Holistic Graph-Based Document Representation and Management for Open Science

Stefano Ferilli(✉), Davide Di Pierro, and Domenico Redavid

University of Bari, 70125 Bari, BA, Italy
stefano.ferilli@uniba.it

Abstract. (Extended Abstract) While most previous research focused only on the textual content of documents, advanced support for document management in Digital Libraries, for Open Science, requires handling all aspects of a document: from structure, to content, to context. These different but inter-related aspects cannot be handled separately, and were traditionally ignored in Digital Libraries. We propose a graph-based unifying representation and handling model based on the definition of an ontology that integrates all the different perspectives and drives the document description in order to boost the effectiveness of document management. We also show how even simple algorithms can profitably use our proposed approach to return relevant and personalized outcomes in different document management tasks.

Keywords: Document Representation · Knowledge Graphs · Document Management · Open Science

1 Introduction

Open Science (OS) is an approach to the scientific process that focuses on making all research knowledge available, so as to build a more replicable and robust science using new technologies, altering incentives, and changing attitudes [11]. Fundamental to OS are the FAIR (Findability, Accessibility, Interoperability, Reusability) principles for data and metadata [10]. The obvious infrastructure to support OS are Digital Libraries (DLs). However, to handle OS issues, the standard realm of DLs must be expanded, in order to describe and/or store the content of the documents (textual or conceptual content, physical, layout and logic structure, semantics), additional information and materials that are external to the publications (datasets, systems, tools, etc.), and their *context*. This expansion requires advanced knowledge handling approaches, but also enables new, high-level functions that support scholars and researchers in their activities.

The solution we propose is to leverage approaches and methods developed in the field of AI, and specifically knowledge representation and handling models based on Knowledge Graphs (KGs). In this direction, a few works have tried

O. Alonso et al. (Eds.): TPDL 2023, LNCS 14241, pp. 3–7, 2023.
https://doi.org/10.1007/978-3-031-43849-3_1

to go beyond simple metadata schemas and proposed the use of ontologies for DLs. Still, there is a lack of infrastructure to support the practices of OS [6]. Some existing taxonomies to describe OS are just organizations of concepts, but cannot be used as schemes of a DL database. Even the data model proposed in OpenAIRE [8] does not fully grasp our idea of context.

The objectives of our work are:

1. crafting an ontology for DLs that: (i) moves from traditional record-based description to a graph-based representation of knowledge; (ii) expands the area of description to both content and context; (iii) can describe concepts that are typical of OS; (iv) may support the FAIR principles on both standard and additional materials;
2. implementing a prototype with an initial set of functions that this ontology may enable, and that may improve the practice.

The core ontology we defined can be extended by each community based on its needs, and that would act as a schema for the knowledge base.

A novel contribution of our approach is the *contextual* perspective. It can establish additional, direct or indirect, non-trivial connections between documents, document components, or pieces of content, based on domain-specific or common-sense knowledge, automatically extracted from, or manually contributed by, external sources.

2 Proposed Representation

The top-level classes (i.e., the immediate subclasses of the universal class) in our ontology are the following: Artifact, Collection, ContentDescription, **Dataset**, **Device**, Document, DocumentDescription, Environment, Event, IntellectualWork, InternetComponent, Item, Organization, Person, Place, **ProcessComponent**, **Project**, **Setting**, **Software**, **System**, TemporalSpecification, **Tool**, User. In bold are those specifically connected to OS and sufficiently general; any specific branch of science may develop, if needed, its own subclasses for these classes. Relationships are also provided to connect items within each of the above classes or across classes. More technically, we adopt an LPG-based approach to ontologies and knowledge graphs, as described in [3], and thus we may also define properties on nodes and arcs.

The portion of ontology dealing with DL concepts is compliant with the Dublin Core Metadata Initiative (DCMI), the IFLA Functional Requirements for Bibliographic Records (FRBR) [7] and the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) standards[1]. The portion of ontology dealing with OS is aligned to OpenAIRE [8]. We expanded this core in several directions: while [5] discusses the DL-related extensions, we considered that the following different but complementary aspects must be considered in DLs to provide a real support to OS:

[1] https://www.openarchives.org/ore/.

- *Textual*, related to the lexical and grammatical features;
- *Layout*, concerning the geometrical structure of documents;
- *Logical*, dealing with the roles played by the document's components;
- *Conceptual*, interested in the meaning conveyed by the documents, both explicitly (e.g., the terms appearing in the text) and implicitly (e.g., the subject dealt with in the document);
- *Contextual*, adding information and creating connections outside what is expressed in each document, or even in the entire collection.

We call our ontology-based approach a 'holistic' one, because it considers and brings to cooperation all these aspects.

The Textual, Layout, Logical and Conceptual aspects concern the content, and may describe the documents as a whole or their single (layout, logical, or grammatical) components. Concepts are typically organized in taxonomies. E.g., the WordNet ontology [9], the Dewey Decimal Classification (DDC) system [1], and the ACM Computing Classification System (CCS)[2]. Several taxonomies can be stored, inter-connected and expanded with additional user-defined and/or domain-specific items. Any instance of these classes can be used to tag individuals of other classes, possibly with different weights. Instances of the various relationships enable forms of associative reasoning, such as graph traversal, that leveraging textual, semantic and contextual information allow finding non-trivial paths between the documents and their contents.

Contextual description of documents relies on general and domain-specific classes provided by the ontology, and not strictly related to document structure, content or management. It may also involve DL-related classes in the ontology, but using them in additional and different relationships than in bibliographic records. Even classes to express users and their profiles, useful for personalization purposes, may be included. Together with the textual-semantic portion of the ontology, the contextual portion acts as a hub to interconnect pieces of information that would otherwise be disconnected, e.g. two documents using the same dataset that do not explicitly mention each other. This can help in carrying out some research tasks: in scholarly research, supporting or even suggesting investigation directions not explicitly present in any of the single documents, but emerging from their direct or indirect relationships; in document clustering, improving the quality of similarity computation, by leveraging information that is, again, not present in any of the available documents; in document classification, improving performance by expanding and integrating the information present in the document with related information coming from the background knowledge or from other documents; in document indexing, allowing to retrieve documents that do not explicitly contain the search parameters set by the user; in query answering, allowing to find more source documents, indirectly related to the question posed by the user but relevant to answer it.

[2] https://dl.acm.org/ccs.

3 Prototype Implementation

For a first implementation of our proposal, we leveraged a number of previous works and systems from our past research, as described in [2], and specifically GraphBRAIN [4] for knowledge storage and management.

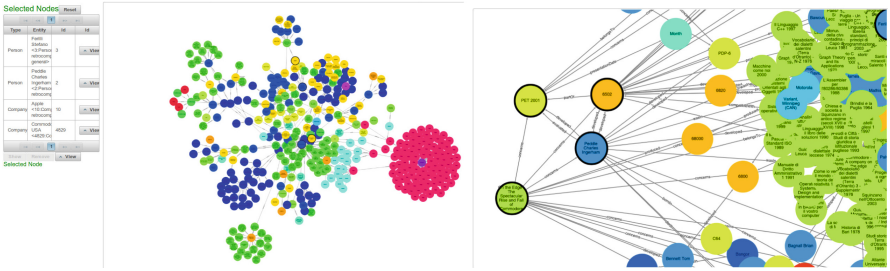

Fig. 1. Overall and zoomed portion of GraphBRAIN's knowledge base

The prototype included a few demonstrative functions:

Subgraph Extraction Starting from a set of nodes provided as input, returns a selected portion of the knowledge graph which is more relevant to these nodes (see Fig. 1, where the starting items are dragged on the side).

Information Retrieval based on both (lexical or conceptual) content and on context, for extending the set of results compared to traditional approaches.

Question Answering based on identifying a subgraph including the answer and translating it into natural language.

Instance Clustering where clusters are emerging aggregations of related items that may involve instances of any kind (see Fig. 1).

Recommendation based on both closeness in the graph and compatibility with the user's profile.

Support for Scholarly Research through automatic extraction (by applying network analysis algorithms) or manual browsing (by expanding portions of the graph at need, and exploring the properties of the nodes and arcs).

The prototype was tested in various domains: history of Computer Science, Cultural Heritage, LAM (Libraries/Archives/Museums), Tourism and Food, but also including linguistic, ontological and contextual information. Each of these sub-domains is organized according to its specific ontology, and these ontologies are connected to each other.

Our proposal responds to the five 'schools of thought' of OS. For democracy, we guarantee access to all types of users and provide functions for searching and question answering. From a pragmatic point of view, we bring different people together through links between works and authors. Concerning infrastructure, the information we store about the structures, tools and technologies used in a given context allow to share and reuse ideas on how to build infrastructure. For

integration with the public, our system does not pose any technological barrier to entry. The interface is simple, secure and does not distinguish users with specific knowledge from others. Through cooperation and the amount of data available, different metrics can be shared to evaluate any solution from several viewpoints and a more accurate overview can be obtained.

References

1. Dewey, M.: A classification and subject index for cataloguing and arranging the books and pamphlets of a library. Amherst, Massachusetts (1876)
2. Ferilli, S.: An automatic intelligent system for document processing and fruition. Trans. Mach. Learn. Data Min. **11**, 43–62 (2018)
3. Ferilli, S.: Integration strategy and tool between formal ontology and graph database technology. Electronics **10**(2616) (2021)
4. Ferilli, S., Redavid, D.: The GraphBRAIN system for knowledge graph management and advanced fruition. In: Helic, D., Leitner, G., Stettinger, M., Felfernig, A., Raś, Z.W. (eds.) ISMIS 2020. LNCS (LNAI), vol. 12117, pp. 308–317. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59491-6_29
5. Ferilli, S., Redavid, D.: An ontology and knowledge graph infrastructure for digital library knowledge representation. In: Ceci, M., Ferilli, S., Poggi, A. (eds.) IRCDL 2020. CCIS, vol. 1177, pp. 47–61. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39905-4_6
6. Hocker, J., Schindler, C., Rittberger, M.: Participatory design for ontologies: a case study of an open science ontology for qualitative coding schemas. Aslib J. Inf. Manage. **72**, 671–685 (2020)
7. IFLA Study Group on the FRBR: Functional requirements for bibliographic records - final report. Tech. rep., International Federation of Library Associations and Institutions (2009)
8. Manghi, P., et al.: The openaire research graph data model (2019). https://doi.org/10.5281/zenodo.2643199
9. Miller, G.: WordNet: a lexical database for English. Commun. ACM **38**, 39–41 (1995)
10. Mons, B., et al.: Cloudy, increasingly fair; revisiting the fair data guiding principles for the European open science cloud. Inf. Serv. Us **37**, 49–56 (2017)
11. Spellman, B.A., Gilbert, E.A., Corker, K.S.: Open Science, pp. 1–47. John Wiley & Sons, Ltd (2018). https://doi.org/10.1002/9781119170174.epcn519