Constructing a Knowledge Graph for Italian Cinema Divas' Autobiographies

Giorgio Corona, Dario Guidotti and Laura Pandolfo*

DUMAS, Università degli Studi di Sassari, via Roma 151, Sassari, 07100, Italy

Abstract

Autobiographical writings are invaluable for research, offering relevant insights into personal experiences and cultural contexts. This is particularly true for Italian actresses, whose autobiographies, while rich with information, have been relatively underexplored in academic research. The Women Writing around the Camera (WOW) project addresses this gap by developing a semantic portal dedicated to these autobiographical texts. The WOW portal will reveal the dynamics between the actresses' writings, their private lives, their artistic careers, and the cultivation of the diva image. As a first step towards this goal, this paper presents the WOW knowledge graph (KG), which maps the personal and professional networks related to the divas' lives. The KG was built starting from the actresses' autobiographies, guided by a taxonomy of themes curated by domain experts. Although still under development and expansion, the KG provides a solid foundation for future enhancements.

Keywords

Linked Open Data, Semantic Web, Cultural Heritage, Digital Humanities, Film and Literary Studies

1. Introduction

Autobiographies are not merely collections of biographical data. They are intricate narratives that integrate personal experiences, emotions, and reflections, offering a depth that goes beyond mere factual recounting. As such, they serve as invaluable resources for researchers seeking to explore and understand historical contexts and personal histories [1]. Traditionally, autobiographies are analysed through close reading [2], a manual method that offers detailed insights into personal narratives without the aid of computational tools or automated analysis. However, there is a growing interest in publishing autobiographical information as Linked Open Data [3] and leveraging Knowledge Graphs (KGs) [4]. This approach not only facilitates computational analyses typical of distant reading but also leverages Semantic Web tools and technologies, enabling a wider range of analytical and semantic methods [5, 6, 7]. By structuring autobiographical data as Linked Data, researchers can uncover patterns, relationships, and insights that might be missed through close reading alone, thus enriching the understanding of the text and expanding data-driven exploration opportunities. However, converting autobiographical information into Linked Data presents significant challenges due to the unique and complex nature of such data, necessitating careful handling and representation.

In this paper, we present our efforts in developing a knowledge graph within the context of the Women Writing around the Camera (WOW) project [8], which aims to create a semantic portal in the field of humanities that collects and organises data related to the autobiographies of Italian actresses, known as *divagrafie* [9]. The presented knowledge graph serves as the foundational knowledge base of the portal, mapping the personal and professional networks associated with the lives of Italian cinema divas, e.g., Sophia Loren, Monica Vitti, and Franca Valeri. By focusing on the dynamics between actresses' writings and films in which they performed, our work aims to integrate isolated datasets into a cohesive and harmonised KG. The data we use includes biographical and bibliographic information,

along with a comprehensive taxonomy of concepts and keywords that encapsulate the key themes for studying the lives of Italian actresses. To construct the KG, we began with semi-automatically generated annotations made by domain experts on the autobiographical texts, in a manner similar to that applied in [10, 11]. These annotations were then transformed into triples in Resource Description Framework (RDF) format, integrating with existing data sources from the Linked Data Cloud. This amount of data allows scholars to explore the lives and careers of these actresses from a broader perspective.

This paper is organised as follows. Section 2 presents relevant examples of datasets and structured knowledge in the cultural heritage domain concerning literary texts and biographical materials. Section 3 presents the methodology used for implementing the KG, by describing source datasets, data model and the population of the dataset. An analysis of the data quality and usefulness is given in Section 4, while conclusion is provided in Section 5.

2. Related Work

The digitisation and semantic enrichment of corpora have significantly transformed the accessibility and analysis of biographical and literary data. Several initiatives have developed datasets to facilitate cataloguing and study of such data in a variety of research contexts. Some examples include the following:

- **BiographyNet** [12] is a project that enables semantic analysis of biographical dictionaries from the 18th century to today. Its main goals are to facilitate prosopographical analysis of social groups, examine biography topics, and provide insights into historiographical approaches across countries and languages.
- **BiographySampo** [13] is a semantic portal based on Finland's National Biography Collection, transforming biographical texts into Linked Data. It offers advanced tools for data visualization, analysis, and knowledge discovery, specifically designed for prosopographical research in the humanities.
- **The World Literature Knowledge Graph** [14] features over 194,000 writers and 971,000 works, enabling exploration of global literature and authors. It integrates data on the reception of these works across reader communities within a unified semantic model.
- **WarSampo** [15] represents a significant application of knowledge graphs in historical research. By integrating diverse sources like Finnish national archives, war diaries, veterans' magazines, photographs, and prisoner records, it offers a comprehensive view of World War II, facilitating both large-scale and micro-historical research using Semantic Web technologies.
- **arkivo** [16, 17, 18, 19, 20] catalogues the extensive collection of archives, manuscripts, photographs, and artefacts held by the Józef Piłsudski Institute of America. By linking these resources to Linked Open Data databases such as DBpedia [21] and Wikidata [22], the arkivo dataset enriches the available information, facilitating further research into Polish history.
- **Memorata Poetis** [23] focuses on preserving and exploring ancient poetic traditions from Greek, Latin, Italian, and Arabic literature. Its database incorporates semantic web features and is enriched with Linked Open Data resources, such as Pleiades for geographic data and DBpedia for various entities, providing enhanced tools for textual analysis.

By integrating and harmonising complex datasets, these initiatives provide rich data for exploring historical and literary information.

3. The WOW Knowledge Graph

The WOW project's KG captures the personal and professional relationships of Italian cinema divas from their autobiographies. The following sections detail its development, starting with the source datasets. We then describe the data model, which structures information to represent complex relationships. Finally, we explain the workflow for populating the model, detailing how raw data was transformed into RDF, linked to external datasets, and incorporated into the KG.

3.1. Source Datasets

The WOW KG uses data from three primary sources: divas' autobiographies dataset, DBpedia, and Wikidata. The divas' autobiographies dataset has been developed from the annotations made by domain experts, which catalogued 103 autobiographies written by 59 actresses, identifying 479 quoted passages and 19 scholarly themes centred around 232 keywords. Additionally, 374 individual names mentioned in the texts were annotated, along with their occurrences. This data was provided in CSV format. From DBpedia, we obtained supplementary information on actress-writers and the individuals they mention, as well as details about the films in which these actresses participated, including 448 film entities. Additionally, we collected information on the professionals involved in these films, as well as the actresses' relatives and romantic partners, adding 1,049 names to the dataset. From Wikidata, we incorporated user IDs for the divas across major social networks.

3.2. Data Model

The annotations were converted into RDF triples and integrated with other data sources into a cohesive data model [1]. Concepts from the source datasets are described using metadata schemas [24], such as DCMI Metadata Terms, and vocabulary models like SKOS and RDF Schema. Key entities and properties were extensively reused, applying established reuse strategies [25] to promote interoperability and reusability. Table 1 lists the imported ontologies and their namespace prefixes, while Figure 1 illustrates the schema of classes, subclasses, and properties that comprise the data model.

| Ontology	Prefix Name	Expansion
arkivo	arkivo:	http://purl.org/arkivo/ontology#
DBpedia	dbo:	http://dbpedia.org/ontology/
DBpedia	dbp:	http://dbpedia.org/property/
DCMI	dcterms:	http://purl.org/dc/terms/
DoCo	doco:	http://purl.org/spar/doco/
FOAF	foaf:	http://xmlns.com/foaf/0.1/
Pattern	po:	http://www.essepuntato.it/2008/12/pattern#
RDF Schema	rdfs:	http://www.w3.org/2000/01/rdf-schema#
Schema	schema:	http://schema.org/
SKOS	skos:	http://www.w3.org/2004/02/skos/core#

Table 1
Details on the re-used ontologies in the data model. It includes the ontology name, the corresponding prefix used, and the expansion of each prefix into its full *Internationalized Resource Identifier* (IRI).

For clarity, our data model can be categorised into three main areas. The first area, bibliographic and thematic, focuses on the autobiographies as literary works, incorporating bibliographic details and thematic content. The second area addresses the professional aspects of the actresses' lives, detailing their film roles and career-related information. The third area relates the personal aspects, including information about the actresses' family and sentimental lives. Each of these areas is described in detail below.

3.2.1. Bibliographic Data and Themes

The bibliographic data and classification system by subject and keyword form a key part of our dataset. Autobiographies are catalogued on the basis of their first edition, with parameters such as title, publisher,

---

[1]The data model is available at this link: https://github.com/AIMet-Lab/PRIN-WOW/blob/main/wow_schema.ttl. To produce a human-readable form of the schema, the pyLODE documentation can be found here: https://aimet-lab.github.io/PRIN-WOW/
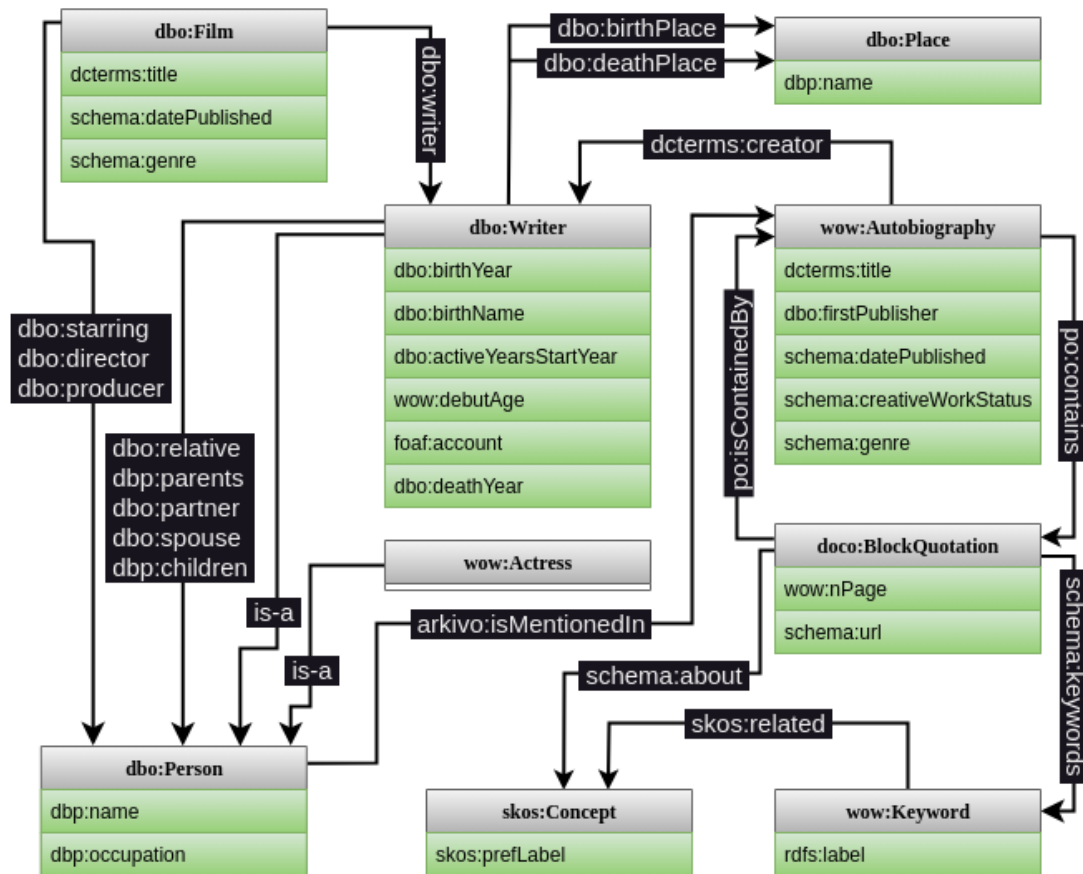
Figure 1: Schema of the WOW data model, illustrating the classes, properties and relationships.

year of publication, and genre (including several subcategories of the autobiography genre). These data are supplemented by information on the publication status (whether still on the market or not), represented by the `schema:creativeWorkStatus` property.

Excerpts from autobiographies, selected and analysed by scholars, are indexed in the WOW knowledge base by page number (`wow:nPage`), stored into the `doco:BlockQuotation` class (imported from the DoCO [26] ontology), and linked to the texts from which they were extracted through `po:contains` and `po:isContainedBy` properties from the Pattern ontology[2]. The content of each quoted passage is classified by one or more pairs of terms formed by a theme (`skos:Concept`) and a keyword, e.g. `:Melancholy` (theme) and `:Loneliness` (keyword), or `:Body` and `:Mirror`. Usually each keyword belongs to only one theme, but rarely a keyword can belong to more than one theme: `:Desire and attraction` can refer to both `:Relationship with men` and `:Relationship with women` themes (as well as `:Disappointment`). The above mentioned non-hierarchical semantic relationship between themes and keywords was implemented through the use of the `skos:related` property.

Names referenced in autobiographies are linked to the texts in which they are mentioned through the `arkivo:isMentionedIn` property, which has been imported from the arkivo ontology. It is worth noting the implementation of two further annotation properties not present in the above schema's depiction. These are: `wow:nTimes`, which enumerates the number of mentions of names in autobiographies; and `wow:artistAge`, which notes the age of the divas at the time of the publication of the autobiographies.

3.2.2. Filmographic Data and Divas' Professional Lives

To gain insight into the professional lives of the actresses and delineate a timeline, we have started collecting information on their year of debut (`dbo:activeYearsStartYear`) and the age of debut (`wow:debutAge`). We then gathered their filmographies from DBpedia: these included not only performances as actresses (`dbo:starring`) but also roles as directors, producers, or screenwriters (`dbo:writer`). After annotating these properties with the personal age of the divas at the time using the `wow:artistAge` annotation property, we provided each film entity with information about genre and release year (using the same properties as for the autobiographies). For each movie, we then gathered information about cast, director, and producer: the retrieved names were incorporated into the WOW dataset, along with the names of individuals referenced in the autobiographical texts. For each name in the dataset, we then collected information about their occupations through the `dbp:occupation` property. To date, the WOW dataset contains information on 448 films and an additional 1,116 names of colleagues and others who were part of the actresses' professional lives.

In the intersection between public and private, social networks provide valuable data, as noted by experts. For each diva in the database, efforts were made to retrieve their account IDs on major platforms (Facebook, Instagram, X). A search via Wikidata produced 36 results, recorded using the `foaf:account`, opening opportunities for future research.

3.2.3. Divas' Private Lives

In regard to the personal histories of the writing divas, domain experts provided data on their dates of birth and death, which we supplemented with the corresponding geographical information (`dbo:birthPlace` and `dbo:deathPlace`), as well as with birth names. A further attempt was made to reconstruct the private lives of the divas by collecting data about parents, children, relatives, partners and husbands. For each name collected, as before for names mentioned in texts or collected from filmographic research, professions were then recorded. This kind of research presents a significant challenge for domain experts, who are therefore highly interested in the results. To date, 64 entities with different degrees of relatedness have been documented. However, this number is expected to grow as more data from Wikidata is integrated.

3.3. Populating the Data Model

The names of individuals mentioned in the texts were extracted using Named Entity Recognition (NER) and stored in CSV files. We then assigned semantic meaning to these names, including those of the writer divas, through Named Entity Linking (NEL). This process generated the IRIs necessary for subsequent searches. These IRIs were used to perform SPARQL queries against the DBpedia endpoint[3] and to query the Wikidata endpoint[4] to obtain the social network account IDs of the divas. The individuals representing persons, movies, and geographic locations are imported from DBpedia, as well as the classes that contain them: `dbo:Person`, `dbo:Film` and `dbo:Place`. Actresses are further catalogued, according to their profession, into the `wow:Actress` subclass of the `dbo:Person` class, and writer divas into both the `wow:Actress` and the `dbo:Writer` subclasses. The `doco:BlockQuotation` class and the `skos:Concept` class were imported and populated with individuals representing text excerpts and their themes, while the `wow:Autobiography` class and the `wow:Keyword` class have been implemented. Figure 2 illustrates a representation of the data transformation pipeline and the connections established through NEL, as described above. The WOW dataset currently comprises 2,792 individuals, 1,266 of which are derived from annotations by domain experts and 1,526 have been imported from external resources. The most relevant properties and their frequency are reported in Table 2. The KG is available on Zenodo, with an associated citation [27], and is licensed with the open Creative Commons BY 4.0 license.
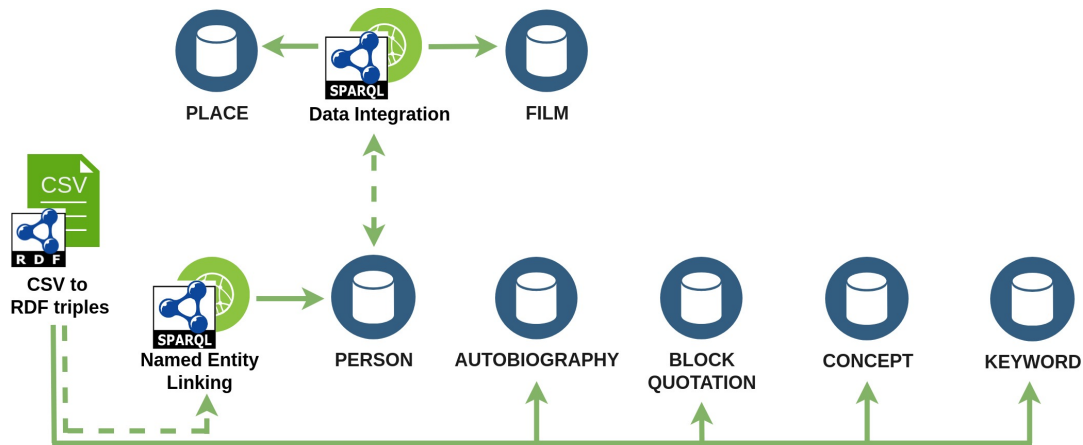
---

[3]https://dbpedia.org/sparql/
[4]https://query.wikidata.org/sparql

Property Name	Usage
dbp:occupation	1,929
dbo:starring	1,516
schema:keywords	1,435
schema:about	1,014
dbo:producer	292
dbo:director	237

Table 2
Usage count of main property linking class instances.

Figure 2: Graphical representation of the data flow process. The cylinders represent RDF data divided by class, dashed arrows indicate NEL and data integration steps, while solid arrows indicate the population of classes. In the initial phase, annotations were converted from CSV to RDF triples. Entities of the class dbo:Person underwent NEL and data integration processes, while the classes dbo:Place and dbo:Film were populated with data from DBpedia. All other classes were populated from the annotations.

4. Data Quality & Usefulness

The WOW knowledge graph is based on heterogeneous data that are considered high quality, as they are ensured through multiple layers of validation and enrichment processes. First, domain experts meticulously annotate autobiographical texts, identifying key themes, quoted passages, and references, which are then transformed into RDF triples. This process leverages both manual and semi-automated techniques to enhance accuracy. Additionally, integration with external datasets such as DBpedia and Wikidata enriches the KG by providing supplementary information and establishing robust links between entities.

In terms of fairness and data rights management, the knowledge graph leverages publicly available information and datasets, including published autobiographies and data from Wikidata and DBpedia. These sources are open and accessible, which supports transparency and reproducibility in research. The validation and enrichment processes ensure that the data is accurate and representative, while the use of open data adheres to principles of equitable access and responsible data usage.

The usefulness of the KG is evaluated based on its ability to facilitate meaningful research and insights. Its Linked Data structure allows complex queries that reveal patterns and relationships not visible in unstructured texts. For example, researchers can explore connections between Italian cinema divas' autobiographies and their film roles. By integrating bibliographic details, filmographies, and personal histories, the KG supports diverse research objectives and provides a comprehensive view of the subject.

5. Conclusion

The development of the presented KG marks an advancement in representing actresses' autobiographies within the humanities. By converting autobiographical texts into Linked Data and integrating them with external datasets, this preliminary version of the KG offers a comprehensive, semantically rich resource for examining the lives and careers of Italian cinema divas. However, it is important to note that the KG is still a work in progress. As the dataset continues to evolve, efforts will focus on expanding coverage and exploring additional research applications. Future developments aim to maximize the graph's impact by enriching the autobiographical data and its connections to broader cultural contexts.

Acknowledgments

References

[1] K. Thomas, Changing Conceptions of National Biography: The Oxford DNB in Historical Perspective, Cambridge University Press, 2005. URL: http://dx.doi.org/10.1017/cbo9780511497582.001. doi:10.1017/cbo9780511497582.001.

[2] F. Moretti, Distant Reading, Verso Books, 2013.

[3] E. Hyvönen, Publishing and Using Cultural Heritage Linked Data on the Semantic Web, volume 3, Morgan & Claypool Publishers, 2012. URL: https://doi.org/10.1007/978-3-031-79438-4. doi:10.1007/978-3-031-79438-4.

[4] C. Gutierrez, J. F. Sequeda, Knowledge Graphs, Commun. ACM 64 (2021) 96–104. URL: https://doi.org/10.1145/3418294. doi:10.1145/3418294.

[5] E. Hyvönen, Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery, Semantic Web 11 (2020) 187–193. URL: http://dx.doi.org/10.3233/SW-190386. doi:10.3233/SW-190386.

[6] G. Adorni, M. Maratea, L. Pandolfo, L. Pulina, An Ontology for Historical Research Documents, in: Web Reasoning and Rule Systems - 9th International Conference, RR, volume 9209 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 11–18. URL: https://doi.org/10.1007/978-3-319-22002-4_2. doi:10.1007/978-3-319-22002-4_2.

[7] G. Adorni, M. Maratea, L. Pandolfo, L. Pulina, An ontology-based archive for historical research, in: Proceedings of the 28th International Workshop on Description Logics, volume 1350 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015. URL: https://ceur-ws.org/Vol-1350/paper-41.pdf.

[8] L. Pandolfo, L. Cardone, L. Cutzu, R. Perna, B. Seligardi, G. Simi, The WOW Project: Bridging AI and Cultural Heritage for Actress Writings, in: Proceedings of the 2nd Workshop on Artificial Intelligence for Cultural Heritage (IAI4CH 2023) co-located with AIxIA 2023, volume 3536 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 34–41. URL: https://ceur-ws.org/Vol-3536/04_paper.pdf.

[9] M. Rizzarelli, Il doppio talento dell'attrice che scrive. per una mappa delle "divagrafie", Cahiers d'études italiennes (2021). URL: http://dx.doi.org/10.4000/cei.9005. doi:10.4000/cei.9005.

[10] 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.

[11] L. Pandolfo, L. Pulina, Adnoto: A self-adaptive system for automatic ontology-based annotation of unstructured documents, in: S. Benferhat, K. Tabia, M. Ali (Eds.), 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*, Springer, 2017, pp. 495–501. URL: https://doi.org/10.1007/978-3-319-60042-0_54. doi:`10.1007/978-3-319-60042-0\_54`.

[12] A. Fokkens, S. ter Braake, N. Ockeloen, P. Vossen, S. Legêne, G. Schreiber, V. de Boer, BiographyNet: Extracting Relations Between People and Events, arXiv e-prints (2018) arXiv–1801. URL: https://doi.org/10.48550/arXiv.1801.07073. doi:`10.48550/arXiv.1801.07073`.

[13] M. Tamper, P. Leskinen, E. Hyvönen, R. Valjus, K. Keravuori, Analyzing Biography Collections Historiographically as Linked Data: Case National Biography of Finland, Semantic Web 14 (2023) 385–419. doi:`10.3233/SW-222887`.

[14] M. A. Stranisci, E. Bernasconi, V. Patti, S. Ferilli, M. Ceriani, R. Damiano, The World Literature Knowledge Graph, in: T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, J. Li (Eds.), The Semantic Web – ISWC 2023, Springer Nature Switzerland, Cham, 2023, pp. 435–452. doi:`10.1007/978-3-031-47243-5\_24`.

[15] M. Koho, E. Ikkala, P. Leskinen, M. Tamper, J. Tuominen, E. Hyvönen, Warsampo Knowledge Graph: Finland in the Second World War as Linked Open Data, Semantic Web 12 (2021) 265–278.

[16] L. Pandolfo, L. Pulina, M. Zielinski, Towards an ontology for describing archival resources, in: Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) co-located with ISWC, volume 2014 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017, pp. 111–116. URL: https://ceur-ws.org/Vol-2014/paper-12.pdf.

[17] L. Pandolfo, L. Pulina, M. Zielinski, ARKIVO: An Ontology for Describing Archival Resources, in: Proceedings of the 33rd Italian Conference on Computational Logic, 2018, volume 2214 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 112–116. URL: https://ceur-ws.org/Vol-2214/paper12.pdf.

[18] L. Pandolfo, L. Pulina, Building the Semantic Layer of the Józef Piłsudski Digital Archive With an Ontology-Based Approach, Int. J. Semantic Web Inf. Syst. 17 (2021) 1–21. URL: https://doi.org/10.4018/ijswis.2021100101. doi:`10.4018/IJSWIS.2021100101`.

[19] L. Pandolfo, L. Pulina, M. Zielinski, Exploring Semantic Archival Collections: The Case of Piłsudski Institute of America, in: 15th Italian Research Conference on Digital Libraries, IRCDL 2019, volume 988 of *Communications in Computer and Information Science*, Springer, 2019, pp. 107–121. URL: https://doi.org/10.1007/978-3-030-11226-4_9. doi:`10.1007/978-3-030-11226-4\_9`.

[20] L. Pandolfo, L. Pulina, ARKIVO dataset: A benchmark for ontology-based extraction tools, in: Proceedings of the 17th International Conference on Web Information Systems and Technologies, WEBIST, SCITEPRESS, 2021, pp. 341–345. URL: https://doi.org/10.5220/0010677000003058. doi:`10.5220/0010677000003058`.

[21] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al., DBpedia–A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia, Semantic web 6 (2015) 167–195. doi:`10.3233/SW-140134`.

[22] D. Vrandečić, M. Krötzsch, Wikidata: a Free Collaborative Knowledgebase, Communications of the ACM 57 (2014) 78–85. URL: http://dx.doi.org/10.1145/2629489. doi:`10.1145/2629489`.

[23] F. Khan, S. Arrigoni, F. Boschetti, F. Frontini, Restructuring a Taxonomy of Literary Themes and Motifs for More Efficient Querying, MATLIT: Materialidades da Literatura 4 (2016) 11–27. URL: http://dx.doi.org/10.14195/2182-8830_4-2_1. doi:`10.14195/2182-8830_4-2_1`.

[24] R. Gartner, Metadata: Shaping Knowledge from Antiquity to the Semantic Web, Cham, Springer International, 2017. doi:`https://doi.org/10.1007/978-3-319-40893-4`.

[25] V. A. Carriero, M. Daquino, A. Gangemi, A. G. Nuzzolese, S. Peroni, V. Presutti, F. Tomasi, The Landscape of Ontology Reuse Approaches, in: Applications and Practices in Ontology Design, Extraction, and Reasoning, IOS Press, 2020, pp. 21–38. URL: http://dx.doi.org/10.3233/ssw200033. doi:`10.3233/ssw200033`.

[26] A. Constantin, S. Peroni, S. Pettifer, D. M. Shotton, F. Vitali, The Document Components Ontology (DoCO), Semantic Web 7 (2016) 167–181. URL: https://doi.org/10.3233/SW-150177. doi:`10.3233/SW-150177`.

[27] L. Pandolfo, G. Corona, D. Guidotti, Women Writings Around the Camera Knowledge Graph, [Dataset], 2024. doi:`10.5281/zenodo.13784082`.