Mehran University Research Journal of Engineering and Technology
Vol. 39, No. 3, 595 - 602, July 2020
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.2003.14

Assessing Large-Scale, Cross-Domain Knowledge Bases for Semantic Search

Aatif Ahmad Khan[1a], Sanjay Kumar Malik[1b]

RECEIVED ON 19.12.2018, ACCEPTED ON 03.03.2020

ABSTRACT

Semantic Search refers to set of approaches dealing with usage of Semantic Web technologies for information retrieval in order to make the process machine understandable and fetch precise results. Knowledge Bases (KB) act as the backbone for semantic search approaches to provide machine interpretable information for query processing and retrieval of results. These KB include Resource Description Framework (RDF) datasets and populated ontologies. In this paper, an assessment of the largest cross-domain KB is presented that are exploited in large scale semantic search and are freely available on Linked Open Data Cloud. Analysis of these datasets is a prerequisite for modeling effective semantic search approaches because of their suitability for particular applications. Only the large scale, cross-domain datasets are considered, which are having sizes more than 10 million RDF triples. Survey of sizes of the datasets in triples count has been depicted along with triples data format(s) supported by them, which is quite significant to develop effective semantic search models.

Keywords: Semantic Search, Knowledge Base, Cross-Domain Dataset, RDF Triples, Linked Open Data Cloud

1. INTRODUCTION

Retrieval of concerned specific information from available repositories based on an input query is called search. The information that is retrieved, also known as the result set for specified query, may or may not be relevant to the user due to lack of context understanding on the machine part. That is, result set may contain highly irrelevant responses if intent of the query is not understandable by underlying search mechanisms. Semantic Search refers to search mechanisms considering meaning of query terms and its context as a whole. For making a transition towards semantic search, information retrieval mechanisms are exploiting Semantic Web technologies along with NLP (Natural Language Processing) techniques to process the query in a machine understandable way. RDF based representation of data along with schema description using Ontology is transforming traditional information processing into knowledge processing.

To process queries intelligently, machines require proper formatting of data, large Knowledge Bases, and powerful ambiguity resolution techniques (for multi-meaning terms used in query). RDF representation of data often with XML (eXtensible Markup Language) formatting (collectively referred as RDF/XML), makes information machine interpretable. Also, with the availability of web scale knowledge repositories such as DBpedia (KB behind Wikimedia Projects), Google Knowledge Graph *etc.*, approaches are being actively developed to exploit this global range of

[1] University School of Information, Communication & Technology, Guru Gobind Singh Indraprastha University, Delhi, India. Email: [a]aatif1992@gmail.com (Corresponding Author), [b]sdmalik@hotmail.com

knowledge for processing information needs specific to them. And fortunately for the ambiguity resolution part, effective techniques such as Word Sense Disambiguation (WSD) are being advanced rapidly in the NLP domain. WSD resolves multi-meaning mappings of query terms by considering overall context of the query and deriving best mapping to meaning by analyzing rest of the query terms [1].

Data present in KBs needs to be valid across multiple domains for approaching a true web scale semantic search. This is to make sure that knowledge vocabulary for one domain should not collide with another. e.g. query term "mean" has differing interpretations in Linguistics and Mathematics domains. In the former it corresponds to the "meaning" and for the Mathematics it represents the "average of sum of numbers". Thus, there is a need for development and utilization of cross-domain KBs.

In this paper an assessment of large scale and cross-domain KBs is presented with the motivation of a formal comparative analysis of such datasets for suitability towards semantic search applications. This study is essentially a prerequisite to model and develop effective semantic search approaches of global scale, as these KBs are the backbone for deriving knowledge in ways machine can understand. For keeping the discussion compact and useful, we have shortlisted only the largest KBs in terms of data size i.e. the 25 largest datasets with more than 10 million semantic triples are only considered. In section 2, semantic search is introduced, also discussing its necessity in new age information retrieval. We have also discussed the need of KBs in semantic search process in this section. In section 3, technical discussion regarding KBs on Linked Open Data Cloud and various RDF serialization formats is presented. In section 4, we have concisely tabulated descriptions of KBs from the perspective of their usage in information retrieval and specifically in semantic search applications. Then, two key parameters regarding KB sizes and their support for serialization formats are surveyed, analyzed and depicted with pictorial representations. Finally, we conclude the paper in section 5 along with future work in this direction.

1.1 Contributions

Availability of Linked Open Data (LOD) is a practical measure of realization progress towards the Semantic Web (Web 3.0). There exists research and literature in the direction of comparative growth analysis of LOD as a whole over time (e.g. growth in number of datasets, triples counts over the years) [2]. But, no efforts are yet made to assess the LOD datasets individually to the best of our knowledge. Following are the significant contributions of this article.

- This article surveys and concisely summarizes two key parameters of Knowledge Size and Knowledge Representation for 25 largest LOD cross-domain KBs.
- Analyzes all the available RDF triples formats for their usability scope towards specific applications.
- Guides application developers to suitably select and exploit particular RDF serialization to overcome constrains such as Storage, Network Bandwidth, Universal Character Set support, Web application support etc. It further lists the supported KBs they may utilize.
- Expands the research prospects towards some less popular but global scale KBs summarizing the nature of knowledge present in them.

2. SEMANTIC SEARCH

Search approaches where machines are capable of analyzing the meaning of query and information are referred as Semantic Search approaches. Typically, semantic search includes the usage of Semantic Web Technologies such as RDF, Ontology etc. as knowledge repositories in order to make content machine interpretable, effectively improving the efficiency of search. NLP techniques such as Part of Speech Tagging, Named Entity Recognition etc. are also used to preprocess the search query. Semantic search is different from keyword-based searching in the way that it actually analyzes the concepts behind the query and its context, while keyword-based searching rely only on the effectiveness of string matching algorithms. In the literature, keyword-based searching is also referred as navigational search, and the searching with conceptual clarity as research search [3].

2.1 Need of Semantic Search

The World Wide Web (WWW) introduced the searching on the internet with approaches based on keyword matching. As the web expanded, approaches are modified in terms of efficiency but still maintaining the keyword-centric methodology. But, at its present Big Data age, information is overloaded on the web with issues of inconsistency and redundancy. Now, if keyword-based approaches are used alone, result set will suffer in terms of precision of results. Hence, modern web search providers (including search engine giants Google and Bing) have started to use semantic search elements as additional parameters in their web search offerings. Table 1 tabulates the issues with keyword-based approaches on web scale information retrieval and their remedies with semantic search.

2.2 Knowledge Bases for Semantic Search

Semantic search approaches utilize machine interpretable knowledge contained in KBs to process the query; get context out of the query; use the derived context to search conceptually similar information on target repositories; and finally, present the retrieved results. Most of the KBs contains knowledge in the form of RDF triples, making it machine understandable. As RDF data triples are represented in <subject, predicate, object> form, machine processing has an extra formal metadata in the form of predicate to derive conceptual relationships among query terms and other concepts. By matching the target results conceptually, semantic search yields increased precision and hence, relevancy in result set for that specified search query.

3. KNOWLEDGE BASES

KB are data repositories containing machine interpretable information i.e. the knowledge. KBs utilizing Semantic Web technologies represent data in the form of RDF triples, which are often structured in XML, and also in other serialization formats as shown in Section 3.2

TABLE. 1: ISSUES WITH KEYWORD-BASED SEARCH VS. SEMANTIC SEARCH

Issue	Keyword-based Search	Semantic Search
Tremendous Information Availability	Keyword-based information retrieval produces low precision results due to availability of tremendous information in the ever increasing web.	Semantic search does not depends on the size of target information repositories, instead it analyzes the concepts in search query.
Inconsistent Information	Inconsistent information at multiple sources provoke the need of trustworthiness of information sources.	Semantic search relies on the data facts as available on underlying KBs. Hence, it has very little scope for knowledge inconsistencies.
Redundant Information	Availability of similar information at multiple sources effectively doesn't improve the quality of result set. It just increase its size.	Availability of similar information at multiple sources doesn't make it semantically different. They resolve to very same concepts.
Usage of Ambiguous terms in queries	This is the key issue for irrelevant results in the result set. Often, machines fails to interpret the correct conceptual usage of terms that have mapping to multiple meanings at different contexts. (E.g. "Mean").	Resolving the ambiguity among concepts is a primary step in semantic search processing. Typically, NLP techniques are utilized here extensively.
Usage of linguistic variations (synonyms, plurals etc.)	Result set of keyword driven search approaches heavily depends on spellings of keywords. It also changes tremendously on usage of plurals or synonyms in search query.	For semantic search approaches, linguistic variations in terms are preprocessed and resolved to corresponding concepts. Also, popular KBs already have synonym and spell variation mappings.

3.1 Linked Open Data Cloud

KBs specific to a single domain are of very little use for large scale semantic search. Also, it is highly unlikely to describe all the domain using a single schema definition due to domain modelling constraints. Hence, the need to interlink vocabularies and data across multiple datasets was identified very early in the form of LOD. LOD contains inter-dataset linkage Internationalized Resource Identifiers (IRIs) in addition to multiple KBs. LOD Cloud represents the published KBs in Linked Data format. As of June 2018, this massive cloud has more than 1200 KBs with around 16000-KB linkages. LOD Cloud has numerous cross-domain KBs along with a large number of domain specific KBs pertaining to geography, governments, life sciences, linguistics, media, publications, social networking and user generated datasets. Interlinking IRIs among these sets essentially shows the intent of heavy reuse of knowledge available in them.

3.2 Dataset Formats

RDF triples are represented in various data serialization formats depending on constraints on storage and processing power. Triple formats specified by WWWC World Wide Web Consortium (WWWC) are tabulated in Table 2 also listing popular KBs that represent their data using these formats.

3.2.1 Significance of RDF Triples Serialization Formats

Representation of RDF triples in various formats is a result of need of efficiently processing huge amount of data by different applications. These format limits the nature of applications that may use these datasets. Major factors include size of data, Unicode support, bandwidth requirements and web application support. Storage and processing of such huge amount of data is a constraint for any system (e.g. Freebase KB, the fifth largest dataset in our list has over 220 GB (Giga Bytes) of data). Hence, for accessing such amount of data, developer may develop web applications and utilize web-friendly XML and JSONLD formats for efficient processing. Further, some formats are human readable making it easier for developers to debug their application code.

For most applications, RDF/XML is preferred as it is supported by most programming languages, further reducing the size of triples using namespaces instead of full Universal Resource Identifiers (URIs). Turtle is more developer friendly in terms of readability and hence, debugging. Also, it is much efficient for low bandwidth connections over RDF and supports Unicode character set. JSONLD is the most convenient and efficient format for processing in JavaScript web applications. N-Triples is easily for

TABLE 2: RDF TRIPLES SERIALIZATION FORMATS

Format	Extn.	Description	Popular KBs
Turtle	.TTL	Compact plain text format. <Subject (S)> <Predicate (P)> <Object (O)>	WikiData, YAGO, NPM
N-Triple	.NT	Line-based & plain text format extended from Turtle format. <S> <P> <O> <Full Stop (.)>	Freebase, DBkWik, EPA (all)
Notation3 (N3)	.N3	Superset of RDF with assertions and logic. Further, adds formulae, variables, logic and functional constructs.	Muninn World War 1
N-Quad	.NQ	Line-based & plain text format. <S> <P> <O> <IRI for graph triple belongs to>	DBpedia, EventKG, WebIsALOD
JSON-LD	.JSONLD	JavaScript Object Notation (JSON) based format intended to use Linked Data for interoperable Web Services.	GND, WikiData
RDF/XML	.RDF	XML based formatting for RDF.	Data.gov, Open Library
TriG	.TRIG	Human readable natural text format abbreviating datatypes and usage patterns.	WikiPathways

parser efficiency as it does not contain any file starts or line endings. But, it does not support Unicode, hence some characters present in triples may get escaped.

4. ASSESSMENT OF CROSS-DOMAIN KNOWLEDGE BASES ON LOD CLOUD

Availability of global knowledge repositories is a prerequisite to large-scale semantic search, as we need semantic relationships between search query keywords with other keywords present in target web documents. Such KB are populated over significant time and are maintained by experts. Domain of web search being generic, can only be fulfilled using cross domain datasets. In this assessment, we have shortlisted 25 largest of the cross-domain KBs on LOD Cloud that have more than 10 million triples. Selection is made considering the amount of knowledge available in them for targeting large-scale semantic search. Their brief description is tabulated in Table 3.

A true web scale semantic search solution needs to process most of the knowledge available on web by exploiting the global repositories of interlinked cross-domain LOD. These repositories contains machine understandable descriptions for entities and their semantic relationships with other entities. Semantic search approaches utilize such relationships to derive semantic similarity between search query keywords and other keywords present in web pages. Web pages having larger semantic similarity are flagged as relevant responses to search query, and are returned as top results. We have surveyed and analyzed two important parameters for evaluating a dataset:

(a) Amount of knowledge present (measured in Semantic RDF Triple counts)
(b) Representation of knowledge present (represented by RDF triples serialization format).

Former increases the domain and range of search, and the latter is required for designing and developing applications considering storage and processing constraints. Fig. 1 depicts the survey of comparative sizes of all 25 surveyed KBs. DBpedia is indeed the most valuable dataset being generic as well as multilingual. Data.gov catalogue being second largest in size provides data in divided sets as categorized by US government. WikiData is the third largest cross-domain KB, and is heavily used in real world search applications due to its very close proximity to human readable Wiki articles.

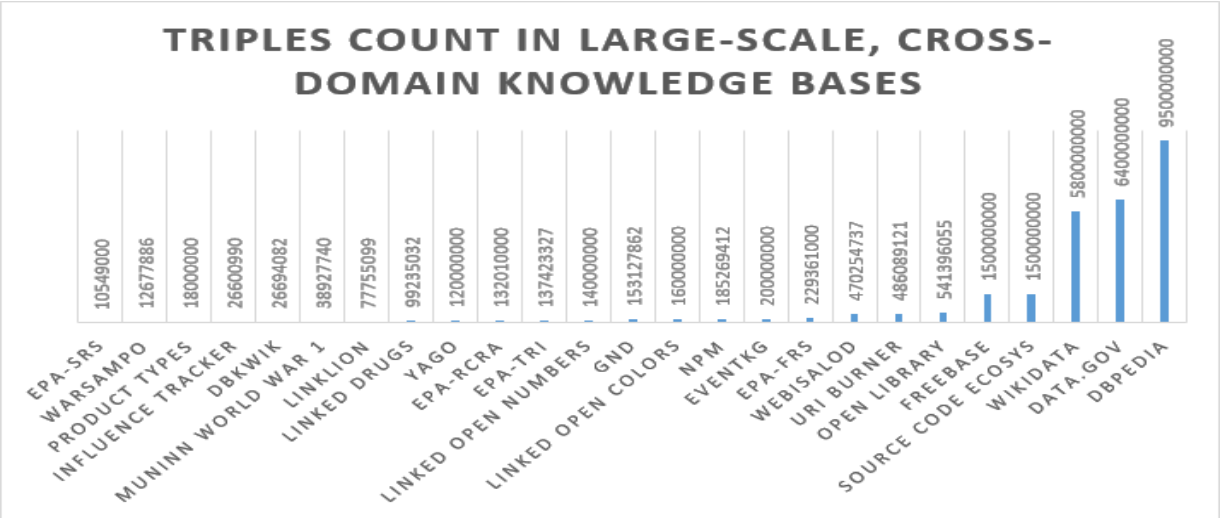

FIG. 1. SURVEY DEPICTING SIZES OF LARGE SCALE CROSS-DOMAIN DATASETS IN TRIPLES COUNT (ONLY THE DATASETS WITH TRIPLES COUNT MORE THAN 10 MILLION ARE CONSIDERED).

TABLE 3: 25 LARGEST CROSS-DOMAIN KBS ON THE LOD CLOUD

Dataset	Description
DBpedia	It is a community-driven dataset populated by extracting structured information available in multiple Wikimedia projects. It is largest multilingual cross-domain KB which is actively exploited by range of semantic search approaches due to its generic vocabulary and largest domain-specific data collection also [4].
Data.gov	US government documents converted to RDF and categorized into 417 datasets pertaining to different aspects. It is the largest Open Government Dataset [5].
WikiData	WikiData KB focused on structuring and linking of data extracted from Wikipedia, the free encyclopedia. It maintains facts from data present in Wikipedia articles [6].
Source Code Ecosystem	KB of collected facts about source code from open source projects on the web. Facts are extracted at different levels of syntax and semantics of the code [7].
Freebase	Freebase KB was designed as wiki for structured content on the web. At present, its data is migrated to WikiData. Its last data dump is still one of the largest KBs available. Hence, applications use it for knowledge which is time invariant [8].
Open Library	It contains structured data about most of the books ever published globally. Knowledge can be derived about authors, editions etc.
URIBurner	KB populated through conversion of global databases into Linked Data Objects.
WebIsALOD	KB containing collections of hypernymy relations. Hypernymy is property of a superordinate to a subordinate. E.g. Color is hypernym for Red and Blue.
EventKG	KB with facts pertaining to various events as recorded in Wikipedia articles [9].
NPM	KB containing structured data from NPM repository, the largest registry of software containing tons of reusable code packages especially for JavaScript [10]
Linked Open Colors	Structured data repository for facts about colors.
Gemeinsame Normdatei (GND)	Catalogue for authority files pertaining to people, corporations, Geographic information, works, events etc. It is derived from German Integrated Authority File and has data from German National Library on these subjects [11].
Linked Open Numbers	KB containing billions of facts about numerals. These include numeral usage in multiple languages and relations with other number systems (binary, hex etc.) [12]
EPA (FRS, RCRA, SRS and TRI)	It contains datasets about biomedical chemicals manufactured and their recorded effects for protection of human health and the environment. This KB is majorly used in medicine domain but also has cross linkages to other global KBs.
YAGO	It is a massive semantic repository for people, organization and geographic data.
LinkedDrugs	Structured data about medicines (drugs) from 23 countries [13].
LinkLion	Central KB for storing links of resources available on Linked Open Data [14].
Muninn World War 1 Dataset	Multi-disciplinary and multi-national KB with millions of investigation records from World War 1 archives.
DBkWik	Single consolidated KB derived out of thousands of Wikipedia articles [15].
Influence Tracker	Social Networking knowledge repository for tracking influence of individual users on Twitter microblogging website [16].
Product Types Ontology	Repository providing definitions to 0.3 million products described in various Wikipedia articles [17].
WarSampo	LOD KB resulted by transforming Finnish World War 2 data archives [18].

Second parameter, the representational format for KBs limits the scope of applications that may utilize these sets. Most of these large KBs have Web-based APIs (Application Programming Interfaces) for querying and retrieving knowledge. Hence, the repositories providing data dumps in XML or JSON formats are more Web Services friendly. Also the availability of datasets in multiple formats increases the range of applications that may exploit knowledge in them. Fig.2 depicts the survey of availability and support for various RDF serialization formats by all of surveyed KBs. Although, WikiData has lesser size in triples, but availability of its datasets in multiple formats makes it open to be used by applications of varying nature (e.g. web-based) and varying computation processing constraints. It should also be noticed that although

most datasets use Turtle and RDF/XML formats, but focus is shifting towards their conversion into N-Triples and N-Quads. This is due to the fact most KBs needs to maintain linking compatibility with ever increasing DBpedia triples which are exclusively available in N-Triple and N-Quad format.

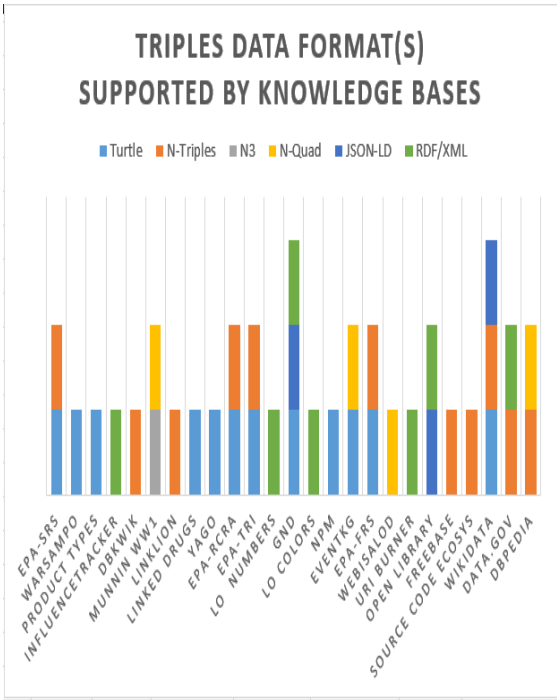

FIG. 2: SURVEY DEPICTING TRIPLE DATA FORMAT(S) SUPPORTED BY KB

Different RDF triples serialization formats have certain support and efficiency constraints hence may not be utilized for all kind of Semantic Search applications as described in Section 3.2.1. Storage and processing of such huge amount of data is itself a primary constraint for end user system. Hence, most of the LOD KBs support web-based APIs along with SPARQL Protocol and RDF Query Language (SPARQL) endpoints for querying these datasets. But, developers must target the proper serialization format and hence the supported KBs for optimizing bandwidth, network latency and internationalization (Unicode Character Set) support.

5. CONCLUSION AND FUTURE WORK

KB are the backbone for deriving semantic relationships among keywords used during web search. In this paper, a concise assessment across 25 largest cross-domain knowledge bases available as Linked Open Data is presented. Surveyed datasets are analyzed and compared across two key parameters of Knowledge Size (in triples count) and Knowledge Representation (RDF triples Serialization format). Knowing the nature of data available and their efficiency constrains may aid application developers to target their applications for suitable formats and datasets. Survey results are analyzed and depicted with pictorial representations in Section 4. DBpedia KB is found to be most valuable for web scale semantic search applications being the largest and having maximum linkages from other KBs. Also, WikiData KB, has wider application support due to availability of its multi-format data dumps.

LOD Cloud is not just limited to cross-domain knowledge bases but also has linkages with datasets pertaining to specialized domains of geography, governments, life sciences, linguistics, media, social networking, and publications among others. As part of the future work, this work can be expanded towards a comprehensive survey across all the knowledge bases available and linked on the LOD Cloud. This work may also be expanded to study various large scale semantic search applications to analyze state of art research towards global semantic search solutions.

ACKNOWLEDGMENTS

This publication is an outcome of the R&D work undertaken project under the Visvesvaraya Ph.D. Scheme of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation.

REFERENCES

[1] Stevenson, M., and Wilks, Y., “Word sense disambiguation”, The Oxford Handbook of Comp. Linguistics, pp. 249-265, 2003.

[2] Ermilov I., Martin M., Lehmann J., and Auer S. “Linked Open Data Statistics: Collection and Exploitation”, In: Klinov P., Mouromtsev D. (eds) Knowledge Engineering and the Semantic Web. KESW 2013. Communications in Computer and

Information Science, Vol. 394. Springer, Berlin, Heidelberg, 2013.

[3] Guha, R., McCool, R., and Miller, E., "Semantic search", In Proceedings of the 12th international conference on World Wide Web, ACM, pp. 700-709, May, 2003.

[4] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S. and Bizer, C., "DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia", Semantic Web Journal, Vol. 6, No. 2, pp. 167-195, 2015.

[5] Ding, L., DiFranzo, D., Graves, A., Michaelis, J., Li, X., McGuinness, D. L., and Hendler, J., "Data-gov Wiki: Towards Linking Government Data", In AAAI Spring Symposium: Linked data meets artificial intelligence, Vol. 10, pp. 1-1, March, 2010.

[6] Vrandečić, D., and Krötzsch, M., "Wikidata: a free collaborative knowledgebase", Communications of the ACM, Vol. 57, No. 10, pp. 78-85, 2014.

[7] Keivanloo, I., Forbes, C., Rilling, J., and Charland, P., "Towards sharing source code facts using linked data", In Proceedings of the 3rd International Workshop on Search-Driven Development: Users, Infrastructure, Tools, and Evaluation, ACM, pp. 25-28, 2011.

[8] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J., "Freebase: a collaboratively created graph database for structuring human knowledge", In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, ACM, pp. 1247-1250, 2008.

[9] Gottschalk, S., and Demidova, E., "Eventkg: A multilingual event-centric temporal knowledge graph", In European Semantic Web Conference, Springer, pp. 272-287, 2018.

[10] Van Herwegen, J., Taelman, R., Capadisli, S., and Verborgh, R., "Describing configurations of software experiments as Linked Data", In ISWC2017, the 16e International Semantic Web Conference, Vol. 1931, pp. 1-8, 2017.

[11] Haslhofer, B., and Isaac, A., "data. europeana. eu-The Europeana Linked Open Data Pilot", 2011.

[12] Vrandecic, D., Krötzsch, M., Rudolph, S., and Lösch, U., "Leveraging non-lexical knowledge for the linked open data web", Review of April Fool's day Transactions (RAFT), Vol. 5, pp. 18-27, 2010.

[13] Jovanovik, M., and Trajanov, D., "Consolidating drug data on a global scale using Linked Data", Journal of biomedical semantics, Vol. 8, No. 1, p. 3, 2017.

[14] Nentwig, M., Soru, T., Ngomo, A. C. N., and Rahm, E., "Linklion: A link repository for the web of data", In European Semantic Web Conference, Springer, pp. 439-443, 2014.

[15] Hofmann, A., Perchani, S., Portisch, J., Hertling, S., and Paulheim, H., "DBkWik: Towards knowledge graph creation from thousands of wikis", In Proceedings of the International Semantic Web Conference (Posters and Demos), pp. 21-25, 2017.

[16] Razis, G., and Anagnostopoulos, I., InfluenceTracker: Rating the impact of a Twitter account, In IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, pp. 184-195, September, 2014.

[17] Hepp, M., "Products and services ontologies: a methodology for deriving OWL ontologies from industrial categorization standards", International Journal on Semantic Web and Information Systems (IJSWIS), Vol. 2, No. 1, pp. 72-99, 2006.

[18] Hyvönen, E., Tuominen, J., Mäkelä, E., Dutruit, J., Apajalahti, K., Heino, E., Leskinen, P. and Ikkala, E., "Second World War on the Semantic Web: The WarSampo Project and Semantic Portal", In International Semantic Web Conference (Posters & Demos), 2015.