A Generic Framework for Adding Semantics to Digital Libraries

Muhammad Ahtisham Aslam[1(✉)], Naif Radi Aljohani[1], Rabeeh Ayaz Abbasi[1], Miltiadis D. Lytras[2], and Muhammad Ashad Kabir[3]

[1] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
{maaslam,nraljohani,rabbasi}@kau.edu.sa
[2] The American College of Greece, Athens, Greece
mlytras@acg.edu
[3] School of Computing and Mathematics, Charles Sturt University, Sydney, NSW, Australia
akabir@csu.edu.au

Abstract. The World Wide Web (WWW) is emerging as the Web of Data, providing information on various domains. A vast number of scientific documents such as books, articles and journals can be found through many publisher's websites, portals and XML exports. The challenge here is that the data about these scientific documents can not be explored collectively, as they are published as a bounded group of sources organized by different publishers. To address these limitations, we have developed a generic framework termed as Linked Open Publications Data Framework (LOPDF) that facilitates crawling, processing, extracting and producing machine-processable data that is open and linked to other open datasets. We also demonstrate the RDF datasets produced by using LOPDF framework and describe statistics of different datasets entities.

1 Introduction

With the growth of the Web, it has become preferred platform on which to publish documents and organizational data [4]. Traditionally, data published on the Web has been made available in formats such as HTML, XML, CSV, and tables, resulting in loss of data's structure and semantics [3]. Moreover, a significant amount of information is encoded in structured forms [7] by using information templates. Specific to the scientific domain, information about scientific publications is encoded by using common terms such as *title, author, ISBN*, and structured by using various publisher-specific templates. This provides better representation and human understanding, yet it prevents the information from being processed by machines. As a result, data of scientific publications is silenced and fails to connect to scientific works put out by other publishers.

In response, a generic framework termed as Linked Open Publications Data Framework (LOPDF) is described in this paper. The LOPDF framework can be used to crawl, parse, extract and produce LOD of scientific publications. We

I. Ciuciu et al. (Eds.): OTM 2016 Workshops, LNCS 10034, pp. 277–281, 2017.
DOI: 10.1007/978-3-319-55961-2_28

implemented and used the LOPDF framework to extract huge datasets of about three hundred million RDF triples providing information on over nine million scientific documents from *SpringerLink*[1] as source of data.

The remainder of this paper is organized as follows. Related work is discussed in Sect. 2. In Sect. 3 we describe the architecture of our framework. The knowledge extraction algorithm is discussed in Sect. 4. Section 5 demonstrates the results in terms of datasets extracted by using LOPDF framework. Finally, we conclude and outline future directions for our work in Sect. 6.

2 Related Work

For more than two decades, the semantic Web and LOD communities have been working on information integration [8] and semantic Web technologies [3]. This has led to the development of different frameworks and approaches for extracting and producing Linked Open Data (LOD). For example, a template-based extract transform load method is described in [2] to publish archaeological linked data from excavation archaeological datasets. The tool introduced in [2] minimizes the load to understand schema and mapping rules and maximizes the auto mapping of data to linked data. Another domain-independent framework for extracting linked data from tables is presented in [7]. The proposed framework can be used to interpret tables and thus produce linked data. Graphical models and probabilistic reasoning theories are used to extract the content of the columns and, to some extent, their data. DBpedia [6] is a key project, acting as the nucleus of the Web of open data. Datasets of the dbpedia knowledge base have been extracted using the dbpedia extraction framework, resulting in the addition of more than five million RDF triples to the Web of open data. A generic language (i.e. RML) for mapping relational databases to RDF is introduced in [5]. RML is basically an extension to R2RML and can be used to extract and map heterogeneous data to RDF.

3 Framework Architecture

Data on scientific publications is expressed by using terms (e.g. *title, ISBN*, etc.). Links between parent and child documents (e.g. link of a book with chapters) are established through publisher-specific templates. LOPDF architecture is designed to crawl across all parent/child documents and to extract and triplify metadata as well as links between documents. LOPDF architecture consists of four modules (as shown in Fig. 1) which are described below:

Crawler. The crawler is the main component of the LOPDF framework and it crawls through the data source in such a way that the relational information between documents is preserved. That is why, while crawling, if it finds that a document is a parent document (e.g. book or journal) that might have child

[1] http://link.springer.com/.

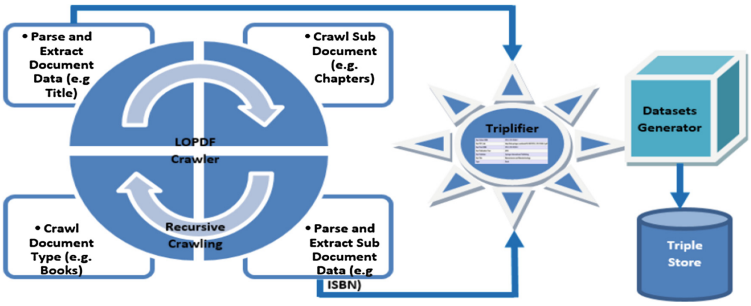

Fig. 1. *Linked open publications data framework (LOPDF)* architecture.

documents (e.g. chapters or articles), the crawler goes through all related child documents so that neither is left unprocessed.

Parser. This component is used to parse and extract metadata (e.g. *title, author, etc.*) about every document. Relational metadata is also parsed and extracted by this component to preserve the relational information between various entities. During the crawling process, when the crawler finds an information template, this component is activated to extract available information at that particular stage.

Triplifier. The triplifier component is used to process the information extracted by the parser, and to generate RDF triples for all metadata and to store them in to relevant data models. Before triplification, this component also takes care of the data types of the extracted data by considering the nature of values of properties as object or data type properties. The values of object properties are mapped to the relevant classes and values of data type properties are mapped to the appropriate literal data types.

RDF Generator. This component takes the data models created by the triplifier as its input and processes them to generate RDF datasets in N-Triple format. The resulting datasets contain complete information (as RDF triples) about metadata, as well as links between various entities, which enables researchers to put semantically enriched queries to these datasets through SPARQL endpoint.

4 Data Extraction Algorithm

The LOPDF data extraction algorithm adopts a recursive approach (as shown in Algorithm 1) to crawl, parse, extract and produce scientific publication's LOD. In its first step, the algorithm starts to traverse the data source from the first discipline (e.g. computer science, engineering, etc.) and crawls across every discipline.

In its next step, the algorithm starts crawling inside a particular content type (e.g. within a book or journal) and crawls through all documents with in a particular content type (e.g. all chapters in a book, all volumes, issues and articles in a journal, etc.). While crawling inside a particular content type, information about all documents that are part of a particular content type (i.e. relational information

between content type and document) and metadata about a particular document (e.g. title, isbn, etc.) are extracted and added to the data model. Then recursive approach takes the extraction process back to the next content type, and then to the next discipline. By using this recursive approach, all disciplines, content types and document types are processed, and the data models are made ready for the triplification of the extracted data and the production of final datasets in N-Triple format.

```
Data: Publisher's portal/ Web site home as data source
Result: Semantically enriched publications data in N-Triple format
while Not reach end of disciplines do
    Crawl disciplines;
    while Not reach end of content types (e.g. book, journal, etc.) do
        Crawl content types ;
        while Not reach end of content type with in a discipline do
            Parse and extract content metadata (e.g. title, abstract, etc.) ;
            Add metadata information in to data model;
            Crawl sub-documents with in a particular content type;
            while Not reach end of sub-documents do
                Extract metadata of sub-documents ;
                Add metadata information in to data model;
            end
        end
    end
    Triplify extracted information;
    Generate RDF datasets in N-Triple format;
end
```

Algorithm 1. *LOPDF* data extraction algorithm.

5 Results and Discussion

As a part of LOPDF implementation, we undertook minor customizations to the endpoint triggers of the generic framework and applied it to the extraction and production of semantically enriched data, using *SpringerLink* as our source of data. We were able to extract data from about nine million documents including 5.476 million articles, 3.24 million chapters, 0.478 million reference work entries. Resulting RDF datasets consist of about three hundred million RDF triples (all together) and are published on the project website[2] for download in .nt format. We also created a knowledge base (termed as SPedia [1]) by using these datasets and a SPARQL endpoint that can be used to put sophisticated queries to *SPedia* datasets, either by making use of semantic Web techniques or connecting the semantic Web browser to the SPARQL endpoint. Table 1 shows some sample statements from RDF datasets gernerated by using LOPDF framework.

[2] http://wo.kau.edu.sa/Pages-SPedia.aspx.

Table 1. Sample RDF statements from datasets generated by using LOPDF.

Subject	Predicate	Object
spedia:Coordinate_Metrology	spedia:has_Title	"Coordinate_Metrology".
spedia:Coordinate_Metrology	rdf:type	"Book".
spedia:Coordinate_Metrology	spedia:has_Online_ISSN	"978-3-662-48465-4".

6 Conclusion and Future Work

In this article we presented a generic framework that can be used to add semantics to digital libraries and produce LOD on scientific publications. We described the architecture of our framework and the generic recursive algorithm on which we implemented it. We outlined our results, comprised of RDF datasets that we extracted from *SpringerLink* as source of data. These datasets consist of about three hundred million RDF triples while providing information on about nine million scientific documents. As part of future enhancement, we are working on customizing the LOPDF framework to extract semantically enriched data from other well-known publishers as well.

References

1. Aslam, M.A., Aljohani, N.R.: SPedia: a semantics based repository of scientific publications data. In: Cui, B., Zhang, N., Xu, J., Lian, X., Liu, D. (eds.) WAIM 2016. LNCS, vol. 9658, pp. 479–490. Springer, Heidelberg (2016). doi:10.1007/978-3-319-39937-9_37
2. Binding, C., Charno, M., Jeffrey, S., May, K., Tudhope, D.: Template based semantic integration: from legacy archaeological datasets to linked data. Int. J. Semant. Web Inf. Syst. **11**(1), 1–29 (2015)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semant. Web Inf. Syst. **5**(3), 1–22 (2009)
4. Ceri, S., Bozzon, A., Brambilla, M., Valle, E., Fraternali, P., Quarteroni, S.: Web information retrieval. In: Chapter Publishing Data on the Web, pp. 137–159. Springer, Heidelberg (2013)
5. Dimou, A., Sande, M.V., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: R.M.L: a generic language for integrated RDF mappings of heterogeneous data. In: Proceedings of the 7th Workshop on Linked Data on the Web, CEUR Workshop Proceedings, vol. 1184, April 2014
6. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semant. Web **6**(2), 167–195 (2015)
7. Mulwad, V., Finin, T., Joshi, A.: A domain independent framework for extracting linked semantic data from tables. In: Search Computing Broadening Web Search, pp. 16–33 (2012)
8. Wiederhold, G.: Intelligent integration of information. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD 1993, pp. 434–437. ACM, New York (1993)