An Evaluation and Critique of India's Open Government Data

Cesar P. Malenab Jr.
College of Computer Studies
De La Salle University
Manila, Philippines
cesar_malenabjr@dlsu.edu.ph

Emmanuel Pedernal
College of Computer Studies
De La Salle University
Manila, Philippines
emmanuel_pedernal@dlsu.edu.ph

***Abstract*—Open data refers to information that is freely accessible for reuse and redistribution, fostering transparency, innovation, accountability, and economic growth. India, with its pioneering open data policies, serves as a significant example in this field. This paper evaluates India's data governance practices through the DAMA-DMBOK framework and focusing on the open data portal and data quality metrics. The analysis demonstrates that India operates under a federated data governance model, with the Non-Personal Data Authority playing a crucial role in balancing innovation with data privacy. The portal is equipped with a robust architecture ensuring the anonymity of personally identifiable information while offering public access through cloud-based APIs. It is designed to be user-friendly, incorporating intuitive navigation, data visualizations, blogs, and infographics. The paper also critiques the portal for its lack of multilingual support and the absence of comprehensive tutorials for navigating its features, suggesting that addressing these gaps could further enhance the portal's accessibility and usability for a diverse user base.**

Keywords— Open data, data governance, open data portal, data quality, DAMA-DMBOK

I. INTRODUCTION

As the world economy transitions to a data-centric model, the ability to harness and maximize the value of data has become a crucial determinant of power across political, social, cultural, and economic spheres [1]. Government data, due to its extensive reach, comprehensive coverage, and authority as a primary source of information, plays a particularly pivotal role. Governments generate and collect extensive datasets through routine activities such as managing pensions, tax collection, traffic monitoring, and issuing official documents. This vast repository of data holds significant potential for societal benefit, provided it is made accessible and usable.

Open data, defined as information anyone can freely use, reuse, and redistribute, plays a vital role in this landscape. Open data are generally available in raw form, necessitating proper citation and republishing to maintain integrity and accuracy. For governments, open data represents a significant opportunity to foster transparency, drive innovation, and stimulate economic growth. The ability to access and utilize government data empowers citizens, enhances public services, and promotes civic engagement.

One of the primary challenges of open data is ensuring its accessibility through well-designed open data portals. These portals feature user-friendly dashboards for viewing, downloading, and accessing data via APIs, organized into searchable catalogs with user-defined tags. Operated by various entities such as government agencies or citizen initiatives, these portals reflect the availability of data for public disclosure. Identifying the correct dataset with relevant variables, time frames, and categories is essential for users. Additionally, the data must be clearly described and of high quality to be converted into valuable insights.

The main value of open data lies in its potential to make government operations more transparent and efficient. Transparency not only builds trust between the government and its citizens but also fosters engagement by enabling public participation in policy discussions and decision-making processes. For example, the Chicago open data portal exemplifies how open data can be used to promote civic engagement. It offers a range of datasets in machine-readable formats, along with tools for filtering, categorization, and real-time updates. This model allows citizens to contribute to the development of applications, participate in feedback processes, and access resources that enhance public services.

However, challenges persist in making open data truly effective. Data published in unstructured formats can severely limit its usability, rendering it as ineffective as data that is not open at all. Many governments, particularly in developing countries, have made strides towards transparency but often fall short by failing to publish data in ways that are conducive to innovative use and reuse. Issues related to data quality, accuracy, privacy, and institutional politics around data ownership further complicate the effective deployment of open data initiatives.

Economically, open data holds substantial promise. The European Union estimates annual economic gains of 40 billion Euros from open data, with additional value derived from addressing societal challenges, achieving efficiency gains, and fostering citizen participation [2].

Active open data policies foster a dynamic private sector that drives economic growth and innovation. This, in turn, enhances public budgets through increased tax revenues from the expanding data industry. With this, the number of open data initiatives has grown from two to over three hundred between 2009 and 2014, and membership in the Open Government Partnership (OGP) has risen from eight countries in 2011 to sixty-nine in 2016. In India, the open data landscape is evolving, with the country ranked highly in global indices such as the

Global Open Data Index (GODI) and the Open Data Barometer (ODB). In addition, according to the findings of Heimstadt et al., India is the second country following the UK which provides the highest quality datasets, followed by the US and Australia [3]. The expansion of open data initiatives globally, coupled with India's progress, underscores the profound impact of open data on shaping policies, driving innovation, and improving public services.

This paper aims to provide a comprehensive review of the various aspects of India's open government data, including its conceptual framework, historical development, legal foundation, and the usability of the open data portal. An in-depth examination of the components of India's open government data is presented, which can serve as a model for developing countries to identify adaptable elements and areas for improvement. Furthermore, the paper includes recommendations by the authors, offering valuable insights for the enhancement of open government data initiatives globally.

II. Conceptual Framework

A. Data Governance Framework

The establishment of control and authority over all data assets by the government, intended for public sharing, necessitates a robust data governance model. This model must address key areas such as data security, modeling, design, storage, and accessibility. Decision-making through policies and control measures is articulated within the adapted data governance framework, which delineates compliance measures for data management procedures. The DAMA-DMBOK Framework [4], represented by the DAMA Wheel, places governance at the core, surrounded by ten knowledge areas that encapsulate the comprehensive data management practices of an enterprise or government sector. This document elucidates major functions and covers topics such as data strategy, policy, quality, issue management, and data asset valuation. The DAMA-DMBOK framework's advantage lies in its scalability and adaptability, making it suitable even for large-scale entities like the Government of India. To review India's data governance practices, this paper will focus on six topics that describe the country's major data governance functions.

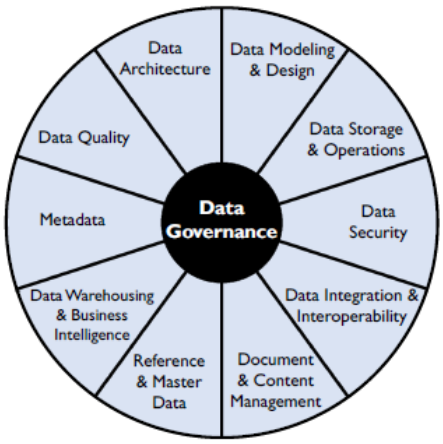


Fig. 1. The DAMA Wheel

a) Business Drivers

Two primary drivers for data governance in an enterprise are regulatory compliance and the complexity of datasets, as outlined in the McKinsey Framework [5]. Legislation such as the General Data Protection Regulation (GDPR) imposes stringent measures that industries must adhere to, ensuring the protection and privacy of personal data [6]. Additionally, the complexity of data handled by enterprises necessitates a rigorous data governance function to maintain interoperability between datasets, ensuring data accuracy and usability. Business initiatives should guide data governance practices, leading to improved data management by analyzing data access patterns, supporting artificial intelligence projects that enhance business functions, and meeting data quality requirements to support analytical measures.

b) Goals and Principles

To enable data to serve as a valuable asset for any organization, the DAMA-DMBOK framework highlights principles that establish a robust foundation for data governance programs.

- *Shared Responsibility:* This principle emphasizes the collaborative effort of both technical and non-technical personnel in maintaining data integrity across all stakeholders.
- *Multi-layered*: A comprehensive data governance program should operate at multiple levels, from data entry at the local level to the enterprise level, and should include intermediate levels as necessary.
- *Framework-based*: This principle advocates for an operating model that organizations should adopt to ensure accountability across all departments.
- *Principle-based*: This involves best practices that form the basis for data policies, with a detailed definition of business drivers

Properly defining these principles is crucial in minimizing complications and effectively guiding the enterprise's efforts in data governance.

c) Data Governance Operating Model

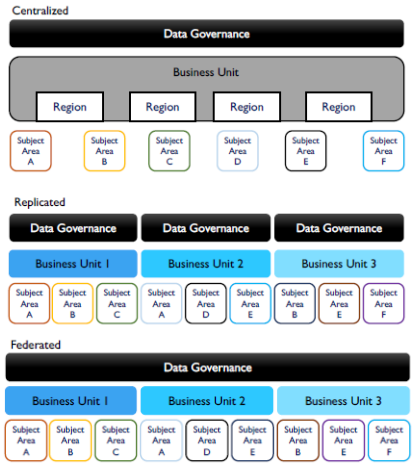


Fig. 2. The DAMA DMBOK Data Governance Operating Models

The DAMA-DMBOK framework presents three data governance operating models: centralized, replicated, and federated.

- *Centralized Model*: In this model, a single entity oversees all data governance functions, ensuring uniformity and consistency in data management practices across the organization.
- *Replicated Model*: This model features multiple committees that handle data governance for different business areas, allowing for specialized governance tailored to specific needs and functions within each area.
- *Federated Model*: The federated model is a hybrid approach that combines elements of both the centralized and replicated models. In this model, a central governing body oversees multiple business units, each responsible for specific subject areas, ensuring both centralized control and localized expertise.

Each of these models offers distinct advantages depending on the organizational structure and specific data governance needs.

d) Data Stewardship

Data stewards are responsible for handling, managing, storing, and processing data within an organization. Their primary tasks include managing data catalogs and data dictionaries, which serve as documentation for data qualities and standards. Additionally, data stewards address data issues through their daily data governance operations, supporting the overall goals of the organization. According to the DAMA-DMBOK framework, several data steward positions are differentiated by their specific focus:

- *Data Owner*: Holds authority for decisions within their domain.
- *Coordinating Data Steward*: Acts as a representative of the business domains from which data is produced.
- *Business Data Steward*: Defines and controls the data produced by data owners.
- *Enterprise Data Steward*: Oversees the overall governance of data across business functions.
- *Chief Data Steward*: Supervises the entire data governance structure within an enterprise.

e) Data Policies

The specific implementation of data governance is realized through data policies, which articulate the standards dictated by data governance principles. These data policies outline the expected outcomes and, as recommended by the DAMA-DMBOK framework, should be concise and direct to effectively convey the core principles of data governance. This brevity ensures clarity and facilitates the adherence to and enforcement of these policies across the organization.

f) Data Architecture

System design is a critical component in executing data governance principles, detailing the technical elements and relationships at various levels. Data architecture typically encompasses enterprise applications, technology, and the technical expertise necessary to fulfill data management requirements. According to the DAMA-DMBOK framework, a comprehensive data architecture includes data and metadata definitions, entity-relationship diagrams, and governing business logic. Properly implemented data architecture ensures the standardization of all data assets across the organization.

Defining data requirements based on priority business initiatives necessitates various data management practices, including data lineage and dataflow tracing, data lifecycle management, and data modeling. Data architecture that supports data governance initiatives must also be designed to be future-proof, accommodating evolving technological requirements and trends. Additionally, these technologies should comply with regulatory requirements, such as data security, to ensure that personally identifiable information is protected against potential compromises affecting data owners and principals.

B. Characteristics of Open Data Portal

To critically assess the quality level of an open data portal, several criteria must be considered. Data accessibility is crucial, discoverability and free use of the data portal with an existing accessibility tool, sharing of results, re-use of data and support for the end user.

a) Discoverability

Discoverability pertains to the ease with which users can locate and access a website and its content. For optimal discoverability, the website should rank well in search engine results, allowing users to find it using relevant keywords and minimizing the risk of navigating to erroneous sites. The website should feature a clear and intuitive structure, enabling users to effortlessly navigate and locate the information they need. Robust internal search tools are essential for users to quickly pinpoint specific datasets or information. Furthermore, descriptive information and tags should be used to categorize content, facilitating the discovery of related information. Including links to other websites, databases, and portals can enhance the website's visibility and integration within the broader data ecosystem.

b) Free Use

Data provided by the portal should be accessible, downloadable, and usable without financial constraints or restrictive licensing issues. Open licensing is crucial as it allows users to reuse data without legal repercussions. It is important to clearly communicate to users that the data can be used for various purposes, including research, analysis, or commercial activities. The data should be accompanied by detailed instructions and comprehensive information about the subject matter, which will facilitate its use without requiring extensive documentation or specialized tools.

According to Machova and Lnenicka, open data is defined as content or information that is freely available for use, reuse, and redistribution, subject only to requirements such as attribution and sharing alike. Most open data is raw, and while republishing the data requires citing the original source, this practice ensures that the data remains unmodified and accurately represented [7]. The open data concept posits that

governmental data should be accessible to anyone and redistributable in any format without copyright restrictions.

Initially, the term "open data" was used in academic contexts to advocate for the free availability of academic data. Over time, it gained political significance with initiatives such as data.gov in the United States [8], aimed at enhancing government transparency and effectiveness. The primary advantage of open data lies in its potential to improve government transparency and encourage civic engagement by providing the public with access to official documents. This access allows citizens to participate in discussions on how to address their needs more effectively.

Data sets released in machine-readable formats can be processed by independent developers for e-government projects. Advances in technology can facilitate local information sharing and peer-to-peer networking platforms, which leverage open data to empower citizens and influence decision-making processes. Such open data projects that foster direct civic participation exemplify the significant value that open data can bring to communities [8].

c) Number of Available Categories and Dataset

Driven by the National Data Sharing and Accessibility Policy (NDSAP), the National Informatics Centre (NIC) has established the Open Government Data (OGD) Platform India [9] to facilitate proactive and open access to data from various Ministries, Departments, and Organizations of the Government of India. This platform is part of Pillar 6 (Information for All) of the Digital India initiative. The current version, OGD 2.0, employs a Microservices-Based Architecture leveraging cloud technology.

The comprehensiveness of an open data platform is often reflected in the number of categories and datasets it offers. A well-organized portal should categorize data across a broad spectrum of topics, including health, education, transportation, and the environment. A greater number of categories and datasets enhances the portal's value as a resource for diverse research and analysis.

As of today, the Open Government Data Portal India hosts 620,707 datasets across various sectors, domains, departments, web services, and APIs.

d) User Interface

The user interface (UI) of an open data portal encompasses the elements that users interact with to access menus, controls, and navigation features. Designed to be intuitive and user-friendly, a well-crafted UI includes key features such as intuitive navigation, a responsive design compatible with various devices, and interactive functionalities like filters, charts, and maps that facilitate effective data visualization. A clean and organized layout, combined with quick loading times, significantly enhances the user experience.

The UI acts as a crucial link between users and the system, allowing users to interact with the computer or machine to perform tasks. As noted, the UI serves as a conduit between users and systems, enabling effective interaction with the system [10].

The open data portal organizes datasets in various machine-readable formats—such as tables, plain text, or maps—and features simple navigation and personalization options. Users can filter datasets alphabetically, chronologically, or by popularity, and the portal also supports data organization, search functionalities, and automatic RSS feeds [8].

Despite the inherent complexity of such systems, a proficient system effectively manages this complexity through a well-designed UI. Effective communication between the user and the system is essential, achieved through an intuitive user interface [11].

e) Application Programming Interface (API)

An open data portal's API (Application Programming Interface) allows developers to programmatically interact with the website to retrieve data. This functionality is crucial for integrating the portal's information with other development tools, automating data retrieval, and enabling more sophisticated analysis and presentation of the data. Providing clear and comprehensive API documentation is essential to help users understand how to make API calls and utilize the data effectively.

An open data portal aggregates information from various sources and presents it on dashboards accessible via an API. Users can select tags to identify datasets based on catalogs maintained by government agencies, citizen initiatives, and other entities. While the catalog record typically does not include the actual dataset, it generally provides a download link or web page link where the dataset can be accessed [7].

For effective data utilization, it is necessary for applications to consume linked open data programmatically. APIs offer interface specifications that enable machine-to-machine data queries, ensuring that transparent information is made available as a product, potentially for financial gain. Representational State Transfer (REST) is a popular and effective approach for delivering web APIs in open data initiatives due to its ease of deployment and consumption. JSON, often used with REST APIs, has become a preferred format among developers and API users for its simplicity and accessibility. The REST API supports data in two primary formats: Extensible Markup Language (XML) and JavaScript Object Notation (JSON), facilitating the accessibility and integration of open data [12].

f) Projects created by users

Highlighting projects developed by users can effectively showcase the practical applications of the data provided by the open data portal. These projects demonstrate how the data can be utilized to make informed decisions, foster innovation, and inspire creativity. By sharing user-generated projects, the platform can illustrate the utility and impact of the data, highlighting its value to governments, businesses, and the broader community.

According to Zuiderwijk and Jansen, the reuse of data has not been sufficiently emphasized in generating new analyses for the public sector or government [13]. Consequently, the public sector has implemented processes to ensure there are no restrictions on data reuse. Making information accessible online promotes greater openness, engagement, and creativity within the open data community.

Open data publication platforms must meet specific legal, administrative, and technical requirements. In practice, accessing raw data, contextualizing it, and extracting valuable insights can be challenging. Over recent years, various solutions have been developed to streamline the data reuse lifecycle, including data discovery, cleaning, integration, processing, and visualization. The open data process encompasses all activities from data creation to dissemination, including publishing, finding, and reusing data. Open data involves a range of participants, such as facilitators, brokers, citizens, businesses, and legislators [7].

g) Linked Data

Linked Data refers to a methodology for publishing structured data that facilitates its interconnection and enhances its utility through Semantic Web technologies. This involves using Uniform Resource Identifiers (URIs), Resource Description Framework (RDF), and various vocabularies and ontologies to create a network of related data. By integrating datasets with other data sources, linked data makes the information richer and more accessible, thereby simplifying analysis and understanding.

An open data portal should support linked data to enable seamless connection and integration with other datasets. This interoperability ensures that the data is readily accessible and usable across diverse platforms and applications. The utilization of cloud-based storage combined with linked data technologies further enhances the accessibility and consumption of data.

The value of information increases as it is reused and linked to other sources. Providing detailed and illuminating information about each dataset can facilitate and encourage this linking process. This approach led to the development of Linked Open Data (LOD), which combines linked data principles with open data practices to organize and make data accessible for reuse without restrictions. This practice is highly recommended for reducing technological and cost barriers associated with data aggregation processes [7].

Tim Berners-Lee, the creator of the World Wide Web, developed a five-star rating system to encourage the implementation of linked datasets. According to his system, linked data represents the highest level of open data, promoting more effective and efficient data reusability [12].

h) Metadata

Metadata refers to the information that describes specific data, providing essential details such as its origin, methodology, descriptive attributes, and limitations. Comprehensive metadata helps users understand the context of the data and ensures more effective utilization. Effective and thorough metadata is crucial for maintaining transparency and credibility.

Historically, metadata was primarily a concern for information professionals involved in cataloging, classification, and indexing. However, with the increasing number of creators and consumers of digital content, metadata has become essential for ensuring that datasets can be efficiently cataloged and accessed. Without sufficient metadata, such as descriptions or tags, datasets can be challenging to locate through manual or automatic searches. Typically, metadata includes the dataset's name, description, and the URL of the actual sources, such as files or service endpoints. This information facilitates users in finding and filtering the data they need.

The Data Catalogue Vocabulary (DCAT), developed by the World Wide Web Consortium (W3C), is a significant RDF vocabulary that data portals should adopt to enhance interoperability. Using DCAT to describe datasets improves discoverability and makes it easier for users to access data from various catalogs [7].

As the quantity and diversity of data sources continue to grow, it is imperative to develop proficient metadata—such as detailed descriptions, geographical coverage, and limitations—to enable stakeholders without domain expertise to efficiently search and utilize data. Quality in an open government environment can encompass various aspects. Low-quality metadata can adversely affect the discovery and use of datasets within and across multiple portals. For instance, inadequate metadata impairs search and discovery services' ability to locate relevant datasets, while inaccurate descriptions can create bottlenecks in data processing and integration [7].

i) Email Alerts

Email alerts are automated notifications or reminders designed to inform users about specific events or changes. These alerts can keep users updated on new datasets, modifications to existing data, and other significant updates within the portal. By providing timely information through email, users can stay informed about the latest developments without needing to continuously check the website.

j) Feedback forms and FAQs

Feedback forms and FAQs are essential tools for enhancing user experience on open data portals. Feedback forms enable users to provide opinions, report issues, or suggest improvements, helping entities refine their services based on user input. These forms are valuable for understanding user needs and enhancing the portal's functionality. FAQs, on the other hand, address common questions or obstacles encountered by users, offering readily available answers that reduce the need for users to search for information or seek intervention.

A well-designed FAQ section can resolve frequent inquiries quickly, thereby improving customer support efficiency. Feedback forms and FAQs together help users resolve issues and find answers more efficiently, contributing to a better overall user experience.

Authorities responsible for open data portals should consider the diverse needs of individuals seeking transparency-related data and provide appropriate tools and structures to meet these needs. Enhancing the functionality of existing open government data (OGD) sources with more efficient tools for data discovery can help users locate datasets of interest more easily and quickly. Data visualizations, such as maps or charts, can provide an intuitive understanding of datasets, allowing users to assess whether further analysis is warranted.

User feedback is crucial for improving the quality of datasets and addressing weaknesses or gaps. Collaboration between OGD users and providers is essential for generating

value from the data, as feedback helps identify necessary improvements and additional datasets needed [7].

k) Security

The protection of both data and the open data portal itself is paramount. This involves safeguarding user information, securing API access, and preventing unauthorized access or data breaches. Implementing robust security measures such as encryption, authentication, and regular security audits is essential for maintaining the integrity and trustworthiness of the open data portal.

According to the Open Government Data of India Report, the Indian government recognizes the importance of investing in electronic security and privacy frameworks to support growth and protect data. In response, the government is considering new regulations for data protection. The Data Security Council of India (DSCI), established by NASSCOM to enhance the trustworthiness of Indian companies as global service providers, is currently working with the government on these regulations. The DSCI focuses not only on ensuring the economic value of data flows but also on safeguarding individuals' rights to their personal information [14].

C. Characteristics of Datasets

a) Data Quality

Data quality is a critical measure of how accurate and reliable the data on an open data portal is. It reflects the extent to which data collected from sources accurately represents the original values without errors or biases. To ensure data quality, open data portals must implement rigorous data verification processes with their sources, maintaining accuracy and reliability. Users should be able to trust that the data available is free from errors that could compromise analysis.

Data quality attributes can be assessed from various perspectives and contexts. From the user's point of view, data quality is often defined as the suitability of data for its intended use, meaning it meets user demands effectively. Quality measures typically focus on aspects such as consistency, accuracy, and completeness. Batini et al. provided a comprehensive description of methodologies for assessing and improving data quality, highlighting various approaches and criteria for evaluation. They emphasize the importance of assessing data quality to understand its value when used appropriately. Consistency, including contextual, temporal, and operational aspects, is a key component in evaluating the quality of heterogeneous datasets [7].

In the open data domain, data quality standards and measures are increasingly used to assess the quality of datasets and online platforms. To meet high-quality informational requirements, open data portals must address a broad range of entities and requirements within each country's institutional framework. Challenges such as semantics and language can impact data quality. Key characteristics for evaluating open data portals include quality, completeness, accessibility and visibility, usability and comprehension, timeliness, value and utility, granularity, and comparability.

Success in open data initiatives largely depends on the quality of the data provided. Users must understand the nature of the data to assess its quality effectively. Since data producers cannot anticipate all potential uses, providing high-quality metadata is crucial. This includes detailed descriptions and contextual information that aid users in evaluating the data. Government organizations are also focusing on improving data availability and quality while encouraging innovation through the use of open data [15].

b) Completeness

Completeness refers to the extent to which datasets include all required data without missing values or gaps. For an open data portal, completeness means that datasets should provide comprehensive coverage, ensuring users have unrestricted access to the full scope of information. Incomplete data can lead to inaccurate analyses and misguided decisions. Therefore, the open data portal should clearly state metadata that describes the scope, coverage, and limitations of the data, including any areas where information may be lacking.

Missing values are a significant issue that can hinder accurate assessments and analyses. In statistical studies, missing values often arise from responses like "don't know," "refused," or "unintelligible." Handling missing values effectively is crucial to avoid introducing bias and to ensure reliable results. One common method is the deletion of records containing missing values, which excludes incomplete data from analysis but can lead to the loss of valuable information and potential selection bias if the proportion of missing values is substantial.

An alternative approach is imputation, where missing values are estimated and substituted based on available data. This method avoids the loss of data but requires careful handling to prevent introducing bias. Statistical procedures, such as the expectation-maximization algorithm developed by Dempster et al., are used to improve the estimation of missing values and enhance data completeness [15].

Emran identifies several types of completeness measurements:

- *Null-based Completeness*: Focuses on handling missing values by substituting them with null values or placeholders.
- *Tuple-based Completeness*: Concerned with ensuring that tuples, or ordered sequences of values, are complete, where values can be repeated but must be finite.
- *Schema-based Completeness*: Ensures that all attributes and entities described in the schema are filled, addressing completeness at the attribute and entity level.
- *Population-based Completeness*: Measures completeness by comparing missing individuals or data points against the total population, ensuring that the dataset represents the entire population accurately.

By addressing completeness through these categories, open data portals can better manage and fill missing values, thus improving the reliability and usability of the data provided.

c) Consistency

Data consistency pertains to the uniformity and reliability of data across various datasets, systems, and databases. This includes handling issues such as missing values, null values, and duplicates through methods like imputation and other validation techniques. An open data portal should provide clear instructions on the methods employed to achieve consistency throughout its datasets, facilitating better information analysis. A common challenge arises from the disconnect between open data publishers and users, where providers focus on facilitating data accessibility without fully understanding users' needs for data formats and publishing practices. Lourenço emphasizes that to improve transparency and accountability, data catalogs must meet specific prerequisites, including data quality, completeness, consistency, and timeliness. These aspects pertain to the organization of datasets, the variety of available information, and the strategies for information extraction.

There are two primary strategies for enhancing data quality: data-driven and process-driven. The data-driven approach involves altering data contents to correct errors or standardize information, while the process-driven approach focuses on improving data creation and modification processes. Addressing data quality issues, such as ensuring accurate and complete catalog records, is crucial. Inaccuracies in terms of use or broken links can lead to unintended violations or inaccessibility of datasets. Kucera et al. highlight that catalog records should accurately reflect the associated data to ensure that end users receive precise and useful information [10].

d) Timely Details

Timeliness is a critical dimension of data management, particularly within open data portals and big data environments. It underscores the necessity for datasets to be current and reflective of the latest developments to support effective decision-making and analysis. According to McGivray (2010), "timeliness" pertains to the rapid creation and dissemination of information, emphasizing that data must be accessible within a reasonable timeframe to be of practical use [16].

For open data portals, maintaining timeliness involves several key practices. Regular updates are essential to ensure that data remains current and relevant. Version control is also crucial, as it helps maintain data integrity and provides a historical record of changes. Additionally, ensuring prompt availability minimizes delays between data collection and accessibility, thereby enhancing the usability of the data.

The management of big data introduces further complexities in maintaining timeliness. The rapidly evolving nature of big data necessitates ongoing advancements in processing tools and practices to keep pace with the changes. As highlighted in [7], timeliness is a component of the Availability dimension, which encompasses the prompt arrival and usability of data, regular updates, and effective version control.

Addressing these aspects of timeliness is essential for mitigating the risks associated with outdated or erroneous information. By ensuring that data is current and accurately versioned, organizations can enhance the accuracy and reliability of their analyses and decision-making processes.

III. HISTORY AND LEGAL BASIS

A. Origins and Principles of Open Data

The advocacy for transparency and access to information gained momentum during the mid-20th century, notably through the efforts of a committee of the United States National Research Council, which called for an international system of "full and open exchange" of data to address complex global challenges. This period saw the emergence of a broader movement advocating for the freedom of information and greater transparency in government record management.

The principle that public data should be openly available was further reinforced by the advent of information technology, which made data collection, storage, and dissemination more feasible and cost-effective. The push for open data was also driven by the need to tackle global issues such as climate change, economic development, and public health crises, which required robust data sharing across borders.data.

B. Open Government Data Principles

The modern principles of open government data were significantly shaped by a group of thinkers and activists who convened in Sebastopol, California, in 2007. This meeting included prominent figures such as Lawrence Lessig, Carl Malamud, Aaron Swartz, and Tim O'Reilly. They articulated the eight Open Government Data (OGD) Principles, which have since served as a foundation for open data initiatives worldwide. These principles define open government data as data that is:

1. *Complete*: All public data is made available, subject to valid privacy, security, or privilege limitations. This ensures that data sets are not selectively published and that the public can access a comprehensive range of information.

2. *Primary*: Data is collected at the source with the highest possible level of granularity, rather than in aggregate or modified forms. This principle emphasizes the importance of raw data for accuracy and utility in various applications.

3. *Timely*: Data is made available as quickly as necessary to preserve its value. Timeliness is crucial for data to be actionable and relevant, especially in areas like emergency response and policymaking.

4. *Accessible*: Data is available to the widest range of users. Accessibility ensures that data is usable by diverse stakeholders, including researchers, businesses, and the general public.

5. *Machine processable*: Data is reasonably structured to allow automated processing. Machine readability is essential for the efficient analysis and integration of data into various applications.

6. *Non-discriminatory*: Data is available to anyone, with no registration requirement. This principle promotes equal access to data, preventing barriers that could limit its use.

7. *Non-proprietary*: Data is available in a format over which no entity has exclusive control. Open formats prevent vendor lock-in and promote interoperability.

8. *License-free*: Data is not subject to copyright, patent, trademark, or trade secret regulation, allowing for reasonable privacy, security, and privilege restrictions. This ensures that

data can be freely used, modified, and shared without legal constraints.

These principles have evolved as governments worldwide have implemented open data initiatives. While there are variations in the definitions and implementations of open data, most institutions reflect a shared premise of enabling the accessibility and use of government data by the general public.

Table 1: Definitions and Conceptions of Open Data
Source: Author's representation based on cited sources

Institution / Instrument	Conception of Open Data	Key Concepts
Open Knowledge Foundation[11]	Data that can be freely used, modified, and shared by anyone for any purpose.	• Free use / Usability • Sharing • Modification
US Open Data Policy (Part I (Definitions))[12]	Publicly available data structured in a way that enables the data to be fully discoverable and usable by end users.	• Free use/ Usability • Discoverability • Accessibility (public availability) • Data format (structured)
EU Open Data Directive[13] (Recital 16)	Open data as a concept is generally understood to denote data in an open format that can be freely used, re-used, and shared by anyone for any purpose.	• Free use • Sharing • Data format (open format)
UK National Data Strategy (Glossary)[14]	Data that can be freely used, re-used, and re-distributed by anyone, subject only, at most, to the requirement to attribute and share alike	• Free use • Redistribution • Attribution
Government Open Data License - India (Part 1. Preamble)[15]	Structured data available in open format and open license for public access and use.	• Accessibility (public access and use) • Data format (structured, open format) • Attribution (open license)

While the core principles of open data are widely accepted, their implementation can vary significantly between countries. For instance, the UK's national data strategy and India's National Data Sharing and Accessibility Policy (NDSAP) require an open license to access data, justified by the need to prevent misuse or misinterpretation. This contrasts with the license-free and non-discriminatory vision of the OGD Principles. Such differences highlight the ongoing debate about how best to balance openness with control and security concerns [17]

C. Open Data in India

a) Vision

To foster a transparent and inclusive data environment where government-held data is freely accessible to all stakeholders. It aims to promote innovation and economic growth by enabling the use of government data in diverse applications while ensuring robust privacy protections. Enhancing data transparency and accessibility will empower citizens, researchers, businesses, and policymakers with reliable information for informed decision-making. It envisions a data-driven governance approach that improves efficiency within government operations and strengthens public trust through accountable data management practices.

b) Mission

India's open data initiatives aim to enhance access to shareable data and information owned by the Government of India in both human-readable and machine-readable formats. It seeks to proactively and periodically update this information, ensuring it aligns with existing policies, acts, and rules of the Government of India. This approach facilitates broader accessibility and utilization of public data and information across the country. This mission is underpinned by the belief that open data can drive socio-economic development, improve public service delivery, and support evidence-based policymaking.

c) Early Developments

In India, the concept of open government data developed over time, influenced by both international trends and domestic needs. The Indian Council for Social Science Research (ICSSR), established in 1969, played a significant role in creating research institutes across the country. Although these institutes were privately owned, they received grants from the ICSSR, which facilitated data collection. However, the responsibility for data collection remained primarily with central and state governments, conducted through agencies like the Census of India and the National Sample Survey Office (NSSO) [18].

The adoption of computers by the central government in 1975 and the establishment of the National Informatics Centre (NIC) marked a significant shift from manual, paper-based systems to automated processes. This transition was crucial for the development of a robust data infrastructure, laying the groundwork for future open data initiatives.

d) Economic Reforms and the RTI Movement

Following the economic reforms of the early 1990s, private and non-profit organizations increasingly took on data collection, storing local and contextual data on their focus issues, and interacting more frequently with the government [19]. India's open data policies emerged in a context of heightened public demand for government accountability, influenced by the right to information (RTI) movement. This grassroots campaign, primarily led by marginalized laborers and rural communities, sought to overturn colonial-era laws that limited access to official records. The movement culminated in the landmark Right to Information Act in 2005, which is similar to freedom of information legislation in many countries. The RTI Act is cited in the preamble to India's open data policy, NDSAP, as a key motivation.

e) Establishment of NDSAP and Digital India

The Indian government approved the National Data Sharing and Accessibility Policy (NDSAP) in 2012, advocating the proactive disclosure of "all sharable, non-sensitive datasets in open formats" by various government entities. The NIC, in collaboration with the US government, created the Open Government Data Platform India (data.gov.in), enabling Indian government departments and ministries to publish datasets for easy access by citizens. In 2015, 85 government ministries, departments, and agencies contributed over 12,000 datasets across various sectors [20].

The Indian government launched the Digital India program, aiming to ensure that people could access government services online. This initiative was complemented by the establishment of the Development Monitoring and Evaluation Office (DMEO) in 2015, which has been providing the government with rigorous, data-driven, citizen-centric, and outcomes-driven program management and policymaking since its inception.

In 2017, the Government of India introduced a revised version of the National Data Sharing and Accessibility Policy

(NDSAP). This policy aimed to enhance the accessibility and usability of government data by emphasizing two critical aspects: machine-readable formats and non-discriminatory access. The shift towards machine-readable formats such as CSV, JSON, and XML facilitated easier data integration and analysis for researchers, businesses, and citizens alike. Non-discriminatory access ensured that data was accessible to all stakeholders without bias or undue restrictions, promoting transparency and accountability across government departments.

D. Challenges and Community Efforts

India's open data policy includes provisions allowing government departments to decide which datasets to share, meaning each department can determine the shareability of its data. Additionally, access to shared data may be subject to registration, reflecting tensions between aspirations of openness and government control over data releases.

Another distinct aspect of India's policy is the assertion of government ownership of public datasets. Although the NDSAP recognizes that data is gathered through public investment, it frames data as an "asset" and includes provisions for pricing datasets. In 2022, the Indian government proposed new frameworks to replace the NDSAP, continuing to assert government ownership and control of data, with provisions to charge fees for maintaining open data services. This framing is similar to policy documents from other countries. The EU's Data Strategy, the UK's national data strategy, the US Federal Data Strategy, and China's five-year plans all describe data as a strategic asset or resource essential for economic growth, innovation, and societal progress.

However, the growth of the OGD platform in India has faced challenges. Many data-rich public agencies and departments have refrained from contributing datasets or updating past contributions. Data shared is often in inaccessible formats or incomplete. Even before the NDSAP, small communities of NGOs and individuals had been experimenting with aggregating and sharing data from various sources. Organizations like DataMeet, WikiData, and OpenStreetMap have become prominent, filling gaps left by the official OGD portal by building data pipelines and tools to enhance accountability and governance.

During the COVID-19 pandemic, community-led open data efforts were particularly significant, as volunteers sourced, verified, and presented data independently due to the lack of streamlined government information [17].

E. Recent Developments

In 2018, India underwent significant reforms with its Geospatial Data Policy, transitioning from stringent regulations to a more liberalized approach. The policy aimed to streamline access to geospatial data and maps, empowering Indian entities to utilize location-based information more effectively across various sectors. According to the analysis by Y Nithiyanandam and Satyam Kushwaha [21], these reforms were pivotal in enhancing decision-making in urban planning, agriculture, disaster management, and infrastructure development, thereby fostering innovation and economic growth. However, the analysis also highlights several challenges that may impede the policy's full realization. These include issues related to data quality and accuracy, interoperability among different datasets, concerns over privacy and security of sensitive geospatial information, and the need for capacity building among users to effectively utilize the available data resources. Moreover, the article suggests that despite the policy reforms, there remains a gap in comprehensive governance frameworks and standardized practices for managing and sharing geospatial data effectively across sectors.

In 2019, the Ministry of Electronics and Information Technology (MeitY) introduced a draft data policy aimed at unlocking government data for broader accessibility and innovation. The policy proposed measures to facilitate easier access to government data by businesses, researchers, and citizens. It emphasized the importance of data as a strategic asset, promoting transparency and accountability in governance while encouraging economic growth through data-driven innovation. However, according to the Economic Times [22], one of the challenges anticipated was the need for robust data security measures to safeguard sensitive information while ensuring open access. Additionally, ensuring compliance and adoption across diverse government entities posed logistical and technical challenges. Despite these hurdles, the policy aimed to unlock the potential of government data for fostering digital innovation and improving public service delivery.

NITI Aayog proposed the development of a National Data and Analytics Platform (NDAP) in 2020 and aimed at integrating data across various sectors for policy-making and public use. The platform was envisioned to centralize and standardize data from sectors such as healthcare, education, agriculture, and the economy [23]. This initiative aimed to enhance governance effectiveness by providing policymakers with comprehensive and reliable data insights for evidence-based decision-making. NDAP also aimed to facilitate public access to data, fostering transparency and accountability in government operations.

Introduced in 2021, India introduced the Data Empowerment and Protection Architecture (DEPA), aimed at empowering individuals with control over their personal data while ensuring its secure and transparent use. According to Jain [24], DEPA enables secure sharing of personal data across various service providers, enhancing consumer trust in digital transactions and fostering innovation in data-driven services. However, challenges exist, including technical implementation complexities, ensuring robust data security measures to prevent misuse or breaches, and balancing data openness with privacy protection concerns. Additionally, achieving widespread adoption and compliance across diverse sectors and stakeholders poses a challenge, requiring effective regulatory frameworks and stakeholder engagement to address these concerns.

In 2022, the Indian government introduced the Non-Personal Data Governance Framework and the Digital Personal Data Protection Bill to enable data sharing. These frameworks include several recommendations for data-sharing purposes [25]:

9. *Sovereign purpose*: Data may be requested for national security, law enforcement, legal, or regulatory purposes.
10. *Core public interest purpose*: Data may be requested for community uses/benefits or public goods, research and innovation, policy development, better delivery of public services, etc.
    a. Specific data with commercial importance may be recognized as high-value datasets.
    b. Utilize data for research purposes.
    c. Consider the health sector as a pilot use case for the Non-Personal Data Governance Framework.
11. *Economic purpose*: Data may be requested to encourage competition and provide a level playing field or encourage innovation through start-up activities (economic welfare purpose) or for a fair monetary consideration as part of a well-regulated data market.
    d. Data requests by start-ups/businesses.
    e. Data requests by data trustees/governments.
    f. Setting up data and cloud innovation labs

IV. Evaluation and Critique of Data Governance and Management

A. Data Governance

a) Business Drivers, Goals, and Principles

The vision and mission of India's open data initiatives, as previously discussed, are fundamentally driven by two key factors: enhancing government accountability and fostering technological advancement through analytics and artificial intelligence.

The Right to Information (RTI) Act has been instrumental in advocating for greater transparency in governance. This legislative framework laid the groundwork for the development of an open data platform, aiming to foster trust and accountability by ensuring that reliable information is accessible to Indian constituents. By making data available in both human-readable and machine-readable formats, the initiative seeks to promote inclusivity, catering to both technical and non-technical audiences. This transparency is expected to increase public awareness of governmental projects and contribute to more informed decision-making processes. Moreover, the availability of comprehensive and accessible data is anticipated to enhance the efficiency of government operations and ensure that both existing and future policies are grounded in robust data evidence.

In addition to promoting transparency, innovation and technological advancement are central to India's open data strategy. The initiative envisions leveraging government data across various applications to harness the potential of citizens in transforming raw data into valuable, actionable insights. By embracing advanced technologies, such as analytics and artificial intelligence, the platform aims to facilitate the creation of sophisticated data products that can drive further innovation and enhance overall governance. This dual focus on transparency and technological advancement underscores India's commitment to utilizing open data as a means of both enhancing public trust and stimulating technological growth.

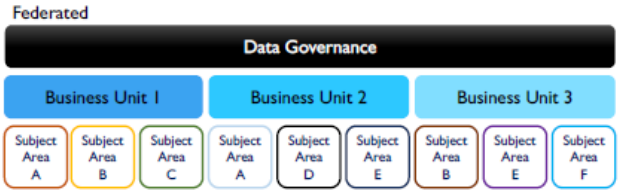


Fig. 3. Federated Data Governance Model

b) Data Governance Operating Model

Among the data governance operating models outlined by the DAMA-DMBOK, the federated model bears significant resemblance to India's framework. This alignment is evident in the definition of "Data Business" as described in the "Reports by the Committee of Experts on Non-Personal Data Governance Framework," published by the Ministry of Electronics and Information Technology.

In India's framework, a business is classified as a Data Business if it exceeds a specific threshold set by the Non-Personal Data Authority. This classification does not represent an additional industry sector but rather a horizontal categorization that mandates compliance with certain requirements, including access to metadata. This approach extends beyond private sectors to encompass government and non-government organizations as well. Key sectors identified for this classification include banking, transportation, research labs, and consumer goods, among others.

India's classification system exemplifies a federated data governance framework. Within this model, entities that possess expertise in their respective fields are categorized as Business Units, each governed under a centralized data governance body. This structure allows for specialized management and oversight of data within various sectors while ensuring adherence to overarching governance standards.

c) Data Stewardship

Similar to the data steward roles outlined in Chapter Two of this paper, India has established specific roles for managing non-personal data. These roles, as detailed in the table below, align with the corresponding roles defined in the DAMA-DMBOK framework.

TABLE I. KEY ROLES IN INDIA'S DATA GOVERNANCE

DAMA-DMBOK	***India Data Governance***	***Description***
Data Owner	Data Principal	Natural person to whom data relates
Coordinating Data Stewards	Data Trustees	Exercise data rights on behalf of community
Business Data Stewards	Data Custodian	Undertakes collection, storage, processing, and use of data
Enterprise Data Stewards	Data Trusts	Defines rules and protocols for containing and sharing data
Chief Data Stewards	Non-Personal Data Authority	Governs all data roles and regulations.

Non-personal data encompasses more than merely anonymized personal data; it can also derive from communities or exist as public data. A pertinent example of community data

includes information generated by farmers using IoT devices, where no single individual possesses exclusive rights to the data. Instead, the community collectively retains economic rights. In such scenarios, the community can appoint a data trustee to act as their representative and safeguard their collective interests.

The data trustee, serving on behalf of the community, can designate a data custodian responsible for developing and implementing practices and guidelines for the collection and storage of the data. Data trusts may then seek the approval of data principals to access this data for specific purposes that benefit society. These trusts can utilize the data managed by custodians for economic objectives. Ultimately, the Non-Personal Data Authority has the authority to make final decisions regarding data sharing, particularly if the intended use is deemed beneficial. This authority can override the reluctance of data principals to share their data, ensuring that decisions align with broader societal interests.

d) Data Policies

Several data policies warrant discussion, with four key policies being the focus of this section: the definition of non-personal data, the rights of data principals, the requirements for data businesses, and data sharing policies.

Definition of Non-Personal Data: Non-personal data can be categorized into public, community, and private data. Public non-personal data includes information collected by the government, such as land records and public health data. Community non-personal data refers to data produced by a collective of individuals, such as data from municipal corporations or private entities like telecom companies. Private non-personal data is generated by individuals and is not personally identifiable or comes from an entity's private efforts.

Rights of Data Principals: Regardless of how non-personal data is collected, data principals have the ultimate right over how their data is used and consumed. For example, in government-conducted censuses, the policy in India grants full data rights to the individuals to whom the data pertains, even if collected by the government. Similarly, for community data, the producers, or specific organizations responsible for collecting the data, retain ultimate rights over their data.

Requirements for Data Businesses: Organizations identified as data businesses are required to disclose the data they collect, store, and process. Metadata should be made publicly available, provided it does not raise data privacy concerns. This policy aims to foster the integration of multiple datasets from various industries and sectors to stimulate innovative products through data interaction. Compliance requirements for data businesses are mandated to be fulfilled entirely through digital means and are compulsory by law.

Data Sharing Policies: Data businesses may be required to share data for several reasons, categorized into sovereign purposes, public interests, and economic purposes. For sovereign purposes, such as national security, disease mapping, and law enforcement, companies are required to share data like geospatial and telecommunications data. For public interests, companies must disclose data related to infrastructure, public services, and some high-value datasets like agriculture and education. Under economic purposes, companies may be required to provide data, such as road data from transportation companies, to better address public transport demand.

e) Data Architecture

The data architecture illustrated above supports the data governance framework envisioned by India. Data collected from various sources will be processed through differential privacy algorithms, which anonymize all datasets according to the thresholds and standards specified by the Non-Personal Data Authority. These differential privacy algorithms will be developed by experts from academia and industry. Anonymized datasets will then be made available either as aggregated datasets or through secure non-personal data clouds. Cloud providers will create APIs for registered institutions to access this data for developing machine learning models or gaining insights through analytics.

A key component of this architecture is the policy switch, which streamlines the compliance requirements for data principals or data trustees by consolidating them under a single authority. This policy switch will function as a digital clearinghouse for non-personal data. The policies will be established by the Non-Personal Data Authority, ensuring data security while maintaining economic incentives across India.

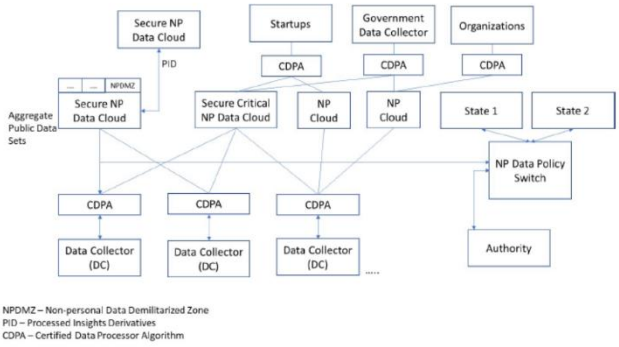


Fig. 4. Architecture of different stakeholders, data and control flow

B. *User Interface Features*

a) Navigation

On the home page, users can access various options including "Home," "Catalog," "API," "Sector," "Chief Data Officer," and "Metrics." Additionally, there is a font size option allowing users to adjust the text size according to their preference. This option also includes a dark mode feature for more comfortable browsing.

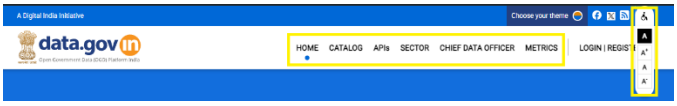


Fig. 5. Homepage

Upon selecting "Catalog," users are presented with a "Filter by" option to search for specific datasets or groups of datasets within the catalog. Users can filter their searches by "Most

Recent," "Relevance," or "Updated," and choose between fuzzy or exact matches for the datasets.

From the catalog, users can refine their searches using the following options:

- Domain: Allows selection based on regions in India.
- Sector: Includes all government sectors in India such as Census, Agriculture, Transportation, Tourism, Science and Technology, Defense, and Economy.
- Ministry/Department: Filters datasets based on the specific ministry or department that provided the data.
- Asset Jurisdiction: Provides a dropdown of locations to filter datasets by specific locations.
- Group: Enables a pseudo group-by command for the dataset, allowing grouping based on categories chosen by the user.
- Catalog/API and High Value: Allows boolean filtering to determine if the dataset contains an API or is considered high value.

For each dataset, metadata details are displayed, including the number of datasets within the group, API availability, number of views, number of downloads, last update date, and publication date.

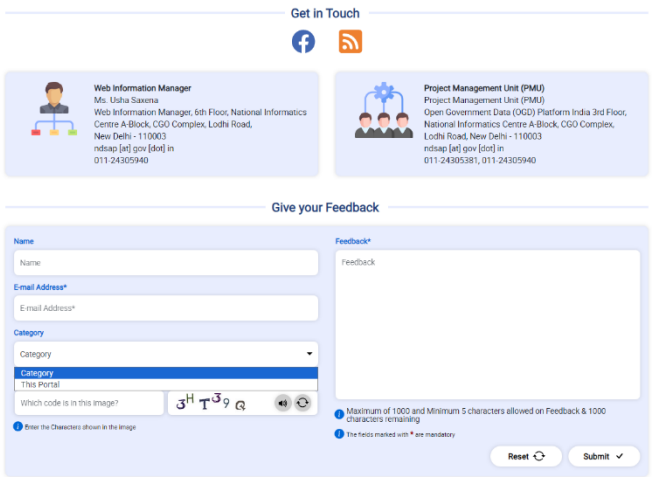


Fig. 6. Catalogue

The "Chief Data Officer" option allows users to select a department or ministry associated with a specific Chief Data Officer (CDO). This section provides details such as the work location of the CDO, their office phone number, email address, and mailing address. This information enables users to directly contact the CDO with concerns or queries related to datasets specific to that CDO's ministry.

b) Feedback Forms, Email Alerts, User Account Creation

Feedback forms are an essential information-gathering tool that allow the governing authority to obtain user opinions for the improvement of the model. These forms are available under the website's support category, which also includes the contact information for the Web Manager and Project Manager, who provide chat support for any issues or feedback regarding the portal.

To give feedback, users need to complete the form by providing their name, email address, selected category, and their comments (within specified minimum and maximum character limits). To prevent automated submissions, users must also complete a CAPTCHA before submitting the form.

Email alerts provide users with updates such as newsletters, new developments on the portal, or updates to specific datasets they are interested in. To receive these updates, users must sign up with the Open Data Portal of India. Account creation is managed through a partnership with the third-party service provider, Meri Pehchaan. Users can create an account either through existing government-based accounts or by linking their social media and email accounts to the portal.

The newsletters received by users contain monthly updates to the Open Data Portal, including information about the top-performing ministry for the portal. Additionally, the newsletter provides numerical updates, such as the number of updated datasets, new catalogues, visualizations, APIs, download counts, views per month, and the total number of new users of the open data portal.

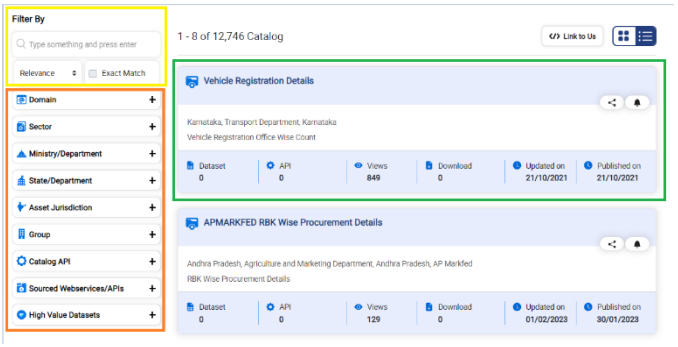


Fig. 7. Homepage

c) Data Visualizations

Data visualization is the graphical representation of information and data, using tools such as charts, graphs, and maps to make trends, anomalies, and recurring patterns easy to understand. It provides an effective way for employees or business owners to present data to non-technical audiences clearly.

On the Open Data Portal of India, users have three options for data visualizations:

- Latest Visualization: Displays the most recently created visualizations regardless of the category or location of the datasets.
- Recent Visualization: Shows recently created visualizations, similar to the "Latest Visualization," without filtering by category or location.

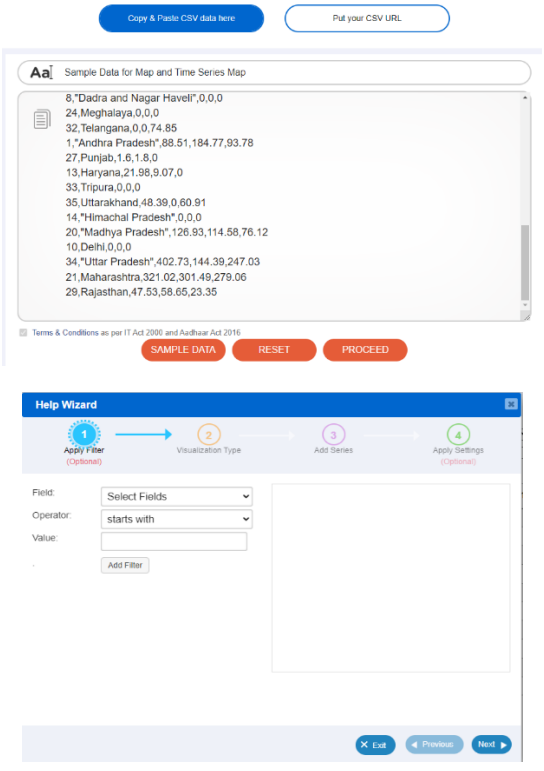
Selecting either of these options opens a new tab displaying the visualization and its settings. Users can report bugs, download, embed, and share the data visualizations using the help tools provided as shown in Figure 7.

- Create Your Own Data Visualization: Users can create custom visualizations from their own data as shown in Figure 8. After selecting this option, users are

prompted to enter details such as the visualization title and can directly input CSV data. Once the CSV file is loaded, the portal allows users to create panel data similar to Excel, using columns and rows to build their visualization. Users can choose from various visualization types, including timeseries data, maps, bar charts, stacked charts, and pie charts.

The website provides a help wizard to guide users through the creation process, offering options and customizations such as attribute filters, sorting methods, visualization types, focus attributes, and details to be displayed in the final visualization.

Fig. 8. Sample Data Viz



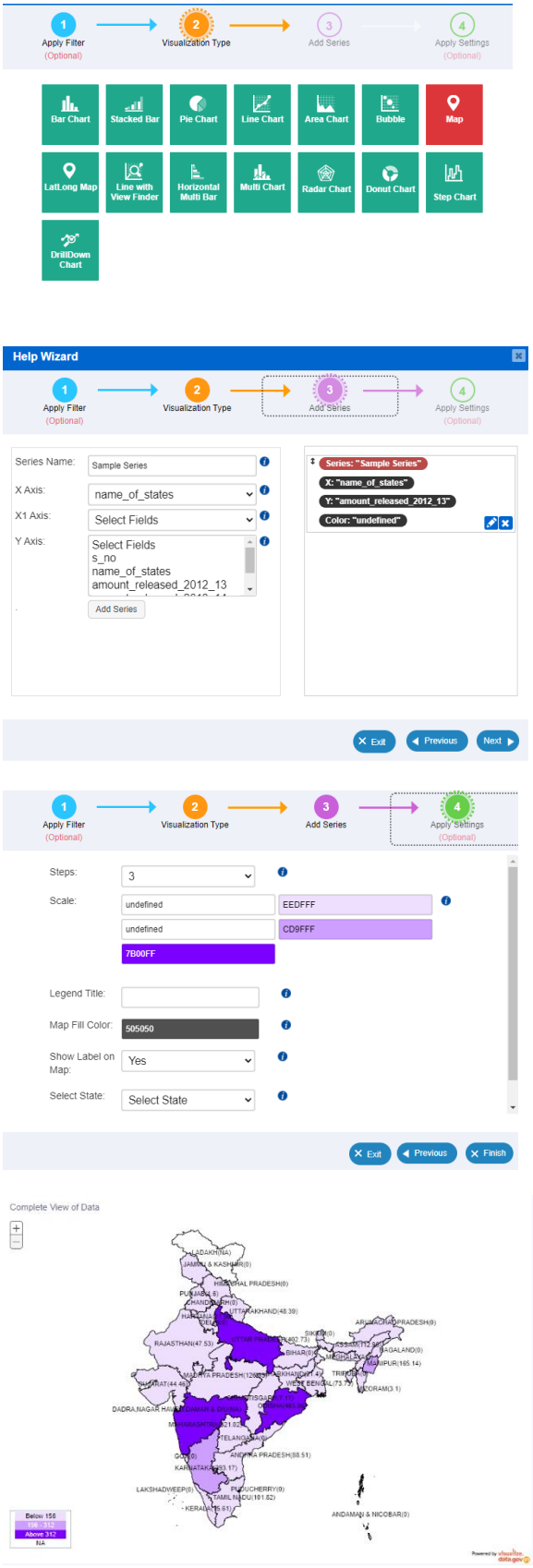


Fig. 9. Creating a Data Viz

C. Data Architecture

a) No. of datasets and categories

According to the Open Data Portal, the platform currently hosts 620,709 datasets across a range of categories, including departments, ministries, and locations throughout various regions of India. Users have the option to download or preview these datasets, which are available in multiple formats such as XML, JSON, XLS, and ODS. The datasets also include additional information such as options to request API access, visualize the data, and details regarding file size, granularity, frequency of downloads and views, as well as publication and last update dates.

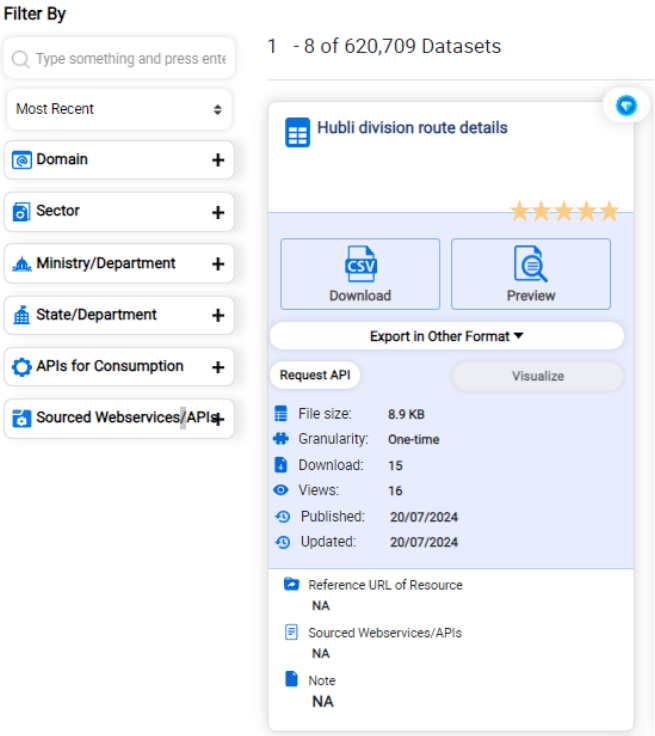


Fig. 10. Creating a Data Viz

b) API and Metadata Availability

Application Programming Interfaces (APIs) facilitate communication between applications, such as the Open Data Portal of India and its users. APIs enable users to efficiently fetch data and visualizations, providing a streamlined method for data retrieval and analysis.

The API option on the portal includes a help wizard to guide users in interacting with the API. The wizard provides pre-filled parameters for dataset queries, including the API key, output format (e.g., JSON, XML, CSV), offset (number of skipped rows), limit (maximum number of rows to retrieve), and filters to apply to the dataset.

We conducted tests with two methods for accessing the API: using the built-in API feature of the Open Data Portal and accessing the API through Postman. Both methods yielded the same results [26].

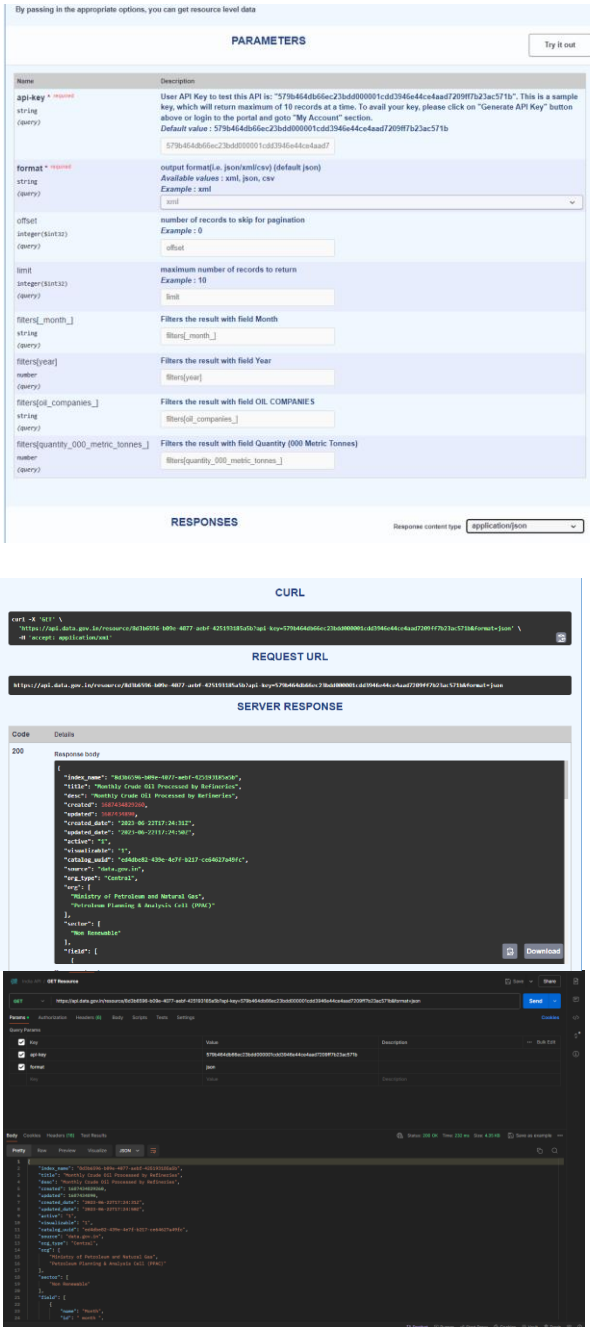


Fig. 11. API testing

c) Data Quality

To assess data quality, we sampled a dataset retrieved from the API and applied several criteria to evaluate its quality. The first metric used was timeliness, which involved verifying whether the dataset was current and updated. We confirmed that the dataset was up-to-date by checking the latest available year, which was correctly labeled as 2024 (see Fig 13).

The second metric, completeness, involved checking for any incomplete values in all attributes, potential imputations, and the presence of duplicates. We used the filter function to identify missing values and discovered that some attributes

were missing, but the quantity was minimal, suggesting these could be null values for specific months.

To examine duplicates, we employed Microsoft Excel's "check for duplicates" function on the CSV-formatted data. This function highlighted possible duplicate values, which we found to be primarily due to total and sub-total amounts. Despite these findings, the dataset still met the completeness criteria.

Month	Year	OIL COMPANIES	Quantity (000 Metric Tonnes)
January	2024	IOCL-BARAUNI,BIHAR	517.57
January	2024	IOCL-KOYALI, GUJARAT	1191.57
January	2024	IOCL-GUWAHATI,ASSAM	95.74
January	2024	BPCL-TOTAL	3476.48
January	2024	MRPL-MANGALORE,KARNATAKA	1527.48
February	2024	IOCL-KOYALI, GUJARAT	1249.94
February	2024	IOCL-PANIPAT, HARYANA	693.25
February	2024	IOCL-GUWAHATI,ASSAM	99.02
February	2024	BPCL-TOTAL	3174.59

Fig. 12. API testing

Month	Year	OIL COMPANIES	Quantity (000 Metric Tonnes)
March	2024	CPCL-MANALI, TAMILNADU	1072.31
January	2024	IOCL-PARADIP,ODISHA	1363.85
January	2024	CPCL-MANALI, TAMILNADU	1001.24
February	2024	CPCL-TOTAL	1013.93
February	2024	CPCL-MANALI, TAMILNADU	1013.93
January	2024	CPCL-TOTAL	1001.24
March	2024	CPCL-TOTAL	1072.31

Fig 13 Sample Dataset Completeness

V. Conclusion and Recommendations

India's Open Government Data represents a prime example of an effectively implemented open data initiative, promoting transparency and driving innovation through its extensive datasets and user-centric design. An evaluation using the DAMA-DMBOK framework reveals that India employs a federated data governance model, supported by critical roles that enhance the data governance framework. The portal's user-friendly features and high data quality align with India's objective of advancing societal progress through technological engagement, enabling citizens to derive value from interoperable data. This approach underscores the potential of open data to drive societal and technological advancements, reflecting India's commitment to leveraging data for transparency and innovation.

The Government of India has made strides in promoting transparency and public engagement through its Open Government Data (OGD) portal. As user numbers continue to grow, it is essential to enhance data accessibility for a diverse group of analysts. This critique proposes two key recommendations: establishing comprehensive tutorials and implementing multilingual capabilities. These measures aim to empower analysts across a broader range of users.

1. Comprehensive Tutorials: To support users, the portal should offer detailed tutorials covering essential topics such as navigating the portal's user interfaces, retrieving data, and applying data analysis and visualization techniques. These tutorials should be available in multiple formats, including video guides and written documentation, to cater to different learning preferences.
2. Multilingual Capabilities: Incorporating multilingual support into the OGD portal will significantly enhance user engagement and accessibility. By offering the portal in major Indian languages such as Hindi, Bengali, Tamil, Telugu, Marathi, Gujarati, Urdu, and Kannada, as well as considering additional languages based on user surveys, the portal can better serve non-English-speaking users. The user interface should be designed to allow easy language switching, and the added languages should be integrated into the portal's operational framework.

Implementing these recommendations will improve the accessibility and usability of the OGD portal, ensuring that it effectively serves all users, regardless of their technical expertise or language background.

References

[1] Sonia, Buchholtz, et al. Big and Open Data in Europe - a Growth Engine or a Missed Opportunity? demosEUROPA, 2014.

[2] Zillner S, Rusitschka S, Skubacz M, (2014) Big Data Story: Demystifying Big Data with Special Focus on and Examples from Industrial Sectors, Whitepaper, Siemens AG, https://www.bibsonomy.org/bibtex/2ec1ca5231e88bd230216bfe5c4cb6a7f/bigfp7

[3] Heimstädt, Maximilian, et al. "Conceptualizing Open Data Ecosystems: A Timeline Analysis of Open Data Development in the UK." Econstor.eu, 2014, www.econstor.eu/handle/10419/96627, http://hdl.handle.net/10419/96627. Accessed 5 Aug. 2024.

[4] Earley, Susan. DAMA-DMBOK : Data Management Body of Knowledge. 2nd ed., Basket Ridge, New Jersey, Technics Publications, 2017.

[5] Petzold, Bryan, et al. "Designing Data Governance That Delivers Value | McKinsey." Www.mckinsey.com, 26 June 2020, www.mckinsey.com/capabilities/mckinsey-digital/our-insights/designing-data-governance-that-delivers-value.

[6] GDPR. "General Data Protection Regulation (GDPR)." General Data Protection Regulation (GDPR), 2018, gdpr-info.eu/.

[7] Máchová, R., & Lnénicka, M. (2017). "Evaluating the Quality of Open Data Portals on the National Level." Journal of theoretical and applied electronic commerce research, 12(1), 21-41.

[8] Kassen, M. (2013). "A promising phenomenon of open data: A case study of the Chicago open data project." Government Information Quarterly, 30(4), 508-513.

[9] "Open Government Data (OGD) Platform India." Open Government Data (OGD) Platform India, data.gov.in.

[10] Kučera, J., Chlapek, D., & Nečaský, M. (2013). "Open Government Data Catalogs: Current Approaches and Quality Perspective." Lecture Notes in Computer Science, 152–166. doi:10.1007/978-3-642-40160-2_13

[11] Pratama, T., & Cahyadi, A. (2020). "Effect of User Interface and User Experience on Application Sales." IOP Conference Series: Materials Science and Engineering, 879. 012133. 10.1088/1757-899X/879/1/012133

[12] Idowu, L. L., Ali, I. I., & Abdullahi, U. G. (2018). "A Model and Architecture for Building a Sustainable National Open Government Data (OGD) Portal," Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance, Galway, Ireland.

[13] Zuiderwijk, Anneke, and Marijn Janssen. "Open Data Policies, Their Implementation and Impact: A Framework for Comparison." Government Information Quarterly, vol. 31, no. 1, Jan. 2014, pp. 17–29, https://doi.org/10.1016/j.giq.2013.04.003.

[14] Cai, L and Zhu, Y 2015 The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14: 2, pp. 1-10, DOI: http://dx.doi.org/10.5334/dsj-2015-002

[15] Emran, N. A. (2015). "Data Completeness Measures." Advances in Intelligent Systems and Computing, 117–130. doi:10.1007/978-3-319-17398-6_11

[16] McGilvray, D. (2010) Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information, Beijing: Publishing House of

Electronics Industry

[17] R. Raghavan, "Keywords: Open Data," Data & Society, Apr. 24, 2024. [Online]. Available: https://datasociety.net/wp-content/uploads/2024/04/Keywords_OpenData_Raghavan_04242024.pdf.

[18] S. Deo and A. Basrur, "Towards Evidence-Based Policymaking: India's Open Data Initiatives," ORF, Apr. 2024. [Online]. Available: https://www.orfonline.org/research/towards-evidence-based-policymaking-indias-open-data-initiatives

[19] Kuldeep Mathur. Public Policy and Politics in India : How Institutions Matter. New Delhi, Oxford University Press, 2016.

[20] Parihar, Isha. “On the Road to Open Data: Glimpses of the Discourse in India.” World Bank Blogs, 17 Feb. 2015, blogs.worldbank.org/en/digital-development/road-open-data-glimpses-discourse-india.

[21] Nithiyanandam, Y, and Satyam Kushwaha. “India’s National Geospatial Policy: Analysing Progress and Charting the Future.” The New Indian Express, The New Indian Express, 28 Dec. 2023, www.newindianexpress.com/web-only/2023/Dec/28/indias-national-geospatial-policy-analysing-progress-and-charting-the-future-2645728.html. Accessed 5 Aug. 2024.

[22] ETtech. “MeitY’s Draft Data Policy Looks to Unlock Govt Data for All.” The Economic Times, Economic Times, 22 Feb. 2022, economictimes.indiatimes.com/tech/tech-bytes/meitys-draft-data-policy-looks-to-unlock-govt-data-for-all/articleshow/89732650.cms. Accessed 5 Aug. 2024.

[23] Risha Chitlangia. “Niti Aayog Plans Data and Analytics Platform.” Hindustan Times, Hindustan Times, 16 Apr. 2022, www.hindustantimes.com/india-news/niti-aayog-plans-data-and-analytics-platform-101650074280003.html. Accessed 5 Aug. 2024.

[24] Jain, Shreya. “All You Need to Know about Data Empowerment and Protection Architecture - IPleaders.” IPleaders, 30 Oct. 2021, blog.ipleaders.in/all-you-need-to-know-about-data-empowerment-and-protection-architecture/. Accessed 5 Aug. 2024.

[25] KPMG India, "India’s open data initiative: Opportunity for states," Apr. 2023. [Online]. Available: https://assets.kpmg.com/content/dam/kpmg/in/pdf/2023/04/india-open-data-initiative-opportunity-for-states.pdf

[26] “Resource | Open Government Data (OGD) Platform India.” Data.gov.in, 21 Jan. 2022, www.data.gov.in/resource/monthly-crude-oil-processed-refineries#api. Accessed 5 Aug. 2024.