SDG Aligned Datalake Framework over Open Data

Apurva Kulkarni, Ph.D. Student, International Institute of Information Technology, Bangalore (corresponding author)
apurva.kulkarni@iiitb.ac.in
26/C Hosur Road, Electronic City Phase 1,
Bangalore, Karnataka, 560100
Chandrashekar Ramanathan, Professor, International Institute of Information Technology, Bangalore
Vinu E Venugopal, Professor, International Institute of Information Technology, Bangalore

Abstract

Open-Data constitutes data published by governmental bodies, including statistics, budgets, legislation, public services, and geospatial data (maps, satellite imagery). Domain users, including individuals, organizations, and governments, can leverage open data for decision-making purposes. Open data can play a vital role in advancing the SDGs by enabling evidence-based decision-making, facilitating collaboration, and measuring progress. While open data promotes the availability and accessibility of a wide range of data, it may not always be necessary or practical to work with the entire dataset. Identifying appropriate subset of data can significantly enhance decision-making by providing relevant and meaningful information while reducing complexity and noise. In our study, we propose a datalake framework for open data that assists domain users, policymakers, domain specialists, and government entities in identifying relevant documents. The relevant documents used for decision-making help to improve focus, minimize computational complexity, and make better use of limited resources. Given the progress made in AI applications, policymakers around the globe have a keen interest in utilizing AI models for policy-making. The suggested framework allows AI models to utilize pertinent data during their training process, thereby enhancing the credibility of their outcomes. Let's consider an instance where policymakers are aiming to create a policy regarding the use of fertilizers. To prevent a prolonged food crisis, policymakers or an AI model should consider various factors such as crop production, import-export activities, and land quality while planning. However, it is difficult to identify specific factors or location-specific documents from the vast amount of open data available. To address this challenge, researchers at IIIT Bangalore (Karnataka, India) conducted a study to develop a framework that can host and analyze domain-specific semantic data with location awareness. This proposed framework serves as a foundation for decision-makers, data analysts, policymakers, subject experts, domain users and AI applications to identify relevant datasets and contribute to reliable decision-making. The research proposes a solution for a document retrieval framework based on Sustainable Development Goals (SDGs), which utilizes domain knowledge to bridge the gap between user queries and search results. Although the study focuses on the SDG-2 domain and data from the Government of Karnataka (India), the findings have broader applicability. The researchers examined 170 carefully selected input queries that were specifically chosen by domain experts to reflect the interests of policymakers. These queries consist of combinations of search terms, location tags, and semantic interests. Three baseline systems were used to evaluate the performance of these queries, and the results were verified by domain experts to assess the effectiveness of the system. The experimental findings demonstrate promising improvements compared to the baseline systems with promising results.

Keywords: Heterogeneous Datalake, Open Data, SDG aligned Framework

1. Introduction

Achieving the Sustainable Development Goals necessitates the adoption of sustainable policies, which entails a crucial task of policy formulation. Policymakers must consider multiple factors, such as the present development status, which can be assessed using diverse data points, as well as the policy's impact and an action plan outlining its implementation [4]. These data points typically originate from various sources, primarily governmental bodies encompassing statistics, budgets, legislation, public services, and geospatial data (maps, satellite imagery), along with domain users comprising individuals, organizations, and governments [5].

The primary component of the data utilized in the decision-making process is open data, which significantly contributes to the process. With technological advancements, AI tools are leveraged to facilitate decision-making, aiding policymakers in formulating improved policies [6]. The typical process of employing AI tools involves training them on data, where the model learns to establish a correlation between input features and their corresponding output labels. This process entails adjusting the internal parameters (weights and biases) of the model using an optimization algorithm, such as gradient descent, to minimize the discrepancy between predicted and actual outputs. It is crucial to recognize that the accuracy of the model depends on the quality and relevance of the data employed during training.

The utilization of open data for training AI models can give rise to potential issues that may result in misleading results. These issues include:

- **Generic Model:** When a large amount of data is used, it can sometimes provide a generalized understanding of the information, causing the model to overlook specific and detailed characteristics present in the data.
- **Validity:** Since open data is sourced from various providers, it becomes challenging to ascertain the reliability and validity of the data. The accuracy and trustworthiness of the data source can significantly impact the quality of the results.
- **Completeness:** The data used for training may be incomplete at the time of training, and it may not account for new information or emerging patterns that can arise in the future. This limitation can result in the model missing out on important insights or making inaccurate predictions.
- **Relevance:** Identifying the relevant subset of data that truly contributes to enhancing the model's training can be a complex task. Determining which data points are significant for the given problem at hand can be challenging, potentially leading to sub-optimal training outcomes.

Addressing these issues requires careful consideration of data quality, source validation, continuous data updates, and thoughtful selection of relevant data subsets for training. In the proposed research, the focus is on retrieving relevant documents based on domain information and semantics. Further sections discuss the architecture and implementation details of the proposed approach focusing on SDG2 [8].

2. Proposed Approach

Figure 1 depicts the architecture of the proposed approach. The system has mainly two layers- Data Ingestion Layer and Data Access Layer. The data Ingestion Layer is responsible for accommodating heterogeneous data sources and aligning them according to SDGs. The Data Access Layer interacts with AI Tools and provides a relevant subset of data from a huge pile of Open Data. Conventionally, there are two ways to handle the heterogeneous data - either by integrating them into a common format or by federation where an orchestration mechanism facilitates seamless access to heterogeneous distributed data [7].

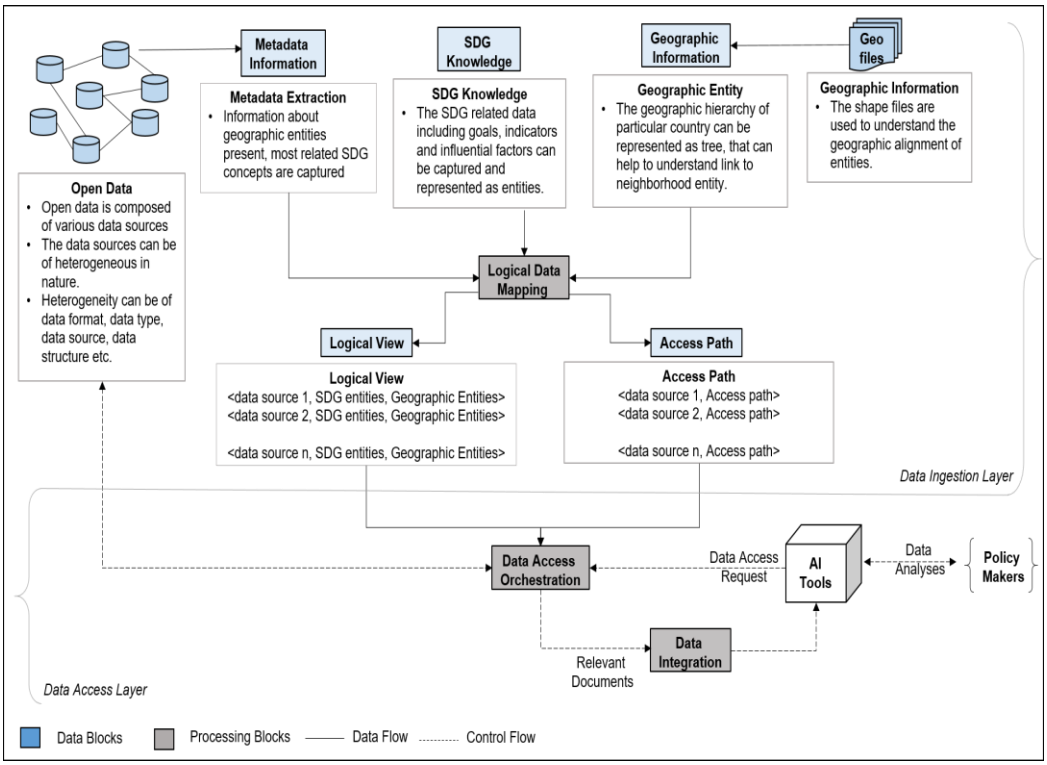
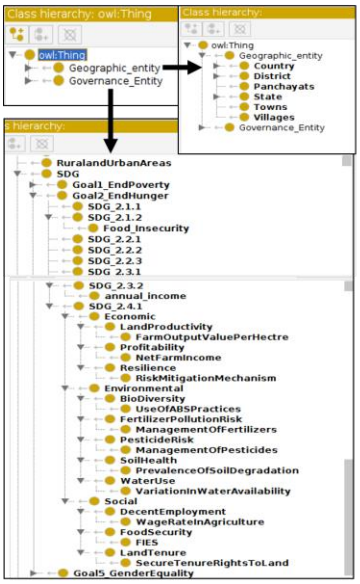


Figure 1. SDG aligned Datalake Framework for Open Data

In this work, we focus on the latter approach where the data sources are intact in their original formats without having the overhead of conversion and transformation. Individual data sources are considered for metadata extraction. The metadata gives additional information about the data sources depending on their type. All structured-tabular data sources (.XLS, .CSV, .SQL) are treated to extract information about attributes, unique values of the attribute, type of attributes and description of the file. On the other hand, unstructured data sources (.PDF, .DOC, .TXT, .JPEG/.PNG) generate metadata about the textual description of the file and frequently occurring words in the file.



a. Domain information in OWL format

subject	object	predicate
agriculture	agricultural area or land use	has_subclass
agriculture	agricultural cooperatives	has_subclass
agriculture	agricultural education	has_subclass
agriculture	agricultural export	has_subclass
agriculture	agricultural implements and machinery	has_subclass
agriculture	agricultural indices	has_subclass
agriculture	agricultural input output	has_subclass
agriculture	agricultural institutional finance	has_subclass
agriculture	agricultural insurance	has_subclass
agriculture	agricultural land holding	has_subclass
agriculture	agricultural marketing	has_subclass
agriculture	agricultural offtake allocation and procurement	has_subclass
agriculture	agricultural prices	has_subclass
agriculture	agricultural production	has_subclass
agriculture	agricultural schemes	has_subclass
agriculture	agricultural stocks	has_subclass
agriculture	agricultural subsidy	has_subclass
agriculture	agricultural wages	has_subclass
agriculture	agricultural workforce	has_subclass
agriculture	animal husbandary and livestock	has_subclass
agriculture	cold storage	has_subclass
agriculture	consuption of fertilisers	has_subclass

b. Domain information as triplets in CSV format

Figure 2. Domain Information captured as an OWL file or CSV file formats

Another input to the system is domain information. In this work, we are considering domain information focusing on SDG2. The information about the agriculture domain is gathered from domain experts. The hierarchical structure of SDGs allows for capturing relationships among goals, concepts, and indicators. Each unique term in the domain information is represented as an entity tagged with its description. The policy decisions are often variable depending on the geographic condition. The value of indicators impacting goals varies from location to location. To understand the data points with respect to location, it is crucial to consider data of that location or neighboring locations (if data on the required location is not available). To enable data access depending on location, we propose to use geographic information to align the data sources. The shape files are considered to identify the geographic boundaries of an entity like Villages, Talukas[1] and States. The geographic information helps to identify neighboring entities. Figure 2 elaborates the domain information and geographic information employed for the proposed work. The Part (a) in Figure 2 shows the domain information in an OWL format and Part (b) highlights the triplet form of domain information capturing relations between two SDG entities (subject and object) as predicate in CSV file format. The proposed architecture supports both the formats to plug-in the domain information.

The metadata information, domain information and geographic information are forwarded to Logical Mapping Block. This block processes all the information and maps each file with the relevant SDG entities and geographic entities. The mapping generates a logical view that facilitates data access via SDGs and geographic location. To retrieve the data files across data sources, the complete path is obtained from the Access Path file. The primary goal of a Data Ingestion Layer is to provide seamless access to disparate data sources to retrieve relevant files.

The Data Access Layer is responsible for interacting with data sources depending on the requests from AI Tool. The Policy Maker can interact with AI Tool for analysis. Depending on the SDGs entities and location of policy implementation, AI Tool requests a Data Access Orchestration block. Based on the SDG entities and geographic entities

[1] The generic geographic hierarchy is Country-State-County-Village, here Taluk can be considered equivalent to County

relevant data files are selected from the Mapping File. The complete access path to retrieve these files over a heterogeneous data environment is obtained using Access Path File. The Data Access Orchestration block uses the access path to retrieve the files. The set of relevant files is given to the Data Integration Block. The Data Integration Block links files together and forwards them to AI Tool for further analysis.

3. Implementation of the proposed approach

The proposed approach is implemented on Agriculture data from the Government of Karnataka, India. The system is evaluated against 170 data requests for identifying relevant documents. The search requests cover the search interests of domain users, domain experts, and some requests are generated using AI Tools. Compared to three baseline systems Lucene [1], Doc2Vec [2], and ElasticSearch [3], the proposed approach retrieves more relevant results [5]. There are mainly two categories of users for the system. The AI Tools can directly interact with the system to retrieve relevant data and another set of users can be domain users or policymakers. The users can download a subset of data and can perform some exploratory analytics on relevant data using the user interface.

Figure 3 depicts the interface to retrieve the documents relevant to the search interest. The relevant files can be selected for further processing or downloading. The system provides metadata information about the file and sample data view. The metadata information helps the user to understand a file to decide on selection.

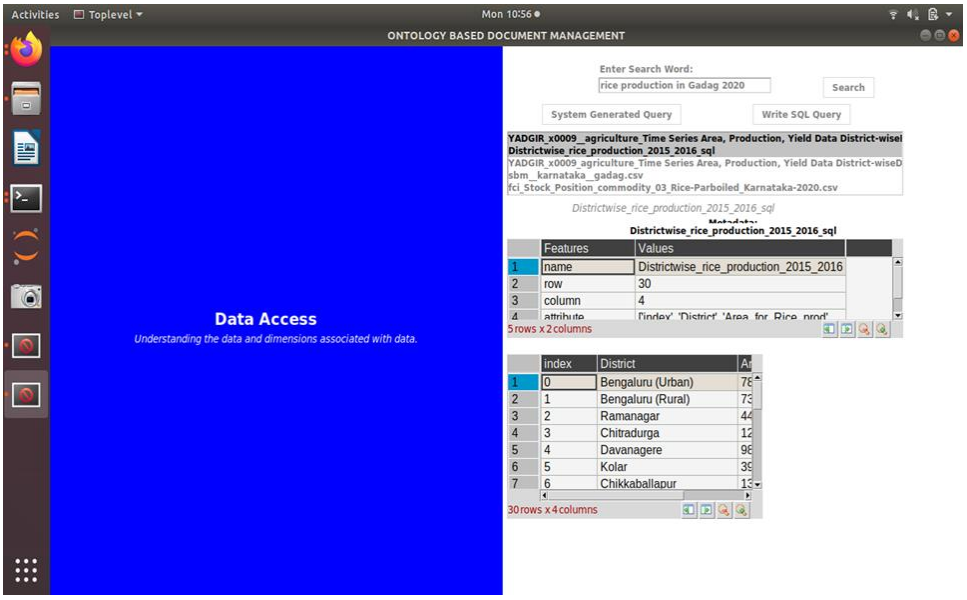


Figure 3. User Interface to retrieve relevant data using SDG aligned Datalake Framework for Open Data

The selected files can be considered for further analysis where the user can query data and explore the results. Figure 4 showcases the query interface where the user can write SQL queries over files and get the results.

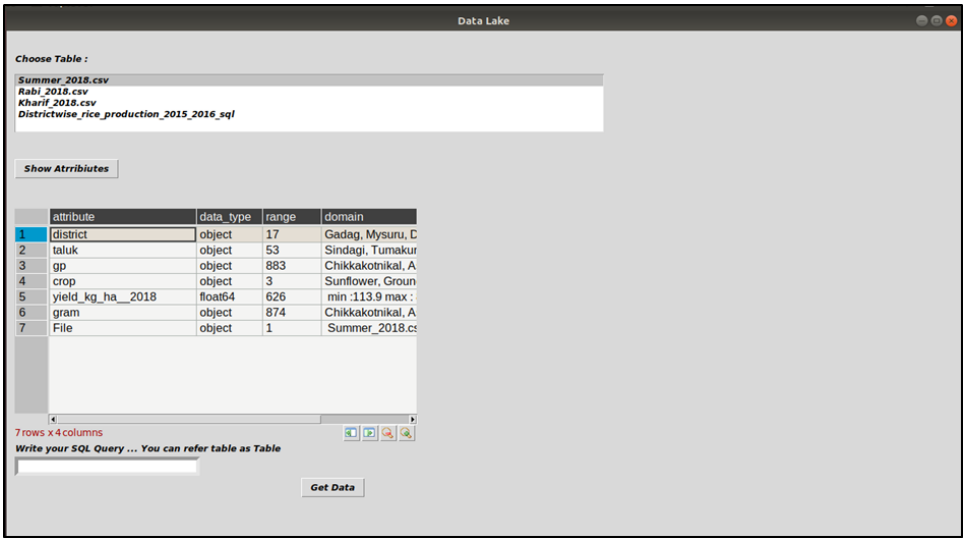


Figure 4. User Interface to query relevant data

The generated results are highlighted on a map showcasing the geographic entities in the result set. The interface also allows the User to select a subset of results and visualize the graphical representation as shown in Figure 5.

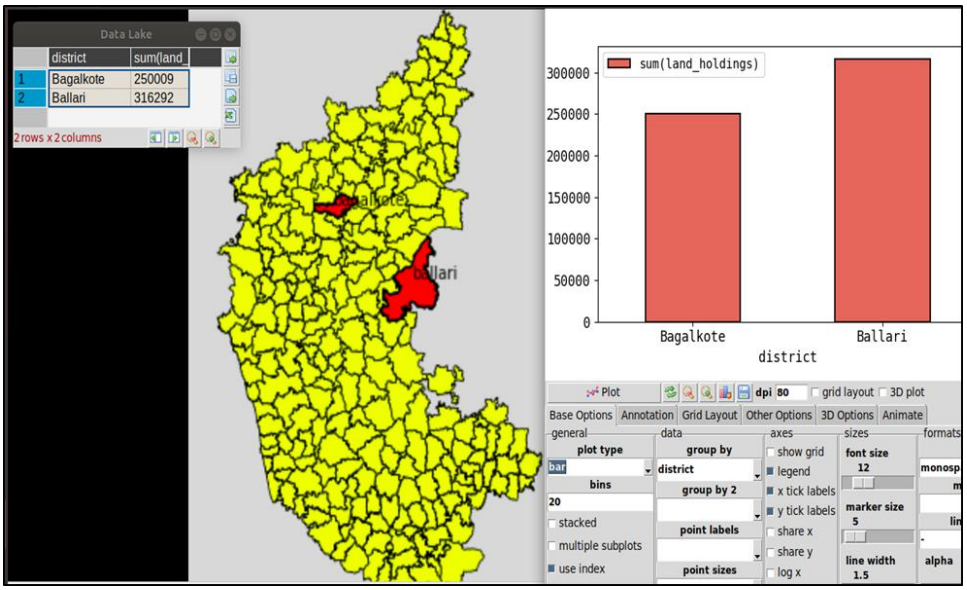


Figure 5. Graphical representation of the result highlighting geographic entities

4. Conclusion and Future Work

The proposed work focuses on retrieving a relevant set of documents over Open Data. The framework is developed to retrieve SDG-aligned data files across data sources considering geographic information into account. The current implementation accommodates SDG-2, while the research team at IIIT-Bangalore is working towards

widening the implementation for other SDGs. The implementation is evaluated by comparing it with three baseline systems: Lucene [1], Doc2Vec [2], and ElasticSearch [3]. This evaluation aims to assess the relevance of the retrieved documents. The primary results indicate superior performance, particularly in terms of achieving true positive results. These findings suggest a promising potential for future advancements in proposed work.

5. Acknowledgement

This work was supported by Karnataka Innovation & Technology Society, Dept. of IT, BT and S&T, Govt. of Karnataka, India, vide GO No. ITD 76 ADM 2017, Bengaluru; Dated 28.02.2018. The research team is also grateful to the Government of Karnataka, the IIIT-Bangalore Center for Open Data Research (CODR) and the Public Affairs Center (PAC), Bengaluru, India, for their significant data and domain expertise collaboration.

References

[1] Andrzej Białecki, Robert Muir, Grant Ingersoll, and Lucid Imagination. "Apache Lucene 4." In SIGIR 2012 Workshop on Open Source Information Retrieval, p. 17, 2012.

[2] Clinton Gormley and Zachary Tong. Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine. O'Reilly Media, Inc., 2015.

[3] Jey Han Lau and Timothy Baldwin. "An Empirical Evaluation of Doc2Vec with Practical Insights into Document Embedding Generation." arXiv preprint arXiv:1607.05368, 2016.

[4] Bassin, Pooja, Niharika Sri Parasa, Srinath Srinivasa, and Sridhar Mandyam. "Big data management for policy support in sustainable development." In International Conference on Big Data Analytics, pp. 3-15. Cham: Springer International Publishing, 2021.

[5] Kulkarni Apurva, Chandrashekar Ramanathan, and Vinu E. Venugopal. "Semantics-aware Document Retrieval for Government Administrative Data." International Journal of Semantic Computing (2023).

[6] Kulkarni, Apurva, Chandrashekar Ramanathan, and Vinu E. Venugopal. "Ontology Mediated Document Retrieval for Exploratory Big Data Analytics." In 2023 IEEE 17th International Conference on Semantic Computing (ICSC), pp. 100-103. IEEE, 2023.

[7] Kulkarni, Apurva, and Chandrashekar Ramanathan. "CDEF: Conceptual Data Extraction Framework for Heterogeneous Data." In 2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS), pp. 329-331. IEEE, 2022.

[8] Kulkarni, Apurva, Pooja Bassin, Niharika Sri Parasa, Vinu E. Venugopal, Srinath Srinivasa, and Chandrashekar Ramanathan. "Ontology Augmented Data Lake System for Policy Support." In International Conference on Big Data Analytics, pp. 3-16. Cham: Springer Nature Switzerland, 2022.