ISSN (Print) : 0972-2467
ISSN (Online) : 0976-2477
Journal of Information and Knowledge, Vol 60(2), April 2023, p.113-125
DOI: 10.17821/srels/2023/v60i2/170971

Data Wrangling from Socio-Academic Web-Space: Designing a Meta Model

Amit Nath* and Sibsankar Jana

Department of Library and Information Science, University of Kalyani, Kalyani –741235, West Bengal, India; amitnathdlis@gmail.com

Abstract

Data carpentry is an emerging field in the domain of LIS and has opened new possibilities for information professionals to survive in the age of data-intensive information services. However, library professionals face the challenges of information overload because of the free availability of data, both in terms of quantity and variety. The role of library professionals is moving from tech-savvy to data-savvy. This research discusses the possibilities of ODbL-based data sources that offer freely accessible data through API calls and proposes a meta-model for fetching and extracting datasets from these databases using an open-Source Data Wrangling Tool (OpenRefine). Further, it discusses the possible application of data wrangling techniques from diverse sources in libraries and how information professionals can take advantage of openly available data to provide value-added information services to users. The practical implications of an array of databases are projected through two case studies: Case Study I deals with measuring the productivity of individual institutions through different metrics, and Case Study II projects a coverage comparison among the ODbL-based citation and Altmetric databases. This meta-model will aid in understanding the potential application of data wrangling techniques to an array of library services.

Keywords: Data Wrangling, Data Carpentry, OpenRefine, ODbL Databases, Socio-Academic Data, REST/API

1. Introduction

Data is considered an asset in the digital economy, right from government organizations to small-scale industries. The power of any individual organization or country is measured by its data quality: how much data do we have? What kind of data do we have? People are generating different kinds of data by simply clicking or leading their daily routines. As a result, the quantity, nature, and complexity of data are growing day-by-day. In the meantime, the emergence of open science technologies (standards, protocols, software) takes away the attention of worldwide researchers to develop datasets (mainly ODbL-based data sources) that are openly available for information seekers. The term ODbL stands for Open Data Commons Open Database License; it means we are free to share the database (use, copy, and distribute), to produce different services from the databases, and to modify and transform the database as per needs. In this paper, we discuss the ODbL-based socio-academic databases that provide different kinds of data to the researchers and develop a meta-approach for extracting data from the large array of databases through REST/API calls. Further, we examine how researchers and information professionals can take advantage of openly available data for their research work, and to provide day-to-day library services.

2. Data Wrangling: What and Why

Data wrangling is the process of acquiring, transforming, and enriching raw data that can be used for further analysis and visualization. This process includes following steps like exploring and cleaning the raw data, structuring the data into tables, standardizing data formats, and dealing with data curation issues like typos, identifying missing values, changes in units of measurement, etc.

*Author for correspondence

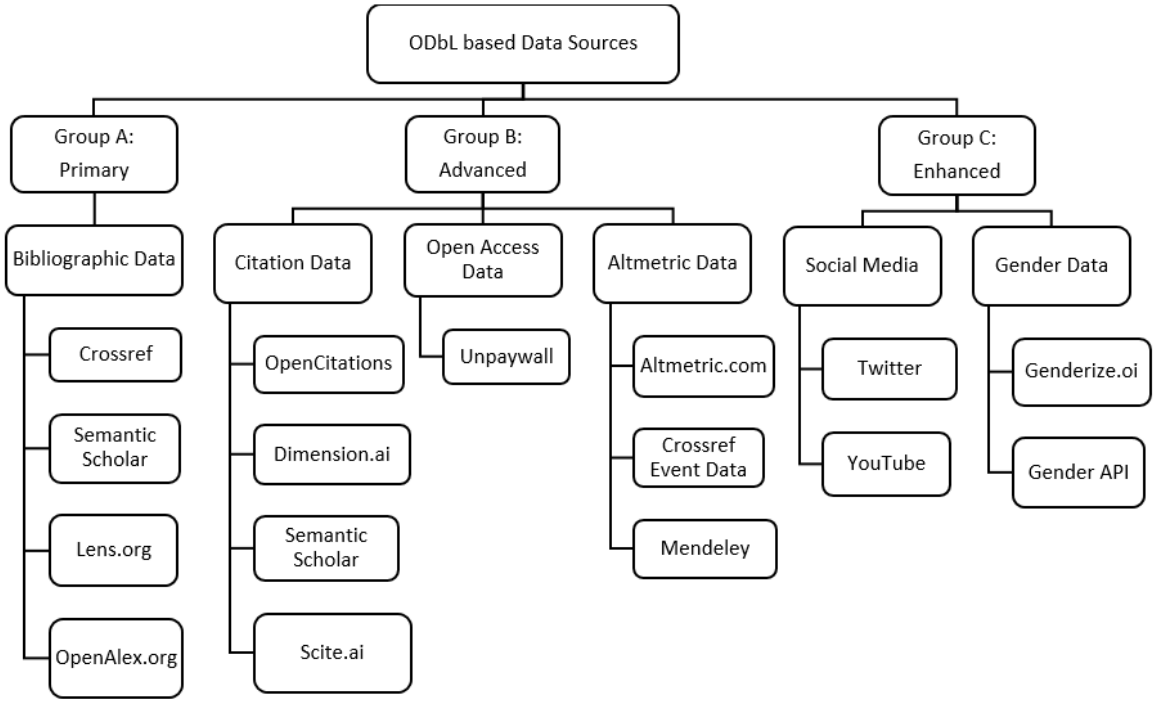

Figure 1. Categorization of ODbL databases.

Library professionals face the challenge of information overload because of freely available data, both in terms of quantity and variety (Robinson and Bawden, 2017). They deal with different kinds of data in day-to-day library services like bibliographic data, research data, usage data, social media data, authority data, and so on. Traditional data processing techniques, however, can no longer manage the data due to the enormous volume and diversity of the data (Kandel *et al.*, 2011; Eberendu, 2016). So, information professionals should be capable of handling data of all types confidently for their own goals, such as using data analytics to enhance their library services, even if they do not specifically assist their users in dealing with data.

3. Data Wrangling: Tools

OpenRefine is an open-source data wrangling tool (an extended version of Google Refine). When Google discontinued funding the project in 2012, a committed group of international volunteers took up project maintenance and renamed as OpenRefine. It launched with the tagline, "a powerful free, open-source tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data." While it runs like a database in a web browser, OpenRefine has the appearance of a spreadsheet. It offers an alternative to Microsoft Excel for data cleaning and a minimal entry barrier. The application runs on three major operating systems: Windows, Linux, and Mac, and the installation and configuration process are very easy for computers with embedded Java. However, it works best on WebKit browsers, namely Google Chrome, Opera, Microsoft Edge, Chromium, and Safari (Firefox has some performance and rendering issues). In addition, there are different features like fetching data from external databases, clustering of datasets, data reconciliation, and so on (Mukhopadhyay *et al.*, 2021). The latest release of OpenRefine is 3.7.2 (as of April, 2023), with an array of new features.

4. Categorization of ODbL-based Databases

A large array of ODbL-based socio-academic data sources are available in the academic domain. In this study, we cover only sixteen databases under three broad groups, namely Group A: Primary, Group B: Advanced, and Group C: Enhanced (Figure 1), and sub-groups under the basic categories. The categorization of databases is based on the complexity of data acquisition through OpenRefine and

REST/API calls. For example, the primary data includes the bibliographic metadata related to publications like journals, books, conference proceedings, etc. We can extract the advanced data based on primary data and enhance data based on the advanced data.

4.1 Group A: Primary Data

There are four ODbL-based databases available in the bibliographic domain, namely Crossref, Semantic Scholar, Lens.org, and OpenAlex.org (Table 1), that offer metadata of published documents through REST/API calls (Dimensions.ai is omitted because it provides API through subscription).

4.1.1 Crossref

Crossref is a non-profit organization run by PILA (Publishers International Linking Association). It was developed in 2000 with the stated purpose: "to promote the development and cooperative use of new and innovative technologies to speed and facilitate scientific and other scholarly research." The main goal is making scholarly content more easily accessible, discoverable, and reusable through collaboration among publishers and other stakeholders. Crossref provides free API[1] to researchers for access to the data. Anyone can search the database using the query parameters like container-title, author, editor, chair, translator, contributor, bibliographic, and affiliation, which supports different sorting and filtering options.

4.1.2 Semantic Scholar

Semantic Scholar was developed in 2015 as a non-profit and AI-powered research database by the Allen Institute for AI. The main aim is to help researchers to discover and understand the relevant literature more effectively and combat the problem of information overload (Kinney *et al.*, 2023). It has many features, like personalized libraries, auto-generated author pages, paper recommendations, auto summarizations, and so on. The data can be downloaded through open-source libraries and API services. The database has indexed more than 210 million publications from all fields of science (as of March 12, 2023). Semantic Scholar provides free APIs[2] for different purposes, namely (i) Graph API: the most recent data from Academic Graph. Users may obtain publications using internal and external IDs (arXiv, PubMed, DOI) by author search, keyword search, and so on. (ii) Datasets API: It provides monthly snapshots of knowledge graphs in a collection of zipped, JSON-formatted files. The dataset file includes metadata of authors, abstracts, papers, citations, etc. (iii) Recommendations API: It provides recommendations based on positive or negative annotations. (iv) Peer Review API: It helps the process of peer reviewer matching/detection of conflict of interest.

4.1.3 Lens.org

Lens is another non-profit large-scale bibliographic database, previously known as Patent Lens, launched in 2000. It has collected scholarly data from different sources, namely Microsoft Academic, Crossref, ORCID, PubMed, ImpactStory, CORE and patent data from the European Patent Office, United States Patent and Trademark Office, IP Australia, and WIPO. It provides API[3] in two formats, i.e., Scholarly API (for accessing scholarly contents) and Patent API (for accessing patent information), including 60+ searchable fields and enhances the search result with sorting, filtering, and paging. The API service is free on a trial basis for fourteen days with some access limitations. The trial API includes 5K API requests and up to 5 million records/month, 10 API requests/minute, and up to 1K records/request.

4.1.4 OpenAlex

OpenAlex is the youngest among the bibliographic databases, officially launched in 2022. It is the most comprehensive database and has supported the API[4] calls through seven different formats, like works (query by DOI, OpenAlex ID, keywords, etc.), author (query by author name and identifiers), sources (hosting location of the work), Institutions (afflation search), Concepts (based on Wikidata ID), Publishers (from where the work is published), and Geo (geographic location of the work).

Researchers can use these databases for different purposes, such as i. Measuring the productivity of an institution; ii. Development of institutional repository (most digital library software, like DSpace, Omeka are supported by CSV plugins, so that anyone can create an IR by extracting institutional research products and

1 https://github.com/CrossRef/rest-api-doc
2 https://api.semanticscholar.org/api-docs/
3 https://docs.api.lens.org/request-scholar.html
4 https://docs.openalex.org/

Table 1. REST/API Syntax for fetching bibliographic data from ODbL-based databases

Databases	Resultant API call	Note
Crossref	"https://api.crossref.org/works?query=data+deluge&filter=from-pub-date:2022-01-01,until-pub-date:2022-12-31,type:journal-article&select=DOI,title,abstract,container-title,author,score&rows=1000&offset=0"	It will retrieve all published journal articles containing the term "data deluge" during 2022 with metadata elements like DOI, title, abstract, author(s), etc.
Semantic Scholar	"https://api.semanticscholar.org/graph/v1/paper/search?query=data+deluge&year=2022&fields=paperId,externalIds,title,authors,abstract,year,citationCount&publicationTypes=journal+article&limit=100&offset=0"	
Lens.org	"https://api.lens.org/scholarly/search?token=hrkU2EAAP4dkMs7AblSt5DbRAXIXBAplMEGz1coAQHfvkoV0Tyk8&size=1000&query=data+deluge&include=authors,lens_id,title,external_ids,year_published=2022&type=journal&sort=desc(date_published)" N.B: token= <your access token>	
OpenAlex	"https://api.openalex.org/works?search=data+deluge&filter=publication_year:2022,type:journal-article&sort=relevance_score:desc&select=doi,title"	

uploading it into the digital repository); iii. Recommender Services (anyone can make an API-based recommender service to alert users), and iv. Scientometrics/Bibliometrics Studies (researchers can use these databases for any studies related to publication data).

4.2 Group B: Advanced Data

The advanced data is basically based on the primary data. Therefore, we can enrich our primary datasets by fetching different kinds of advanced-level data. This group is categorized under sub-groups, namely citation data, open access data, and altmetric data.

4.2.1 Citation Data

There are many ODbL-based databases that offer citation data freely through the REST/API calls (Table 2), like OpenCitations Corpus (OCC), Dimensions.ai, Semantic Scholar, and Scite.ai.

4.2.1.1 OpenCitations Corpus (OCC)

The official launch of the OCC[5] took place in 2010 at the University of Oxford as a JISC-funded one-year Open Citations Project, which was later extended for an additional six months (Peroni *et al.*, 2017). The main objective was to develop an RDF-based citation database that harvested openly available literature from PubMed Central and was freely available under the Creative Commons (CC0) license. The OCC supports REST/API calls for obtaining citation counts against the DOI.

4.2.1.2 Dimensions.ai

Dimensions.ai is the product of Digital Science, a profit-making organization created in 2018. It is an excellent alternative source of citation data to subscription-based databases like Scopus and WoS. It provides free Metrics API[6] for fetching citation data against article identifiers, like Dimensions ID, PubMed ID, and DOI. It provides four types of data elements, i.e., times_cited (total citations till date), recent_citations (total citations received in the past two years), relative_citation_ratio (the citation rate of an article by comparing to other articles in the same research area), and field_citation_ratio (the relative citation of an article by comparing other similar-aged articles in the same research area).

4.2.1.3 Semantic Scholar

Semantic Scholar is a large bibliographic database (as we discussed earlier). It doesn't support any direct API calls for obtaining citation data because, in every case, it requires a paper ID (like DOI, PubMed ID, URL, MAG ID, and so on.) to obtain any details about the paper and also support citation data as fields search. For example, to obtain the citation data of a specific paper, users can search the paper using any paper ID by adding 'fields=citationCount.'

4.2.1.4 Scite.ai

Scite is a young player among the ODbL-based citation databases, and is still in the developing stage. It uses

5 http://opencitations.net/

6 https://figshare.com/articles/journal_contribution/Dimensions_Metrics_API_Documentation/5783694

Table 2. REST/API Syntax for fetching citation data from ODbL-based databases

Data Sources	REST/API Calls	Data Elements
OpenCitations Corpus	"https://opencitations.net/index/api/v1/citation-count/"+value	count
Dimensions.ai	"https://metrics-api.dimensions.ai/doi/" +value	times_cited
		recent_citations
		relative_citation_ratio
		field_citation_ratio
Semantic Scholar	"https://api.semanticscholar.org/graph/v1/paper/DOI:"+ value+"?fields=citationCount"	citationCount
Scite.ai	"https://api.scite.ai/tallies/" + value	total
		supporting
		contradicting
		mentioning
		unclassified
		citingPublications

Value = DOI

artificial intelligence technology to produce "Smart Citations." It classifies citations according to the citation statement/context and supports API[7] calls free of cost up to 500 requests at a time against the DOI. The data elements include total count (total citations received by the DOI), supporting (the citations that supported the DOI), contradicting (the citations that not supported the DOI), mentioning (only mentioned the DOI by other researchers), unclassified (the citations are not classified by algorithm), and citing Publications (the number of articles which cite the DOI).

4.2.2 Open Access Data

In the open access domain, we discuss only one database.

4.2.2.1 Unpaywall

Unpaywall[8], a project of OurResearch group, officially started in 2017 as a non-profit organization. It is the extended version of the oaDOI.org service. It is the largest storehouse of open science and harvest data, mainly from DOAJ, Crossref, and 50K+ content hosting locations like journals, institutions and disciplinary repositories across the world. The database supports the REST/API (version 2) calls against the DOI endpoints, and it is quite easy, but limited to one lac requests per day. It includes large metadata like generic metadata (journal_is_oa, journal_is_in_doaj, is_oa, oa_status, has_repository_copy, etc.), best_oa_location, first_oa_location, and so on (Table 3).

4.2.3 Altmetric Data

These three databases are freely available in the domain of altmetric data (Table 4).

4.2.3.1 Altmetric.com

Altmetric.com is a large social media aggregator service launched in 2011 as a product of Digital Science. It crawls many news, blog, and social media sites to gather opinions, mentions, and other feedback about academic articles. Altmetric.com provides free API[9] against the article ID for scientometric research in two ways. A researcher can access the API without a license key, but the access rate is limited (1 call/second or 86,400 calls/day). For unlimited access, researchers can apply for a

Table 3. REST/API Syntax for fetching open access data from ODbL-based databases

REST/API Calls	Data Elements
"https://api.unpaywall.org/v2/"+value+"?email=<YOUR MAIL>"	OA Status Best OA location First OA location

Value = DOI

7 https://api.scite.ai/docs#section/

8 https://unpaywall.org/products/api

9 https://api.altmetric.com/

license key by submitting a simple application form[10] (for non-commercial use only). It provides JSON-formatted data as a response that includes the number of mentions in different social media platforms, including the mention links and aggregated weightage-based Attention Score.

4.2.3.2 Crossref Event Data (CED)

CED is the youngest player in the altmetric domain, started in 2016 and officially launched in 2017 with the statement, "we provide the unprocessed data-you decide how to use it." CED provides a public API[11] for accessing social media data freely. Compared to altmetric.com, the CED does not display any metrics, it only provides the social media linked to any DOI.

4.2.3.3 Mendeley

Mendeley is a free referencing and sharing tool accrued by Elsevier in 2014. It provides free API[12] to researchers for obtaining different kinds of information (like catalog, files, folders, groups, identifier_types, institutions, and so on) about the document. But the authentication process is complex, requiring an account-based access token to obtain the data. In this paper, we only focused on readership counts, and it includes readership data in different formats such as reader_count (total reader), reader_count_by_academic_status, reader_count_by_subject_area, and reader_count_by_country.

These advanced databases can be used by researchers for different purposes, like metrics studies, metrics integration in institutional repositories (Indian thesis repository, Sodhganga, already integrated document level metrics), measuring open access status, and social media engagement with publications.

Table 4. REST/API Syntax for fetching altmetric data from ODbL-based databases

Data Sources	REST/API	Data Elements
Altmetric.com	"https://api.altmetric.com/v1/fetch/doi/"+value+"?key=<API KEY>"	Altmetric Attention Score
		Individual Altmetric Events
		Mention Links
Crossref Event Data	"https://api.eventdata.crossref.org/v1/events?mailto=amitnathdlis@gmail.com&obj-id="+value	Altmetric Mention Links
Mendeley.com	"https://api.mendeley.com/catalog?doi="+value+"&view=stats"	Reader Counts

Value = DOI

4.3 Group C: Enhanced Data

This group is again divided into two subgroups. The fetching of enhanced data is based on advanced data; we can use the advanced data elements to enrich our datasets by collecting data from social media platforms, gender databases, and so on.

4.3.1 Social Media Data

Many social media tools offer desired data through REST/API calls. In this section, we introduce only two, i.e., Twitter and YouTube (Table 5).

4.3.1.1 Twitter

Twitter is a microblogging company that offers free API[13] in two formats: Search API (designed for historical data) and Streaming API (designed for real-time data) to researchers for accessing twitter data. You need to create an app using the developer account, but, have some access limitations[14] for using free APIs.

4.3.1.2 YouTube

YouTube is the largest video-sharing platform since its inception in 2005 by Google Inc. It provides free search API[15] to search using different analogous searches (keywords, channel IDs, video IDs, video category, publish date, view counts, and so on), but a license key is required. Anyone can register through a Gmail account to access the license key, but one limitation is that it cannot provide a large amount of data and retrieve only 50 documents at a time.

10 https://forms.monday.com/forms/a472e298bb388b1b7b00b4dd22fb282f?r=use1
11 https://www.crossref.org/documentation/event-data/use/
12 https://api.mendeley.com/apidocs/docs
13 https://developer.twitter.com/en/docs/twitter-api
14 https://developer.twitter.com/en/docs/twitter-api/rate-limits
15 https://developers.google.com/youtube/v3/docs/search

Table 5. REST/API Calls fetching social media data from ODbL-based databases

Data Sources	REST/API Calls	Data Elements
Twitter	"https://api.twitter.com/2/tweets/"+value	Tweet Text
		Tweet Statistics
YouTube	"https://www.googleapis.com/youtube/v3/videos?id="+value+"&key=<YOUR KEY>&part=snippet,contentDetails,statistics,status"	Description
		Content Details
		Status
		Statistics

Value = Tweet ID (Twitter) and Video ID (YouTube) respectively.

Table 6. REST/API Calls fetching gender data from ODbL-based databases

Data Sources	REST/API	Data Elements
Gender API	"https://genderapi.com/get?name="+value+"&country=IN&key=<Your KEY>"	Gender
		Accuracy
		Samples
Genderize.io	"https://api.genderize.io/?name="+value+"&country_id=IN"	Gender
		Probability

Value = Author First Name

4.3.2 Gender Data

Determining gender in authorship is an important aspect in research. There are many services that offer gender data free of cost using API calls, such as Gender API, Genderize.io (Table 6).

5. Case Study I

In this section, we discuss the process of data wrangling in the domain of metrics studies by collecting bibliographic data for a specific institution (i.e., University of Kalyani), and then enrich the datasets by applying data wrangling tools (OpenRefine) (Kusumasari, 2016) and techniques (REST/API calls).

5.1 Objectives and Research Questions

This case study aims to measure a specific university's productivity in terms of publications and social media engagement. Apart from counts, it also explores the data mining techniques from social media platforms (Twitter and YouTube). Based on the objectives described, the following research questions have been developed:

RQ1: What are the socio-academic databases that offer bibliographic metadata freely? Are these datasets accessible through REST/API? If yes, what is the REST/API syntax for gathering data from ODbL-based sources?

RQ2: Can OpenRefine fetch advanced and enhanced level data based on the bibliographic metadata from external databases against REST/API calls? If yes, what is the methodological approach for fetching data from the socio-academic web-space?

5.2 Methodology

The entire methodology part of the Case Study I is divided into a few subsequent parts:

A) Gathering Bibliographic Dataset

There are four ODbL-based databases available in the bibliographic domain. Here, we consulted OpenAlex.org for two reasons. i) It supported institution search through the ROR (Research Organization Registry) ID; ii) It has indexed many leading data sources like Crossref, MAG, PubMed, and so on. Here, we selected the University of Kalyani (https://ror.org/03v783k16) as an example, and entire publications of the said university were fetched from OpenAlex for the year 2022 through REST/API (searched in a web browser). The database returns 485 records (including journals, conferences, book chapters, datasets, etc.) in JSON format. In JSON format, we convert the data into CSV format using an online converter called Konklone[16]. Finally, the entire dataset is uploaded in OpenRefine to enrich datasets from socio-academic web-spaces.

B) Fetching Altmetric Data

There are three altmetric data providers who offer free data against DOI-based REST/API calls (as discussed earlier). Here, we selected altmetric.com for three reasons: i) It is the most exclusive and comprehensive database in the altmetric domain, provides mention counts as well as links; ii) This database also provides readership data from three referencing tools; iii) The dataset includes individual altmetric counts as well as weightage based aggregated altmetric count (called as Altmetric Attention Score). The detailed API syntax is presented in Table 7.

16 https://konklone.io/json/

Table 7. API Response in OpenRefine from altmetric.com

API Syntax	Response in JSON	Parsing GREL
"https://api.altmetric.com/v1/fetch/doi/10.1016/j.jaim.2019.05.002?key=<YOUR KEY>"	"{"altmetric_id":65290839,"counts":{"**readers**":{"citeulike":"0","**mendeley**":"49","connotea":"0"},"total":{"posts_count":23},"**twitter**":{"unique_users_count":13,"**posts_count**":20},"**facebook**":{"unique_users_count":1,"unique_users":["767174926717500"],"**posts_count**":1},"**video**":{"unique_users_count":2,"unique_users":["Dr Nambison","Venkata Chaganti"],"**posts_count**":2}},{"altmetric_id":65290839,"counts":{"readers":{"citeulike":"0","mendeley":"49","connotea":"0"},"total":{"posts_count":23},"twitter":{"unique_users_count":13,"posts_count":20},"facebook":{"unique_users_count":1,"unique_users":["767174926717500"],"posts_count":1},"video":{"unique_users_count":2,"unique_users":["Dr Chaganti"],"posts_count":2}}"**altmetric_score**":{"score":12.249999999999998,"score_history":{"1y":2.85,"6m":0.85,"3m":0.85,"1m":0,"1w":0,"6d":0,"5d":0,"4d":0,"3d":0,"2d":0,"1d":0,"at":12.249999999999998}"	**Reader Counts** value.parseJson().counts.readers.mendeley **Twitter Post** value.parseJson().counts.twitter.posts_count **Facebook Post** value.parseJson().counts.facebook.posts_count **Video Post** value.parseJson().counts.video.posts_count **Altmetric Score** value.parseJson().altmetric_score.score

Table 8. API response from altmetric.com database

API Response	Parsing GREL
"geo":{"twitter":{"US":2,"IN":6}}},"posts":{"twitter":[{"license":"gnip","citation_ids":[65290839],"author":{"tweeter_id":"1141340430514003968"},"tweet_id":"**1201979733321207808**"},{"license":"gnip","citation_ids":[65290839],"author":{"tweeter_id":"2569229502"},"tweet_id":"**1226385290203258880**"},{"license":"gnip","citation_ids":[65290839],"author":{"tweeter_id":"17804715"},"tweet_id":"**1226390838680391681**"},{"license":"gnip","citation_ids":[65290839],"author":{"tweeter_id":"939827120003457024"},"tweet_id":"**1240922712979079169**"},{"license":"gnip","citation_ids":[65290839],"author":{"tweeter_id":"485579201"},"tweet_id":"**1244167474209161221**"}"	forEach(value.parseJson().posts.twitter,v,v.tweet_id).join("|") Then Split Multi-valued cells
""video":[{"title":"Traditional Ayurvedic fumigation for disinfection","url":"https://www.youtube.com/watch?v=Gst9hAzsFWM","image":"https://i.ytimg.com/vi/Gst9hAzsFWM/hqdefault.jpg","license":"public","citation_ids":[65290839],"posted_on":"2020-05-29T16:34:17+00:00","summary":"Traditional Ayurvedic fumigation for disinfection\nEver thought of disinfecting every corner of your home, including roof and other surfaces?\nIn a research publised in www.sciencedirect.com Ayurvedic traditional fumigation with natural plant product found t","embed":{"type":"youtube","youtube_id":"**Gst9hAzsFWM**"},"author":{"name":"Dr Nambison","id_on_source":"UCSGyqGqLzxLVpZ84DSmEJ3Q"}},{"title":"In 1400 AD world wanted to find India. Now Jail and $50,000 fine if they return from INDIA - Shame!","url":"https://www.youtube.com/watch?v=RUOicd6rM9E","image":"https://i.ytimg.com/vi/RUOicd6rM9E/hqdefault.jpg","license":"public","citation_ids":[74763844,2650373,3162597,4078549,65290839],"posted_on":"2021-05-05T00:30:05+00:00","summary":"Even in the pandemic era of closed borders, Australia's latest travel restriction stands out: Anyone, including Australians citizens, who arrives in the country after visiting India in the previous 14 days can face up to five years in jail, a $50,000 fine","embed":{"type":"youtube","youtube_id":"**RUOicd6rM9E**"},"author":{"name":"Venkata Chaganti","id_on_source":"UCxYVMBYeIJF-CtBxTfTU3Cw"}}]}"	forEach(value.parseJson().posts.video,v,v.embed.youtube_id).join("|") Then Split Multi-valued cells

C) Getting Social Media Data through Text Mining

Besides full counts, altmetric.com also provides the mention link where an article is mentioned (Table 8). For example, if an article is twitted five times by five users, the database stores and provides all details of the tweet (including tweet ID and tweeter ID). We can fetch the original tweet text and tweet statistics through REST/API calls from Twitter. Similarly, we can extract the video details, including description, content details, statistics etc. from the YouTube API.

Table 9. Top 20 publications from the University of Kalyani according to attention score

SL	DOI	Attention Score	Facebook	Twitter	Readers
1	10.1093/ofid/ofac603	134.7	1	15	4
2	10.1002/ptr.7461	53.45	1	86	30
3	10.1093/mmy/myac072.p482	36.5	0	59	0
4	10.1155/2022/6873874	10.25	0	2	12
5	10.2174/1570159x20666220927121022	10.25	0	2	1
6	10.1039/d2ob00401a	7.65	0	14	2
7	10.1007/s13224-022-01653-8	7.5	0	5	4
8	10.1016/j.revpalbo.2022.104633	5.7	0	12	7
9	10.1177/02537176221130252	5.65	0	7	0
10	10.1039/d2cc02378d	4.55	0	8	0
11	10.1039/d2ob00109h	4.1	0	6	9
12	10.1016/j.jphotochem.2021.113565	4.05	0	6	5
13	10.1002/widm.1462	4	0	9	43
14	10.1101/2022.06.29.115436	3.85	0	7	0
15	10.1021/acs.iecr.2c02902	3.85	0	8	4
16	10.1007/s12010-022-04155-5	3.5	0	10	4
17	10.1089/ars.2021.0174	2.95	1	3	3
18	10.2147/dmso.s346097	2.75	0	5	33
19	10.1021/acs.jpcb.2c06897	2.6	1	4	1
20	10.2147/jir.s343236	2.5	0	7	6

5.3 Results

Altmetric.com tracked 165 records from the University of Kalyani against the 485 API requests. Finally, we parsed the required fields like attention score, Twitter counts, Facebook counts, reader counts, etc. from the JSON-formatted response data using the GREL functions. Table 9 reveals the top ten publications from the university according to attention score.

Table 10 represents the original tweets extracted through tweet ID from Twitter API (Tweet Lookup) as indexed by altmetric.com (here we presented only five tweets against a DOI). These tweets can be used for sentiment analysis to categorize the text automatically according to sentiment score (positive or negative). We can also extract the profiles of Twitter users (User Lookup).

Similarly, YouTube API provides the qualitative data like content details, status, descriptions, and tags, and quantitative data (statistics) about the videos those are mentioned in a research product (Table 11).

6. Case Study II

Citations are considered as a key element for evaluating research products. Many subscription-based (Scopus and Web of Science) and ODbL-based (such as OpenCitation Corpus, Semantic Scholar, Scite, and Dimensions) data sources are available for collecting citation data. In this section, we make a comparative study among ODbL-based citation and altmetric data sources.

6.1 Objectives

The objective of the Case Study II is to demonstrate a systematic coverage comparison of four citation data sources and three altmetric data provides across a large sample of GSCP (Google Scholar Classic Paper). More specifically, we selected a set of seed papers to make a comparative coverage among the citation databases.

Table 10. Fetching original tweets using Twitter API

Tweet ID	Original Tweets	Refined Tweet Text*
1201979733321207808	{"data":{"edit_history_tweet_ids":["1201979733321207808"],"id":"1201979733321207808","text":"Validation of environmental disinfection efficiency of traditional Ayurvedic fumigation practices. - PubMed - NCBI https://t.co/NLdHPI37tg"}}	Validation of environmental disinfection efficiency of traditional Ayurvedic fumigation practices. - PubMed - NCBI https://t.co/NLdHPI37tg
1226385290203258880	{"data":{"edit_history_tweet_ids":["1226385290203258880"],"id":"1226385290203258880","text":"Fumigation techinque has been tested and proven to kill bacteria.\nhttps://t.co/fXVXwkVoLQ\n~3~"}}	Fumigation techinque has been tested and proven to kill bacteria. https://t.co/fXVXwkVoLQ ~3~
1226390838680391681	{"data":{"edit_history_tweet_ids":["1226390838680391681"],"id":"1226390838680391681","text":"RT @HeritageCitizen: Fumigation techinque has been tested and proven to kill bacteria.\nhttps://t.co/fXVXwkVoLQ\n~3~"}}	RT @HeritageCitizen: Fumigation techinque has been tested and proven to kill bacteria. https://t.co/fXVXwkVoLQ ~3~
1240922712979079169	{"data":{"edit_history_tweet_ids":["1240922712979079169"],"id":"1240922712979079169","text":"@moayush @PMOIndia @shripadynaik @NITIAayog @mygovindia @PIB_India @NHPINDIA @AyushmanNHA @MoHFW_INDIA @AIIA_NDelhi @ccrschennai https://t.co/0pjH6l1mIa\n\nfumigation with garlic peel, turmeric, ajwain seed and loban powder respectively indicating that fumigation with these botanical agents significantly decreased cfu/m3 of air. These Ayurvedic fumigants significantly improved microbiological air quality."}}	@moayush @PMOIndia @shripadynaik @NITIAayog @mygovindia @PIB_India @NHPINDIA @AyushmanNHA @MoHFW_INDIA @AIIA_NDelhi @ccrschennai https://t.co/0pjH6l1mIa fumigation with garlic peel, turmeric, ajwain seed and loban powder respectively indicating that fumigation with these botanical agents significantly decreased cfu/m3 of air. These Ayurvedic fumigants significantly improved microbiological air quality.
1244167474209161221	{"data":{"edit_history_tweet_ids":["1244167474209161221"],"id":"1244167474209161221","text":"Validation of environmental disinfection efficiency of traditional Ayu... https://t.co/LaP3GauOlL"}}	Validation of environmental disinfection efficiency of traditional Ayu... https://t.co/LaP3GauOlL

*Parsing GREL: value.parseJson().data.text

6.2 Methods

6.2.1 Selection of Seed Paper

The primary datasets include the selection of seed paper. We selected the highly cited publications according to Google Scholar, referred to as the Google Scholar Classic Paper[17] (GSCP). The GSCP includes ten highly cited publications for each narrow subject category of Google Scholar. A total of 2515 publications (one discipline, French Studies, indexed only five publications) are indexed by Google Scholar for the year 2006. Anyone can download the datasets from the OSF repository[18]. We imported the datasets into OpenRefine for further data collection process.

6.2.2 Secondary Dataset

The collection of secondary datasets is mainly based on DOI. So, we checked the DOI of the seed publications.

17 https://scholar.google.com/citations?view_op=list_classic_articles&hl=en&by=2006

18 https://osf.io/gnb72/?view_only=

Table 11. Fetching YouTube Data using REST API call.

Video ID	Resultant API Response	Parsing GREL

RUOicd6rM9E	"categoryId": "27",
	"liveBroadcastContent": "none",
	"localized": {
	"title": "Traditional Ayurvedic fumigation for disinfection",
	"description": "Traditional Ayurvedic fumigation for disinfection\nEver thought of disinfecting every corner of your home, including roof and other surfaces?\nIn a research publised in www.sciencedirect.com Ayurvedic traditional fumigation with natural plant product found to disinfect our environment. \nNatural plant products such as \ngarlic (Allium sativum) peel, \nturmeric (Curcuma longa) powder, \nCarom (Trachyspermum ammi) seeds (Ajwain) \nand Loban (resin of Styrax benzoin and Boswellia species). \ncan effectively disinfect our environment all you have to do is to burn these \nnatural products and allow it to reach all the corners specially the surface that cannot be wiped Traditional Ayurvedic fumigation with natural plant products is effective in reducing air-borne bacteria and disinfecting inanimate surfaces. \n\nSo use the natural method to disinfect your environment\nBe safe, be healthy!\n\nLink to the research paper\nhttps://www.sciencedirect.com/science/article/pii/S0975947618306582?via%3Dihub\n#drnambison #disinfection #ayurvedicfumigation"
	}
	},
	"contentDetails": {
	"duration": "PT2M",
	"dimension": "2d",
	"definition": "hd",
	"caption": "false",
	"licensedContent": false,
	"contentRating": {},
	"projection": "rectangular"
	},
	"status": {
	"uploadStatus": "processed",
	"privacyStatus": "public",
	"license": "youtube",
	"embeddable": true,
	"publicStatsViewable": true,
	"madeForKids": false
	},
	"statistics": {
	"viewCount": "221",
	"likeCount": "11",
	"favoriteCount": "0",
	"commentCount": "1"
	}
		Description:
		value.parseJson().items[0].snippet.description
		
		viewCount:
		value.parseJson().items[0].statistics.viewCount
		
		Video Dimension:
		
		value.parseJson().items[0].contentDetails.dimension

Table 12. Coverage of citation data in diverse databases.

Data Sources	Coverage (%)	Total Citations
OCC	2468 (99.64)	229,331
Dimensions	2460 (99.31)	2,450,543
Semantic Scholar	2403 (97.01)	2,819,693
Scite	2465 (99.52)	2,091,228

Table 13. Coverage of altmetric data in diverse sources

Events	Altmetric.com		Crossref Events Data		Mendeley	
	No. of Doc.	%	No. of Doc.	%	No. of Doc.	%
News	825	33.30	229	9.25	--	--
Reddit	467	18.85	150	6.05	--	--
Blogs	871	35.16	178	7.19	--	--
F1000	602	24.30	311	12.56		
Wikipedia	1034	41.74	1027	41.46	--	--
Readers	2205	89.02	--	--	2319	93.62

We found 2477 publications with active DOIs. Then we fetched the citation data as well as altmetric data from diverse sources through the REST/API calls. A JSON-formatted response is received against each request, and the response data is parsed using GREL functions.

6.3 Results

The descriptive statistics of coverage of citation data for GSCP highly cited publications are systematically presented in Table 12. The result indicates the Open Citation Corpus (OCC) has the most extensive coverage with 2468 (99.64%) publications, followed by Scite with 2465 (99.52%) publications, Dimensions with 2460 (99.31%) publications, and Semantic Scholar with 2403 (97.01%) publications. On the other hand, the largest number of citations is received in Semantic Scholar, followed by Dimensions, Scite, and OpenCitation Corpus. Altmetric.com covers the largest percentage of GSCP publications among the altmetric providers (Table 13).

7. Conclusion

Data wrangling is an essential step in the data science workflow, as it ensures that the obtained data is appropriate, accurate, and in consistent format for analysis (Mani *et al.*, 2021). Library professionals are no expectations as they often deal with large amount of data, including bibliographic, circulation, patron, and more. This research work limited is to some ODbL-based data sources. The case studies cover the possible application of data wrangling tool and techniques in the domain of metrics studies and highlight the key functionality of ODbL data sources that anyone can use for their ends. Future work will also include some advanced-level applications of data wrangling from subscription-based databases, data reconciliation, and more.

8. References

Eberendu, A. C. (2016). Unstructured Data: an overview of the data of Big Data. International Journal of Computer Trends and Technology, *38*, 46-50. https://doi.org/10.14445/22312803/IJCTT-V38P109

Kandel, S., Heer, J. and Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. Information Visualization, *10*(4), 271-288. https://doi.org/10.1177/1473871611415994

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., et al. (2023). The Semantic Scholar Open Data Platform. arXiv. Available at: http://arxiv.org/abs/2301.10140

Kusumasari, T. F. (2016). Data profiling for data quality improvement with OpenRefine. In 2016 International Conference on Information Technology Systems and Innovation (ICITSI 2016), 24-27 Oct 2016, Bandung-Bali, Indonesia. p. 1-6. https://doi.org/10.1109/ICITSI.2016.7858197

Mani, N. S., Cawley, M., Henley, A., Triump, 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

Mukhopadhyay, P., Mitra, R. and Mukhopadhyay, M. (2021). Library carpentry: Towards a new professional dimension (Part I–Concepts and case studies). SRELS Journal of Information Management, *58*(2), 67-80. https://doi.org/10.17821/srels/2021/v58i2/159969

Peroni, S., Shotton, D. and Vitali, F. (2017). One year of the OpenCitations Corpus: Releasing RDF-based scholarly citation data into the public domain. In 16th International Semantic Web Conference, 21-25 Oct 2017, Vienna, Austria, edited by C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, and J. Heflin, 2012, p. 184-192. https://doi.org/10.1007/978-3-319-68204-4_19

Robinson, L. and Bawden, D. (2017). "The story of data": A socio-technical approach to education for the data librarian role in the CityLIS library school at City, University of London. Library Management, *38*(6/7), 312-322. https://doi.org/10.1108/LM-01-2017-0009