

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

ArchMine: Learning from non-machine-readable documents for additional insights

Mariana Ferreira Dias



Mestrado em Engenharia Informática e Computação

Supervisor: Carla Teixeira Lopes

March 31, 2023

ArchMine: Learning from non-machine-readable documents for additional insights

Mariana Ferreira Dias

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Sérgio Nunes

Referee: Prof. Carla Teixeira Lopes

Referee: Prof. Pedro Henriques

March 31, 2023

Resumo

Dados Ligados são uma nova forma de estruturar e conectar dados que tem vindo a ser implementada em várias áreas, incluindo a área do património cultural. Os arquivos têm como missão proteger, valorizar e partilhar património cultural. Estas instituições são silos de informação que detêm e conservam conhecimento nos seus catálogos e bases de dados. O paradigma de Dados Ligados abriu novas oportunidades de exploração e enriquecimento de coleções, o que pode melhorar a qualidade dos arquivos digitais e promover o seu acesso ao disponibilizar aos utilizadores um conhecimento mais aprofundado das coleções e melhores ferramentas de pesquisa. Neste contexto, os arquivos têm vindo a usar Dados Ligados como uma maneira de descobrir informação e de melhorar as descrições arquivísticas com o aumento da granularidade de informação. Isto deve-se ao uso de ligações semânticas em Dados Ligados, uma vez que cada componente numa descrição de um registo pode ser agregada num grafo em que nós e arcos representam conjuntos de triplos (sujeito, predicado, objeto) que podem ser acedidos individualmente. Em 2019, o projeto EPISA começou a explorar o uso de Dados Ligados nos Arquivos Portugueses.

Descrever manualmente objetos de património cultural pode ser uma tarefa exigente e morosa, dificultando o trabalho dos arquivistas de descrever documentos ou coleções de forma mais detalhada. Assim, a existência de uma ferramenta semiautomática de descrição de objetos para um modelo de Dados Ligados pode acelerar o trabalho dos arquivistas. Inclusive, alguns registos arquivísticos possuem não só descrições arquivísticas, como também representações digitais de artefactos físicos. Estas representações digitais podem ser exploradas para propor valores de descrição. Desta forma, o objetivo da nossa tese é auxiliar o trabalho de arquivistas ao extrair informação a partir de representações digitais associadas a documentos e sugeri-la como valores de descrição de registos arquivísticos.

A maioria das representações digitais disponíveis nos Arquivos Portugueses encontram-se na forma de imagens digitalizadas. Consequentemente, torna-se necessário converter as representações para um formato legível por máquina para aproveitar o potencial de Dados Ligados. O Reconhecimento Óptico de Caracteres (OCR) é uma tecnologia que permite reconhecer texto a partir de imagens e traduzi-lo para texto legível por máquina. Neste enquadramento, avaliamos o impacto de métodos de processamento de imagem e configuração de parâmetros em OCR com documentos de património cultural datilografados. A nossa abordagem formula a configuração de parâmetros dos métodos como um problema de otimização multi-objetivo que minimiza a distância de edição de *Levenshtein* e maximiza o número de palavras corretamente identificadas com o algoritmo *Non-Dominated Sorting Genetic Algorithm* (NSGA-II). Os resultados da avaliação demonstram que a parametrização por tipologia de representações digitais beneficia o desempenho de algoritmos de processamento de imagem em OCR. Além disso, os nossos resultados sugerem que o uso de algoritmos de processamento de imagem em OCR pode ser mais adequado para tipologias em que a tarefa de reconhecimento de texto sem pré-processamento não produz bons resultados. Em particular, *Adaptive Thresholding*, *Bilateral Filter* e *Opening* são os algoritmos com o melhor desempenho para as capas de peças de teatro, cartas e para o conjunto de dados no

geral, respetivamente, e devem ser aplicados antes do OCR para melhorar o seu desempenho.

Adicionalmente, propomos uma *pipeline* de extração de informação baseada em ontologias que usa modelos de aprendizagem máquina de extração de informação para extrair entidades e relações entre entidades a partir do resultado de OCR do conteúdo textual presente em representações digitais e um povoamento inicial de uma ontologia para arquivos, ArchOnto, criada no contexto do projeto EPISA. Com essa finalidade, utilizamos programação orientada a ontologias para instanciar classes e propriedades da ontologia a partir dos conceitos identificados por modelos de extração de informação de redes neuronais e relações extraídas por *Conditional Random Fields* (CRFs). Além disso, desenvolvemos uma implementação em Python da avaliação da povoação de ontologias baseada em métricas hierárquicas e lexicais. Os resultados da avaliação dos modelos de extração de informações são semelhantes a outros trabalhos de pesquisa. No entanto, os resultados da avaliação do povoamento da ontologia com texto resultante de OCR não são satisfatórios. A avaliação inicial do povoamento da ontologia revela pistas relevantes para trabalho futuro.

A fase final da nossa solução proposta é integrar a *pipeline* de extração de informação baseada em ontologias na plataforma EPISA. Com esse intuito, desenvolvemos *mockups* da funcionalidade de sugestão de metadados descritivos na página de criação da descrição de um novo registo.

Abstract

Linked Data defined a new way of structuring and connecting data that has seen its implementation among various fields, including cultural heritage. Cultural heritage institutions have the mission of protecting, valuing, and sharing national inheritances. They are information silos, holding knowledge stored in catalogs and databases. The Linked Data paradigm opened up new opportunities for exploring and enriching collections that can improve the quality of digital archives and promote access to cultural archives by providing users with a more in-depth knowledge of collections and better search tools. Cultural heritage institutions have lately been using linked data to improve archival descriptions with greater descriptions' granularity and information discovery. This is due to the use of semantic links in Linked Data, because each component in a record's description can be aggregated into a graph where nodes and edges represent sets of triples (subject, predicate, object) that are individually accessible. In 2019, the EPISA project started exploring the use of linked data in the Portuguese Archives.

Manually describing cultural heritage objects can be taxing and time-consuming, making it challenging for archival professionals to describe documents or collections in finer detail. The existence of a semi-automated object description tool for a linked data model would accelerate the work of cultural heritage professionals. Furthermore, some archival records possess not only archival descriptions but also digital representations of physical artifacts. These digital representations can be explored to propose description values. Our goal with this thesis is to assist archival professionals by extracting information from digital representations associated with cultural documents and suggesting it as description values of archival records.

Most of the available digital representations of archival records in the Portuguese Archives are in the form of scanned images. To fully take advantage of the potential of Linked Data, we need to convert the representations from non-machine-readable into a machine-readable format. Optical Character Recognition (OCR) recognizes text in images and translates it into machine-encoded text. We evaluated the impact of image processing methods and parameter tuning in OCR applied to typewritten cultural heritage documents. The approach uses a multi-objective problem formulation to minimize Levenshtein edit distance and maximize the number of words correctly identified with a Non-Dominated Sorting Genetic Algorithm (NSGA-II) to tune the methods' parameters. Evaluation results show that parametrization by digital representation typology benefits the performance of image pre-processing algorithms in OCR. Furthermore, our findings suggest that employing image pre-processing algorithms in OCR might be more suitable for typologies where the text recognition task without pre-processing does not produce good results. In particular, Adaptive Thresholding, Bilateral Filter, and Opening are the best-performing algorithms for the theatre plays' covers, letters, and overall dataset, respectively, and should be applied before OCR to improve its performance.

We propose an ontology-based information extraction pipeline that uses machine-learning information extraction models to extract entities and relations between entities from OCR-ed textual content of digital representations and an initial ontology population of a cultural heritage ontology,

ArchOnto, created in the EPISA context. We used ontology-oriented programming to instantiate ontology classes and properties using concepts identified by neural networks and relations extracted with conditional random fields. Moreover, we developed a Python implementation for the ontology population evaluation based on hierarchy and lexical evaluation measures. Evaluation results of information extraction models are in-line with other research works. However, ontology population evaluation results applied to OCR text are not satisfactory. The initial ontology population reveals relevant clues for future research.

The final phase of our proposed solution is the integration of the ontology-based information extraction and ontology population pipeline into the EPISA platform. We developed mockups of the functionality of suggesting descriptive metadata on the screen of creating a new record description.

Agradecimentos

This work was developed within the scope of the EPISA project (DSAIPA/DS/0023/2018) funded by National Funds through FCT - Portuguese Foundation for Science and Technology, I.P.

A special word of gratitude to my supervisor, Prof. Carla Teixeira Lopes, for her support throughout this arduous process. Thanks to all the observations and thoughtfulness, it was possible to go further and beyond in my knowledge and surpass my expectations on this thesis.

Thank you to all the members of the EPISA project for the companionship and helpful suggestions. A word of gratitude to Margarida Falcão for creating a dataset of transcriptions of typewritten digital representations from the Portuguese National Archives for her curricular internship. To have worked with you on this theme was my pleasure. Another special thanks to Inês Koch for always being available to answer my, believe me, many questions about the ArchOnto model.

Thank you to DGLAB and to the archival professionals who participated in the consensual description of a dataset of digital representations, Dra. Ana Maria Rodrigues and Dra. Lucília Runa. Thank you for your time, patience, and commitment. You both helped me dive into archives and made me admire the hard work archivists put into descriptions, not only in contextualizing bodies of work but also with the historical research behind them.

To my parents, thank you for all the free coffee provided during the writing of this thesis. Thank you for just being there for the good and the bad times, for all the words of encouragement and for never giving up.

To my sister, thank you for always being there for me as my personal life advisor and the angel on my shoulder. One day you will learn Python! I believe in you just as much as you believed in me in this thesis. Thank you for making me a slightly more normal human being.

To my friends, thank you for all the nights out, video calls, and gaming nights always accompanied by laughter and craziness. Above all, thank you for always listening to me.

To my grandparents, even though they did not have the opportunity to accompany me through this journey, thank you for giving me the best memories and allow me to continue to make you proud. Thank you for everything.

Mariana Dias

“A vida não é um privilégio pessoal: é alguma coisa que nos ultrapassa como indivíduos; que pertence à natureza, à história e à sociedade. É dos homens como “todo”, no tempo e no espaço, que se faz a verdadeira eternidade. E é só entre eles, na sociedade, na consciência e na acção, que somos e nos sentimos reais.”

José Rodrigues Miguéis

Contents

1	Introduction	1
1.1	Context	2
1.2	Motivation	2
1.3	Objectives	3
1.4	Contributions	4
1.5	Document Structure	6
2	Background and Related Work	7
2.1	Optical Character Recognition	7
2.1.1	OCR Software	8
2.1.2	Image Processing	9
2.1.3	Parameter Optimization	9
2.1.4	Evaluation of OCR systems	12
2.1.5	Works exploring image pre-processing to improve OCR	12
2.2	Ontology-Based Information Extraction	16
2.2.1	Linguistic rules and dictionary	16
2.2.2	Statistical Machine Learning	17
2.2.3	Deep Learning	18
2.2.4	Annotated Datasets	20
2.2.5	Related Work in Cultural Heritage	21
2.3	Ontology Population	22
2.3.1	Ontologies for Cultural Heritage	22
2.3.2	Ontology Programming Interfaces	23
2.3.3	Evaluation of Ontology Learning systems	24
2.3.4	Related Work	25
3	Proposed Solution	27
3.1	Architecture	27
3.2	Optical Character Recognition	28
3.3	Ontology-Oriented Information Extraction	29
3.4	Ontology Population	29
4	Optimizing Character Recognition in Cultural Heritage Typewritten Documents	30
4.1	Methodology	30
4.1.1	Cultural Heritage Dataset	31
4.1.2	Parameterization and Evaluation Samples	33
4.1.3	Image Processing Pipeline	36
4.1.4	Evaluation	36

4.2	Parameter Tuning of Image Processing Algorithms	37
4.2.1	Problem Formalization and Optimization	37
4.2.2	Best Parameters Per Digital Representation Typology	38
4.2.3	Effect of Parameterization on OCR Performance	38
4.3	Effect of Image Pre-Processing on OCR Performance	39
4.3.1	Overall	40
4.3.2	Letters	40
4.3.3	Non-structured reports	41
4.3.4	Process covers	41
4.3.5	Structured reports	41
4.3.6	Theatre plays' covers	42
4.4	Discussion	42
5	Ontology-Based Information Extraction	45
5.1	Corpora Transformation	45
5.1.1	Dataset Pre-Processing	45
5.1.2	Dataset Parsing	57
5.1.3	Characterization of Transformed Corpora	59
5.2	Named Entity Recognition	59
5.3	Relation Extraction	61
5.4	Evaluation	61
5.4.1	Results	62
5.5	Discussion	64
6	Ontology Population	66
6.1	Implementation	66
6.2	Evaluation	67
6.2.1	Consensual Dataset of ArchOnto Representations	70
6.2.2	Results	70
6.3	Discussion	73
7	User Interface for Description Values Suggestion	77
7.1	Requirements	77
7.2	Mockups	78
7.2.1	Record Creation Screen	78
7.2.2	Record View Screen	83
8	Conclusions and Future Work	84
8.1	Conclusions	84
8.1.1	Optimization of Image Processing Algorithms for Character Recognition in Cultural Heritage Typewritten Documents	84
8.1.2	Ontology-Based Information Extraction	85
8.1.3	Ontology Population	86
8.2	Further Work	86
	References	88
A	HAREM Corpora Information	96
B	Optimization of Image Processing Algorithms for OCR	99

C User Interface Mockups

110

List of Figures

1.1	Atomization of an ISAD(G) record in Linked Data [39].	3
1.2	From DigitArq to DigitArq+ [39].	4
2.1	Taxonomy of NER with deep learning [53].	18
2.2	Overview of main RiC-CM v0.2 entities.	22
2.3	ArchOnto architecture [47].	23
3.1	An overview of the execution pipeline.	27
3.2	An overview of the architecture of the proposed ontology information-extraction and ontology population system.	28
4.1	Digital representations from the typewritten dataset.	33
4.2	Overview of the different steps of the proposed methodology.	36
5.1	Overview of the different steps of the corpora transformation.	46
7.1	Individual view of the creation of a description record.	79
7.2	Individual view of a digital representation of a record's document uploaded to the UI.	80
7.3	Validation scenario: warning message.	81
7.4	Edit scenario: modified title field.	81
7.5	Edit scenario: edit a relation between two entities.	82
7.6	Edit scenario: edit a relation between two entities by associating the source entity with an existing entity.	82
7.7	Edit scenario: semantic network section with the corrected relations between entities.	83
B.1	Distribution of OCR performance measures in the different typologies of digital representations.	109
C.1	Individual view of a description record [39].	111
C.2	Individual view of the suggested archival description from a digital representation.	112
C.3	Individual view of the created record.	113

List of Tables

2.1	Comparison of four OCR software applications.	8
2.2	Description of binarization algorithms in the OpenCV library.	9
2.3	Description of binarization algorithms' parameters in the OpenCV library.	10
2.4	Description of image smoothing algorithms in the OpenCV library.	10
2.5	Description of image smoothing algorithms' parameters in the OpenCV library.	10
2.6	Description of morphological transformation algorithms in the OpenCV library.	11
2.7	Description of morphological transformation algorithms' parameters in the OpenCV library.	11
2.8	Comparison of text-based OCR performance measures.	12
2.9	Comparison of research papers related to image pre-processing in OCR or parameter tuning.	14
2.10	Comparison of research papers related to information extraction in the Portuguese language.	20
2.11	Comparison of state-of-the-art NER results evaluated on test set (MiniHAREM).	20
4.1	Characterization of the dataset's archival records and digital representations throughout the series' description level. The digital representations were split by writing format.	32
4.2	Characterization of the parameterization, evaluation, and population datasets' digital representations.	35
4.3	Average time to parameterize the image processing algorithms.	39
4.4	Image processing algorithms that improved the OCR performance per digital representation typology and overall and were selected as the most significant.	43
5.1	Description of HAREM tags.	46
5.2	Mapping between relevant concepts identified in the First HAREM corpus and ontology classes that differs from the mapping done with the Second HAREM corpus.	48
5.3	Meaning and purpose of HAREM relationships kept in the corpus.	49
5.4	Mapping between relevant concepts identified in the Second HAREM corpus and ontology classes.	50
5.5	Characterization of original and transformed HAREM corpora.	59
5.6	Distribution of recognized named entities in transformed HAREM corpora.	60
5.7	Distribution of relations between recognized named entities in transformed Second HAREM corpus.	60
5.8	Distribution of explicit relations in corpora annotated with relations between the entities Group, Person, Place, and Title.	60
5.9	Hyperparameter settings for named entity classification.	61

5.10	CoNLL evaluation results of BiLSTM-CRF using default Flair Embeddings and FlairBBP.	62
5.11	Evaluation results of BiLSTM-CRF using default Flair Embeddings and FlairBBP with entity-type matching.	63
5.12	Results of 5-folds cross validation for relation extraction evaluation in corpora with annotations of affiliation, authorship, and placement.	64
5.13	Results of relation extraction evaluation in Collovini et al. [18] corpora with annotations of affiliation and placement.	64
6.1	Mapping of extracted concepts to property assertions in ArchOnto.	68
6.2	Mapping of extracted relations between recognized named entities to property assertions in ArchOnto.	69
6.3	Numbering of dataset's archival record and corresponding Digitarq URL.	71
6.4	Performance results of ontology population evaluation on selective and total scenarios with OCR'ed dataset and their transcriptions.	74
A.1	Entity types in Second HAREM's GC split by categories, types, and subtypes. . .	96
A.2	Relation types in Second HAREM's GC [35].	98
B.1	Parameters set for the image processing algorithms tuned by NGSa-II for classified types of digital representations.	100
B.2	Performance comparison between image processing algorithms tuned with Wilcoxon pairwise comparisons using default, global and digital representation-type parameters.	102
B.3	Summary of Friedman test results and Wilcoxon pairwise comparisons for post-hoc analysis of OCR performance of image processing algorithms with the overall dataset.	103
B.4	Summary of Friedman test results and Wilcoxon pairwise comparisons for post-hoc analysis of OCR performance of image processing algorithms with the letters sample.	104
B.5	Summary of Friedman test results and Wilcoxon pairwise comparisons for post-hoc analysis of OCR performance of image processing algorithms with the non-structured reports sample.	105
B.6	Summary of Friedman test results and Wilcoxon pairwise comparisons for post-hoc analysis of OCR performance of image processing algorithms with the process covers sample.	106
B.7	Summary of Friedman test results and Wilcoxon pairwise comparisons for post-hoc analysis of OCR performance of image processing algorithms with the structured reports sample.	107
B.8	Summary of Friedman test results and Wilcoxon pairwise comparisons for post-hoc analysis of OCR performance of image processing algorithms with the theatre plays' covers sample.	108

Listings

5.1	Extracted temporal expressions from the First and Second HAREM GC that showcase the differences in named entity classification between the corpora.	46
5.2	Recognized named entity from the Second HAREM GC that showcases the annotation of relationships between entities.	47
5.3	Extracted sentences from the Second HAREM GC that showcase the use of the ownership relationship in different contexts.	51
5.4	Example of a vague named entity that has a different relation notation.	52
5.5	Resolved vague named entity with a textual excerpt that uses the example from Listing 5.4.	52
5.6	Excerpt of an ALT case with an entity tagged as an E54 Dimension. Textual expression reads: "In this time span of 15 years". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.	53
5.7	Excerpt of an ALT case with an entity of a person. Textual expression reads: "Catherine of Aragon". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.	54
5.8	Excerpt of an ALT case with a person's role in a group. Textual expression reads: "director of the National Historical Museum". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.	54
5.9	Excerpt of an ALT case with a person's religious role. Textual expression reads: "Rabbi of Antwerp Jacob Kruger". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.	55
5.10	Excerpt of an ALT case with an entity identified as a place. Textual expression reads: "Belém Cultural Center". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.	55
5.11	Excerpt of an ALT case with a temporal expression. Textual expression reads: "At twenty minutes past three in the afternoon of June 13, 1888". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.	56
5.12	Excerpt of an ALT case with the title of a document and its author. Textual expression reads: "Ninety-five Theses by Martin Luther". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.	56

5.13	Example of manual relation annotation of relation descriptors between named entities from the transformed Second HAREM corpus.	57
5.14	Output of the CoNLL format for the NER classification task using the example sentence.	58
5.15	Output of the corpus with POS tagging and dependency parsing with SpaCy for the RE classification task with the affiliation relation (vinc_inst).	59
6.1	Example of affiliation relationship extracted from a letter in the dataset.	72
6.2	Example of authorship relationship extracted from a structured report in the dataset.	72
6.3	Example of placement relationship extracted from a structured report in the dataset.	72
6.4	Matching instances based on data property content for record 1.	73

Abreviaturas e Símbolos

ANTT	Arquivo Nacional da Torre do Tombo
BDM	Balanced Distance Metric
BERT	Bidirectional Encoder Representations
BIO	Beginning, Inside and Outside
BIOES	Beginning, Inside, Outside, End and Single
Bi-LSTM	Bidirectional Long Short-Term Memory
CER	Character Error Rate
CIDOC	International Committee for Documentation
CNN	Convolutional Neural Network
CoNLL	Conference on Natural Language Learning
CRF	Conditional Random Field
CRM	Conceptual Reference Model
DGFP	Direcção Geral da Fazenda Pública
DGLAB	Direção-Geral do Livro, dos Arquivos e das Bibliotecas
DNN	Deep Neural Networks
EPISA	Entity and Property Inference for Semantic Archives
FCT	Foundation for Science and Technology
FlairBBP	Flair Embeddings Backwards and Forward
HMM	Hidden Markov Model
IberLEF	Iberian Languages Evaluation Forum
ICA	International Council for Archives
IE	Information Extraction
INESC TEC	Institute for Systems and Computer Engineering, Technology and Science
ISAD(G)	General International Standard for Archival Description
JSON	JavaScript Object Notation
LD	Linked Data
LSTM	Long Short-Term Memory
NER	Named Entity Recognition
NLP	Natural Language Processing
NSGA-II	Non-dominated Sorting Genetic Algorithm II
OBIE	Ontology-Based Information Extraction
OCR	Optical Character Recognition
OP	Ontology Population
OWL	Web Ontology Language
POS	Parts-of-speech
RDF	Resource Description Framework
RE	Relation Extraction
Ric-O	Records in Contexts-Ontology

RNN	Recurrent Neural Network
SNI	Secretariado Nacional de Informação
SVM	Support Vector Machine
TSV	Tab-Separated Values
UI	User Interface
WER	Word Error Rate
XML	Extensible Markup Language

Chapter 1

Introduction

The World Wide Web has revolutionized global access to information. While hyperlinks only connect web pages, semantic links can connect various kinds of resources. Linked Data (LD) is a concept created by Tim Berners-Lee to describe a web of data that behaves much in the same way as the World Wide Web is a network of hypertext documents connected by hypertext links [9]. Instead, data is connected through semantic links that can be expressed as Resource Description Framework (RDF) predicates. Semantic links are more meaningful than hypertext links because hyperlinks do not specify how two web pages are connected, while RDF triples show how a subject and an object are connected through a predicate [66].

Cultural heritage institutions have the role of protecting, preserving, and facilitating access to cultural knowledge. In the current digital era, cultural material can be digitally preserved and remotely accessed by the public. Cultural heritage institutions have been adopting Linked Data to improve archival descriptions and information discovery [66]. Both cultural heritage institutions and Linked Data share the purpose of describing and providing access to artifacts. Archivists have the task of describing objects to identify and handle them. They also provide historical context for sets of documents. Descriptions make cultural heritage material findable and more easily accessible to users [26]. Linked Data allows users to find extra information since data resources, while decentralized, can be aggregated in graphs through semantic links. The use of semantic links improves the descriptions' granularity and interoperability because each component in a record's description can be aggregated into an RDF graph where each node (subject and object) and edge (predicate) are individually accessible [66]. This means that records' metadata, artifacts, and concepts can be linked to improve data exploration [9].

Overall, the use of Linked Data not only brings advantages to archivists but also to users. Archivists can make more thorough descriptions and links to information from other Linked Data sources [62]. In turn, users will have better knowledge of the archives and each record with information aggregated from different institutions.

1.1 Context

This thesis emerged in the context of the EPISA project (Entity and Property Inference for Semantic Archives), funded by FCT (Foundation for Science and Technology) and proposed by INESC TEC (Institute for Systems and Computer Engineering, Technology and Science) in partnership with DGLAB (Direção-Geral do Livro, dos Arquivos e das Bibliotecas) and the University of Évora. It began in January 2019 with the intent of designing an open-source knowledge platform that represents archival data on a Linked Data model [46]. The project developed new description tools that are more fine-grained, flexible, and machine-actionable to account for the growing amount of information and the need for interoperability. EPISA also intended to migrate the current descriptions by extracting relevant entities and properties from the existing records.

Currently, contents of the Portuguese National Archives [46] are described using the ISAD(G) standard (General International Standard for Archival Description). The ISAD(G) [69] is an archival description standard with a multilevel structure. It uses descriptions from general to specific to represent the context and hierarchical structure of the fonds and their parts.

One of the main tasks of EPISA was the development of a Linked Data model, ArchOnto, which is an ontology-based model for archives based on the CIDOC-CRM (Conceptual Reference Model). It was created because some ISAD(G) information cannot be turned into CIDOC-CRM [46]. The CIDOC-CRM is a formal ontology developed by the International Committee for Documentation (CIDOC) to facilitate the integration, mediation, and exchange of heterogeneous sources of cultural heritage information.

Another main task was the migration from the ISAD(G) multilevel description to a Linked Data description [38], after the development of ArchOnto. The migration process from the ISAD(G) standard to LD descriptions can be seen in Figure 1.1.

A visual overview of the expected system, DigitArq+, is presented in the rich picture of Figure 1.2. It portrays the flow of information from the data sources, e.g., institutions and archives, through systems for migrations and interfacing to the end users [39]. There is data migration from the ISAD(G) standard to the CIDOC-CRM standard at the interface level, and at the infrastructure level, there is a migration from a relational database to a graph database.

1.2 Motivation

The task of describing cultural heritage objects can be taxing and time-consuming, making it challenging to describe documents or collections in finer detail. Furthermore, the migration to a more fine-grained description can exacerbate that difficulty. While the migration task solves this challenge for archival records that already have descriptions, records that have no description need to be described using ArchOnto. Our task will be useful in the creation of new descriptions of records that will be represented directly with ArchOnto.

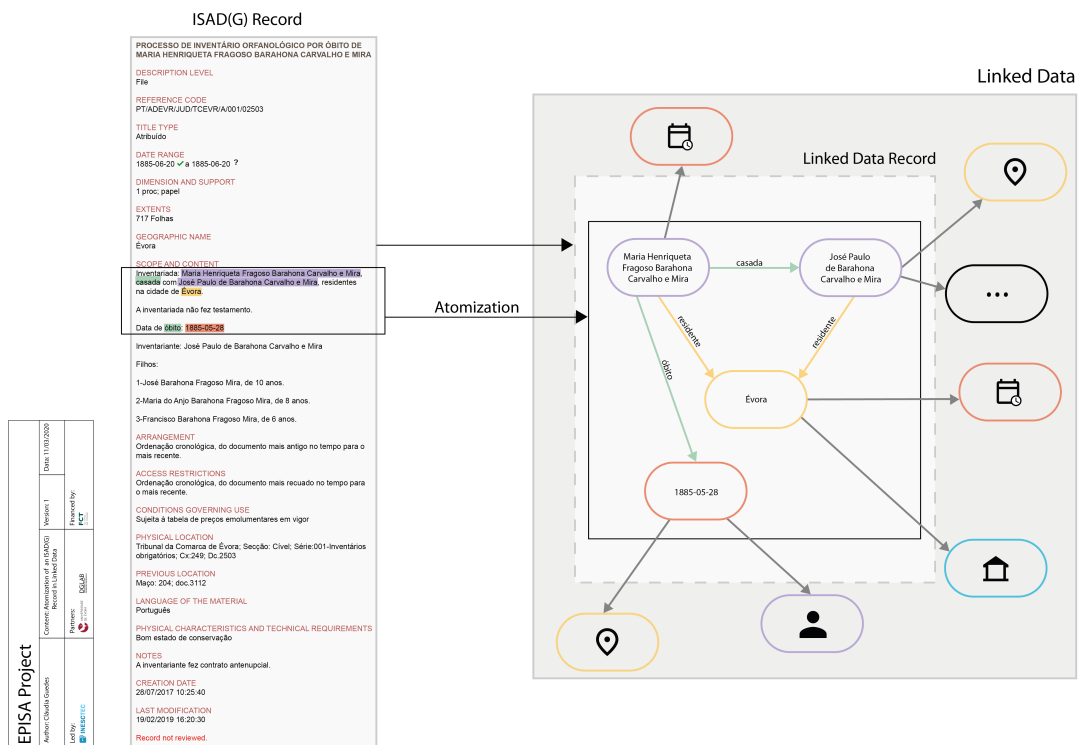


Figure 1.1: Atomization of an ISAD(G) record in Linked Data [39].

1.3 Objectives

It would greatly benefit the work of cultural heritage professionals to have a semi-automated description tool for the ArchOnto model. This tool would automatically propose instance values for the entities and properties in descriptions from digital representations of records' documents. The suggestions should be validated or modified by archival professionals through the EPISA platform. However, most of the available digital representations of archival records in the Portuguese Archives are in the form of scanned images. Because of this, we also need to convert records' digital representations from non-machine-readable into a machine-readable format.

To achieve the goal of implementing an ontology population tool to aid the curators' job of describing historical documents, we defined the following objectives:

1. Create a module that converts non-machine-readable digital representations of documents in an archival record to machine-readable ones;
2. Train information extraction models that extract entities and relations of the ArchOnto ontology;
3. Instantiate the extracted information in the ArchOnto ontology with an Ontology Population module.

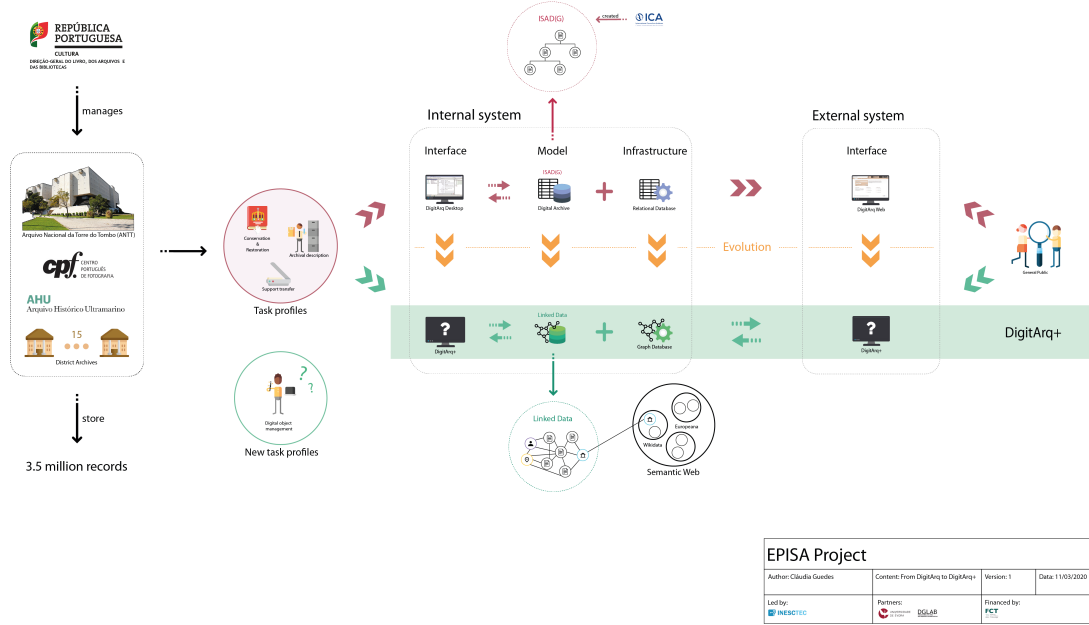


Figure 1.2: From DigitArq to DigitArq+ [39].

We also defined the following research objectives:

1. Analyze the best methodologies and tools to convert non-machine-readable digital representations to machine-readable ones in the context of historical documents;
2. Explore methodologies and existing tools to extract information from unstructured data in the Portuguese language and the cultural heritage domain;
3. Explore methodologies and existing tools to instantiate ontologies from structured data.

1.4 Contributions

This work contributed to the scientific progress in optimizing OCR with the parametrization of image processing algorithms, exploring available solutions in Portuguese to the problem of classifying entities and relations in textual contents, and developing an ontology population and evaluation approach. The main contributions are summarized as follows:

- Analysis of the impact of NSGA-II on the parameterization of image processing algorithms;
- Analysis of the impact of parameterization by digital representation typology;

- Analysis of the impact of image pre-processing on OCR considering global and digital representation typology;
- Proposal of an approach for the automatic population of the ArchOnto ontology from textual data that applies information extraction techniques to identify and classify ontology instances and properties;
- Proposal of a methodology to evaluate ontology population implemented in Python based on GATE's¹ Java implementation of BDM (Balanced Distance Metric) score and lexical evaluation.

Moreover, there are few resources in Portuguese in the scientific areas we are exploring. This work contributed to the progress of these areas for Portuguese with the publication of the following datasets to the INESC TEC research data repository:

- [23] Dias, M., & Teixeira Lopes, C. (2022). Typewritten Digital Representations of Portuguese Cultural Heritage Documents from the 20th century [Data set]. INESC TEC. <https://doi.org/10.25747/ZC25-1531>
- [30] Falcão, M., Dias, M., & Teixeira Lopes, C. (2022). Manual Transcriptions of Typewritten Digital Representations of Portuguese Cultural Heritage Documents from the 20th Century [Data set]. INESC TEC. <https://doi.org/10.25747/WPNA-JE39>
- [24] Dias, M., & Teixeira Lopes, C. (2023). Consensual ArchOnto representation of 13 Portuguese Historical Archival Records based on their Digital Representations [Data set]. INESC TEC. <https://doi.org/10.25747/EADP-M943>

We also published and submitted the following articles:

- [22] Mariana Dias and Carla Teixeira Lopes. Mining Typewritten Digital Representations to Support Archival Description. Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries – Workshops and Doctoral Consortium, ISBN: 1613-0073, pages 70-76. 2022.
- Mariana Dias and Carla Teixeira Lopes. Optimization of Image Processing Algorithms for Character Recognition in Cultural Typewritten Documents. J. Comput. Cult. Herit. [Under Major Revision]
- Mariana Dias and Carla Teixeira Lopes. Automatic Ontology Population from Cultural Heritage Typewritten Documents. Semantic Web. [Being prepared for submission]

¹<https://gate.ac.uk/>

1.5 Document Structure

This document is composed of eight chapters including the current introduction chapter, which contextualizes this work's theme. Chapter 2 explores the current state-of-the-art tools, methodologies, and related work for Optical Character Recognition, Ontological-Based Information Extraction, and Ontology Population. Chapter 3 describes the system's workflow and architecture. In Chapter 4, we analyze the impact of image processing algorithms and parameter tuning to optimize the OCR task. Chapter 5 describes our approach to training ontology-based information extraction models, and in Chapter 6, we explain our implementation of ontology population and evaluate the process using the previously trained information extraction models. Chapter 7 presents the mockups of our proposed solution of integrating an ontology-information extraction and population pipeline into the EPISA platform. In Chapter 8, we present an overview of the work done and what needs to be done in the future.

Chapter 2

Background and Related Work

In this chapter, we present the state-of-the-art related to the main processes addressed in this work: Optical Character Recognition (OCR), Ontology-Based Information Extraction (OBIE), and Ontology Population (OP).

In Section 2.1, we explore different OCR software, image processing algorithms to optimize the OCR performance, and evaluation measures. In Section 2.2, we analyze common approaches to the ontology-based information extraction task, while in Section 2.3 we detail the ontology population process and evaluation approaches.

2.1 Optical Character Recognition

Digitalization methodologies are increasingly being applied to allow remote access to archival content. As digitalization in cultural heritage contents usually results in non-machine-readable representations, an initial conversion to a machine-readable format is needed to fully explore the potential of these representations in Linked Data implementations.

The process of recognizing handwritten, typewritten, or printed text in digital images and translating them into machine-readable documents is called Optical Character Recognition (OCR). Historical documents present various challenges to the OCR process. Most engines require additional training to recognize historical fonts and languages. Additionally, the success of the OCR process depends on the quality of the scanned image, and it is common for heritage documents to suffer some degree of degradation over time. Digital representations of historical printed records are often noisy because of the papers' degradation due to aging, such as smears and faded text, and other printing noise like varying kerning and leading, i.e., different spaces between letters and line-break hyphenation amongst others [40]. Image processing methods can be applied before the text recognition phase to improve image quality and overall text extraction. Various parameters

can tune image processing algorithms. Using default values for parameters may yield bad performance results in comparison to the optimized combination of parameters. Therefore, the task of assigning values to parameters is essential.

2.1.1 OCR Software

There are currently four prominent OCR software applications: ABBYY FineReader, Google Cloud Vision API, Ocropy, and Tesseract. Both Tesseract and Ocropy are open-source engines, while ABBYY FineReader and Google Cloud Vision API are commercial engines. In terms of approaches, all engines use Deep Neural Networks (DNN) to train their models, except for Ocropy which uses shallow networks. A DNN can improve accuracy by adding training data [29], unachievable with shallow networks.

Furthermore, while Ocropy is a language-independent engine, ABBYY FineReader, Google Cloud Vision API, and Tesseract support dictionaries and language modeling. It is not easy to give a general definitive comparison of the different engines based on accuracy because the characteristics of different datasets influence the results. Engines are also affected by languages and font types. Moreover, recent versions of some engines have additional or updated training models and data that improve the performance and invalidate past research conclusions over which software is the best. In Table 2.1, a comparison between the different highlighted technologies is available. One of the comparison criteria is whether the OCR software supports the Portuguese language because this work uses a dataset of Portuguese cultural heritage documents.

Table 2.1: Comparison of four OCR software applications.

OCR Software	Approach	Platforms	License	Languages
ABBYY FineReader	Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) [70]	Mac, Windows, Online	Proprietary	Supports Portuguese with dictionary support [71]
Google Cloud Vision API	CNN-based with language models [44]	Online	Proprietary	Supports Portuguese through Latin scripts [44]
Ocropy	One-dimensional Long Short-Term Memory (LSTM) without language models [29]	Linux	Apache	Supports languages that use Latin scripts [5]
Tesseract	LSTM-based engine with language models	Mac, Windows, Linux	Apache	Portuguese language model with dictionary package [96]

2.1.2 Image Processing

Image processing techniques aim to visually enhance images and make them more suitable for human or machine interpretation. The strategies highlighted in this thesis are image binarization and noise reduction.

Binarization consists of transforming colored pixels into black-and-white ones. Most recent implementations of binarization techniques are based on thresholding [90]. Thresholding is a segmentation method that sets pixels to black or white depending on whether the grey value is above or below a threshold value to evidence contrast between the background and foreground.

There are various noise reduction techniques, such as filtering and morphological transformations. Filters can be grouped into smoothing and sharpening. Morphological transformations are used to thin characters, connect broken strokes, and smooth contours [50].

OpenCV is an open-source computer vision Python library launched in 1999 by Intel [61] with various image processing algorithms, specifically for binarization, image smoothing, and morphological transformations. The library offers four binarization operations: adaptive threshold, Otsu threshold, simple threshold, and triangle threshold (see Tables 2.2 and 2.3); four different smoothing operations: bilateral filter, Gaussian filter, homogeneous blur, and median blur (see Tables 2.4 and 2.5); seven different morphological transformation operations: black hat, closing, dilation, erosion, morphological gradient, opening, and top hat (see Tables 2.6 and 2.7). Tables 2.2, 2.4, and 2.6 describe the meaning of each image processing algorithm from the OpenCV library. Tables 2.3, 2.5, and 2.7 describe the meaning and typology of each image processing algorithm's parameters.

Table 2.2: Description of binarization algorithms in the OpenCV library.

Algorithm	Description
Adaptive thresholding	Calculates the threshold value within the pixels' neighborhood area
Otsu thresholding	Automatically selects a threshold value based on interclass variance maximization
Simple thresholding	Applies the same global threshold value to every pixel
Triangle thresholding	Histogram-based algorithm that automatically selects a threshold value based on interclass variance maximization

2.1.3 Parameter Optimization

To select appropriate values for the algorithms' parameters, we can use default values or manual configurations based on literature, experience, or trial-and-error [78]. OpenCV's documentation presents some examples of implementations using recommended assigned values. For instance, in thresholding algorithms, 255 is the typical value assigned to the maxVal parameter. However, these configurations do not guarantee an optimal parameterization of the algorithms as the parameters' values should be adjusted to the images' state. As an alternative, parameter tuning

Table 2.3: Description of binarization algorithms' parameters in the OpenCV library.

Parameter Name	Parameter Type	Meaning
adaptiveMethod	nominal	Adaptive thresholding method. Takes the values: mean (0); and gaussian (1)
blockSize	discrete	Size of a pixel's neighborhood area
constant	continuous	Arbitrary constant
maxValue	continuous	Maximum value assigned to pixels that exceed the threshold value
thresh	continuous	Threshold value
thresholdType	nominal	Thresholding type. Takes the values: threshold binary (0); and threshold binary inverted (1)
type	nominal	Threshold type. Takes the values: threshold binary (0); threshold binary inverted (1); truncate (2); threshold to zero (3); and inverted (4)

Table 2.4: Description of image smoothing algorithms in the OpenCV library.

Algorithm	Description
Bilateral filter	Uses Gaussian filter, but its weight has one more component: the difference in intensity
Gaussian blur	Blurs an image using Gaussian filter
Homogeneous blur	Blurs an image using normalized box filter
Median blur	Blurs an image by replacing each pixel with a median of its neighborhood pixels

Table 2.5: Description of image smoothing algorithms' parameters in the OpenCV library.

Parameter Name	Parameter Type	Meaning
anchor	discrete	Anchor point
borderType	nominal	Border type. Takes the values: constant (0); replicate (1); reflect (2); default (3); and isolated (4)
d	discrete	Diameter of each pixel neighborhood
ksize	discrete	Aperture linear/Gaussian kernel size
sigmaColor	continuous	Filter sigma in the colour space
sigmaSpace	continuous	Filter sigma in the coordinate space
sigmaX	discrete	Gaussian kernel standard deviation in X direction
sigmaY	discrete	Gaussian kernel standard deviation in Y direction

can be done using a combinatorial search approach. According to the literature, the genetic and Hill-Climbing algorithms are the two best-tailored approaches to handle various parameters and their values [49].

Table 2.6: Description of morphological transformation algorithms in the OpenCV library.

Algorithm	Description
Black hat	Difference between input image and closing of the image
Closing	Dilation followed by erosion
Dilation	Enlarges boundaries of regions of foreground pixels
Erosion	Erodes boundaries of regions of foreground pixels
Morphological gradient	Difference between input image and opening of the image
Opening	Erosion followed by dilation
Top hat	Difference between input image and opening of the image

Table 2.7: Description of morphological transformation algorithms' parameters in the OpenCV library.

Parameter Name	Parameter Type	Meaning
anchor	discrete	Anchor point
borderType	nominal	Border type. Takes the values: constant (0); replicate (1); reflect (2); default (3); and isolated (4)
iterations	discrete	Number of times technique is applied
kernel	continuous	Structuring element that determines the shape of a pixel neighborhood

A genetic algorithm is a randomized search algorithm inspired by Charles Darwin's theory of natural selection [99]. It has the premise that individuals pass traits to their offspring, and the individuals with the best characteristics, tend to survive and have more offspring [100]. Reproduction is the process in which the fitness function, i.e., the objective function, dictates the probability of having offspring for the next generation [80]. In a minimization problem, lower fitness values mean a higher likelihood of offspring. Crossover is a process in which offspring are generated by combining traits from two parents, randomly selected members of the last population. Mutations are used to ensure a generation's diversity onto the next one. The hill-climbing method uses neighbors in every iteration to evaluate a current solution candidate [74]. It greedily moves towards the neighbor with the highest value until there is no better neighbor. Only the genetic algorithm can manipulate nominal parameter types in a meaningful way from the two algorithms. In contrast, the hill-climbing algorithm requires a notion of neighborhood.

As some of the OpenCV image processing algorithms have nominal parameters, we use genetic algorithms. Genetic algorithms have many advantages in dealing with complex problems, and parallelism as multiple offspring that act as independent agents are ideal for parallel exploration of the search space. The main disadvantage is the difficulty in formulating the fitness function and parameters, such as the population size, crossover, and mutation rates [99].

2.1.4 Evaluation of OCR systems

There are two ways to evaluate OCR systems: via text or layout. A text-based evaluation is based on the comparison of the OCR text result against the ground truth. Recent research has four prominent text-based measures: Character Error Rate (CER), Word Error Rate (WER), Index/Count-based Bag of Words, and Flex Character Accuracy. In Table 2.8, a comparison between the different measures is available. Traditionally, text-based metrics use the edit distance between the ground truth and the OCR output. That is the case for the indicators CER and WER which are also dependent on reading order. Reading order refers to a reading route for textual elements that can be sequentially ordered lists for simple layouts, i.e., a book page, or more complex in the case of newspapers where there isn't necessarily a left-to-right order [14]. Its reading dependence makes them unreliable when applied to complex page layouts since the performance is influenced by page segmentation and text detection [14].

Table 2.8: Comparison of text-based OCR performance measures.

Measure	Description	Formula
Character Error Rate (CER) [77]	Minimum number of character edition operations (insertions, deletions, and substitutions) needed to obtain the ground truth based on Levenshtein distance.	$\frac{C_{\text{insertions}} + C_{\text{deletions}} + C_{\text{substitutions}}}{\text{Number of characters in text}}$
Word Error Rate (WER) [77]	Minimum number of word edition operations (insertions, deletions, and substitutions) needed to obtain the ground truth based on Levenshtein distance.	$\frac{W_{\text{insertions}} + W_{\text{deletions}} + W_{\text{substitutions}}}{\text{Number of words in text}}$
Index-based Bag of Words [4]	Uses the number of words recognized at least once and does not account for every occurrence.	$\frac{\text{Number of correct words}}{\text{Number of words in text}}$
Count-based Bag of Words [4]	Considers the correct count of the number of occurrences of recognized words.	$\frac{\text{Number of words with correct occurrences}}{\text{Number of words in text}}$
Flex Character Accuracy [15]	Based on CER but adapted to not depend on reading order. Edit operations are determined by text alignment and line matching, where non-matched lines are insertions or deletions.	$\frac{L_{\text{insertions}} + L_{\text{deletions}}}{\text{Number of chunks in text}}$

In contrast, a layout-based evaluation is based on the document page's structure. Measures analyze the document's segmentation process in different regions, classification of the areas, and reading order. The most popular ones are segmentation error, misclassification and reading order error.

2.1.5 Works exploring image pre-processing to improve OCR

A popular way to improve the OCR results in historical documents is to apply an image pre-processing step to the OCR task. In the literature, various approaches combine image process-

ing algorithms. Harraj et al. [42] applied illumination adjustment, grayscale conversion, unsharp masking, and Otsu binarization to counteract image distortions and improve OCR accuracy. Similarly, Ntogas et al. [68] trained a model for Finnish Fraktur fonts and analyzed different combinations of pre-processing image algorithms. Koistinen et al. [48] show a process for image binarization with manual image condition classification and image processing methods. The approach has five stages: image acquisition, image preparation, filtering methods, threshold methods, and image refinement. Bui et al. [11] use a Convolution Neural Network (CNN) to automatically select image processing algorithms according to each input document image. Most of the previously mentioned works have historical documents as their dataset. Ntogas et al. [68] used document images from the Holy Monastery of Dousiko, Meteora, Greece, while Koistinen et al. [48] used Finnish historical documents.

According to the overall results of the previously mentioned research, the use of image processing algorithms as a pre-processing step to the OCR process improves the accuracy of the engines. Koistinen et al. [48] improved on word level of 27.5% compared to FineReader 7 or 8 and 9.2% for FineReader 11. Harraj et al. [42] registered an improvement of OCR accuracy of between 2% to 6.8%. [11] succeeded in proving the performance of engines with their automatic pre-processing method selection, with Tesseract OCR having an improvement of 25%.

As previously discussed in Section 2.1.3, there are various approaches to the parameter tuning of image processing algorithms. Lund et al. [56] selected seven values for thresholding from the range of 0 to 255. There is an equal incremental difference between the values and an equal division of the grayscale spectrum. This experimental method reduced the error rate to 39.1% in comparison to the original performance of the engine. To our knowledge, there is no literature on combinatorial search algorithms applied to tuning parameters of image processing algorithms in OCR.

In contrast, there are some examples of implementations of machine learning algorithms to optimize OCR accuracy. For instance, Sporici et al. [86] developed an adaptive image pre-processing method using a reinforcement learning model to minimize the edit distance between recognized text and ground truth. The technique improved the F1-score of Tesseract 4.0 from 16.3% to 72.9%.

Table 2.9 presents the comparison of the different approaches, algorithms, and results for the discussed articles.

Table 2.9: Comparison of research papers related to image pre-processing in OCR or parameter tuning.

Article	Methodology	Image processing algorithms	Evaluation dataset	Results
[68]	Analyzed the performance of a pipeline of image processing algorithms according to a manual image condition classification	Filtering algorithms: Mean, Median, Wiener, Bullerworth, and Gaussian	Documents from the Holy Monastery of Dousiko, Meteora, Greece	Binarization algorithms Sauvola and Niblack had the best image quality
		Binarization algorithms: Bernsen, Niblack, Otsu, and Sauvola		
[56]	Parameter tuning of thresholding image processing methods by selecting parameter values with basis on an even distribution in a range	Global thresholding, Otsu, and Sauvola	Collection of 19th Century Mormon Article	Reduction in error rate of 39.1% in comparison to the original performance of the OCR engine
[42]	Pipeline of image pre-processing methods	Illumination adjustment, grayscale conversion, unsharp masking, and Otsu binarization	Standard dataset	OCR accuracy improved between 2% to 6.8%
[48]	Analyze the performance of the combination of different thresholding algorithms	Otsu thresholding, Gaussian blur, Sauvola, WolfJolion, Contrast Limited Adaptive Histogram Equalized (CLAHE), and Linear Normalization	Historical Finnish documents	Combination of four methods (Linear Normalization + WolfJolion, CLAHE + WolfJolion, original image and WolfJolion) had the best results with an improvement of word error of 1.91% in comparison to FineReader 11, and 7.21% in comparison to FineReader 7 and 8

Table 2.9 continued from previous page

Article	Methodology	Image processing algorithms	Evaluation dataset	Results
[11]	Automatic selection of image pre-processing algorithms with a Convolution Neural Network (CNN)	Otsu binarization, Su et al. binarization, Gaussian filtering, non-local means of noise reduction, and Singh image sharpening	Documents of supermarket receipts captured by mobile devices	Selection method improved performance of Tesseract by 25%. The neural network had a performance from 93% to 96%
	Reinforcement learning model			
	that minimizes edit distance for an adaptive image pre-processing method	Adaptive thresholding	Brmo Mobile OCR Dataset [45]	F1-score of Tesseract 4.0 improved from between 0.163 to 0.729

2.2 Ontology-Based Information Extraction

Ontology-Based Information Extraction systems use existent information extraction techniques with an ontology as a guide to the extraction process [97]. For domain-specific ontologies, defining particular concepts to extract is crucial to the process. For example, in the cultural heritage domain, descriptions are event-driven, which signifies that it is vital to extract concepts like events.

OBIE systems use ontologies to define entities and relations to extract as classes, properties, and instances. Named Entity Recognition (NER) and Relation Extraction (RE) are subtasks of Information Extraction that extract entities and relationships between entities. There are three main extraction techniques: linguistic, machine learning, and deep learning approaches.

2.2.1 Linguistic rules and dictionary

Linguistic approaches can either be rule-based, where linguistic patterns identify entities and relations, or dictionary-based, where nouns are matched to concepts from a gazetteer.

Most pipelines of linguistic approaches include tokenization, parts-of-speech (POS) tagging, and text pre-processing. Tokenization is the process of splitting words in a document. The identified tokens are grammatically categorized into POS tags, e.g., noun (N), verb (V), preposition (P), or adjective (A). The most common forms of text pre-processing are stemming, lemmatization, and stop word removal. Stemming is a process that reduces inflectional forms of a word to its stem, while lemmatization reduces derivationally related forms of a word to a common base form known as the lemma using vocabulary and morphologic analysis of words [57]. Stop words are common words that are excluded from the vocabulary seeing that the intended focus is on essential words and keywords, and their removal improves performance.

2.2.1.1 Related Work

Relation Extraction approaches for the Portuguese language often apply linguistic rules [17]. Chaves [13] developed an ontology enrichment system, SEI-Geo, that only extracts the Location entity and its relations. The system uses patterns based on Geo-ontologies and the translated Hearst pattern in Portuguese to identify entities and relations.

Ferreira et al. [31] were the only participants of IberLEF on the task Relation Extraction for Named Entities. They employed a set of rules based on the sequence of Parts-of-Speech tags between named entities bootstrapped from examples provided by the organization. The approach achieved an F1-score of 72%. However, the authors expressed their doubts about the meaningfulness and utility of the relations extracted, as most relations are vague, e.g., "de" (of) or "em" (at).

Gonçalves et al. [36] evaluated available free tools that perform POS tagging in the Portuguese language. They considered the following NLP libraries: FreeLing, NLTK, OpenNLP, NLPyPort, SpaCy, Polyglot, Stanford NLP, TreeTagger, and LinguaKit. They also performed a qualitative study on dependency parsers from SpaCy and Stanford NLP.

2.2.2 Statistical Machine Learning

The most popular machine learning methods use supervised learning. Supervised learning depends on annotated data to train a model from machine learning algorithms that gives predictions from unseen data. Supervised learning algorithms in IE include Hidden Markov Models (HMM), Conditional Random Fields (CRF), and Support Vector Machines (SVM).

Hidden Markov models (HMMs) are statistical sequence models. They use Markov chains with transition probabilities between states where the next state is independent of the past, but state-to-state transitions are guided by joint transition and emission probability distributions [91]. The model is applied to the information extraction context to perform POS tagging and NER where labeling a word does not depend only on the word itself but also on the sequence of labels that preceded it [92]. However, HMMs struggle to model overlapping, non-independent features. For example, in NER an HMM is only dependent on the feature of the word's identity, but often other interdependent features, e.g., capitalization, suffixes, neighboring words, etc, carry important information that the generative nature of HMMs cannot easily model [89].

In contrast, Conditional Random Fields (CRF) are discriminative models that can model overlapping, non-independent features [89]. Whereas HMM defines a joint probability, linear-chain CRF defines a conditional probability. Defining and selecting relevant features is essential to CRF modeling in order to take advantage of its ability to incorporate many non-independent features.

2.2.2.1 Related Work

Milidiú et al. [60] proposed seven machine learning approaches to NER that use HMM and SVM for the Portuguese language. The SVM-based approach had a better performance with an F1-score of 88.11%, but the HMM-based approach has the advantage of having good enough results while not needing expert knowledge.

Pires et al. [76] assessed the performance of well-established NER tools such as Stanford CoreNLP, OpenNLP, SpaCy, and NLTK, using the Portuguese dataset HAREM for the evaluation. The authors performed 10-fold cross-validation on the corpus to evaluate the models, which prevents the direct comparison of the results obtained with other approaches. However, the authors concluded that results achieved with Stanford CoreNLP matched the best results of Portuguese state-of-the-art tools presented in HAREM conferences at the time with an F1-score of 56.1% on the total scenario.

Gonçalves et al. [36] evaluated available free tools that perform NER in the Portuguese language. The authors used the models trained on the SIGARRANEWS corpus by Pires et al. [76] to evaluate the NLTK, OpenNLP, and StanfordNLP tools. They trained three models using NLTK: Decision Tree, Naive Bayes, and Maximum Entropy.

Among RE systems in Portuguese, Collovini et al. [18] evaluated a CRF model for open-relation extraction between named entities Organization-Person and Organization-Place, as well as with pre-defined relation types between the same entity pairs. The authors used subsets of the HAREM's GC as a corpus. In comparison to two other relation extraction systems that used the

Second HAREM’s corpus, the authors achieved a better result in the placement relation with an F1-score of 57%, an increase of 26% from the state-of-the-art with the Second HAREM task.

2.2.3 Deep Learning

Deep learning approaches are more commonly used in recent research as they have been advancing the state-of-the-art in both NER and RE [65]. Methods include Conventional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and variants, such as Bi-LSTM (bidirectional LSTM). Most recent state-of-the-art information extraction approaches combine neural networks with word embeddings to improve text analysis.

The architecture of a typical deep learning-based system formulated as a sequence labeling task for NER is shown in Figure 2.1 [53]. The distributed representations for the input phase typically use word and character-level embeddings to map tokens from source sentences into dense vector representations that are read by the encoder and encoded into a contextual representation [52]. The tag decoder produces a translated sentence from the contextual representation by predicting tags for the tokens.

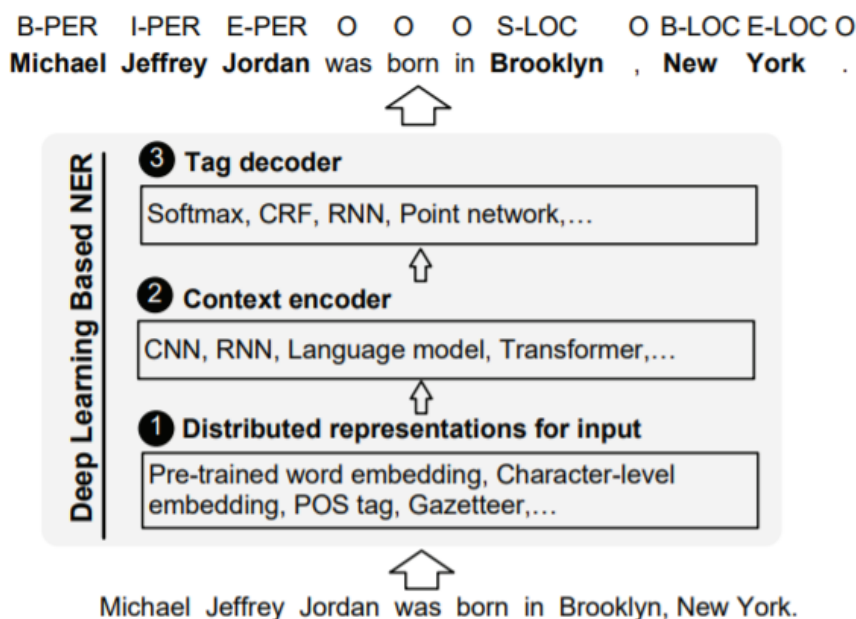


Figure 2.1: Taxonomy of NER with deep learning [53].

Word embeddings are vectors of representations of words that capture semantic meaning [63]. The contextual similarity between words is expressed by closer proximity in the vector space. Common static word embedding models are Word2Vec [59], fastText [7] and GloVe [72].

Akbik et al. [3] created a new type of embeddings called contextual string embeddings, Flair embeddings. The embeddings are contextualized by the surrounding text such that the same word can have different embeddings depending on its contextual use. It is a character-based language model with two independently trained models, a forward and backward model.

Devlin et al. [21] created BERT (Bidirectional Encoder Representations), contextual word embeddings that know the previous and succeeding context at the same time [10]. In contrast, Flair embeddings only know the previous context. The character-based nature of Flair embeddings has an advantage in capturing morphological and semantic structures, while BERT embeddings have the advantage of learning associations between both tokens and sentences.

2.2.3.1 Related Work

We provide an overview of works that used different methodologies to achieve state-of-the-art results on Portuguese NER with the golden HAREM collection. The authors of the works we will describe evaluated the systems using the MiniHAREM corpus as a test set with the CoNLL 2003 evaluation script¹. MiniHAREM was an intermediate event that repeated the First HAREM evaluation.

Santos et al. [28] extended the neural network architecture for sequential classification by applying character-level embeddings on Portuguese corpora. The authors used a language model [81] trained on a corpus that contains texts from the Portuguese Wikipedia, the CETENFolha², and the CETEMPUBLICO³ corpora. Castro et al. [12] implemented a Bidirectional Long Short-Term Memory with Conditional Random Fields (BiLSTM-CRF) approach on a Portuguese corpus. Santos et al. [83] proposed a BiLSTM-CRF model with contextual embeddings. The language model, FlairBBP, consists of a forward model, Flairs Forward⁴, and a backward model, Flair Backwards⁵. It was trained on three contemporary Portuguese corpora: BlogSet-BR⁶, which consists of texts from Brazilian Portuguese web blogs, brWaC⁷ [95], and ptwiki-20190301⁸ that contains texts extracted from Wikipedia. Souza et al. [85] trained BERT (Bidirectional Encoder Representations from Transformers) models and applied a BERT-CRF architecture to the NER task. The BERT models, BerTimbau⁹, were trained on the brWaC corpus and were made available [84]. Table 2.10 shows the comparison of the different approaches, resources, and results for the discussed articles.

Table 2.11 presents the results of state-of-the-art approaches over the years in two evaluation scenarios: total and selective. The total scenario considers all ten categories from the First HAREM, while the selective scenario focuses on five categories: Organization, Location, Time, Person, and Value.

Santos et al. [28] outperformed the previous state-of-the-art by 7.9% in the F1-score on the total scenario, and by 7.2% in the F1-score on the selective scenario. Castro et al. [12] improved

¹<https://www.cnts.ua.ac.be/conll2000/chunking/conllev.txt>

²<http://www.linguatca.pt/cetenfolha/>

³<http://www.linguatca.pt/cetempublico/>

⁴https://drive.google.com/file/d/1UXri84BUH3p_DD0j71GEls6m8qM5v4Pw/view?usp=sharing

⁵https://drive.google.com/file/d/11X-STlWBqvdp1k_Q2VMs31XRJ0PoQOrf/view?usp=sharing

⁶<https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/blogset-br/>

⁷https://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC#Current_version

⁸<https://dumps.wikimedia.org/ptwiki/20190301/>

⁹<https://github.com/neuralmind-ai/portuguese-bert>

Table 2.10: Comparison of research papers related to information extraction in the Portuguese language.

Approach	Language Model	Corpora for LM
CharWNN [28]	Character-level embeddings	Portuguese Wikipedia, CETENFolha, CETEMPUBLICO
BiLSTM-CRF [12]	-	-
BiLSTM-CRF+FlairBBP [83]	Flair embeddings	BlogSet-BR, brWac, ptwiki-20190301
BERTimbau Large+CRF [85]	BERT embeddings	brWaC

Table 2.11: Comparison of state-of-the-art NER results evaluated on test set (MiniHAREM).

Approach	Total Scenario			Selective Scenario		
	Precision	Recall	F1-score	Precision	Recall	F1-score
CharWNN [28]	67.2%	63.7%	65.4%	74.0%	68.7%	71.2%
BiLSTM-CRF [12]	72.8%	68.0%	70.3%	78.3%	74.4%	76.3%
BiLSTM-CRF+FlairBBP [83]	74.9%	74.4%	74.6%	83.4%	81.2%	82.3%
BERTimbau Large+CRF [85]	79.6%	77.4%	78.5%	84.9%	82.5%	83.7%

the NER F1-score by 5% in the total scenario. Santos et al. [83] outperformed the Portuguese state-of-the-art by 6% on the selective scenario, and 4.3% on the total scenario. They also participated in IberLEF2019, achieving an F1-score of 88.15% in the police dataset, 28.81% in the clinic dataset, and 66.66% in the general dataset [19]. They outperformed the other participants in the results for the general dataset. Finally, Souza et al. [84] improved the F1-score by 1% on the selective scenario, and by 4% on the total scenario.

2.2.4 Annotated Datasets

The standard datasets used to evaluate NER models in Portuguese are the HAREM golden collections¹⁰. HAREM is a Portuguese NER evaluation contest with two golden collections. The First HAREM collection [82] includes 1,202 documents with 5,132 named entities. It has two levels of named entity annotations: 10 categories and 41 types. In addition, the most recent golden collection (GC), the Second HAREM [35], contains both manually annotated named entities and semantic relations. The collection includes 1,040 documents with 7,847 named entities and 4,803 relations. It has three levels of named entity annotations: 10 categories, 44 types, and 21 subtypes. The named entity categories are the following: Person (PESSOA), Location (LOCAL), Organization (ORGANIZACAO), Time (TEMPO), Value (VALOR), Abstraction (ABSTRACCAO), Event (EVENT), Thing (COISA), Work (OBRA), and Other (OUTRO). Table A.1 shows the annotations per category, type, and subtype of the Second HAREM corpus. Furthermore, it contains 38 relation labels: twelve are symmetric relations, and the others are direct/indirect relations.

¹⁰https://www.linguatca.pt/aval_conjunta/HAREM/harem_ing.html

Table A.2 details the relation types in the GC. The relation types that the participants of the contest had to annotate are the following: identity (ident), inclusion (inclui/incluido), and location (ocorre_em/sede_de).

The IberLEF (Iberian Languages Evaluation Forum) is an evaluation forum of natural language processing on the Iberian languages Spanish, Portuguese, Catalan, Basque, and Galician. IberLEF 2019¹¹ had a Portuguese Named Entity Recognition and Relation Extraction task [19]. The tasks were performed on five new datasets. There are three NER datasets: a general dataset with 10,004 named entities composed of a portion of the Second HAREM corpus with entities of the Value category, and the SIGARRA¹² corpus [75] with the categories Person, Place, Organization, and Time; a clinic dataset with the category Person that contains 77 named entities; and a police dataset with the annotation Person that includes 916 annotations. The clinical and police datasets were not published to the public due to the sensitive nature of official police documents and clinical notes. The RE task used a dataset composed of a subset of the golden collections of the two HAREM conferences manually annotated with relations between entities by the coordinators, and the Summ-it++ [32] corpus annotated with semantic relations between named entities using crowdsourcing [20]. The resultant dataset has 752 instances of relations between named entities.

2.2.5 Related Work in Cultural Heritage

There are some implementations of information extraction approaches in the context of Portuguese cultural heritage. For instance, Vieira et al. [93] automatically annotated a cultural dataset with the entities Person, Location, and Organization from an existing dataset of transcribed handwritten documents¹³ extracted from the National Archives of Torre do Tombo, Parish Memories-Alentejo (1758-1761)¹⁴. Two researchers manually annotated a portion of the dataset for an evaluation set, while the rest was annotated automatically with a NER model trained on neural networks. The authors performed a comparative analysis of the NER performance of a CNN using the spaCy framework and a BiLSTM-CRF with contextual string embeddings trained by Santos et al. [83] using the Flair framework. The BiLSTM-CRF model was the best-performing model, with an F1-score of 46%.

Zilio et al. [101] automatically annotated a cultural dataset of Portuguese 18th century medical texts. The authors compared three off-the-shelf NER models: a spaCy model pre-trained with a small language model (spaCy sm), a spaCy model pre-trained with a large language model (spaCy lg), and a BERT-CRF model [85]. The spaCy models were trained on WikiNER annotation, while the BERT-CRF model was trained on the HAREM corpus. The models were evaluated on a gold-standard sample of three texts from the time period that were transcribed with modern spelling and manually annotated with entity labels. The BERT-CRF model had a better performance than the spaCy models in the modernized adaptation of the spelling of the texts and on the original texts.

¹¹<https://www.inf.pucrs.br/linatural/wordpress/iberlef-2019/>

¹²<https://rdm.inesctec.pt/dataset/cs-2017-004>

¹³<http://www.cidehusdigital.uevora.pt/portugal1758/memorias>

¹⁴<https://digitarq.arquivos.pt/details?id=4238720>

2.3 Ontology Population

Ontologies are a way to represent machine-readable semantic knowledge. In information systems, they can be defined as "an explicit account or representation of a conceptualization" [87]. The main ontology components are concepts, taxonomic relations, and non-taxonomic relations.

Ontology population consists of adding instances of classes and properties to those defined by an ontology [58]. It is a fundamental part of OBIE systems that use ontologies to guide the information extraction process and store its output.

2.3.1 Ontologies for Cultural Heritage

Several international standards define how to describe cultural heritage objects. We are going to highlight two ontologies focused on expressing archival records: RiC-O¹⁵, and ArchOnto¹⁶.

2.3.1.1 RiC-O

RiC-O (Records in Contexts-Ontology) is an ontology developed to describe archival record resources by the International Council of Archives (ICA). It is an international standard that replaces current ICA description standards, such as ISAD(G). It does not incorporate any components from existing ontologies. The architecture is presented in Figure 2.2.

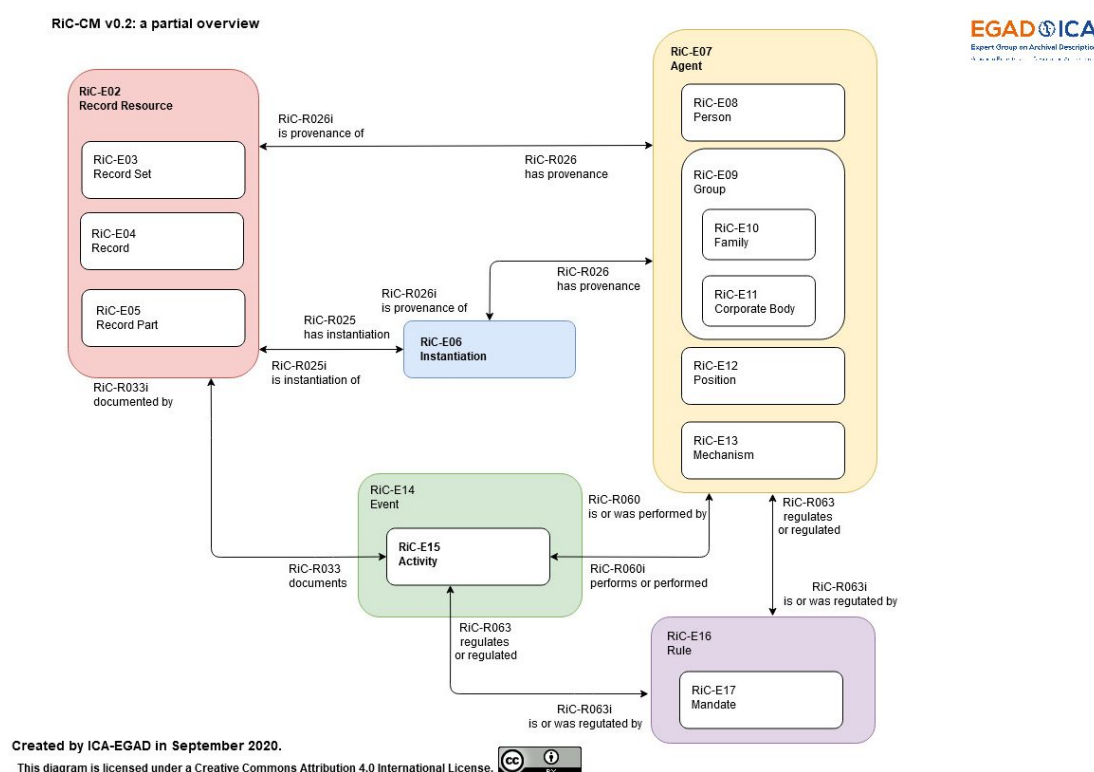


Figure 2.2: Overview of main RiC-CM v0.2 entities¹⁷.

¹⁵https://www.ica.org/standards/RiC/RiC-O_v0-2.html

¹⁶<https://purl.archive.org/episa/archonto>

2.3.1.2 ArchOnto

On the other hand, ArchOnto is an extension of CIDOC-CRM for archives developed by the EPISA project. CIDOC-CRM (CIDOC Conceptual Reference Model)¹⁸ is an ontology developed for museums by the International Committee for Documentation of the International Council of Museums (CIDOC). With CIDOC-CRM as a core ontology, ArchOnto incorporates other four ontologies as satellite ontologies: N-ary to represent non-binary relationships proposed by CIDOC, DataObject for data validation and data properties that CRM does not provide, ISAD to include textual information from ISAD(G), and Link2DataObject to link DataObject and CIDOC-CRM. ArchOnto also includes extensions to represent archival objects further. The architecture is presented in Figure 2.3.

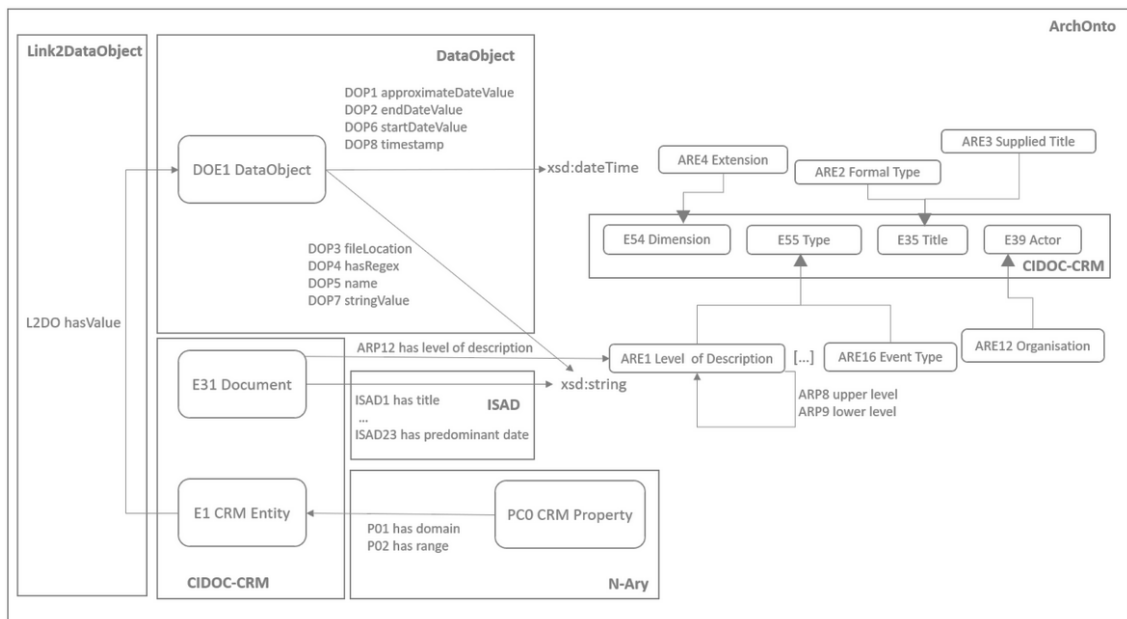


Figure 2.3: ArchOnto architecture [47].

2.3.2 Ontology Programming Interfaces

There are three ways of accessing an ontology in a programming language: with a query language, such as SPARQL [79], with an Application Programming Interface (API), such as OWL API [43] and Jena3 in Java, or with an ontology oriented programming language, such as Owlready [51] in Python.

Generally, object-oriented approaches in ontology oriented programming are more convenient for manipulating elements of an ontology, considering that query language approaches require writing queries for each access to an ontology. While APIs are widely used, they do not take full advantage of the object-oriented paradigm and focus more on performance rather than ease of use. In contrast, ontology-oriented programming approaches model objects after ontologies. Ontology

¹⁸<https://www.cidoc-crm.org/versions-of-the-cidoc-crm>

classes, properties, and individuals can be accessed as classes, class attributes, and instances of Python objects, which makes it easier to access and manipulate ontologies.

2.3.3 Evaluation of Ontology Learning systems

There are four main approaches to the evaluation of ontology learning tools: golden standard, application based, data driven, and human based [73]. The technique most suitable to automatically evaluate an ontology-based extraction system is the golden standard approach.

The golden standard evaluation method compares the ontology that has to be evaluated to the golden standard created by domain experts. While convenient, this approach is also subjective considering that there needs to be a manual construction of a golden standard. The quality of the results and conclusions derived is affected by the quality of the matching between the golden standard and the ontology that is to be evaluated. Two standard evaluation metrics are precision and recall calculated in Equations 2.1 and 2.2 [57]. They can be measured at lexical and relational levels to determine the similarity of concepts and how many relations are correct. At the lexical level, string similarity measures, such as the Levenshtein edit distance, can be used to measure the similarity between concept and relation names.

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (2.1)$$

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (2.2)$$

Along with traditional Precision, Recall and F1-score measures, there are also cost-based measures take into account semantic distance. Learning Accuracy (LA) [41] is a distance-based metric that calculates how well an instance was added at the right level of the ontology. However, it has the limitations of not taking the density of the hierarchy into consideration and not being symmetric, i.e., the comparison of two concepts in an ontology gives different results depending on which concept is the key or response.

Balanced Distance Metric (BDM) [58] is another measure that calculates the semantic distance between concepts. It gives a real number between 0 and 1 where the BDM score is higher the closer the concepts are. It was formulated by Maynard et al. [58] as shown in Equation 2.3.

$$BDM = \frac{BR(CP/n1)}{BR(CP/n1) + (DPK/n2) + (DPR/n3)} \quad (2.3)$$

where:

CP = shortest length from root to the most specific common parent

DPK = shortest length from the most specific common parent to the Key concept

DPR = shortest length from the most specific common parent to the Response concept

$n1$ = average chain length of all ontological chains containing Key and Response

$n2$ = average chain length of all ontological chains containing Key

$n3$ = average chain length of all ontological chains containing Response

BR = branching factor of relevant concepts divided by average branching factor of all branch nodes

Based on the BDM measure, Maynard et al. [58] created the Augmented Precision, Recall, and F1-score cost-based ontology population evaluation measures based on the traditional Precision, Recall, and F1-score metrics formulated in Equations 2.4, 2.5 and 2.6.

$$AP = \frac{BDM}{BDM + Spurious} \quad (2.4)$$

$$AR = \frac{BDM}{BDM + Missing} \quad (2.5)$$

$$F1 = \frac{AP * AR}{0.5 * (AP + AR)} \quad (2.6)$$

2.3.4 Related Work

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 that employ linguistic techniques of information extraction.

Witte et al. [98] and Vlachidis et al. [94] 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. The conversion of the scanned material to machine-readable text was performed manually as the source material proved to be too complex to allow an automatization of the process. The authors used a combination of two ontologies to model both the historical documents' structure and their content.

Goy et al. [16] presented an automatic algorithm to define mappings and an information extraction process between the Wikidata semantic model and the HERO ontology. The authors [37] previously created the ontology HERO (Historical Event Representation Ontology) in the context

of the PRiSMHA (Providing Rich Semantic Metadata for Historical Archives) project. The objective of the approach is to use information extracted from Wikidata to provide suggestions to users in an ontology-driven annotation platform of documents from historical archives.

Chapter 3

Proposed Solution

The goal of this thesis is to build an ontology-based information extraction tool that represents unstructured data in ArchOnto. This chapter gives an overview of the proposed system's architecture and describes each module's challenges.

3.1 Architecture

We propose an architecture divided into three modules: Optical Character Recognition (OCR), Ontology-Based Information Extraction (OBIE), and Ontology Population (OP). The execution pipeline is described in Figure 3.1.

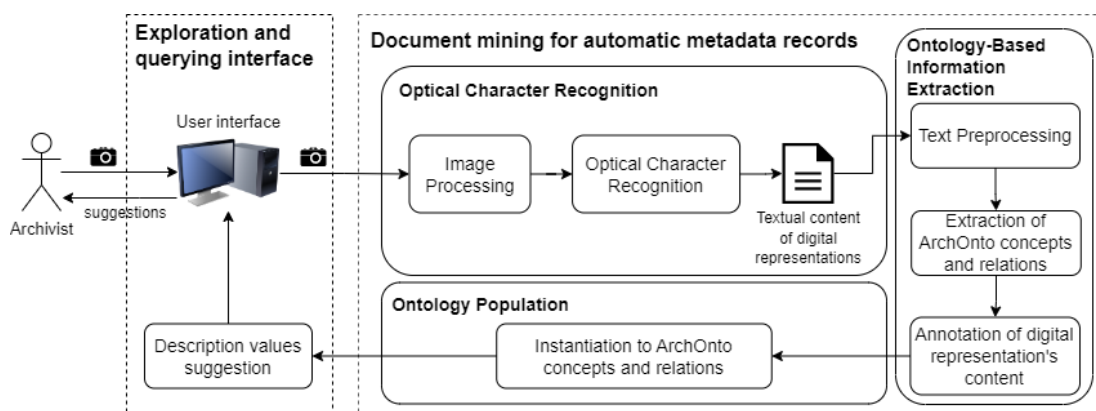


Figure 3.1: An overview of the execution pipeline.

It 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 OBIE module uses trained NER and RE models to predict and annotate concepts and relations from the processed textual content. The extracted information is then 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.

We define the methodology of the image processing pipeline with OCR in Section 4.1.3. As the Ontology-Based Information Extraction and Ontology Population modules are intrinsically linked, we describe the phases to our work’s pipeline for the modules in Figure 3.2.

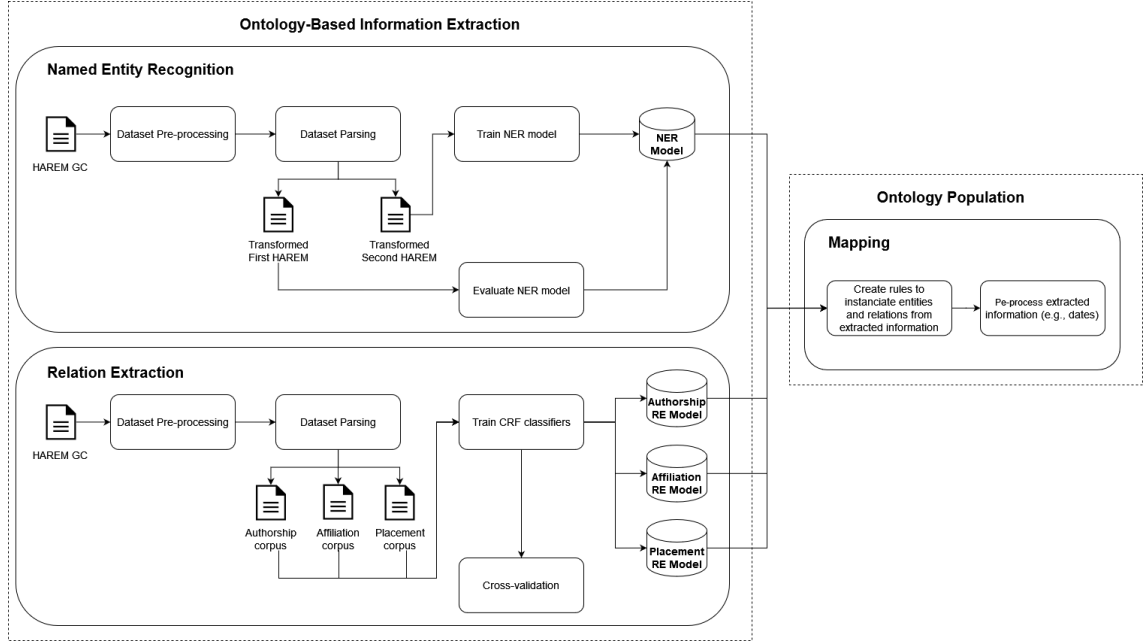


Figure 3.2: An overview of the architecture of the proposed ontology information-extraction and ontology population system.

3.2 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 image quality and overall text extraction.

Our main challenge is determining whether applying an image pre-processing phase to OCR can improve the tasks’ performance in the context of historical documents. Moreover, we want to explore how the parametrization of image processing algorithms can impact the OCR performance on documents with different structures and image quality.

Thus, we propose to conduct an optimization experiment of different image algorithms and parametrization using the OpenCV [8] library and a non-dominated sorting genetic algorithm (NSGA-II) [6] to determine the impact of image processing algorithms on the OCR performance. Converting digital representations to a machine-readable format was executed with Tesseract [1]. The methodology is further explained in Section 4.1.

3.3 Ontology-Oriented Information Extraction

We need an annotated Portuguese dataset to train supervised information extraction models. 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 HAREM collections [34], were pre-processed to suit the ArchOnto ontology's classes. Annotations that were not relevant to 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). The steps for corpora transformation, i.e., Dataset Pre-processing and Dataset Parsing from Figure 3.2, are further detailed in Section 5.1.

We propose to train NER models using a Bidirectional Long Short-Term Memory Network with a Conditional Random Field Layer (BiLSTM-CRF) Neural Network model with pre-trained contextual string embeddings. Implementation details are described in Section 5.2. As for the RE task, we propose to train CRF classifiers on semi-automatically annotated datasets from the Second HAREM. Implementation details are described in Section 5.3.

3.4 Ontology Population

After annotating the text extracted from the digital representations using the trained models, we must map the extracted information to a domain-specific ontology, ArchOnto.

Our solution is to define mapping rules to instantiate concepts and relations and implement them using ontology-oriented programming with the Python package OwlReady2 [51]. We report implementation details in Section 6.1.

Chapter 4

Optimizing Character Recognition in Cultural Heritage Typewritten Documents

This chapter explores several image processing algorithms (Adaptive Thresholding, Simple Thresholding, Triangle Thresholding, Otsu Thresholding, Median Blur, Bilateral Filtering, and Opening), using OpenCV, an image processing Python library with the Tesseract OCR engine.

To solve the parameters' optimization problem, we applied a non-dominated sorting genetic algorithm (NSGA-II) that instantiates the values of parameters of the image processing algorithms. The optimization problem is formalized as a multi-objective optimization problem that maximizes the number of correctly extracted words and minimizes the Levenshtein distance in characters between the extracted text and the ground truth. Each image processing algorithm has a solution set for the optimization problem for each type of document in the dataset. There will be a comparison of combinations of image processing algorithms and the use of default vs. global vs. specific parameter values in the evaluation stage.

4.1 Methodology

The methodology can be divided into four main steps: define and collect a dataset of cultural heritage digital representations, explored in Section 4.1.1, split the dataset into parametrization and evaluation subsets in Section 4.1.2, showcase the image processing pipeline in Section 4.1.3, and present the evaluation method in Section 4.1.4.

4.1.1 Cultural Heritage Dataset

The training of image processing algorithms' parameters and evaluation of the OCR optimization process requires a dataset to conduct various experiments. In this work, we use records from the National Archives of Torre do Tombo (ANTT), a Portuguese central archive with millions of documents that date back to the 9th century. This archive contains more than 40 million digital representations and their respective archival records [25] that can be shared and accessed through an information system called *Digitarq*¹.

4.1.1.1 Creation

To build a dataset of digital representations of archival records, we first had to identify printed or typewritten representations in the overall set of representations. We avoided using handwritten digital representations due to the lower accuracy in text recognition with handwritten documents. Exploring the database, we found that the document classification of the digital representations into typewritten or handwritten is usually done in the fields "Dimension and Support", "Scope and Content", and "Note". The description forms varied from mentions to the Portuguese word for "typewritten" ("datilografado"), its abbreviations ("dact.") or alternative ways of spelling, i.e., form of the word before and after the Portuguese Language Orthographic Agreement of 1990, "dactilografado", and "datilografado". However, we could not systematically find records with printed documents, even though the archive possesses them. This is because we could not establish a pattern in its textual description's content in the same manner we did to records with typewritten documents.

In collaboration with DGLAB, we searched the database for records whose descriptions in the previously mentioned fields contained different forms of the term "typewritten". The query results returned the name of the digital representation files and its dissemination URL that was later used to retrieve each digital representation.

4.1.1.2 Characterization

The dataset [23] has 26,812 one-page digital representations from 8,146 records of two fonds from the 20th century: the General Administration of National Treasury (DGFP) and the National Secretariat of Information (SNI). DGFP was a central department of the Ministry of Finance created during the First Portuguese Republic regime (1910-1926) that managed public assets, including movable and fixed assets that had been under the responsibility of the former royal family. SNI was a public agency created during the authoritarian *Estado Novo* regime (1933-1974) to supervise political propaganda, press organs, popular culture, and entertainment, and promote national tourism. It is characterized in Table 4.1.

¹<https://digitarq.arquivos.pt/>

Table 4.1: Characterization of the dataset’s archival records and digital representations throughout the series’ description level. The digital representations were split by writing format.

Series’ Reference Code	Series’ Name	Number of records	Number of digital representations		
			Blank	Handwritten	Typewritten Overall
PT/PNA/DGFP/0001	Royal Palace’s Listings	59	59	16	57
PT/PNA/DGFP/0002	List of deliveries made to various institutions and personalities	114	158	2	132
PT/PNA/DGFP/0003	Deliveries made to the Royal Family and the respective property appraisal processes	27	43	3	62
PT/PNA/DGFP/0005	Work documentation related to the listings	50	76	9	96
PT/PNA/DGFP/0006	Correspondence	111	147	6	139
PT/TT/SNI-DGE/1	Censorship cases for theatre plays	119	971	324	9,366
PT/TT/SNI-DGE/3	Minutes of the Censorship Commissions’ Sessions	6	19	6	1,479
PT/TT/SNI-DSC/9	Subscription of Press Records	17	9	165	1,007
PT/TT/SNI-DSC/13	Information about forbidden books and period publications	6	58	1,015	2,499
PT/TT/SNI-DSC/19	Records on the apprehension and prohibition of publications	2	1	1	39
PT/TT/SNI-DSC/22	Records on new publications	2	0	13	26
PT/TT/SNI-DSC/24	Nonperiodic publications	3	0	7	34
PT/TT/SNI-DSC/35	Reports on censored books	7,630	1	114	8,653
Overall		8,146	1,542	1,681	23,589
					26,812

The digital representations were classified by writing format (handwritten, typewritten, or blank), as some archival records containing handwritten or blank documents are not relevant to our work. The dataset has 23,589 typewritten digital representations, 1,681 handwritten digital representations, and 1,542 blank digital representations. Furthermore, the dataset was classified into ten types of digital representations. The typewritten dataset has 3,264 letters, 6,560 structured reports, 1,970 non-structured reports, 82 covers of processes, 6 covers of minutes, 1,473 contents of minutes, 19 books' covers, 182 books' contents, 165 theatre plays' covers, and 8,845 theatre plays' contents. Figure 4.1 shows examples of different types of typewritten digital representations from the dataset. Identifying digital representation typologies relevant to the OCR task was carried out by observing existing textual descriptions of the records and document layouts of the digital representations. The classification of the dataset was performed manually.

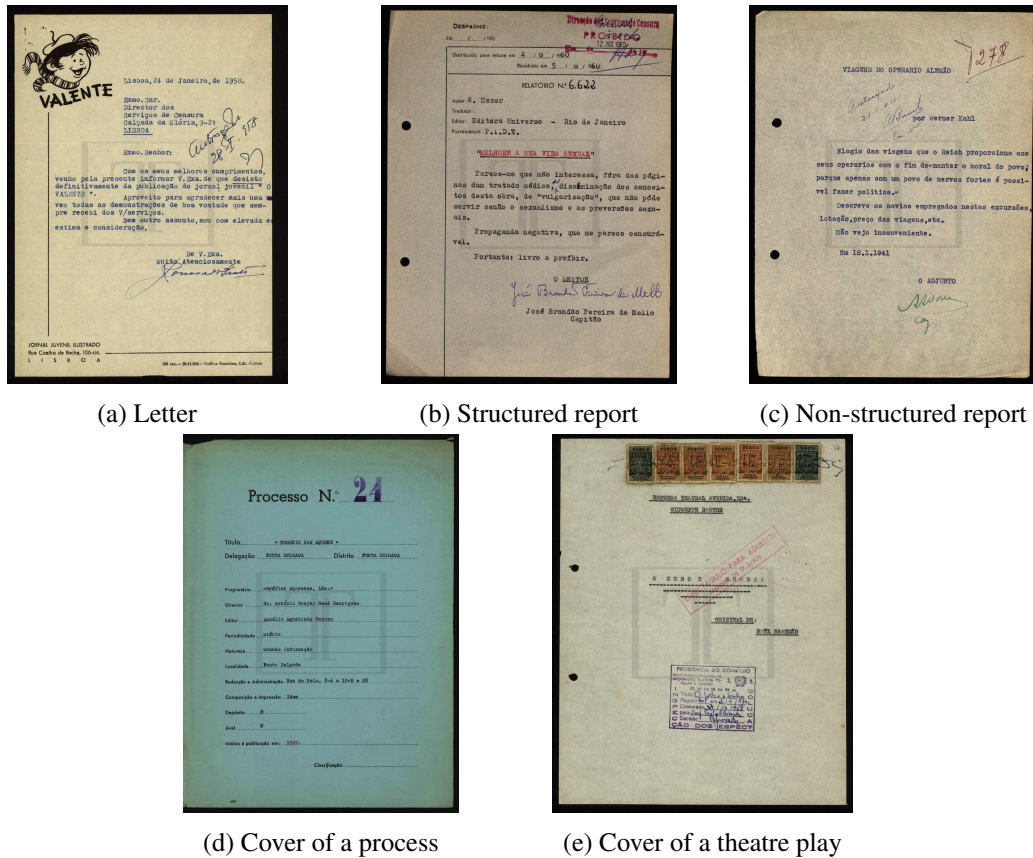


Figure 4.1: Digital representations from the typewritten dataset.

4.1.2 Parameterization and Evaluation Samples

To conduct an unbiased evaluation of the OCR optimization process, we need to use different samples of the cultural heritage dataset to find the best parameter values for image processing algorithms and to assess the algorithms' performance in the text recognition task. The best-performing

image processing algorithms with tuned parameters will be used in the pre-processing phase of the OCR task with the population dataset to extract its textual content.

The dataset of documents we are working with refers to historical records with secondary value, i.e., evidential, and informative values, that no longer possess administrative value [27]. Historical archives consider the documents' information as a whole, unlike current archives that detail the information contained in each document to facilitate quick access to the data. For this reason, we selected the digital representation typologies that contain the most informative value and that can be the most relevant to the description of historical records. Therefore, from the 10 digital representation typologies, we chose the covers of documents (processes and theatre plays' covers), letters, structured, and non-structured reports as the most informative.

The first step of creating the samples was to select 708 typewritten digital representations from the five digital representation types (processes and theatre plays' covers, letters, structured, and non-structured reports). We randomly selected 5% of the digital representations of each archival series associated with each typology. Theatre plays and processes' covers are the only exceptions because they have a small number of digital representations. For the theatre plays' covers typology, since the digital representations are only part of one series, we defined a minimum selection of 60 digital representations. As for the process covers' typology, to match the previous typology with a minimum of 60 digital representations, we considered 75% of the digital representations of each archival series.

The second step was to split the sample into two: one to perform the parameter optimization of image processing algorithms (parameterization dataset) and another to evaluate the OCR performance using the algorithms with tuned parameters (evaluation dataset). The parameterization and evaluation datasets both have 81 letters, 164 structured reports, 49 non-structured reports, 30 covers of processes, and 30 theatre plays' covers. Table 4.2 characterizes the parameterization and evaluation datasets in comparison to the population dataset regarding the number of one-page digital representations through different typologies. The digital representations of the samples were manually transcribed to provide a ground truth in the evaluation stage [30].

Table 4.2: Characterization of the parameterization, evaluation, and population datasets' digital representations.

Series' Reference Code	Series Name	Parameterization / Evaluation / Population Datasets				
		Letter	Process cover	Structured report	Theatre play cover	Non-structured report
PT/PNA/DGFP/0001	Royal Palace's Listings	1/2/57	-	-	-	-
PT/PNA/DGFP/0002	List of deliveries made to various institutions and personalities	4/2/122	-	-	-	-
PT/PNA/DGFP/0003	Deliveries made to the Royal Family and the respective property appraisal processes	1/0/29	-	-	-	-
PT/PNA/DGFP/0005	Work documentation related to the listings	1/0/29	-	-	-	-
PT/PNA/DGFP/0006	Correspondence	1/5/113	-	-	-	-
PT/TT/SNI-DGE/1	Censorship cases for theatre plays	7/0/138	18/22/54	0/1/20	30/30/165	0/0/1
PT/TT/SNI-DGE/3	Minutes of the Censorship Commissions' Sessions	-	-	-	-	-
PT/TT/SNI-DSC/9	Subscription of Press Records	16/18/699	8/5/17	0/1/6	-	0/2/6
PT/TT/SNI-DSC/13	Information about forbidden books and period publications	41/48/1,782	2/1/5	3/0/66	-	0/0/2
PT/TT/SNI-DSC/19	Records on the apprehension and prohibition of publications	0/2/34	0/1/2	-	-	-
PT/TT/SNI-DSC/22	Records on new publications	1/0/23	1/0/1	-	-	-
PT/TT/SNI-DSC/24	Non-periodic publications	0/1/22	1/1/3	0/0/1	-	-
PT/TT/SNI-DSC/35	Reports on censored books	8/3/216	-	161/162/6,467	-	49/49/1,961
Overall		81/81/3,264	30/30/82	164/164/6,560	30/30/165	49/49/1,970

4.1.3 Image Processing Pipeline

We used the Tesseract engine for the OCR task with the Pytesseract wrapper for Python, and the available `tessdata_best` [96] training data. We did not train a language model, since a Portuguese language model is already available, and literature claims that the engine is prepared to handle typewritten documents.

For image processing, the OpenCV library described in Table 2 was used. Three image processing techniques were selected: binarization, smoothing and morphological. The selection of the methods was based on the state of the images in the dataset. The images' degradation stems mainly from contrast variation, bleed-through, faded ink, smears, and thin text [88], e.g., there is no angled digital representation; therefore, there is no need for a skew correction algorithm. This work's pipeline is shown in Figure 4.2.

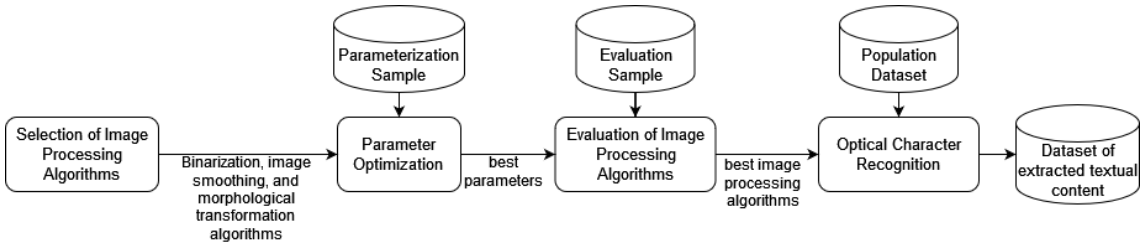


Figure 4.2: Overview of the different steps of the proposed methodology.

4.1.4 Evaluation

Two research questions are associated with this experiment. The first is whether applying parameter tuning to image processing algorithms benefits the OCR performance of cultural heritage digital representations. The second is whether applying image processing algorithms to the cultural heritage digital representations before the text recognition task improves the OCR performance. Our methodology includes two evaluations that address these questions: the effect of parameter tuning of image processing algorithms on the OCR task in Section 4.2.3 and the effect of the algorithms themselves in Section 4.3.

We used performance metrics based on the CER and count-based Bag of Words presented in Table 2.8. The CER-based metric is presented as character accuracy in Equation 4.1 to facilitate the comparison between the measure as success rates. The count-based Bag of Words metric is presented as an F-measure in Equation 4.4, based on the notions of precision in Equation 4.2 and recall in Equation 4.3.

$$\text{Character Accuracy} = (1 - \text{CER}) * 100 \quad (4.1)$$

$$\text{Precision} = \frac{\text{number of words with correct occurrences}}{\text{number of words in OCR output text}} \quad (4.2)$$

$$\text{Recall} = \frac{\text{number of words with correct occurrences}}{\text{number of words in ground truth text}} \quad (4.3)$$

$$\text{F1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (4.4)$$

4.2 Parameter Tuning of Image Processing Algorithms

The process of formalizing and solving the problem of parameter tuning into a multi-objective optimization is detailed in this section. We identify image processing algorithms with potential impact on the digital representations in Section 4.2.1, determine image processing algorithms' best parameter values using genetic algorithms in Section 4.2.2, and evaluate the effect of image pre-processing on OCR performance, which is analyzed in Section 4.3.

4.2.1 Problem Formalization and Optimization

The autotuning of image processing algorithms is defined as a constrained multi-objective problem with two objective functions, $f_1(x)$ formalized by Equation 4.5 [33], and $f_2(x)$, formulated by Equation 4.6. The first objective function, $f_1(x)$, is the Levenshtein edit distance between the ground truth and the output. The edit distance calculates the minimum number of characters' edit operations, such as insertions, substitutions, and deletions, to transform a string a into b with lengths i and j , respectively. In contrast, the second objective function, $f_2(x)$, is the number of words with the same word occurrences between the ground truth and the output, where i is the index of matched words between strings a and b , and $n_a(i)$ is the number of occurrences of the matched word i in string a . The goal is to find solutions for the parameter values that satisfy each image processing algorithm's constraints and are as good as possible for both objective functions [6].

$$\text{lev}_{a,b}(i, j) = \begin{cases} \max(i, j) & , \min(i, j) = 0 \\ \min \begin{cases} \text{lev}_{a,b}(i-1, j) + 1 \\ \text{lev}_{a,b}(i, j-1) + 1 \\ \text{lev}_{a,b}(i-1, j-1) + 1 \end{cases} & , \text{otherwise} \end{cases} \quad (4.5)$$

$$\text{bow}_{a,b}(i) = \begin{cases} 1 & , n_a(i) = n_b(i) \\ 0 & , \text{otherwise} \end{cases} \quad (4.6)$$

We used the Non-dominated Sorting Genetic Algorithm (NSGA-II) [6] to perform parameter tuning of image processing algorithms. The NSGA-II modifies the processes of mating and survival selection. Since not all individuals can survive, individuals are first compared by rank and crowding distance. It uses a binary tournament mating selection to determine rank and the crowding distance uses the Manhattan Distance to select solutions.

For the formalization of the constraints, $g(x)$, the algorithms' equality constraints are transformed into inequality constraints $g(x) \leq 0$. All the constraints needed for the selected image processing algorithms are of the type "a parameter must be odd", which can be expressed through the equality constraint $|\text{parameter}\%2| = 1$ and transformed into the inequality $|\text{parameter}\%2 - 1| \leq 0$, where $g(x) = |\text{parameter}\%2 - 1|$. The algorithms Adaptive Thresholding, Homogeneous Blur, Median Blur, and Gaussian Blur all have a parity constraint on the parameters `blockSize`, in the case of the thresholding algorithm, and on the parameter `ksize` for the rest.

4.2.2 Best Parameters Per Digital Representation Typology

We applied the NSGA-II algorithm to tune image processing algorithms using the parameterization dataset. We considered two parametrization scenarios: global and by digital representation typology. Global parameter tuning consisted of using the overall parametrization sample instead of subsets of the sample by digital representation typology. We found better trends with digital representation-type parametrization than with global parametrization. For this reason, we will only be considering digital representation-type parametrization from now on for which the results are shown in Table B.1. The default values are the ones suggested in the OpenCV library's documentation. As seen, the obtained parametrization values are often quite different from the default parameter values. It should be noted that the description of nominal parameters' values, such as type, adaptiveMethod, thresholdType, and borderType can be consulted in Tables 2.3, 2.5, and 2.7. Moreover, the dilation algorithm has no results for letters because the average result of the objective functions was not valid due to the OCR result of applying dilation with the retrieved parameter values only containing stop words or not containing discernible words.

We performed the experiment on a PC with 3.20 GHz, AMD Ryzen 7 5800H, and 16GB RAM. Table 4.3 shows the average time to parameterize the image processing algorithms averaged over the digital representation typologies. The table shows that parameter tuning for each digital representation took the least time, in the range of 2 minutes, with the image processing algorithms Dilation, Erosion, and Morphological Gradient. In contrast, the parameterization for each digital representation took the longest with the algorithms Black Hat, Opening, and Otsu Thresholding, within 10 minutes.

4.2.3 Effect of Parameterization on OCR Performance

To evaluate the effect of default, global and digital representation-type parametrization on the OCR performance, we compared the performances of image processing algorithms in OCR with default parameter values to algorithms with parameter values obtained by our parametrization. Table B.2 displays the comparison of each image processing algorithm using default, global and digital representation-specific parameter values. To compare the character accuracy and F1-score results of each method using parameters tuned by digital representation typologies and with the default ones, we use the Wilcoxon signed-rank test [67] with the Bonferroni correction. The table presents the results of the Wilcoxon pairwise comparisons between the same algorithms using

Table 4.3: Average time to parameterize the image processing algorithms.

Algorithm	Average parameter tuning time (min:s)
Adaptive Thresholding	06:46
Otsu Thresholding	10:52
Simple Thresholding	06:00
Triangle Thresholding	08:52
Bilateral Filter	09:11
Gaussian Blur	07:12
Homogeneous Blur	04:33
Median Blur	08:39
Black Hat	10:29
Closing	05:31
Dilation	02:02
Erosion	02:59
Morphological Gradient	02:35
Opening	10:35
Top Hat	07:11

default, global or digital representation-specific parameter values. We used this non parametric test because the assumptions for the t-test were not verified. One-sided hypothesis tests were conducted to determine whether the OCR performance using digital representation-type parameterization was better than the OCR performance using default or global parameter values.

By analyzing the statistical outcomes, the OCR performance using digital representation-type parameterization on the image processing algorithms shows statistically significant improvements in both measures compared to the parametrization using default values. Similarly, the OCR performance using global-type parametrization also shows significant improvement in both measures compared to using default values, except for Simple Thresholding and Dilation where the performance is worse. For most algorithms there is either not a significant difference between global and digital representation-type typology or the specific parametrization performs better in both measures, except for the algorithms Otsu Thresholding, Homogeneous Blur, Median Blur, and Morphological Gradient where it performs worse in at least one of the measures.

4.3 Effect of Image Pre-Processing on OCR Performance

While in the previous section we studied the effect of parameterization of image processing algorithms on the OCR performance, this section aims to analyze the effect that the different image processing algorithms have on the OCR performance.

We evaluate image processing algorithms with the digital representation-type parameterization because it achieved a significantly better performance in the OCR task than with default parameter values and tendentially better results than the global parameter values. For this analysis, we will

consider the evaluation dataset. We will conduct our analysis in the overall set of documents and by digital representation typology, that is, letters, process covers, structured reports, theatre plays' covers, and non-structured reports. We used the Friedman test to determine significant differences between the OCR performance using the image processing algorithms, and the one-sided Wilcoxon signed-ranked test with the Bonferroni correction as a post-hoc test to discover which algorithms perform better in relation to each other and to not applying any. We used these non parametric tests because the assumptions for repeated measures ANOVA and t-test were not verified.

4.3.1 Overall

In this section, we will present the results regardless of the digital representation typologies. Table B.3 presents the results of the statistical tests performed on image processing algorithms for the overall dataset. The results show that, in the overall dataset, Bilateral Filter and Opening are the best-performing algorithms. Both algorithms significantly improve the OCR performance in terms of character accuracy, with the Opening algorithm reaching a more significant improvement. In contrast, Median Blur and Morphological Gradient are the lowest-performing algorithms in character accuracy, and Morphological Gradient is the lowest-performing algorithm in F1-score. Four algorithms worsen the OCR performance in terms of character accuracy: Otsu Thresholding, Triangle Thresholding, Median Blur, and Morphological Gradient. Similarly, nine algorithms worsen the OCR performance in terms of F1-score: Adaptive Thresholding, Otsu Thresholding, Simple Thresholding, Triangle Thresholding, Gaussian Blur, Homogeneous Blur, Median Blur, Morphological Gradient, and Black Hat.

Looking at the results for the binarization algorithms, the Simple Thresholding algorithm is the best-performing algorithm. However, it does not improve the OCR performance. For the smoothing algorithms' results, Bilateral Filter is the best-performing smoothing algorithm. Every smoothing algorithm except Bilateral Filter deteriorate F1-score when used to pre-process the images before OCR. The results for the morphological transformation algorithms show that Opening is the best-performing algorithm in comparison to not applying an image pre-processing algorithm and other binarization, smoothing, and morphological transformation algorithms.

4.3.2 Letters

In this section, we will present the results regarding the letters typology. Table B.4 presents the results of the statistical tests performed on image processing algorithms for the letters sample. The results for the letters sample show that, similarly to the results for the overall dataset, Bilateral Filter is again the best-performing algorithm. Triangle Thresholding and Morphological Gradient are the worst-performing algorithms. Moreover, the mean results of character accuracy and F1-score are higher than in the overall dataset.

For the binarization methods' results, all the algorithms, except Adaptive Thresholding, decrease character accuracy. The results for F1-score are akin to the overall dataset. As for the

results for the smoothing methods, Bilateral Filter is still the best-performing algorithm in character accuracy. However, the results for the morphological transformation algorithms indicate that, in comparison to the overall dataset, Opening no longer improves the OCR performance.

4.3.3 Non-structured reports

In this section, we will present the results regarding the non-structured reports typology. Table B.5 presents the results of the statistical tests performed on the image processing algorithm for the non-structured reports sample. The results for the sample of non-structured reports show that the mean results of the OCR performance are slightly higher than in the overall dataset and no algorithm improves the OCR performance. Black Hat and Opening are the best-performing algorithms and Homogeneous Blur, Median Blur, and Morphological Gradient are the worst-performing algorithms.

The binarization methods Otsu and Triangle Thresholding are the worst-performing algorithms as they deteriorate the OCR performance. As for the smoothing algorithms' results, Bilateral Filter is the best-performing algorithm. All the smoothing algorithms except Bilateral Filter have significantly worse results than not using any image processing algorithm. The results for the morphological transformation methods indicate that Black Hat and Opening are the best-performing algorithms. In sum, most morphological transformation algorithms improve the OCR performance in comparison to binarization and smoothing algorithms in the non-structured reports typology.

4.3.4 Process covers

In this section, we will present the results regarding the process covers typology. Table B.6 presents the results of the statistical tests performed on image processing algorithms for the process covers sample. The results show that, in the process covers sample, the mean OCR performance values are lower than in the overall dataset, and no algorithm improved the OCR performance.

There are no significant differences between the binarization algorithms. Moreover, unlike in the previous sections, binarization algorithms do not impair the OCR performance. As for the smoothing methods, we can observe that most algorithms have a better performance in character accuracy and F1-score than Black Hat, Morphological Gradient, and Otsu Thresholding. Finally, the results with morphological transformation algorithms indicate that Black Hat and Morphological Gradient are the lowest-performing algorithms and deteriorate the OCR performance.

4.3.5 Structured reports

In this section, we will present the results regarding the structured reports typology. Table B.7 presents the results of the statistical tests performed on image processing algorithms for the structured reports sample. The results for the structured reports sample show that no algorithm improved the OCR performance in comparison to not using any image processing algorithm.

We can observe that, from the performance results of binarization methods, Adaptive Thresholding and Otsu Thresholding impair the OCR performance. For the smoothing algorithms' results, we see that Bilateral Filter is the best-performing algorithm. However, it does not improve the OCR performance. On the other hand, Median Blur and Gaussian Blur worsen the OCR performance. The results for morphological transformation algorithms indicate that Morphological Gradient and Black Hat deteriorate the performance.

4.3.6 Theatre plays' covers

In this section, we will present the results regarding the theatre plays' covers typology. Table B.8 presents the results of the statistical tests performed on the image processing algorithm for the theatre plays' covers sample. The results for the theatre plays' covers sample show a significantly lower mean OCR performance than in the overall dataset. Furthermore, unlike the results in the overall dataset and the previous typologies where binarization algorithms worsened the OCR performance, in this sample it is a binarization algorithm, Adaptive Thresholding, that is the best-performing algorithm. In sum, three algorithms improved character accuracy: Adaptive Thresholding, Black Hat, and Top Hat.

4.4 Discussion

It is important to stress that special care for parameterization influences the impact of image processing algorithms on OCR, that is, not only are the parameter values very different with digital representation-type parametrization, but the results are significantly better. Despite the difference between global and digital representation-type parametrization not being significant, parameterization by typology seems to be beneficial when applying image pre-processing algorithms, even with a small dataset. It also has the advantage of being faster to process.

From the previous sections, we can conclude that there are no significant improvements in the OCR performance in the overall dataset when using binarization algorithms in an image pre-processing OCR phase in comparison to not using them. Unlike the binarization method, the smoothing and morphological transformation methods significantly improved the OCR performance with Bilateral Filter and Opening, respectively.

The results show that it is possible to improve the text recognition task using tuned image processing algorithms in the overall dataset and the subsets of letters and theatre plays' covers. However, no image processing algorithms significantly improved the OCR performance with the samples of process covers, structured reports, and non-structured reports. In Table 4.4, we list the image processing algorithms that tended to improve the OCR performance, that is, the algorithms that had a better mean performance than not applying an image processing algorithm. From these algorithms, we selected the ones that significantly improved the OCR performance without an image pre-processing phase for the overall dataset and each digital representation typology. The table shows the mean performance values for applying the selected significant algorithms in the OCR task.

Table 4.4: Image processing algorithms that improved the OCR performance per digital representation typology and overall and were selected as the most significant.

Typology	Improved algorithms	Selected algorithm	Mean character accuracy	Mean F1-score
Overall	Bilateral Filter, Opening and Top Hat	Opening	67.8%	49.2%
Letters	Bilateral Filter and Black Hat	Bilateral Filter	78.6%	60.2%
Non-structured reports	Black Hat, Erosion, Opening and Top Hat	None	69.2%	53.8%
Process covers	Simple Thresholding, Bilateral Filter, Gaussian Blur, Homogeneous Blur, Median Blur, Opening and Top Hat	None	62.4%	35.3%
Structured reports	Bilateral Filter	None	69.7%	50.1%
Theatre plays' covers	Adaptive Threshold, Black Hat, Opening and Top Hat	Adaptive Threshold	35.6%	21.8%

The mean OCR performance values refer to the employment of the the selected image pre-processing algorithms.

We present the distribution of the performance measures for the overall dataset and the digital representation-type samples in Figure B.1. We conclude that the digital representation typology with the best results is the letters sample, with a median character accuracy above 75% and a median F1-score above 60% for most algorithms. In contrast, the digital representation typology with the worst performance results is the theatre plays' covers sample, with a median character accuracy below 25% and a median F1-score below 20%. Image processing algorithms appear to have a greater impact on the OCR performance of digital representations from the theatre plays' covers sample, where the OCR performance without applying image processing algorithms is low.

Process covers and theatre plays' covers are the samples with the lowest mean OCR performance values. They are also the samples with the most algorithms that tend to improve the OCR performance or do so in a significant way. These findings suggest that image processing algorithms tend to have a greater impact on digital representation typologies that have a lower OCR performance without an image pre-processing phase.

The results also show that morphological transformation algorithms tend to be the most beneficial to every typology. In contrast, binarization algorithms only benefit the theatre plays' covers typology and tend to do so for the sample of process covers. On the other hand, while smoothing algorithms tend to be beneficial to every typology, like morphological transformation algorithms,

only one algorithm, Bilateral Filter, is responsible for that performance improvement.

Chapter 5

Ontology-Based Information Extraction

Machine learning information extraction techniques are often trained on annotated text. However, there are no available archival data sets in Portuguese and manually creating a large annotated corpus is a time-consuming task.

This chapter will describe the pre-processing of an available Portuguese corpus, the HAREM golden collection, to suit the ArchOnto representation. We will use the transformed corpora to train and test NER and RE models.

5.1 Corpora Transformation

The annotated corpora provided by Linguateca, the HAREM GCs, follow an XML format that preserves the original text while adding semantic information. However, to train IE models, we must convert the datasets to a TSV format and remove entities and relations irrelevant to ArchOnto, since the extracted information will be structured and stored following the knowledge base.

The following sections detail the steps of the corpora transformation. The pipeline is shown in Figure 5.1.

5.1.1 Dataset Pre-Processing

The HAREM's annotation syntax has five tags: DOC, P, EM, ALT, and OMITIDO. Table 5.1 describes the meaning of the tags.

The EM tag is a classification label with more than one possible attribute. The First HAREM collection's EM tag has two attributes: CATEGORIA (category) and TIPO (type). In contrast, the Second HAREM collection's EM tag has three attributes: CATEGORIA (category), TIPO (type), and SUBTIPO (subtype). All attributes are optional. Listing 5.1 contains two identified temporal expressions from the First and Second HAREM collections, respectively, that showcase the differences in classification. The entity extracted from the First HAREM GC has the attribute values TEMPO (time) for category and DATA (date) for type. On the other hand, the entity

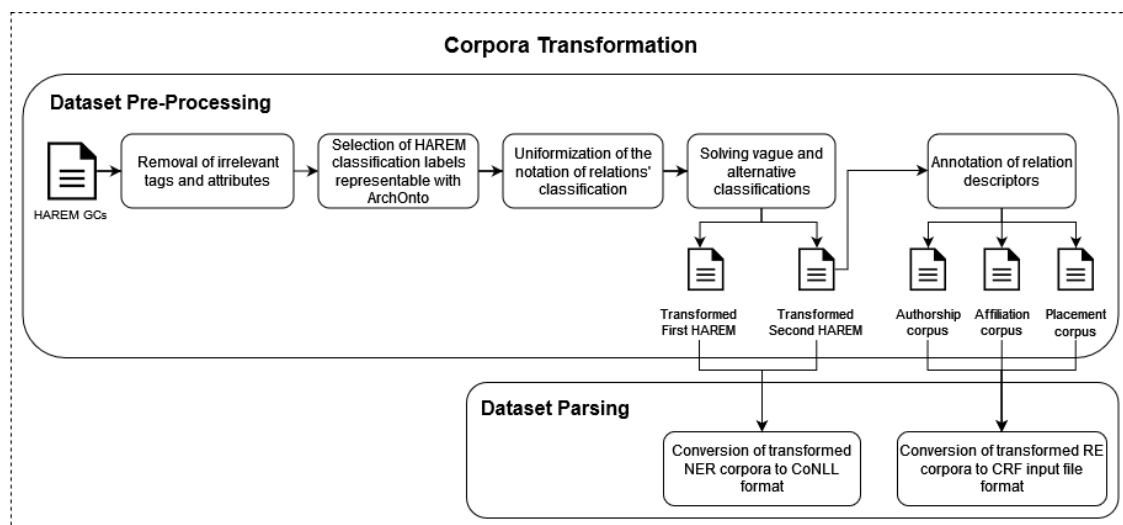


Figure 5.1: Overview of the different steps of the corpora transformation.

Table 5.1: Description of HAREM tags.

HAREM Tag	Meaning	Description
DOC	Document	Identifies different documents
P	Paragraph	Identifies sentences that might contain an entity
EM	Recognized entity	Identifies recognized entities
ALT	Alternative analysis	Identifies alternative classification analysis of textual segments
OMITIDO	Omitted text	Identifies a lack of majority consensus on classifying textual segments during the golden collection annotation process

extracted from the Second HAREM GC has the attribute values TEMPO (time) for the category, TEMPO_CALEND (calendar time) for the type, and DATA (date) for the subtype.

```

<EM ID="5" CATEG="TEMPO" TIPO="DATA">Junho de 1992</EM>

<EM ID="H2-dftre765-52" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">em
  Novembro de 1534</EM>

```

Listing 5.1: Extracted temporal expressions from the First and Second HAREM GC that showcase the differences in named entity classification between the corpora.

In the Second HAREM corpus with annotated relations, each named entity identified with the EM tag has a unique ID that allows annotators to link entities through relations. The EM tags indicate relations with the attributes COREL, which contains the ID of the related entity, and TIPOREL, which classifies the type of relationship. Moreover, a recognized entity can be associated with more than one entity through several relations. List 5.2 shows an example of an

identified named entity, "Ministério da Cultura" (Ministry of Culture), that contains a relationship of identity ("ident") to a named entity with the ID "relkj7666-222".

```
<EM ID="relkj7666-231" CATEG="ORGANIZACAO" TIPO="ADMINISTRACAO" COREL="relkj7666-222" TIPOREL="ident">Ministério da Cultura</EM>
```

Listing 5.2: Recognized named entity from the Second HAREM GC that showcases the annotation of relationships between entities.

The EM tag may also contain comments provided by the annotators for future studies with the attribute COMENT. The attribute takes the following values¹:

- 2/3 - indicates named entity classifications that were not consensual and were, instead, the result of a majority vote;
- DUVIDA_DIRECTIVASTEMPO - signals temporal expression for which the annotators were uncertain about their classification, usually because of different interpretations of the annotation guidelines;
- Duvida_ReReLEM - indicates uncertainties in the classification of relations;
- INDEP - specifies relations that annotators could not infer with the textual information available to them;
- ReReLEM_passado - indicates relationships that happened in the past;
- ReReLEM_futuro - marks relationships that will only happen in the future.

5.1.1.1 Removal of irrelevant tags and attributes

We removed the OMITIDO tag since the HAREM contest did not use it in its evaluation. Instead, we kept the text inside the tag without annotations. We also removed the COMENT attribute from the annotated entities. Particularly, we chose to keep the relations in annotated entities with the INDEP comment, even though they mark relationships that cannot be inferred from the textual information available in the corpus. This is because there were only eight instances where the annotators used the INDEP comment, and in some instances, the comments contained a question mark. Furthermore, in the case of annotators making this comment on an entity with more than one relationship, it is unclear which one they were referring to in their assessment.

5.1.1.2 Selection of HAREM classification labels representable with ArchOnto

We selected HAREM classification labels that describe entities and relationships between entities that we can represent with the ArchOnto model. We focused on the classes and properties from the ArchOnto and CIDOC-CRM ontologies for this task. Table 5.4 details the mapping between

¹https://www.linguateca.pt/aval_conjunta/HAREM/Readme.txt

concepts from the Second HAREM and ArchOnto classes from the CIDOC-CRM core and extension. The mapping for the First HAREM is the same as the one shown in the previous table without the HAREM subtype, except for temporal expressions mapped according to Table 5.2.

Table 5.2: Mapping between relevant concepts identified in the First HAREM corpus and ontology classes that differs from the mapping done with the Second HAREM corpus.

Concept	HAREM category	HAREM type	Class
Expression of time	Tempo	Data, Periodo	E52 Time-Span (CIDOC-CRM)

It is important to note that we additionally processed recognized temporal entities to include only full or partial dates. We removed temporal expressions such as "passada sexta-feira" (last Friday), "mais tarde" (later), "ontem" (yesterday), and "há quinze anos" (fifteen years ago) from the corpora. We chose to extract dates because, in the archival context, it is a vital piece of information to have. It might clue archivists in on the nature of a document or serve to situate events in time, or eventually be used as metadata to search for other documents or entities in the same temporal timeframe.

For the corpus solely destined for the Named Entity task, we replaced the EM tag for each recognized entity with the ontology-based classification. However, for the corpus that contains relationships between entities, we kept the EM tag since the Relation Extraction task requires us to link entities through identifiers and specify entity and relation classifications. Instead, we replaced the attributes category, type, and subtype with the attribute ENT for the named entity classification. For each named entity we removed, we checked if they were part of a relationship connection to other entities. If so, we also remove the reference to the deleted named entity from the TIPOREL and COREL attributes of the other entities.

We maintained the Second HAREM's terminology for relations. However, to simplify the future process of processing the corpus to train an RE model, we replaced the alphanumeric IDs with incremental IDs starting with 0 for each document. We also updated the references to the identifiers in the attribute COREL. We list the relationships we selected to use in our work in Table 5.3.

The other relations not listed in Table 5.3 were either stripped from the corpus or replaced. We replaced the relations "localizado_em / localizacao_de" (located in / place of) between the HAREM categories Obra (Title) and Local (Place) with the relation "inclui / incluido" (includes / included by) between entities of the CIDOC-CRM class E53 Place. Similarly, we replaced the relations "propriedade_de / proprietario_de" (ownership) with the relations "inclui / incluido" (includes / included by) and "produtor_de / produzido_por" (producer of / produced by) dependent on the context.

List 5.3 exemplifies two different sentences that used the relation of ownership in different contexts. In the first sentence, a Museum is the owner of the Sophia Jobim collection through a donation, while in the second sentence the sales and car manufacturer company BMW owns the company Rover. The first sentence represented in ArchOnto would result in the use of the

Table 5.3: Meaning and purpose of HAREM relationships kept in the corpus.

HAREM relation	Meaning	Purpose
autor_de / obra_de	authorship	Connects an author to their work
data_nascimento	date of birth	Links a person to their date of birth
data_morte	death date	Links a person to the date on which they died
datado_de / data_de	dated from / date of	Provides the date of events and foundational date of groups and works' titles
ident	identity	Links identical entities
incluir / incluído	includes / included by	Expresses inclusion between instances of the same entity
local_morte	place of death	Links a person to the place they died at
natural_de / local_nascimento_de	place of birth	Links a person to the place they were born on
ocorre_em / sede_de	placement	Indicates the location of groups and events
participante_em / ter_participacao_de	participation in	Expresses the participation of people in works or events
periodo_vida	lifetime	Links a person to the period of their lifetime, i.e., date of birth and death
produtor_de / produzido_por	producer of / produced by	Expresses the act of executing or producing works or events by groups or people
residente_em / residencia_de	place of residence	Connects a person to the place they live in
vinculo_inst	affiliation	States an affiliation between people and groups

CIDOC-CRM class E10 Transfer of Custody with the ArchOnto class ARE15 Acquisition Type donation as a type. Our solution in cases such as this is simply to remove the relation between entities. There are two reasons for this decision. First, our objective with this work is not to further annotate the HAREM corpora with new classification labels, but simply to use the existent annotations in the HAREM corpora and understand how they work in the archival context. Second, it would not be reasonable to create a new relation to portray the activity of transfer of custody, since most of the content related to museums in the Second HAREM comes from a document with extracted content from the National Historical Museum² in Brazil from 2008. Therefore, there are not many instances where this relation could be annotated. On the other hand, the second sentence represented in ArchOnto could be expressed as *BMW has current or former member (P107) Rover*. The CIDOC-CRM property *P107 has current or former member* can be portrayed

²<http://www.museuhistoriconacional.com.br>

Table 5.4: Mapping between relevant concepts identified in the Second HAREM corpus and ontology classes.

Concept	HAREM category	HAREM type	HAREM subtype	Class
Person	Pessoa	Individual, Membro	-	E21 Person (CIDOC-CRM)
Organization	Organizacao	Administracao, Empresa,	-	E74 Group (CIDOC-CRM)
		Instituicao, Outro	-	
	Pessoa	GrupoInd, GrupoMembro	-	
Location	Local	Humano	Rua, Pais, Divisao, Regiao, Outro	E53 Place (CIDOC-CRM)
		Fisico	Ilha, Aguacurso, Planeta, Regiao, Relevo, Aguamassa, Outro	
			-	
			-	
Expression of time	Tempo	Tempo_Calend	Data, Intervalo	E52 Time-Span (CIDOC-CRM)
Event	Acontecimento	Efemeride, Evento, Organizado, Outro	-	E5 Event (CIDOC-CRM)
Quantity	Valor	Quantidade	-	E54 Dimension (CIDOC-CRM)
Role of a Person in an Event	Pessoa	Cargo, GrupoCargo	-	ARE8 Role Type (ArchOnto)
Title assigned to texts, artworks, and musical pieces	Obra	Arte	Pintura	ARE2 Formal Title (ArchOnto)
		Plano	-	
		Reproduzida	Livro, Programa, Musica, Outro	

in the HAREM corpus as an inclusion relationship. Thus, our solution is to change the relationship from ownership to inclusion between groups.

```
Toda a coleção
<EM ID="152" COREL="145" TIPOREL="ident" ENT="ARE2">Sophia Jobim</EM>
é doada ao
<EM ID="153" COREL="152 12 135 123 103" TIPOREL="proprietario_de ident
    proprietario_de proprietario_de proprietario_de" ENT="E74">Museu</EM>
após a sua morte.

Quantos milhões de libras pagou a
<EM ID="9" COREL="7" TIPOREL="ident" ENT="E74">BMW</EM>
pela
<EM ID="10" COREL="7" TIPOREL="propriedade_de" ENT="E74">Rover</EM>?
```

Listing 5.3: Extracted sentences from the Second HAREM GC that showcase the use of the ownership relationship in different contexts.

5.1.1.3 Uniformization of the notation of relations' classification

In the cases of vague classifications of named entities, multiple classifications are annotated and split with the vertical bar symbol "|". If the vague entities contain relations, the relations' classification is explicitly referenced in the TIPOREL attribute by prefacing the relation with the category of the vaguely named entity (e.g., *CURRENT_CATEGORY**relation*) and finishing it with the ID of the related entity and its classification (e.g., *CURRENT_CATEGORY**relation**related_id**-RELATED_CATEGORY*).

We uniformized the notation of the relations' classification by removing the explicit references to the vague named entities classification in the TIPOREL attribute. List 5.4 gives an example of an identified vague entity, a Museum, annotated with two classifications: category LOCAL (place), type HUMANO (human), and subtype CONSTRUCAO (construction); and category OBRA (work), type ARTE (art), and subtype EDIFICIO (building). The entity classified as a place has a relation of inclusion with an entity with the ID "relkj7666-205". Moreover, the entity classified as a work of art has two relations of inclusion and placement with the entities identified with "relkj7666-205" and "relkj7666-206". The first named entity is the output of the Second HAREM GC and the second is the output of our corpus transformation up to this phase.

```

<EM ID="relkj7666-207" CATEG="LOCAL|OBRA" TIPO="HUMANO|ARTE" SUBTIPO="CONSTRUCAO|
EDIFICIO" COREL="relkj7666-205 relkj7666-205 relkj7666-206" TIPOREL="LOCAL**
inclui**relkj7666-205**LOCAL OBRA**inclui**relkj7666-205**OBRA OBRA**
localizado_em**relkj7666-206**LOCAL">Museu</EM>

<EM ID="2" COREL="0 0 1" TIPOREL="inclui inclui localizado_em" ENT="E53-ARE2">Museu
</EM>

```

Listing 5.4: Example of a vague named entity that has a different relation notation.

5.1.1.4 Solving vague and alternative classifications

The corpora still contain vague entities and the ALT tag to indicate alternative classifications for the named entities. We chose a manual approach to select a suitable classification for the archival context and to keep relevant relations between entities.

Listing 5.5 shows how we resolved the case of the vague entity presented in Listing 5.4. The sentence reads: "From the Fortress of Saint James, at the Calaboose Point, the evolution of the architectural complex of the Museum...". Annotators marked the Fortress of Saint James and Calaboose Point as places. They also indicated a relation of inclusion between them, with the Calaboose Point included by the Fortress of Saint James. In this context, it makes sense to classify the entity Museum as a place related to the Fortress by inclusion.

```

A partir do
<EM ID="0" ENT="E53">Forte de Santiago</EM>
, na
<EM ID="1" COREL="0" TIPOREL="inclui" ENT="E53">Ponta do Calabouço</EM>
, a evolução do conjunto arquitetônico do
<EM ID="2" COREL="0" TIPOREL="inclui" ENT="E53">Museu</EM>

```

Listing 5.5: Resolved vague named entity with a textual excerpt that uses the example from Listing 5.4.

We identified some patterns in our task of solving the issue of having alternative classifications for the same named entity:

- Dimension - we do not include temporal expressions. Listing 5.6 shows an example where "15 years" can be classified as a dimension or not be identified as a named entity, for which we opt for the second option;
- Group - we split a group's name from other entities (person's name, another group, or a place) unless they are part of the actual name of the group;
- Person - for historical figures, we do not split a title from their name. Listing 5.7 expresses an example of a named entity of a person, Catherine of Aragon. We can classify the entity in

its completion as a person or split it into her first name, Catherine, and the place she ruled, Aragon. We chose to keep the reference to the nation she ruled in her name as it is relevant to identify her as a person;

- Person's role - we split the role from the group a person belongs to or from the person's name if it makes sense. Listing 5.8 shows an example of a person identified through their role in a group. We chose to classify the role separately to the group and added a relation of affiliation between them. On the other hand, Listing 5.9 exhibits an example of a person identified through their religious role. We opted to classify the religious title separately to the city and the person's name. Additionally, we included a relation of identity between the role and the person, along with a relation of affiliation between the group and the person;
- Place - we split places' names from other places unless their existence is vital to identify the place. Listing 5.10 shows an example where we identify the name of a venue. We rejected to split the entities into the venue "Cultural Center" and its location "Belém" because we consider the location to be a part of the name of the place;
- Temporal expression - we only keep information regarding the date. Listing 5.11 shows an example where we identify a temporal entity as the date "June 13, 1888" and do not include the information that it was in the afternoon;
- Work's title - we split the title from the author. Listing 5.12 indicates an example of textual expression from which we identify a document's title and its author.

```
Neste arco temporal
<ALT>
  de 15 anos
  |
  de <EM ID="13" ENT="E54">15 anos</EM>
</ALT>

Neste arco temporal de 15 anos
```

Listing 5.6: Excerpt of an ALT case with an entity tagged as an E54 Dimension. Textual expression reads: "In this time span of 15 years". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.

```

<ALT>
  <EM ID="31" COREL="16" TIPOREL="natural_de" ENT="E21">Catarina de Aragão</EM>
  |
  <EM ID="32" COREL="16" TIPOREL="natural_de" ENT="E21">Catarina</EM> de <EM ID="
    33" COREL="16" TIPOREL="incluido local_nascimento_de" ENT="E53-E74">Aragão
  </EM>
</ALT>

<EM ID="29" COREL="14" TIPOREL="natural_de" ENT="E21">Catarina de Aragão</EM>

```

Listing 5.7: Excerpt of an ALT case with an entity of a person. Textual expression reads: "Catherine of Aragon". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.

```

<ALT>
  diretor do <EM ID="57" COREL="14" TIPOREL="ident" ENT="E74">Museu Histórico
    Nacional</EM>
  |
  <EM ID="58" ENT="ARE8">diretor do Museu Histórico Nacional</EM>
</ALT>

<EM ID="52" ENT="ARE8">diretor</EM>
do
<EM ID="53" COREL="12 52" TIPOREL="ident vinculo_inst" ENT="E74">Museu Histórico
  Nacional
</EM>

```

Listing 5.8: Excerpt of an ALT case with a person's role in a group. Textual expression reads: "director of the National Historical Museum". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.

```

<ALT>
  <EM ID="21" ENT="E21">Rabino de Antuérpia Jacob Kruger</EM>
  |
  <EM ID="22" ENT="ARE8">Rabino de Antuérpia</EM>
  <EM ID="23" ENT="E21">Jacob Kruger</EM>
  |
  <EM ID="24" ENT="ARE8">Rabino</EM>
  de
  <EM ID="25" ENT="E53">Antuérpia</EM>
  <EM ID="26" ENT="E21">Jacob Kruger</EM>
</ALT>

<EM ID="21" ENT="ARE8">Rabino</EM>
de
<EM ID="22" ENT="E74">Antuérpia</EM>
<EM ID="23" COREL="21 22" TIPOREL="ident vinculo_inst" ENT="E21">Jacob Kruger</EM>

```

Listing 5.9: Excerpt of an ALT case with a person's religious role. Textual expression reads: "Rabbi of Antwerp Jacob Kruger". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.

```

<ALT>
  <EM ID="102" COREL="40" TIPOREL="ident" ENT="E53">Centro Cultural de Belém</EM>
  |
  <EM ID="103" ENT="E53">Centro Cultural</EM> de <EM ID="104" COREL="40 103 5 0 3
    " TIPOREL="inclui inclui incluído sede_de sede_de" ENT="E53">Belém</EM></
  ALT>

<EM ID="91" COREL="37" TIPOREL="ident" ENT="E53">Centro Cultural de Belém</EM>

```

Listing 5.10: Excerpt of an ALT case with an entity identified as a place. Textual expression reads: "Belém Cultural Center". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.

```

<ALT>
  Às três horas e vinte minutos da tarde
  <EM ID="1" COREL="4" TIPOREL="data_nascimento" ENT="E52">de 13 de Junho de
    1888</EM>
  |
  Às três horas e vinte minutos
  <EM ID="2" COREL="4" TIPOREL="data_nascimento" ENT="E52">da tarde de 13 de
    Junho de 1888</EM>
</ALT>

Às três horas e vinte minutos da tarde
<EM ID="1" COREL="3" TIPOREL="data_nascimento" ENT="E52">de 13 de Junho de 1888</EM
>

```

Listing 5.11: Excerpt of an ALT case with a temporal expression. Textual expression reads: "At twenty minutes past three in the afternoon of June 13, 1888". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.

```

<ALT>
  <EM ID="7" ENT="ARE2">95 Teses de Martinho Lutero</EM>
  |
  <EM ID="8" ENT="ARE2">95 Teses</EM>
  de
  <EM ID="9" COREL="8 5 0" TIPOREL="autor_de natural_de participante_em" ENT="E21
    ">Martinho Lutero</EM>
</ALT>

<EM ID="6" ENT="ARE2">95 Teses</EM>
de
<EM ID="7" COREL="6 4 0" TIPOREL="autor_de natural_de participante_em" ENT="E21">
  Martinho Lutero</EM>

```

Listing 5.12: Excerpt of an ALT case with the title of a document and its author. Textual expression reads: "Ninety-five Theses by Martin Luther". The first expression corresponds to the processed Second HAREM GC still with ALT tags, and the second corresponds to the final output of the processing without ALT tags.

5.1.1.5 Annotation of relation descriptors

By analyzing Table 5.7 that characterizes the transformed Second HAREM corpus with annotated relations, we can conclude that the dataset is highly imbalanced. Moreover, the number of annotations for some relations is too small to efficiently train a deep learning model. To solve this

issue, we opted to train a machine learning model with a corpus of annotated relation descriptors following the work of Collovini et al. [18].

To build the corpora, we selected relation descriptors that have the biggest amount of explicit relation annotations. Therefore, we only considered relations in the transformed corpus that have more than 150 annotations. However, we excluded the relations of identity and inclusion, because the majority of the annotations were either implicit relations or relations that could not be inferred from the text. In sum, we selected the entities Group, Person, Place, and Title with the relations "autor_de / obra_de", "ocorre_em / sede_de", and "vinculo_inst".

We developed a semi-automatic approach where we automatically selected relations between entities that were in the same sentence and manually annotated the relation descriptor. Following the work of Li et al. [54], we replaced multi-token named entities with ARG1 and ARG2. For each pair of entities, ARG1 is treated as a Title and ARG2 is treated as a Person with the authorship relation, ARG1 is treated as a Person and ARG2 is treated as an Organization with affiliation relationships, and ARG1 is treated as an Organization and ARG2 is treated as a Place in placement relationships. In cases where sentences had multiple triples, we either split the sentence into multiple fragments to include all the relations or we grouped the named entities together as one. List 5.13 is an example of a manual annotation of the relation descriptor "professor da" (professor of) that expresses an affiliation relationship between a person (Hugo Doménech) and a group (Universidade Jaume de Castellón).

```
<EM ID="8" ENT="E21">Hugo Doménech</EM>,
<EM ID="9" COREL="8" TIPOREL="ident" ENT="ARE8">professor</EM> da
<EM ID="10" COREL="8" TIPOREL="vinculo_inst" ENT="E74">Universidade Jaume de
  Castellón</EM>

ARG1, <relation>professor da</relation> ARG2
```

Listing 5.13: Example of manual relation annotation of relation descriptors between named entities from the transformed Second HAREM corpus.

Collovini et al. [18] created corpora³ of explicitly annotated triples with the relations of affiliation and placement, i.e., any kind of relationship between the entities Organization and Person, and Organization and Place, respectively. We joined their corpora to ours by adding triples that were missing from our corpora.

5.1.2 Dataset Parsing

We used the Flair NLP library [2] to train NER models with the result of the transformation of the HAREM corpora. As it is, Flair requires that the input dataset is in the CoNLL format. The CoNLL format consists of a TSV file where each line represents a single word and its classification label. We used the BIO (Beginning, Inside and Outside) annotation scheme to label each token

³<https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/textos-e-listas-de-relacoes/>

in a chunk with a predicted tag according to the token's relation to a named entity. With the BIO scheme, each token is tagged with a label that represents if it is at the beginning (B), inside (I), or outside (O) of a named entity.

The process of transforming the corpora had the following steps: iterate the documents and perform sentence splitting and word tokenization; for each token, assign it the O symbol; for the named entity tags, if the content had a single-token, then set it as B; otherwise, assign the first token the symbol B, and the remaining ones the symbol I; finally, if the named entity tag has a tail, its content is tokenized and assigned the symbol O.

We will be using the previous excerpt from the corpus as an example of the parsing of the NER task: "Hugo Doménech, professor da Universidade Jaume de Castellón [...]". The sentence in English reads: "Hugo Doménech, professor at the University Jaume de Castellón [...]". Listing 5.14 shows the sentence in the CoNLL format for the NER classification task where we identify three entities: "Hugo Doménech" as a person (E21), "professor" as a role type (ARE8), and "Universidade Jaume de Castellón" as a group (E74).

```
Hugo B-E21
Doménech I-E21
, O
professor B-ARE8
da O
Universidade B-E74
Jaume I-E74
de I-E74
Castellón I-E74
[...]
```

Listing 5.14: Output of the CoNLL format for the NER classification task using the example sentence.

On the other hand, for the RE task we used a CRF. We used SpaCy to perform tokenization, part-of-speech tagging, and dependency parsing for each sentence with the pre-trained *pt_core_news_lg* model. The output of the process is a file where each line represents a single word with the following fields: a word, a POS tag, a dependency tag, and a relation classification label. Listing 5.15 shows the output for the previously used sentence fragment where we identify a relation of affiliation between the person "Hugo Doménech" (ARG1) and group "Universidade Jaume de Castellón" (ARG2) with the relation descriptor "professor da".

```

610 ARG1  PROPN nsubj O
611 ,      PUNCT punct O
612 professor NOUN appos I-relation
613 da     ADP  case  I-relation
614 ARG2  PROPN nmod  O
[...]

```

Listing 5.15: Output of the corpus with POS tagging and dependency parsing with SpaCy for the RE classification task with the affiliation relation (vinc_inst).

5.1.3 Characterization of Transformed Corpora

The result of the transformation of the HAREM corpora is characterized in Table 5.5 along with the original HAREM corpora not considering inverse relations.

Table 5.5: Characterization of original and transformed HAREM corpora.

Corpora	# tokens	# entities	# entity categories	# relations	# relation categories
HAREM I	165,582	5,065	10	-	-
HAREM II	98,499	7,847	10	4,803	24
Transformed HAREM I	165,582	4,208	8	-	-
Transformed HAREM II	98,499	5,880	8	3,554	13

Table 5.6 shows the distribution of recognized named entities in the transformed HAREM corpora. Table 5.7 shows the distribution of relations between recognized named entities in transformed Second HAREM corpus. Table 5.8 characterizes the three corpora with explicit relations between the entities Person and Title (authorship), Group and Place (affiliation), and Person and Place (placement). The number of relation instances is similar to that of the work of Collovini et al. [18] that had 61 positive affiliation relations and 40 positive placement relations.

5.2 Named Entity Recognition

In this experiment, we trained and tested two BiLSTM-CRF Neural Network models with Flair Embeddings [3] for entity classification. We used the Flair NLP framework with the following pre-trained language models: Flair Embeddings Backwards and Forward (FlairBBP)⁴, and a Flair Embedding⁵ for Portuguese made available by the library that we will refer to in this work as the default.

We set the training hyperparameters from the similar experiment presented in Santos et al. [83] to evade the computationally expensive cost of optimizing hyperparameters. Table 5.9 shows the

⁴<https://github.com/jneto04/ner-pt>

⁵https://github.com/ericclief/language_models

Table 5.6: Distribution of recognized named entities in transformed HAREM corpora.

Entities	# entities	
	Transformed HAREM I	Transformed HAREM II
Formal Title	200	435
Role Type	89	160
Event	129	281
Person	833	1,583
Time-Span	379	603
Place	1,160	1,213
Dimension	350	229
Group	1,056	1,376
Total	4,196	5,880

Table 5.7: Distribution of relations between recognized named entities in transformed Second HAREM corpus.

Relations	# relations
autor_de / obra_de (authorship)	152
data_nascimento (date of birth)	5
data_morte (date of death)	10
datado_de / data_de (dated from / date of)	93
ident (identity)	1,777
inclui / incluido (includes / included by)	717
local_morte (place of death)	4
natural_de / local_nascimento_de (place of birth)	116
ocorre_em / sede_de (placement)	195
participante_em / ter_participacao_de (participation in)	142
periodo_vida (lifetime)	5
produtor_de / produzido_por (producer of / produced by)	39
residente_em / residencia_de (place of residence)	15
vinculo_inst (affiliation)	289
Total	3,554

Table 5.8: Distribution of explicit relations in corpora annotated with relations between the entities Group, Person, Place, and Title.

Relations	# relations
authorship	51
affiliation	89
placement	54

training hyperparameters for the NER task. We set the initial learning rate to 0.1 with an annealing factor of 0.5 on every three epochs without improvement. The training was stopped after 150 epochs or if the learning rate was below 0.0001.

Table 5.9: Hyperparameter settings for named entity classification.

Parameter	Value
Batch size	32
Hidden size	256
Learning rate	0.1
Training epochs	150
Optimizer	SGDW

5.3 Relation Extraction

We trained and tested three CRF classifiers for relations between named entities of Group, Person, Place, and Title. We applied a linear-chain CRF using the Python library *sklearn_crfsuite*⁶. We considered two labels: I-relation and O, where a word labeled with I-relation is inside of a relation descriptor, and a word labeled with O is outside of a relation descriptor. We used the following set of features: the current word and its POS and dependency tags, the succeeding word at position (+1) and its POS and dependency tags, and the preceding word at position (-1) and its POS and dependency tags.

For the training, we used the default algorithm, gradient descent with Limited-memory BFGS (L-BFGS), with Elastic Net (L1 + L2) regularization of $c1=0.1$ and $c2=0.01$.

5.4 Evaluation

We formulated the tasks of Named Entity Recognition and Relation Extraction as classification tasks. For this reason, we used metrics like Precision, Recall, and F1-score to evaluate the tasks' performance. The metrics for the NER and RE tasks are presented in Equations 5.1, 5.2, and 5.3.

$$Precision = \frac{\text{number of correctly identified entities/relations}}{\text{total number of identified entities/relations}} \quad (5.1)$$

$$Recall = \frac{\text{number of correctly identified entities/relations}}{\text{actual number of identified entities/relations}} \quad (5.2)$$

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (5.3)$$

⁶<https://sklearn-crfsuite.readthedocs.io/en/latest/>

5.4.1 Results

For the evaluation of the NER task, we used the corpora characterized in Table 5.6: the transformed First HAREM corpus as the testing set and the transformed Second HAREM corpus as the training set. We included evaluation types based on CoNLL with the Python library *seqval* [64] and SemEval-2013 with the Python library *nerevaluate*⁷. Lignos et al. [55] recommended using *seqval* on NER evaluation after assuring that it produces the same BIO results as *conlleval*. It has three main advantages: supports BIO and BIOES encoding, provides greater numerical precision than *conlleval* by not truncating scores, and can be called directly from Python. We also included an entity-type evaluation where an entity is considered correct if the type of entity is correct regardless of the BIO labeling.

As for the evaluation of the RE task, due to the reduced data size of the corpora we described in Table 5.8, we opted to apply cross-validation. We performed 5-fold cross-validation on the transformed corpora that identify authorship, affiliation, and placement relations.

5.4.1.1 Named Entity Recognition

We evaluated the BiLSTM-CRF model with Flair Embeddings available on the NLP library (default) and FlairBBP. Table 5.10 presents the results obtained for the NER models for each entity and overall.

Table 5.10: CoNLL evaluation results of BiLSTM-CRF using default Flair Embeddings and FlairBBP.

Entities	Default			FlairBBP		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Dimension	63.98%	33.33%	43.83%	49.36%	46.10%	47.67%
Event	39.42%	26.28%	31.54%	39.37%	40.38%	39.87%
Group	69.20%	55.43%	61.55%	70.05%	69.56%	69.80%
Person	72.57%	72.24%	72.41%	72.23%	79.22%	75.56%
Place	75.80%	81.21%	78.41%	81.67%	87.00%	84.25%
Role	21.19%	27.17%	23.81%	37.04%	54.35%	44.05%
Time-Span	15.09%	13.42%	14.20%	14.08%	12.13%	13.03%
Title	44.85%	22.85%	30.27%	39.18%	28.46%	32.97%
Overall	65.36%	57.92%	61.42%	65.42%	66.42%	65.92%

Bold values indicate an overall better performance on the transformed First HAREM corpus.

It is not possible to directly compare the results we obtained with other research works that used the HAREM corpora for the NER task. The main reason is that we adapted the corpora by modifying the classifications of named entities. Another is that most state-of-the-art works that apply supervised learning methods to NER and adopt the HAREM collections as training and

⁷<https://github.com/MantisAI/nerevaluate>

testing corpora use the First HAREM corpus as the training set and MiniHAREM as the testing set.

The BiLSTM-CRF model with FlairBBP embeddings performed better than the default Flair Embeddings on nearly all named entity categories with an overall improvement of 4.5% on the F1-score. We performed a one-way analysis of variance (ANOVA) on the two models with the F1-score measure. Our goal was to determine if the variance between the Flair embeddings and the labeling was statistically significant. The results show that there is not a significant difference between the two models. The difference between the means of the two models is 0.06 and the adjusted p-value for the multiple comparisons is 0.6.

Predictably, the entities with the best results were E21 Person, E53 Place, and E74 Group, which are the entities with the most annotations in the corpora. In contrast, the performance of recognizing the entity E52 Time-Span on the transformed First HAREM corpus was unexpectedly low with an F1-score of 13.03%.

To better understand the reason for the less good outcomes, we used the *nervaluate* library to perform a more in-depth analysis. We focused our analysis on entity-type matching, i.e., when there are mismatches on the boundaries of the BIO tagging of the prediction with the ground truth, but the entity type is correct. Table 5.11 shows the evaluation of the NER models with entity-type matching.

Table 5.11: Evaluation results of BiLSTM-CRF using default Flair Embeddings and FlairBBP with entity-type matching.

Entities	Default			FlairBBP		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Dimension	86.17%	44.89%	59.03%	75.12%	68.91%	71.89%
Event	42.47%	39.74%	41.06%	43.32%	51.92%	47.23%
Group	70.30%	61.90%	65.84%	70.79%	73.77%	72.25%
Person	89.90%	85.45%	87.62%	87.08%	91.52%	89.25%
Place	85.59%	85.17%	85.38%	85.65%	88.47%	87.03%
Role	26.71%	42.39%	32.77%	40.38%	68.48%	50.81%
Time-Span	97.13%	80.52%	88.05%	90.65%	82.17%	86.20%
Title	41.81%	27.72%	33.33%	48.27%	36.70%	41.70%
Overall	79.89%	70.79%	75.07%	78.19%	79.40%	78.79%

Bold values indicate an overall better performance on the transformed First HAREM corpus.

The results illustrate that the main problem of our models for recognizing the entity E52 Time-Span is boundary errors. The previous F1-score of 13.03% with FlairBBP embeddings improves to 86.20% with entity-type matching. However, after we performed a qualitative analysis of the ground truth and the predictions for the temporal entity, we believe the issue is not with the model itself but rather with differences between the training and testing corpora’s annotations. The Second HAREM annotations for temporal expressions include adverbs that the First HAREM leaves

out. For example, for the date "em Junho de 1992" (in June of 1992), the First HAREM classifies the entity as "Junho de 1992", while the Second Harem classifies the entity as "em Junho de 1992". This issue did not arise in other research that used the HAREM corpora to train a NER model that achieved good results, i.e., the First HAREM and MiniHAREM, because these corpora were annotated in a similar style.

5.4.1.2 Relation Extraction

The results of the 5-fold cross-validation of relation extraction in corpora with annotations of affiliation, authorship, and placement are shown in Table 5.12.

Table 5.12: Results of 5-folds cross validation for relation extraction evaluation in corpora with annotations of affiliation, authorship, and placement.

Relations	Precision	Recall	F1-score
affiliation	64.96%	45.07%	52.95%
authorship	71.13%	47.64%	55.00%
placement	77.51%	60.51%	66.20%

The RE evaluation results are in-line with those of Collovini et al. [18] that used a CRF classifier with relation-specific features for Portuguese. They obtained an F1-score 3% higher on affiliation relationships and 9% lower on placement relationships. The table of results from the article is shown in Table 5.13.

Table 5.13: Results of relation extraction evaluation in Collovini et al. [18] corpora⁸ with annotations of affiliation and placement.

Relations	Precision	Recall	F1-score
affiliation	65%	49%	56%
placement	73%	47%	57%

5.5 Discussion

The entity-type NER evaluation results for temporal expressions suggest that further normalization of the HAREM corpora is necessary. Not only to improve results but also to potentially filter irrelevant information in named entities in the context of cultural heritage description. Currently, we filter the content of recognized entities in the ontology population module. For example, temporal expressions are represented as a DateTime structure in ArchOnto through the DataObject ontology so that we filter words in the recognized entities that do not correspond to a date, and we do not represent expressions that only have the information of the date's day since we cannot narrow down a temporal range.

We can conclude that there are no significant differences between applying FlairBBP and the Flair Embeddings made available by the FlairNLP library with a BiLSTM-CRF on the corpora we transformed. However, in our work, we will be using FlairBBP as it is the most cited in other bodies of work in Portuguese.

The RE models we trained produced results that are in line with other works in Portuguese. However, this methodology does not allow us to extract implicit relations between named entities. Combined with the reduced size of available annotated data, this limits the potential of extracting other relations that we can represent with ArchOnto.

Nevertheless, it would be interesting to perform a more in-depth analysis of the impact of using other features, such as word embeddings, for the Portuguese language on the RE task. Other word embeddings could also be trained and tested for the NER task using the Flair NLP framework.

Chapter 6

Ontology Population

In this chapter, we will be describing the implementation of the automatic ontology population using ontology-oriented programming given the ontology-based information extraction models we trained, and evaluating the results against a golden standard built with the collaboration of archival professionals.

6.1 Implementation

We developed the ontology population module with ontology-oriented programming through the Python module Owlready2 [51]. As the library does not have a feature that deals with the redirection of imports when loading an ontology, we implemented a parser according to the XML Catalog standard used by Protégé.

The identified entities resulting from the ontology-based information extraction process were stored in JSON files. The data is stored in key:value pairs with ArchOnto classes as the keys and arrays of the identified entities as the values. We removed duplicate values within each class. As the structured data cannot be directly exported into the ontology as instances, we defined mapping rules to create individuals and property assertions from extracted entities.

The following naming conventions were applied to instances:

- instances use camel case (e.g., descriptiveDate1);
- concepts are instantiated with their type (event, person, descriptiveDate, place, dimension, group, title, or role) followed by an auto-incremented identifier (e.g., place4);
- an appellation of a term is instantiated with the term’s type, followed by the term “name” and an auto-incremented identifier (e.g., personName1);
- instances from the CIDOC-CRM and ArchOnto extension ontologies are linked to the DataObject ontology with instances that add an underscore (e.g., title1_);

- instances of appellations and DataObject entities related to temporal entities are instantiated as the date value in the format yyyy-mm-dd (e.g., 1910-11-30). Instances might only include a year or the year and the month (e.g., 1910 or 1910-11).

The identified concepts were directly instantiated as classes of the CIDOC-CRM and ArchOnto extension ontologies following the previous naming conventions. We validated the content of the extracted entities through DataObject ontology classes. Except for measurement units and temporal expressions, data property assertions were performed directly. We filtered and validated temporal entities to extract valid information about the date with the Python *datetime* module (e.g., E52 entities without a year are invalid). As our NER model does not extract information on time (hours, minutes, seconds) and most temporal expressions do not contain that information, all temporal entities were instantiated as DOE11 Interval with a start date "00:00:00" and end date "23:59:59". We also filtered dimensional entities that contain measurement units. The content of the entities was split through regular expressions into a numeral section (quantity) followed by a textual section (measurement unit). We instantiated the numerical part as an E54 Dimension and the textual part as an E58 Measurement. Entity E58 Measurement is the only entity we do not validate through the DataObject ontology to avoid redundancy. Table 6.1 presents the mapping of extracted concepts to property assertions in ArchOnto.

Furthermore, we defined mapping rules to create individuals and property assertions from extracted relations between named entities. Table 6.2 presents the mapping of extracted relations between recognized named entities to property assertions in ArchOnto.

6.2 Evaluation

The evaluation measures of ontology learning approaches are split into the ontology layers they assess due to the structural complicity of ontologies as knowledge bases. Due to the use of the DataObject ontology in ArchOnto to validate literal values, we decided to implement a hybrid approach to hierarchy evaluation that performs a lexical evaluation on data property values to decide on correctly matched instances. The lexical evaluation is based on Levenshtein edit distance, where two instances of classes belonging to the DataObject ontology are assigned a similarity score. We use the measure of character accuracy formulated in Equation 4.1 to allow us the use of threshold percentage values as a way to determine whether two instances are similar enough to be considered matched. Thus, above a threshold of 50%, two instances from the ground truth and our system are considered matched.

The evaluation adopted a bottom-top approach where the most specific concepts that contain data property assertions were compared and used to trace back to less specific individuals connected to them. This evaluation only contemplates relations of identity, "P1_is_identified_by" and "L2DO_hasValue", and the numerical representation of a dimension, "P90_has_value" and 'P91_has_unit".

The evaluation uses the Augmented Precision, Recall, and F1-score measures presented in Section 2.3.3. We developed a Python implementation of the evaluation script based on GATE's

Table 6.1: Mapping of extracted concepts to property assertions in ArchOnto.

Extracted Concepts	Property Assertions
E5	event[id] (<i>E5_Event</i>), 'P1_is_identified_by', eventName[id] (<i>E41_Appellation</i>)
	eventName[id] (<i>E41_Appellation</i>), 'L2DO_hasValue', eventName[id]_ (<i>DOE8_String</i>)
	eventName[id]_ (<i>DOE8_String</i>), 'DOP7_stringValue', value (<i>xsd:string</i>)
E21	person[id] (<i>E21_Person</i>), 'P1_is_identified_by', personName[id] (<i>E41_Appellation</i>)
	personName[id] (<i>E41_Appellation</i>), 'L2DO_hasValue', personName[id]_ (<i>DO17_PersonName</i>)
	personName[id]_ (<i>DO17_PersonName</i>), 'DOP5_name', value (<i>xsd:string</i>)
E52	descriptiveDate[id] (<i>E52_Time-Span</i>), 'P1_is_identified_by', yyyy-mm-dds (<i>E41_Appellation</i>)
	yyyy-mm-dd (<i>E41_Appellation</i>), 'L2DO_hasValue', yyyy-mm-dd_ (<i>DOE11_Interval</i>)
	yyyy-mm-dd_ (<i>DOE11_Interval</i>), 'DOP6_startDateValue', value (<i>xsd:dateTime</i>)
	yyyy-mm-dd_ (<i>DOE11_Interval</i>), 'DOP2_endDateValue', value (<i>xsd:dateTime</i>)
E53	place[id] (<i>E53_Place</i>), 'P1_is_identified_by', placeName[id] (<i>E41_Appellation</i>)
	placeName[id] (<i>E41_Appellation</i>), 'L2DO_hasValue', placeName[id]_ (<i>DOE8_String</i>)
	placeName[id]_ (<i>DOE8_String</i>), 'DOP7_stringValue', value (<i>xsd:string</i>)
E54	dimension[id] (<i>E54_Dimension</i>), 'P90_has_value', value (<i>xsd:string</i>)
	dimension[id] (<i>E54_Dimension</i>), 'P91_has_unit', value (<i>E58_Measurement_Unit</i>)
E74	group[id] (<i>E74_Group</i>), 'P1_is_identified_by', groupName[id] (<i>E41_Appellation</i>)
	groupName[id] (<i>E41_Appellation</i>), 'L2DO_hasValue', groupName[id]_ (<i>DOE8_String</i>)
	groupName[id]_ (<i>DOE8_String</i>), 'DOP7_stringValue', value (<i>xsd:string</i>)
ARE2	title[id] (<i>ARE2FormalTitle</i>), 'L2DO_hasValue', title[id]_ (<i>DOE8_String</i>)
	title[id]_ (<i>DOE8_String</i>), 'DOP7_stringValue', value (<i>xsd:string</i>)
ARE8	roleType[id] (<i>ARE8_RoleType</i>), 'L2DO_hasValue', roleType[id]_ (<i>DOE8_String</i>)
	roleType[id]_ (<i>DOE8_String</i>), 'DOP7_stringValue', value (<i>xsd:string</i>)

Table 6.2: Mapping of extracted relations between recognized named entities to property assertions in ArchOnto.

Extracted Relations	Entities	Property Assertions
authorship	E21 & ARE2	pc14_[pc14Id] (<i>PC14_Carried_Out_By</i>), 'P14.1_in_the_role_of', roleType[roleId] (<i>ARE8_RoleType</i>)
		pc14_[pc14Id] (<i>PC14_Carried_Out_By</i>), 'P01_has_domain', creation[creationId] (<i>E65_Creation</i>)
		pc14_[pc14Id] (<i>PC14_Carried_Out_By</i>), 'P02_has_range', person[personId] (<i>E21_Person</i>)
		roleType[roleId] (<i>ARE8_RoleType</i>), 'L2DO_hasValue', roleType[roleId]_ (<i>DOE8_String</i>)
		roleType[roleId]_ (<i>DOE8_String</i>), 'DOP7_stringValue', "Autor" (xsd:string)
		document[docId] (<i>E31_Document</i>), 'P102_has_title', title[titleId] (<i>ARE2_FormalTitle</i>)
affiliation	E74 & E21	document[docId] (<i>E31_Document</i>), 'P94_was_created_by', creation[creationId] (<i>E65_Creation</i>)
		group[groupId] (<i>E74_Group</i>), 'P107_has_current_or_former_member', person[personId] (<i>E21_Person</i>)
		pc14[pc14Id] (<i>PC14_Carried_Out_By</i>), 'P01_has_domain', activity[activityId] (<i>E7_Activity</i>)
		pc14[pc14Id] (<i>PC14_Carried_Out_By</i>), 'P02_has_range', person[personId] (<i>E21_Person</i>)
		pc14[pc14Id] (<i>PC14_Carried_Out_By</i>), 'P14.1_in_the_role_of', roleType[roleId] (<i>ARE8_RoleType</i>)
placement	E74 & E53	group[groupId] (<i>E74_Group</i>), 'P107_has_current_or_former_member', person[personId] (<i>E21_Person</i>)
		group[groupId] (<i>E74_Group</i>), 'P74_has_current_or_former_residence', place[placeId] (<i>E53_Place</i>)

Java implementation. Our evaluation only considers entities from the ground truth that belong to classes that can be extracted with our system to produce more meaningful results.

6.2.1 Consensual Dataset of ArchOnto Representations

The evaluation of our ontology-based information extraction and ontology population approaches requires a dataset as a golden standard to compare to the output of our system. Hence, we have developed a dataset of 13 consensual ArchOnto representations of Portuguese Historical Archival Records based on their digital representations. The archival records were selected from a previously created dataset¹ of typewritten digital representations of Portuguese records from the 20th century extracted from the Arquivo Nacional da Torre do Tombo.

We requested the help of two archivists from Direção-Geral do Livro, dos Arquivos e das Bibliotecas (DGLAB) to perform the ArchOnto descriptions. The main criteria for the description is that it must be solely focused on textual content visible in the digital representations. The output of the descriptions is a set of textual files with rows of triples of relationships between two individuals or between individuals and data values. After thoroughly analyzing the differences between the two descriptions, we held a meeting on 2022/11/17 to uniformize the description of similar concepts to reach a consensual description for the 13 records.

Upon consensus, the textual descriptions were transformed to OWL format. We removed information that could not be extracted from the digital representations and information that could only be extracted from handwritten sections of the representations (e.g., signatures and annotations).

The records described in the dataset are accessible through their reference code in Digitalq. We will be referencing them in this article through the numbering listed in Table 6.3.

6.2.2 Results

This section showcases the evaluation of the ontology population on a sample of digital representations of 13 Portuguese archival records.

We employed the trained NER and RE models to the dataset of digital representations of 13 Portuguese archival records. We split the analysis of the annotations with OCR'ed text into two scenarios: selective and total. The selective scenario includes digital representations classified into letters and documents' covers, while the total scenario includes every digital representation not classified as blank in its writing type classification. We chose these two scenarios because some archival records possess hundreds of digital representations, which makes it important to figure out which document typologies contain information that is relevant to archival description, not only from the perspective of improving the utility of our system to produce description suggestions that archivists find pertinent as an output but also to decrease the potential computational complexity of applying OCR, information extraction models and ontology population to a large body of digital representations in real-time. Furthermore, we also included manual transcriptions

¹<https://rdm.inesctec.pt/dataset/cs-2022-004>

Table 6.3: Numbering of dataset’s archival record and corresponding Digitarq URL.

#	Reference Code	Digitarq URL
1	PT/PNA/DGFP/0001-002/0015/00053/ 00002	https://digitarq.arquivos.pt/ details?id=4684059
2	PT/PNA/DGFP/0002/0001/00021/ 00005	https://digitarq.arquivos.pt/ details?id=4684160
3	PT/PNA/DGFP/0003/0002/00005/ 00003	https://digitarq.arquivos.pt/ details?id=4684697
4	PT/PNA/DGFP/0005/0002/00036/ 00011	https://digitarq.arquivos.pt/ details?id=4685629
5	PT/PNA/DGFP/0006/0001/00029	https://digitarq.arquivos.pt/ details?id=4684945
6	PT/TT/SNI-DGE/1/5583	https://digitarq.arquivos.pt/ details?id=4319971
7	PT/TT/SNI-DGE/3/5	https://digitarq.arquivos.pt/ details?id=4323545
8	PT/TT/SNI-DSC/13/812/A-2-12	https://digitarq.arquivos.pt/ details?id=4642954
9	PT/TT/SNI-DSC/19/81	https://digitarq.arquivos.pt/ details?id=4330356
10	PT/TT/SNI-DSC/22/930	https://digitarq.arquivos.pt/ details?id=4331332
11	PT/TT/SNI-DSC/24/203	https://digitarq.arquivos.pt/ details?id=4331528
12	PT/TT/SNI-DSC/35/3/00632	https://digitarq.arquivos.pt/ details?id=6030720
13	PT/TT/SNI-DSC/9/241	https://digitarq.arquivos.pt/ details?id=4327291

of the dataset to understand how OCR errors may impact the Ontology-Based Information Extraction pipeline and, consequently, the ontology population process. Margarida Falcão provided us with transcriptions [30] of letters and documents’ covers from the sample of typewritten digital representations of records used in the ontology-based extraction evaluation.

We successfully extracted entities from the OCR’ed content of the dataset’s digital representations on the two scenarios and from manual transcriptions. The only archival record that did not produce results was record 9 with the sample of manual transcriptions. This record only possesses the manual transcription of a process cover.

For future work, it would be interesting to expand the sample of typewritten digital representations to include documents of other genres. We should also expand the corpora of relation annotations to include a wider variety of documents' genres and domain-specific relationships (e.g., transfer of custody).

In contrast, while the CRF models successfully extracted relation descriptors, they were not able to extract relations between named entities within a sentence with this dataset. There might be various reasons for this. First of all, the classifiers might be biased toward the corpora they were trained on as there is a small number of relation descriptors. Moreover, most of the corpora's texts are extracted from news articles. Naturally, news articles have a different linguistic requirement to letters and reports, which are the genre of the majority of the documents present in our sample. For example, structured reports and documents' covers tend to not have any syntactical relation between named entities. Most relationships in the dataset are either implicit or not expressed in the same sentence. We give examples of identified relation descriptors with the affiliation relationship between an E21 and an E74 in Listing 6.1, with the authorship relationship between an ARE2 and an E21 in Listing 6.2, and the placement relationship between an E74 and an E54 in Listing 6.3.

```
Ao Exmo. <relation>Presidente</relation> da <E74>Comissão</E74> d'Arrolamento dos
  Bens
do Palacio das Necessidades.
```

Listing 6.1: Example of affiliation relationship extracted from a letter in the dataset.

Listing 6.1 shows the phrase fragment "To Hon. Chairman of the Commission for the Inventory of Assets/ of the Palace of Necessidades.". The NER model identifies President ("Presidente") as an ARE8 entity and Commission ("Comissão") as an E74 entity, while the RE model identifies President ("Presidente") as the relation descriptor. However, the sentence does not contain an explicit relation between a Person and an Organization.

```
LA MENACE ROUGE
por <E21>Jacques Valdour</E21>
O <relation>autor</relation>,observando a vida dos operarios[...]
```

Listing 6.2: Example of authorship relationship extracted from a structured report in the dataset.

Listing 6.2 shows the phrase fragment "LA MENACE ROUGE/ by Jacques Valdour/ The author, observing the life of workers [...]". The NER model identifies Jacques Valdour as an E21 entity, but it does not extract the title "LA MENACE ROUGE", which is in a different language than Portuguese. The RE model identifies the descriptor author ("autor"). Had the NER model correctly identified the book's title, the correct relation descriptor to extract would be "by" ("por").

```
Delegação PORTO <relation>Distrito</relation> <E54>PORTO</E54>
```

Listing 6.3: Example of placement relationship extracted from a structured report in the dataset.

Listing 6.1 shows the phrase fragment "Delegation PORTO District PORTO". The NER model identifies PORTO as an E53 entity, while the RE model identifies District ("Distrito") as the relation descriptor. There is no E74 entity identified so we cannot form a triple. Moreover, the relation we could extract in this sentence should be between the Delegation ("Delegação") and PORTO, and not with the descriptor District.

Therefore, we only used the extracted entities in our evaluation of the ontology population, as it was not possible to extract relationships between the recognized entities.

We performed a comparison between the ontologies of the ground truth and our system considering the selective and total scenarios with the OCR'ed results of the dataset of digital representations and their transcriptions. Table 6.4 presents the performance results.

The evaluation shows that, as expected, overall the ontology-based information extraction and population system has a better performance with manual textual transcriptions than with an OCR input. There are three records with an Augmented Precision score above 50%: record 2 with the selective scenario, record 3 with transcriptions, and record 12 with the selective scenario and transcriptions. The worst-performing results with the OCR input are records 3, 6, and 7. We manually analyzed the annotation and ontology population results and concluded that the bad results in record 3 and 6 on the selective and total scenarios stem from OCR errors. There are two main reasons that may explain the results for record 7. The first is the inability of our NER model to extract entities that are relevant to the archival description on digital representations that are highly informative yet contain less textual data, such as documents' covers. The second reason is that while our NER model extracts more information on records with textually rich digital representations, such as letters, this does not translate to good precision results since most of the extracted information is not deemed relevant to the archival description.

Taking record 1 as an example, Listing 6.4 shows the matching instances found between the ground truth and our system in the selective scenario with a similarity score. We can see that the group entity *group2* in the ground truth "Biblioteca da Ajuda" is matched to the group entity *group1* in our system "Bibliotheca da Ajuda", the person entity *person2* "José Duarte Ramalho Ortigão" is matched to the person entity *person3* in our system "JosS Duarte Ramalho", and the person entity *person3* "Costa" is matched to the place entity *place1* in our system "Costa .".

```
{PT-PNA-DGFP-0001-002-0015-00053-00002.groupName2_: [registro1.groupName1_, 0.95],
  PT-PNA-DGFP-0001-002-0015-00053-00002.personName2_: [registro1.personName3_,
    0.5263157894736843], PT-PNA-DGFP-0001-002-0015-00053-00002.personName3_: [
    registro1.placeName1_, 0.7142857142857143]}
```

Listing 6.4: Matching instances based on data property content for record 1.

6.3 Discussion

The object-oriented focus of OwlReady2 is a valuable asset for an easy way to build source code for the ontology population task where properties are treated as class attributes with dot notation.

Table 6.4: Performance results of ontology population evaluation on selective and total scenarios with OCR’ed dataset and their transcriptions.

Record	Selective Scenario			Total Scenario			Transcriptions		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
1	32.11%	20.57%	25.08%	29.50%	20.57%	24.24%	42.11%	28.07%	33.68%
2	55.56%	33.33%	41.67%	28.54%	20.57%	23.91%	39.98%	27.76%	32.77%
3	0.00%	0.00%	-	0.00%	0.00%	-	50.00%	28.57%	36.36%
4	33.33%	9.30%	14.55%	33.33%	9.30%	14.55%	26.18%	27.22%	26.69%
5	28.57%	13.33%	18.18%	28.57%	13.33%	18.18%	19.35%	10.34%	13.48%
6	0.00%	0.00%	-	0.00%	0.00%	-	16.67%	23.53%	19.51%
7	0.00%	0.00%	-	0.35%	47.06%	0.70%	0.00%	0.00%	-
8	9.69%	34.44%	15.13%	9.69%	34.44%	15.13%	0.00%	0.00%	-
9	2.20%	35.71%	4.15%	2.12%	35.71%	4.00%	-	-	-
10	9.80%	40.00%	15.75%	10.05%	40.00%	16.06%	11.00%	42.31%	17.46%
11	4.72%	29.41%	8.13%	1.87%	33.25%	3.53%	22.22%	13.79%	17.02%
12	53.33%	32.00%	40.00%	44.44%	18.18%	25.81%	66.67%	32.00%	43.24%
13	4.88%	30.51%	8.41%	3.08%	32.76%	5.63%	36.36%	25.23%	29.79%
Overall	18.02%	21.43%	19.10%	14.73%	23.48%	14.81%	27.55%	19.90%	27.00%

Precision = Augmented Precision; Recall = Augmented Recall; F1-score = Augmented F1-score
Bold values highlight best-performing evaluation values among the results for the selective scenario, total scenario and transcriptions for each record.

However, in our case, the ontology was previously built and had some clashes with the object-oriented paradigm. One such case is the CIDOC-CRM property "P14.1_in_the_role_of" which is not compatible with dot notation. Nevertheless, the OwlReady2 library has a feature that allows the addition of a Python alias to properties through the "python_name" annotation which allowed us to overcome this.

On the ontology population evaluation process, we adopted a bottom-top approach to the taxonomy where the most specific concepts that contained data property assertions were compared and used to trace back to less specific individuals connected to them. However, tracking assertions is not a trivial task with dot notation, requiring previous knowledge of the properties we want to access. This affects the abstraction of the evaluation implementation that needs domain-specific knowledge of ontologies to function.

Since ontology-based information extraction evaluation performed on ontologies is not often done in research works, the difficulty of the evaluation task is increased. Most works compare annotated textual data and not the output of the OBIE system itself, i.e., a populated ontology. This makes it difficult to compare results and improve the evaluation techniques themselves in the context of information extraction.

From this initial evaluation, we can see that document typologies greatly impacted the performance of Relation Extraction as the models we trained used similar document structures with a small number of annotations that cannot be successfully applied to this evaluation dataset. The capitalization of some documents might have also hindered the Information Extraction results, as, for example, the only document analyzed in the transcriptions scenario for record 8 contains capitalized information. For future work we should expand the sample of typewritten digital representations to include documents of other genres. We should also expand the corpora of relation annotations to include a wider variety of documents' genres and domain-specific relationships (e.g., transfer of custody).

We conclude that our ontology-based information extraction and population system performed better with manual textual transcriptions than with an OCR input. We can identify three causes for the bad results on the ontology population evaluation: persistence of OCR errors that hinder information extraction and lexical evaluation; inability of our NER model to extract relevant entities in document typologies that are highly informative yet contain less textual data; and the lack of relevance of some extracted information in document typologies that are textually rich.

It would be interesting to further this study to establish a correlation between document typologies and ontology population performance. It should allow us to explore the relevance of the extracted information in archival descriptions. For example, this evaluation dataset leads us to re-evaluate the relevance of extracting the E54 Dimension from digital representations. This is because the initial ArchOnto representations made by the archivists use this class to represent information that cannot be extracted from digital representations, such as digital files' sizes and the documents' number of pages. However, automatically judging whether a document is relevant or not to describe is difficult, as each historical document has value in itself and in relation to the aggregation of documents they are a part of. While some documents may stand out from the rest

if they contain information that possesses evidential or historical value, to make this assessment, archivists must have extensive knowledge on the fonds the records belong.

Chapter 7

User Interface for Description Values Suggestion

In the context of the EPISA project, previous works have researched and proposed design solutions that connect Linked Data to the archival context. One of those projects is a thesis by Guedes [39] that created the mockups and design patterns used to develop the project's platform.

The final phase of our proposed solution is the integration of the ontology-based information extraction pipeline into the EPISA platform. Our objective is to incorporate the functionality of suggesting descriptive metadata through the pipeline we conceived into the existing design solution of creating a record.

This chapter provides requirement specifications and non-functional mockups as the results of the design solution.

7.1 Requirements

Guedes [39] gathered the requirements for the individual view of descriptive records and entities in collaboration with one of the elements of the EPISA project and DGLAB. They are listed as follows:

- Create descriptive metadata that complies with the ArchOnto model but uses appropriate labels (e.g., natural language) in the interface;
- View all entities related to a resource;
- Filter the viewing of entities related to a resource (e.g., only view places).

The requirements included a graph view of semantic networks that we will disregard for visual simplification of the exploration of scenarios of our ontology-based information extraction

system. Our design implementation will display semantic networks as tables. To offer description suggestions in the creation of a record in the interface, we added the following requirements:

- Fill-in inputs of descriptive metadata from images associated with a record (digital representations);
- Validate suggested metadata (for example, accept or dismiss title suggestion);
- Alter suggested metadata.

7.2 Mockups

We focused on developing mockups for the functionalities of creating and accessing records. We created the mockups based on screens designed by Guedes [39] using Figma, an interface design tool. Figure C.1 shows the individual view of a created record entitled "Carta de sentimento do Papa Bento XIV ao rei D. José I sobre o terramoto" developed by Guedes [39].

We created the mockups using information extracted from our system with a transcription of a document's digital representations and the gold-truth description in ArchOnto created by archival professionals. However, to better exemplify the various possible uses of the UI, we added suggestions of relations between entities that we could not extract using the CRF models we trained.

7.2.1 Record Creation Screen

This section explores the feature of the creation of archival records in the platform. Figure 7.1 shows the screen where archival professionals can create the description of a record. Each description section has the following fields:

- Digital Representations: allows archivists to associate digital representations of documents to a record;
- Description Level: users can assign a description level to the record in the fond hierarchy (e.g., collections, fonds, sections, subfonds, series, installation units, and items) [69];
- Identifiers: internal identifiers (e.g., reference code or original numbering);
- Titles: succinct summary supplied by an archivist of the unit of description (supplied title) or explicitly stated in the unit of description's documentation (formal title);
- Existence: range of dates of the records' production;
- Dimensions: quantity, bulk or size of the unit of description [69];
- Supports: materials of physical artifacts (e.g., clay, paper, papyrus) [69];
- Languages: languages of the material;

- Access: state of access of record (e.g., available or unavailable).

ArchClient
Search
Hierarchy
Create ▼
Account

Create Record

Archival Description
Conservation & Restoration
Support Transfer

Digital Representations


Select a file or drag and drop here
Supported formats: JPEG, PNG, TIFF

Select File


PT-TT-SNI-DSC-24-203_m0001.tiff
Preview
232 KB


PT-TT-SNI-DSC-24-203_m0002.tiff
175 KB

Description Level
Identifiers
Titles
Existence
Dimensions
Supports
Languages
Access

Cancel
Create Record

© 2022 EPISA ArchClient

Figure 7.1: Individual view of the creation of a description record.

Our system will automatically fill in the Titles and Existence sections with suggestions of information extracted from the uploaded digital representations. Furthermore, the creation screen also has a section of a Semantic Network view that shows the identified entities and relations between them.

We will be using as a scenario the extraction and instantiation of information from a digital representation of a structured report from the Censorship Service Department (Direção Geral de Censura) about a non-periodic journal publication. Figure 7.2 shows the view of the digital representation on the platform. It contains the following extractable information: the publication's title,

"Petróleo no Mundo" (Oil Worldwide); the owner of the publication, a group called "Esso Standard Portugal, Inc"; the publication's director, "Fernando Faro Valadas"; the publication's editor, "Fernando Faro Valadas"; and the headquarters of the journal's redaction and administration "Rua Rosa Araújo, 43 - r/c Lisboa".

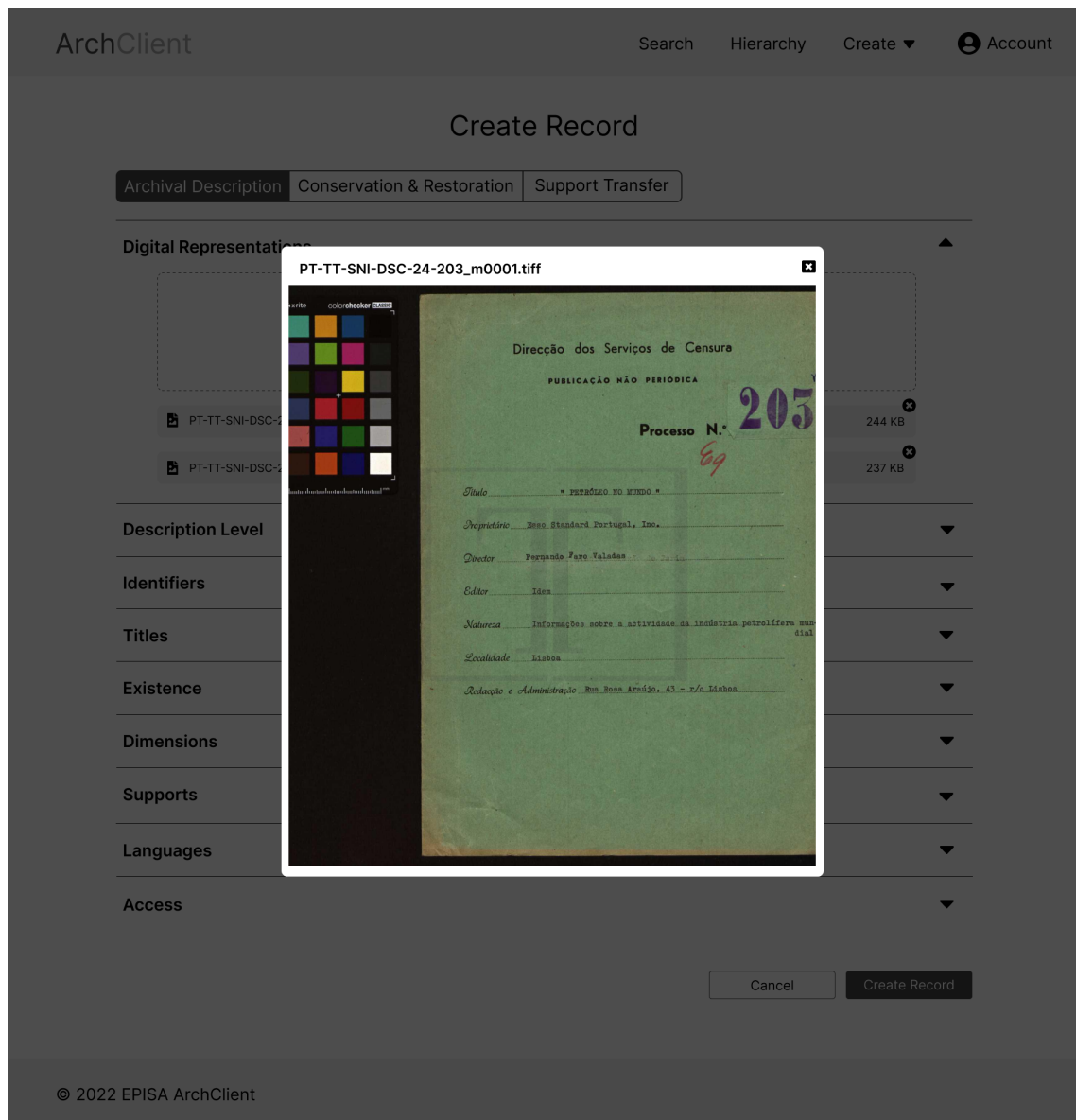


Figure 7.2: Individual view of a digital representation of a record's document uploaded to the UI.

Figure C.2 shows the result of our information extraction system with the suggested description values. The suggested value for the Title section is "PETRÓLEO NO MUNDO" as a formal title. The Semantic Network view shows the extracted entities: a group ("Esso Standard Portugal"), two people ("Fernando Faro Valadas" and an untitled person), and two roles ("director" and "editor").

7.2.1.1 Validation Scenario

The interface gives users the option to accept a suggestion or discard it with interactive buttons. Our validation scenario is the archivist not validating the system's suggestions. In this case, the platform presents a red warning above the invalidated fields, alerting the user to the need to either approve or discard them. Figure 7.3 shows an example of this scenario with the validation of the record's title.

Titles ▲

Title has not been validated.

Title

PETRÓLEO NO MUNDO

Title Type

Formal

✕ Discard

✓ Add

+ 🔍

Figure 7.3: Validation scenario: warning message.

7.2.1.2 Edit Information Scenario

Archival professionals might want to modify the suggestions to include contextual information that is more descriptive or to correct information. Figure 7.4 shows an example of this scenario with the modification of the suggestion for the record's title. The title was altered by an archival professional from "PETRÓLEO NO MUNDO" (Oil Worldwide) to "Processo relativo à publicação Petróleo no mundo" (Process regarding publication "Oil Worldwide") as a supplied title.

Titles ▲

Title

Processo relativo à publicação "Petróleo no mundo"

Title Type

Supplied

+ ✎ 🔍

Figure 7.4: Edit scenario: modified title field.

Users might also want to edit the semantic network section. For example, if we look at the suggestion for the role of editor in Figure C.2, it was associated with an unnamed entity person. However, the document identifies the editor as "Idem" (same author). Figure 7.5 shows the pop-up that allows users to edit the relation.

The user can either remove the current relationship and associate an existing entity or create a new entity, as seen in Figure 7.6. An archivist can correct this information by associating the

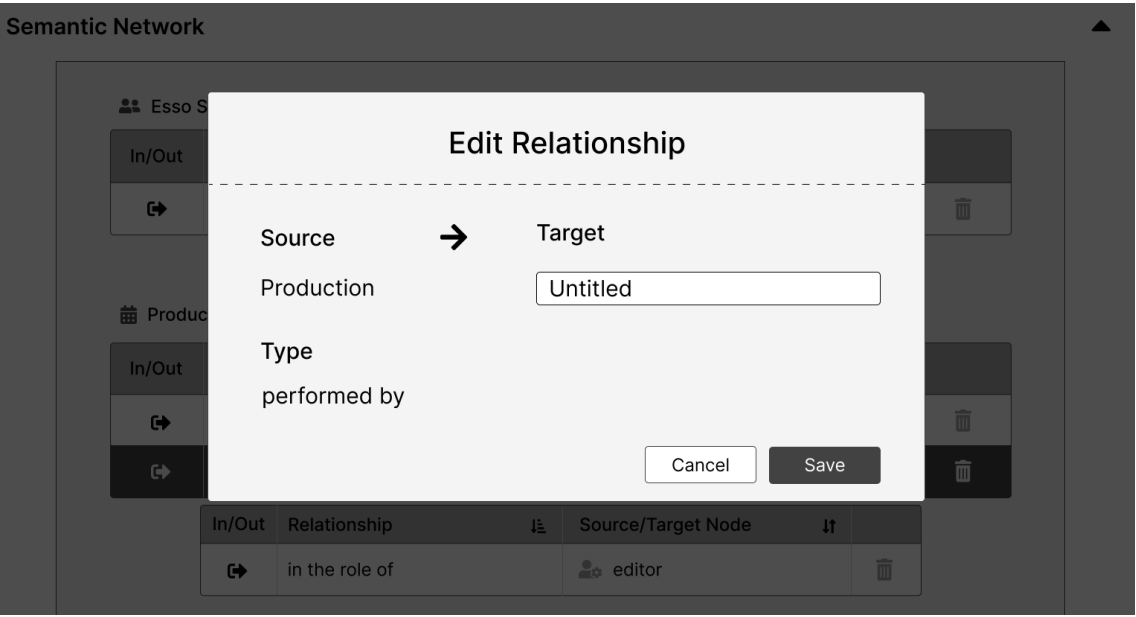


Figure 7.5: Edit scenario: edit a relation between two entities.

person entity Fernando to the role of editor in the interface by selecting the information’s row on the table. Figure 7.7 shows the semantic network section with the corrected information.

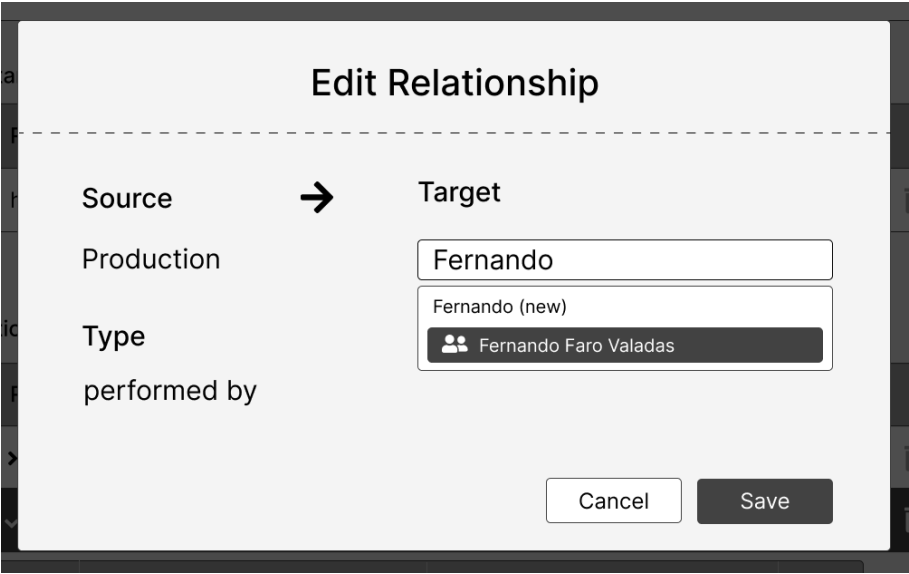


Figure 7.6: Edit scenario: edit a relation between two entities by associating the source entity with an existing entity.

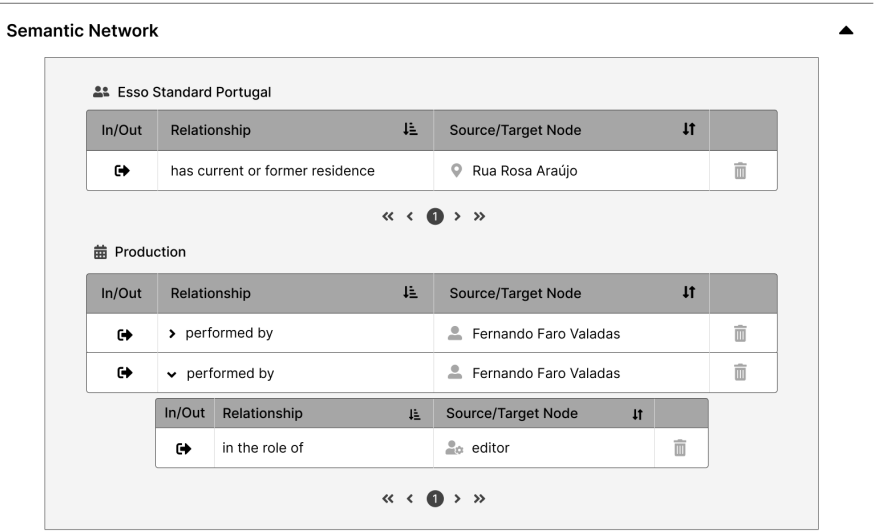


Figure 7.7: Edit scenario: semantic network section with the corrected relations between entities.

7.2.2 Record View Screen

This section explores the feature of viewing the created archival records in the platform. Figure C.3 shows the screen where archival professionals can view the created record’s description.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

This chapter presents the conclusions of this thesis regarding the work we developed and the results we obtained in the optimization of OCR with the parametrization of image processing algorithms, the training of ontology-based information extraction models, and the implementation of an ontology population approach. We will be describing our objectives, contributions, and work to be done for each module.

8.1.1 Optimization of Image Processing Algorithms for Character Recognition in Cultural Heritage Typewritten Documents

The main goal was to understand how to improve the OCR performance on cultural heritage digital representations using image processing algorithms. We evaluated the use and parameterization of image processing algorithms for text recognition tasks using Tesseract and OpenCV methods.

The key takeaway is that parametrization by digital representation typology benefits the employment of image pre-processing techniques in OCR, even if only for a small parametrization dataset. The analysis of the performance of image processing algorithms leads us to conclude that some algorithms improve the OCR performance significantly, while several tend to improve it and it is up to debate whether to apply them or not, and others even worsen the performance. On top of that, some algorithms have more impact in some samples of the evaluation dataset than others. A binarization algorithm positively impacted the OCR performance in the sample of theatre plays' covers. On the other hand, smoothing algorithms improved the performance of the overall dataset and the samples of letters and theatre plays' covers. Likewise, morphological transformation algorithms also improved the OCR performance of the overall dataset and the samples of theatre plays' covers. These findings suggest that the employment of image pre-processing algorithms in OCR might be more suitable for typologies where the text recognition task without pre-processing does

not produce good results, as is the case of the typology for theatre plays' covers. This argumentation leads us to question whether the OCR engine we are working with usually produces good results without image pre-processing in documents of this nature. Factually, the algorithms that have been proven to improve the OCR task compared to not applying any are Adaptive Thresholding, Bilateral Filter, Black Hat, Opening, and Top Hat. Having more information on the nature of the digital representations where this positive impact is registered could help us trace patterns between the results in these typologies and, for example, the image quality of the digital representations or the amount of text they contain, or the existence of stamps and annotations. Ultimately, there are big differences in the OCR performance by digital representation typology, which suggest that we should adopt the optimization of the OCR task for typologies that have weaker results without employing image pre-processing algorithms.

The outcomes of this work are the source code of the parametrization of image processing algorithms and two datasets. One is a cultural heritage dataset of typewritten Portuguese documents from the 20th century that includes digital representations of documents and a classification file that identifies the digital representations' typologies and writing format. The other is a dataset of manual transcriptions containing a sample of digital representations from the previous dataset and its corresponding transcriptions. Both datasets are available to consult for other scholars or researchers to use. Likewise, we will provide the source code of the parametrization of image processing algorithms using NSGA-II. We also submitted an article on this module.

8.1.2 Ontology-Based Information Extraction

The main goal was to develop an ontology-based information extraction system with a domain-specific ontology in Portuguese Cultural Heritage. We adapted existing annotated Portuguese corpora to the ArchOnto notation and used it to train Information Extraction models that extract entities and relations that can be mapped to ArchOnto.

The quality and size of the available annotated corpora greatly impacted our methodology choice for the information extraction process. There are still not many rich corpora in the Portuguese language for NER and RE, especially for relation extraction. The initial analysis of the performance of BiLSTM-CRF with two word embedding models leads us to conclude that the approach fell short of expectations in classifying temporal expressions. Further analysis applying an entity-type matching evaluation made us realize that the corpora we used for training and testing annotated temporal expressions in different ways, which lead to low results. The main takeaway is that there is not a significant difference between using a BiLSTM-CRF model with the available word embeddings in Portuguese from the FlairNLP library and the word embeddings FlairBBP published as Portuguese NER state-of-the-art. Moreover, the use of transformed corpora that only includes relevant information to the ArchOnto ontology did not seem to alter the performance results in comparison to the works and tools we followed.

The outcomes of this work are the transformed annotated Portuguese corpora, and the information extraction models we trained.

8.1.3 Ontology Population

The main goal was to represent structured data of extracted entities and relations in ArchOnto. We used the models we trained in the previous section to extract information. We evaluated the approach on a corpus of thirteen archival records manually represented in the ontology by two archival professionals based on their digital representations.

The main takeaway is that our ontology-based information extraction and population system performs better with manual textual transcriptions than with an OCR input. The analysis of the ontology population results leads us to identify three main problems that may have impacted the performance. The main problem with OCR'ed content is OCR errors that impact the information extraction performance and the lexical evaluation. The other two issues are related to the relevance of the extracted information in the context of archival description. Our NER model is unable to extract relevant entities in document typologies that are highly informative yet contain less textual data, such as documents' covers, and textually rich document typologies, such as letters, with a lot of extracted information lack relevance. However, the reduced size of our sample does not allow us to conclude on the impact of document typologies and their relevance to archival descriptions.

The outcomes of this work are a domain-specific approach for the automatic population of the ArchOnto ontology from textual data that applies information extraction techniques to identify and classify ontology instances and properties, a released dataset on the consensual ArchOnto representation of a cultural heritage dataset of thirteen Portuguese historical archival records based on textual information from digital representations, and the proposal of a methodology to evaluate ontology population based on the taxonomy evaluation from GATE's¹ Java implementation of the BDM score and lexical evaluation. We are also preparing to submit an article on the work developed on this module and on previous one.

8.2 Further Work

Future research will address the following aspects:

Optical Character Recognition (OCR) Module: other image processing algorithms and libraries should be evaluated and tuned. Optionally, the OCR engine should be trained with page analysis for different classified documents by using document analysis tools to manually segment images by regions where text is located. This step would improve the performance of the text recognition task for documents with complex layouts like inventories.

In future research, we would like to analyze the effect of the combination of image processing algorithms in the text recognition task. Furthermore, we would like to explore the detection and removal of stamps and annotations in digital representations.

Ontology-Based Information Extraction (OBIE) Module: there are two main goals in mind, expanding term extraction and performing triple extraction. For that, a new corpus from

¹<https://gate.ac.uk/>

the extracted dataset of digital representations from the 20th century should be manually annotated with high-level ArchOnto concepts and relations to be used as a gold standard. This would allow us to satisfactorily train a RE system with our current approach of using CRFs or, given a large enough amount of annotated relations, we could apply a deep learning RE model. Moreover, we could also explore domain-specific relations that are not possible to obtain using supervised learning with the available corpora.

Ontology Population (OP) Module: expand the current sample of ArchOnto representations and transcriptions to allow a more significant analysis of the suggested system. We should establish a correlation between document typologies and ontology population performance to understand the relevance of extracted information in archival descriptions. We should also implement co-reference resolution to eliminate redundant instances and maintain ontology consistency. Additionally, we should incorporate a measure in the ontology population evaluation to assess semantic relations, i.e., object properties.

References

- [1] GitHub - tesseract-ocr/tesseract: Tesseract Open Source OCR Engine (main repository). <https://github.com/tesseract-ocr/tesseract>, 2022.
- [2] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.
- [3] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.
- [4] Apostolos Antonacopoulos, Christian Clausner, Christos Papadopoulos, and Stefan Pletschacher. ICDAR 2013 Competition on Historical Book Recognition (HBR 2013). In *2013 12th International Conference on Document Analysis and Recognition*, pages 1459–1463, 08 2013.
- [5] Konstantin Baierer. Models · OCROPUS/ocropy Wiki, 2020.
- [6] Julian Blank and Kalyan Deb. Pymoo: Multi-Objective Optimization in Python. *IEEE Access*, PP:1–1, 04 2020.
- [7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information, 2016.
- [8] Gary Bradski. The OpenCV Library. *Dr. Dobb’s Journal of Software Tools*, 2000.
- [9] David Bromage. Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata. *Archives and Manuscripts*, 44:172–174, 10 2016.
- [10] Ann Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. To BERT or not to BERT - Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of four Types of Speech, Thought and Writing Representation. In *SwissText/KONVENS*, 2020.
- [11] Quang Anh Bui, David Mollard, and Salvatore Tabbone. Selecting Automatically Pre-Processing Methods to Improve OCR Performances. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 169–174, Kyoto, Japan, 2017. IEEE Computer Society.
- [12] Pedro Castro, Nadia Felix, and Anderson Soares. Portuguese Named Entity Recognition using LSTM-CRF. 07 2018.

- [13] Marcirio Silveira Chaves. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM. 2008.
- [14] Christian Clausner, Stefan Pletschacher, and Apostolos Antonacopoulos. The Significance of Reading Order in Document Recognition and Its Evaluation. In *2013 12th International Conference on Document Analysis and Recognition*, pages 688–692, Washington, DC, USA, 2013. IEEE Computer Society.
- [15] Christian Clausner, Stefan Pletschacher, and Apostolos Antonacopoulos. Flexible character accuracy measure for reading-order-independent evaluation. *Pattern Recognition Letters*, 131:390–397, 2020.
- [16] Davide Colla, Annamaria Goy, Marco Leontino, Diego Magro, and Claudia Picardi. Bringing Semantics into Historical Archives with Computer-Aided Rich Metadata Generation. *J. Comput. Cult. Herit.*, aug 2021.
- [17] Sandra Collovini, Tiago Bonamigo, and Renata Vieira. A review on Relation Extraction with an eye on Portuguese. *Journal of the Brazilian Computer Society*, 19, 11 2013.
- [18] Sandra Collovini, Gabriel Machado, and Renata Vieira. A Sequence Model Approach to Relation Extraction in Portuguese. In *International Conference on Language Resources and Evaluation*, 2016.
- [19] Sandra Collovini, Joaquim Francisco Santos Neto, Bernardo Scapini Consoli, Juliano Terra, Renata Vieira, Paulo Quaresma, Marlo Souza, Daniela Barreiro Claro, and Rafael Glauber. IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks. In *IberLEF@SEPLN*, 2019.
- [20] Sandra Collovini, Bolivar Pereira, Henrique dos Santos, and Renata Vieira. *Annotating Relations Between Named Entities with Crowdsourcing*, pages 290–297. 01 2018.
- [21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.
- [22] Mariana Dias and Carla Teixeira Lopes. Mining Typewritten Digital Representations to Support Archival Description. In *TPDL Workshops*, 2022.
- [23] Mariana Dias and Carla Teixeira Lopes. Typewritten Digital Representations of Portuguese Cultural Heritage Documents from the 20th century. Data set, 2022.
- [24] Mariana Dias and Carla Teixeira Lopes. Consensual ArchOnto representation of 13 Portuguese Historical Archival Records based on their Digital Representations. Data set, 2023.
- [25] DigitArq. Arquivo Nacional Torre do Tombo, n.d.
- [26] Martin Doerr, Georgios Markakis, Maria Theodoridou, and M. Tsikritsis. DIATHESIS: OCR based semantic annotation of newspapers. June 2007.
- [27] Instituto dos Arquivos Nacionais/Torre do Tombo. Orientações para a gestão de documentos de Arquivo no contexto de uma reestruturação da Administração Central do Estado, 09 2016.
- [28] Cícero Nogueira dos Santos and Victor Guimarães. Boosting Named Entity Recognition with Neural Character Embeddings. *CoRR*, abs/1505.05008, 2015.

- [29] Senka Drobac and Krister Lindén. Optical character recognition with neural networks and post-correction with finite state methods. *International Journal on Document Analysis and Recognition (IJ DAR)*, 23:1–17, 12 2020.
- [30] Margarida Falcão, Mariana Dias, and Carla Teixeira Lopes. Manual Transcriptions of Type-written Digital Representations of Portuguese Cultural Heritage Documents from the 20th Century. Data set, 2022.
- [31] João Ferreira, Hugo Gonçalo Oliveira, and Ricardo Rodrigues. NLPyPort: Named Entity Recognition with CRF and Rule-Based Relation Extraction. In *IberLEF@SEPLN*, 2019.
- [32] Evandro Fonseca, André Antonitsch, Sandra Collovini, Daniela Amaral, Renata Vieira, and Anny Figueira. Summ-it++: an Enriched Version of the Summ-it Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 2047–2051, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).
- [33] Lionel Fontan, Isabelle Ferrané, Jérôme Farinas, Julien Pinquier, and Xavier Aumont. Using Phonologically Weighted Levenshtein Distances for the Prediction of Microscopic Intelligibility. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, pages 650–654, San Francisco, CA, United States, 9 2016. ISCA.
- [34] Cláudia Freitas, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira, and Paula 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)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).
- [35] Cláudia Freitas, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira, and Paula Carvalho. Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese. 01 2010.
- [36] Matilde Gonçalves, Luisa Coheur, Jorge Baptista, and Ana Mineiro. Avaliação de recursos computacionais para o português. *Linguamática*, 12:51–68, 01 2021.
- [37] Anna Goy, Rossana Damiano, Fabrizio Di Loreto, Diego Magro, Stefano Musso, Daniele Paolo Radicioni, Cristina Accornero, Davide Colla, Antonio Lieto, Enrico Mensa, Marco Rovera, Dunia Astrologo, Bruno Boniolo, and Matteo D’Ambrosio. PRiSMHA (Providing Rich Semantic Metadata for Historical Archives). In *JOWO*, 2017.
- [38] ICOM/CIDOC Documentation Standards Group. Definition of the CIDOC Conceptual Reference Model. 2019.
- [39] Cláudia Guedes. Designing User Interaction with Linked Data in Historical Archives. Master’s thesis, Faculdade de Engenharia da Universidade do Porto, 2020.
- [40] Maya Gupta, Nathaniel Jacobson, and Eric Garcia. OCR binarization and image pre-processing for searching historical documents. *Pattern Recognition*, 40:389–397, 02 2007.
- [41] Udo Hahn and Klemens Schnattinger. Towards Text Knowledge Engineering. pages 524–531, 07 1998.

- [42] Abdeslam Harraj and Naoufal Raissouni. OCR Accuracy Improvement on Document Images Through a Novel Pre-Processing Approach. *Signal & Image Processing : An International Journal*, 6, 09 2015.
- [43] Matthew Horridge and Sean Bechhofer. The OWL API: A Java API for OWL ontologies. *Semantic Web*, 2:11–21, 2011.
- [44] Ashok C. Popat Jake Walker, Yasuhisa Fujii. A Web-Based OCR Service for Documents. In *13th IAPR International Workshop on Document Analysis Systems (DAS)*, volume 1, Vienna, Austria, 2018.
- [45] Martin Kiss, Michal Hradis, and Oldrich Kodym. Brno Mobile OCR Dataset. *CoRR*, abs/1907.01307, 2019.
- [46] Inês Koch, Cristina Ribeiro, and Carla Teixeira Lopes. ArchOnto, a CIDOC-CRM-Based Linked Data Model for the Portuguese Archives. *Digital Libraries for Open Knowledge Lecture Notes in Computer Science*, page 133–146, 2020.
- [47] Inês Koch, Carla Teixeira Lopes, and Cristina Ribeiro. Moving from ISAD(G) to a CIDOC-CRM-based Linked Data Model in the Portuguese Archives. *ACM Journal on Computing and Cultural Heritage*, 2022.
- [48] Mika Koistinen, Kimmo Kettunen, and Tuula Pääkkönen. Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 277–283, Gothenburg, Sweden, 05 2017. Association for Computational Linguistics.
- [49] Lars Kotthoff. *Algorithm Selection for Combinatorial Search Problems: A Survey*, pages 149–190. Springer International Publishing, Cham, 12 2016.
- [50] Gaurav Kumar. Analytical Review of Preprocessing Techniques for Offline Handwritten Character Recognition. *International Journal of Advances in Engineering Sciences*, 3:14–22, 2013.
- [51] Jean-Baptiste Lamy. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artificial Intelligence in Medicine*, 80, 08 2017.
- [52] Jangwon Lee, Jungi Lee , Minho Lee , and Gil-Jin Jang. Named Entity Correction in Neural Machine Translation Using the Attention Alignment Map. *Applied Sciences*, 11(15), 2021.
- [53] Jing Li, Aixin Sun, Ray Han, and Chenliang Li. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1, 03 2020.
- [54] Yaliang Li, Jing Jiang, Hai Leong Chieu, and Kian Ming Adam Chai. Extracting Relation Descriptors with Conditional Random Fields. In *International Joint Conference on Natural Language Processing*, 2011.
- [55] Constantine Lignos and Marjan Kamyab. If You Build Your Own NER Scorer, Non-replicable Results Will Come. pages 94–99, 01 2020.

- [56] William Lund, Douglas Kennard, and Eric Ringger. Combining Multiple Thresholding Binarization Values to Improve OCR Output. In *Document Recognition and Retrieval XX*, volume 8658. SPIE, 02 2013.
- [57] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [58] Diana Maynard, Yaoyong Li, and Wim Peters. NLP Techniques for Term Extraction and Ontology Population. *Ontology learning and population: bridging the gap between text and knowledge*, 167:107–127, 06 2008.
- [59] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, 2013.
- [60] Ruy Luiz Milidiú, Julio Cesar Duarte, and Roberto Cavalcante. Machine Learning Algorithms for Portuguese Named Entity Recognition. *Inteligencia Artif.*, 11:67–75, 2007.
- [61] Mumtazimah Mohamad. A Review on OpenCV, 08 2015.
- [62] Eric Lease Morgan and LiAM. Linked Archival Metadata: A Guidebook. Technical report, 2014.
- [63] Marwa Naili, Anja Habacha Chaibi, and Henda Hajjami Ben Ghezala. Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science*, 112:340–349, 2017. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.
- [64] Hiroki Nakayama. seqeval: A Python framework for sequence labeling evaluation, 2018. Software available from <https://github.com/chakki-works/seqeval>.
- [65] Zara Nasar, Syed Waqar Jaffry, and Muhammad Malik. Named Entity Recognition and Relation Extraction: State of the Art. *ACM Computing Surveys*, 54, 02 2021.
- [66] Jinfang Niu. Linked data for archives. *Archivaria*, 2016:83–110, 01 2016.
- [67] Gottfried E. Noether. *Introduction to Wilcoxon (1945) Individual Comparisons by Ranking Methods*, pages 191–195. Springer New York, New York, NY, 1992.
- [68] Nikolaos Ntogas and Dimitrios Ventzas. Binarization of pre-filtered historical manuscripts images. *International Journal of Intelligent Computing and Cybernetics*, 2:148–174, 03 2009.
- [69] International Council on Archives. ISAD(G): General International Standard Archival Description. *International Council on Archives*, 2000.
- [70] FineReader PDF. How AI powers PDF Software & Technology Trends: FineReader Blog, 2019.
- [71] FineReader PDF. Technical Specifications and System Requirements | FineReader PDF, 2021.

- [72] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [73] Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Anastasia Krithara, and Elias Zavitsanos. Ontology Population and Enrichment: State of the Art. pages 134–166, January 2011.
- [74] Philip Pfaffe, Martin Tillmann, Sigmar Walter, and Walter Tichy. Online-Autotuning in the Presence of Algorithmic Choice. In *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 1379–1388, Lake Buena Vista, FL, 06 2017. IEEE Computer Society.
- [75] André Pires. SIGARRA News Corpus. Data set, 2017.
- [76] André Pires, José Luís Devezas, and Sérgio Nunes. Benchmarking Named Entity Recognition Tools for Portuguese. In *Proceedings of the Ninth INForum: Simpósio de Informática*, pages 111–121, 2017.
- [77] Stefan Pletschacher, Christian Clausner, and Apostolos Antonacopoulos. Europeana Newspapers OCR Workflow Evaluation. In *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*, pages 39–46, New York, NY, USA, 2015. Association for Computing Machinery.
- [78] Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. Tunability: Importance of hyperparameters of machine learning algorithms. *J. Mach. Learn. Res.*, 20(1):1934–1965, jan 2019.
- [79] Eric Prud’hommeaux and Andy Seaborne. SPARQL query language for RDF, W3C Recommendation, 01 2008.
- [80] Wilfried Roetzel, Xing Luo, and Dezhen Chen. *Chapter 6 - Optimal design of heat exchanger networks*, pages 231–317. Academic Press, 2020.
- [81] Cicero Dos Santos and Bianca Zadrozny. Learning Character-level Representations for Part-of-Speech Tagging. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1818–1826, Beijing, China, 22–24 Jun 2014. PMLR.
- [82] Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA).
- [83] Joaquim Santos, Bernardo Consoli, Cicero dos Santos, Juliano Terra, Sandra Collonini, and Renata Vieira. Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition. In *Proceedings of the 8th Brazilian Conference on Intelligent Systems*, pages 437–442, 2019.
- [84] Fábio Souza, Rodrigo Nogueira, and Roberto de Alencar Lotufo. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Brazilian Conference on Intelligent Systems*, 2020.

- [85] Fábio Souza, Rodrigo Frassetto Nogueira, and Roberto de Alencar Lotufo. Portuguese Named Entity Recognition using BERT-CRF. *CoRR*, abs/1909.10649, 2019.
- [86] Dan Sporici, Elena Cuşnir, and Costin-Anton Boiangiu. Improving the Accuracy of Tesseract 4.0 OCR Engine Using Convolution-Based Preprocessing. *Symmetry*, 12:715, 05 2020.
- [87] Xiaomeng Su and Jon Gulla. Semantic Enrichment for Ontology Mapping. volume 3136, pages 217–228, June 2004.
- [88] Alaa Sulaiman, Khairuddin Omar, and Mohammad Faizul Nasrudin. Degraded Historical Document Binarization: A Review on Issues, Challenges, Techniques, and Future Directions. *Journal of Imaging*, 5, 04 2019.
- [89] Charles Sutton and Andrew McCallum. An Introduction to Conditional Random Fields, 2010.
- [90] Chris Tensmeyer and Tony Martinez. Historical Document Image Binarization: A Review. *SN Computer Science*, 1, 05 2020.
- [91] Joo Chuan Tong and Shoba Ranganathan. 5 - Computational T cell vaccine design. In Joo Chuan Tong and Shoba Ranganathan, editors, *Computer-Aided Vaccine Design*, Woodhead Publishing Series in Biomedicine, pages 59–86. Woodhead Publishing, 2013.
- [92] Jake Vasilakes, Sicheng Zhou, and Rui Zhang. Chapter 6 - Natural language processing. In Subhi J. Al’Aref, Gurpreet Singh, Lohendran Baskaran, and Dimitris Metaxas, editors, *Machine Learning in Cardiovascular Medicine*, pages 123–148. Academic Press, 2021.
- [93] Renata Vieira, Fernanda Olival, Helena Cameron, Joaquim Santos, Ofélia Sequeira, and Ivo Santos. Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities. *Journal of Open Humanities Data*, 7, 09 2021.
- [94] Andreas Vlachidis and Douglas 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, 04 2015.
- [95] Jorge Wagner, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. The brWaC Corpus: A New Open Resource for Brazilian Portuguese. 05 2018.
- [96] Stefan Weil. Tesseract-OCR/langdata: Source training data for Tesseract for lots of Languages, 2021.
- [97] Chameera Wimalasuriya and Dejing Dou. Ontology-based information extraction: An Introduction and a survey of current approaches. *J. Information Science*, 36:306–323, 05 2010.
- [98] René Witte, Ralf Krestel, Thomas Kappler, and Peter C. Lockemann. Converting a Historical Architecture Encyclopedia into a Semantic Knowledge Base. *IEEE Intelligent Systems*, 25, 2010.
- [99] Xin-She Yang. *Chapter 6 - Genetic Algorithms*, pages 91–100. Academic Press, second edition edition, 2021.

- [100] Ming Yu, Nawapak Eua-Anant, A. Saudagar, and Lalita Udp. Genetic algorithm approach to image segmentation using morphological operations. In *1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*, volume 3, pages 775–779, Chicago, Illinois, USA, 11 1998. IEEE Computer Society.
- [101] Leonardo Zilio, Maria Finatto, and Renata Vieira. Named Entity Recognition Applied to Portuguese Texts from the XVIII Century. 2022.

Appendix A

HAREM Corpora Information

Table A.1: Entity types in Second HAREM’s GC split by categories, types, and subtypes.

Categories	Types	SubTypes
Abstracao	Disciplina	
	Estado	
	Ideia	
	Nome	
	Outro	
Acontecimento	Efemeride	
	Evento	
	Organizado	
	Outro	
Coisa	Classe	
	MembroClasse	
	Objecto	
	Substancia	
	Outro	
Local	Fisico	Ilha, AguaCurso, Planeta, Regiao, Relevo, AguaMassa, Outro
	Humano	Rua, Pais, Divisao, Regiao, Construcão, Outro
	Virtual	ComSocial, Sitio, Obra, Outro
	Outro	
Obra	Arte	Edificio, Pintura, Outro
	Plano	
	Reproduzida	Filme, Livro, Programa, Musica, Outro
	Outro	
Organizacao	Administracao	
	Empresa	
	Instituicao	

Table A.1 continued from previous page

Categories	Types	SubTypes
	Outro	
Pessoa	Cargo	
	GrupoCargo	
	GrupoInd	
	GrupoMembro	
	Individual	
	Membro	
	Povo	
	Outro	
Tempo	Duracao	
	Frequencia	
	Generico	
	Tempo_Calend	Hora, Intervalo, Data, Outro
	Outro	
Valor	Classificacao	
	Moeda	
	Quantidade	
	Outro	
Outro		

Table A.2: Relation types in Second HAREM's GC [35].

Relation type	#
autor_de/obra_de (authorship)	142
causador_de (agent)	22
consequencia_de (result_of)	1
data_de/datado_de (date of)	105
data_morte (death date)	10
data_nascimento (birth date)	5
ident (identity)	2229
inlui/incluido (inclusion)	854
local_nascimento_de/natural_de (birth_place)	142
localizado_em/localizacao_de (place of)	24
nome_de/nomeado_por (name-of)	56
ocorre_em/sede_de (location)	358
outra_edicao (other edition)	3
outrarel (other relation)	93
participante_em/ter_participacao_de (participation-in)	153
periodo_vida (lifetime)	5
personagem_de (character of)	14
praticado_em/pratica_se/ praticante_de/praticado_por (practicing)	99
produtor_de/produzido_por (manufacturing)	50
proprietario_de/propriedade_de (ownership)	39
relacao_familiar (kinship relation)	88
relacao_profissional (professional relation)	17
residente_de/residencia_de (place of residence)	19
vinculo_inst (affiliation)	275
TOTAL	4803

Appendix B

Optimization of Image Processing Algorithms for OCR

Table B.1: Parameters set for the image processing algorithms tuned by NGSa-II for classified types of digital representations.

Algorithm	Parameter	Default	Global	Type of digital representation			
				Letter	Process	Structured	Non-structured
					cover	report	cover
Adaptive Thresholding	maxValue	255	217	24	72	13	152
	adaptiveMethod	0 or 1	1	0	0	0	1
	thresholdType	0	0	0	0	0	1
	blockSize	11	33	39	65	43	57
	c	2	25	30	29	43	18
Bilateral Filtering	d	9	4	4	2	2	1
	sigmaColor	75	10	4	85	36	6
	sigmaSpace	75	231	31	52	100	191
Black Hat	kernel	5	10	6	229	139	30
	iterations	1	7	10	7	9	2
	borderType	3	4	4	2	3	1
Closing	kernel	5	1	1	1	1	1
	iterations	1	2	2	4	9	4
	borderType	3	3	1	0	1	2
Dilation	kernel	5	32	-	1	1	1
	iterations	1	8	-	4	7	9
	borderType	3	2	-	3	2	1
Erosion	kernel	5	1	1	1	1	1
	iterations	1	5	7	1	5	2
	borderType	3	3	2	2	1	3

Table B.1 continued from previous page

Algorithm	Parameter	Default	Global	Type of digital representation				
				Letter		Process		
					cover	Structured report	Theatre play cover	Non-structured report
Gaussian Blur	ksize	5	3	3	3	3	3	3
	borderType	3	1	0	1	1	1	1
Homogeneous Blur	ksize	5	1	1	3	1	3	3
	borderType	3	1	2	3	1	0	3
Median Blur	ksize	5	3	3	3	5	3	3
Morphological Gradient	kernel	5	2	2	2	2	105	2
	iterations	1	1	1	2	1	2	2
	borderType	3	2	2	0	1	1	2
Opening	kernel	5	2	1	2	1	2	2
	iterations	1	2	2	1	9	1	1
	borderType	3	0	4	1	4	0	3
Otsu Thresholding	maxValue	255	199	24	251	92	34	14
	type	0	3	0	0	3	3	0
Simple Thresholding	thresh	127	152	110	33	82	240	224
	maxValue	255	45	4	10	221	3	42
	type	0	1	3	3	3	4	4
Top Hat	kernel	5	90	51	217	233	139	204
	iterations	1	5	7	5	2	5	2
	borderType	3	2	0	0	2	2	4
Triangle Thresholding	maxValue	255	15	74	248	220	174	238
	type	0	3	3	2	2	3	0

Table B.2: Performance comparison between image processing algorithms tuned with Wilcoxon pairwise comparisons using default, global and digital representation-type parameters.

Technique	Algorithm	Character Accuracy				F1-score			
		Default vs		Global vs		Default vs		Global vs	
		Global	Typology	Global	Typology	Global	Typology	Global	Typology
Binarization	Adaptive Thresholding (a)	<<<	<<<	<<<	<<<	<<<	<<<	<<<	<<<
	Adaptive Thresholding (b)	<<<	<<<	<<<	<<<	<<<	<<<	<<<	<<<
	Otsu Thresholding	<<<	<<<	>>>	>>>	<<<	<<<	>>>	>>>
	Simple Thresholding	>>>	<<<	<<<	<<<	>>>	<<<	<<<	<<<
	Triangle Thresholding	0.65	<<<	<<<	<<<	<<<	<<<	<<<	<<<
Image smoothing	Bilateral Filtering	<<<	<<<	0.17	<<<	<<<	<<<	1	1
	Gaussian Blur	<<<	<<<	1	<<<	<<<	<<<	1	1
	Homogeneous Blur	<<<	<<<	0.19	<<<	<<<	<<<	>>	>>
	Median Blur	<<<	<<<	>>>	<<<	<<<	<<<	>>>	>>>
	Black Hat	<<<	<<<	0.2	<<<	<<<	<<<	0.24	0.24
Morphological transformation	Closing	<<<	<<<	0.25	<<<	<<<	<<<	0.26	0.26
	Dilation	>>>	<<<	<<<	<<<	>>	<<<	<<<	<<<
	Erosion	<<<	<<<	0.38	<<<	<<<	<<<	0.38	0.38
	Morphological Gradient	<<<	<<<	>	<<<	<<<	<<<	0.19	0.19
	Opening	<<<	<<<	<<<	<<<	<<<	<<<	<<<	<<<
	Top Hat	<<<	<<<	1	<<<	<<<	<<<	0.44	0.44

(a) Adaptive Thresholding using adaptiveMethod=0;

(b) Adaptive Thresholding using adaptiveMethod=1;

For < or >, $P < 0.05$; for << or >> $P < 0.01$; for <<< or >>> $P < 0.001$

Displays the p -values of Wilcoxon pairwise comparisons with symbol "<" for significantly worse performances and ">" for significantly better performances.

Table B.3: Summary of Friedman test results and Wilcoxon pairwise comparisons for post-hoc analysis of OCR performance of image processing algorithms with the overall dataset.

	Character Accuracy				F1-score			
	Mean	Significantly worse results	#	Mean	Significantly worse results	#		
None	N	66.4	OT TT MB MG	4	48.8	AT OT ST TT GB HB* MB MG BH*	9	
Adaptive Thresholding	AT	68.9	TT MB MG	3	47.1	TT MB MG	3	
Otsu Thresholding	OT	65.7	TT MB MG	3	45.0	TT MB MG	3	
Simple Thresholding	ST	65.8	TT MB MG	3	47.1	OT TT MB MG	4	
Triangle Thresholding	TT	61.2	MB MG	2	41.1	MB MG	2	
Bilateral Filter	BF	68.0	N* OT ST TT GB HB MB C* MG	9	49.1	AT OT ST TT GB HB** MB BH** MG	9	
Gaussian Blur	GB	64.2	TT MB MG	3	45.9	MB MG TT	3	
Homogeneous Blur	HB	64.8	OT TT MB MG	4	47.0	OT TT GB** MB MG	5	
Median Blur	MB	45.7		0	30.3	MG	1	
Black Hat	BH	68.7	TT MB MG	3	49.0	AT OT TT MB MG	5	
Closing	C	66.8	OT TT GB* MB MG	5	48.7	AT OT ST** TT GB HB* MB BH* MG	9	
Erosion	E	67.0	OT TT MB MG	4	48.7	AT OT ST* TT GB MB MG	7	
Morphological Gradient	MG	49.3		0	30.3		0	
Opening	O	67.8	N** OT ST** TT GB** HB MB C* E* MG	10	49.2	AT OT ST TT GB HB MB BH** MG	9	
Top Hat	TH	67.6	OT* TT MB MG	4	48.9	AT OT ST* TT GB** MB MG	7	
χ^2		905.97			1170.5			
p-value		<2.2e-16			<2.2e-16			

*P<0.05; **P<0.01; P<0.001 for the rest of the values

The bold values represent the image processing algorithms and average results that improved the OCR performance.

Table B.4: Summary of Friedman test results and Wilcoxon pairwise comparisons for post-hoc analysis of OCR performance of image processing algorithms with the letters sample.

Character Accuracy				F1-score			
	Mean	Significantly worse results	#	Mean	Significantly worse results	#	
None	N	OT ST** TT MB MG	5	60.1	AT OT ST TT GB** MB MG	7	
Adaptive Thresholding	AT	TT MB* MG	3	57.6	TT MG	2	
Otsu Thresholding	OT	TT MG	2	52.7	TT MG	2	
Simple Thresholding	ST	TT MG	2	55.3	OT TT MG	3	
Triangle Thresholding	TT		0	37.7		0	
Bilateral Filter	BF	N* OT ST TT HB* MB C* E* MG O*	10	60.2	AT OT ST TT GB** MB MG	7	
Gaussian Blur	GB	TT MB* MG	3	60.1	OT TT MB MG	4	
Homogeneous Blur	HB	OT ST** TT MB MG	5	60.1	AT OT ST TT GB** MB MG	7	
Median Blur	MB	MG* TT	2	54.2	MG TT	2	
Black Hat	BH	TT MB MG	3	60.7	AT OT TT MB** MG	5	
Closing	C	OT ST** TT MB MG	5	60.1	AT OT ST TT GB** MB MG	7	
Erosion	E	OT ST** TT MB MG	5	60.1	AT OT ST TT GB** MB MG	7	
Morphological Gradient	MG		0	36.6		0	
Opening	O	OT ST** TT MB MG	5	60.1	AT OT ST TT GB** MB MG	7	
Top Hat	TH	TT MG	2	59.2	AT* OT TT MB* MG	5	
χ^2				395.56			
p-value				<2.2e-16			

*P<0.05; **P<0.01; P<0.001 for the rest of the values

The bold values represent the image processing algorithms and average results that improved the OCR performance.

Table B.5: Summary of Friedman test results and Wilcoxon pairwise comparisons for post-hoc analysis of OCR performance of image processing algorithms with the non-structured reports sample.

	Character Accuracy			F1-score		
	Mean	Significantly worse results	#	Mean	Significantly worse results	#
None	N	OT GB HB MB MG	5	53.8	OT TT* GB HB MB MG	6
Adaptive Thresholding	AT	TT GB* HB* MB** MG	5	53.7	TT MG	2
Otsu Thresholding	OT	GB** HB** MB** MG	4	44.0	MG	1
Simple Thresholding	ST	OT GB HB MB MG	5	53.1	OT TT* GB HB MB MG	6
Triangle Thresholding	TT	MG	1	46.5	MG	1
Bilateral Filter	BF	GB HB MB MG	5	53.8	OT TT* GB HB MB MG	6
Gaussian Blur	GB	HB*	1	44.3	MG	1
Homogeneous Blur	HB		0	41.1	MG	1
Median Blur	MB		0	42.2	MG	1
Black Hat	BH	OT TT GB HB MB E* MG TH	8	58.6	AT** OT TT GB HB MB MG	7
Closing	C	OT GB HB MB MG	5	53.8	OT TT* GB HB MB MG	6
Dilation	D	OT GB HB MB MG	5	53.8	OT TT* GB HB MB MG	6
Erosion	E	GB** HB** MB** MG	4	54.1	OT TT** GB* HB** MB MG	6
Morphological Gradient	MG		0	22.6		0
Opening	O	OT** TT* BF* GB HB MB MG	7	56.0	OT TT** GB HB MB MG	6
Top Hat	TH	GB* HB** MB* MG	4	55.4	OT TT GB HB MB MG	6
χ^2	161.69			199.79		
p-value	<2.2e-16			<2.2e-16		

*P<0.05; **P<0.01; P<0.001 for the rest of the values

The bold values represent the average results that improved the OCR performance.

Table B.6: Summary of Friedman test results and Wilcoxon pairwise comparisons for post-hoc analysis of OCR performance of image processing algorithms with the process covers sample.

	Character Accuracy				F1-score	
	Mean	Significantly worse results	#	Mean	Significantly worse results	#
None	N	62.4	MG	1	35.3	BH** MG
Adaptive Thresholding	AT	63.1	MG	1	33.0	MG
Otsu Thresholding	OT	61.6	MG	1	31.5	MG
Simple Thresholding	ST	64.5	BH* MG	2	36.5	BH** MG
Triangle Thresholding	TT	56.3	MG*	1	32.8	MG
Bilateral Filter	BF	67.1	BH** MG	2	38.9	OT** BH MG
Gaussian Blur	GB	67.1	OT** BH MG TH*	4	39.2	OT** TT* BH** MG
Homogeneous Blur	HB	67.4	OT* TT* BH MG	4	38.8	OT** BH** MG
Median Blur	MB	65.8	BH* MG	2	39.3	OT BH** MG
Black Hat	BH	54.6	MG	1	28.0	MG
Closing	C	62.4	MG	1	35.3	BH** MG
Dilation	D	62.4	MG	1	35.3	
Erosion	E	62.4	MG	1	35.3	BH** MG
Morphological Gradient	MG	34.9		0	11.4	
Opening	O	65.1	BH** MG	2	36.1	OT* BH** MG
Top Hat	TH	62.7	MG	1	35.6	MG
χ^2		104.27			126.29	
p-value		7.183e-16			<2.2e-16	

*P<0.05; **P<0.01; P<0.001 for the rest of the values

The bold values represent the average results that improved the OCR performance.

Table B.7: Summary of Friedman test results and Wilcoxon pairwise comparisons for post-hoc analysis of OCR performance of image processing algorithms with the structured reports sample.

	Character Accuracy				F1-score		
	Mean	Significantly worse results	#	Mean	Significantly worse results	#	
None	N	69.7	OT MB MG	3	50.1	AT GB MB BH** MG	
Adaptive Thresholding	AT	68.5	MB MG	2	47.2	MB MG	
Otsu Thresholding	OT	69.4	MB MG	2	49.6	AT MB MG	
Simple Thresholding	ST	69.0	MB MG	2	49.1	AT MB MG	
Triangle Thresholding	TT	65.4	MB MG	2	47.2	MB MG	
Bilateral Filter	BF	70.7	AT ST* TT** GB** MB BH*	6	50.3	AT GB MB BH** MG	
Gaussian Blur	GB	68.3	MB MG	2	48.0	MB MG	
Homogeneous Blur	HB	69.7	OT** MB MG	3	50.1	AT GB MB BH** MG	
Median Blur	MB	30.0		0	16.7	MG	
Black Hat	BH	69.5	MB MG	2	49.0	AT** MB MG	
Closing	C	69.7	OT** MB MG	3	50.1	AT GB MB BH** MG	
Dilation	D	69.7	OT** MB MG	3	50.1	AT GB MB BH** MG	
Erosion	E	69.7	OT** MB MG	3	50.1	AT GB MB BH** MG	
Morphological Gradient	MG	52.7	MB	1	23.3	MB	
Opening	O	69.7	OT** MB MG	3	50.1	AT GB MB BH** MG	
Top Hat	TH	69.5	MB MG	2	49.8	AT GB* MB MG	
χ^2		746.29			888.56		
p-value		<2.2e-16			<2.2e-16		

*P<0.05; **P<0.01; P<0.001 for the rest of the values

The bold values represent the average results that improved the OCR performance.

Table B.8: Summary of Friedman test results and Wilcoxon pairwise comparisons for post-hoc analysis of OCR performance of image processing algorithms with the theatre plays' covers sample.

	Character Accuracy				F1-score	
	Mean	Significantly worse results	#	Mean	Significantly worse results	#
None	N	18.2	MG	1	16.6	MG
Adaptive Thresholding	AT	35.6	N OT* ST* BF* HB* C* D* E* MG	9	21.8	HB* MB* MG
Otsu Thresholding	OT	22.6	MG	1	14.1	MG
Simple Thresholding	ST	21.4	MG	1	14.6	MG
Triangle Thresholding	TT	26.7	MG	1	16.0	MG
Bilateral Filter	BF	23.4	MG	1	16.1	MG
Gaussian Blur	GB	21.4	MG	1	14.0	MG
Homogeneous Blur	HB	21.2	MG	1	12.3	MG
Median Blur	MB	22.8	MG	1	11.8	MG
Black Hat	BH	31.8	N* MG	2	22.2	MG
Closing	C	22.4	MG	1	14.7	MG
Dilation	D	22.4	MG	1	14.7	MG
Erosion	E	22.4	MG	1	14.7	MG
Morphological Gradient	MG	0.47		0	0	0
Opening	O	25.7	MG	1	16.3	MG
Top Hat	TH	29.3	N* MG	2	18.9	MG
χ^2	92.122				81.827	
p-value	1.506e-13				1.295e-11	

*P<0.05; **P<0.01; P<0.001 for the rest of the values

The bold values represent the image processing algorithms and average results that improved the OCR performance.

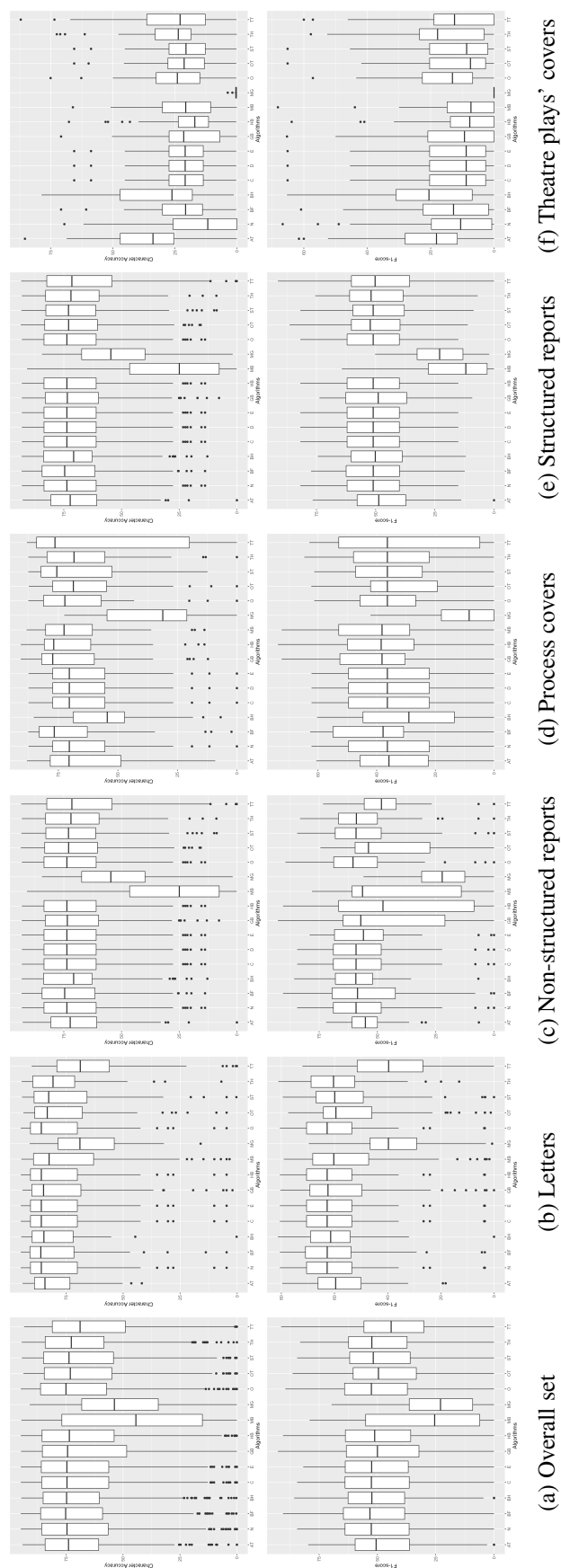


Figure B.1: Distribution of OCR performance measures in the different typologies of digital representations. The first row of results refers to the character accuracy performance measure, and the second row refers to the F1-score performance measure.

Appendix C

User Interface Mockups

DigitArq+
Search
Hierarchy
Create
My Bookmarks
Luana Sá

Home / Search Results / D_João V / D_José I / Carta de sentimento do Papa Bento XIV ao rei D. José I sobre o terramoto

ds Carta de sentimento do Papa Bento XIV ao rei D. José I sobre o terramoto
Edit
History
Add Bookmark
Download
Print

Archival Description
Conservation & Restoration
Support Transfer

Description Level
Description Level
Item

Internal Identifiers
DigitArq+ Identifier
0126373
DigitArq+ Identifier Type
General
DigitArq+ Identifier
PT/TT/BUL/0062/04
DigitArq+ Identifier Type
Reference Code
DigitArq+ Identifier
Bulas, mc. 62, no. 4
DigitArq+ Identifier Type
Current Quote

External Identifiers
Aggregator
No data available
External Identifier
No data available

Titles
Title
Carta de sentimento do Papa Bento XIV ao rei D. José I sobre o terramoto
Title Type
Assigned

Existence
Date
1755-12-10
Date Type
Production

Dimensions
Value
1
Physical Unit
doc.

Supports
Support
Paper

Languages
Language
Latin
Language
Italian

Access
Access
Available

Hierarchy
Search in the hierarchy
C1 Bulas 1179/1882
iii Maço 62
ds Carta de sentimento do Papa Bento XIV ao rei D. José I

Semantic Network/Table View
Search for nodes or relations
Options
Legend
Degree of separation: 0 1 2 3 4

Creation Details
Creator
jbraga
Creation Date
2013-04-22 11:57:39
Modified by
ptremoceiro
Last Modification
2013-04-22 11:57:39
Revised Content
Descrição feita por Joana Braga (Torre do Tombo). Data da descrição: 2013-04-22.
Status
Published
Version
2

Keywords
Letter

© 2020 DigitArq+

Figure C.1: Individual view of a description record [39].

ArchClient
Search
Hierarchy
Create
Account

Create Record

Archival Description
Conservation & Restoration
Support Transfer

Digital Representations

Select a file or drag and drop here
Supported formats: JPEG, PNG, TIFF
Select File

PT-TT-SNI-DSC-24-203_m0001.tiff Preview 232 KB
PT-TT-SNI-DSC-24-203_m0002.tiff Preview 175 KB
PT-TT-SNI-DSC-24-203_m0003.tiff Preview 305 KB
PT-TT-SNI-DSC-24-203_m0004.tiff Preview 253 KB
PT-TT-SNI-DSC-24-203_m0005.tiff Preview 308 KB

Load More

Description Level

Identifiers

Titles

Title PETRÓLEO NO MUNDO
Title Type Formal
Discard Add

Existence

Dimensions

Supports

Languages

Access

Semantic Network

Esso Standard Portugal

In/Out	Relationship	Source/Target Node
	has current or former residence	Rua Rosa Araújo

Production

In/Out	Relationship	Source/Target Node
	performed by	Fernando Faro Valadas
	in the role of	director
	performed by	Untitled
	in the role of	editor

Cancel
Create Record

© 2022 EPISA ArchClient

Figure C.2: Individual view of the suggested archival description from a digital representation.

ArchClient

SearchHierarchyCreate▼Account

Processo relativo à publicação “Petróleo no mundo”

EditDownloadPrint

Archival DescriptionConservation & RestorationSupport Transfer

Description Level		▲
Description Level	No data available	
Identifiers		▲
Identifier Type	Identifier	
No data available	No data available	
Titles		▲
Title Type	Title	
Supplied	Processo relativo à publicação “Petróleo no mundo”	
Existence		▲
Date Type	Date	
No data available	No data available	
Dimensions		▲
Material	Value (Measurement Unit)	
No data available	No data available	
Supports		▲
Support	No data available	
Languages		▲
Language	No data available	
Access		▲
Access	No data available	

Semantic Network

Esso Standard Portugal						
In/Out	Relationship	Source/Target Node				
↔	has current or former residence	📍 Rua Rosa Araújo				
« < 1 > »						
Production						
In/Out	Relationship	Source/Target Node				
↔	▼ performed by	👤 Fernando Faro Valadas				
	In/Out	Relationship	Source/Target Node			
	↔	in the role of	👤 director			
	↔	in the role of	👤 editor			
« < 1 > »						

© 2022 EPISA ArchClient

Figure C.3: Individual view of the created record.