Reimagining open data ecosystems: a practical approach using AI, CI, and Knowledge Graphs*

Umair Ahmed[1,*,†]

[1]University of Camerino (UNICAM), 9 Via Madonna delle Carceri, Camerino, Italy

Abstract

Open data promotes transparency, facilitates innovation, and enables informed decision-making. In the information age, despite the abundance of data, issues related to findability, accessibility, usability, and value creation continue to be significant challenges. This study focuses on ways to tackle those challenges within the realm of open data ecosystem. It particularly investigates the utilization of AI (Artificial Intelligence) and CI (Collective Intelligence) to enhance the open data ecosystem in the aforementioned aspects. It also navigates through fitting knowledge representation methodologies for open data, which promote semantic reasoning and make it conducive for AI and CI to work more effectively. Given the main objectives of this study, in the preliminary stage, apart from the literature review, we surveyed multiple open data portals to find the state of functional traits currently. We found the state to be significantly lacking in terms of the aforementioned objectives. Initially, we focused on the problem of missing metadata. We explored state-of-the-art AI methodologies such as BERT, YAKE, RAKE, TextRank, ChatGPT and proposed BRYT (a hybrid methodology) for automated metadata extraction. We proposed to extract keywords, themes/categories, and descriptions of data sets using AI to fill in the missing metadata and recommend them to publishers while uploading new data sets. Following metadata extraction, we proposed to explore the idea of constructing a representative knowledge graph from open data sets and investigate how it aids with the objectives of this study. To address this, we chose Open Street Maps as our source of geographical open data and GTFS as a layer of mobility data on top of it. Following it, we propose to employ intuitive search algorithms and recommender systems on top of it to enhance the open data ecosystem. In association with OSM and GTFS, we also plan to focus on the problem of optimal route recommendation, improved navigation, and emergency response planning. The objectives of this study mainly focus on enhancing the open data ecosystem, and by employing the previously stated methods, we intend to advance it and evaluate the impact.

Keywords

Open Data, Artificial Intelligence, Collective Intelligence, Knowledge Graphs, LLM, Metadata Extraction, Search Engine, Open Street Map,

1. Introduction

Open data ecosystems represent the foundation of a transparent and interconnected digital landscape, fostering knowledge-sharing, empowering innovation, and driving societal progress through participation [1] [2]. Although they represent and foresee a well-connected transparent digital environment, the transformation toward them does not come without its fair share of challenges [3]. A massive amount of data, while being open, is not discoverable, accessible,

BIR-WS 2023: BIR 2023 Workshops and Doctoral Consortium, 22nd International Conference on Perspectives in Business Informatics Research (BIR 2023), September 13-15, 2023, Ascoli Piceno, Italy

umairahmedq@gmail.com (U. Ahmed)

usable, value-driven, or conducive to participation.

This study employs Artificial Intelligence (AI), Collective Intelligence (CI), and Knowledge Graphs as methodologies to address key challenges in open data. Our aim is to facilitate a transformation toward an ecosystem distinguished by enhanced data discoverability, improved data accessibility, and optimized data utilization. This transformed ecosystem will focus on value generation, aiming to make open data systems and portals more inclusive, circular, and value-driven. These characteristics of inclusivity, circularity, and value generation are the foundation of any robust open data ecosystem In the preliminary stage of our research, an extensive evaluation of multiple data portals revealed a notable lack of data discoverability. This deficit directly affects the accessibility and overall usability of datasets, thereby limiting their potential for value creation.

The research commences by focusing on the fundamental aspect of findability, namely the provision of comprehensive and accurate metadata that adequately represents the subject matter. In order to tackle this issue, the research conducted an investigation into different metadata attributes that facilitate the process of data discovery, as well as various methods for extracting them automatically. The specific metadata attributes that receive particular attention in this study include keywords, themes/categories, and descriptions. In this study, we suggest utilizing advanced natural language processing (NLP) methodologies such as BERT [4], YAKE [5], RAKE [6][7], TEXTRANK [8], ChatGPT [9] and BRYT (hybrid) to extract the desired information and assess their efficacy.

In addition, we suggest the implementation of an intuitive knowledge representation, specifically knowledge graphs, which facilitate intricate reasoning, intelligent search mechanisms, and efficient recommender systems in order to augment data exploration [10]. Our proposal involves the construction of knowledge graphs utilizing open data, along with the establishment of a semantic relationship framework and contextual comprehension. Subsequently, we aim to develop artificial intelligence solutions based on this foundation. Our primary emphasis is on search engine and recommender systems that effectively facilitate data discovery and utilization without contravening the fundamental principles of open data. We have chosen geographical and mobility open data as our use cases when it comes to constructing knowledge graphs. This study particularly focuses on Open Street Maps (OSM) as our geographical open data source and GTFS as a layer of mobility data on top of it. Once a well-connected knowledge graph using these sources is constructed, we plan to experiment with intelligent routing, navigation, and optimized mobility solutions framework on top of it. In addition to our existing research questions, we also propose the development of a sophisticated reasoning mechanism by leveraging Language Learning Models (LLMs) in conjunction with a highly interconnected knowledge graph.

The overall objective of this Ph.D. project is to enhance the open data ecosystem by leveraging the capabilities of artificial intelligence, computational intelligence, and knowledge graphs. Through the improvement of metadata extraction techniques, the creation of knowledge graphs, and the advancement of intelligent functionalities, we facilitate the more effective utilization of data for researchers, policymakers, and data practitioners. In conclusion, the results of our research make a significant contribution towards enhancing the open data ecosystem, harnessing the potential of open data, and fostering innovation in various fields.

In the following sections, we briefly navigate through the related literature that aids our re-

search questions, the methodology that we adopted or proposed, the results of our methodology or expected results, and a discussion about the way ahead.

2. Related Works

In this section, we examine the existing literature that serves as the foundation for our research and helps derive its key questions. We conducted a literature review regarding Open Data, AI, CI, and Knowledge Graphs, their interconnectivity, and how they aid each other. Following is a brief overview of the literature review conducted regarding our research questions.

2.1. Open Data, AI and CI

In this age of digitization, there is an abundance of data generated every second, which provides stakeholders with distinct assets, namely, data and connected people. It allows them the opportunity to leverage AI and CI use and build solutions on top of it to tackle administrative challenges [11]. It was noted that when considering using CI along with AI, it often comes with concerns of unfairness, biases, lack of transparency, and inefficiencies in the implementations [12]. AI and humans, albeit think differently, largely complement each other in converging towards effective solutions to practical problems [13]. Collective Intelligence in this context refers to multiple software bots or people working collectively or sometimes both [14]. The municipality of Slagelse in Denmark implemented a pilot initiative that utilized crowd-sourcing to enhance citizen engagement in policy-making. Additionally, artificial intelligence (AI) was employed to filter and prioritize relevant inputs, thereby improving the efficiency and effectiveness of the process. However, it is important to acknowledge that this approach was not without challenges, including potential biases and unfairness. [15]. Crowdlaw is another such initiative, which reflects how the government and local governments propose to employ crowd-sourcing and include citizens in the process of establishing policies and laws. [16].

2.2. Open Data and Automated Metadata Extraction

As per our analysis of 15 open data portals, it is evident that the majority of these portals exhibit limited search capabilities and encounter significant challenges in terms of dataset discoverability. The concept of findability encompasses the utilization of search mechanisms and parameters, which are closely intertwined with metadata. However, it is crucial to note that the majority of datasets suffered from a significant deficiency in representative metadata. Consequently, the effectiveness of any search mechanism is severely hindered due to the absence of adequate metadata. The current state of Low-quality metadata poses a serious risk to an effective open data ecosystem [17] [18]. The most representative metadata attributes that aid in findability are keywords, theme/category, and description [19]. If the metadata is not defined properly, it highly affects the discoverability and in effect, the usability of any given resource [20] [21] [22].

When it comes to automated keyword extraction, supervised algorithms, most of the time, are ineffective and require a huge amount of representative domain data to be trained [23]. Unsupervised methods, on the other hand, do not require a huge amount of data; instead, they

make statistical inferences from any given document and extract keywords accordingly [24]. The current optimal approach for addressing such natural language processing challenges using unsupervised approaches is the use of transformers [25]. CERMINE [26] and ROOMBA [27] are two such proposed systems that target the problem of automated metadata extraction and highlight the extensive challenges towards an efficient extraction system. This study focuses on a data set of Brazilian legal documents and theme prediction for it [28]. It encourages the exploration of more techniques on this dataset. Another such study fine-tuned BERT from its base level to modify it, whilst also keeping it light, for the task of document classification into multiple categories [29]. This study briefly discusses textual descriptions and their generation from merely given images. It also proposes to combine LLMs with Reinforcement Learning [30]. Another study discusses how a contemporary LLM ChatGPT fails when it comes to accurately specifying evidence-based medicine. Regardless of its lack of accuracy, the study suggests some ways to avoid pitfalls in such tasks [31].

2.3. Open Data and Knowledge Graphs

A Knowledge Graph is defined as an object for describing entities and their interrelations, by means of a graph [32]. It comes with its own set of challenges, such as entity link prediction problem, which is potentially addressed by machine learning [29]. Knowledge graph faces challenges with entity prediction, relation extraction, and its overall construction, even within the same domain [33]. When it comes to interconnecting graph construction of different domains, it faces even more challenges as a lot of information and vocabulary might not be unified for representation [33]. One of the studies discusses the possible role of knowledge graphs in promoting explainable AI to enhance openness, which eludes the current landscape [34]. Another study constructed a knowledge graph from the perspective of covid using data (24 million publications) from PubMed [35]. The study employed deep learning to identify repurposable drugs from the corpus in regard to covid [35].

This study constructed a rich geographical knowledge graph from Open Street Maps (OSM) dataset [36]. It currently consists of 100 million geographic entities and more than 800 million triples, although it only considers the state of the OSM dataset at the time of construction of the KG [36]. It was noted that when it comes to well-defined methodologies for constructing knowledge graphs, particularly related to health, the literature is quite limited [33].

2.4. Open Data, Search Engines, and Recommendation Systems

Open data is spread over a large number of repositories, sources and catalogs, which necessitate virtual searching mechanisms to discover pertinent data for it to be useful [37]. Search engines related to open dataset search mostly used simplified keywords-based search mechanism currently [37] [38]. This study from Google proposes a dataset search engine that is scalable over a variety of metadata to enhance data discovery[39]. Another study from NYU discusses how Auctus dataset search engine addresses the technical challenges posed by other conventional dataset search engines and how it supports more complex queries for data discovery which are not limited to keywords-based search [37]. Another study makes employs table join and table union search solutions to make search over a million attributes more effective and speedy [40].

Although it is targeted more towards data integration [40]. One such study combines semantics and machine learning to address the problem of data discovery. It focuses on connecting datasets with similar concepts through an ontology for the search to be more intuitive rather than being based on keywords [38].

Apart from search mechanisms, the literature pertinent to recommendation systems in the context of open data portals is near to nonexistent. Although they have seen their fair share of applications in the fields of e-commerce, entertainment, social media, and service personalization [41] [42]. This study explores the application of deep learning in the recommendation of representative news given its dynamic nature [43]. Another such study explore the category of entertainment and navigates through all aspects of the recommendation system employed by Netflix [44]. The most prevalent recommendation system in this context is Amazon's, this study explores its formation and its technical intricacies [45]. Although there has been significant research on recommender systems, their application in open data portals is yet to be explored in more detail.

3. Methodology Conception

Our conceived methodology contains three parts, namely, Automated Metadata Extraction, Knowledge Graph Construction, and Search Engine and Recommender Systems. Our methodology follows a threefold lifecycle. Initially, we focus on metadata, following it we focus on extensive data representation using knowledge graphs, and consequently, we employ AI(Recommender Systems/Search Engines) to improve data findability. This section has three parts covering all stages and workflow of our methodology. In the first section, we describe our methodology regarding metadata (keywords, theme/category, description) extraction. In the second section, we propose to construct knowledge graphs using OSM and GTFS. In the last section, we propose to build a search engine and recommender system on top of it.

3.1. Automated Metadata Extraction

We found three metadata attributes to be the most prominent when it comes to findability, namely, keywords, themes, and descriptions. We used NLP methodologies to extract or generate all of them. For our experimentation, we used the European Data Portal (EDP) as our data source [46]. We used EDP SPARQL endpoints to extract experimentation data from it and applied our methodologies to build intuitive solutions.

3.1.1. Automated Keyword Extraction

Initially, we focused on the most prominent metadata feature when it comes to findability: keywords. Our methodology of automated keyword extraction, as shown in Figure 1, reflects that initially, the methodology contained five steps, namely, Data Gathering, Data Cleaning, Data Transformation, Keyword Extraction, and Matching and Evaluation.

Data Gathering In the phase of data gathering, we gathered pertinent metadata of datasets from each of the 12 listed themes using the SPARQL endpoint that EDP exposed. The metadata

points that we gathered were initially 69423. After we gathered the metadata, we stored it in an SQL database for further processing.

Data Cleaning In the phase of data cleaning, we removed the rows with missing data, duplicate data, garbage data, and nonrepresentative data. It significantly decreased the size of the source data, but nevertheless, it was mostly representative. After data cleaning, the metadata sample was reduced to 1393 representative data points.

Data Transformation In the phase of data transformation, we translated the data into English. Most of our experimentation base was predicated on the English language. Following EDP, we also used the Etranslation API for translation. To access the API, we built two Java API handlers: API Request Handler and API Response Handler. Following the translations, we updated the latest data in our sql database.

Keyword Extraction In the phase of keyword extraction, we employed five contemporary NLP methodologies to extract keywords from descriptions of datasets. The methodologies that we employed were BERT, RAKE, YAKE, TextRank, and ChatGPT. We also proposed a hybrid model, BRYT, which combined the results of the first four methods: B.R.Y.T. ChatGPT was not a part of our hybrid model B.R.Y.T, as the other four complemented each other better than chatgpt. Our proposed model reevaluated and re-scored the results of the four methods using cosine similarity and chose the top 10 results [47].

Matching and Evaluation In the phase of matching and evaluation, we used gestalt pattern matching and jaccard similarity to evaluate the similarity between predicted keywords and original keywords [48] [49]. We then used the similarity measure to evaluate which method extracted more similar or representative keywords.

3.1.2. Automated Theme Extraction

In automated theme extraction, we used the same EDP data that we previously gathered. As reflected in Figure 2, we employed BERT and ChatGPT to predict the theme or category of a dataset based on its title and description. Both of these methods predict multiple categories from the 12 categories predefined by EDP.

3.1.3. Automated Description Generation

In automated description generation, as in Figure 3, we have three steps. In the first step, we choose fifty random rows from any given dataset. Following it, we engineer a prompt for ChatGPT, including the selected 50 rows and prompt text to request that it generates a representative dataset description for it.

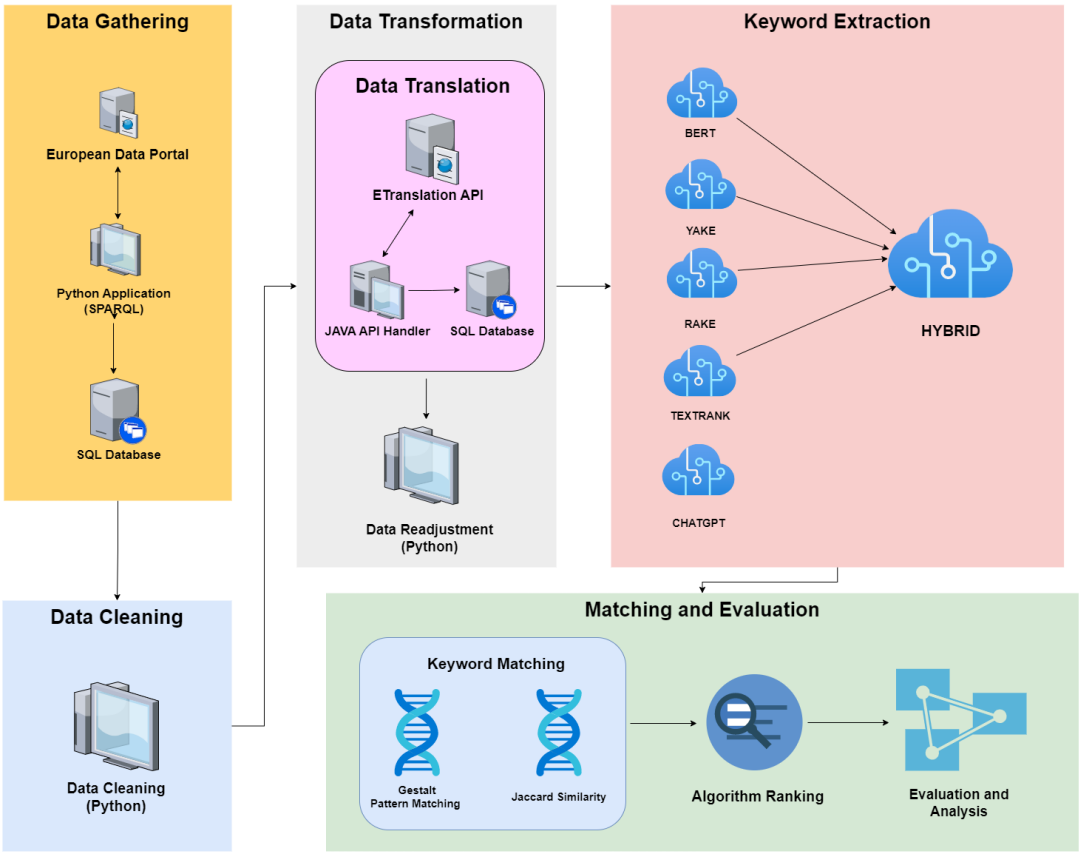

Figure 1: BRYT (Automated Keyword Extraction)

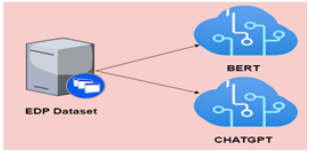

Figure 2: Automated Theme Extraction

3.2. Knowledge Graph Construction

In the stage of knowledge graph construction, we have considered two categories of data that complement each other, namely, geographical data and mobility data. In geographical data, we have proposed to focus on Open Street Maps (OSM), and in mobility data, we have proposed to focus on GTFS, which is a de facto standard. This study proposes to construct a knowledge graph out of both these data repositories and combine to build more intuitive solutions.

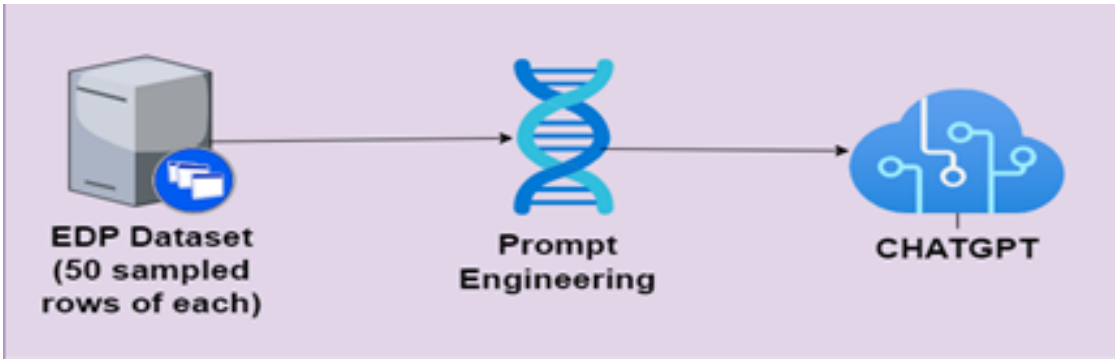

Figure 3: Automated Description Generation

Open Street Map (OSM) Open Street Map is a global geographical database that has most of the geographical entities mapped, and it grows at an increasing pace every passing moment [50]. The data inside it, albeit filled, is not well connected in terms of semantics, which brings about a motivation to connect it inside a knowledge graph.

GTFS Data GTFS is a de facto standard for the mobility industry to publish their data [51]. It is world wide popular, particularly for transportation data to be published. There are various transportation companies that subscribe to it. We proposed to use GTFS data along with OSM to complement each other.

Knowledge Graph Construction In the stage of knowledge graph construction, we proposed to engineer a prompt for LLMs such as ChatGPT to create RDF triples out of OSM and GTFS data. Given the output from it, we plan to construct a knowledge graph from the LLM output and layer it together to be able to build more intuitive solutions from well-connected different data sources inside a knowledge graph.

3.3. Search Engine and Recommender System

In this section, we propose to build an intuitive search engine and recommendation system on top of the constructed knowledge graph.

Search Engine It will allow users to interact with the knowledge graph to get their desired results. It will enable semantic search, geospatial search, faceted search and filtering, and query expansion and similarity using a well-connected knowledge graph.

Recommender System The recommender system will utilize a user's query history, whether personalized or not, to generate informed recommendations. In our specific case, when applied to OSM (OpenStreetMap) and GTFS (General Transit Feed Specification) knowledge graphs, it will recommend places and routes. In case of open data portals, it will suggest relevant datasets.

Additionally, it will take advantage of the intricate semantic relationships embedded in the knowledge graphs to offer context-aware recommendations.

4. Results and Future Directions

In this section, we list the results we acquired, the analysis we conducted, or the results and directions we expect in the future, given our proposed methodologies. In our survey regarding open data portals, our analysis revealed the current state of their functionalities. In Automated Metadata Extraction, we hypothesized and evaluated the performance of our methodologies. While the rest of them are yet to be experimented with extensively and evaluated.

4.1. Analysis of Open Data Portals

We surveyed 15 different open data portals and found there to be a significant lack of metadata, basic keyword search mechanisms, and non existent recommender system. We have listed these portals in Table 1. The search mechanisms that all of these portals had were basic word-to-word matching and search, which did not adhere to the context of the query. Some of them had autocomplete functionality with the search query, but it was matching the exact dataset name. Apart from that, we found there to be no recommendation system of any kind, although there sure was a tab of related datasets that were based on category or basic similarity of the same author/organization. This motivated us to work toward data discovery, findability and usability. The preliminary step for working toward it was exploring ways to improve metadata.

Table 1
Data Portals Technical Analysis

Data Portal	Region	Search Engine	Autocomplete	Recommendations	Multilanguage
https://data.europa.eu/en	Europe	Basic	No	No	Yes
https://www.dati.piemonte.it/	Italy	Basic	Yes	No	No
https://www.dati.gov.it/	Italy	Basic	No	No	No
https://www.data.gouv.fr/en/	France	Basic	No	No	Yes
https://opendata.paris.fr/	France	Basic	No	No	No
https://data.overheid.nl/	Netherlands	Basic	Yes	No	No
https://www.data.gov.uk/	UK	Basic	No	No	No
https://data.gov/	US	Basic	No	No	No
https://data.gov.ie/	Ireland	Basic	No	No	No
https://africaopendata.org/	Africa	Basic	No	No	Yes
https://dataportal.asia/home	Asia	Basic	No	No	Yes
http://data.un.org/	UN World	Basic	No	No	No
https://data.nasa.gov/	US	Basic	Yes	No	No
https://data.gov.be/	Belgium	Basic	Yes	No	Yes
https://www.data.gv.at/	Austria	Basic	No	No	Yes

4.2. Evaluation of Automated Metadata Extraction

In automated metadata extraction, we mostly employed LLMs, but we proposed to have an appropriate evaluation which was fitting to the situation.

4.2.1. Automated Keyword Extraction

During this stage, the performance of our algorithms is assessed using the following two metrics:

- Gestalt matching: The Ratcliff/Obershelp algorithm employs the concept of gestalt pattern matching, which identifies patterns as distinct functional units with their own unique characteristics. It measures the similarity between two one-dimensional patterns like text strings by tallying the shared sub-patterns between them. [49]:

$$Similarity = \frac{2k_m}{\|S_1\| + \|S_2\|} \tag{1}$$

- Jaccard similarity: The Jaccard similarity quantifies the similarity between two sets by determining the proportion of their intersecting elements relative to the total elements found in their union [48].

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

In order to evaluate the resemblance between the extracted keywords and the original keywords, we use these metrics to compute a similarity score for each method, comparing the extracted keywords to the original ones. Keywords with a similarity score equal to or above 0.6 for each measure are classified as matches. We used 0.6 as the threshold to ensure the matching of more than half of the keyword even if it has additives towards the end or start. The evaluation of algorithm performance in each example is determined by assessing the number of matches using Gestalt matching and Jaccard similarity techniques. The matches were classified into three distinct categories: major match, minor match, and no match. We classified them as such to streamline the process of interpreting results and deciding the relevance of individual keywords and whether to consider them. The categories in question are delineated as follows:

- Major Match: The result was classified as a major match under that algorithm if, when applying a certain algorithm, more than 50% or five keywords matched in any given instance.
- Minor Match: A result was classified as a minor match if, when applying a certain algorithm, less than 50% or 5 keywords matched in any given instance.
- No Match: It was classified as a no match if none of the keywords met the predetermined criteria.

Using this criteria, according to our overall findings (Fig. 4), 69.1% of cases were major matches utilizing any given method, whereas 24.7% were minor matches and 6.2% had no matches at all. Considering major matches, we observed that our suggested hybrid approach BRYT outperformed other algorithms when comparing the matches of each algorithm, with

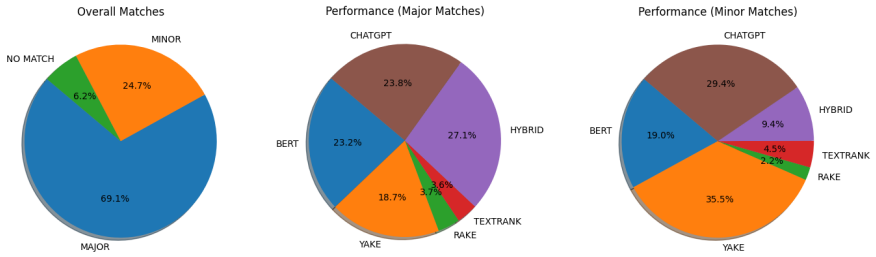

Figure 4: Keyword extraction evaluation

27.1% of times having more efficient or equivalent efficiency matches, whereas CHATGPT was a close second with 23.8%. This percentage also implies that there were cases in which two algorithms were equal matches, and in such cases, it considered both methods as efficient/winning algorithms in that specific situation.

When analyzing the minor matches produced by each algorithm, we observed that YAKE performed well (35.5%) whereas CHATGPT came in second (29.4%). Because major matches made up 69.1% of the total data, we concentrated our research more on them because our suggested technique performed better in those cases, making it seem like a more appropriate assessment for this issue and in the unique context of open data. We also plan to employ human evaluation for keyword extraction by conducting a survey containing open data portal users.

4.2.2. Automated Theme Extraction

In automated theme extraction, we predicted themes for each dataset given its title and description. For its evaluation, we propose to compare the predicted and original themes and evaluate the similarity between the both. Moreover, we propose to design human evaluation for this task to evaluate its functional effectiveness.

4.2.3. Automated Description Generation

In automated description generation, we engineered a prompt and employed chatgpt to access a sample of the dataset and generate a representative description for the dataset. For its evaluation, we propose to involve domain-aware humans to be able to rank the efficiency of description generation. We propose to design the evaluation criteria for the sample group to estimate the efficiency of chatgpt in any given instance.

4.3. Knowledge Graph Construction

During the stage of knowledge graph construction, we plan to use two open data sources name OSM and GTFS. Also, we plan to use LLM (ChatGPT) to aid in triple creation for knowledge graph construction. As an outcome, we expect a well-connected knowledge graph constructed on OSM and GTFS, which is conducive to intuitive solutions pertaining to geographical or mobility challenges.

4.4. Search Engine and Recommender System

In the phase of the Search Engine and Recommender System, we expect an intuitive search engine that is context-aware and has access to the semantic connections of the knowledge graph to be able to query efficiently in aspects such as semantic, geospatial, faceted, or similarity search. Moreover, we expect to have a prototype of a recommender system that utilizes the semantic connects of well-connected knowledge graphs and make intelligent, personalized recommendations.

5. Discussion and Conclusion

In conclusion, this research focuses on improving open data ecosystems by using Knowledge Graphs, CI (Collective Intelligence), and AI (Artificial Intelligence). The main goal is to deal with the difficulties in finding, making accessible, making use of, and creating value from open data. The paper suggests methods for automated information extraction and notes the dearth of representative metadata as a serious problem.

The paper suggests BRYT, a hybrid approach that incorporates several NLP (Natural Language Processing) techniques, such as BERT, RAKE, YAKE, TextRank, and ChatGPT, to extract keywords, themes/categories, and descriptions of datasets using automated metadata extraction. When compared to other approaches, BRYT greatly outperforms them in terms of major matches.

The research suggests creating a knowledge graph using data from Open Street Maps (OSM) and GTFS (General Transit Feed Specification), in addition to extracting information. This knowledge graph attempts to improve recommender systems, intelligent search engines, and data exploration. The knowledge graph will provide more logical answers for the best route suggestions, enhanced navigation, and emergency response planning by linking geographical and mobility data.

The research's overarching objective is to improve the open data ecosystem by using AI, CI, and Knowledge Graphs. Researchers, politicians, and data practitioners will be able to use data more effectively, as a result encouraging innovation across a variety of industries and creating a more connected and transparent digital environment.

Declarations

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955569

References

[1] E. Ruijer, S. Grimmelikhuijsen, J. Van Den Berg, A. Meijer, Open data work: understanding open data usage from a practice lens, International Review of Administrative Sciences 86 (2020) 3–19.

[2] K. Braunschweig, J. Eberius, M. Thiele, W. Lehner, The state of open data, Limits of current open data platforms 1 (2012) 72–72.
[3] B. Van Loenen, A. Zuiderwijk, G. Vancau-Wenberghe, F. J. Lopez-Pellicer, I. Mulder, C. Alexopoulos, R. Magnussen, M. Saddiqa, M. D. De Rosnay, J. Crompvoets, et al., Towards value-creating and sustainable open data ecosystems: A comparative case study and a research agenda, JeDEM-eJournal of eDemocracy and Open Government 13 (2021) 1–27.
[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[5] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, Yake! keyword extraction from single documents using multiple local features, Information Sciences 509 (2020) 257–289.
[6] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic keyword extraction from individual documents, Text mining: applications and theory (2010) 1–20.
[7] G. Campobello, A. Segreto, S. Zanafi, S. Serrano, Rake: A simple and efficient lossless compression algorithm for the internet of things, in: 2017 25th European Signal Processing Conference (EUSIPCO), IEEE, 2017, pp. 2581–2585.
[8] R. Mihalcea, P. Tarau, Textrank: Bringing order into text, in: Proceedings of the 2004 conference on empirical methods in natural language processing, 2004, pp. 404–411.
[9] M. Song, H. Jiang, S. Shi, S. Yao, S. Lu, Y. Feng, H. Liu, L. Jing, Is chatgpt a good keyphrase generator? a preliminary study, arXiv preprint arXiv:2303.13001 (2023).
[10] M. Färber, D. Lamprecht, The data set knowledge graph: Creating a linked open data source for data sets, Quantitative Science Studies 2 (2021) 1324–1355.
[11] S. G. Verhulst, Where and when ai and ci meet: exploring the intersection of artificial and collective intelligence towards the goal of innovating how we govern, AI & society 33 (2018) 293–297.
[12] J. Danaher, M. J. Hogan, C. Noone, R. Kennedy, A. Behan, A. De Paor, H. Felzmann, M. Haklay, S.-M. Khoo, J. Morison, et al., Algorithmic governance: Developing a research agenda through the power of collective intelligence, Big data & society 4 (2017) 2053951717726554.
[13] M. M. Peeters, J. van Diggelen, K. Van Den Bosch, A. Bronkhorst, M. A. Neerincx, J. M. Schraagen, S. Raaijmakers, Hybrid collective intelligence in a human–ai society, AI & society 36 (2021) 217–238.
[14] D. S. Weld, C. H. Lin, J. Bragg, Artificial intelligence and collective intelligence, Handbook of collective intelligence (2015) 89–114.
[15] C. A. Hallin, N. Lipka, Reinventing local government through collective intelligence and artificial intelligence, The Routledge Handbook of Collective Intelligence for Democracy and Governance (2023).
[16] B. S. Noveck, Crowdlaw: Collective intelligence and lawmaking, Analyse & Kritik 40 (2018) 359–380.
[17] J. Umbrich, S. Neumaier, A. Polleres, Quality assessment and evolution of open data portals, in: 2015 3rd international conference on future internet of things and cloud, IEEE, 2015, pp. 404–411.
[18] R. Máchová, M. Lněnička, Evaluating the quality of open data portals on the national level, Journal of theoretical and applied electronic commerce research 12 (2017) 21–41.
[19] P. Škoda, D. Bernhauer, M. Nečaskỳ, J. Klímek, T. Skopal, Evaluation framework for search

methods focused on dataset findability in open data catalogs, in: Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services, 2020, pp. 200–209.
[20] J. Nogueras-Iso, J. Lacasta, M. A. Ureña-Cámara, F. J. Ariza-López, Quality of metadata in open data portals, IEEE Access 9 (2021) 60364–60382.
[21] S. Neumaier, J. Umbrich, A. Polleres, Automated quality assessment of metadata across open data portals, Journal of Data and Information Quality (JDIQ) 8 (2016) 1–29.
[22] S. Kubler, J. Robert, S. Neumaier, J. Umbrich, Y. Le Traon, Comparison of metadata quality in open data portals using the analytic hierarchy process, Government Information Quarterly 35 (2018) 13–29.
[23] S. Beliga, Keyword extraction: a review of methods and approaches, University of Rijeka, Department of Informatics, Rijeka 1 (2014).
[24] J. Baruni, J. Sathiaseelan, Keyphrase extraction from document using rake and textrank algorithms, Int. J. Comput. Sci. Mob. Comput 9 (2020) 83–93.
[25] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, AI Open (2022).
[26] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, Ł. Bolikowski, Cermine: automatic extraction of structured metadata from scientific literature, International Journal on Document Analysis and Recognition (IJDAR) 18 (2015) 317–335.
[27] A. Assaf, A. Senart, R. Troncy, Roomba: Automatic validation, correction and generation of dataset metadata, in: Proceedings of the 24th international conference on World Wide Web, 2015, pp. 159–162.
[28] P. H. L. De Araujo, T. E. de Campos, F. A. Braz, N. C. da Silva, Victor: a dataset for brazilian legal documents classification, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 1449–1458.
[29] A. Adhikari, A. Ram, R. Tang, J. Lin, Docbert: Bert for document classification, arXiv preprint arXiv:1904.08398 (2019).
[30] Y. Ge, W. Hua, J. Ji, J. Tan, S. Xu, Y. Zhang, Openagi: When llm meets domain experts, arXiv preprint arXiv:2304.04370 (2023).
[31] Y. Peng, J. F. Rousseau, E. H. Shortliffe, C. Weng, Ai-generated text may have a role in evidence-based medicine, Nature Medicine (2023) 1–2.
[32] P. A. Bonatti, S. Decker, A. Polleres, V. Presutti, Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371), in: Dagstuhl reports, volume 8, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.
[33] B. Abu-Salih, Domain-specific knowledge graphs: A survey, Journal of Network and Computer Applications 185 (2021) 103076.
[34] F. Lecue, On the role of knowledge graphs in explainable ai, Semantic Web 11 (2020) 41–51.
[35] X. Zeng, X. Song, T. Ma, X. Pan, Y. Zhou, Y. Hou, Z. Zhang, K. Li, G. Karypis, F. Cheng, Repurpose open data to discover therapeutics for covid-19 using deep learning, Journal of proteome research 19 (2020) 4624–4636.
[36] A. Dsouza, N. Tempelmeier, R. Yu, S. Gottschalk, E. Demidova, Worldkg: A world-scale geographic knowledge graph, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 4475–4484.
[37] S. Castelo, R. Rampin, A. Santos, A. Bessa, F. Chirigati, J. Freire, Auctus: A dataset search

engine for data discovery and augmentation, Proceedings of the VLDB Endowment 14 (2021) 2791–2794.
[38] S. Jiang, T. F. Hagelien, M. Natvig, J. Li, Ontology-based semantic search for open government data, in: 2019 ieee 13th international conference on semantic computing (icsc), IEEE, 2019, pp. 7–15.
[39] D. Brickley, M. Burgess, N. Noy, Google dataset search: Building a search engine for datasets in an open web ecosystem, in: The World Wide Web Conference, 2019, pp. 1365–1375.
[40] R. J. Miller, Open data integration, Proceedings of the VLDB Endowment 11 (2018) 2130–2139.
[41] J. Lu, D. Wu, M. Mao, W. Wang, G. Zhang, Recommender system application developments: a survey, Decision support systems 74 (2015) 12–32.
[42] F. Ricci, L. Rokach, B. Shapira, Recommender systems: introduction and challenges, Recommender systems handbook (2015) 1–34.
[43] G. Zheng, F. Zhang, Z. Zheng, Y. Xiang, N. J. Yuan, X. Xie, Z. Li, Drn: A deep reinforcement learning framework for news recommendation, in: Proceedings of the 2018 world wide web conference, 2018, pp. 167–176.
[44] C. A. Gomez-Uribe, N. Hunt, The netflix recommender system: Algorithms, business value, and innovation, ACM Transactions on Management Information Systems (TMIS) 6 (2015) 1–19.
[45] B. Smith, G. Linden, Two decades of recommender systems at amazon. com, Ieee internet computing 21 (2017) 12–18.
[46] European data portal, https://data.europa.eu/en, ???? Accessed: 2023-07-31.
[47] A. R. Lahitani, A. E. Permanasari, N. A. Setiawan, Cosine similarity to determine similarity measure: Study case in online essay assessment, in: 2016 4th International Conference on Cyber and IT Service Management, IEEE, 2016, pp. 1–6.
[48] S. Niwattanakul, J. Singthongchai, E. Naenudorn, S. Wanapu, Using of jaccard coefficient for keywords similarity, in: Proceedings of the international multiconference of engineers and computer scientists, volume 1, 2013, pp. 380–384.
[49] J. W. Ratcliff, D. Metzener, et al., Pattern matching: The gestalt approach, Dr. Dobb's Journal 13 (1988) 46.
[50] J. Bennett, OpenStreetMap, Packt Publishing Ltd, 2010.
[51] A. Antrim, S. J. Barbeau, et al., The many uses of gtfs data–opening the door to transit and multimodal applications, Location-Aware Information Systems Laboratory at the University of South Florida 4 (2013).