Mining Typewritten Digital Representations to Support Archival Description (short paper)

Mariana Dias[1], Carla Teixeira Lopes[1]

[1]Faculty of Engineering of the University of Porto and INESC-TEC, Portugal

Abstract

Linked Data is used in various fields as a new way of structuring and connecting data. Cultural heritage institutions have been using linked data to improve archival descriptions and promote findability. The required detail in manual descriptions of cultural heritage objects can be taxing and time-consuming. Given this, in EPISA, a research project on this topic, we propose to use the contents of the digital representations associated with the objects to assist archivists in their description tasks. More specifically, to extract information from the digital representations useful for an initial ontology population that should be validated or edited by the archivist. We apply optical character recognition in an initial stage to convert the digital representation to a machine-readable format. We then use ontology-oriented programming to identify and instantiate ontology concepts using neural networks and contextual embeddings.

Keywords

Cultural Heritage, Information Extraction, Optical Character Recognition, Ontology Population, Semantic Web

1. Introduction

Cultural heritage institutions have the mission of protecting, valuing, and sharing national inheritances. Linked Data has provided the possibility to improve the quality of archival descriptions and promote access to enriched cultural archives by providing users with a more in-depth knowledge of collections. However, manually describing cultural heritage objects can be taxing and time-consuming, making it challenging to describe documents or collections in finer detail. The automatic extraction of information from digital representations relevant to the archival description can ease the work of cultural heritage professionals.

EPISA (Entity and Property Inference for Semantic Archives) is a research project that explores the use of Linked Data in the context of the Portuguese National Archives. ArchOnto[1] [1], a linked data model for archives, was proposed in this project. This paper presents an overview of what is being done in an EPISA task that aims to extract concepts and relations from digital representations of archival records and map them to ArchOnto. This task's final goal is to provide these mappings as suggestions in the user interface to speed-up future descriptions. It is not the purpose of this paper to go into detail about each stage of the process, including

TPDL2022: 26th International Conference on Theory and Practice of Digital Libraries, 20-23 September 2022, Padua, Italy

up201606486@up.pt (M. Dias); ctl@fe.up.pt (C. T. Lopes)

0000-0002-4202-791X (C. T. Lopes)

[1]https://purl.archive.org/episa/archonto

their evaluation, which we will describe in other articles. Our goal with this paper is to report the ongoing work supporting linked data archival description in a research project, share our experience, and obtain feedback from others participating in the workshop.

2. Related Work

Recently, research works in the cultural heritage field have used semantic representations of cultural heritage objects to improve accessibility to the resources and information extraction techniques to represent information related to the objects that sustain the knowledge bases. Projects with the goal of extracting information from non-machine-readable historical documents into ontologies are presented in several works [2, 3, 4, 5]. Witte et al. [2] and Vlachidis and Tudhope [4] developed automatic ontology-based information extraction approaches based on linguistic NLP techniques for a 19th-century encyclopedia of compiled architectural knowledge and archaeological grey-literature reports, respectively. Packer and Embley [3] created an automatic tool, ListReader, to extract information from lists in OCRed documents using a wrapper induction technique that populates a user-defined ontology. Goy et al. [5] presented a proof-of-concept prototype [6] of a crowdsourcing platform using non-machine-readable documents from the Istituto Gramsci dating from 1968 to 1969. Experts participate in the semantic annotation process guided by the ontology and supported by suggestions provided by automatic Information Extraction techniques.

3. Proposed Approach

We propose an architecture divided into three modules: Optical Character Recognition (OCR), Information Extraction (IE), and Ontology Population (OP). The execution pipeline is described in Figure 1 that shows the relation of the project task in which we are working on *Document mining for automatic metadata records* with the task *Exploration and querying interface*. Upon a non-machine-readable archival digital representation uploaded to the user interface, the OCR module extracts its textual content using a pre-processing image phase. The IE module processes the textual content and uses a trained NER model to predict and annotate concepts. The concepts are mapped to the ArchOnto ontology creating candidate concepts and relations suggested as description values to the archivist. We detail each of these modules in the following subsections.

3.1. Optical Character Recognition

The success of text recognition depends on the quality of digital representations, and it is common for heritage documents to suffer some degree of degradation over time. From uneven illumination to erased characters and angled digital representations, image processing methods can be applied before the text recognition phase to improve the image quality and the overall text extraction. We conducted an optimization experiment of different image algorithms' and parametrization using the OpenCV [7] library and a non-dominated sorting genetic algorithm

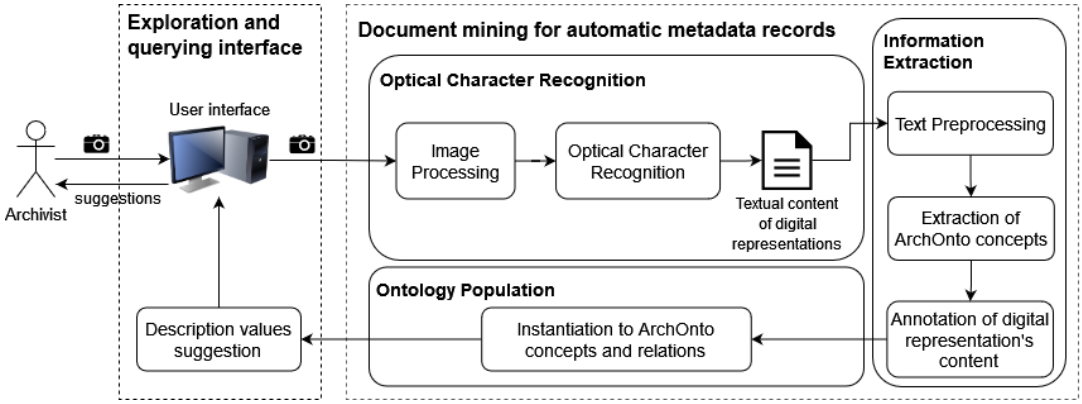

Figure 1: Architecture of the proposed approach.

(NSGA-II) to determine the impact of image processing algorithms on the OCR performance. Converting digital representations to a machine-readable format is executed with Tesseract [8].

To better illustrate the process, we show an excerpt of the OCR output of a typewritten letter in Figure 2.

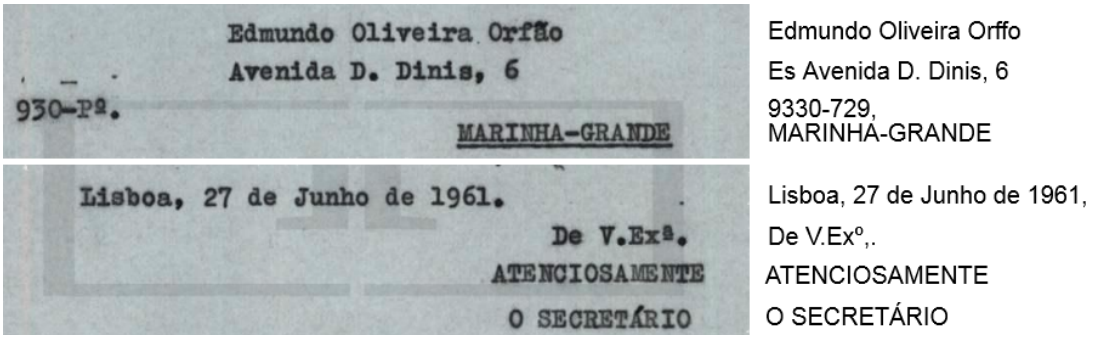

Figure 2: Extracted content from the heading and closing of a typewritten letter.

3.2. Information Extraction

This module firstly processes the textual content extracted from digital representations in the OCR module. This pre-processing includes removing characters and punctuation introduced by OCR that are not part of the original works, removing multiple empty lines, and rejoining hyphenated words.

We trained a NER model using a Bidirectional Long Short-Term Memory Network with a Conditional Random Field Layer (BiLSTM-CRF) Neural Network model and pre-trained contextual string embeddings by Santos et al. [9]. An annotated Portuguese dataset is necessary for model training with the Sequence Tagger. However, there are no available annotated archival

collections in Portuguese and manually creating a large annotated corpus is time-consuming. As an alternative, an available Portuguese corpus for named entities, the Second HAREM collection [10], was pre-processed to suit the ArchOnto ontology's classes. The collection includes 1,040 documents with 7,847 named entities. Annotations that are not relevant for the ontology population task were stripped from the annotated corpus, and suitable annotations were mapped to ArchOnto classes. For instance, textual elements tagged with labels relating to the concept of a person were mapped to a CIDOC-CRM person (class E21 Person), and elements tagged with labels relating to the concept of the role of a person in an event were mapped to the ArchOnto role (class ARE8 Role Type).

We adopted the BIOES annotation scheme, a variant of the BIO [11] scheme, to label multi-token named entities. Each token is firstly tagged with a label that represents if it is at the beginning (B), inside (I), or end (E) of a named entity, if it is a single-token entity (S), or if it is outside (O) of a named entity. The transformed corpus was used as training data. We trained the final model on the entire collection with a split of 80% training data and 20% validation data.

Taking the previous example of the OCR output of a typewritten letter from Figure 2 as an example, the passage refers to the heading and closing of a letter that contains the recipients' name and address, the date, and the sender's name. Listing 1 contains the list of labeled concepts identified in the letter.

```
Edmundo <B-E21> Oliveira <I-E21> Orffo <E-E21>
Avenida <B-E53> D. <I-E53> Dinis <E-E53>
Lisboa <S-E53>
27 <B-E52> de <I-E52> Junho <I-E52> de <I-E52> 1961 <E-E52>
SECRETÁRIO <S-ARE8>
```

Listing 1: Annotated concepts extracted from the typewritten letter presented in Figure 2.

3.3. Ontology Population

After annotating the text extracted from the digital representations, we must map each annotation to ArchOnto. The mapping and instantiation of concepts and relations were implemented with the Python package for ontology-oriented programming OwlReady2 [12]. Work on the ontology rules is still ongoing with the instantiation of concepts and linking nominative relations. With the concepts extracted in the IE module, Listing 2 presents the result of the instantiation.

This paper presents an automatic approach to populate the ArchOnto ontology. However, this approach can be generalized to other Linked Data models, such as RiC-O[2] (Records in Context Ontology), with the adaptation of mapping and instantiation of concepts and relations of the ontology.

4. Evaluation

For the evaluation of each of this work's modules, we first created a dataset that contains typewritten Portuguese documents from the 20th century [13]. Along each task, we used

[2]https://www.ica.org/standards/RiC/ontology

```
person1 (E21_Person), 'P1_is_identified_by', personName1 (E41_Appellation)
personName1, 'L2DO_hasValue', personName1_ (DOE17_PersonName)
personName1_, 'DOP5_name', ``Edmundo Oliveira Orffo''  (xsd:string)

place1 (E53_Place), 'P1_is_identified_by', placeName1 (E41_Appellation)
placeName1, 'L2DO_hasValue', placeName1_ (DOE8_String)
placeName1_, 'DOP7_stringValue', "Avenida D. Dinis"  (xsd:string)
place2 (E53_Place), 'P1_is_identified_by', placeName2 (E41_Appellation)
placeName2, 'L2DO_hasValue', placeName2_ (DOE8_String)
placeName2_, 'DOP7_stringValue', "Lisboa"  (xsd:string)

descriptiveDate1 (E52_Time-Span), 'P1_is_identified_by', descriptiveDateName1 (E41_Appellation)
descriptiveDateName1, 'L2DO_hasValue', descriptiveDateName1_ (DOE10_Instant)
descriptiveDateName1_, 'DOP8_timestamp', ``1961-06-27T00:00:00'' (xsd:dateTime)

pc14_1 (PC14_Carried_Out_By), 'P01_has_domain', person2 (E21_Person)
pc14_1 (PC14_Carried_Out_By), 'P14.1_in_the_role_of', secretario (ARE8_RoleType)
```

Listing 2: Instantiation of concepts and relations from the annotation presented in Listing 1.

subsets of this dataset to suit the different experiments' goals and requirements. For the OCR module, we transcribed 708 typewritten digital representations extracted from the original dataset. This new dataset [14] will be split in two, one for parameter optimization and another for evaluation. The model trained in the IE module will be evaluated using the First HAREM collection [15]. The corpus will be transformed similarly to the approach detailed with the Second HAREM corpus and used as testing data. For the evaluation of the OP module, we will ask two archivists to provide ArchOnto representations for a subset of thirteen aggregated digital representations of archival records that were manually transcribed. From these descriptions, a consensual representation will later be defined and used for a comparison with the output of our automatic approach.

5. Conclusions and Future Work

This paper presented our approach for the automatic population of a domain-specific ontology with information extracted from non-machine-readable digital representations of cultural heritage documents. We will extract further information from the documents in the future to create rich relations between the concepts identified in the NER module. Additionally, we will evaluate each module to determine the quality of the results. We will also develop two APIs: one for the OCR module that extracts the content of digital representations and another for the OP module that populates the ArchOnto ontology given a textual file. Furthermore, we will integrate the pipeline we developed into the EPISA interfaces by suggesting concepts and relations when an archivist uploads a digital representation.

Acknowledgments

This work is financed by National Funds through FCT - Foundation for Science and Technology I.P., within the scope of the EPISA project - DSAIPA/DS/0023/2018.

References

[1] I. Koch, C. T. Lopes, C. Ribeiro, Moving from ISAD(G) to a CIDOC-CRM-based Linked Data Model in the Portuguese Archives, Journal on Computing and Cultural Heritage (2021).

[2] R. Witte, R. Krestel, T. Kappler, P. C. Lockemann, Converting a Historical Architecture Encyclopedia into a Semantic Knowledge Base, IEEE Intelligent Systems 25 (2010).

[3] T. Packer, D. Embley, Cost effective ontology population with data from lists in OCRed historical documents, 2013, pp. 44–52. doi:10.1145/2501115.2501132.

[4] A. Vlachidis, D. Tudhope, A knowledge-based approach to information extraction for semantic interoperability in the archaeology domain, Journal of the Association for Information Science and Technology 67 (2015). doi:10.1002/asi.23485.

[5] A. Goy, R. Damiano, F. D. Loreto, D. Magro, S. Musso, D. P. Radicioni, C. Accornero, D. Colla, A. Lieto, E. Mensa, M. Rovera, D. Astrologo, B. Boniolo, M. D'Ambrosio, PRiSMHA (Providing Rich Semantic Metadata for Historical Archives), in: JOWO, 2017.

[6] D. Colla, A. Goy, M. Leontino, D. Magro, C. Picardi, Bringing semantics into historical archives with computer-aided rich metadata generation, J. Comput. Cult. Herit. (2021). URL: https://doi.org/10.1145/3484398. doi:10.1145/3484398, just Accepted.

[7] G. Bradski, The OpenCV Library, Dr. Dobb's Journal of Software Tools (2000).

[8] GitHub - tesseract-ocr/tesseract: Tesseract Open Source OCR Engine (main repository), https://github.com/tesseract-ocr/tesseract, 2022.

[9] J. Santos, B. Consoli, C. dos Santos, J. Terra, S. Collonini, R. Vieira, Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition, in: Proceedings of the 8th Brazilian Conference on Intelligent Systems, 2019, pp. 437–442.

[10] C. Freitas, C. Mota, D. Santos, H. G. Oliveira, P. Carvalho, Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/412_Paper.pdf.

[11] L. A. Ramshaw, M. P. Marcus, Text Chunking using Transformation-Based Learning, CoRR cmp-lg/9505040 (1995). URL: http://arxiv.org/abs/cmp-lg/9505040.

[12] J.-B. Lamy, Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies, Artificial Intelligence in Medicine 80 (2017). doi:10.1016/j.artmed.2017.07.002.

[13] M. Dias, Typewritten Digital Representations of Portuguese Cultural Heritage Documents from the 20th century, Data set, 2022. doi:10.25747/ZC25-1531.

[14] M. Falcão, Manual Transcriptions of Typewritten Digital Representations of Portuguese Cultural Heritage Documents from the 20th Century, Data set, 2022. doi:10.25747/WPNA-JE39.

[15] D. Santos, N. Seco, N. Cardoso, R. Vilela, HAREM: An Advanced NER Evaluation Contest for Portuguese, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, 2006. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/59_pdf.pdf.