Aragon Open Data: A NLP-Based ontology population for Public Administration

Paula Peña
Big Data and Cognitive Systems
Technological Institute of Aragon (ITAINNOVA)
Zaragoza, Spain
ppena@itainnova.es

Rafael del Hoyo
Big Data and Cognitive Systems
Technological Institute of Aragon (ITAINNOVA)
Zaragoza, Spain
rdelhoyo@itainnova.es

Rocío Aznar
Big Data and Cognitive Systems
Technological Institute of Aragon (ITAINNOVA)
Zaragoza, Spain
raznar@itainnova.es

Rosa Montañés
Big Data and Cognitive Systems
Technological Institute of Aragon (ITAINNOVA)
Zaragoza, Spain
rmontanes@itainnova.es

Abstract—Governments are one of the main producers of information and initiatives of making government open data are continuously gaining interest, recently. In this sense, how to provide relevant elements to improve the structuring of institutional information and deepen its standardization to make easier its access and make use of the information has become an important challenge and allows a great value to initiatives that leverage the potential of Open Data. Within this context, given the amount of unstructured information related to the Government of Aragon published on Internet at this time, with slightly or no standardization and decentralized, the need to gather it systematically to be served from a single access point in a public and structure way emerges. In this paper, an approach based on the use of web crawling, Natural Language Processing and ontology-based techniques is proposed. Experimental results based on disperse, non-homogeneous, uncontrolled and non-exploitable institutional information validate that the proposed approach improves the structuring and standardization of data that can be analyzed together, be accessible, be browsed, be exploited and offered to all interested collectives. The generated ontology is populated with institutional data, affording and securing new possibilities of generating valuable and consistent ontology-based knowledge services.

Index Terms—Open Data, Ontologies, Natural Language Processing

I. Introduction

Open Data is now a worldwide movement whose philosophy aims to provide data openness and availability to citizens. Many countries have an Open Data initiative[1]. Even in countries where that data is not yet fully open, civil society and the technology community are using government data. In Europe, Open Data has been a focus for policymakers for over a decade[2]. Today, almost all European countries have an Open Data portal, and these portals are becoming more advanced, being used more frequently and creating more benefits for citizens. As part of a national Open Data initiative, regional and provincial governments and by councils at city level have not been unresponsive to these initiatives.

Aragon Open Data is a project to open data by agreement of July 17, 2012 of the Government of Aragon, and its Internet portal[3] was presented with the objectives of creating economic within the ICT sector through the reuse of public information, increasing transparency in the Administration, promoting innovation, improving information systems of the Administration, adopting technical standards in the information society field and generating data interoperability between public sector websites. Throughout this time, there have been numerous works that allow the incorporation of new data and information available to third parties (citizens, companies, infomediary sector, developers, etc.).

In this context, on one hand, given the volume of the data that begins to exist in the Aragonese public administration, are beginning to have a special relevance all tools that help in the improvement of the structuring the information and in the standardization of data. The idea of generating a set of technical and legal rules that allow deepening in that standardization arises. On the other hand, within the line of difficulty of access and use of information, the complex casuistry presented by the regional public administration in the generation of data and information is reflected in the proliferation of a large number of websites under its root domain: `aragon.es`. These circumstances make it difficult to access and make use of information by third parties as well as public administration services of Aragon, generating the popular sense of certain lack of transparency.

This work introduces an approach to allow institutional information that is in a dispersed, non-homogeneous, non-controlled and non-exploitable way become into structured data that can be analyzed, accessible, reused, browsed, ex-

---

[1]Barometer Global Report, https//opendatabarometer.org, last accessed 2018/05/25

[2]The Re-use of Public Sector Information Regulations, http://www.legislation.gov.uk/uksi/2015/1415/made, last accessed 2018/09/10

[3]Aragon Open Data, http://opendata.aragon.es, last accessed 2018/09/10

ploited and served in a controlled mode under a single point. A semantic ontology has been designed and created to structure and standardize public administration information. This ontology, known as *Interoperable Information Scheme of Aragon (EI2A)*, allows modeling organizational structures, natural persons, places, territories, temporal properties of resources, events, websites, etc. In addition, a prototype has been implemented to recover open data related to the organizational chart of the Government of Aragon and all the information offered on institutional web to structure it according to EI2A ontology. Information conversion into knowledge and useful data is performed through the application of techniques and technologies such as *web crawling* or *spider*, intelligent data processing, Natural Language Processing (NLP), semantic technologies for the representation of knowledge and Big Data architecture. Specifically, the approach exploits text processing and automatic population technologies to extract knowledge from institutional content and organize it conceptually in EI2A ontology.

This paper is organized as follows: Related work on the use of semantic and NLP techniques to extract knowledge and structure it is presented in Section 2. Section 3 describes the architecture used and presents a high level description of proposed approach. EI2A ontology used is described in section 4. We discuss results in section 5 and highlight implications and of the work made. Finally, section 6 describes conclusions and discusses ideas for future work.

II. RELATED WORK

Initiatives of making government open data are continuously gaining interest, recently. While more countries are embracing the Open Government paradigm, among the researchers working with those data there is an increasing awareness in using semantic techniques to represent them. Applying semantic web technologies enables data integration among different organization and established links to interconnect data on the Web [1]. Integration of raw data gathered from different sources, and formally and semantically represented and based on ontology, leads to opportunities for information exchange, analysis on combined datasets, simplicity in creation of mash-ups, and exploration of innovative ways to use data creatively. Enhancing with application of semantic web technologies to link data, and provide unexpected and unexplored insights into different domains and problem areas is the real value of open government data [2].

Application of semantic web technologies in government not only contributes to opening data to the community but also contributes to creating a knowledge network map that interconnects different sources that actually share data [3]. By making government data discoverable and accessible to the public, government shows that they are dedicated to the application of innovative e-services to improve accessibility, reusability and easy consumption of their data [4]. This leads to extends characteristics of publicly available data [5]. On the other hand, ontologies have become a major tool for developing semantically rich application. They provide for a standardized means of modeling, querying, and reasoning over large knowledge bases. Specialized knowledge services require tools such as [6] that can search and extract specific knowledge directly from unstructured text on the Web, guided by ontology. Other researchers have used ontology to support knowledge extraction [7], [8].

In a world in which the majority of knowledge is encoded in natural language text, automating the population of these ontologies using results obtained from the application of NLP techniques is becoming increasingly important and a major challenge [9]. The process of automatic or semi-automatic construction, enrichment, population and adaptation of ontologies is known as ontology learning. Ontology population is the task of adding new instances of concepts. Different approaches exit to export or populate results of NLP analyses into ontology such as [10], [11]. In addition, NLP techniques for term extraction and ontology population, using a combination of rule-based approaches and machine learning is described in [12]. In [13] links a knowledge-extraction tool with ontology to achieve continuous knowledge support and guide information extraction.

This paper aims to focus on the new possibilities afforded by semantic web and natural language technologies in the area of knowledge management and open government data to allow institutional information from very diverse, scattered and different sources to be structured formally and semantically to facilitate its analysis, accessibility, reuse and exploitation from a single access point in a public way.

III. PROPOSED APPROACH

The high-level architecture of the proposed approach (see Fig. 1) is primarily focused on information retrieval. Particularly, textual information of websites under `aragon.es` domain using *web crawling*, *spidering* or *spider* techniques and data related to the organizational chart of the Government of Aragon are retrieved. Secondly, the extracted textual information is analyzed/processed using NLP techniques for lowercase conversion, lemmatization, stopwords filtering, the dictionaries application, cleaning tasks, recognition and classification of concepts and summarization. Lastly, results are stored into NoSQL databases that allow the indexation of rich text and structuring information according to the EI2A ontology for later consultation through the Aragon Open Data access point.

To implement the functionalities of the proposed approach, a set of sub-processes has been developed through an own framework called *Moriarty*[4] [14]. Moriarty is and advanced Artificial Intelligence software solution framework for Big Data, developed by ITAINNOVA[5]. It is based on two basic concepts: *workitem* (class that implements an atomic function, and that can be used in multiple contexts reuse concept) and *workflow* (composed of workitems or other workflows that receive some inputs and perform transformations in them generating and returning some outputs).

---

[4]Moriarty, http://www.ita.es/moriarty, last accessed 2018/09/10

[5]Instituto Tecnológico de Aragón, http://www.itainnova.es, last accessed 2018/09/10

Fig. 1. High-level architecture.

Through the execution of a periodic process, open data related to the Government of Aragon organizational chart (legislatures, entities, roles) are accessed in reusable formats (*xml*, *csv*, *json*) from Aragon Open Data portal with the aim of structuring them according to the EI2A ontology. It is initially populated with this data and is available as RDF (*Resource Description Framework*). Since web information could change regularly and new web pages may appear, a web crawling process is executed periodically (see Fig. 2), analyzing the new webs that emerge or reprocessing the webs that have changed, and thus to have the information updated. URL information of websites in an updated *csv* file is considered as the seed of this approach. *ExtractURLsFile* sub-process extracts them in order to analyze and process them. Initialized the necessary variables to carry out the entire web crawling process in *CsvToText* sub-process, a list of URLs filtered to avoid images, *css* or unauthorized links is created with the depth of analysis, the maximum number of pages to analyze and the number of crawling-threads desired in *Crawling* workitem. It uses *crawler4j*[6] library, presented as the most suitable available for the recovery of institutional information offered in websites, after a thorough analysis of the current state on research trends and technological development in relation to web crawling techniques and after developing and following a methodology with the objective of applying selection criteria.

With the aim of granting a unique identifier to the new pages instances to be processed and analyzed, *GetMaxURLID*

sub-process returns the maximum identifier inserted in *Openlink Virtuoso* database. Url-to-url, *ExtractTextFromURL* sub-process extracts the text information whether it is HTML or a file (pdf, doc or docx). A series of rules are applied that allow obtaining a clean text. *CheckWebPagesChanges* is responsible for detecting if a website has changed and it is necessary to reprocess. For this purpose, a CRC *(Cyclic Redundancy Check)* code is calculated from the text. *LoadDataURL* sub-process incorporates real information to a web, subdomain or portal as RDF triplets in *Virtuoso* (e.g. *ei2a:idWebPage rdf:type ei2a:WebPage, ei2a:idWebPage rdf:type owl:NamedIndividual, ei2a:idWebPage ei2a:URL url*). In *UpdateText* sub-process, after text preprocessing, in which some common task is performed such as lowercase transformation, stemming or stopwords filtering, the main NLP techniques used on textual data involve semantic classification based on thesaurus, text summary and recognition and classification of named entities (NERC). The text summary task offers a synthesis of the textual information with the most relevant sentences by means of graph-based ranking algorithms [15]. This avoids having to read the whole web to know what it is about. The NERC task, implemented in *PredictionNerSpark* sub-process, identifies, extracts and classifies implicit information of the texts related to people, organizations and locations that are named in them. This is possible thanks to the use of neural network algorithms [16]. In this case, the neural network known as the *Multi-Layer Perceptron (MLP)* has been used. To apply the mathematical model in charge of recognizing and classifying named entities, textual information is transformed

---

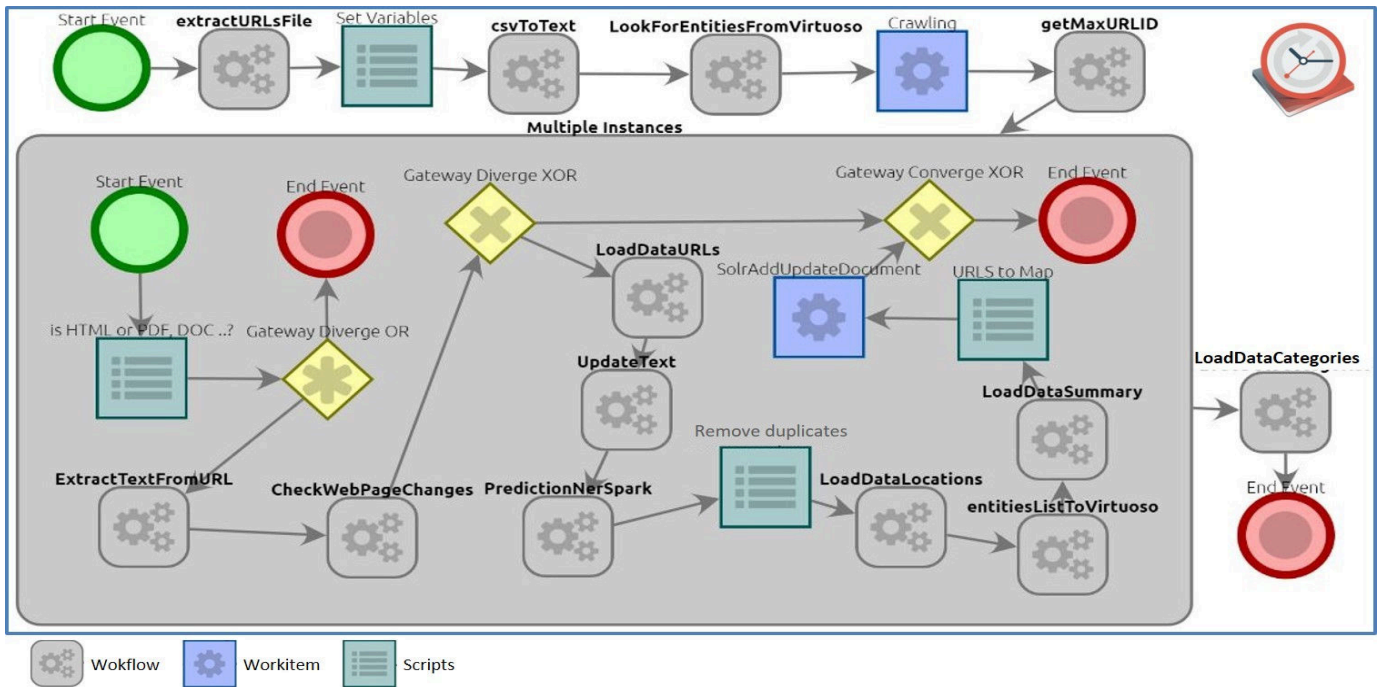[6]Crawler4j library, https://github.com/yasserg/crawler4j, last accessed 2018/09/10

Fig. 2. Web crawling process.

into numerical representation, post-processed tasks are applied and a training set (or corpus) of formal informative type (news) has had to be built so that the model learns.

Identified the technical-legal framework applicable to the use of web crawling techniques, a methodology has been elaborated and followed for the extraction, treatment, storage, publication and reuse of institutional web information. The methodology takes into account aspects of crawling exclusion (file *robot.txt* and terms of use), intellectual property and protection of personal data. In fact, due to legal aspects, people extracted from the NERC technique, which do not belong to the organizational chart of the Government of Aragon (persons obtained through *LookForEntitiesFromVirtuoso* sub-process), is anonymized in the summary generated with asterisks. After that, extracted knowledge is stored in a structured and controlled manner in *Openlink Virtuoso* and *Solr* NoSQL databases for further exploitation. In *Openlink Virtuoso*, *LoadDataLocations* sub-process inserts through triplets all the locations previously extracted with the NERC technique, *entitiesListToVirtuoso* sub-process adds the citations, and that is, information regarding an entity (person, organization or location) is cited on a web, subdomain or portal, and *LoadDataSummary* sub-process introduces information related to the website summary. Finally, the url, CRC, categories in which a website is classified, the original text, the processed text, the current date, the summary and the persons, organizations and locations identified in the processed text, is stored in *Solr* database through the *SolrAddUpdateDocument* sub-process. Finally, the *LoadDataCategories* sub-process inserts in *Openlink Virtuoso* the data relative to the categories in which a website is classified.

IV. EI2A ONTOLOGY

In the government of Aragon an important volume of data begins to exist, and those elements such as the ontologies that contribute to the improvement of the structuring of the information and to the standardization of the data are increasingly important and relevant in the field of automation in information management systems. EI2A emerges with the main idea of generating a framework (a set of technical and legal rules) in which the open data and regional government data in general can begin to be automated in a much deeper way.

Borrowed from philosophy, ontology is a term that refers to an explicit and formal specification of a shared conceptualization [17]. It details the concepts and relationships of a domain allowing analyzing, sharing and reusing common knowledge. Ontology populated with high-quality instantiations provides a consistent knowledge base. With the knowledge of regional public administration experts, EI2A ontology has been defined and created using a process that follows the methodological guidelines contained in [18], [19]. Based on the philosophy of the Semantic Web, well-known ontologies, schemes and vocabularies such as [20]–[25] endorsed by European directives (INSPIRE)[7] and International Consortium (W3C)[8] have been reused to model EI2A ontology.

Institutional data of common interest identified and modeled are focused on concepts related to describe organizational structures, natural persons, any places in terms of its name, address or geometry, geolocations, territories, events, temporarily and documents (including web pages). In this way,

---

[7]Inspire, http://www.idee.es/europeo-inspire, last accessed 2018/09/10
[8]W3C, https://www.w3.org/, last accessed 2018/09/10

the ontology provides concepts and relationships to support the representation of a very broad range of organizational structures. It is possible to describe the notion of an organization, the decomposition into sub-organizations, departments and units and the purpose and classification of organizations. In addition, membership and reporting structure with an organization: roles, posts and relationships between people and organizations. Furthermore, locations information at which organizations exist (sites or buildings) and organizational history (merger, renaming). A Government of Aragon domain-specific extension has been added to model the nature of an organic unit or office in Aragon (level of administration, public or private character, etc.). EI2A model has been enriched with aspects and metadata of DIR[9] and ENI[10]. On the other hand, motivated by the need of automatic way to extract, structure and standardize information from the huge amount of textual content available on the institutional websites, EI2A ontology provides concepts and relationships to specify semantically information related to a recognized entity (person, organization and/or location) has been cited on a web classified under a categorization of Government of Aragon themes.

V. Results

To evaluate our work of gathering, analyzing, processing, storing and keeping updated unstructured institutional information to be offered to all interested collectives from a si ngle access point in a public and structured way, 667 websites are been crawled in order to extract text content. In this section, results from a sampling of five websites (http://www.educaragon.org, http://transparencia.aragon.es, https://www.turismodearagon.com, https://www.saludinforma.es and https://inaem.aragon.es/), considered of greater relevance for the Government of Aragon and as seed of this approach, are presented as part of a larger experiment designed. The system is ready to add as websites as required. Through a periodic web crawling process, with a depth of 4 and 10 crawling-threads, 3,963 urls (HTML or a file) have been processed. Data extracted has been cleaned applying customized meta-data removing rules such as headers, footnotes or indexes, and the texts have been processed (lowercase transformation, lemmatization or stopwords filtering) to prepare them for later application of NLP techniques (thesaurus-based semantic classification, named entities recognition and classification and, summarization).

To apply EI2A scheme on real data, the ontology has been populated with information from each institutional website. *Person*, *Organization*, *Site*, *Address*, *Location*, *Temporal Entity*, *WebPage* and *Citation* concepts have been populated with more than 95,978 new instances and relationships generated as triplets (subject-relation-object). Some of them provided by the NERC process. For example, information related to a recognized entity (person, organization and/or location) has been cited on a web classified under a categorization is specified semantically. In addition, data related to the url, the date of textual web content capture and the summary is added to EI2A. On the other hand, EI2A has been populated with information of the organizational chart of Government of Aragon extracted from a dataset[11] of Aragon Open Data portal. In this way, semantic information is added to indicate the nature of a person′s membership of an organization, that is to say, that a person belongs to a unit or department with a specific role in a valid time interval. Results from extracting knowledge of web textual content are stored in a structured and controlled mode into *Solr* and *Openlink Virtuoso* NoSQL databases allowing future open access and simple data exploitation. Thanks to a web interface, as a semantic search engine in which it is able to understand through natural language type questions and generate appropriate responses to the context, it is possible to search information on the generated knowledge base and to browse through the ontological model EI2A. For example, Fig. 3 shows information about where a specific person is working and on which websites is cited: 'Francisco Javier Lambán' is President of Aragon since July 4, 2015 to the present, in the department 'The presidency' that has its headquarters in Paseo María Agustín, 36. This person has been cited on a pdf document classified in the *Dependency* and *SocialCareAndDependency* categories.

Experimental results based on institutional information validate that the proposed approach improves the structuring and standardization of decentralized data that can be processed and offered to all interested collectives through the application of web crawling, NLP and ontology-based techniques. EI2A has been populated with many high-quality instantiations, affording and securing new possibilities of generating valuable and consistent ontology-based knowledge services.

VI. Conclusions and Future Work

Governments are one of the main producers of information, which, in the exercise of their functions, create, collect, treat, store, distribute and disseminate large amounts of information of different fields. How to provide relevant elements that allow improving the structuring of institutional information and deepening its standardization to make easier its access and make use of the data has become an important challenge. In this paper, an approach based on ontology and the use of NLP techniques is proposed.

Despite dealing with texts with a great diversity of domains and formats, the work carried out manages to integrate a generic system capable of fulfilling the expectations presented at the beginning. Although improvements have been identified related to define new entities according to the context, the application of a more complex text pre-processing and the generation of different mathematicians models according to the text context would also imply the improvement in the extraction of information and in the quality of the structured and stored data, results obtained are significantly satisfactory.

[9]DIR, https://administracionelectronica.gob.es/ctt/dir3, last accessed 2018/09/10

[10]ENI, https://administracionelectronica.gob.es/ctt/eni, last accessed 2018/09/10

[11]Organigram, https://opendata.aragon.es/datos/catalogo/dataset/organigrama-del-gobierno-de-aragon, last accessed 2018/09/10
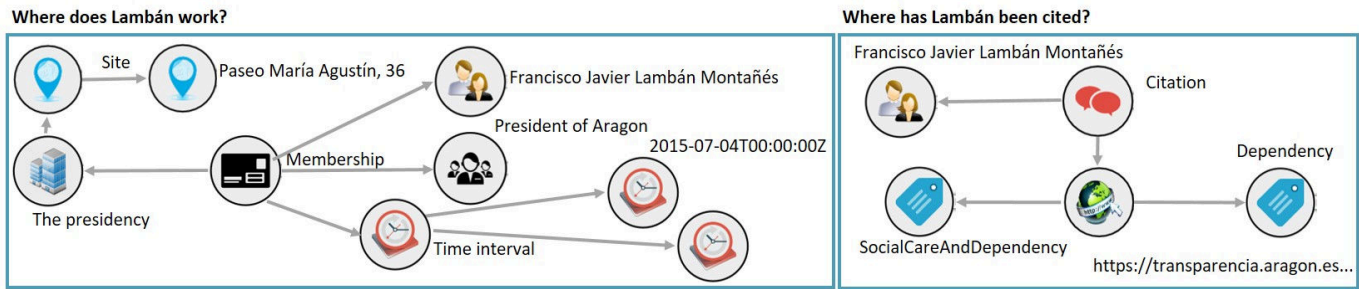
Fig. 3. Browser for EI2A ontology.

In this context, the viability of the proposed approach has been verified and new aspects have been detected in which it is necessary to continue exploring. The aim of the future work is to deploy the solution over the public Aragonese infrastructures in order to develop on top of this system new natural language recognition services with the challenge of deepening more in the understanding questions asked by a user and knowing what needs to be answered (for example, semantic search engine and assistant BOT). Moreover, to investigate new services in the line of extracting knowledge from the unstructured information that the Government of Aragon has, and to continue expanding and evolving the EI2A schema with the definition of new concepts and relationships based on the information processed as a consequence of the indicated actions.

References

[1] J. Hoxha and A. Brahaj, "Open Government Data on the Web: A Semantic Approach," *Emerging Intelligent Data and Web Technologies (EIDWT)*, 2011, pp. 107–113.

[2] A. Gugliotta, V. Tanasescu, J. Domingue, R. Davies, L. Gutiérrez-Villarías, M. Rowlattt, et al. "Benefits and challenges of applying semantic web services in the E-government domain". In Proceedings of *International Conference on Semantic Technologies (SEMANTICS 2006)*. Vienna, Austria.

[3] R. Cyganiak (2014). Retrieved 2016, from The linking open data cloud diagram: http://lod-cloud.net

[4] S. Bogdanović-Dinić, N. Veljković and L.Stoimenov (2014). "How open are public government data? An assessment of seven open data portals". In M. P. Rodríguer-Bolívar (Ed.), *Measuring E-government efficiency*, vol. V, pp. 25–44.

[5] P. Milic, L.V. Stoimenow and N.Veljkovic, "Semantic Technologies in e-government: Towards Openness and Transparency". M.P. Rodríguez Bolívar (Ed.), *Smart Technologies for Smart Governments.Public Administration and Information Technology*

[6] H. Alani, S. Kim, D. Millard, M. Weal, W. Hall, P. Lewis and N.Shadbolt. "Automatic Ontology-Based Knowledge Extraction from Webs Documents," *IEEE Intelligent Systems*, vol. 18, pp.14–21.

[7] S. Handschuh, S. Staab, and F. Ciravegna. "S-CREAM Semi-Automatic Creation of Metadata," *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web (EKAW 02)*. Lecture Notes in Artificial Intelligence, no. 2473, Springer-Verlag, Heidelberg, Germany, 2002, pp.358–372.

[8] M.Vargas-Vera et al. "Knowledge Extraction Using an Ontology-Based Annotation Tool," *Workshop on Knowledge Markup and Semantic Annotation*. ACM, pp.5–12

[9] P. Cimiano. "Ontology Learning and Population from Text: Algorithms, Evaluation and Applications". Springer-Verlag New or, Inc., Secaucus, NJ. USA, 2006.

[10] R. Witte, N. Khamis and J. Rilling, "Flexible Ontology Population from Text: The OwlExporter," *Int. Conf. on Language Resources and Evaluation*. 2010.

[11] J. Makki, "OntoPRiMa: A prototype for automating ontology population," *International Journal of Web/Semantic Technology (IJWesT)*. Vol. 8, October 2017.

[12] D. Maynard, Y. Li and W. Peters, "NLP Techniques for Term Extraction and Ontology Population." Vol.167. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pp. 107–127.

[13] J. Makki, A-M Alquier and V. Prince, "Ontology Population via NLP Techniques in Risk Management," *International Journal of Industrial and Manufacturing Engineering*. 2008.

[14] P. Peña, R. del Hoyo, J. Vea-Murguía, J. I. Calvo and J. M. Martín, "Moriarty: Improving 'time to market' in Big Data and Artificial Intelligence applications," *Proccedings of the International Conference on Big Data*. Alicante (Spain). International Journal of Design & Nature and Ecodynamics, 11(3), pp. 230-238, 2016.

[15] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, 22:457-479, 2004.

[16] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*". 2015.

[17] T. R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," *International Journal of Human and Computer Studies*, 43 (5-6), pp.907–928, 1992.

[18] M. Fernández-López, A. Gómez-Pérez and N. Juristo, "METHONTOLOGY: From Ontological Art Towards Ontological Engineering," Spring Symposium on *Ontological Engineering of AAAI*. Stanford University, California, pp 33-40, 1997.

[19] N. F. Noy and D. L. McGuinness, "Desarrollo de Ontologías-101: Guía Para Crear Tu Primera Ontología." Standford University. United States, September 2005.

[20] The Organization Ontology, https://www.w3.org/TR/vocab-org/, last accessed 2018/09/10

[21] Simple Knowledge Organization System, https://www.w3.org/2004/02/skos/, last accessed 2018/09/10

[22] Dublin Core Metadata Initiative, http://dublincore.org/documents/dcmi-terms/, last accessed 2018/05/18

[23] ISA Person Vocabulary, https://www.w3.org/ns/person,last accessed 2018/09/10

[24] ISA Location Vocabulary, https://www.w3.org/ns/locn, last accessed 2018/09/10

[25] The Event Ontology, http://motools.sourceforge.net/event/event.html, last accessed 2018/09/10