Unlocking Historical Insights: Developing a Dataset from Historical Archives

Laura Pandolfo*,†, Luca Pulina†

DUMAS, University of Sassari, via Roma 151, Sassari, Italy

Abstract

The proliferation of data on the Web has resulted in an increased need for effective techniques to extract relevant and valuable knowledge from this data. The intersection of the fields of Information Extraction and Semantic Web has created new opportunities to improve ontology-based information extraction tools. However, the development and evaluation of such systems have been hampered by the scarcity of annotated documents, particularly in historical domains. This article discusses the current state of our work in creating a large RDF dataset that aims to support the development of ontology-based extraction tools. The dataset was created through manual annotation by domain experts as part of the `arkivo` project and contains approximately 300,000 triples, which are freely available. This dataset can be used as a benchmark to evaluate systems that automatically extract entities and annotate documents.

Keywords

Semantic Web, Linked Open Data, Cultural Heritage, Ontology

1. Introduction

Historical texts are a crucial resource for scholars in the field of Digital Humanities [1]. They provide a unique perspective on the social, cultural, and political contexts of the past, providing valuable insights into the evolution of human thought and behavior. By examining them, researchers can gain a better understanding of how people interacted with each other during different historical periods [2, 3].

The digitization of historical texts has made them more accessible than ever before. Often these collections are preserved in digital libraries and digital archives that allows researchers to search and analyze across different time periods and geographic regions, by enabling them to identify patterns and trends that would have been difficult to discern using traditional methods. In fact, by leveraging digital tools and methods, it is possible to explore historical texts and gain new insights into the complex history of humanity. While the mere digitization of historical texts has now become an obvious and common practice, using computational techniques to automatically analyze them is a complex and challenging process. Historical texts, in fact, present unique challenges that must be addressed in order to automatically extract meaningful

CILC'23: 38th Italian Conference on Computational Logic, June 21–23, 2023, Udine, Italy

*Corresponding author.

†These authors contributed equally.

lpandolfo@uniss.it (L. Pandolfo); lpulina@uniss.it (L. Pulina)

0000-0002-5785-5638 (L. Pandolfo); 0000-0003-0258-3222 (L. Pulina)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3428/paper17.pdf

and accurate information. In the following, we report some of the main issues when working with this kind of documents:

1. **Quality**. It is not uncommon that historical texts that may be damaged, discolored, or written in a non-standard fonts. This poses challenge for OCR (Optical Character Recognition) technology which may fail to interpret some characters, resulting in errors that are not easy to rectify.
2. **Variability**. This kind of texts may use archaic language, obsolete spelling and syntax, which can make it difficult to automatically analyze them. In this regard, the lack of models trained on outdated languages is felt.
3. **Inconsistency**. Historical texts often contain inconsistencies and errors that can hinder the process of extracting accurate key information.
4. **Lack of standardization**. They often use different protocols for formatting, citation, and referencing, making it difficult to compare and analyze texts from different sources.
5. **High expertise**. Historical texts regularly contain cultural and historical references that may not be easily understood by modern-day readers, thus requiring high expertise for a proper understanding of the text. For example, historical texts may reference specific events, customs, or practices that are no longer relevant or well-known in modern times.

To cope with all these issues, specialized domain experts should be involved in the manual annotation process in order to extract relevant data and make it accessible in a structured way. However, it is clear that manual annotation cannot be an affordable solution, since it represents a time-consuming and expensive task.

For many years, research in the fields of Text Mining [4], Information Extraction [5] and Natural Language Processing [6] have been working on developing techniques able to automatically extract structured information from historical documents with high precision. However, despite significant progresses in these fields, computers are still far from being able to have a complete semantic understanding of the human language [7]. Methods to automatically extract information have been a core topic also in the context of the Semantic Web field, in which information extraction techniques are especially useful to populate the semantic knowledge-bases. On the other hand, Semantic Web resources, such as ontologies, languages, data, tools, have been used to guide and improve the information extraction process [8]. In particular, the application of ontology has proven to be beneficial in the field of information extraction, as it provides a formal and explicit definition of domain concepts. This has led to the emergence of Ontology-Based Information Extraction as a sub-discipline of knowledge extraction [9]. Ontology-based techniques are used to enhance the performance of systems by guiding algorithms for efficient and relevant information extraction [10]. Additionally, the use of formal ontologies enables standard inference engines to reason over extracted entities, thus allowing to infer additional information that may not be explicitly stated in the original text [11].

Promising developments in the field of Machine Learning [12] for historical texts have been made. Recent studies demonstrated how Neural Networks (NNs) could be effective in processing historical texts since they effectively support many NLP tasks that are relevant for Digital Humanities research, such as Named Entity Recognition (NER), Entity Linking (EL), Relation Extraction (RE) and other tasks – see, e.g. [13, 14, 15]. Nonetheless, the use of NNs seems to be rather limited for different reasons. Firstly, historical texts often lack consistent annotations for

named entities such as people, places, and organizations. Currently, the amount of annotated corpus available to train models is very scarce, making it difficult to train models that are able to accurately recognize and extract entities from historical texts. Developing these labeled corpora is not a straightforward task as it requires a great amount of resources in terms of time, budget, and expertise. Secondly, the lack of off-the-shelf tools represents a potential hindrance for scholars in the digital humanities who could provide a valuable contribution in this field.

In this paper, the authors present the `arkivo` dataset, which was built from archival historical documents previously manually annotated by domain experts. Currently, the dataset contains around 300,000 triples and is freely available for use. In addition to its intrinsic value as historical artifacts, this dataset could also be of great importance for the development of information extraction tools and methods as a benchmark to evaluate systems that automatically annotate entities in unstructured documents, such as places, persons, and organizations. In fact, one of the critical aspects in the development of this type of system is the evaluation phase, which requires 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. This dataset also could serve as a rich source of training data for machine learning algorithms in order to allow researchers to create more accurate and efficient natural language processing systems. In this respect, historical data provides a valuable testbed for evaluating the effectiveness of these models, as they present unique challenges that must be overcome in order to extract meaningful information. Additionally, the ontology schema of the `arkivo` dataset is based on the OWL 2 DL profile, which makes it suitable for ontology benchmarking purposes, as there is a shortage of expressive ontologies and language element combinations available. The dataset was created as part of the largest `arkivo` project.

The rest of the paper is organized as follows. Section 2 describes the research baseline for this work, including background on ontology and Linked Data, and it discusses some related works. Section 3 presents the source datasets and the ontology model, while Section 4 describe the `arkivo` dataset and its usefulness. Section 5 concludes and presents planned future work.

2. Background

In the field of philosophy, ontology is often understood as the study of the nature of being, including the nature of entities or substances, their properties, and their relations. In recent years, the term ontology has also been used in other fields, including computer science, where it refers to the study of conceptual models of a particular domain, and in information science, where it refers to the study of the nature of information and knowledge. In the field of computer science, an ontology is commonly defined as a formal and explicit specification of the concepts, entities, and relationships within a particular domain [16]. The goal of the ontology is to create a shared understanding of a particular domain among people and machines, enabling effective communication and knowledge sharing. There are different formal languages and frameworks used to develop ontologies in computer science, such as the Resource Description Framework Schema (RDFS) [17] and the Web Ontology Language OWL [18]. These languages provide a way of representing knowledge in a machine-readable format, thus enabling automated reasoning

and inference.

The most recent version of OWL is OWL 2 [19], which tackles the issue of complexity by establishing profiles, namely fragments. In particular, OWL 2 has the following profiles: OWL 2 EL, OWL 2 QL, OWL 2 RL, OWL 2 DL, and OWL 2 Full. Each profile varies in terms of their expressivity and reasoning complexity. The first three profiles (OWL 2 EL, OWL 2 QL, OWL 2 RL) are tractable fragments of OWL 2 having polynomial reasoning time. Reasoning over OWL 2 DL ontologies has a complexity of N2EXPTIME, while OWL 2 Full is undecidable. OWL 2 DL, the version of OWL we focus on, is based on Description Logics (DL) [20], a group of formal languages for knowledge representation that model concepts, roles, individuals, and their relationships. In DL, a database is called a knowledge base. In particular, if $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ is a knowledge base, then the Tbox $\mathcal{T}$ is a set of *inclusion assertions*, i.e. concept descriptions in $\mathcal{AL}$ or some of its extensions, whereas the Abox is a set of *membership assertions* of the form $A(x)$ and $R(x, y)$ where A is some atomic concept, R is some atomic role and x, y are objects of a domain. Some OWL 2 constructors with the corresponding DL syntax are listed in Table 1.

Table 1
The constructors of OWL 2 and the corresponding DL syntax.

Constructor	DL syntax
class	C
subClassOf (*C D*)	$C \sqsubseteq D$
equivalentClasses (C_1...C_n)	$C_1 \equiv ... \equiv C_n$
disjointClasses (C_1...C_n)	$C_i \sqcap C_j \sqsubset \bot$ $_{i \neq j\ i,j \in \{1,...,n\}}$
disjointUnion (C C_1...C_n)	$C = C_1 \sqcup ... \sqcup C_n$
objectProperty	R
datatypeProperty	T
objectInverseOf (R)	R^-
objectSomeValuesFrom(R C)	$\exists$R.C

The standard query language is SPARQL [21], whose latest version, namely SPARQL 1.1, includes many new language features such as aggregates, sub-queries, a new suite of built-in functions, and path expressions. SPARQL queries typically consist of various clauses and blocks, which specify basic graph patterns to be matched along with keywords that join, filter and extend the solution sequences to these patterns.

For a more detailed description of SPARQL syntax and its operators, please refer to [21].

2.1. Linked Data

The term Linked Open Data (LOD) refers to tools or platforms that allow for the collection and integration of data from various sources or formats, which can be accessed by both machines and humans [22]. These tools typically enable users to search for information using predefined languages or mechanisms (e.g. SQL, HTML, SPARQL, DL, RDFS, etc.). Linked Open Data is technically defined as a knowledge graph that is represented as a semantic web or schema using ontologies to interconnect data. The goal of LOD is to establish a worldwide data environment that can be accessed, shared, and resused by anyone, from any location, and for any objective.

The state-of-the-art of LOD is constantly evolving, driven by advances in technologies, standards, and best practices [23, 24]. In particular, it is characterized by a growing emphasis on interoperability, quality, and reuse, as well as a focus on developing new applications and use cases that can leverage the vast amount of interconnected data available on the Web. In the last years, the number of datasets in the LOD cloud [1] has been growing steadily, from a few hundred in 2007 to over 1,300 as of 2021.

As LOD datasets become more widely used, there is a growing emphasis on ensuring their quality and provenance. This includes efforts to develop standards and best practices for specific application domains concerning data cleaning, enrichment, as well as mechanisms for tracking the source and history of data [25]. In the cultural heritage domain, for examples, several studies have been conducted in order to analyze the state of application of linked data and to develop specific actions and recommendations to be implemented to improve their effective usage [26, 27].

In the last year, the use of knowledge graphs has been considered relevant for representing and linking data in a structured and semantic way [28]. In fact, by using knowledge graphs, it is possible to connect different data sources and extract meaningful insights from large and complex datasets. Currently, many LOD applications are using knowledge graphs to provide more sophisticated search, recommendation, and analysis capabilities.

Also the integration with AI and machine learning has been investigated in the LOD field [29, 30]. For example, linked data is increasingly being used as a source of training and validation data for AI and machine learning applications.

2.2. Related Work

Cultural heritage data is an important resource for research, study and analysis in different fields, including history, social sciences, and humanities in general. In the last decade, several historical and cultural heritage datasets have been published in the LOD cloud, covering a wide range of topics and formats. Very often these datasets have been built as part of the development of semantic digital libraries, as for example Europeana [2], which is a platform that provides access to millions of digitized cultural heritage items from museums, archives, and libraries across Europe. Its datasets can be downloaded for free and it also offers a range of APIs. Data is represented in the Europeana Data Model (EDM), which re-use some of the reference ontologies already available (e.g. CIDOC-CRM [3], SKOS [4], FOAF [5], Dublin Core [6]). It enables interoperability without affecting the source data models. Provides queries to multiple linked metadata from European institutions

The Smithsonian Institution [7] offers more than 3 million digital images and assets through their open access platform, allowing users to explore and download high-resolution images of

[1] https://lod-cloud.net/
[2] https://pro.europeana.eu/data
[3] https://www.cidoc-crm.org/
[4] https://www.w3.org/TR/2009/REC-skos-reference-20090818/
[5] http://xmlns.com/foaf/0.1/
[6] https://www.dublincore.org/
[7] https://www.si.edu/openaccess

their collection items. More than 230,000 museum object records from across the 14 institutions have been converted to LOD and are available to be explored through one interface.

Open Heritage 3D [8] is a project that aims to digitize and make available 3D models of cultural heritage sites and artifacts. Their datasets can be downloaded for free and used for research, education, and preservation purposes.

The British Library [9] published some of its collections as LOD by using existing RDF vocabularies and ontologies, including BIBO, BIO, Dublin Core, FOAF, GeoNames, Schema.org and SKOS. Over the years, the British Library extended its services and it actually provides access to over 150 digital collections, some of which can be queried from the SPARQL endpoint, such as the British National Bibliography.

WarSampo [10] is a shared semantic infrastructure and a LOD service for publishing data about WW2, with a focus on Finnish military history, which contains over 14 million triples. It uses some existing vocabularies such as CIDOC-CRM, Dublin Core and SKOS.

Enslaved [11] aims to create a LOD portal that allow users to easily query and inspect integrated historical data related to the slave trade. To effectively combine the vast array of diverse data sources commonly found in historical research communities, an ontology schema has been developed using the OWL 2 DL profile.

3. Materials & Methods

3.1. Source Dataset

The source dataset of `arkivo` derives from the Józef Piłsudski Institute of America (JPIA), which is a non-profit organization dedicated to preserving and promoting the legacy of Józef Piłsudski, a prominent Polish statesman and military leader who played a key role in the history of Poland during the early 20th century. The Institute was founded in 1943 by a group of Polish-American intellectuals and scholars who sought to honor Piłsudski's contributions to Polish independence and sovereignty.

Located in New York City, the JPIA is home to an extensive collection of archives, manuscripts, photographs, and artifacts related to Piłsudski's life and career, as well as the history of Poland and the Polish-American community. Its collections and resources are open to the public, and it welcomes visitors and researchers from all over the world. 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. Table 2 reports all the heterogeneous source of data of `arkivo` dataset. The source data are in different formats, such as PDF documents, printed texts, letters, photographs, video, digital images, and spreadsheets.

In the last five years, these JPIA's collections of historical materials 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 the following stages.

[8] https://artsandculture.google.com/project/openheritage
[9] https://www.bl.uk/
[10] https://www.ldf.fi/dataset/warsa
[11] https://enslaved.org/

Table 2
Data source, period and topic of reference, and format.

N	Data source	Period	Format
	Valerian Platonov Archive	Zabory 1795-1918	PDF documents, photos
1	Józef Piłsudski Archive	Second Republic 1918-1939	printed texts, PDF documents, digital images, spreadsheet, photos, video
2	Legation of the Republic of Poland in Rio de Janeiro	Second Republic 1918-1939	printed texts, PDF documents
3	National Defense Committee	Second Republic 1918-1939	printed texts, PDF documents
4	Adjutant General of the Commander-in-Chief	Second Republic 1918-1939	printed texts, PDF documents
5	Polish Legions, Polish Military Organization, Supreme National Committee, Rifle Association	Second Republic 1918-1939	PDF documents
6	Files of the Chief of the General Staff of the Polish Army, General Tadeusz Rozwadowski	Second Republic 1918-1939	PDF documents
7	Liquidation Commission of General Lucjan Żeligowski	Second Republic 1918-1939	PDF documents
8	Judicial and Honorary Cases of Generals and Higher Commanders	Second Republic 1918-1939	PDF documents
9	Ukrainian Military Mission in Poland	Second Republic 1918-1939	PDF documents, letters
10	Collection on the President of the Republic of Poland Ignacy Mościcki	Second Republic 1918-1939	PDF documents
11	ESTEZET Independent Intelligence Facility	Events of the Second World War	PDF documents, printed texts, spreadsheet
12	Embassy of the Republic of Poland in the Vatican	Events of the Second World War	PDF documents, letters
13	Polish government in exile	Events of the Second World War	PDF documents
14	National Committee of Americans of Polish Descent	Poland and the world after 1945	PDF documents

1. Archive workers from JPIA annotated documents manually with relevant entities – such as title, author, date of creation, mentioned persons and/or event – and reported the annotations into Excel spreadsheets.
2. All annotations have been rigorously inspected by domain experts in order to assess the accuracy. The annotation of historical texts is a specialized activity that requires not only knowledge related to the cultural context, but also linguistic knowledge related to the evolution of language. One of the additional difficulties in this work is due to the fact that the considered dataset is multilingual.
3. Using OpenRefine [31], Excel tables have been transformed in order to clean and reshape the data as our reference data model needed. Then data has been converted in CSV format.
4. Using Tarql [12], which requires a practical knowledge of SPARQL, we mapped CSV data into RDF format.
5. RDF data has been stored on Stardog [32], a commercial RDF triple store with fast SPARQL query, transactions, and OWL reasoning support.

3.2. Ontology Data Model

The source datasets reported in Table 2 were harmonized and transformed in RDF format in order to populate the `arkivo` ontology. This ontology was developed in order to provide a common reference schema able 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, it captures the standard levels of archival structure, ranging from the highest level of

[12] "SPARQL for Tables: Turn CSV into RDF using SPARQL syntax", https://tarql.github.io/examples/, accessed: 2023-03-03.

a *collection*, which can consist of other collections or individual items, down to the smallest indivisible unit, or single *item*. Additionally, the ontology includes modeling of certain historical elements referenced in archival documents, and serves as a reference schema for publishing them as LOD. `arkivo` ontology was developed using a top-down approach, which involves identifying the most general concepts in the domain before proceeding to the more specific ones. This methodology is closely aligned with the approach outlined in [33] and enables the development of simple, modular, and reusable ontologies that can easily adapt to future changes and expansions.

The ontology was developed using the OWL 2 DL profile, which provides support for constructs such as universal quantification, inverse object properties, and disjunctions. This language was chosen because it allows domain experts to encode the knowledge deemed important for the ontology. Additionally, OWL 2 DL allows for reasoning over the ontology to ensure consistency [34]. The current version of the ontology is composed of 46 classes, 26 object properties, 34 data properties, and 280,282 axioms. In the following, we pinpoint some of the main classes and properties of arkivo ontology.

During the ontology development, some parts of existing ontologies were reused for the scope of this study, since it is widely acknowledged that promoting the integration and reuse of existing standard metadata and vocabularies is one of the best practices in the Semantic Web field. This approach can accelerate the ontology design process and ensure extensibility and interoperability with other resources and applications [35]. Table 3 lists the vocabularies employed as core ontologies of `arkivo`.

Table 3
Existing ontologies reused.

Prefix	Ontology	Description
`bibo`	The Bibliographical Ontology	It provides main concepts and properties for describing citations and bibliographic references (http://bibliontology.com).
`dc` or `dct`	Dublin Core	An upper-level ontology that describes a broad class of information objects. (http://dublincore.org).
`foaf`	Friend Of A Friend	It describes people, their activities and their relations to other people and objects (http://www.foaf-project.org).
`geo`	GeoNames Ontology	This ontology provides geospatial semantic information in OWL (http://www.geonames.org/ontology/documentation.html).
`lode`	Ontology for Linking Open Descriptions of Events	It is an ontology for publishing descriptions of historical events as linked data (http://linkedevents.org/ontology/).
`schema`	Schema.org	It is a large vocabulary counting hundreds of terms coming from multiple domains (http://schema.org.).

The main classes of `arkivo` are `Collection`, which represents the set of documents or collections, and `Item`, which is the smallest indivisible unit of an archive. In order to describe the structure of the archive, different subclasses of the class `Collection` are modeled, namely the subclasses *Fonds*, *File* and *Series*. Using existential quantification property restriction (`owl:someValuesFrom`), we defined that the class `Item` as the class of individuals that are linked to individuals in the class `Fonds` by the `isPartOf` property, as shown below using the DL syntax.

$$Item \sqsubseteq \exists\, isPartOf.Fonds$$

This means that there is an expectation that every instance of `Item` is part of a collection, and that collection is a member of the class `Fonds`. This is useful to capture incomplete knowledge.

For example, if we know that the individual `701.180/11884` is an item, we can infer that it is part at least of one collection.

Moreover, we defined some union of classes for those classes that perform a specific function on the ontology. In this case, we used `owl:unionOf` constructor to combine atomic classes to complex classes, as we describe in the following:

$$CreativeThing \equiv Collection \sqcup HistoricalEvent \sqcup Item$$

This class denotes things created by agents and it includes individuals that are contained in at least one of the classes `Collection`, `HistoricalEvent` or `Item`.

$$NamedThing \equiv Place \sqcup Date \sqcup Agent$$

It refers to things, such as date, place and agent, that are related to individuals in the `CreativeThing` by the object property *isMentionedIn*, and it includes individuals that belong to at least one of the classes `Place`, `Date` or `Agent`.

The full ontology documentation is available at https://github.com/arkivoTeam/arkivo, and the `.owl` file is available under a Creative Commons CC BY 4.0 license. The latest version builds upon and extends some previous contributions, i.e. [36, 37, 38, 39].

4. Dataset

`arkivo` ontology were used to describe 12,848 collections and 28,644 items of archival holdings of the JPIA. Taking advantage of the reference schema provided by the ontology for publishing LOD, an integration process of data coming from different sources was carried out. This allows us to link the resources of Piłsudski Archival Collections to external datasets of the LOD cloud in order to enrich the information provided with each resource. We referred the most common datasets for identifying people, organizations and historical events, such as Wikidata, DBpedia, and VIAF (Virtual International Authority File).

Figure 1 reports an example of individuals and properties of the `arkivo` dataset and highliths how these data have been linked to external resources, such as Wikidata (`wd` prefix) and DBpedia (`dbo` prefix). In this particular example, 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*, which is in turn linked to other external instances and data in the LOD cloud.

The dataset is freely available under a Creative Commons CC BY 4.0 license at https://github.com/ArkivoTeam/ARKIVO [40]. A typical use case is the discovery of historical data for a more comprehensive and interconnected understanding of historical events, movements, political and cultural developments. Also, 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. `arkivo` dataset could be especially useful to conduct named entity extraction and linking task, which refers to task devoted to identify mentions of entities in a text and linking them to a reference knowledge base provided as input [8]. This process is also known as entity disambiguation since it typically requires annotating a potentially ambiguous

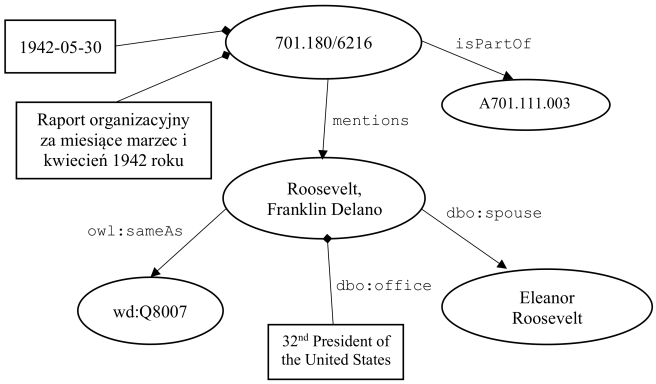


Figure 1: `arkivo` instances and linkage between them. Instances are drawn as labelled ellipses, object properties as labelled edges, while boxes represent data properties.

entity mentioned with a link to an identifier that describes a unique entity [41]. 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:Q295688*), which has its own name data property *Great Britain*. This example is graphically depicted in Figure 2.

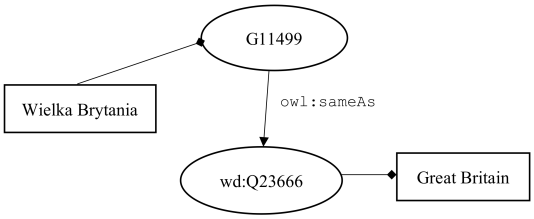


Figure 2: Example of annotated entities and relationships in the `arkivo` dataset in order to avoid potential ambiguous entity mentions.

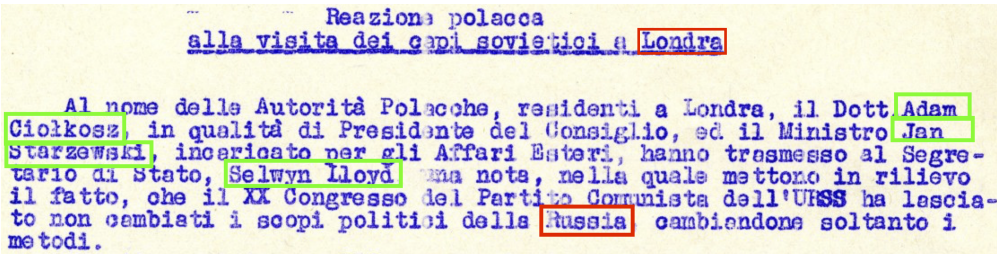
The collections of archival historical texts from which `arkivo` dataset originated are available in PDF and published online at http://archiwa.pilsudski.org/index.php#1.

In the following, we report a straightforward 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 of documents, including the one represented in Figure 3. In the depicted excerpt, the entities that our NER tool should be able to detect and extract are marked in green (person entities) and in red (place entities) colours.

Using our dataset as benchmark, we can obtain the actual named entities in the document by querying it using simple SPARQL queries, such as the one reported in Figure 4.

In Table 4, the obtained names' entities and the class to which they belong are reported. Note that the SPARQL query results refer to the whole document and not to the only excerpt depicted above.

Furthermore, considering the lack of expressive ontologies and language element combina-


Reazione polacca
alla visita dei capi sovietici a Londra

Al nome delle Autorità Polacche, residenti a Londra, il Dott Adam Ciołkosz, in qualità di Presidente del Consiglio, ed il Ministro Jan Starzewski, incaricato per gli Affari Esteri, hanno trasmesso al Segretario di Stato, Selwyn Lloyd una nota, nella quale mettono in rilievo il fatto, che il XX Congresso del Partito Comunista dell'URSS ha lasciato non cambiati i scopi politici della Russia, cambiandone soltanto i metodi.

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

```
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` dataset.

Table 4
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

tions, `arkivo` can also be used for ontology benchmarking purposes, such as those presented in [42], since it provides good coverage of the OWL 2 language constructs.

5. Conclusion

This paper presented the approach we applied for building a dataset focused on historical texts coming from the JPIA. The aim is to create a dataset that can be used not only for its intrinsic value as historical artifact, but also a benchmark to evaluate information extraction tools and methods devoted to automatically annotate entities in unstructured documents, such as places, persons, and organizations. Moreover, the presented dataset can also be used for ontology benchmarking purposes.

The main obstacle of the whole work 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 [43, 44]. 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.

References

[1] P. Svensson, Humanities Computing as Digital Humanities, in: Defining Digital Humanities, Routledge, 2016, pp. 175–202.

[2] G. Adorni, M. Maratea, L. Pandolfo, L. Pulina, An Ontology-Based Archive for Historical Research, in: Description Logics, 2015.

[3] G. Adorni, M. Maratea, L. Pandolfo, L. Pulina, An Ontology for Historical Research Documents, in: Web Reasoning and Rule Systems: 9th International Conference, RR 2015, Berlin, Germany, August 4-5, 2015, Proceedings. 9, Springer, 2015, pp. 11–18.

[4] T. Jo, Text Mining, Studies in Big Data (2019).

[5] R. Grishman, Information Extraction, IEEE Intelligent Systems 30 (2015) 8–15.

[6] K. Chowdhary, K. Chowdhary, Natural Language Processing, Fundamentals of Artificial Intelligence (2020) 603–649.

[7] K. Adnan, R. Akbar, An analytical study of information extraction from unstructured and multidimensional big data, Journal of Big Data 6 (2019) 1–38.

[8] J. L. Martinez-Rodriguez, A. Hogan, I. Lopez-Arevalo, Information extraction meets the semantic web: a survey, Semantic Web (2020) 1–81.

[9] D. C. Wimalasuriya, D. Dou, Ontology-based information extraction: An introduction and a survey of current approaches, 2010.

[10] A. Konys, Towards knowledge handling in ontology-based information extraction systems, Procedia computer science 126 (2018) 2208–2218.

[11] D. A. de Araujo, S. J. Rigo, J. L. V. Barbosa, Ontology-based information extraction for juridical events with case studies in brazilian legal realm, Artificial Intelligence and Law 25 (2017) 379–396.

[12] I. El Naqa, M. J. Murphy, What is Machine Learning?, Springer, 2015.

[13] O. Suissa, A. Elmalech, M. Zhitomirsky-Geffet, Text Analysis Using Deep Neural Networks in Digital Humanities and Information Science, Journal of the Association for Information Science and Technology 73 (2022) 268–287.

[14] O. Dereza, Lemmatization for Ancient Languages: Rules or Neural Networks?, in: Artificial Intelligence and Natural Language: 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings 7, Springer, 2018, pp. 35–47.

[15] C. Yan, Q. Su, J. Wang, Mogcn: Mixture of Gated Convolutional Neural Network for Named Entity Recognition of Chinese Historical Texts, IEEE Access 8 (2020) 181629–181639.

[16] N. Guarino, D. Oberle, S. Staab, What is an Ontology?, Handbook on Ontologies (2009) 1–17.

[17] B. McBride, The Resource Description Framework (rdf) and its Vocabulary Description Language rdfs, Handbook on Ontologies (2004) 51–65.

[18] G. Antoniou, F. v. Harmelen, Web Ontology Language: OWL, Handbook on Ontologies (2009) 91–110.
[19] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, S. Rudolph, et al., Owl 2 Web Ontology Language Primer, W3C Recommendation 27 (2009) 123.
[20] F. Baader, I. Horrocks, U. Sattler, Description Logics, Foundations of Artificial Intelligence 3 (2008) 135–179.
[21] S. Harris, A. Seaborne, E. Prud'hommeaux, Sparql 1.1 Query Language, W3C Recommendation 21 (2013) 778.
[22] C. Bizer, T. Heath, K. Idehen, T. Berners-Lee, Linked Data on the Web (ldow2008), in: Proceedings of the 17th international conference on World Wide Web, 2008, pp. 1265–1266.
[23] M. T. Kone, State of the art in semantic organizational knowledge, Encyclopedia of Organizational Knowledge, Administration, and Technology (2021) 1762–1773.
[24] M. Mountantonakis, Y. Tzitzikas, Large-Scale Semantic Integration of Linked Data: A Survey, ACM Computing Surveys (CSUR) 52 (2019) 1–40.
[25] D. Feitosa, D. Dermeval, T. Ávila, I. I. Bittencourt, B. F. Lóscio, S. Isotani, A Systematic Review on the Use of Best Practices for Publishing Linked Data, Online Information Review 42 (2018) 107–123.
[26] L. Zhang, Empowering Linked Data in Cultural Heritage Institutions: A Knowledge Management Perspective, Data and Information Management 6 (2022) 100013.
[27] E. Davis, B. Heravi, Linked Data and Cultural Heritage: a Systematic Review of Participation, Collaboration, and Motivation, Journal on Computing and Cultural Heritage (JOCCH) 14 (2021) 1–18.
[28] J. Z. Pan, G. Vetere, J. M. Gomez-Perez, H. Wu, Exploiting Linked Data and Knowledge Graphs in Large Organisations, Springer, 2017.
[29] P. Ristoski, G. K. D. De Vries, H. Paulheim, A Collection of Benchmark Datasets for Systematic Evaluations of Machine Learning on the Semantic Web, in: The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II 15, Springer, 2016, pp. 186–194.
[30] M. Mountantonakis, Y. Tzitzikas, How Linked Data can Aid Machine Learning-Based Tasks, in: Research and Advanced Technology for Digital Libraries: 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings, Springer, 2017, pp. 155–168.
[31] R. Verborgh, M. De Wilde, Using OpenRefine, Packt Publishing Ltd, 2013.
[32] S. Union, Stardog, 2018.
[33] E. Blomqvist, K. Hammar, V. Presutti, Engineering Ontologies with Patterns-The eXtreme Design Methodology, Ontology Engineering with Ontology Design Patterns (2016) 23–50.
[34] D. Riboni, C. Bettini, Owl 2 Modeling and Reasoning with Complex Human Activities, Pervasive and Mobile Computing 7 (2011) 379–395.
[35] M. Katsumi, M. Grüninger, Choosing Ontologies for Reuse, Applied Ontology 12 (2017) 195–221.
[36] L. Pandolfo, L. Pulina, M. Zielinski, Towards an Ontology for Describing Archival Resources, in: WHiSe@ ISWC, 2017, pp. 111–116.
[37] L. Pandolfo, L. Pulina, M. Zielinski, Arkivo: an ontology for describing archival resources, in: CILC, 2018, pp. 112–116.

[38] L. Pandolfo, L. Pulina, M. Zieliński, Exploring Semantic Archival Collections: the Case of Piłsudski Institute of America, in: Digital Libraries: Supporting Open Science: 15th Italian Research Conference on Digital Libraries, IRCDL 2019, Springer, 2019, pp. 107–121.

[39] L. Pandolfo, L. Pulina, Building the Semantic Layer of the Józef Piłsudski Digital Archive with an Ontology-Based Approach, International Journal on Semantic Web and Information Systems (IJSWIS) 17 (2021) 1–21.

[40] L. Pandolfo, L. Pulina, ARKIVO Dataset: a Benchmark for Ontology-based Extraction Tools, in: WEBIST, 2021, pp. 341–345.

[41] L. Derczynski, D. Maynard, G. Rizzo, M. Van Erp, G. Gorrell, R. Troncy, J. Petrak, K. Bontcheva, Analysis of Named Entity Recognition and Linking for Tweets, Information Processing & Management 51 (2015) 32–49.

[42] O. Zamazal, A Survey of Ontology Benchmarks for Semantic Web Ontology Tools, International Journal on Semantic Web and Information Systems (IJSWIS) 16 (2020) 47–68.

[43] L. Pandolfo, L. Pulina, Adnoto: A self-adaptive system for automatic ontology-based annotation of unstructured documents, in: Advances in Artificial Intelligence: From Theory to Practice: 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Springer, 2017, pp. 495–501.

[44] L. Pandolfo, L. Pulina, G. Adorni, A Framework for Automatic Population of Ontology-Based Digital Libraries, in: AI*IA 2016: Advances in Artificial Intelligence, volume 10037 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 406–417. URL: https://doi.org/10.1007/978-3-319-49130-1_30. doi:10.1007/978-3-319-49130-1_30.