Preprint of:
M. Mountantonakis and Y. Tzitzikas, LODsyndesisIE: Entity Extraction from Text and Enrichment using Hundreds of Linked Datasets (Demo Paper), 17th Extended Semantic Web Conference (ESWC'2020)June 2020, Heraklion, Crete

LODsyndesis$_{IE}$: Entity Extraction from Text and Enrichment using Hundreds of Linked Datasets

Michalis Mountantonakis[1,2][0000−0002−1951−0241] and Yannis Tzitzikas[1,2][0000−0001−8847−2130]

[1] Institute of Computer Science - FORTH-ICS, Greece,
[2] Computer Science Department - University of Crete, Greece
{mountant, tzitzik}@ics.forth.gr

Abstract. We shall demonstrate LODsyndesis$_{IE}$, which is a research prototype that offers Entity Extraction from text and *Entity Enrichment* for the extracted entities, using several Linked Datasets. LODsyndesis$_{IE}$ exploits widely used Named Entity Extraction and Disambiguation tools (i.e., DBpedia Spotlight, WAT and Stanford CoreNLP) for identifying the entities of a given text, and enriches each identified entity with hyperlinks to LODsyndesis, which offers various services for millions of entities by leveraging hundreds of Linked Datasets. LODsyndesis$_{IE}$ brings several benefits to the entity extraction task: the user can a) annotate the entities of a given text by selecting different entity recognition tools, b) retrieve all the URIs and facts of each recognized entity from multiple datasets, and c) discover the K most relevant datasets (e.g., datasets containing the most facts) for each entity. The demo is available at https://demos.isl.ics.forth.gr/LODsyndesisIE/.

Keywords: Information Extraction, Linked Data, Multiple Datasets

1 Introduction

There is a large proliferation of approaches that perform named entity extraction (or recognition), linking and disambiguation [1, 13] from textual sources, which is an important task of any *Information Extraction* (IE) process. These approaches use pure NLP methods (e.g., Stanford CoreNLP [8]), methods based on a knowledge base (KB), e.g., DBpedia Spotlight [9], and others [1]. They usually associate each recognized entity with links (i.e., URIs) either to a single or to a few KBs (see more details in a recent survey [1]), i.e., for reasons of disambiguation and/or for extracting more information from the corresponding KB. For instance, *DBpedia Spotlight* [9] annotates each entity with a link to DBpedia [2] and WAT [12] provides links to Wikipedia. Since these approaches link each entity to a few number of knowledge bases, it is not trivial to find all the related URIs (and to collect all the triples) for each entity from multiple sources, e.g., for aiding users to select the URI that is more desirable for a given task or the URI that corresponds to the desired meaning of the word occurrence. This could be achieved by using approaches such as LODsyndesis [11]

and sameAs.cc [3], which provide all the available URIs for an entity. However, such systems are not connected with *Entity Extraction* tools, therefore the user has to use two or more systems: one *Entity Extraction* tool and one system that provides all the URIs of a given entity.

For facilitating this process, we present LODsyndesis$_{IE}$ (*IE* stands for Information Extraction), which provides fast access to all data related to a recognized entity by leveraging data coming from 400 RDF datasets. The approach is depicted in Figure 1. It takes as input a text of any length, like the text about the greek writer "Nikos Kazantzakis". As an output, it offers a) the initial text enriched with hyperlinks to LODsyndesis for each entity, by using three popular *Entity Recognition* tools (i.e., DBpedia Spotlight, WAT and Stanford CoreNLP), and b) an HTML table containing links to LODsyndesis, for extracting more information for each entity (e.g., related URIs and facts) from 400 RDF datasets. The output of LODsyndesis$_{IE}$ is offered in several formats (e.g., RDF, JSON, and HTML), either through its web interface, or by using its REST API. As we shall see, many tasks could be benefited from LODsyndesis$_{IE}$, including *Data Enrichment*, *Annotation*, *Data Integration*, *Data Discovery* and *Data Veracity*.

The rest of this demo paper is organized as follows: Section 2 describes related work, Section 3 introduces the steps of LODsyndesis$_{IE}$, and Section 4 reports use cases for demonstration. Finally, Section 5 concludes the demo paper.

2 Related Work

First, there are available several *Entity Extraction* systems over knowledge graphs, i.e., see a recent survey for more details [1], whereas a comparison of such approaches (through the benchmark *GERBIL*) is given in [13]. Moreover, there are tools such as *WDAqua* [4] and *LODQA* [5], which support *Entity Extraction* for offering Question Answering over Knowledge bases (more tools are surveyed in [6]). Comparing to these approaches, we neither focus on proposing a new *Entity Extraction* system (e.g., [8, 9, 12]) nor a new Question Answering system (e.g., [4, 5]). We focus on combining existing *Entity Extraction* tools and LODsyndesis [11] for facilitating the extraction of *more* information for the entities of a given text from hundreds of linked datasets.

3 The Steps of LODsyndesis$_{IE}$

LODsyndesis$_{IE}$ consists of two major steps, i.e., *Entity Recognition* (see Section 3.1) and *Entity Enrichment* (see Section 3.2).

3.1 Entity Recognition Step

The user can select to use *DBpedia Spotlight*, *Stanford CoreNLP*, *WAT*, or any combination of these tools (see Step 2 of Figure 1). Concerning *DBpedia Spotlight* and *WAT*, both tools produce a set of entity-URI pairs. In particular, for each

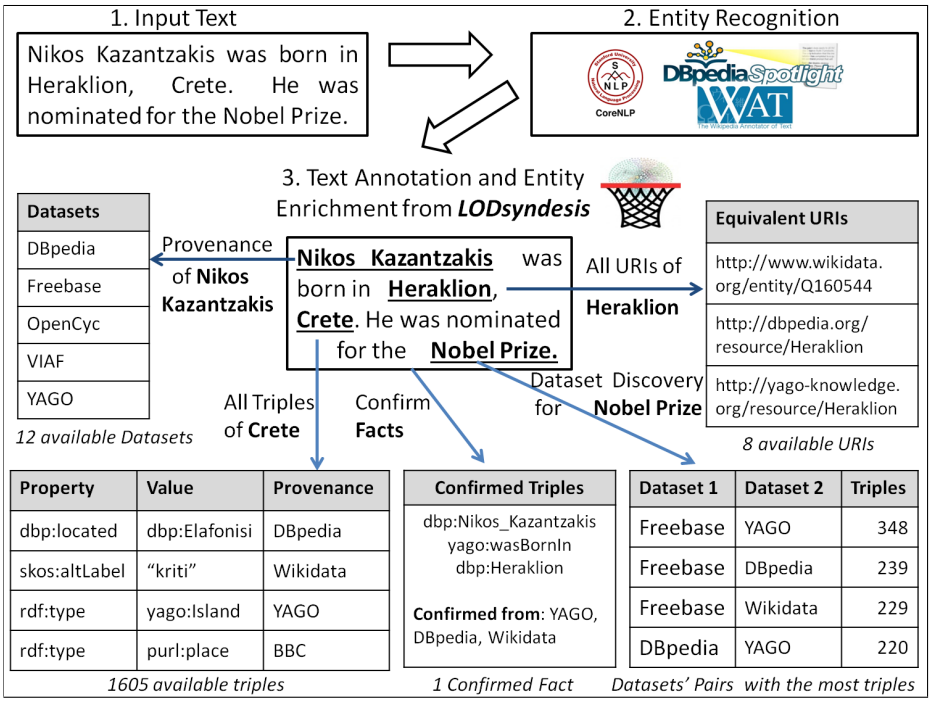

Fig. 1: The process of LODsyndesis$_{IE}$

recognized entity *DBpedia Spotlight* provides its corresponding DBpedia URI [2], whereas *WAT* offers its corresponding *Wikipedia* URI. However, for being able to compare the URIs derived from these tools, we replace each *Wikipedia* URI with its equivalent DBpedia URI. On the contrary, *Stanford CoreNLP* produces just a unique word for each entity (and not a URI to a knowledge base). For this reason, we take such words and we use LODsyndesis to find the most relevant DBpedia URI for each of these words (this approach is described in [5]), and then we create the desired set of entity-URI pairs. When the user has selected to use two or more tools, we take the union of all the recognized entity-URI pairs, produced by these tools. In case of conflicts and for reasons of disambiguation, i.e., if two or more tools identified different DBpedia URIs for the same entity e, we just keep the URI whose suffix, i.e., the last part of the URI (e.g., the suffix of "http://dbpedia.org/resource/Crete" is "Crete"), has the minimum Levenhstein distance with e. Therefore, in all cases the output of this step is a single entity-URI pair, for each recognized entity.

3.2 Entity Enrichment Step

For the set of recognized entity-URI pairs, we replace the URI of each entity with a hyperlink to LODsyndesis (i.e., we annotate the text with this hyperlink), for making it feasible to browse all the available triples for the given entity (i.e.,

see step 3 of Figure 1), whereas we also retrieve and show an image for each recognized entity. Moreover, the user can extract more information for each entity through `LODsyndesis`, which supports cross-dataset identity reasoning (i.e., computation of transitive and symmetric closure of `owl:sameAs` relationships), for offering semantics-aware indexes and services for 400 Linked Datasets, 412 million entities and 2 billion triples.

`LODsyndesis`$_{IE}$ exploits the aforementioned indexes and services and offers six options for the user, which can be accessed either through its web interface, or by using its REST API. In particular, one can browse or download (in JSON and RDF format), i) the URLs of the datasets containing each recognized entity e, ii) all the URIs that refer to e, and iii) all the triples (and their provenance) for e. Moreover, iv) `LODsyndesis`$_{IE}$ exploits the *Dataset Discovery* service of `LODsyndesis` for discovering for each entity e, "the K datasets maximizing the number of triples for e", and "the K datasets having the most common triples that contain e". The results of the aforementioned service can be exported in CSV format. For instance, we can provide answers for queries like "Give me the 4 datasets (i.e., quad of datasets) that maximize the number of triples for Nikos Kazantzakis". Since the number of possible combinations of datasets (e.g., quads) is given by the binomial coefficient formula, our services rely on incremental algorithms that are quite efficient for such a problem [11]. Moreover, the user is able to use these services for all the recognized entities together, e.g., one can download the triples of all the recognized entities in a single RDF file.

The user can also v) verify the correctness of the facts that are included in the text, i.e., `LODsyndesis`$_{IE}$ shows all the triples that connect the "key" entity of the text with any of the rest entities, e.g., ⟨"Nikos Kazantzakis","was born in","Heraklion"⟩. By default, the "key" entity is the entity which was recognized first in the text (e.g., "Nikos Kazantzakis" in Figure 1), however, the user can select any other entity, as the "key" one. In this way, it is feasible to find all the relationships between any pair of recognized entities. Finally, one can vi) export the annotated text in *HTML+RDFa* format, i.e., we store for each entity in the output file, its DBpedia URI, its type (e.g., "Person"), its corresponding URI to `LODsyndesis`, and all its related URIs, by using the *schema.org* vocabulary [7].

Example. Figure 1 shows an example of the output offered by `LODsyndesis`$_{IE}$, for each of the four recognized entities of the input text, i.e., "Nikos Kazantzakis", "Heraklion", "Crete" and "Nobel Prize". In this example, we selected to find all the datasets for "Nikos Kazantzakis" (12 datasets contain triples for this entity from the 400 available ones), all the URIs of "Heraklion" (8 URIs), all the triples for "Crete" (in total 1,605 triples), and the pairs of datasets offering the most triples for the entity "Nobel Prize". In particular, the union of $\{FreeBase, YAGO\}$ offers 348 triples for this entity. Moreover, we can see that the fact "Nikos Kazantzakis was born in Heraklion" is verified from 3 datasets.

Demo and REST API. The demo is accessible at `https://demos.isl.ics.forth.gr/LODsyndesisIE/`. For making it feasible to integrate `LODsyndesis`$_{IE}$ with external services, the demo website also offers a REST API and a REST client for JAVA. The backend of this website is implemented using Java technolo-

gies, whereas the front-end is based on common web technologies (Javascript). Finally, a demo video is available at `https://youtu.be/i52hY57dRms`.

4 Demonstration of Use Cases

We present four use cases, the first one corresponds to the *Entity Recognition* step, and the three remaining ones to the *Entity Enrichment* step.

Use Case 1. Comparison of Entity Recognition tools. By exploiting LODsyndesis$_{IE}$, it is feasible to compare the effectiveness of each tool performing *Entity Recognition* (or any combination of them) for different scenarios, i.e., for different texts.

Use Case 2. Data Integration and Enrichment. Suppose that a user wants to integrate data for one or more entities of the given text, for enriching their content, e.g., for creating either a Mediator or a Semantic Warehouse. In the case of Mediator (i.e., the data remain at their sources [10]), the user can collect and use any subset of the available URIs for each entity (e.g., through the RDFa file), or can find which datasets contain information about these entities (and probably their SPARQL endpoints). On the contrary, for constructing a Semantic Warehouse (i.e., the data should be pre-collected [10]), one can directly download and use all (or any subset of) the available triples for each entity.

Use Case 3. Dataset Discovery and Selection. The number of available datasets for a single or multiple entities can be large, e.g., for the entity "Greece" there are 40 available datasets in LODsyndesis. However, in many cases the user desires to keep only K (e.g., five) datasets, since the cost of integrating several datasets, can be huge as the number of datasets grows [10]. LODsyndesis$_{IE}$ can aid the user to discover and select the K most relevant datasets for one or more entities. In particular, one can discover the K datasets that maximize the available information for a set of entities, i.e., the union of these K datasets contains the maximum number of triples for the given entities, comparing to any other combination of K datasets.

Use Case 4. Data Veracity. The user has the opportunity to explore the relationships between any pair of recognized entities (which are included in the text), i.e., whether there is a property (or edge) that connects these two entities. In this way, the user can see which facts that occur in the given text, can also be confirmed from one or more datasets (which are indexed from LODsyndesis).

5 Conclusion

In this paper, we presented the research prototype LODsyndesis$_{IE}$, which exploits existing *Entity Recognition* tools (i.e., *DBpedia Spotlight*, *Stanford CoreNLP* and *WAT*) for recognizing the entities of a given text, and offers *Entity Enrichment* through LODsyndesis. We introduced the steps of LODsyndesis$_{IE}$, and we showed several use cases where LODsyndesis$_{IE}$ could be useful, including *Data Enrichment*, *Annotation*, *Data Integration*, *Data Discovery* and *Data Veracity*. As a future work, we plan to extend LODsyndesis$_{IE}$ for covering more tasks

of the *Information Extraction* process, e.g., to extract also the properties of a given text and link them to `LODsyndesis`. Finally, we plan to enrich the produced RDFa file by including more information (i.e., more triples for each entity).

Acknowledgements. The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under the HFRI PhD Fellowship grant (GA. No. 166).

References

1. T. Al-Moslmi, M. G. Ocaña, A. L. Opdahl, and C. Veres. Named entity extraction for knowledge graphs: A literature overview. *IEEE Access*, 8:32862–32881, 2020.
2. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.
3. W. Beek, J. Raad, J. Wielemaker, and F. van Harmelen. sameas. cc: The closure of 500m owl: sameas statements. In *ESWC*, pages 65–80. Springer, 2018.
4. D. Diefenbach, K. Singh, and P. Maret. WDAqua-core1: A Question Answering service for RDF Knowledge Bases. In *Companion Proceedings of the The Web Conference 2018*, pages 1087–1091, 2018.
5. E. Dimitrakis, K. Sgontzos, M. Mountantonakis, and Y. Tzitzikas. Enabling efficient question answering over hundreds of linked datasets. In *International Workshop on Information Search, Integration, and Personalization*, pages 3–17. Springer, 2019.
6. E. Dimitrakis, K. Sgontzos, and Y. Tzitzikas. A survey on question answering systems over linked data and documents. *Journal of Intelligent Information Systems*, pages 1–27, 2019.
7. R. V. Guha, D. Brickley, and S. Macbeth. Schema. org: evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51, 2016.
8. C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.
9. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight: shedding light on the web of documents. In *SEMANTiCS*, pages 1–8. ACM, 2011.
10. M. Mountantonakis and Y. Tzitzikas. Large Scale Semantic Integration of Linked Data: A survey. *ACM Computing Surveys (CSUR)*, 52(5):103, 2019.
11. M. Mountantonakis and Y. Tzitzikas. Content-based union and complement metrics for dataset search over rdf knowledge graphs. *Journal of Data and Information Quality (JDIQ)*, 12(2), 2020.
12. F. Piccinno and P. Ferragina. From tagme to wat: a new entity annotator. In *Proceedings of workshop on Entity recognition & disambiguation*, pages 55–62, 2014.
13. M. Röder, R. Usbeck, and A.-C. Ngonga Ngomo. Gerbil–benchmarking named entity recognition and linking consistently. *Semantic Web*, 9(5):605–625, 2018.