.

Semantic Label Property Graph Ontologies: A Methodology for Enhanced Data Management in Digital Libraries

Eleonora Bernasconi[1,*], Stefano Ferilli[1]

[1]University of Bari Aldo Moro, Department of Computer Science, Via Orabona 4, Bari, Italy

Abstract

Ontologies are crucial for managing and integrating diverse datasets in digital libraries, where data heterogeneity poses ongoing challenges. This paper presents a novel framework specifically designed to address the unique needs of digital libraries using Semantic Label Property Graphs. Our methodology aligns with semantic web standards, offering a sophisticated approach to data management that enhances integration, querying, and visualization of complex datasets. The proposed framework supports automated ontology generation, advanced semantic integration, and seamless visualization, leveraging the structural efficiency of Property Graphs with semantic annotations to optimize resource discovery, management, and retrieval. We detail the architecture and core functionalities of the framework, demonstrating its adaptability in managing complex ontologies and improving workflows for researchers and practitioners. Empirical evaluations reveal significant performance improvements in data management and linked data integration, underscoring the framework's potential to streamline workflows and enhance semantic interoperability. This innovative approach addresses the evolving challenges of large-scale data management, positioning the framework as a valuable tool for the future of digital libraries.

Keywords

Digital Libraries, Semantic Ontologies, Label Property Graph, Schema Management, Semantic Web, Artificial Intelligence, Large Language Models

CEUR-WS.org/Vol-3865/11_paper.pdf

1. Introduction

Digital libraries are essential platforms for storing, managing, and providing access to vast collections of cultural, historical, and academic resources. These collections are diverse, encompassing textual documents, multimedia, complex metadata, and intricate relationships between various entities such as authors, works, genres, and historical events. As these libraries continue to grow in size and complexity, traditional data management systems increasingly struggle to handle the heterogeneity, scale, and interconnected nature of the data, leading to significant challenges in data integration, retrieval, and usability [1].

Digital libraries face several critical challenges that impede their ability to effectively manage, integrate, and provide access to their vast collections:

1. **Data heterogeneity and integration:** Digital libraries typically aggregate data from a multitude of sources, each employing distinct metadata standards and formats, such as Dublin Core and MARC [2]. This diversity creates significant obstacles to data integration, as the lack of a unified framework complicates efforts to harmonize these varied datasets. Consequently, seamless access and resource discovery are often hindered, affecting the overall usability of digital library systems.

3nd Italian Workshop on Artificial Intelligence for Cultural Heritage (IAI4CH 2024, https://ai4ch.di.unito.it/), co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024). 26-28 November 2024, Bolzano, Italy

[*]Corresponding author.

[†]These authors contributed equally.

eleonora.bernasconi@uniba.it (E. Bernasconi); stefano.ferilli@uniba.it (S. Ferilli)
0000-0003-3142-3084 (E. Bernasconi); 0000-0003-1118-0601 (S. Ferilli)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

2. **Complex relationships and semantic enrichment:** The relationships within digital library collections—such as the impact of an author on a literary genre or the historical significance of a work—are complex and multifaceted. Traditional keyword-based search systems frequently fail to capture these nuanced connections, leading to a superficial exploration of the data [3]. There is a growing need for advanced methods that can identify, represent, and leverage these intricate relationships, enriching the user's ability to explore and discover information in more meaningful ways.
3. **Scalability and performance:** As digital libraries expand their collections, maintaining efficient performance in data retrieval and querying becomes increasingly challenging. The sheer volume and complexity of data require sophisticated storage solutions and indexing mechanisms that can handle large-scale, semantically rich queries. Without these, performance bottlenecks can severely limit the practical utility of digital libraries, particularly when dealing with extensive datasets.
4. **Interoperability and data reusability:** The ability to easily share and reuse data across various platforms is essential, especially in collaborative settings involving multiple institutions, archives, and research bodies [4]. However, the absence of interoperable standards poses significant barriers to data exchange, reducing the potential of digital libraries to function as interconnected and accessible information hubs. Overcoming these interoperability challenges is crucial to enhancing the collective value and accessibility of digital library resources [5].

1.1. Role of semantic ontologies and graph databases in Digital Libraries

To address these challenges, the integration of Semantic Web technologies and graph-based data models has emerged as a critical area of research [6]. The Semantic Web, built on standards such as RDF (Resource Description Framework) and OWL (Web Ontology Language), aims to create a web of data that is both machine-readable and semantically meaningful. These technologies allow digital libraries to represent complex relationships between data points, enhancing searchability, interoperability, and the overall user experience [7].

Graph databases, particularly those based on the Label Property Graph (LPG) model, offer a complementary approach by efficiently managing the interconnected nature of digital library data. LPGs provide a flexible structure for modeling entities and their relationships, supporting advanced queries and visualizations that are crucial for exploring complex datasets. The hybrid approach of combining Semantic Web technologies with graph databases promises to overcome current limitations, offering a powerful solution for managing, integrating, and retrieving semantically enriched data in digital libraries [8].

2. Related work

Hybrid approaches that integrate Semantic Web technologies with graph-based data models have become increasingly relevant in the digital library domain, enhancing data management, integration, and retrieval capabilities. Previous research has demonstrated the advantages of combining Semantic Web standards, such as RDF (Resource Description Framework), with graph-based models like Label Property Graphs (LPGs) [9]. This synergy significantly improves data interoperability [10] and querying functionalities [11], which are crucial for developing adaptable and efficient data management systems [12] for digital libraries and humanities research [13].

Nguyen et al. [14] introduced the Singleton Property Graph to add a semantic web abstraction layer to graph databases, enabling more sophisticated data interactions. Angles et al. [15] explored methods for mapping RDF to property graph databases, enhancing the flexibility and utility of hybrid systems. Hristovski et al. [16] demonstrated the practical application of these approaches in knowledge discovery by implementing semantic literature-based discovery using graph databases.

Graph databases such as Neo4j[1], ArangoDB[2], and Amazon Neptune[3] have incorporated RDF and LPG capabilities, supporting the semantic integration of diverse datasets [17]. These platforms allow digital libraries to capture the semantic relationships between resources [18], enhancing the searchability and discoverability of content [19]. However, existing solutions often face challenges related to scalability, standardization, and performance when handling large-scale, semantically enriched data, which can limit their effectiveness in practical digital library applications.

2.1. Gaps in existing research

Despite the progress made in integrating Semantic Web technologies and graph databases in digital libraries, several critical gaps remain unaddressed. One of the key challenges is the limited adoption of hybrid models that combine RDF (Resource Description Framework) and Label Property Graphs (LPGs). Although these hybrid models offer significant advantages in terms of data integration and semantic enrichment, their implementation is still relatively rare in digital libraries. This limited adoption is largely due to technical challenges, such as the complexity of integrating these technologies [20], and the lack of standardized tools [21] and frameworks that can facilitate their widespread use.

Scalability also remains a significant issue in the current landscape [22]. As digital libraries continue to grow, both in terms of the size and complexity of their datasets, existing solutions often struggle with performance bottlenecks, particularly in areas like querying and data storage. These bottlenecks can significantly hinder the practical implementation of RDF and LPG-based systems in large-scale digital libraries, limiting their ability to efficiently handle the large volumes of interconnected data that these institutions manage.

Moreover, there is a pressing need for more robust semantic enrichment tools that can automatically generate and manage semantic annotations within graph-based models. The current lack of sophisticated tools in this area hampers the ability to create richer data connections and enhance user interactions with library content. Such tools are essential for facilitating deeper exploration of digital library collections, allowing users to navigate complex relationships and discover new connections within the data.

This study is motivated by the need to develop and refine methodologies that leverage the strengths of Semantic Web technologies and graph databases to address the ongoing challenges faced by digital libraries. By creating a framework that integrates SLPGs, this research aims to favour the management of complex datasets, improve semantic interoperability, and optimize the user experience for researchers, students, and practitioners in the field of cultural heritage and beyond.

3. Methodology

This research utilizes the OntoBuilder tool [23] to construct and automatically populate an ontology tailored for the digital library domain. The methodology consists of several key phases:

The process begins with constructing the SLPG ontology schema, where entities (e.g., books, authors, genres) and their relationships (e.g., authoring, publication, thematic links) are defined. Each node represents an entity, while edges define relationships, and both can have associated properties. This structure is flexible, enabling the modeling of complex and heterogeneous data in digital libraries.

Semantic web standards such as RDF, RDF Schema (RDFS), and OWL (Web Ontology Language) are integrated within the graph-based framework. This allows the representation of rich metadata, enhancing the system's ability to understand and manage relationships between entities. RDF triples are mapped into the LPG structure, enriching nodes and edges with semantic meaning that aligns with global standards for data interchange and reuse.

Concepts and relationships are extracted from structured and unstructured data sources within the digital library. Using predefined rules and algorithms, entities like book titles, author names, and

[1] https://neo4j.com

[2] https://arangodb.com

[3] https://aws.amazon.com/it/neptune

publication dates are automatically identified and incorporated into the SLPG. This automation reduces the manual effort required to build ontologies while ensuring consistency and accuracy in the ontology creation process.

Once the SLPG is constructed, it supports advanced querying capabilities through graph traversal techniques combined with semantic reasoning. Users can query the ontology to retrieve complex information, such as identifying all works by a particular author within a specified timeframe or discovering thematic connections across different collections. The LPG's inherent flexibility in managing relationships allows for efficient querying, while semantic annotations ensure the relevance and precision of search results.

The methodology supports continuous evolution and scaling of the ontology. As new data is ingested into the digital library, the ontology can be updated dynamically, allowing for the seamless integration of additional metadata and relationships. This adaptability ensures that the ontology remains relevant as the digital library grows and evolves over time.

By adhering to RDF and other semantic web standards, the SLPG-based ontologies ensure that data can be shared and integrated across different systems. This interoperability is crucial for digital libraries that rely on external data sources, such as linked open data initiatives, to enrich their collections. The ability to interlink resources across various libraries and cultural heritage institutions enhances the discoverability and usability of digital assets.

The final SLPG ontology can be exported into RDF format for integration with other semantic web technologies or maintained in the LPG format (e.g., Neo4j) to optimize graph-specific features and performance. This flexibility allows digital libraries to choose the most appropriate format based on their specific needs, balancing semantic richness with operational efficiency.

This methodology provides a robust and flexible approach for enhancing data management within digital libraries, offering improvements in metadata organization, search functionality, and interoperability. By integrating SLPGs with semantic web standards, it enables more efficient handling of complex datasets, addressing the challenges of scalability, data integration, and advanced querying in modern digital libraries.

3.1. Use case

In the evolving landscape of digital libraries, effective organization and retrieval of information hinge on the development of robust ontologies. In this section, we outline a systematic approach to constructing an ontology that captures essential entities within a digital library framework, such as books, authors, topics, publishers, locations, and contributors. Using existing semantic resources, this process not only enhances the richness of the data, but also facilitates better user interactions and knowledge discovery. A key feature of our approach is the automatic population of the ontology from textual documents. This allows for the seamless extraction and categorization of information, ensuring that the ontology remains up-to-date and reflective of the latest publications and research. The following steps detail the approach, beginning with the integration of DBpedia as a foundational reference for creating our ontology.

3.1.1. Using DBpedia for guided creation

The first step is to use guided creation using DBpedia as a reference. For instance, we select the class `Book` as the base entity. From DBpedia, we incorporate several properties that describe books, including:

- `comment`: A brief description or annotation of the book.
- `author`: The name of the person who wrote the book.
- `subject`: The main topics or themes addressed in the book.
- `publisher`: The company or organization responsible for publishing the book.
- `publicationDate`: The date when the book was first published.
- `isbn`: The International Standard Book Number, a unique identifier for books.

Additionally, we define custom properties specific to the digital library, such as:

- `bookStatus`: Indicates the current availability status of the book (e.g., available, checked out).
- `editingContributors`: Names of the individuals involved in the editing process of the book.

Once the class `Book` is created with these properties, it is named and linked to both the DBpedia resources and the internal digital library resources. DBpedia properties maintain links to external URIs, while custom properties are linked to specific URIs representing the internal reference environment of the digital library.

3.1.2. Manual creation example

To describe the manual creation process, consider the book "The Name of the Rose" by Umberto Eco. The following details how an instance of the class `Book` would be created and populated:

Creation of the class `Book`

- Name: `Book`
- DBpedia Properties:
  - `comment`: A description of the book.
  - `author`: Umberto Eco.
  - `subject`: Historical novel, Mystery.
  - `publisher`: Bompiani.
  - `publicationDate`: 1980.
  - `isbn`: 978-88-452-1523-5.
- Custom Properties:
  - `bookStatus`: Available.
  - `editingContributors`: Maria Bonfantini.

Manual insertion of properties

- `comment`: "The Name of the Rose is a historical novel and mystery written by Umberto Eco, set in a Benedictine monastery in the 14th century."
- `author`: Umberto Eco.
- `subject`: Historical novel, Mystery.
- `publisher`: Bompiani.
- `publicationDate`: 1980.
- `isbn`: 978-88-452-1523-5.
- `bookStatus`: Available.
- `editingContributors`: Maria Bonfantini.

3.1.3. Automatic creation example

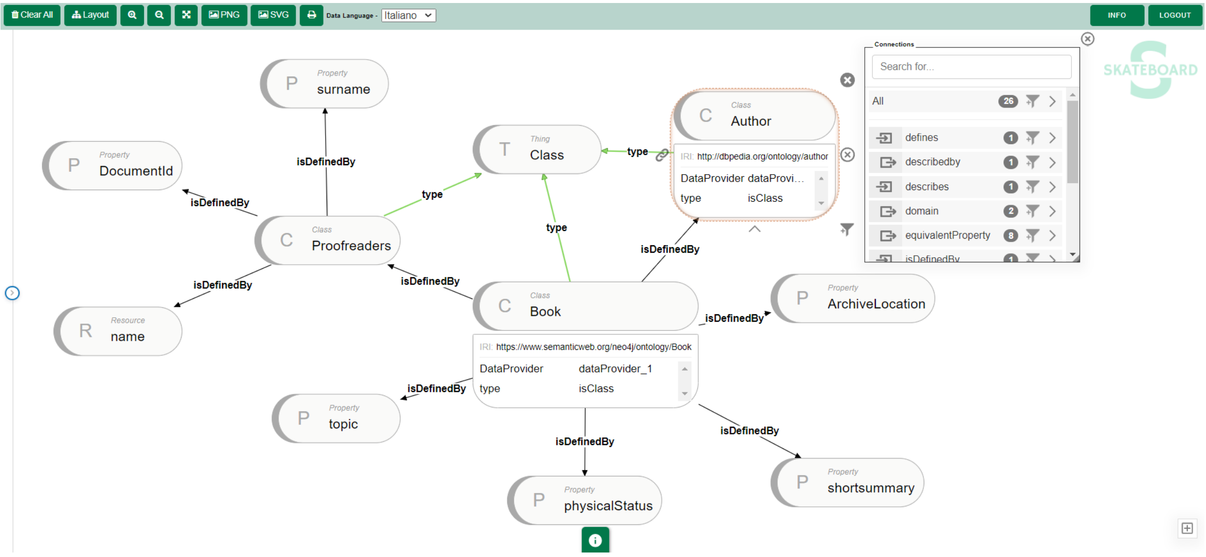
During the ontology population phase, instances of classes can be added either manually or automatically. In the manual mode, users input values for class properties, as demonstrated above. In the automatic mode, an advanced language model is employed to process text, identifying and categorizing entities and properties. For example, given the text:

> "The Name of the Rose, written by Umberto Eco, is a historical novel published by Bompiani in 1980. The book deals with themes such as faith, truth, and heresy and is available in our library."

In automatic insertion mode, the language model parses the text, identifying "The Name of the Rose" as a book, extracting the author "Umberto Eco," the publisher "Bompiani," the publication date "1980," and other relevant information. This information is then used to automatically create class instances in the ontology.

3.1.4. Linking resources and visualization

Once properties are associated with DBpedia and custom resources, the ontology is enriched with semantic data. This enriched data can be visualized using graph-based tools [24]. For example, the SKATEBOARD interface [25] can be used to visualise the connections between books, authors, and other entities, allowing dynamic exploration and further semantic augmentation of the ontology (see Figure 1).



Data Language - Italiano

IRI: http://dbpedia.org/ontology/author
DataProvider dataProvi...
type isClass

IRI: https://www.semanticweb.org/neo4j/ontology/Book
DataProvider dataProvider_1
type isClass

Figure 1: SKATEBOARD Interface showing ontological relationships.

4. Conclusion

This study highlights the importance of integrating Semantic Web technologies and graph databases, specifically through Semantic Label Property Graphs, to enhance data management in digital libraries. By addressing challenges such as data heterogeneity, complex relationships, scalability, and interoperability, our methodology demonstrates improved organization and retrieval of diverse datasets.

An empirical evaluation of the methodology applied, as detailed in [23], reveals that the proposed methodology significantly enhances ontology management and linked data integration. The results indicate strong user satisfaction, particularly in areas like integration compatibility and ontology representation accuracy, underscoring the framework's effectiveness in real-world applications.

The case study illustrates how our framework enables efficient ontology creation and automated metadata integration, fostering richer user interactions. The ability to dynamically update ontologies ensures ongoing relevance and usability.

Future research can focus on enhancing semantic enrichment tools and addressing scalability challenges to further improve user experience and collaboration among institutions. Overall, our work contributes both theoretical insights and practical methodologies for optimizing digital library management, positioning libraries to better serve cultural and academic resources.

Acknowledgments

This research was partially supported by projects CHANGES "Cultural Heritage Active Innovation for Sustainable Society" (PE00000020), Spoke 3 "Digital Libraries, Archives and Philology" and FAIR "Future AI Research" (PE00000013), spoke 6 "Symbiotic AI", funded by the Italian Ministry of University and Research NRRP initiatives under the NextGenerationEU program.

References

[1] O. Diseiye, S. E. Ukubeyinje, B. D. Oladokun, V. V. Kakwagh, Emerging technologies: Leveraging digital literacy for self-sufficiency among library professionals, Metaverse Basic and Applied Research 3 (2024) 59–59.

[2] J. V. Krefft, Z. Du, R. Bakker, From dublin core to marc-crosswalking etd metadata from digital commons to the library catalog (2020).

[3] H. Lee, S. Yoon, Z. Park, "semantic" in a digital curation model, Journal of Data and Information Science 5 (2020) 81–92.

[4] K. Shahzad, S. A. Khan, Factors affecting the adoption of integrated semantic digital libraries (sdls): a systematic review, Library Hi Tech 41 (2023) 386–412.

[5] P. Fafalios, K. Petrakis, G. Samaritakis, K. Doerr, A. Kritsotaki, Y. Tzitzikas, M. Doerr, Fast cat: collaborative data entry and curation for semantic interoperability in digital humanities, Journal on Computing and Cultural Heritage (JOCCH) 14 (2021) 1–20.

[6] B. O'Neill, L. Stapleton, Digital cultural heritage standards: from silo to semantic web, Ai & Society 37 (2022) 891–903.

[7] S. Auer, A. Oelen, M. Haris, M. Stocker, J. D'Souza, K. E. Farfar, L. Vogt, M. Prinz, V. Wiens, M. Y. Jaradeh, Improving access to scientific literature with knowledge graphs, Bibliothek Forschung und Praxis 44 (2020) 516–529.

[8] E. Bernasconi, M. Ceriani, S. Ferilli, Lpg semantic ontologies: A tool for interoperable schema creation and management, Information 15 (2024) 565.

[9] H. Moon, Z. Zhao, J. Choi, S. Han, A novel property graph model for knowledge representation on the web, International Journal of Engineering & Technology 7 (2018) 187–190.

[10] S. Ferilli, R. Basili, F. Esposito, Hybrid approaches to semantic data management, Journal of Data Semantics 12 (2023) 123–145.

[11] T. Liebig, M. Opitz, V. Vialard, M. Wenzel, Scalable no-code knowledge graph exploration and querying with semspect., in: SEMANTiCS (Posters & Demos), 2023.

[12] S. Ferilli, E. Bernasconi, D. Di Pierro, D. Redavid, A graph db-based solution for semantic technologies in the future internet, Future Internet 15 (2023) 345.

[13] D. Di Pierro, S. Ferilli, D. Redavid, Lpg-based knowledge graphs: A survey, a proposal and current trends, Information 14 (2023) 154.

[14] V. Nguyen, H. Y. Yip, H. Thakkar, Q. Li, E. Bolton, O. Bodenreider, Singleton property graph: Adding a semantic web abstraction layer to graph databases., BlockSW/CKG@ ISWC 2599 (2019) 1–13.

[15] R. Angles, H. Thakkar, D. Tomaszuk, Mapping rdf databases to property graph databases, IEEE Access 8 (2020) 86091–86110.

[16] D. Hristovski, A. Kastrin, D. Dinevski, T. C. Rindflesch, Towards implementing semantic literature-based discovery with a graph database, DBKDA 2015 (2015) 190.

[17] D. Fernandes, J. Bernardino, et al., Graph databases comparison: Allegrograph, arangodb, infinitegraph, neo4j, and orientdb., Data 10 (2018) 0006910203730380.

[18] J. J. Miller, Graph database applications and concepts with neo4j, in: Proceedings of the southern association for information systems conference, Atlanta, GA, USA, volume 2324, 2013, pp. 141–147.

[19] J. Cheng, Graph feature management: Impact, challenges and opportunities, in: Proceedings of the 6th Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA), 2023, pp. 1–1.

[20] S. Khayatbashi, S. Ferrada, O. Hartig, Converting property graphs to rdf: a preliminary study of the practical impact of different mappings, in: Proceedings of the 5th ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA), 2022, pp. 1–9.

[21] H. V. Thakker, On Supporting Interoperability between RDF and Property Graph Databases, Ph.D. thesis, Universitäts-und Landesbibliothek Bonn, 2021.

[22] E. Iglesias, M.-E. Vidal, D. Collarana, D. Chaves-Fraga, Empowering the sdm-rdfizer tool for scaling

up to complex knowledge graph creation pipelines 1, Semantic Web (2024) 1–28.

[23] E. Bernasconi, M. Ceriani, S. Ferilli, Lpg semantic ontologies: A tool for interoperable schema creation and management, Information 15 (2024). URL: https://www.mdpi.com/2078-2489/15/9/565. doi:10.3390/info15090565.

[24] E. Bernasconi, M. Ceriani, D. Di Pierro, S. Ferilli, D. Redavid, Linked data interfaces: A survey, Information 14 (2023). URL: https://www.mdpi.com/2078-2489/14/9/483. doi:10.3390/info14090483.

[25] E. Bernasconi, D. Di Pierro, D. Redavid, S. Ferilli, Skateboard: Semantic knowledge advanced tool for extraction, browsing, organisation, annotation, retrieval, and discovery, Applied Sciences 13 (2023). URL: https://www.mdpi.com/2076-3417/13/21/11782. doi:10.3390/app132111782.