Exploring multiple knowledge graphs in Digital Humanities

Alberto Morvillo[1,*], Massimo Mecella[1]

[1]Sapienza Università di Roma, Dipartimento di Ingegneria informatica, automatica e gestionale Antonio Ruberti (DIAG), via Ariosto, 25, 00185 Rome, Italy

Abstract

In the field of digital humanities, the mode of information consumption constitutes a fundamental factor in the quality of research. The structuring of data into knowledge graphs provides a valuable tool for navigating concepts and exploring new ideas. However, the information are often spread across multiple sources with different data organizations (schema, taxonomy, etc.) if not, in some cases, even different data formats. These differences between sources generate a fragmentation of knowledge, and, in order to obtain an effective quality consultation, they have to be explored alltogether. In this paper, an approach for exploring multiple knowledge graphs using visual representation is discussed.

Keywords

Knowledge graphs, Digital humanities, Information integration

1. Introduction

In the field of digital humanities, the use of knowledge graphs to represent data, structurally and graphically, to both address the data source and data consumption has already been demonstrated [1]. However, as these are often unrelated systems, there is a fragmentation of knowledge mostly caused by the use of different schemas or semantics in each domain, although data format standards are often in common, and the integration of multiple data sources is a topic still widely discussed today.

Another important topic in the field of content exploration through knowledge graphs, is the visual representation of the knowledge in the graphs, which is still an open field. Although there are known search tools based on semantic engines, these are often generic systems, leaving room for research on potential methodologies for specialized visualization and navigation.

A final key point consists of cross-references, i.e. contents that can be referred to each other, and implicit concepts, i.e., concepts which are taken for granted. Current Large Language Models (LLMs) proved to address those points, shifting the boundary from how concepts are expressed to the quantity of information (number of tokens required to process the source). However, the result of this extraction may need to be expanded with knowledge from other existing sources, further increasing the fragmentation and complicating the exploration.

The contribution of this article consists in a proposal for exploring multiple knowledge graphs using visual navigation.

The following of this paper is as it follows. After considering relevant work in Section 2, in Section 3, the use of knowledge graphs for digital humanities is introduced, with its relative advantages and challenges; Section 4 describes NAVIGO, the proposed framework for exploring multiple knowledge graphs; Section 5 analyzes a case study in the field of archaeology. Finally Section 6 presents concluding remarks and possible future directions.

Proceedings of the Joint Ontology Workshops (JOWO) - Episode X: The Tukker Zomer of Ontology, and satellite events co-located with the 14th International Conference on Formal Ontology in Information Systems (FOIS 2024), July 15-19, 2024, Enschede, The Netherlands.

*Corresponding author.

morvillo@diag.uniroma1.it (A. Morvillo); mecella@diag.uniroma1.it (M. Mecella)

0000-0001-5154-6095 (A. Morvillo); 0000-0002-9730-8882 (M. Mecella)

2. Related work

The exploration of knowledge graphs (KGs) and graphical visualization have been the subject of numerous research projects and studies as well as the integration of multiple data sources; in the following sections, we will delve deeper into these topics, discussing the methodologies employed and the results achieved by these projects.

2.1. Extensible systems

The GLOBDEF system [2] operates using pluggable enhancement modules, which are dynamically activated to create on-the-fly pipelines for data enhancement. While the system enriches data with semantics, it does not provide a direct method of consumption. On the other hand, Apache Stanbol[1] offers a suite of components that provide services for semantic enrichment, knowledge graph visualization, and metadata management. Although Stanbol does not offer a standalone ready-to-use system, it seamlessly integrates with existing Content Management Systems (CMS).

2.2. Visualization of semantic data

Metaphactory [3] is a platform designed for building knowledge graph applications by seamlessly integrating with other software infrastructures and utilizes Ontodia [4], a powerful free tool [5] for loading and visualizing data. Ontodia's dashboard displays graph entities along with their properties, interconnected by lines representing their relations. Additionally, Ontodia can be used with existing knowledge graphs such as Wikidata and DBpedia, providing a straightforward way to explore document contents.

2.3. Integration systems

In the theme of integrating multiple knowledge graphs, projects such as Mapping Manuscript Migrations (MMM) [6] and INTAVIA[2] use knowledge graphs generated starting from source metadata and integrated with other existing knowledge graphs through the creation of a single integrated KG.
Similarly, MOVING [7] and Mingei [8] are based on an integrated KG, but with starting data obtained by entity extraction using Large Language Models (LLMs).
All the projects indicated, although they carry out effective data integration, have in common the use of their own internal KG.

3. Knowledge graphs in digital humanities

The goal of our research was to validate both the visual approach to knowledge graph exploration and the effectiveness of using multiple knowledge graphs simultaneously and, throughout our research journey, we have encountered a significant challenge in the form of non-uniformity among platforms that utilize a knowledge graph as their knowledge base. For instance, prominent platforms such as Google, DBpedia, and WikiData each have unique organizational structures for their data and this produces different navigation results, thereby creating a lack of consistency in the information retrieved from these platforms.

Different platforms may use different terminologies, taxonomies, or ontologies, further exacerbating the fragmentation issue. Consequently, integrating data from these disparate sources into coherent search results becomes a complex task and may require sophisticated mapping and alignment techniques.

In addition to the fragmentation issue, the information present in some fields is often of a general nature. This generality can be insufficient for producing precise results and it can also lead to ambiguities and uncertainties in the data, making it difficult for users to derive accurate insights from

[1] Cf. https://stanbol.apache.org
[2] Cf. https://intavia.eu

the knowledge graph.

One possible solution to fragmentation is to explore data from multiple sources in real time to produce the search results. Using instead a batch integration, despite offering various advantages in terms of performance and complexity of the system, requires the generation of a new knowledge graph within the system, with potential redundancy of information and therefore an increase in possible misalignments between sources. Because of this, we decided to focus more on real-time exploration in our research path.

Furthermore, the data must be presented in a way that offers effective navigation and visualization to help users explore and understand the data in an intuitive and user-friendly manner.

3.1. Multiple graphs exploration

In the process of implementing the exploration of multiple knowledge graphs, we have hypothesized two types of scenarios:

- Partially related graphs (with some concepts in common)
- Completely disjoint graphs

In the first scenario, it is feasible to consider one of the graphs as the base and the others as an extensions. Therefore, the navigation results are expanded with concepts from the base graph and the estension graphs. This scenario assumes that there is at least one common schema between the graphs, which can be leveraged to create a more comprehensive and interconnected knowledge base.

The second scenario, involving completely disjoint graphs, necessitates a connecting element that allows for the extension of navigation. In our current hypothesis, this element consists of an additional layer of relationships between the disjoint graphs. This layer serves as a bridge, linking concepts from different graphs and enabling navigation across them. This layer can be of any type, such as a translator between IRIs (e.g., a REST service) or another KG (e.g., an additional KG with the same scheme as both the graphs). By introducing this additional layer, the scenario is effectively transformed into the first case of partially related graphs.

Following this approach, it is possible to explore the contents of multiple knowledge graphs in real time, reducing the information to be processed in batches. However, it also presents challenges, such as the need to ensure the accuracy and consistency of the produced results.

3.2. Information collision

During the alignment phase of multiple knowledge graphs, it is not uncommon to find two concepts with identical definitions but different meanings. In this case, collision management can be:

- automatic;
- human moderated;

Automatic management must be performed at the moment the graphs are related to each other and can use various approaches, from the simplest based on the most recent information (if the entry date is present), to one based on learning systems to recognize which colliding data is the most reliable.

Human moderated management, on the other hand, can be performed at different times:

- if the graphs are linked with each other in a batch operation, then it is possible to introduce a moderation phase to manually validate the collisions;

- if the graphs are linked in real-time, end users themselves can determine which of the sources they consider most reliable for their consultation.

Automatic management, although it significantly reduces the manual data analysis work, introduces a degree of error in the production of results, which varies for each approach, and in some cases, human moderated management might be preferable.

In this paper we will not delve into automated collision management.

4. The NAVIGO framework

A framework called NAVIGO was designed to allow the exploration of contents coming from one or more data sources in real time (therefore without the obligation of a prior batch integration) in order to both extend the quality of the results of the exploration, and to solve the problem of knowledge fragmentation. The framework can make use of both local knowledge graphs (e.g., generated by extracting concepts from non-digital data sources) and external ones (e.g., services such as Wikidata), with the possibility of extension to other possible data sources.

The framework architecture is composed of two types of components (Figure 1):

- a main coordination module;
- a series of management modules for the functions, further divided into
  - research;
  - content elements;
  - relationships between elements.

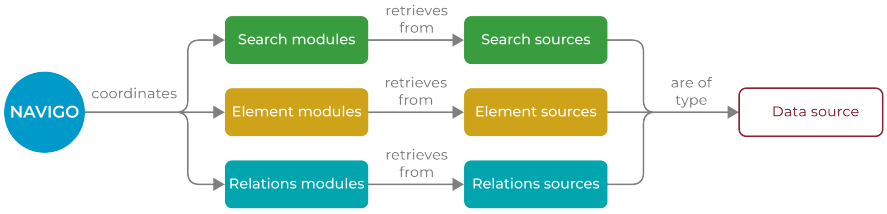

Figure 1: The framework architecture. An orchestrator coordinates modules which independently retrieves data from data sources through their own pipelines. Any source can be of multiple types (i.e., Wikidata is a Search Source, an Element Source and a Relations Source).

5. Case study: Knowledge graphs in archaeological research

In the field of digital humanities, the field of archeology was chosen due to its peculiarity regarding data sources. In fact, in addition to having to draw on data sources of different nature (including, for example, paper texts and maps) which must be cross-referenced with each other, these sources are often of ancient origin, therefore not very suitable for digital conversion, and with concepts implied.

To assess the feasibility of the research endeavor, a prototype named SCIBA [9][10], based on the NAVIGO framework, was developed, targeting the domain of archaeological research. SCIBA's inception was motivated by archaeologists' need to navigate knowledge bases within a geographical framework. It facilitates the discovery of interconnections among various topics associated with a specific location or keyword, thereby fostering the generation of novel insights during content exploration.

This prototype employs semantic and cartographic search methodologies. In contrast to traditional search engines that return a list of texts, documents, and metadata containing the queried keywords, SCIBA's semantic search aims to refine and expand search outcomes.

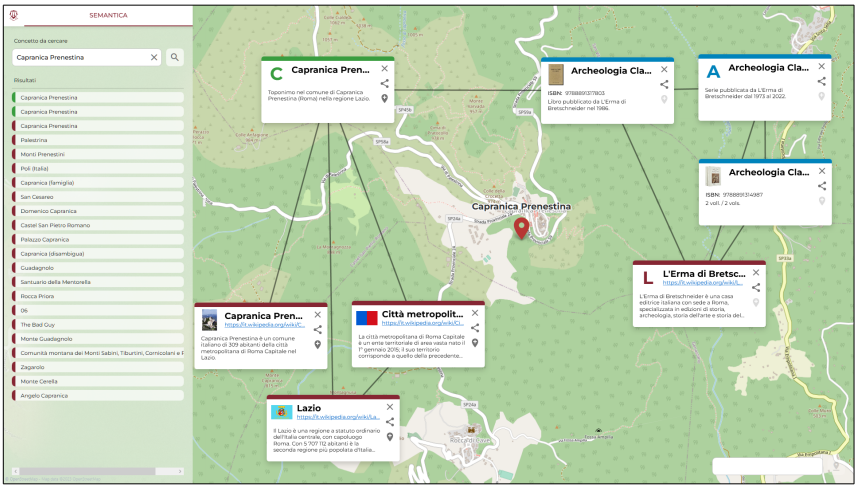

Figure 2: The prototype main screen. The user can perform searches in left bar while floating boxes shows the selected result details and relations, with sources indicated by colour, on an interactive background cartography.

5.1. Data sources

The main data sources used in the prototype consisted of two internal sources:

- Bibliography of books in digital format (with OCR);
- Italian national archive of toponyms (names associated with specific geographical locations);

and two external data sources:

- Wikidata's knowledge graph (Wikimedia foundation);
- Wikipedia articles (Wikimedia foundation);

In addition, as a secondary external source, the DBpedia knowledge graph was added.

Finally, to carry out semantic search, the native semantic engine of Wikipedia was used, as Wikidata was the base graph for all other data sources (except DBpedia).

5.2. Data model and extraction methodology

In a batch processing phase, the internal data sources were converted into a knowledge graph extending the one of Wikidata.

The first data to be converted, using an automated system, were the bibliographic information metadata (title, author, etc.), followed by the toponyms, which required a strong initial moderation phase due to the recurring presence of abbreviations or typos. Subsequently, to associate the content of the books with the concepts represented in the Wikidata KG, the bibliography of books was analyzed using an external Semantic Text Analytics service (DandelionAPI[3]) capable of associating the concepts contained in the books with references to the Wikidata knowledge graph (e.g., the concept "Rome" found in the books was associated with both the Wikidata element Q220, relating to the modern city, and the element Q18287233, relating to the imperial capital). As the last phase, the books were associated with possible toponyms cited in them in an additional internal KG using a query-content matching approach

Regarding the external sources, both Wikidata and DBpedia offer access to a KG, so no intervention was necessary.

In essence, the following local KGs were generated:

[3]Cf. https://dandelion.eu

- **Books KG**, which includes the definition of books derived from metadata;
- **Toponyms KG**, which includes the definition of toponyms in triple format;
- **Book to concepts relationships KG**, which links books to external concepts;
- **Book to toponyms relationships KG**, which links books to the toponyms they cite.

When the user enters a query, starting from the results produced by a semantic reference engine (in this case Wikipedia, managed by one of the NAVIGO modules), the various KGs are queried to produce the results from which to start the exploration of the contents. For the KGs with a schema shared with Wikidata, the use of IRIs is sufficient, while for DBpedia a description-based search was used.

Each source has its unique way of handling data (e.g., Wikipedia API for Wikidata to recover the article details), the data obtained from the sources are further processed by the respective modules in order to produce consultable results (e.g., the Wikidata results are used to recall the associated Wikipedia page).

An example of a workflow used in the prototype is:

1. the researcher inserts a search query into the system;
2. the query is handled by the semantic engine of Wikipedia (the only module with semantic search);
3. using the IRIs or descriptions of the results, possible elements related to the results obtained are searched for.

The results of this workflow generate a list of search results with all possible relationships, extrapolated both from the original KG (Wikidata) and from all the additional KGs accessible from the system. Because the data sources are disjoint, users could view the source of the data and could also choose which source to enable or disable for searching.

5.3. Collision approach

In this case study, the major knowledge source Wikidata with the alternative option of DBpedia are independent of each other and, while beneficial in terms of diversity and coverage, this independence also introduces the potential for data collision.

During the system testing phases, we found that users of the system preferred to exercise their discretion in deciding which results to consider valid and which to discard. This observation led us to forgo the introduction of an automatic collision management system and, while this decision increased the number of ambiguous results, it also presented an opportunity to extend the exploration to a greater number of results.

5.4. Results

During the testing phase, the validity of cross-referencing data to improve archeology search results was confirmed and the system demonstrated the validity of using multiple knowledge graphs instead of a single integrated graph to simplify integration operations. In fact, it was possible to integrate different data sources without the need for burdensome batch integration procedures, also allowing for the possibility of real-time choice.

6. Concluding remarks

The integration of multiple knowledge graphs with other data sources solves the problem of content fragmentation and and opens up potential developments in various fields. With the use of extension graphs dedicated to a specific domain, it allows for precise consultation on topics that are usually addressed in a generic way by other content search systems. This specificity can lead to more accurate and relevant results, thereby improving the efficiency and effectiveness of data-driven research.

The real-time navigation of multiple KGs offers several advantages compared to the generation of a single integrated source, ranging from simplicity of management to the possibility of excluding sources during the search phase, as well as the elimination of batch generation processes.

However, the issue of information collision remains a significant challenge in this context. Information collision occurs when different data sources provide conflicting or overlapping information about the same entity or concept and, in the case of multiple knowledge graphs, each graph may also have its unique representation and interpretation of the same entity or concept. Despite the challenges it presents, information collision also opens up exciting opportunities for future research.

References

[1] B. Haslhofer, A. Isaac, R. Simon, Knowledge graphs in the libraries and digital humanities domain, CoRR abs/1803.03198 (2018). URL: http://arxiv.org/abs/1803.03198. arXiv:1803.03198.

[2] M. Nisheva-Pavlova, A. Alexandrov, Globdef: A framework for dynamic pipelines of semantic data enrichment tools, in: E. Garoufallou, F. Sartori, R. Siatri, M. Zervas (Eds.), Metadata and Semantic Research, Springer International Publishing, Cham, 2019, pp. 159–168.

[3] M. Kejriwal, V. Lopez, J. F. Sequeda, P. Haase, D. M. Herzig, A. Kozlov, A. Nikolov, J. Trame, M. Kejriwal, V. Lopez, J. F. Sequeda, Metaphactory: A platform for knowledge graph management, Semant. Web 10 (2019) 1109–1125. doi:10.3233/SW-190360.

[4] D. Mouromtsev, D. Pavlov, Y. Emelyanov, A. Morozov, D. Razdyakonov, M. Galkin, The simple, web-based tool for visualization and sharing of semantic data and ontologies, in: S. Villata, J. Z. Pan, M. Dragoni (Eds.), The 14th International Semantic Web Conference (ISWC 2015), ceur-ws.org, 2015. URL: http://ceur-ws.org/Vol-1486/paper_77.pdf.

[5] M. Dudáš, S. Lohmann, V. Svátek, D. Pavlov, Ontology visualization methods and tools: a survey of the state of the art, The Knowledge Engineering Review 33 (2018). doi:10.1017/S0269888918000073.

[6] T. Burrows, D. Emery, M. Fraas, E. Hyvönen, E. Ikkala, M. Koho, D. Lewis, A. Morrison, K. Page, L. Ransom, E. Thomson, J. Tuominen, A. Velios, H. Wijsman, Mapping manuscript migrations knowledge graph: Data for tracing the history and provenance of medieval and renaissance manuscripts, Journal of Open Humanities Data (2020). doi:10.5334/johd.14.

[7] V. Bartalesi, G. Coro, E. Lenzi, P. Pagano, N. Pratelli, From unstructured texts to semantic story maps, International Journal of Digital Earth 16 (2023) 234–250. URL: https://doi.org/10.1080/17538947.2023.2168774. doi:10.1080/17538947.2023.2168774.

[8] N. Partarakis, V. Doulgeraki, E. Karuzaki, G. Galanakis, X. Zabulis, C. Meghini, V. Bartalesi, D. Metilli, A web-based platform for traditional craft documentation, Multimodal Technologies and Interaction 6 (2022). URL: https://www.mdpi.com/2414-4088/6/5/37. doi:10.3390/mti6050037.

[9] E. Bernasconi, P. Boccuccia, M. Fabbri, A. Francescangeli, R. Marcucci, M. Mecella, M. Medri, A. Morvillo, M. Pisani, E. Tondi, Sciba - a prototype of the computerized cartographic system of an archaeological bibliography, in: J. Araujo, I. S. de la Vara, Jose Luis adn Brito, N. Condori-Fernandez, L. Duboc, G. Giachetti, B. Marín, E. Serral, A. Bagnato, L. Lopez (Eds.), Joint Proceedings of RCIS 2022 Workshops and Research Projects Track co-located with the 16th International Conference on Research Challenges in Information Science (RCIS 2022), ceur-ws.org, 2022. URL: http://ceur-ws.org/Vol-3144/RP-paper11.pdf.

[10] E. Bernasconi, M. Mecella, A. Morvillo, A proposal for a user-friendly, integrative and collaborative platform for the digital humanities, in: ExICE-Extended Intelligence for Cultural Engagement, 2023.