

# ARKIVO Dataset: A Benchmark for Ontology-based Extraction Tools

Laura Pandolfo <sup>a</sup> and Luca Pulina <sup>b</sup>

*Intelligent System Design and Applications (IDEA) Lab,  
University of Sassari, via Muroni 23A, 07100 Sassari, Italy*

**Keywords:** Semantic Web, Dataset, Benchmark, Ontology, Information Extraction.

**Abstract:** The amount of data available on the Web has grown significantly in the past years, increasing thus the need for efficient techniques able to retrieve information from data in order to discover valuable and relevant knowledge. In the last decade, the intersection of the Information Extraction and Semantic Web areas is providing new opportunities for improving ontology-based information extraction tools. However, one of the critical aspects in the development and evaluation of this type of system is the limited availability of existing annotated documents, especially in domains such as the historical one. In this paper we present the current state of affairs about our work in building a large and real-world RDF dataset with the purpose to support the development of Ontology-Based extraction tools. The presented dataset is the result of the efforts made within the ARKIVO project and it counts about 300 thousand triples, which are the outcome of the manually annotation process executed by domain experts. ARKIVO dataset is freely available and it can be used as a benchmark for the evaluation of systems that automatically annotate and extract entities from documents.


## 1 CONTEXT & MOTIVATION


The Web has grown exponentially in size over the last two decades and today it contains a huge amount of information resources, such as documents, images, audios and videos, which can be accessed anywhere and anytime. Most of this information consists of unstructured or semi-structured free-text documents which makes it overly challenging to search or analyze by users. Therefore, there has been a growing need for effective and efficient techniques for analyzing free-texts in order to aid users to retrieve structured information from unstructured documents and discover valuable and relevant knowledge (Piskorski and Yangarber, 2013). It is evident that manual annotation of documents cannot be an affordable solution, since it represents a time-consuming and expensive task. Moreover, in highly technical and specialized contexts costly domain expertise is required to decide on the correct annotation.

For many years, research in the fields of Information Extraction and Natural Language Processing has been focused on developing techniques able to automatically retrieve – with high precision – structured information from unstructured and/or semi-structured

documents. Despite the great progress in these fields, computers are still far from being able to have a complete semantic understanding of the human language (Adnan and Akbar, 2019). Methods to automatically extract or enhance the structure of various corpora have been a core topic also in the context of the Semantic Web, in which Information Extraction techniques are especially useful to populate the semantic knowledge-bases. On the other hand, using Semantic Web resources, such as ontologies, languages, data, tools, can be used to guide and improve the Information Extraction process (Martinez-Rodriguez et al., 2020). In particular, the use of ontology for formal and explicit specification domain concepts has been helpful in Information Extraction, making Ontology-Based Information Extraction a clear sub-discipline of knowledge extraction (Wimalasuriya and Dou, 2010). In this field, systems exploit ontologies to improve the performance of information extraction, by supporting and guiding algorithms for efficient and relevant IE (Konys, 2018). Also, using formal ontologies allows for applying standard inference engines for reasoning over extracted entities, thus enabling the derivation of further information that is not explicitly contained in texts (de Araujo et al., 2017).

One of the critical aspects in the development of this type of system is the evaluation phase, which re-

<sup>a</sup>  <https://orcid.org/0000-0002-5785-5638>

<sup>b</sup>  <https://orcid.org/0000-0003-0258-3222>

quires a ground truth, i.e., a dataset with all the relevant findings in the documents. Usually, the output of these tools is assessed by comparing it to the reference annotation, in order to compute standard quality metrics, such as recall and precision. However, it is well-known that large scale labeled corpus construction is a laborious and time consuming task (Che et al., 2019) and, for this reason, there is a limited availability of existing annotated documents, especially in domains such as the historical one.

In this paper we present the ARKIVO dataset and the related collections of archival historical documents from which the dataset originated. This dataset is the result of the efforts made within the ARKIVO project and actually it counts about 300 thousand triples, which are the results of the manually annotation process executed by domain experts. ARKIVO dataset is freely available and it can be used as a benchmark for the evaluation of systems that automatically annotate entities, such as places, persons and organizations, in unstructured documents. Since the ontology schema of ARKIVO dataset contains OWL constructs of OWL 2 DL profile (Grau et al., 2008), it also can be used for ontology benchmarking purposes, considering that there is a lack of expressive ontologies and language element combinations.

Our final goal is to achieve the annotation process of the archival historical documents semi-automatically; to do that, we are currently developing an ontology-based information extraction tool able to automatically annotate texts and populate the given knowledge base. The approach that we are going to use mainly rely on a combination of natural language process and information extraction techniques without an extensive involvement of domain experts for the validation of the extracted instances.

The paper is organized as follows. Section 2 provides a brief overview of the ARKIVO project, while in Section 3 we describe the ARKIVO dataset as a benchmark. The concluding remarks and future research are provided in Section 4.

## 2 THE ARKIVO PROJECT

The ARKIVO project stems from the collaboration between the Józef Piłsudski Institute of America and the University of Sassari with the aim of developing the semantic layer of the Piłsudski Institute digital archive (Pandolfo et al., 2019). In the following subsections, we report the main activities implemented to reach the stated goal.

### 2.1 Ontology Modeling and Description

One of the first steps in the development of the semantic layer for the Piłsudski Institute digital archive was the design of a new ontology, which provides a common language to represent not only the hierarchical structure of archival documents, but also some essential data embedded within the textual content of these documents. In fact, the developed ontology represents the typical archival structure levels, from the concept of *collection*, which can contain items or other collections as *fonds*, to the concept of single *item*, which typically is the smallest indivisible unit. Moreover, the ontology models some (historical) elements referenced in the archival documents and provides a reference schema for publishing them as Linked Data.

The ontology has been developed according to a top-down strategy, which consists first in identifying the most abstract concepts of the domain and then in specializing the specific concepts. The adopted methodology, which is closely related to the approach presented in (Blomqvist et al., 2016), allows to build simple, modular and reusable ontologies as well as flexible to future changes and expansions.

The ontology axiomatization is expressed using OWL 2 DL profile. This widely-known profile was chosen as modeling language since it allows to encode the knowledge as determined to be important by domain experts, e.g., it supports constructs such as universal quantification to a class expression, inverse object properties and disjunctions. Moreover, it also allows us to perform reasoning over ontology in order to ensure that ontology is consistent (Riboni and Bettini, 2011). Table 1 shows the number of classes, axioms and properties of the ontology. The full documentation is available at <https://github.com/arkivoTeam/arkivo>, while the ontology is available under a Creative Commons CC BY 4.0 license. The latest ontology version builds on and extends what reported in some previous contributions, i.e., (Pandolfo et al., 2017; Pandolfo et al., 2018; Pandolfo et al., 2019)

Table 1: Ontology metrics.

Classes	46
Axioms	280,282
Object properties	26
Data properties	34

### 2.2 Application and Linked Data

We applied the developed ontology to describe 12,848 collections and 28,644 items of archival holdings of

the Piłsudski Institute of America. The Institute is devoted to collecting, safe-keeping and preserving the documents and other historical memorabilia as well as to make these resources accessible to researchers and visitors by providing support to scholars during archival queries on site. The international character of the archival resources draws the attention of a large number of experts coming from different countries. To give an idea of the importance of the archival material, the collections occupy about 240 linear meters, namely 2 million pages of documents covering mostly the Polish, European and American history of the late 19th and 20th century. The collections include not only documents but also photographs, films, posters, periodicals, books, personal memoirs of diplomats, and political and military leaders, as well as collections of paintings by Polish and European masters. Most of the archival documents are written in Polish, but the number of documents in other languages – including Italian, English, Russian, French, Portuguese – is significant.

In the last five years, the archival collections have been annotated, digitized, full-text indexed, and gradually put online on the website of the Institute - archival collections are available at <http://archiwa.pilsudski.org/index.php>. The manual annotation process of the archival collections has been carried out in two steps. In the first step, archive workers have been manually annotating every document with relevant entities, such as title, author, date of creation, mentioned persons and/or event, etc. In the second step, the annotations have been methodically validated by domain experts and stored into the knowledge base. This process was certainly time and resource consuming and it was the main obstacle of this activity.

Taking advantage of the reference schema provided by the ontology for publishing Linked Data, it carried out a data integration process of combining data residing at different sources with the Piłsudski resources. In this way, the resources of Piłsudski Digital Archival Collections have been linked to external datasets of the linked data cloud in order to enrich the information provided with each resource. We selected, among others, different authority systems such as Wikidata, DBpedia, and VIAF (Virtual International Authority File), since they are the most common source of identifiers of people, organizations and historical events.

In Figure 1, we report an example of individuals and properties stored in the Piłsudski digital archive, and how these data have been linked to external resources, such as Wikidata (*wd* prefix) and DBpedia (*dbo* prefix). Looking at Figure 1, individual

*701.180/6216* of the class *Item* is related to its title and to its date of creation. This item, which is part of the file *A701.111.003*, is linked, via the object property *mentions*, to the person mentioned in it, i.e., *Roosevelt Franklin Delano*. Finally, the internal resource *Roosevelt Franklin Delano* is linked to other external instances and data in the linked data cloud.

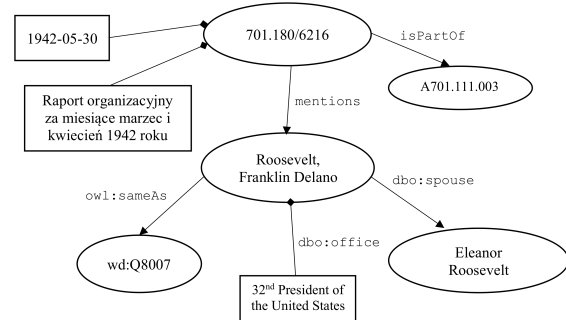


Figure 1: Example of entities and relationships. Classes are drawn as labelled ellipses, object properties between classes are shown as labelled edges, while boxes represent data properties.

### 3 ARKIVO DATASET AND DOCUMENT COLLECTIONS

In this Section, we present the ARKIVO dataset and the related collections of archival historical documents from which the dataset originated. The ARKIVO dataset is the result of the manually annotation process executed by domain experts. The dataset counts about 300 thousand triples and 181,780 of total instances – details of number of instances per class are reported in Table 2.

Table 2: Dataset metrics.

Items	28,644
Collections	12,848
Dates	6,615
Agents	2,093
Places	1,570

The dataset is freely available under a Creative Commons CC BY 4.0 license at <https://github.com/ArkivoTeam/ARKIVO> and it can be used as a benchmark for the evaluation of systems that automatically annotate entities, such as places, persons and organizations, in unstructured documents. In particular, ARKIVO dataset could be especially useful to carry out a named entity extraction and linking task, which refers to identifying mentions of entities in a text and linking them to a reference knowledge base provided as input

(Martinez-Rodriguez et al., 2020). During this task, the entities mentioned are extracted from the text and then they are linked to a specific knowledge base. This process is also known as entity disambiguation since it typically requires annotating a potentially ambiguous entity mentioned with a link to an identifier that describes a unique entity (Derczynski et al., 2015). For example, the ARKIVO dataset’s resource *G11499* is linked to its Polish name *Wielka Brytania* via the `schema:name` data property. In order to provide a disambiguation target, the resource *G11499* is linked via the `owl:sameAs` property to the unique identifier of Wikidata (*wd:Q23666*), which has its own name data property *Great Britain*. This example is graphically depicted in Figure 2.

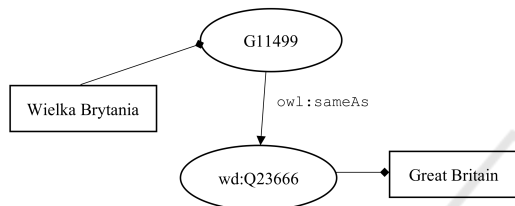


Figure 2: Example of annotated entities and relationships in the ARKIVO dataset in order to avoid potential ambiguous entity mentions.

The collections of archival historical documents from which ARKIVO dataset originated are available in PDF and published online at <http://archiwa.pilsudski.org/index.php#1>. All the documents have been previously scanned and processed by an Optical Character Recognition (OCR) tool.

In the following, we report a simple example to explain how the proposed dataset can be used as a benchmark for named entity extraction. Let suppose that we extracted entities using any Named Entity Recognition (NER) tool from a set an archival documents, including the one represented in Figure 3. In the depicted excerpt, the entities that our NER tool should be able to extract are marked in green (person entities) and in red (place entities) colours.

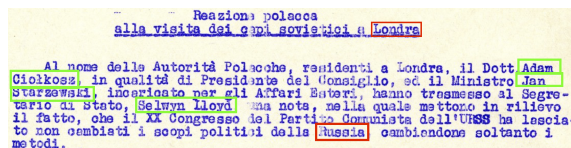


Figure 3: An excerpt from an archival historical document stored in the Pilsudski digital archive.

Using ARKIVO as benchmark, we can obtain the actual named entities in the document by querying the dataset using the SPARQL query depicted in Figure 4. In Table 3, we report the obtained names’ entities and

```
SELECT ?name ?class
WHERE {
  ?entity :isMentionedIn <http://pilsudski.org/
                        resources/701.180/4019>.
  ?entity schema:name ?name .
  ?entity a ?class.
}
```

Figure 4: Example of SPARQL query in ARKIVO.

the class to which they belong. Note that the SPARQL query results refer to the whole document and not to the only excerpt depicted above.

Table 3: SPARQL Query Results.

Entity Name	Entity Class
Ciołkosz, Adam	Person
Stalin, Józef	Person
Starzewski, Jan	Person
Chruszczow, Nikita	Person
Lloyd, Selwyn	Person
Bułganin, Nikołaj	Person
Polska	Place
Rosja	Place
Londyn	Place

Finally, considering the lack of expressive ontologies and language element combinations, ARKIVO can also be used for ontology benchmarking purposes, such as those presented in (Zamazal, 2020), since it provides good coverage of the OWL 2 language constructs.

## 4 CONCLUSION & FUTURE WORK

In this paper we presented the ARKIVO dataset and the related collections of archival historical documents from which the dataset originated. The ARKIVO dataset is the result of the manually annotation process executed by domain experts. The dataset is freely available and it can be used as a benchmark for the evaluation of Ontology-Based information extraction systems, also in unstructured documents. Moreover, ARKIVO can also be used for ontology benchmarking purposes.

The main obstacle of the whole ARKIVO project was represented by the manual annotation activity, which was a very time-consuming process. With this regard, our current research direction consists in the development of a semi-automatic ontology-based annotation process from texts by exploiting some of the techniques presented in (Pandolfo and Pulina, 2017; Pandolfo et al., 2016). The implemented approach will mainly rely on a combination of natural language

process and information extraction techniques without an extensive involvement of domain experts for the validation of the extracted instances.

## ACKNOWLEDGEMENTS

We would like to acknowledge the Józef Piłsudski Institute of America for providing us with the rich archival collections. Also, we would like to thank and commemorate Marek Zieliński, Vice-President of the Piłsudski Institute of America, for his invaluable contribution to both the intellectual and practical side at each stage of the work.

## REFERENCES

- Adnan, K. and Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. *Journal of Big Data*, 6(1):1–38.
- Blomqvist, E., Hammar, K., and Presutti, V. (2016). Engineering ontologies with patterns—the extreme design methodology. *Ontology Engineering with Ontology Design Patterns*, (25):23–50.
- Che, N., Chen, D., and Le, J. (2019). Entity recognition approach of clinical documents based on self-training framework. In *Recent Developments in Intelligent Computing, Communication and Devices*, pages 259–265. Springer.
- de Araujo, D. A., Rigo, S. J., and Barbosa, J. L. V. (2017). Ontology-based information extraction for juridical events with case studies in brazilian legal realm. *Artificial Intelligence and Law*, 25(4):379–396.
- Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., Petrak, J., and Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.
- Grau, B. C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., and Sattler, U. (2008). Owl 2: The next step for owl. *Journal of Web Semantics*, 6(4):309–322.
- Konys, A. (2018). Towards knowledge handling in ontology-based information extraction systems. *Procedia computer science*, 126:2208–2218.
- Martinez-Rodriguez, J. L., Hogan, A., and Lopez-Arevalo, I. (2020). Information extraction meets the semantic web: a survey. *Semantic Web*, (Preprint):1–81.
- Pandolfo, L. and Pulina, L. (2017). Adnoto: A self-adaptive system for automatic ontology-based annotation of unstructured documents. In Benferhat, S., Tabia, K., and Ali, M., editors, *Advances in Artificial Intelligence: From Theory to Practice - 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part I*, volume 10350 of *Lecture Notes in Computer Science*, pages 495–501. Springer.
- Pandolfo, L., Pulina, L., and Adorni, G. (2016). A framework for automatic population of ontology-based digital libraries. In Adorni, G., Cagnoni, S., Gori, M., and Maratea, M., editors, *AI\*IA 2016: Advances in Artificial Intelligence - XVth International Conference of the Italian Association for Artificial Intelligence, Genova, Italy, November 29 - December 1, 2016, Proceedings*, volume 10037 of *Lecture Notes in Computer Science*, pages 406–417. Springer.
- Pandolfo, L., Pulina, L., and Zielinski, M. (2017). Towards an ontology for describing archival resources. In Adamou, A., Daga, E., and Isaksen, L., editors, *Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017*, volume 2014 of *CEUR Workshop Proceedings*, pages 111–116. CEUR-WS.org.
- Pandolfo, L., Pulina, L., and Zielinski, M. (2018). Arkivo: an ontology for describing archival resources. In *CILC*, pages 112–116.
- Pandolfo, L., Pulina, L., and Zielinski, M. (2019). Exploring semantic archival collections: The case of piłsudski institute of america. In Manghi, P., Candela, L., and Silvello, G., editors, *Digital Libraries: Supporting Open Science - 15th Italian Research Conference on Digital Libraries, IRCDL 2019, Pisa, Italy, January 31 - February 1, 2019, Proceedings*, volume 988 of *Communications in Computer and Information Science*, pages 107–121. Springer.
- Piskorski, J. and Yangarber, R. (2013). Information extraction: Past, present and future. In *Multi-source, multilingual information extraction and summarization*, pages 23–49. Springer.
- Riboni, D. and Bettini, C. (2011). Owl 2 modeling and reasoning with complex human activities. *Pervasive and Mobile Computing*, 7(3):379–395.
- Wimalasuriya, D. C. and Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches.
- Zamazal, O. (2020). A survey of ontology benchmarks for semantic web ontology tools. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 16(1):47–68.