. OpenRefine cannot be used for storing or managing data; it is strictly a cleaning and/or standardising tool. As OpenRefine is a different kind of tool, one should consider when it is appropriate to use it versus other tools. A database provides infrastructure aimed at storage capacity and indexing of datasets. Generally, it requires programming skills to edit and is absent of easy visualisation. Excel is a spreadsheet application. It is useful for documenting data and performing operations. And while you can manage your data and have a limited ability to clean it. Data is not always visible, and it lacks powerful visualisation tools. OpenRefine, in contrast, offers multi-cell editing, easy exploration, transformation, and next cooperative visualisation. On the other hand, as was said earlier, it is not aimed at storage, but rather at operating data sets. Here, understand the differences between. Here is a list of useful features that you will find within OpenRefine. Interactive visualisation, multi-cell editing, easy transformation of data, faceting/ filters, use of regular expressions and APIs, desktop app, many import/ exports, large community, free. OpenRefine is software that you install on your computer. It requires the JAVA JRE/ JDK to run. It works on Windows, Mac and Linux. As OpenRefine is free and open source, it is supported by a large community of developers and users. It is easy to find tutorials online on how to use the tool.

Data collection: Data Purging, Removing mistakes and discrepancies. Assuring the integrity of data collection.

Data organising: Data material so that it is regular to find and comprehend. This is making use of information and standardised formats.

Data visualising: Data making graphics that convey information clearly and concisely, making use of resources and methods.

Data mining: Enhancing the data is a crucial phase, whereby additional information is incorporated to augment its utility for analysis and confirmed to guarantee its precision and quality. Data wrangling enhances the accessibility and significance of raw data, allowing analysts and data scientists to extract useful insights with greater efficiency and precision.

Data wrangling: This procedure includes sanitising the data by eliminating or resolving errors, inconsistencies, and redundancies. It also entails organising the data, frequently transforming it into a tabular format that facilitates analysis in programs. It enhanced user productivity by enabling them to conduct their own analyses and interactively explore and manipulate data according to their specific requirements, thereby eliminating the dependence on conventional business intelligence developers for report and dashboard creation, a process that may require days, weeks, or longer. Users can perform ad hoc analysis and run follow-up queries to answer their own questions. (Quinto, 2018). Information wrangling is the method of acquiring, transforming, and enhancing raw information that can be used for further analysis and visualisation.

- All through various bit-by-bit effects, pick up how to get, clean, analyse, and present information effectively. It will learn basic Python syntax, data types and language concepts also
- Work with both machine-readable and human-consumable data
- Scrape websites and APIs to find a bounty of useful information
- Clean and format data to eliminate duplicates and errors in your datasets
- Learn when to standardise data and when to test and script data cleanup
- Explore and analyse your datasets with new Python libraries and techniques
- Use Python solutions to automate your entire data-wrangling process (Kazil & Jarmul, 2016)

Data cleaning: Data fixing mistakes and discrepancies, maintaining the accuracy of the data. OpenRefine, Python, and all are operated by data cleaning development issues.

Data integration: Two methods are included, which are API providers and Database Management Systems (SQL, NoSQL).

API providers: Enhance organisational API research by employing an agility framework to investigate the results of agility following API integration. The qualitative data obtained from a music digital firm indicated four principal agility outcomes: customer agility, characterised by rapid customer feedback; operational agility, manifested through enhanced business processes and reduced delays; partner agility, exemplified by the adoption of flexibility in processes and structures along with ecosystem expansion; and decision agility, reflected in expedited decision-making. (Ofoeda et al., 2024).

Database Management Systems (SQL, NoSQL): A SQL Server database and a non-relational MongoDB database utilising an unstructured data representation in JSON format. A substantial body of work exists comparing various database management tools based on performance, security, and other criteria; nevertheless, there is a paucity of material evaluating these databases based on the specified parameters. (Malik et al., 2020).

Data analysis: Data analytics regards new data and converts it into valuable knowledge. It uses different tools and methods to discover patterns and solve problems with data sets. Data analytics helps businesses make better decisions and grow. Companies worldwide produce substantial amounts of data daily, including log files, web server data, transactional information, and diverse customer-related data. In addition to this, social media websites also generate massive amounts of data interchange. Python, R and others are engaged with their purpose fully in data carpentry in the library.

5. CONCLUSION OF THE STUDY

A library carpentry and data carpentry technique that is considered to be the most effective is data wangling. Data carpentry is the policy of the system renovation of the library. The library is the foundation of great progress and development. The library is the backbone of the educational growth and development system. Data wrangling is another kind of abstracting of the article system of a library. A well-equipped library provides access to books, journals, and digital resources, which help in academic growth and intellectual development. Therefore, the library acts as the backbone of educational advancement and overall development with the data carpentry system.

REFERENCES

1. Atwood TP, Creamer AT, Dull J, Goldman J, Lee K, Leligdon LC, et al. Joining together to build more: the New England Software Carpentry Library Consortium. *J eSci Librarianship*. 2019;8(1):e1161. doi:10.7191/jeslib.2019.1161.
2. Bortruex D. Review of OpenRefine software. *TCB Tech Serv Relig Theol*. 2022;30(2):7. doi:10.31046/tcb.v30i2.3132.
3. Crespo Garrido IDR, Loureiro García M, Gutleber J. The value of an open scientific data and documentation platform in a global project: the case of Zenodo. In: Gutleber J, Charitos P, editors. *The economics of big science 2.0*. Switzerland: Springer Nature; 2025. p. 181–200. doi:10.1007/978-3-031-60931-2_14.
4. Ehrlinger L, Wöß W. A survey of data quality measurement and monitoring tools. *Front Big Data*. 2022;5:850611. doi:10.3389/fdata.2022.850611.
5. Sudhier KG. Mastering data carpentry for open access publications: unlocking the OA research potential. 2024. doi:10.13140/RG.2.2.23025.49768.
6. Kazil J, Jarmul K. *Data wrangling with Python*. 1st ed. Sebastopol: O'Reilly; 2016.

7. Malik A, Burney A, Ahmed F. A comparative study of unstructured data with SQL and NoSQL database management systems. *J Comput Commun*. 2020;8(4):59–71. doi:10.4236/jcc. 2020.84005.
8. Mukhopadhyay P, Roy A. Measuring the open access friendliness of state universities in India through data carpentry. *Ann Libr Inf Stud*. 2022;69(3). doi:10.56042/alis.v69i3.63837.
9. Ofoeda J, Boateng R, Effah J. API integration and organisational agility outcomes in digital music platforms: a qualitative case study. *Heliyon*. 2024;10(11):e31756. doi:10.1016/j.heliyon.2024.e31756.
10. Quinto B. Big data visualisation and data wrangling. In: *Next-generation big data*. New York: Apress; 2018. p. 407–476. doi:10.1007/978-1-4842-3147-0_9.
11. Rüde U, Willcox K, McInnes LC, Sterck HD. Research and education in computational science and engineering. *SIAM Rev*. 2018;60(3):707–754. doi:10.1137/16M1096840.
12. Sterner E. Cleaning collections data using OpenRefine. *Issues Sci Technol Libr*. 2019;(92). doi:10.29173/istl30.
13. Teal TK, Cranston KA, Lapp H, White E, Wilson G, Ram K, et al. Data Carpentry: workshops to increase data literacy for researchers. *Int J Digit Curation*. 2015;10(1):135–143. doi:10.2218/ijdc. v10i1.351.
14. Williamson EP. Fetching and parsing data from the web with OpenRefine. *Program Hist*. 2017;(6). doi:10.46430/phen0065.
15. Wilson G. Software Carpentry: lessons learned. 2013. doi:10.48550/arXiv.1307.5448.

About the corresponding author

Sheuli Hazra is a PhD Research Scholar at RKDF University, Ranchi, Jharkhand, India. Her research interests lie in interdisciplinary studies, focusing on social development, education, and emerging societal challenges. She is committed to academic excellence and contributes to scholarly research through critical analysis and innovative perspectives in her field.