International Journal on Semantic Web and Information Systems
Volume 13 • Issue 1 • January-March 2017

SPedia:
A Central Hub for the Linked Open Data of Scientific Publications

Muhammad Ahtisham Aslam, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Naif Radi Aljohani, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

ABSTRACT

Producing the Linked Open Data (LOD) is getting potential to publish high-quality interlinked data. Publishing such data facilitates intelligent searching from the Web of data. In the context of scientific publications, data about millions of scientific documents published by hundreds and thousands of publishers is in silence as it is not published as open data and ultimately is not linked to other datasets. In this paper the authors present SPedia: a semantically enriched knowledge base of data about scientific documents. SPedia knowledge base provides information on more than nine million scientific documents, consisting of more than three hundred million RDF triples. These extracted datasets, allow users to put sophisticated queries by employing semantic Web techniques instead of relying on keyword-based searches. This paper also shows the quality of extracted data by performing sample queries through SPedia SPARQL Endpoint and analyzing results. Finally, the authors describe that how SPedia can serve as central hub for the cloud of LOD of scientific publications.

KEYWORDS

Digital Libraries, Information Retrieval, Knowledge Extraction, Knowledge Representation, Linked Open Data, Ontologies, Semantic Web

1. INTRODUCTION

The growth of domains of knowledge in our data intensive age depends particularly on the efficiency and sophistication of the processes of data production, distribution and consumption, among the corresponding community (Andriole, 2010). Specific to scientific domain, there is huge amount of data about vast number of scientific documents such as articles, books, reference works, being produced by academia and industry. Unfortunately, these documents are being published as bounded group of publisher specific resources resulting in lake of collaboration and interconnected resources for knowledge sharing. There is an urgent need to publish and share research publications data. This can enable other researchers to interconnect their data to the one that already published. Ultimately this can be used by researchers and practitioners to share their research (Kauppinen de Espindola, 2011) for better collaboration and future analysis.

The set of best practices for publishing and interconnecting distributed data has termed as Linked Open Data (LOD). These best practices are being used by increasing number of data providers (Bizer, Heath Berners-Lee, 2009; Villazón Terrazas, Vilches, Corcho Gómez-Pérez, 2011) such as government (Lebo et al., 2011), education (Lnenicka, 2015), news (Suárez Jiménez-Guarín, 2014), health (Bukhari Baker, 2013), geography (Correndo, Salvadores, Yang, Gibbins Shadbolt, 2010) and by researchers to extract semantically enriched data from different public resources such as Wikis, as

DOI: 10.4018/IJSWIS.2017010108

community effort to publish LOD (Erxleben, Gu¨nther, Krötzsch, Mendez Vrandecic, 2014; Vrandečić Krötzsch, 2014; Lehmann et al., 2015). When it comes to the scientific publications data, very little work has been conducted (e.g. Springer., 2015, Hakimpo-ur, Arpinar Sheth, 2007) to publish LOD of scientific documents. It is also acknowledged (Blmel, Dietze, Heller, Jschke Mehlberg, 2014) that in scientific research, structured data is limited and exposed based on proprietary or less-established schemas resulting in unholistic and inconsistent view on research information.

As a step towards publishing linked open data of scientific publications, in this paper we present SPedia: a semantically enriched knowledge base of scientific publications data. SPedia knowledge base adds three hundred million RDF triples to the Web of data which provide information on about nine million scientific documents published in twenty-four disciplines and four different languages. SPedia knowledge base is populated from the scientific publications data of documents published by Springer and we used SpingerLink[1] as source of data. SPedia datasets are available for download from project Web site[2] and can be used to link other open datasets published in the LOD cloud. In SPedia project we have also established a SPARQL Endpoint that can be used to put semantically enriched queries to SPeida for the intelligent query answering purposes rather than to rely on keyword-based searches on unlinked distributed data.

The work presented in this paper makes the following contributions:

- The extraction of structured information from over 9 million documents available on SpringerLink. The resulting datasets contain information about scientific documents from 24 disciplines (e.g., computer science, engineering, social sciences, etc.) and 6 types of documents (e.g., books, chapters, journals, articles, reference works, and reference work entries), written in four different languages;
- Production of semantically enriched datasets as RDF triples which were extracted from the detailed information (e.g., abstract, DOI, ISBN, pdf link, author, organization, etc.) of scientific documents;
- Extraction and triplification of relational information between various document types (e.g., relationships between book and book chapter, journal and journal article, etc.);
- Customized approach and algorithm for crawling, parsing and extraction of useful information and triples generation;
- Development of a SPARQL Endpoint that can be used to put semantically enriched queries against SPedia datasets that consist of more than three hundred million RDF triples.

The rest of the paper is structured as follows: Related work is discussed in Section 2. In Section 3 we describe data model of information source. Semantics based representation of extracted data is explained in Section 4. Section 5 briefly describes the resulting RDF exports. Then in Section 6 we describe various methods to access, query and browse the resulting SPedia datasets. Section 7 explains that how SPedia can serve as a central hub for the linked open scientific publications data cloud. Finally we conclude our work in section 8 and describe future work.

2. RELATED WORK

Typically, initiatives regarding linked open datasets originate in the publication of online catalogs of raw datasets in different domains such as government (Sheridan Tennison, 2010), geography (Shvaiko et al., 2012), health (Kozák, Něcasky´, Dědek, KÍmek Pokorny´, 2013), news (Stadler, Lehmann, H¨offner Auer, 2012). These feature keyword search and faceted browsing interfaces to

help users find relevant datasets and retrieve corresponding metadata including dataset description and URLs to download (Ding et al., 2011). In this section, we describe the related work in extracting and publishing LOD in different domains. This will highlight the value of publishing LOD. Then we describe related work done in publishing the LOD in the science and education domain which shows that still a vast amount of data on scientific documents needs to be processed and published as a part of the global Web of data.

In (Blmel et al., 2014), authors describe the current state of the art in research information sharing while exposing the door opening for further research to extract and publish LOD specific to scientific research. A solution (i.e. VIVO) to create a semantically enriched network of researcher is described in (Nogales, Sicilia Jrg, 2014; Corson-Rikert Cramer, 2010). VIVO facilitates the sharing and networking of researchers such as investigators, technical staff, students and other stakeholders by making the use of LOD approach and principles. This work does not deeply address the networking of scientific publications and documents that are produced by researchers and scientists from different domains and published by different publishers.

Another system to produce LOD from multimedia contents and to use it in education has been presented in (Chae et al., 2015). Since, multimedia contents don't have hyperlinks that connect related multimedia data, therefore, this approach is very useful even for non-technical people to easily interconnect semantically related multimedia data. Another, benefit of this system is that it could be used to link the multimedia LOD with other open datasets so that it can be used in different educational scenarios.

Similarly, a system (i.e. BibBase) is described in (Xin, Hassanzadeh, Fritz, Sohrabi Miller, 2013; Xin et al., 2010). BibBase can be used to publish the bibliographic data that is available in BibTex files as LOD. The BibBase semantically enriched bibliographic data can be used to link with existing open datasets as well as is available as HTML and RSS feeds that can be integrated into personal Websites. The BibBase system also provide a SPARQL Endpoint that can be used to ask semantically enriched queries by making use of SPARQL protocol.

An approach to produce semantically enriched data about authors and publications from the DBLP knowledge base is presented in (Aleman-Meza et al., 2007). The presented approach makes use of DBLP XML export as source and extracts and produce RDF exports by crawling, parsing and extracting structured information from the source XML document. Key limitation of both approaches that are presented in (Xin et al., 2013, 2010) and (Aleman-Meza et al., 2007) is that they provide only RDF version of bibliographic data while leaving the relational information/links between different documents/publications and authors unprocessed which is very well addressed in our work while considering the Springer publications as our source.

Springer Developer APIs (Springer, 2015, https://dev.springer.com/) is set of APIs which provides programmatic access to the metadata of articles published by Springer. These are available in different types of varying capability, for requests ranging from a particular type of metadata to groups of metadata entities. They also have performance limitations as being working on HTTP protocol as well as they generate output data as JSON or XML which further needs to be processed. In addition to this, Springer has taken an initiative (Springer., 2015, http://lod.springer.com/wiki/bin/view/Linked+Open+Data/About) to publish the LOD on documents published in conferences in computer science only, while leaving vast amount data about millions of other scientific documents unprocessed. This limitation of existing state-of-the-art also created a ground for SPedia knowledge base.

3. THE DATA MODEL OF THE INFORMATION SOURCE

Springer is the leading global scientific and technical publishers which has published over nine million scientific documents. SpringerLink is the portal which provides access to these scientific document including 5.476 million articles, 3.24 million chapters, 0.478 million reference work entries (Springer, 2015). SPedia is a knowledge base which provides semantically enriched information on these scientific documents published by Springer. SPedia is populated from SpringerLink (as source

of data). SpringerLink uses its specific templates to organize and link these vast number of documents and expresses the documents metadata by using standard terms such title, abstract, isbn, doi. In this section we describe the data model of the SpringerLink, being the source of data for SPedia.

3.1. Categorization of Documents

SpringerLink categorizes scientific documents based on two aspects i.e. 1) discipline to which a document belongs and 2) the type of document's contents. These categorizations are further explained below:

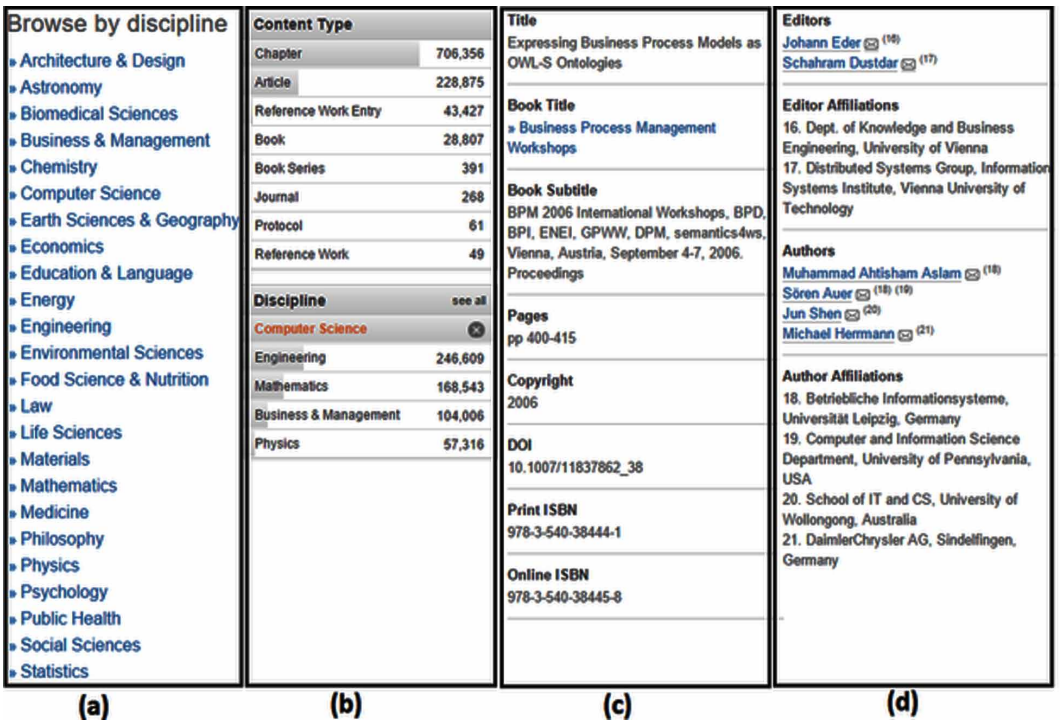
3.1.1. Discipline Based Categorization

Documents on the SpringerLink portal are first categorized based on the discipline to which they belong. SpringerLink identifies twenty four disciplines such as computer science, chemistry, engineering, law for document's categorization. All the published documents belong to anyone of these disciplines. Categorization of documents into disciplines help in key-word based searching as well as in semantic matching of document's ontological belonging which helps in intelligent searching by semantic based applications. Figure 1 (a) shows list of twenty four disciplines in which documents are categorized in the source portal.

3.1.2. Document Type Based Categorization

In SpringerLink templates, documents are categorized in to four major and three sub categories. Four major categories are book, journal, reference work and protocol and three sub categories are chapter, article and reference work entry. These categories are also referred as content types (as shown in Figure 1 (b)). Scientific publications in every discipline are categorized in these content types. So it

Figure 1. SpringerLink template views showing: (a) Discipline categorization; (b) Content types; (c) Document's metadata; and (d) Relation to persons as author/editor



means that every document must have two category types i.e. “discipline” and “content type”. As an example we can say that a document can be an article which is published in computer science discipline.

3.2. Links between Documents and Other Entities

Extracting established links between documents and representing them as RDF triples is a feature that adds key differentiation to SPedia with respect to structured knowledge extracted about bibliographic information. Instead of just bibliographic information (as discussed in (Xin et al., 2013, 2010) and (Aleman-Meza et al., 2007)), SPedia also extracts relations between scientific publications. Different kind of relations/links between documents and other entities are explained below.

3.2.1. Links Between Parent and Child Documents

As discussed above, scientific publications in the source portal are categorized in document/content types which are further grouped as parent and child document types. The documents such as book, journal, and reference work are categorized as parent documents and documents such as chapter, article and reference work entry are categorized as child documents. Every parent document has link/relation to child document and every child document has relation to its parent document. As a general example we can say that every book must have chapter/s and every chapter must belong to some book. Similarly, every article must belong to some journal and every journal has article/s. These parent child documents relations are well organized as parent and child HTML pages which have hyperlinks between each other in addition to parent/child document metadata.

3.2.2. Links Between Documents and Other Entities

In addition to parent/child documents relations, a document is also linked/related to numerous other entities such as author, editor, editor-in-chief; these entities have further relations to sub-entities such as author affiliation (as shown in Figure 1 (d)). All these links are extracted and processed to create different relational properties when the data source is processed to be expressed as semantically enriched RDF datasets. It is because of the reason that semantic enrichment process for a data source corresponds to processing all such links from which properties are extracted.

3.3. Metadata of Documents

Metadata is a key source of information about scientific documents. SpringerLink provides metadata of scientific publications in a very well organized templates and information in these templates is presented by using standard publications metadata terminologies such as title, abstract, print isbn, online isbn, doi, volume, issue, coverage. Figure 1 (c) shows metadata information presented in the SpringerLink metadata specific template.

4. SEMANTICALLY ENRICHED RERESENTATION OF EXTRACTED DATA

In previous sections we describe the data model of the source portal. In this section we describe in detail about how the semi-structured data about scientific documents is processed to be expressed in semantically enriched format so that it can be linked to other open datasets as well as can be processed by machines.

4.1. Expressing Documents and Other Resources

One of the key principles of the semantic Web and LOD is representing the things on the Web as resources. Any concrete or abstract entity can be a resource. Every resource can be described by using its related properties and representing them as RDF statements (Yu, 2007). Following this principle, entities such as books, chapters, journals, articles, authors, are extracted and identified as resources.

Example 1 below shows a Journal identified as resource and used as subject of RDF statements in SPedia knowledge base.

Example 1: Representing subject of RDF statement as a resource:

<http://www.kau.edu.sa/fcit/SPedia#Knowledge_and_Information_Systems>

or, if we have the prefix of the URI i.e.:

PREFIX spedia:<http://www.kau.edu.sa/fcit/SPedia# >

Then we can also present it as:

spedia:Knowledge_and_Information_Systems

4.2. Metadata Representation

Metadata of scientific publications is represented by using standards terms such title, abstract, isbn, doi, author, editor, and is formatted in different publisher specific templates. Such metadata is extracted from the source portal and is represented as resources by mapping it to SPedia ontology classes such as book, chapter, journal, or as literal values such as isbn number, title string, doi number. These resources as well as literal values are triplified as object of resulting RDF statements and added to RDF data models. Table 1 list down some common properties extracted for all documents.

4.3. Linking Extracted Datasets

As discussed above, every document (e.g. book, chapter, journal, article etc.) and every entity (e.g. author, editor, organization etc.) is extracted and identified as a resource in SPedia datasets. These resources, for sure, have strong links/relations with each other. For example, a chapter must be belonging to some book, an article must be belonging to some journal, and a reference work entry must be belonging to some reference work and vice versa. Similarly, every document must be having some relation to some person/s as author, editor, editor-in-chief and vice versa. Also, these authors/editors must be having some affiliation with some organization and vice versa. All these relations inter link the different entities (e.g. documents, persons, organizations) with each other within and across resulting datasets. SPedia datasets are linked with each other as linked data by using linking/relational properties between different entities. Table 2 shows the list of properties, their descriptions as well as sample RDF statements showing the relation between different datasets.

4.4. Expressing Properties and Data Types

Properties play the role of predicate in RDF statements and are used to express the metadata of documents in SPedia. Values of these properties are mentioned as object of resulting RDF statement. Depending on the nature of the property values, they are either mapped to classes (expressing the object as a resources) or to literal values. As an example, the object/value of the property has_Chapter is a resource of type spedia:Chapter, which means object of resulting RDF triple is also a resource (as shown in Example 2).

Example 2: An example of RDF statement in which value of property (i.e. object) is also a resource:

http://www.kau.edu.sa/fcit/SPedia#Digital_Imaging_Primer

Table 1. Description of some common properties extracted for all documents types

Property	Description	Example
type	Provides data about the document type (e.g. type can describe that a document is a Chapter in discipline Computer Science.	spedia:Digital_Imaging_Primer rdf:type spedia:Book. spedia:Digital_Imaging_Primer rdf:type spedia:Computer _Science.
has_Title	Provides the title of the document.	spedia:Digital_Imaging_Primer spedia:has_Title "Digital Imaging Primer".
has_Abstract	This property describes the abstract of the document truncated up to 100 characters from the source abstract.	spedia:Photometry spedia:has_Abstract "In an image, difference of location is marked by.....".
has_DOI	Contains the DOI of the document.	spedia:Digital_Imaging_Primer spedia:has_DOI "10.1007/978-3- 540-85619-1".
has_PDF_Link	Provides the link to the pdf file of the actual document.	spedia:Evaluating_link_prediction_methods" spedia:has_PDF_Link "http://link.springer.com/article/10.1007/s10115-014-0789-0.pdf".
has_Print_ISBN	Contains the print ISBN number of the document.	spedia:Digital_Imaging_Primer spedia:has_Print_ISBN "978-3-540-85617-7".
has_Online_ISBN	Contains the online ISBN number of the document.	spedia:Digital_Imaging_Primer spedia:has_Online_ISBN "978-3-540-85619-1".
has_Pages	Points the number of pages in the source document.	spedia:Photometry spedia:has_Pages "253-267".
has_Publisher	Contains the information about the publisher.	spedia:Photometry spedia:has_Publisher "Springer Berlin Heidelberg".
has_Publication_Year	Provides information about the publication year of the document.	spedia:Photometry spedia:has_Publication_Year "2016".
has_Author	Contains the information about the authors of the document which then further contains his name, email and affiliation.	spedia:Photometry spedia:has_Author spedia:Alan_Parkin.
author/editor type	Describes author/editor/editor-in-chief as a person by using foaf vocabulary.	spedia:Alan_Parkin rdf:type foaf:person.
author/editor organization	Describes author/editor/editor-in-chief's affiliation by using foaf vocabulary.	spedia:Alan_Parkin foaf:organization "London, UK".
has_Editor_In_Chief	Provides information about the editor-in-chief of the document (e.g. editor of book or journal).	spedia:Knowledge_and_Information_Systems spedia:has_Editor_In_Chief foaf:person.
has_References	List all references of the document in string format.	spedia:Photometry spedia:has_References "1. BIPM (2006) The international system of units,.....".

http://www.kau.edu.sa/fcit/SPedia#has_Chapter
<http://www.kau.edu.sa/fcit/SPedia#Photometry >

Similarly, the object/value of the property has_title is a literal value of type string (as shown in Example 3).

Example 3: An example of RDF statement in which value of property is literal value of type string:

http://www.kau.edu.sa/fcit/SPedia#Digital_Imaging_Primer
http://www.kau.edu.sa/fcit/SPedia#has_Title
<xsd:string ”Digital Imaging Primer” >.

In this way properties are categorized as object and datatype properties. Range of object properties are classes of SPedia ontology and Range of datatype properties are of types xsd:int, xsd:string etc.

5. RDF EXPORTS AND DISCUSSION

We crawled and processed the SpringerLink portal as source of data in January 2016 and produced RDF version of the Springer publications data as a semantically enriched knowledge base (named as SPedia). SPedia datasets consist of more than three hundred million RDF triples and are about two hundred and fifty gigabytes in size (all together). These datasets can be downloaded in N- Triple format from project Web site. We extracted and produced SPedia datasets at different levels so that users can consume them as per their requirement and availability of resources. In this section we give necessary details about extracted datasets and different levels of extraction.

5.1. Different Levels of Extracted Datasets

In SPedia project, we extracted datasets about Springer documents at four different levels. Since, SPedia datasets are very huge in size (about two hundred and fifty gigabytes all together) and consist of about three hundred million RDF triples, therefore, keeping user convenience and resources limitations in mind, we generated datasets at four different levels (as shown in Table 3).

These datasets can be downloaded from project Web site[3] at following levels: 1) property level, 2) document type level, 3) document category level, and 4) discipline level. SPedia datasets range from distribution of large number of property level datasets files which are smaller in size to small number of discipline level datasets files which are heavy in size. All these datasets files are available for download in .nt format and can be used for local experiments. Some necessary explanation about these different levels of datasets is as under:

1. **Property Level:** In SPedia knowledge base all documents such books, chapters, journals and other entities such as authors, editors, affiliations are identified as resources. Every resource is described by using its properties such as title, abstract, isbn, doi, issue, email, affiliation. Property level datasets provide information about every property of a document in a separate .nt file. Therefore, every document has a separate dataset file for its every property. For example, every article has a title and the article title.nt file contains title of all articles. Similarly, every book has an ISBN number and the book isbn.nt file contains ISBN numbers of all books. This is the reason that property level datasets consist of large number of property level files and these files are smaller in size;
2. **Document Type Level:** Scientific publications are categorized as book, chapter, journal, article, reference work, reference work entry in SpringerLink portal. Each of these document types are described by their related properties such as title, abstract, isbn, doi, issue. Document type level datasets provide complete information about any particular document type in one file. For example, article.nt file contains all data about all articles in one file. Similarly, chapter.nt file contains all data about all chapters in one file. In other words we can say that document type level datasets are collection of property level datasets which are compiled in one document level *.nt* file. Another point here is that book.nt file and article.nt provides all data about all books and about all articles published in one particular discipline such as computer science, engineering, astronomy;
3. **Document Category Level:** Documents are categorized as Books & Chapters, Journals & Articles and Reference Work & Reference Work Entries. Document category-level datasets provide all information (i.e., property level as well as relational information between documents) in a single

Table 2. Properties used to inter link SPedia datasets

Property	Description	Example
has_Book_Chapter	Links every book with the every chapter published in that book.	spedia:Digital_Imaging_Primer spedia:has_Book_Chapter spedia:Photometry.
is_Book_Chapter_Of	This property links every chapter with the book in which chapter is published.	spedia:Photometry spedia:is_Book_Chapter_Of spedia:Digital_Imaging_Primer.
has_Article	Create a relation/link of a journal with all articles published in that journal.	spedia:Knowledge_and_Information_Systems spedia:has_Article spedia:Evaluating_link_prediction_methods.
is_Article_Of	Links every article with the source journal.	spedia:Evaluating_link_prediction_methods spedia:is_Article_Of spedia:Knowledge_and_Information_Systems.
has_Volume	Links journal and articles publisher in that journal through this property.	spedia:Knowledge_and_Information_Systems spedia:has_Volume xsd:integer "45".
is_In_Volume	Describes the volume of the journal in which an article is published which indirectly links article with journal.	spedia:Evaluating_link_prediction_methods spedia:is_In_Volume xsd:integer "45".
has_Issue	Links journal and articles published in that journal through this property.	spedia:Knowledge_and_Information_Systems spedia:has_Issue xsd:integer "3".
is_In_Issue	Describes the Issue of the journal in which an article is published which indirectly links article with journal.	spedia:Evaluating_link_prediction_methods spedia:is_In_Issue xsd:integer "3".
has_Reference_Work_Entry	Links a reference work with its entries by using this property.	spedia:Encyclopedia_of_Parallel_Computing spedia:has_Reference_Work_Entry spedia:Actors.
is_Reference_Work_Entry_Of	Links reference work entry with reference work.	spedia:Actors spedia:is_Reference_Work_Entry_Of spedia:Encyclopedia_of_Parallel_Computing.
has_Author	Links every type of document (e.g. book, chapter, article etc.) with person (as author).	spedia:Evaluating_link_prediction_methods spedia:has_Author spedia:Nitesh_V_Chawla.
is_Author_Of	Links a person (as author) with any type of document.	spedia:Nitesh_V_Chawla spedia:is_Author_Of spedia:Evaluating_link_prediction_methods.
has_Editor	Links every type of document (e.g. book, chapter, article etc.) with person (as editor).	spedia:Handbook_of_Camera_Monitor_Systems spedia:has_Editor spedia:Anestis_Terzis.
is_Editor_Of	Links a person (as editor) with any type of document.	spedia:Anestis_Terzis spedia: is_Editor_Of spedia:Handbook_of_Camera_Monitor_Systems.
has_Editor_In_Chief	Links between journal and person (as editor-in-chief).	spedia:Knowledge_and_Information_Systems spedia:has_Editor_In_Chief spedia:Xindong_Wu.
is_Editor_In_Chief_Of	Establishes link between person (as editor-in-chief) journal.	spedia:Xindong_Wu spedia:is_Editor_In_Chief_Of spedia:Knowledge_and_Information_Systems.

.nt file. For example in level 3 of the Table 3, books and chapters.nt dataset provide complete information about all Books & Chapters in a particular discipline;

4. **Discipline Level:** As discussed above, SpringerLink divides documents into different types (e.g., book, chapter, journal, article etc.) and various document categories (e.g., book & chapter, journal & article, etc.). Every discipline (e.g., computer science, engineering, etc.) contains documents that belong to these types and categories. Discipline-level datasets provide information about all document types and categories in a single file for every discipline. Discipline-level datasets are huge and contain millions of triples in one .nt file. As an example, the discipline-level dataset for computer science consists of more than sixty million triples and is greater than 6.5 GB in size.

Table 3. Different levels at which SPedia datasets are available for download

I. Property Level Datasets						
	Type	**Title**	**Abstract**	**Print ISBN**	**DOI**	**Other Properties**
Book	book type.nt	book title.nt	book abstract.nt	book print ISBN.nt	book DOI.nt	.... other properties datasets.nt
Chapter	chapter type. nt	chapter title.nt	chapter abstract. nt	chapter print ISBN.nt	chapter DOI. nt	.... other properties datasets.nt
Other documents	other type.nt	other title.nt	other abstract.nt	other print ISBN.nt	other DOI.nt	.... other properties datasets.nt
II. Document Type Level Datasets						
Discipline	**Book**	**Chapter**	**Journal**	**Article**	**Reference Work**	**Reference Work Entry**
Computer Science	books.nt	chapters.nt	journals.nt	articles.nt	Reference works.nt	reference work entries. nt
Other Disciplines	books.nt	chapters.nt	journals.nt	articles.nt	Reference works.nt	reference work entries. nt
III. Document Category Level Datasets						
Discipline	**Books and Chapters**		**Journals and Articles**		**Reference Works and Entries**	
Computer Science	books and chapters.nt		journals and articles.nt		reference work and entries.nt	
Engineering	books and chapters.nt		journals and articles.nt		reference work and entries.nt	
Other Disciplines	books and chapters.nt		journals and articles.nt		reference work and entries.nt	
IV. Discipline Level Datasets						
Architecture and Design	**Astronomy**	**Biomedical Sciences**	**Business and Management**	**Chemistry**	**Computer Science**	**Other Disciplines**
architecture & design all in one.nt	astronomy all in one.nt	biomedical science all in one.nt	business & management all in one.nt	chemistry all in one.nt	computer science all in one.nt	other disciplines all in one.nt

5.2. Statistics of Extracted Datasets

SPedia datasets provide information on more than 9 million scientific publications, including over 3100 journals, 183,000 books, 540 reference works, 5.2 million journal articles, 3.1 million chapters, and 0.63 million reference work entries. In addition to scientific publications metadata, SPedia datasets also provide information about other resources, including 31 million people/persons as authors, editors and/or editors-in-chief of scientific publications and 12 million organizations as affiliation of persons. Table 4 summarizes the statistics of extracted datasets.

SPedia datasets provide semantically enriched information about metadata of scientific documents (e.g. title, abstract, isbn, doi, etc.) published by Springer as well as linking of these documents with each other (e.g. book relation with chapter, article relation with journal and vice versa) as well as link of these documents with other entities such as relation of a person as author, editor with a document and vice versa. All this metadata as well as relational information is dumped in .nt files as RDF triples. Figure 2 provides statistics about number of triples extracted for (a) every property and (b) for every discipline as a group of document types.

5.3. SPedia Live

As one of the world's leading scientific, technical, and medical publishers, Springer frequently brings out new journals and book titles. This creates demand for SPedia datasets to be updated with new titles extracted from the online portal using live data. SPedia datasets are extracted every quarter to keep them a breast of recent publications and updates in the source data. Although SPARQL Endpoint is updated with the latest datasets, those from the past two quarters are also available for download on the project website.

6. ACCESSING, BROWSING AND QUERYING SPEDIA

SPedia is an open knowledge base which makes available the publically accessible metadata of Springer publications as semantically enriched data. SPedia datasets are available as RDF dumps and can be downloaded from the project Web site for local experiments. At the same SPedia knowledge base can be browsed by using third party semantic Web browsers such as Gruff (Franz, 2015), which can also be used to visualize resulting data as well as to browse SPedia datasets in graphical environment. SPedia datasets can be accessed, browsed, queried and used in the following ways: 1) as download able dumps, 2) integrating with home pages 3) querying through SPARQL Endpoint, and 4) browsing in semantic environment. Here, we describe these access methods in detail.

6.1. SPedia Linked Datasets Dumps

The simplest way to use SPedia is to download its datasets which are available as RDF dumps and to use them for local experiments and data acquisition. Users, ranging from basic data needs and limited resources to users with bigger requirements and heavy resources availability can download these dumps which are available as property level (smaller in size with limited data) datasets to discipline level datasets (bigger in size with comprehensive data). These datasets can be loaded in any triple store server such as GraphDB[4], Allegrograph[5] server, and used for user specific needs (Table 5 shows sample URIs of resources when their dump files are loaded in to triple store server). These datasets can also be used to develop different applications and mashups.

6.2. Integrating SPedia with Homepages

As discussed above that all documents published by Springer are available as resources which are represented and accessed by URIs. Scientists and researchers can list down their Springer publications from SPedia via the SPARQL endpoint and integrate it with their professional Web site or home page. Both, human and smart agents can use this listed data to further crawl the SPedia datasets,

Table 4. Statistics of number of triples extracted for each property of different document types

Property	Book	Chapter	Journal	Article	Reference Work	Reference Work Entry
Type	363K	6264K	5.6K	8497K	1.8K	0.86 K
Title	181K	3132K	2.7K	4247K	0.9K	0.43 k
Abstract	-	109K	2.3K	-	-	-
Print ISSN	181K	3120K	2.4K	3634K	0.6K	0.42 k
Online ISBN	181K	3191K	2.7K	3638K	0.9K	0.41 k
Publication Year	181K	3131K	-	-	0.9K	0.42 k
DOI	181k	3131k	-	4247k	0.9k	0.43 k
PDF Link	115k	1909k	-	3566k	0.3k	0.43 k
Pages	-	3131K	-	1.2K	-	716 K
Author	161K	4209K	-	6444K	0.12K	255 K
Editor	185K	57K	-	-	1.9K	337 K
Editor-In-Chief	-	-	1.6K	-	-	-
Affiliation	137K	941K	-	1643K	1.1K	94 K
Organization	111K	614K	-	1012K	0.9K	73 K
Publisher	181K	3131K	2.7K	3691K	0.9K	415 K
Sub/Parent Document	3132K	3132K	4249K	4249K	716K	716 K
Volume	-	-	19K	3238K	-	-
Issue	-	-	19K	3438K	-	-
Coverage	-	-	2.7K	4163K	-	-
Number	-	-	1.3K	-	-	-

when browsed over HTTP protocol. SPedia publications data (as resources) can also be used to create RDF based home page of individual's scientific publications which ultimately makes one's home page part of the LOD cloud which could be crawled and processed by semantic based applications.

6.3. SPedia SPARQL Endpoint

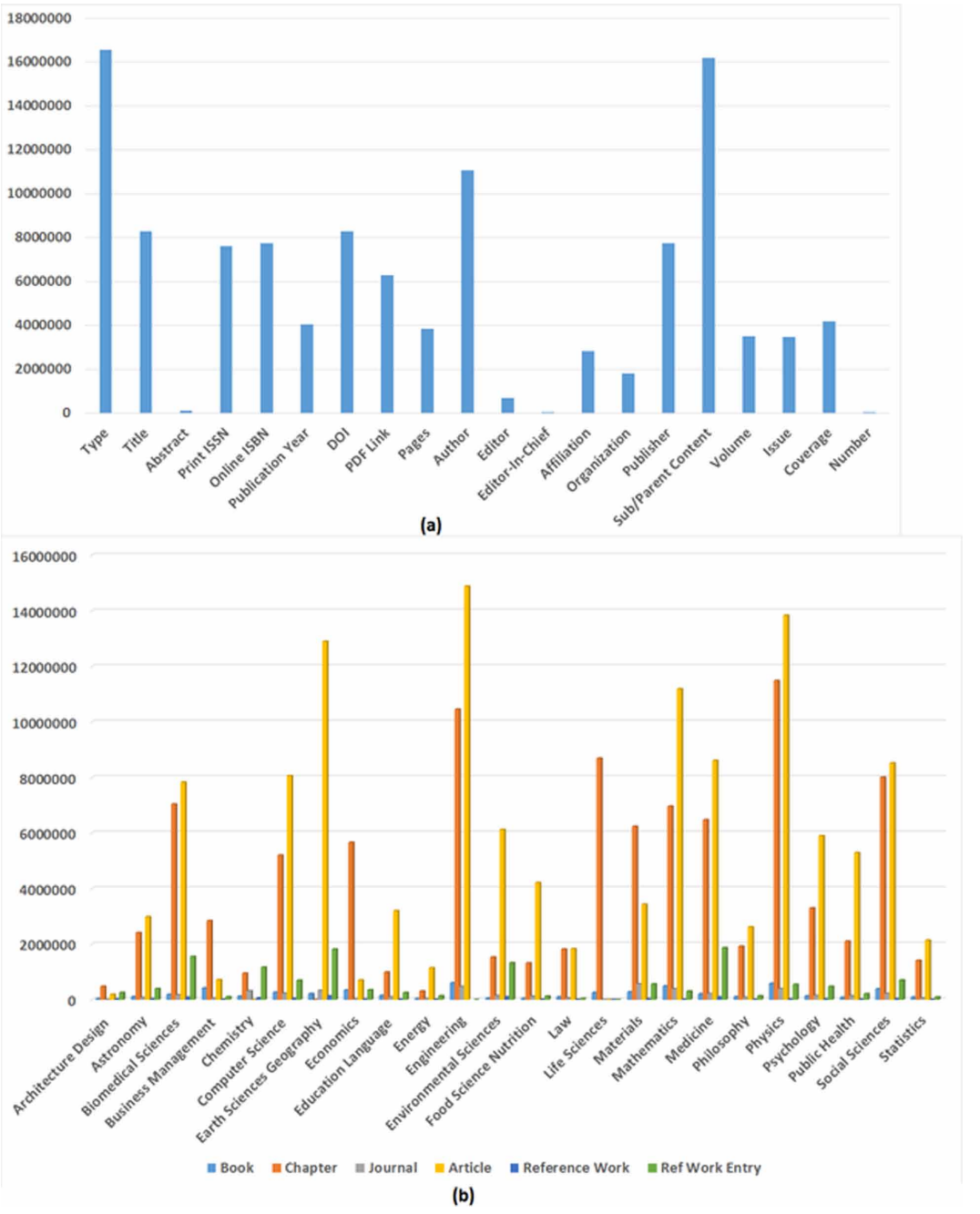
SPARQL Endpoint is the most common and flexible way of accessing and querying the linked open data. In SPedia project, we have also created the SPARQL Endpoint that can be used to query the SPedia datasets. At the same time it can also be used by third party applications as an access point to extracted datasets.

Here, we show some sample SPARQL queries that we execute over the SPedia SPARQL Endpoint and the results that we got from SPedia datasets.

Example 4: Find the Book in Computer Science discipline which has title "A Practical Introduction to Fuzzy Logic using LISP" and all Chapters published in this Book:

```
PREFIX spedia:<http://www.kau.edu.sa/fcit/SPedia# >
select ?book ?chapters where {
?book rdf:type spedia:Book.
```

Figure 2. Statistical graph of number of triples extracted: (a) For commonly used SPedia properties; and (b) For every document type grouped in disciplines

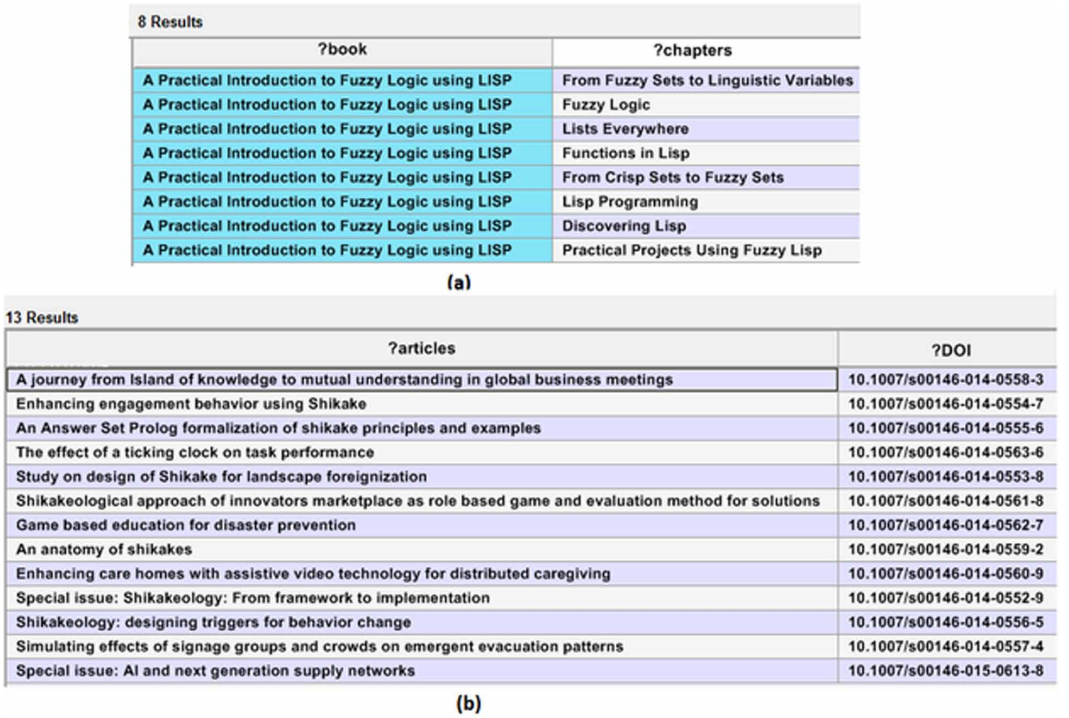


?book rdf:type spedia:Computer_Science.
?book spedia:has_Title ”A Practical Introduction to Fuzzy Logic using LISP”^^xsd:string.
?book spedia:has_Book_Chapter ?chapters.
}

Table 5. Sample URIs from different SPedia dump files

Source	Sample URI
type.nt	http://www.w3.org/1999/02/22-rdf-syntax-ns/type [type of documents].
title.nt	http://www.kau.edu.sa/fcit/SPedia/has_Title [title of documents].
abstract.nt	http://www.kau.edu.sa/fcit/SPedia/has_Abstract [abstract of documents].
organization.nt	http://xmlns.com/foaf/0.1/Organization [affiliation of author/editors].

Figure 3. Results of example queries 4 and 5



8 Results

?book	?chapters
A Practical Introduction to Fuzzy Logic using LISP	From Fuzzy Sets to Linguistic Variables
A Practical Introduction to Fuzzy Logic using LISP	Fuzzy Logic
A Practical Introduction to Fuzzy Logic using LISP	Lists Everywhere
A Practical Introduction to Fuzzy Logic using LISP	Functions in Lisp
A Practical Introduction to Fuzzy Logic using LISP	From Crisp Sets to Fuzzy Sets
A Practical Introduction to Fuzzy Logic using LISP	Lisp Programming
A Practical Introduction to Fuzzy Logic using LISP	Discovering Lisp
A Practical Introduction to Fuzzy Logic using LISP	Practical Projects Using Fuzzy Lisp

(a)

13 Results

?articles	?DOI
A journey from Island of knowledge to mutual understanding in global business meetings	10.1007/s00146-014-0558-3
Enhancing engagement behavior using Shikake	10.1007/s00146-014-0554-7
An Answer Set Prolog formalization of shikake principles and examples	10.1007/s00146-014-0555-6
The effect of a ticking clock on task performance	10.1007/s00146-014-0563-6
Study on design of Shikake for landscape foreignization	10.1007/s00146-014-0553-8
Shikakeological approach of innovators marketplace as role based game and evaluation method for solutions	10.1007/s00146-014-0561-8
Game based education for disaster prevention	10.1007/s00146-014-0562-7
An anatomy of shikakes	10.1007/s00146-014-0559-2
Enhancing care homes with assistive video technology for distributed caregiving	10.1007/s00146-014-0560-9
Special issue: Shikakeology: From framework to implementation	10.1007/s00146-014-0552-9
Shikakeology: designing triggers for behavior change	10.1007/s00146-014-0556-5
Simulating effects of signage groups and crowds on emergent evacuation patterns	10.1007/s00146-014-0557-4
Special issue: AI and next generation supply networks	10.1007/s00146-015-0613-8

(b)

Figure 3(a) shows results of the SPARQL query. The results consist of eight records, where each record shows one chapter (as queried) published in the mentioned book.

Another example of SPARQL query is shown in Example 5. The SPARQL code written in this example queries for a particular journal published in the discipline “Architecture and Design” and all articles published in the particular Volume and Issue of this journal.

Example 5: Find the journal ”AI & SOCIETY” which belongs to the discipline ”Architecture and Design” and then find all articles published in Volume 30 and Issue 4 of this journal and DOI of all articles:

```
PREFIX spedia:<http://www.kau.edu.sa/fcit/SPedia#>
select ?articles ?DOI where {
?document rdf:type spedia:Journal.
?document rdf:type spedia:Architecture_and_Design.
```

```
?document spedia:has_Title ”AI & SOCIETY”^^xsd:string.
?document spedia:has_Article ?articles.
?articles spedia:is_In_Volume ”30”^^xsd:string.
?articles spedia:is_In_Issue ”4”^^xsd:string.
?articles spedia:has_DOI ?DOI.
}
```

Results of the query (written in Example 5) are shown in Figure 3 (b). This figure shows that total 13 articles are published in the Volume 30 and Issue 4 of the journal ”AI & SOCIETY”. Results also show the DOI of every article.

6.4. Browsing SPedia with Client Applications

Client applications and third party applications have been developed by semantic Web and LOD community and are being used for querying and browsing the LOD in semantically enriched environment. Gruff (Franz, 2015) as an example, is a third party application/ semantic Web browser that can be used to load and browse SPedia as a knowledge base of semantically enriched data of scientific publications. It can also be used to connect to SPedia SPARQL Endpoint and execute SPARQL queries over the HTTP protocol. SPedia can also be browsed in visual environment as well as in tabular form by using third party applications and semantic Web browsers. Figure 4(a) shows the textual data in traditional Web browser and Figure 4 (b) shows the semantically enriched data coming from SPedia knowledge base in tabular form when browsed by using semantic Web browser. SPedia datasets can be browsed and explored as long as the interlinked data is available or until the datasets reach to some literal value which have no further link to next data item.

7. CENTRAL HUB FOR LINKED OPEN SIENTIFIC PUBLICATIONS DATA

SPedia is the first effort towards achieving the goal of having linked open data cloud of scientific publications published by well-known publishers. Right now SPedia consists of RDF datasets of scientific documents published by Springer but it can play the role of central hub of scientific publications data by interlinking with LOD of scientific publications published by other publishers. In the next step, as a part of Scientific Publications Pedia (SPPedia) project we are working to produce IPedia, APedia, EPedia, PPedia, MPedia of scientific documents published by IEEE, Amazon, Elsevier, Pearson, McGraw-Hill respectively as shown in Figure 5. SPedia is the central linking point for all

Figure 4. Scientific publication metadata in: (a) Traditional web browser; and (b) In semantic web browser

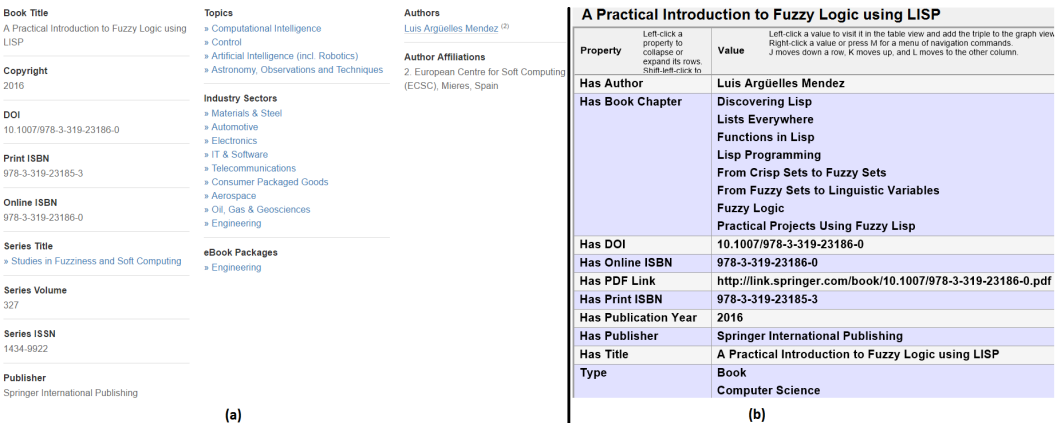
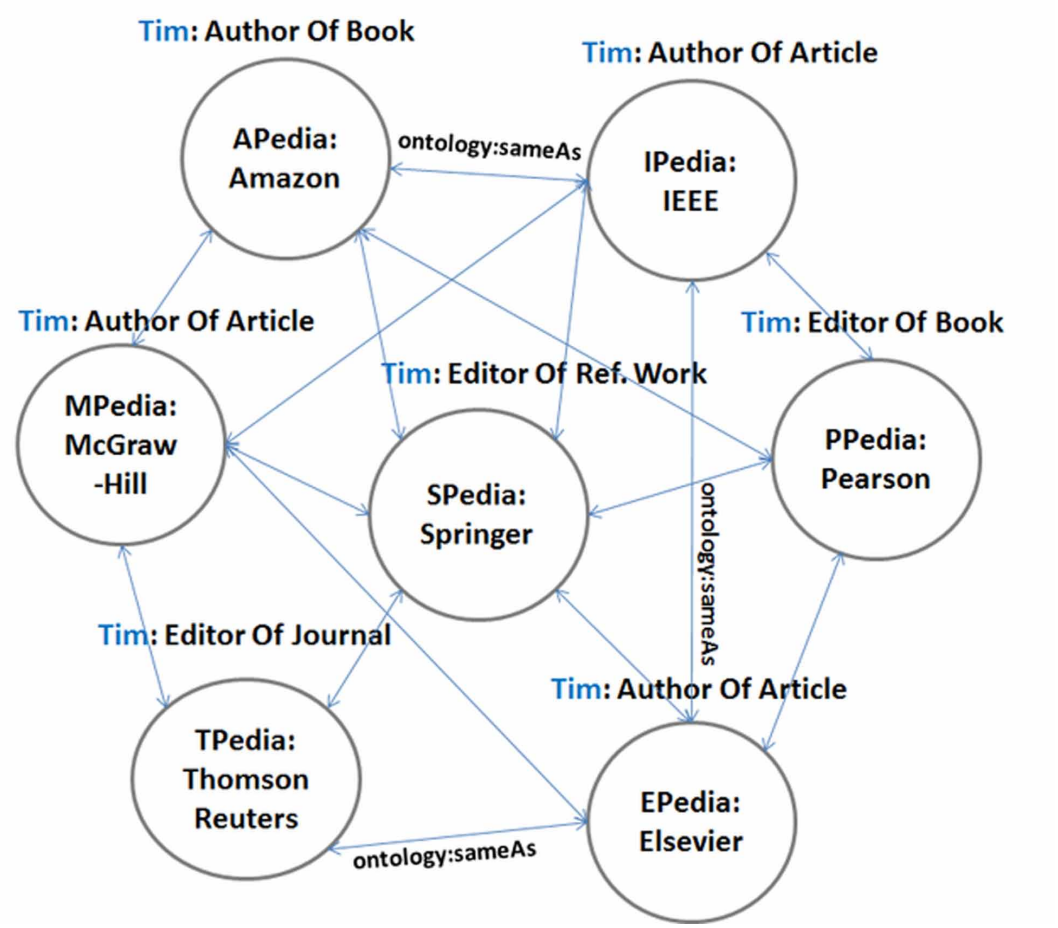


Figure 5. A conceptual view of how SPedia can play the role of central hub for linked open scientific publications data



these repositories in the cloud of scientific publications data and the resulting interlinked cloud of scientific publications data is termed as Scientific Publications Pedia (SPPedia). This cloud of LOD on scientific publications could be used for various purposes such as creating scientific linked open profiles, searching scientists with similar interest and domain, browsing and crawling related documents. The resulting SPPedia datasets can further be interlinked with other datasets available in the LOD cloud.

8. CONCLUSION AND FUTURE WORK

In this paper we have presented SPedia: a semantically enriched knowledge base of scientific publications data which can be used by human as well as smart agents to access, query and browse the linked open data of scientific publications. SPedia consisted of more than three hundred million RDF triples providing information on about nine million scientific documents. We described the knowledge extraction process and algorithm that we used to process the information source and to extract and produce machine processable data. In addition to scientific publications metadata, SPedia datasets also contain a vast amount of data about persons (as author, editor of articles) and organization (as affiliation of persons). SPedia datasets are available on project Web site for download in N-Triple

format (i.e. as .nt files) and can be used for local experiments. We also created a SPARQL Endpoint that can be used to ask semantically enriched queries to the SPedia knowledge base by making use of SPARQL protocol. Keeping in view the semantics aware and unaware users, we described different interaction points that can be used to access, query and browse the SPedia knowledge base. We briefly described and did the quantitative and qualitative analysis of extracted datasets so that it can add confidence to extracted knowledge to be externally linked with other open datasets and used in developing applications that needs machine understandable data. We also showed that how SPedia can play a role of central hub for linked open scientific publications data. SPedia, when interlinked to the LOD, can be used for plenty of purposes: for example it can be used to find related articles from the LOD cloud, it can be used to find researcher with similar domain of interest which ultimately can help in better collaboration and knowledge sharing between researchers.

As part of future work, we are focusing on adding more metadata elements to the SPedia datasets. For example, we are working to process and extract metadata about entities such as impact factor, citations, downloads, in future datasets. An important part of every document that can make the open data really linked, is references. In current SPedia datasets, references are extracted as literal values (i.e. strings). The reason is that these references in the source portal are available as strings only but they can be processed by using some data mining techniques or by using some entity based extraction approach and exporting to corresponding RDF resources. It can enable references to be treated as resources and used for interlinking of datasets which ultimately can enable semantic based agents to crawl across related documents based on every article in references as a next resource. We are also working to extend Springer Developer APIs so that they can be used to produce RDF data (in .nt format) in addition to producing data in XML or JSON. It can help to develop third-party applications as well as new domain specific mashups. As discussed in the Section 7, we are also working on producing semantically enriched datasets from other well-known publishers so that they could be interlinked to create a cloud of linked open scientific publications data.

REFERENCES

Aleman-Meza, B., Hakimpour, F., Arpinar, I. B., & Sheth, A. P. (2007). Swetodblp ontology of computer science publications. *Web Semantics: Science, Services, and Agents on the World Wide Web*, *5*(3), 151–155. doi:10.1016/j.websem.2007.03.001

Andriole, S. J. (2010, December). Business impact of web 2.0 technologies. *Communications of the ACM*, *53*(12), 67–79. doi:10.1145/1859204.1859225

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data the story so far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. doi:10.4018/jswis.2009081901

Blmel, I., Dietze, S., Heller, L., Jschke, R., & Mehlberg, M. (2014). The quest for research information. *Procedia Computer Science*, *33*, 253–260. doi:10.1016/j.procs.2014.06.040

Bukhari, A. C., & Baker, C. J. O. (2013). The Canadian health census as Linked Open Data: towards policy making in public health. In Data integration in the life sciences.

Chae, J., Cho, Y., Lee, M., Lee, S., Choi, M., & Park, S. (2015). Design and implementation of a system for creating multimedia linked data and its applications in education. *Multimedia Tools and Applications*.

Correndo, G., Salvadores, M., Yang, Y., Gibbins, N., & Shadbolt, N. (2010). Geographical service: a compass for the web of data. In C. Bizer, T. Heath, T. Berners-Lee, & M. Hausenblas (Eds.), *Proceedings of the WWW2010 workshop on linked data on the web, LDOW '10*, Raleigh, USA (*Vol. 628*). CEUR-WS.org.

Corson-Rikert, J., & Cramer, E. J. (2010). VIVO: enabling national networking of scientists. *Proceedings of IASSIST '10 – social data and social networking: Connecting social science communities across the globe*, Ithaca, NY, USA.

Ding, L., Lebo, T., Erickson, J. S., DiFranzo, D., Williams, G. T., Li, X., & Hendler, J. A. (2011). TWC LOGD: A portal for linked open government data ecosystems. *Web Semantics: Science, Services, and Agents on the World Wide Web*, *9*(3), 325–333. doi:10.1016/j.websem.2011.06.002

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., & Vrandecic, D. (2014). Introducing wikidata to the linked data web. In *The semantic web ISWC 2014* (pp. 50–65).

Franz. (2015). *Gruff: A grapher-based triple-store browser for allegrograph.* Retrieved from http://franz.com/agraph/gruff/

Kauppinen, T., & de Espindola, G. M. (2011). Linked open science communicating, sharing and evaluating data, methods and results for executable papers. *Procedia Computer Science*, *4*, 726–731. doi:10.1016/j.procs.2011.04.076

Koz'ak, J., Neˇcasky´, M., Dˇedek, J., Kl'ımek, J., & Pokorny´, J. (2013). Linked open data for healthcare professionals. *Proceedings of international conference on information integration and web-based applications & services* (pp. 400–409). New York, NY, USA: ACM. doi:10.1145/2539150.2539195

Lebo, T., Erickson, J. S., Ding, L., Graves, A., Williams, G. T., DiFranzo, D., & Hendler, J. (2011). In D. Wood (Ed.), *Linking government data* (pp. 51–72). New York, NY: Springer New York. doi:10.1007/978-1-4614-1767-5_3

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., & Bizer, C. et al. (2015). DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal*, *6*(2), 167–195.

Lnenicka, M. (2015). An in-depth analysis of open data portals as an emerging public e-service. *International Journal of Social, Behavioral, Educational, Economic. Business and Industrial Engineering*, *9*(2), 589–599.

Nogales, A., Sicilia, M.-A., & Jrg, B. (2014). Combining {VIVO} and google scholar data as sources for {CERIF} linked data: A case in the agricultural domain. *Procedia Computer Science*, *33*, 266–271. doi:10.1016/j.procs.2014.06.042

Phipps, J. B. D. (2008). *Best practice recipes for publishing rdf vocabularies. W3C.* Retrieved from http://www.w3.org/TR/2008/NOTE-swbp-vocab-pub20080828

Sheridan, J., & Tennison, J. (2010). Linking uk government data. In C. Bizer, T. Heath, T. Berners-Lee, & M. Hausenblas (Eds.), Ldow (Vol. 628). CEUR-WS.org.

Shvaiko, P., Farazi, F., Maltese, V., Ivanyukovich, A., Rizzi, V., Ferrari, D., & Ucelli, G. (2012). Trentino government linked open geo-data: a case study. *Proceedings of the 11th international semantic web conference (ISWC'12)* (pp. 196-211). doi:10.1007/978-3-642-35173-0_13

Springer. (2015). LOD for conferences in computer science. Retrieved from http://lod.springer.com/wiki/bin/view/Linked+Open+Data/About)..

Springer. (2015). Springer — biomed central api portal.

Springer. (2015). Springer: Facts and figures.

Stadler, C., Lehmann, J., Höffner, K., & Auer, S. (2012). Linked geodata: A core for a web of spatial open data. *Semantic Web Journal*, *3*(4), 333–354.

Su'arez, D. S., & Jim'enez-Guar'ın, C. (2014). Natural language processing for linking online news and open government data. *Proceedings of Advances in conceptual modeling - ER '14 workshops*, Atlanta, GA, USA (pp. 243 – 252).

Villaz'on-Terrazas, B., Vilches, L., Corcho, O., & G'omez-P'erez, A. (2011). Methodological guidelines for publishing government linked data. In D. Wood (Ed.), *Linking government data*. Springer. doi:10.1007/978-1-4614-1767-5_2

Vrande˘cic, D., & Kr¨otzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85. doi:10.1145/2629489

Xin, R. S., Hassanzadeh, O., Fritz, C., Sohrabi, S., & Miller, R. J. (2013, January). Publishing bibliographic data on the semantic web using bibbase. *Semantic Web, 4*(1), 15–22.

Xin, R. S., Hassanzadeh, O., Fritz, C., Sohrabi, S., Yang, Y., Du, J., & Miller, R. J. (2010, November). Publishing bibliographic data on the semantic web using bibbase. *Proceedings of the 9th international semantic web conference (iswc-10), poster & demo track,* Shanghai, China.

Yu, L. (2007). *Introduction to the semantic web and semantic web services*. CRC Press. doi:10.1201/9781584889342

ENDNOTES

1 http://link.springer.com/
2 http://wo.kau.edu.sa/Pages-SPedia.aspx
3 http://wo.kau.edu.sa/Pages-SPedia.aspx
4 http://ontotext.com/products/graphdb/editions/
5 http://franz.com/agraph/support/documentation/current/agraph-quick-start.html

Muhammad Ahtisham Aslam is currently working as Assistant Professor at Department of Information System, King Abdulaziz University, Jeddah, Saudi Arabia. He has been working as Senior Staff Researcher at Artificial Intelligence Lab, Knowledge Technology Cluster, Malaysian Institute of Microelectronic Systems (MIMOS), Kuala Lumpur, Malaysia. He also has been working as Assistant Professor at COMSATS Institute of Information Technology, Pakistan. He did his PhD from University of Leipzig, Germany. He has several conference and journal publications in the area of semantic web and web services, knowledge extraction and knowledge engineering. His research interests are semantic web, semantic web services, web 2.0, linked open data and knowledge engineering.

Naif R. Aljohani is Assistant Professor at the Faculty of Computing and Information Technology in King Abdul Aziz University, Jeddah, Saudi Arabia. He holds a PhD in Computer Science from the University of Southampton, UK. He received the Bachelor's degree in Computer Education from King Abdul Aziz University, 2005. In 2009, he received the Master degree in Computer Networks from La Trobe University, Australia. His research interests are in the areas of mobile and ubiquitous computing, mobile and ubiquitous learning, learning and knowledge analytic, semantic web, Web Science, technology enhanced learning and human computer interaction.