A proposal for a user-friendly, integrative and collaborative platform for the Digital Humanities

Eleonora Bernasconi[1][0000−0003−3142−3084], Massimo Mecella[2][0000−0002−9730−8882], and Alberto Morvillo[2][0000−0001−5154−6095]

[1] Università degli Studi di Bari Aldo Moro, Department of Computer Science Via Edoardo Orabona, 4, 70125 Bari, Italy
[2] Sapienza Università di Roma, Department of Computer, Control, and Management Engineering Antonio Ruberti (DIAG), via Ariosto, 25, 00185 Rome, Italy

Abstract. Data extraction, retrieval and visualization are some of the key points in Digital Humanities, and together they cover most of the challenges of this field of study, requiring multidisciplinary knowledge. The fruition by users is still today compromised by limited user experiences, problems of integrating data from multiple sources and automatic extraction systems that are not free from errors. The proposal in this document is an approach that, using a simplified graph-based graphical interface, allows the exploration of the contents from multiple data sources and offers a basis for correcting and validating these through social collaboration.

Keywords: Digital library · Cultural heritage · Knowledge graphs · Knowledge discovery.

1 Introduction

Despite the progress in the presentation, management and extraction of data, the complexity of these in the humanities field means that the final content fruition needs to be revised. The effectiveness of using Knowledge Graphs to represent data, structurally and graphically, has already been demonstrated [6]. However, the problem of integrating multiple sources of different natures and errors in the results of automatic systems is still present. To this end, the technology of web semantics [5] comes to the rescue, allowing the structuring of data, independent of the origin, in easily navigable knowledge graphs, while for extraction, the use of AI for linked data recognition [1] facilitates the automation of the process but introduces errors in the correlations often due to the lack of domain-specific knowledge bases and annotated data.

2 Proposed architecture

The proposed approach to addressing the challenges mentioned above consists of three main components: an interface with the composition of a node-link, tabular and map-based interaction paradigms [3] to show the extracted linked data and do search and exploration of the knowledge; a modular framework capable of interfacing with multiple data sources; a collaborative validation platform.

A node-link interaction paradigm is based on the representation of contents using nodes and arcs where nodes are the resources (or concepts) and arcs are the relations that connect one concept to another. This kind of visualization offers, in addition to the search mode using keywords, the possibility of exploration through a logical path.

A modular framework, made up of components for data source interfacing, guarantees high versatility to the system by adding a layer of abstraction and standardization over the data sources technical implementation (pipelines, data formats, ecc.), making the platform data-type independent. In such an architecture, an orchestrator coordinates the modules, which will have the task of independently retrieving the data from the sources.

Social collaboration is a helpful tool for verifying and validating content. There are several existing examples (such as Wikipedia[3] or voice assistants such as Amazon Alexa[4]) and, although it introduces the risk of voluntary alteration or misinformation, therefore the need for moderators, overall it is a tool of interest for Digital Humanities.

We are applying this architecture to **SCIBA** [2], a platform designed to help archaeologists researches by retrieving and discovering multiple information from different Knowledge Graphs, including a digital library and geospatial references, making cross-connection between them. Knowledge extracted from books, books metadata, toponym information and generic knowledge is all sources where the platform can obtain data to provide the users with a visual graph-based interface to search, explore and reference the contents of a digital library on a map. The approach of navigating book contents with the help of a knowledge graph has already been explored by the Arca project [4]. However, SCIBA adds a novel interaction paradigm combining multiple data sources in a geographic-based visualization.

The SCIBA architecture comprises a main orchestrator that synchronizes many static submodules, each with its independent pipeline to retrieve data from a source (figure 1). Involved source types can range from RDF storage to local PDF files; for instance, the Wikidata source is, at the time of this proposal, using a mixed approach between requests to the Wikidata SPARQL endpoint and the Wikimedia REST API. Currently, moderation in SCIBA is still under development. However, thanks to reification, the relationships between concept generated automatically for and by SCIBA can be moderated without the needs to edit the related concept's source.

[3] https://en.wikipedia.org/wiki/Help:Introduction_to_Wikipedia
[4] https://alexaanswers.amazon.com/

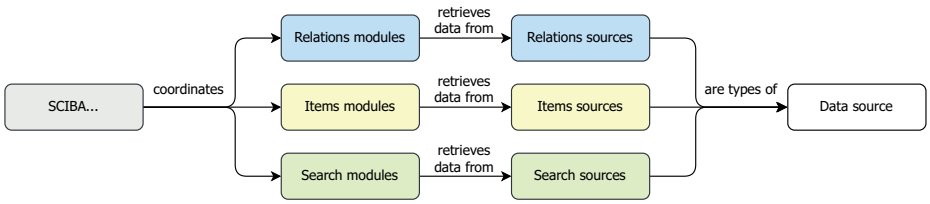

Fig. 1. The SCIBA architecture. An orchestrator coordinates modules which independently retrieves data from data sources through his own pipeline. Any source can be of multiple type (i.e. Wikidata is a Search Source, an Items Source and a Relations Source).

3 Glossary

Knowledge Graph: A knowledge graph, also known as a semantic network, represents a network of real world entities—i.e. objects, events, situations, or concepts—and illustrates the relationship between them. This information is usually stored in a graph database and visualized as a graph structure, prompting the term knowledge "graph."
(from https://www.ibm.com/topics/knowledge-graph)

Semantic web: A development of the World Wide Web in which the data in web pages is structured so that it can be read directly by computers,
(from https://www.oxfordlearnersdictionaries.com/definition/english/the-semantic-web)

RDF: a standard model for data interchange on the Web.
(from https://www.w3.org/RDF/)

Triple: A triple is a statement composed by a subject, a predicate and an object
(from https://www.w3.org/TR/rdf-concepts/#section-triples)

Reification: "Reify" means to take an abstract idea and make it concrete, and in the world of RDF, it means to write RDF statements about RDF statements.
(from https://www.w3.org/wiki/RdfReification)

References

1. Al-Moslmi, T., Gallofré Ocaña, M., L. Opdahl, A., Veres, C.: Named entity extraction for knowledge graphs: A literature overview. IEEE Access **8**, 32862–32881 (2020). https://doi.org/10.1109/ACCESS.2020.2973928
2. Bernasconi, E., Boccuccia, P., Fabbri, M., Francescangeli, A., Marcucci, R., Mecella, M., Medri, M., Morvillo, A., Pisani, M., Tondi, E.: Sciba - a prototype of the computerized cartographic system of an archaeological bibliography. In: Araujo, J., de la Vara, Jose Luis adn Brito, I.S., Condori-Fernandez, N., Duboc, L., Giachetti, G., Marín, B., Serral, E., Bagnato, A., Lopez, L. (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 (May 2022), http://ceur-ws.org/Vol-3144/RP-paper11.pdf
3. Bernasconi, E., Ceriani, M., Mecella, M.: Linked data interfaces: a survey (2023), manuscript submitted for publication
4. Bernasconi, E., Ceriani, M., Mecella, M., Catarci, T.: Design, realization, and user evaluation of the arca system for exploring a digital library. International Journal on Digital Libraries (Dec 2022). https://doi.org/10.1007/s00799-022-00343-0
5. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific american **284**(5), 34–43 (2001)
6. Haslhofer, B., Isaac, A., Simon, R.: Knowledge graphs in the libraries and digital humanities domain. CoRR **abs/1803.03198** (2018), http://arxiv.org/abs/1803.03198