From Data to Knowledge: A Modular Framework for AI-Enhanced Enrichment of Open Data Portals

Tobias Siebenlist [*,†], Jennifer Gnyp [*,†]

Rhine-Waal University of Applied Sciences, Kamp-Lintfort, Germany

Abstract

Open data portals face persistent challenges in making public data accessible and usable for diverse user groups, particularly those without technical expertise. This ongoing research paper presents a modular framework that integrates Retrieval-Augmented Generation (RAG) technologies with open data infrastructures to enhance data accessibility, interpretability, and usability. Following a Design Science Research methodology, the framework combines natural language interactions with automated data processing and support for user-driven visualization while implementing a caching mechanism that sustainably enriches open data portals with generated contextual information. Key components include data integration from existing portals, a RAG engine for accurate information retrieval and generation, and a quality assurance system for generated content. The architecture emphasizes modularity through well-defined interfaces, allowing flexible component exchange as technologies evolve, while supporting digital sovereignty through integration options for local, open-source language models. This research aims to contribute to democratizing data access and enhancing the value of open data for citizens, researchers, and public administrators.

Keywords

Open Data, Retrieval-Augmented Generation, Natural Language Processing, Data Accessibility, Artificial Intelligence, Knowledge Management, Information Enrichment, Digital Sovereignty, Data Portals, Data Infrastructure

1. Introduction

Open data has emerged as a key element of modern governance and a driver for innovation, transparency, and civic participation. International and national initiatives have led to an exponential increase in publicly available datasets. The European Data Portal alone now encompasses over 1.5 million datasets from diverse domains such as environment, transportation, health, and finance [1]. This wealth of data holds enormous potential for evidence-based decision-making in politics and administration, economic innovation, and informed citizen participation.

Despite these positive developments, the actual utilization of open data remains below expectations [2, 3]. This is particularly due to substantial barriers that impede access to and interpretation of the data. Empirical studies show that even skilled users often struggle to find relevant datasets, interpret them, and transform them into valuable insights [4, 5]. For laypeople without corresponding data analysis expertise, these hurdles often represent insurmountable obstacles, leading to a digital divide in data utilization [6].

The identified main barriers can be categorized into three areas:

1. **Findability**: Despite standardized metadata, many datasets remain difficult to locate, as search functions are often focused on exact matches and do not consider semantic concepts [7].

[*] *Proceedings EGOV-CeDEM-ePart conference, August 31-September 4, 2025, University for Continuing Education, Krems, Austria.*

[*] Corresponding author.

[†] These authors contributed equally.

tobias.siebenlist@hochschule-rhein-waal.de (T. Siebenlist); jennifer.gnyp@hochschule-rhein-waal.de (J. Gnyp);

0000-0001-9435-910X (T. Siebenlist); 0009-0007-4812-9910 (J. Gnyp)

2. **Interpretability**: Raw data is frequently provided without sufficient context or explanations, which significantly complicates its interpretation [8].
3. **Usability**: The transformation of raw data into usable information requires technical skills and resources that are not available to many potential users [9].

Recent advances in generative artificial intelligence (AI) offer promising approaches to overcome these barriers. In particular, Large Language Models (LLMs) have demonstrated impressive capabilities in natural language processing, contextualizing information, and generating comprehensible explanations [10, 11]. These developments open new possibilities to fundamentally improve interaction with open data and make it accessible to a broader audience.

This paper presents a framework based on the Retrieval-Augmented Generation (RAG) approach, specifically designed to address the aforementioned barriers in open data utilization. RAG combines the strengths of information retrieval and generative AI by first retrieving relevant information from a knowledge base and subsequently using it to generate contextually appropriate responses [12]. In the context of open data, this approach enables natural language interaction with the data, contextualized explanations, and automatic generation of analyses.

The proposed framework goes beyond existing approaches by not only producing temporary analyses and visualizations but also implementing a systematic mechanism for caching and quality assurance of the generated information. This enriches open data ecosystems continuously with valuable metadata and contextualized information, leading to sustainable improvement of data quality and usability. The framework emphasizes modularity through clearly defined interfaces between components, allowing flexible exchange of individual elements as technologies advance, while also addressing digital sovereignty concerns through support for local, open-source language models.

The results are presented in a new portal, however the data remains in the actual open data portals. Direct integration into existing open data portals is currently not pursued, as these have different core functionalities. This new portal aims to provide users with easy access and insights into the data, without the already identified and well-known difficulties in using open data portals and raw datasets. The management of the datasets remains with the classical open data portals, which are designed for this purpose and are already integrated into established processes at the data providers. References to the portals are implemented as in harvesting, so that regular checks for the existence and updates of datasets must be performed.

The research questions addressed by this paper are:

1. How can a modular framework be designed that effectively integrates generative AI and RAG technologies with open data portals?
2. Which architectural components are necessary to generate high-quality, contextualized information from open data and persistently store it?
3. How can interaction with open data be improved through natural language interfaces and automatically generated visualizations for different user groups?

The paper is structured as follows: After an overview of the current state of research on open data portals and AI-supported information systems (Section 2), the methodological approach to developing the framework is explained (Section 3). The main part of the paper (Section 4) describes the architecture and components of the RAG-based framework. An outlook on future research and implementation work is given in Section 5, followed by a concluding summary (Section 6).

2. Related Work

The scientific exploration of open data and its effective utilization encompasses several intertwined research streams. This section provides an overview of selected research areas relevant to the proposed framework: open data portal architectures, generative AI in the context of open data, and existing approaches to improving open data usability.

2.1. Architectures and Technologies of Modern Open Data Portals

Open data portals have evolved from simple data repositories to complex information systems in recent years. Current platforms implement standardized metadata schemas, open APIs, and increasingly also semantic technologies [13, 14].

The most widely used software solutions for open data portals are CKAN, DKAN, Socrata, and OpenDataSoft [15]. CKAN (Comprehensive Knowledge Archive Network) dominates European space and forms the technological foundation for the European Data Portal as well as numerous national portals [16]. These platforms typically offer functions for cataloging, searching, filtering, and in some cases previewing and visualizing data.

Despite these advances, Nikiforova and McBride [17] identify significant deficiencies: Most portals support only rudimentary search functions without semantic or contextual extensions, making it difficult to find relevant datasets. Additionally, advanced functions for data analysis and interpretation, which would be essential for non-technical users, are often missing [18].

More recent research proposes architectural extensions that integrate semantic technologies into open data platforms. Approaches such as the integration of knowledge graphs and linked data aim to improve data connectivity as well as contextual understanding [19]. Kirstein et al. [20] present an enhanced architecture that integrates SPARQL endpoints and semantic annotations into CKAN-based portals. However, this approach requires extensive manual preparation and specific expertise.

Another promising research direction focuses on user experience (UX) and user-centricity in open data portals. Benitez-Paez et al. [21] and Degbelo et al. [22] develop design principles for user-centered open data platforms that particularly consider the needs of non-experts. These works provide valuable insights into user expectations but do not fully address the technical challenges in automated data interpretation and contextualization.

2.2. Generative AI and RAG in the Context of Open Data

Initial research examines the integration of LLMs into data analysis processes. Hong et al. [23] demonstrate how LLMs can act as natural language interfaces to databases (NLIDBs) by translating user queries into SQL and generating textual summaries of the results (data-to-text). This enables non-technical users to access structured data and receive human-readable explanations. Siebenlist [24] sees language models such as ChatGPT as companions for working with open data in order to enable laypeople to work with it. However, these works primarily focus on structured databases and do not sufficiently address the semantic heterogeneity, varying quality, and contextual ambiguity that characterize typical open data portals.

Burgdorf et al. [25] extend the potential of LLMs by proposing a domain-independent data-to-text generation approach that transforms diverse open data tables into natural language descriptions. Their work highlights how generative models can support the accessibility of heterogeneous datasets by producing contextualized and comprehensible outputs, even in low-resource scenarios. A systematic review by Osuji et al. [26] provides a comprehensive overview of data-to-text generation approaches and highlights the increasing role of transformer-based models, in generating fluent and contextualized outputs from structured data. The review identifies core challenges including hallucination, data sparsity, and limited multilingual capabilities, which are critical when applying generative AI to open data portals. This progression from purely generative models to approaches that integrate external knowledge has led to new strategies for improving the factuality and contextual grounding of generated content.

The RAG approach, first introduced by Lewis et al. [12] combines the strengths of information retrieval and generative models by first retrieving relevant information from a knowledge base and then using it as contextual input for generation. RAG is typically structured into three stages: indexing, retrieval, and generation. A recent survey by Gao et al. [27] categorizes existing RAG systems into naive, advanced, and modular architectures, depending on the degree of integration between retrieval and generation components. Naive systems perform retrieval and generation in

separate steps, while modular RAG architectures enable iterative refinement through feedback loops. This typology provides a foundation for analyzing and comparing different RAG implementations.

While RAG architectures have been widely explored in closed-domain or high-resource settings, their application to open data portals remains limited. Existing approaches rarely address the specific characteristics of public data infrastructures, such as heterogeneity, decentralization, and varying data quality. Moreover, there is a lack of frameworks that systematically integrate RAG with open data while also considering practical requirements like quality assurance, accessibility, and sustainable enrichment of the knowledge base.

2.3. Approaches to Improving Open Data Usability

Improving the usability of open data is a central research topic with various methodological approaches. Zuiderwijk et al. [28] identify critical factors for user acceptance of open data and develop a corresponding acceptance model. They emphasize the importance of perceived ease of use and usefulness, as well as the role of support systems.

An important research stream deals with the development of specific tools and infrastructures for open data. Degbelo et al. [22] present an open smart city toolbox that provides visualization and analysis tools for urban data. Oliveira et al. [29] develop a framework for urban analytics that supports the integration and visualization of heterogeneous urban data. These approaches address specific application domains but do not offer a comprehensive solution for the general improvement of open data usability.

Particularly relevant are studies on intermediary systems that act as bridges between data providers and users. These works identify different types of intermediaries and analyze their role in the open data ecosystem. Vetrò et al. [30] explore the role of intermediaries in promoting open government and data strategies.

An emerging research field is the automated generation of metadata and data descriptions. Wu et al. [31] present an NLP-based approach to automatic metadata generation that aims to improve the quality and completeness of data descriptions. Neumaier et al. [32] develop methods for quality assessment of open data metadata and propose automated improvement mechanisms. However, these works primarily address technical data description rather than user-oriented contextualization and interpretation.

2.4. Research Gap and Contribution of the Proposed Framework

The overview of related research indicates that despite significant advances in the areas of open data portal architectures, generative AI, and usability improvement, a notable research gap remains: There is a lack of integrated frameworks that systematically connect generative AI and RAG technologies with open data portals while also addressing sustainable quality improvement through persisting output from language models and enrichment mechanisms.

The framework proposed in this paper addresses this gap by:

1. **Developing a modular architecture** with well-defined interfaces for integrating RAG technologies with existing open data portals, allowing flexible exchange of components as technologies advance.
2. **Implementing mechanisms for quality assurance and sustainable enrichment** of open data through caching and manual quality checking of generated information.
3. **Designing user-oriented interaction concepts** based on natural language that support diverse user groups.

With this, the framework contributes to the current research debate on the democratization of data access and overcoming technical and cognitive barriers in open data utilization. This is also accompanied by digital sovereignty, which is achieved through the use of local, open-source or open-weights language models.

3. Methodological Approach

The development of the proposed RAG-based framework for open data portals follows a Design Science Research (DSR) methodology [33], which is particularly suited for creating and evaluating innovative IT artifacts that address practical problems. DSR emphasizes the iterative design, construction, and evaluation of artifacts while building theoretical knowledge through the design process.

3.1. Design Science Research Framework

Our research follows the established DSR process model with six iterative activities: problem identification, definition of objectives, design and development, demonstration, evaluation, and communication. The framework artifact addresses the identified problem of limited accessibility and usability of open data for diverse user groups, particularly those without technical expertise.

The research activities completed to date include a first analysis of barriers in open data utilization based on related works and informal communication with data stewards from various governmental organizations. Based on these findings, we formulated three specific research questions addressing modularity in AI integration, quality assurance mechanisms for generated content, and user interaction improvements through natural language interfaces. A first draft of the conceptual framework design with modular architecture and preliminary component specifications represents the current development stage.

Planned research activities encompass the prototypical implementation focusing on specific datasets of an initial domain, followed by comprehensive mixed-methods evaluation including technical assessment and user studies with diverse stakeholder groups. The final phase involves communication and dissemination of findings through academic publications and practical implementation guidelines for open data portal operators. Following the DSR methodology and an agile, iterative development approach, not all planned components will be included in the first prototype. These will be added gradually and the prototype will be adapted and the increments evaluated in each case.

3.2. Theoretical Foundations and Design Principles

The conceptualization of the framework is grounded in three complementary theoretical perspectives that inform different aspects of the design. Information Processing Theory [34] guides the design of cognitive load reduction mechanisms through progressive disclosure and contextualization, ensuring that users are not overwhelmed by complex data structures. The Technology Acceptance Model [35] informs user interface design by focusing on perceived usefulness and ease of use as key factors for adoption across different user groups. Boundary Object Theory [36] shapes the design of interfaces that facilitate collaboration between different stakeholder groups, recognizing that the framework must serve as a mediator between technical data providers and diverse user communities.

Based on these theoretical foundations and insights from existing open data portal research, we derived a comprehensive set of design principles that guide the framework development.

The principle of **modularity** ensures component-based architecture with well-defined interfaces, allowing flexible adaptation to different technical environments and requirements. **Interoperability** focuses on standards-compliant integration with existing infrastructures, while scalability addresses the need to support varying data volumes and user loads. **Transparency** emphasizes provenance tracking and explainable AI mechanisms to maintain trust and accountability. **Quality assurance** integrates multi-stage validation and feedback integration processes, while **user orientation** ensures adaptive interfaces for different expertise levels. Finally, **digital sovereignty** supports local deployment options and open-source models to reduce dependencies on commercial providers.

3.3. Implementation and Evaluation Strategy

The current implementation status includes a proof-of-concept prototype that demonstrates the technical feasibility of the core approach. This prototype focuses on the essential data integration and data-to-text component, the RAG pipeline components and data integration mechanisms, using selected datasets from public repositories to validate the fundamental concepts. Natural language querying, automated data interpretation, and basic visualization generation capabilities will build on this and follow up next.

The planned evaluation approach follows a comprehensive mixed-methods strategy that addresses technical, user-centered, and ecosystem perspectives [37]. Technical evaluation encompasses performance benchmarks measuring response times and accuracy, and integration testing with existing open data platforms. User-centered evaluation involves usability studies with representative user groups including citizens without technical background, researchers from various domains, and public administrators responsible for data publication. Ecosystem evaluation assesses the framework's integration capabilities with existing open data infrastructures and measures the effectiveness of metadata enrichment mechanisms over time.

Evaluation metrics are designed to capture both quantitative and qualitative aspects of framework performance, including task completion rates across different user groups, user satisfaction scores measured through standardized questionnaires, improvements in data findability through comparative studies, and the quality of generated explanations assessed by domain experts. This comprehensive evaluation strategy ensures that the framework meets both technical requirements and user needs while demonstrating sustainable value for the open data ecosystem.

4. The RAG-based Framework: Architecture and Components

The developed framework for AI-supported enrichment of open data portals addresses fundamental challenges in data accessibility through a systematic integration of Retrieval-Augmented Generation technologies with existing open data infrastructures. This chapter presents a requirements-driven analysis of the framework, beginning with identified problems and user needs, followed by the architectural solution and preliminary component specifications.

4.1. Requirements Analysis

4.1.1. Functional Requirements

The framework must address three categories of functional requirements derived from the identified barriers in open data utilization. **Data Discovery and Access Requirements** encompass natural language query processing capabilities that allow users to express information needs in everyday language rather than technical query syntax. The system must provide semantic search functionality that goes beyond keyword matching to understand conceptual relationships and user intent. Cross-portal search capabilities are essential to enable unified access to datasets distributed across multiple open data platforms, while contextual filtering mechanisms should help users narrow down results based on temporal, geographical, or domain-specific criteria.

Data Interpretation and Contextualization Requirements focus on automated generation of plain-language explanations for datasets, including their content, structure, and potential applications. The framework must provide foremost a description of the contents of a dataset in natural language to be able to continue working with it and use the content for further processing. It also must provide statistical summaries and key insights that highlight the most relevant aspects of datasets for different user groups. Domain-specific contextualization is crucial to ensure that explanations are meaningful within specific application areas, while the system should maintain clear provenance tracking to establish the origin and reliability of both source data and generated content.

User Interaction and Collaboration Requirements demand intuitive interfaces that accommodate users with varying levels of technical expertise, from citizens seeking information for personal decisions to researchers requiring detailed analytical capabilities. The framework must support progressive disclosure mechanisms that allow users to access information at appropriate levels of detail. Feedback integration capabilities are essential to enable continuous improvement of generated content based on user experiences and domain expert validation.

4.1.2. Non-functional Requirements

Quality and Reliability Requirements emphasize the critical importance of factual accuracy in generated content, given the public nature of open data and its use in decision-making processes. The framework must implement comprehensive validation mechanisms to prevent confabulations and ensure consistency between generated explanations and underlying data. Versioning and audit trail capabilities are essential for maintaining transparency and enabling continuous quality improvement processes. Manual quality assurance by experts before publication for users is crucial here.

Integration and Interoperability Requirements mandate seamless compatibility with existing open data portal architectures, particularly platforms based on CKAN, DKAN, and similar systems. The framework must support standard metadata schemas including DCAT-AP while providing flexible APIs for integration with external tools and applications. Digital sovereignty requirements necessitate support for local deployment options and open-source or open-weights language models to ensure institutional control over data processing.

Performance and Scalability Requirements specify that the system must handle concurrent users efficiently while maintaining response times suitable for interactive use. The architecture must be scalable to accommodate growing data volumes and user bases across different deployment scenarios. Data processing capabilities should support real-time analysis of moderately sized datasets while providing efficient batch processing for larger analytical tasks. These requirements are excluded from the evaluation of the early prototypes.

4.2. Overall Architecture Design

The framework architecture implements a modular, layered design that addresses the identified requirements while ensuring flexibility and maintainability. The **Data Integration Layer** serves as the foundation, establishing connections with external open data sources and processing metadata as well as additional descriptions from those external sources. This layer implements standardized interfaces to common open data platforms and provides access and processing capabilities for various data formats including JSON, CSV, XML, and RDF structures. In the future normalizing heterogeneous data formats for downstream processing will be added if necessary.

The **Core Processing Layer** contains the central RAG engine with its retrieval and generation components, implementing the primary intelligence of the system. This layer orchestrates the interaction between natural language generation and understanding, data retrieval, contextual analysis, and response generation capabilities. Customized system prompts and templates for responses lead to structured descriptions and answers to ensure appropriate output for different user contexts and data types.

The **Quality Assurance and Enrichment Layer** represents a key innovation of the framework, implementing mechanisms for persistent storage and continuous improvement of generated content. This layer performs validation processes, manages feedback integration, and coordinates the enrichment of open data portals with generated metadata and explanations. The caching mechanism ensures that validated insights become permanent additions to the data ecosystem rather than ephemeral responses.

The **Presentation Layer** provides multiple user interface options including conversational interfaces, interactive dashboards, and programmatic APIs. This layer adapts the framework's

capabilities to different user preferences and integration requirements while maintaining consistency in functionality across access methods.

Figure 1 shows a schematic representation of the framework and its components, oriented around the four layers. At the center are the components implemented in a first development phase for natural language processing and persistent storage of the generated texts.

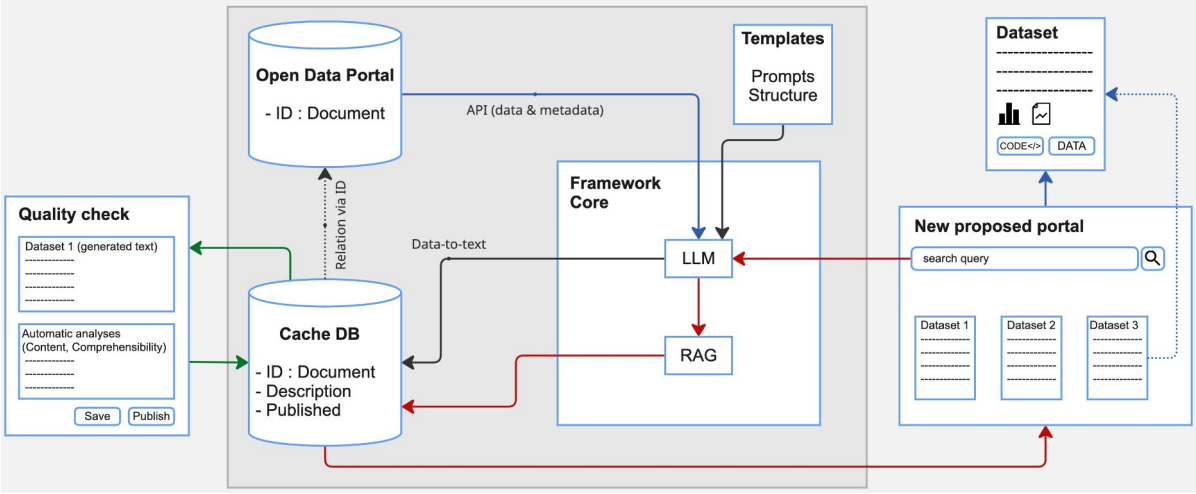

Figure 1: Schematic representation of the framework and its components

The generative AI component transforms retrieved data into accessible, contextually appropriate explanations and analyses using local or API-based Large Language Models. Prompt engineering systems support the generation process by providing structured templates that guide the models in producing contextually appropriate outputs. These templates can be adapted based on user profiles, data characteristics, and intended output formats, offering an initial layer of customization to improve relevance and clarity.

The caching and enrichment component represents the framework's most innovative aspect, transforming temporary AI assistance into permanent improvements to open data ecosystems. Persistent Knowledge Storage maintains validated explanations, analyses, and insights in structured formats that can be queried and reused across sessions and users. This approach ensures that the computational investment in content generation creates lasting value. The quality assurance component implements multiple validation strategies to ensure the reliability of generated content. Automated Consistency Checking compares generated explanations against source data to identify potential confabulations or misrepresentations. Expert Review Workflows facilitate manual validation by domain experts and data stewards, particularly for sensitive or high-impact content.

Multi-modal output generation produces various content types including textual descriptions, statistical summaries, and visualization specifications.

5. Implementation Plans and Research Directions

The proposed RAG-based framework will be implemented in four iterative phases. Phase 1 establishes the core infrastructure, focusing on data integration and RAG-based generation using selected datasets. Phase 2 introduces caching and quality assurance components to support persistent enrichment. In Phase 3, adaptive user interfaces will be developed for different user groups. Finally, Phase 4 integrates the framework with open data portals through automated harvesting and update routines.

The evaluation strategy follows a mixed-methods approach to assess technical, user-centered, and ecosystem dimensions. Technical evaluation includes performance benchmarks and integration

tests. User studies will involve diverse groups—citizens, researchers, and public administrators—to assess usability and perceived value. Long-term ecosystem evaluation will examine the framework's effect on metadata quality and data accessibility.

Future research includes extending the framework to a variety of domains, integrating statistical validation techniques to improve numerical reliability, and experimenting with lightweight community feedback mechanisms. Ethical concerns and digital sovereignty—especially the use of open-source language models and local deployment options—remain guiding principles throughout the development.

The implementation and research program represents a comprehensive approach to validating the framework's practical value while contributing to understanding how AI systems can effectively support democratic data access and public information infrastructures.

6. Conclusion

This ongoing research presents an innovative framework that builds upon Retrieval-Augmented Generation technologies to enable a new type of open data portal addressing fundamental challenges in data accessibility, interpretability, and usability. The framework's modular architecture, combined with sustainable enrichment mechanisms, offers a systematic approach to democratizing access to public data while maintaining quality and reliability standards essential for public information infrastructures.

6.1. Answers to Research Questions

The three research questions formulated in the introduction are directly addressed by the framework design and implementation strategy.

Research Question 1: How can a modular framework be designed that effectively integrates generative AI and RAG technologies with open data portals?

The proposed framework implements a four-layer architecture that separates concerns into data access, processing, validation, and presentation. The Data Integration Layer connects to existing open data platforms (e.g., CKAN, DKAN) and processes heterogeneous data formats through transformation and normalization components. The Core Processing Layer contains the RAG engine, which manages retrieval and generation processes based on user queries and contextual signals. The Quality Assurance and Enrichment Layer validates generated outputs and ensures sustainable reuse through caching and metadata integration. Finally, the Presentation Layer adapts responses into natural language explanations and visualizations for diverse user groups. This modular design enables flexible extension and targeted improvement of individual components as technologies and requirements evolve.

Research Question 2: Which architectural components are necessary to generate high-quality, contextualized information from open data and persistently store it?

The framework identifies five essential components: advanced retrieval mechanisms combining semantic and keyword search, prompt engineering systems that adapt to user contexts and data types, multi-stage validation processes including automated consistency checking and expert review workflows, innovative caching mechanisms that transform temporary assistance into persistent knowledge assets, and metadata enrichment capabilities that enhance the navigability and interpretability of published datasets. The Quality Assurance and Enrichment Layer builds upon the data-to-text paradigm and extends it with mechanisms for validation, persistence, and structured integration, representing a novel implementation focus within the context of open data portals.

Research Question 3: How can interaction with open data be improved through natural language interfaces and automated visualizations for different user groups?

The framework addresses diverse user needs through adaptive presentation mechanisms that accommodate citizens seeking accessible explanations, researchers requiring detailed analytical capabilities, and administrators supporting transparency objectives. Natural language processing

enables intuitive query expression without technical expertise, while progressive disclosure mechanisms provide information at appropriate complexity levels. The framework generates multi-modal outputs including textual descriptions, statistical summaries, and visualization specifications, ensuring that different learning preferences and information needs are addressed effectively.

6.2. Conceptual and Practical Contributions

The framework makes significant contributions across conceptual, methodological, and practical dimensions. Conceptually, the integration of RAG technologies with persistent caching mechanisms represents a novel approach that goes beyond temporary assistance to create lasting value for open data ecosystems. The modular architecture explicitly addresses digital sovereignty concerns while supporting technological evolution, providing a foundation for sustainable AI integration in public sector applications.

Methodologically, the Design Science Research approach combines theoretical foundations from Information Processing Theory, Technology Acceptance Model, and Boundary Object Theory with systematic requirements analysis and user-centered design principles. The comprehensive evaluation strategy addresses technical performance, user acceptance, and ecosystem impact, providing a robust framework for validating AI-enhanced public information systems.

Practically, the conceptual framework demonstrates initial technical feasibility through a proof-of-concept prototype that validates the core RAG pipeline functionality while addressing real barriers identified in open data utilization research. The focus on an initial domain provides a concrete validation scenario with clear public value and available domain expertise for quality assurance.

6.3. Significance for Open Data Ecosystems and Future Directions

The framework has the potential to transform open data ecosystems by democratizing access to public information and enhancing its value for diverse stakeholders. Through natural language interactions and automatically generated and quality-assured explanations, the framework can significantly reduce barriers that currently prevent many citizens from effectively utilizing open data for informed decision-making. The sustainable enrichment mechanism ensures that computational investments in content generation create lasting improvements to data infrastructure rather than ephemeral assistance.

Future research directions focus on validating and extending the current prototype. The immediate next steps include expanding the proof-of-concept to additional datasets, implementing comprehensive user studies with citizens and administrators, and developing robust quality assurance mechanisms for production deployment. Longer-term extensions will explore domain-specific adaptations and enhanced caching mechanisms based on lessons learned from the initial implementation.

The systematic implementation and evaluation strategy outlined in this research provides a roadmap for validating the framework's practical value while contributing to broader understanding of how AI systems can effectively support democratic data access and evidence-based decision-making. As open data continues to grow in volume and complexity, frameworks like this become increasingly essential for realizing the democratic and innovative potential of public information resources.

This ongoing research establishes a foundation for intelligent open data portals that can bridge the gap between technical data provision and meaningful public engagement, ultimately supporting more informed citizenship and evidence-based governance in our increasingly data-driven society.

Declaration on Generative AI

During the preparation of this work, the authors used Claude 3.7 Sonnet for: Grammar and spelling check, improvement of writing style, shortening of the text and a peer review simulation. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] European Data Portal. (2023). Open Data Maturity Report 2023. URL: https://data.europa.eu/en/publications/open-data-maturity/2023

[2] Safarov, I., Meijer, A., & Grimmelikhuijsen, S. (2017). Utilization of open government data: A systematic literature review of types, conditions, effects and users. Information Polity, 22(1), 1-24.

[3] Crusoe, J., & Ahlin, K. (2019). Users' activities for using open government data–a process framework. Transforming Government: People, Process and Policy, 13(3/4), 213-236.

[4] Degbelo, A., Granell, C., Trilles, S., Bhattacharya, D., Casteleyn, S., & Kray, C. (2016). Opening up smart cities: citizen-centric challenges and opportunities from GIScience. ISPRS International Journal of Geo-Information, 5(2), 16.

[5] Benitez-Paez, F., Degbelo, A., Trilles, S., & Huerta, J. (2018). Roadblocks hindering the reuse of open geodata in Colombia and Spain: A data user's perspective. ISPRS International Journal of Geo-Information, 7(1), 6.

[6] Zuiderwijk, A., Shinde, R., & Janssen, M. (2018). Investigating the attainment of open government data objectives: Is there a mismatch between objectives and results? International Review of Administrative Sciences, 85(4), 645-672.

[7] Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated quality assessment of metadata across open data portals. Journal of Data and Information Quality (JDIQ), 8(1), 1-29.

[8] Martin, S., Foulonneau, M., Turki, S., & Ihadjadene, M. (2013). Risk analysis to overcome barriers to Open Data. Electronic Journal of e-Government, 11(2), 348-359.

[9] Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information Systems Management, 29(4), 258-268.

[10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[11] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[12] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

[13] Kučera, J., Chlapek, D., & Nečaský, M. (2013). Open government data catalogs: Current approaches and quality perspective. In Technology-Enabled Innovation for Democracy, Government and Governance (pp. 152-166). Springer.

[14] Máchová, R., & Lnenicka, 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.

[15] Kubler, S., Robert, J., Neumaier, S., Umbrich, J., & Le Traon, Y. (2018). Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. Government Information Quarterly, 35(1), 13-29.

[16] Attard, J., Orlandi, F., & Auer, S. (2016). Data driven governments: Creating value through open government data. In Transactions on Large-Scale Data-and Knowledge-Centered Systems XXVII (pp. 84-110). Springer.

[17] Nikiforova, A., & McBride, K. (2021). Open government data portal usability: A user-centred usability analysis of 41 open government data portals. Telematics and Informatics, 58, 101539.

[18] Crusoe, J., Simonofski, A., Clarinval, A., & Gebka, E. (2019). The impact of impediments on open government data use: Insights from users. In 13th International Conference on Research Challenges in Information Science (pp. 1-12).
[19] Beno, M., Figl, K., Umbrich, J., & Polleres, A. (2017). Open data hopes and fears: Determining the barriers of open data. In 2017 Conference for E-Democracy and Open Government (CeDEM) (pp. 69-81). IEEE.
[20] Kirstein, F., Dittwald, B., Dutkowski, S., Glikman, Y., Schimmler, S., & Hauswirth, M. (2019). Linked data in the European data portal: A comprehensive platform for applying DCAT-AP. In International Conference on Electronic Government (pp. 192-204). Springer.
[21] Benitez-Paez, F., Comber, A., Trilles, S., & Huerta, J. (2018). Creating a conceptual framework to improve the re-usability of open geographic data in cities. Transactions in GIS, 22(3), 806-822.
[22] Degbelo, A., Trilles, S., Kray, C., Bhattacharya, D., Schiestel, N., Wissing, J., & Granell, C. (2016). Designing semantic Application Programming Interfaces for open government data. JeDEM-eJournal of eDemocracy and Open Government, 8(2), 21-58.
[23] Hong, Z., Yuan, Z., Zhang, Q., Chen, H., Dong, J., Huang, F. & Huang, X. (2024) Next-Generation Database Interfaces: A Survey of LLM-based Text-to-SQL. arXiv preprint arXiv:2406.08426.
[24] Siebenlist, T. (2023). Approaches towards using ChatGPT as an open data companion. In Proceedings of the 24th Annual International Conference on Digital Government Research (dg.o '23). Association for Computing Machinery, New York, NY, USA, 674–675.
[25] Burgdorf, A., Barkmann, M., Pomp, A., & Meisen, T. (2022). Domain-independent Data-to-Text Generation for Open Data. In DATA, 95-106.
[26] Osuji, C. C., Ferreira, T. C., & Davis, B. (2024). A systematic review of data-to-text nlg. arXiv preprint arXiv:2402.08496.
[27] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2(1).
[28] Zuiderwijk, A., Janssen, M., & Dwivedi, Y. K. (2015). Acceptance and use predictors of open data technologies: Drawing upon the unified theory of acceptance and use of technology. Government information quarterly, 32(4), 429-440.
[29] Oliveira, M. I. S., Lima, G. D. F. B., & Lóscio, B. F. (2019). Investigations into data ecosystems: a systematic mapping study. Knowledge and Information Systems, 61(2), 589-630.
[30] Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. Government Information Quarterly, 33(2), 325-337.
[31] Wu, J., Ota, K., Dong, M., & Li, J. (2018). Big data analysis-based security situational awareness for smart grid. IEEE Transactions on Big Data, 4(3), 408-417.
[32] Neumaier, S., Umbrich, J., & Polleres, A. (2017). Lifting data portals to the web of data. WWW2017 Workshop on Linked Data on the Web (LDOW2017).
[33] Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS quarterly, 75-105.
[34] Simon, H. A. (1978). Information-processing theory of human problem solving (pp. 271-295). Hillsdale, NJ: Erlbaum.
[35] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly, 319-340.
[36] Star, S. L., & Griesemer, J. R. (1989). Institutional ecology,translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. Social studies of science, 19(3), 387-420.
[37] Creswell, J. W., & Plano Clark, V. L. (2017). Designing and conducting mixed methods research (3rd ed.). SAGE Publications.