A Streamlined Pipeline to Enable the Semantic Exploration of a Bookstore

Miguel Ceriani(✉), Eleonora Bernasconi(✉), and Massimo Mecella(✉)

Sapienza Università di Roma, Rome, Italy
{ceriani,bernasconi,mecella}@diag.uniroma1.it

Abstract. Searching in a library or book catalog is a recurrent task for researchers and common users alike. Thanks to semantic enrichment techniques, such as named-entity recognition and linking, texts may be automatically associated with entities in some reference knowledge graph(s). The association of a corpus of texts with a knowledge graph opens up the way to searching/exploring using novel paradigms. We present a pipeline that uses semantic enrichment and knowledge graph visualization techniques to enable the semantic exploration of an existing text corpus. The pipeline is meant to be ready for use and consists of existing free software tools and free software code contributed by us. We are developing and testing the pipeline on the field, by using it to access the catalog of a bookstore specialized in ancient Rome history.

Keywords: Semantic enrichment · Knowledge graph · Book catalog · Semantic web · Linked data · Pipeline

1 Introduction

Searching in a library or book catalog is a recurrent task for researchers and common users alike. The search tools, once cumbersome physical file cabinets organized by author, topic, etc., are now usually web-based interfaces that allow more search flexibility and are globally accessible from any web-connected device. Nevertheless, the adopted search paradigm is still mainly the same one, albeit with the important addition of free-text search.

In the last years, knowledge graphs gained broad adoption as a way of organizing and exploring a domain of knowledge. They organize information around concepts, which are connected to each other through semantic relationships. If these concepts are interpreted as topics, a knowledge graph is a rich way to organize a set of texts (or other media) by topic. The relationships between concepts (topics) are preserved and can be used to explore/search the corpus in ways that can go beyond the simple classification of media by topics.

This work is partly supported by the project *ARCA* (POR FESR Lazio 2014–2020 - Avviso pubblico "Creatività 2020", domanda prot. n. A0128-2017-17189) and by the *Centro di Eccellenza DTC Lazio* through the project *EcoDigit*.

M. Ceci et al. (Eds.): IRCDL 2020, CCIS 1177, pp. 75–81, 2020.
https://doi.org/10.1007/978-3-030-39905-4_8

We propose a lightweight system that takes advantage of existing technologies to organize a library or book catalog through a knowledge graph with little upfront effort. A visual user interface allows the user to search and explore the graph as a way to access the book corpus. The pipeline is being experimented on the book catalog of an editor specialized in ancient Rome history.

The rest of the paper is organized as follows. Section 2 analyses the related work, while Sects. 3 and 4 describe respectively the proposed system and used data models. Section 5 describes the implementation details and the concrete use case. Finally, Sect. 6 concludes and anticipates future work.

2 Related Work

There has been a large amount of work in literature about visual information seeking [6,13]. Nevertheless, most of the work focus on how to explore and filter items classified by a homogeneous set of properties. For unstructured information like books, exploring and filtering by basic metadata (i.e., author, title, etc.) can be useful but it is often not sufficient. There has hence been recently a lot of research on how to attach semantics to unstructured data [11], through processes like *named-entity recognition and linking (NERL)* [9,12].

Several software tools and research works deal with the issue of such semantic enrichments. Yewno Discover [2] is an integrated system that addresses similar challenges but does not offer flexibility, requiring the development of ad-hoc adjustments to build a specific pipeline. The GLOBDEF system [10] works with pluggable enhancement modules, which are dynamically activated to create on-the-fly pipelines for data enhancement, but it does not provide the management, integration, and visualization of the generated metadata. Apache Stanbol[1] is a set of components able to offer various services for semantic enrichment, visualization of knowledge graph and the management of metadata. It is extremely useful and can be integrated with our system, but on itself, it does not offer a ready to use pipeline. Multiple user interfaces for visualization and exploration of knowledge graphs have been researched [1,4,8], but the question on how to effectively use these extracted semantics is still open.

Although existing work deals with aspects of the pipeline proposed here, our system is novel in being designed from the ground up to offer knowledge graph-based access to an arbitrary corpus of texts. The mechanism of integration of semantic enrichment services, crucial for the adaptivity of the pipeline, is also novel, by being based on simple, actionable, semantic descriptions of the services. Finally, the user interface is novel in adopting visual linked data exploration as a means to search in a corpus of content, rather than just as an end in itself.

3 Scenario and System

In the considered scenario, the responsible of a catalog of books (e.g., an editor or a library) wants to facilitate the search and exploration of its corpus through

[1] https://stanbol.apache.org/.

a specialized knowledge graph. The knowledge graph needs to integrate existing metadata, concepts associated with texts through semantic enriching processes, and relationships between the concepts. Both *generic users* and *domain experts* will be able to interact with the knowledge graph via a visual user interface or via programmatic interfaces which will enable advanced queries, transformations, and integration with further data sources.

The proposed pipeline is shown in Fig. 1. In case of having access only to printed versions of some texts, those are first scanned and go through an OCR. The content of all the books is then stored in electronic form (e.g., PDFs), along with the relevant metadata, in a repository that supports the *linked data container* API, a standard RDF-based REST API [15]. This repository can be maintained by the catalog maintainers (e.g., editors or librarians) through a dedicated frontend application. It can also be directly connected with existing databases/systems for automatic content/metadata insertion/update.

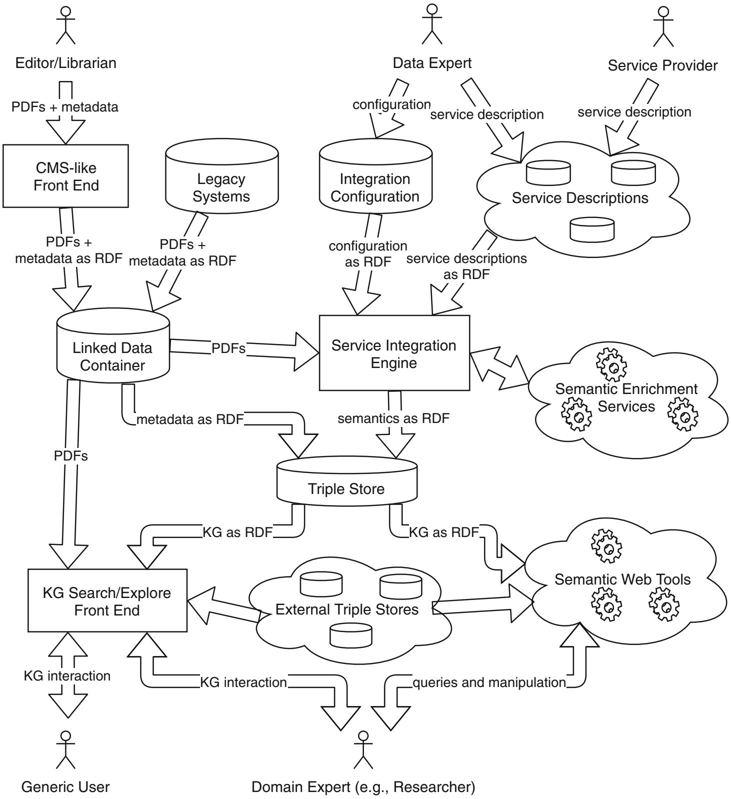
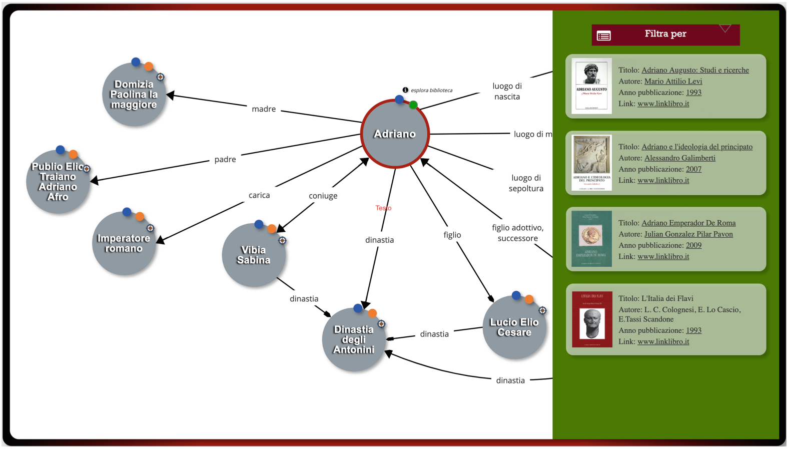


Fig. 1. The proposed pipeline

The content stored in that repository is analysed by some, possibly remote, semantic enrichment services (as NERL) that give as output some knowledge extracted from the content, possibly represented using existing models and knowledge graphs. To allow plugging diverse services, a component called *service integration engine* manages calling and integrating the desired web services based on a global *integration configuration*, which describes which services need to be called, and for each service a specific *service description*, which describes how to adapt it. While the integration configuration is maintained by experts for the specific pipeline, service descriptions are adapters of existing web services that could be developed by the maintainers of this pipeline as well as the service providers or third parties, favouring scalability of the system.

Both the metadata coming from the repository (linked data container) and the extracted knowledge coming from the service integration engine are stored in a triple store, where they can be accessed either through the front end or directly through a SPARQL endpoint. The front end offers a multi-paradigm user interface, in which the knowledge graph visual exploration is coupled with a tabular exposition of the metadata of texts in the corpus (Fig. 2 shows a mockup). Offering the data in RDF format through SPARQL, enables advanced and unanticipated use of the data, through semantic web standards and tools.



Titolo: Adriano Augusto: Studi e ricerche
Autore: Mario Attilio Levi
Anno pubblicazione: 1993
Link: www.linklibro.it

Titolo: Adriano e l'ideologia del principato
Autore: Alessandro Galimberti
Anno pubblicazione: 2007
Link: www.linklibro.it

Titolo: Adriano Emperador De Roma
Autore: Julian Gonzalez Pilar Pavon
Anno pubblicazione: 2009
Link: www.linklibro.it

Titolo: L'Italia dei Flavi
Autore: L. C. Colognesi, E. Lo Cascio, E.Tassi Scandone
Anno pubblicazione: 1993
Link: www.linklibro.it

Fig. 2. A mockup of the user interface

4 Data Models

RDF [3], the basic data model for linked data, is used to represent all the data items in the pipeline, adopting specific vocabularies or ontologies for each type of managed data. The standard query language for RDF, SPARQL [5], is used as a basis to define views and mappings. The structure of the PDF repository is represented thanks to the *linked data platform* vocabulary [15]. Basic metadata about the books are represented through the *Dublin Core Metadata Element Set*[2].

For the description of external services and basic mapping of the inputs/outputs, the *actions* descriptors from the *schema.org*[3] vocabulary are used, as proposed in [14], augmented by the use of the SPARQL Generate language[4] [7] to define non trivial mappings of the output. The semantic annotations (output of the semantic enrichment services) are represented using the Web Annotations Vocabulary[5], which allows to associate properties to the annotation itself (e.g., reliability score) and to identify the exact portion(s) of text an annotation refers to. The annotations associate the text to concepts in some knowledge graph which may be topics (after named entity recognition and linking), moods (after sentiment analysis), etc. The knowledge graphs may adopt different data models.

5 Implementation and Case Study

We are in the phase of implementing the whole pipeline. For each step, there are existing solutions we adopted or new components that we are developing. For the triple store we are using Blazegraph[6], while for the content repository supporting LDP containers we use Fedora Commons[7]. The service integration engine is based on the SPARQL Generate engine[8], which maps the JSON output of each web service to RDF. The web front end is developed using the React framework[9] for modularity, using the JS library Ontodia[10] for the knowledge graph visual user interface. For books that do no exist natively in electronic format, the scanned pages go through the OCR of the software ABBYY FineReader Pro 15[11]. For semantic enrichment, we are using the external *entity extraction* (NERL) web service offered by the Dandelion API[12], which relates segments of the input text to resources in DBpedia, along with a confidence value. Nevertheless, given the flexibility of the service integration mechanism, the system is not tied to this specific service.

The practical case study considered is to implement the idea for “L’Erma di Bretschneider”, an Italian editor with around two thousands publications. “L’Erma” specializes in ancient history, especially ancient Rome history, and it is well-known in the field. The system is being tested on a selected catalog of 198 books, each of them containing from around two hundreds to seven hundreds

[2] http://purl.org/dc/elements/1.1/.
[3] http://schema.org/.
[4] An extension of SPARQL designed to map JSON or XML content to RDF.
[5] http://www.w3.org/ns/oa.
[6] https://www.blazegraph.com/.
[7] https://duraspace.org/fedora/.
[8] https://github.com/sparql-generate/sparql-generate.
[9] https://reactjs.org/.
[10] https://www.ontodia.org/.
[11] https://www.abbyy.com/en-eu/finereader/.
[12] https://dandelion.eu/.

pages and measuring from around one megabyte to 180 megabytes as PDFs. The user interface to the knowledge graph will be publicly available, in order to support the exploration of the catalog of books. The expected users falls in two main categories: casual users willing to explore the catalogue and knowledge on ancient Rome history; expert users that do research in the field and need support to explore and find books relevant to their research topic.

6 Conclusions

This paper presented a concrete lightweight pipeline for enhancing access to a catalog of books through knowledge graph based exploration. The system is based on free software components and meant to be easily deployable for small-medium sized organizations that may not have the technical know-how and resources needed to build and maintain a specifically designed knowledge graph and software system. The development is still in progress but the design analysis and tests carried on so far indicate that the pipeline works without the need for custom coding and it appears useful to the target users. The presented case study will offer a context to thoroughly and formally evaluate the software. The evaluation will include a task oriented analysis as well as a holistic analysis of the impact of the tool on creative processes of research and personal enrichment.

References

1. Bikakis, N., Sellis, T.: Exploration and visualization in the web of big linked data: a survey of the state of the art. arXiv preprint. arXiv:1601.08059 (2016)
2. Bolina, M.: Yewno discover. Nord. J. Inf. Lit. High. Educ. **11**(1) (2019). https://doi.org/10.15845/noril.v11i1.2772
3. Cyganiak, R., Wood, D., Lanthaler, M.: RDF 1.1 concepts and abstract syntax. W3C REC 25 February 2014. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/
4. Dadzie, A.S., Rowe, M.: Approaches to visualising linked data: a survey. Semant. Web **2**(2), 89–124 (2011)
5. Harris, S., et al.: SPARQL 1.1 query language. W3C REC 21 March 2013. http://www.w3.org/TR/2013/REC-sparql11-query-20130321/
6. Keim, D.A.: Information visualization and visual data mining. IEEE Trans. Visual. Comput. Graph. **8**(1), 1–8 (2002)
7. Lefrançois, M., Zimmermann, A., Bakerally, N.: A SPARQL extension for generating RDF from heterogeneous formats. In: Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., Hartig, O. (eds.) ESWC 2017. LNCS, vol. 10249, pp. 35–50. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58068-5_3
8. Marie, N., Gandon, F.: Survey of linked data based exploration systems (2014)
9. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)
10. Nisheva-Pavlova, M., Alexandrov, A.: GLOBDEF: a framework for dynamic pipelines of semantic data enrichment tools. In: Garoufallou, E., Sartori, F., Siatri, R., Zervas, M. (eds.) MTSR 2018. CCIS, vol. 846, pp. 159–168. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-14401-2_15

11. Ristoski, P., Paulheim, H.: Semantic web in data mining and knowledge discovery: a comprehensive survey. J. Web Semant. **36**, 1–22 (2016)
12. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: issues, techniques, and solutions. IEEE Trans. Knowl. Data Eng. **27**(2), 443–460 (2014)
13. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings of 1996 IEEE Symposium on Visual Languages, pp. 336–343 (1996)
14. Şimşek, U., Kärle, E., Fensel, D.: Machine readable web APIs with schema.org action annotations. In: Proceedings of SEMANTiCS 2018, pp. 255–261. Elsevier (2018)
15. Speicher, S., Arwe, J., Malhotra, A.: Linked data platform 1.0. W3C Recommendation 26 February 2015 (2015). http://www.w3.org/TR/2015/REC-ldp-20150226/