

Contextualization
Motivation
Objectives
Documents
Architecture
Implementation
Models and Persistence
Entity Extraction
Relation Extraction
Pipeline Integration
Example
Conclusion

Exploration of documents concerning Foundlings in Fafe along XIX Century

João Costeira Faria Gomes

Universidade do Minho

2022



Outline

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

- 1 Contextualization**
- 2 Motivation**
- 3 Objectives**
- 4 Documents**
- 5 Architecture**
- 6 Implementation**

- Models and Persistence**
- Entity Extraction**
- Relation Extraction**
- Pipeline Integration**
- 7 Example**
- 8 Conclusion**

Contextualization

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

- The Foundling Wheel was introduced to combat the anonymous child abandonment problematic.
- Originally spread across Europe during the XIII century.
- Measure implemented to reduce the high mortality rate, consequent to the anonymous abandonment.

Contextualization

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

- Wheel and Bell components provide a safe abandonment, without compromising the anonymity of the perpetrators.
- These public institutions raised the foundling population, additional support provided by hiring nurses and wet nurses.
- The Foundling Wheel is defunct, replaced by Hospices and family based support.

Contextualization

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

- The Municipal Archive of Fafe preserves documents from these historical institutions.

Case Study

The case study consists in the analysis of archived Foundling Wheel and Hospices documents, dated to the nineteenth century, from the Northern region of Portugal.

Motivation

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

- Preserve and disseminate information contained in these historical documents.
- Historical contextualization to comprehend the effectiveness of these programs and the life conditions provided.
- Current debate towards the reintroduction of anonymous child abandonment mechanisms.

Objectives

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

Research Hypothesis

From the identification of concepts and respective relations over the Foundling Wheel archive, it is possible to develop a knowledge repository to support a digital platform

The main objectives of this project are the design and development of a:

- 1 Ontology
- 2 Knowledge Repository
- 3 Digital Platform

Objectives

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

Research Hypothesis

From the identification of concepts and respective relations over the Foundling Wheel archive, it is possible to develop a knowledge repository to support a digital platform

The main objectives of this project are the design and development of a:

- 1** Ontology
- 2** Knowledge Repository
- 3** Digital Platform

Objectives

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

Research Hypothesis

From the identification of concepts and respective relations over the Foundling Wheel archive, it is possible to develop a knowledge repository to support a digital platform

The main objectives of this project are the design and development of a:

- 1** Ontology
- 2** Knowledge Repository
- 3** Digital Platform

Objectives

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

Research Hypothesis

From the identification of concepts and respective relations over the Foundling Wheel archive, it is possible to develop a knowledge repository to support a digital platform

The main objectives of this project are the design and development of a:

- 1** Ontology
- 2** Knowledge Repository
- 3** Digital Platform

Documents

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion



Figure: Book Cover

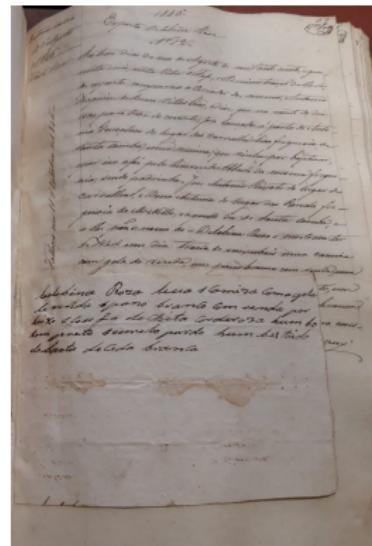


Figure: Document and Signal

Documents

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

Classification of documents according to their main topic:

- Entrance Register
- Departure Register
- Medical Records
- Vaccination Records
- Register of Expenses
- Record of Nurses
- Payment to Nurses
- Foundlings given to Nurses
- Foundlings given to their Mothers
- Record of Letters
- Movement of Foundlings

Documents

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

Main issues found in these documents:

- Linguist issues,
- Degradation issues

Details on the resulting digital representation:

- Documents provided in a Microsoft Word format,
- Additional punctuation to mark degradation issues,
- Key-Value pairs to describe each book

Documents

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

Main issues found in these documents:

- Linguist issues,
- Degradation issues

Details on the resulting digital representation:

- Documents provided in a Microsoft Word format,
- Additional punctuation to mark degradation issues,
- Key-Value pairs to describe each book

Architecture

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

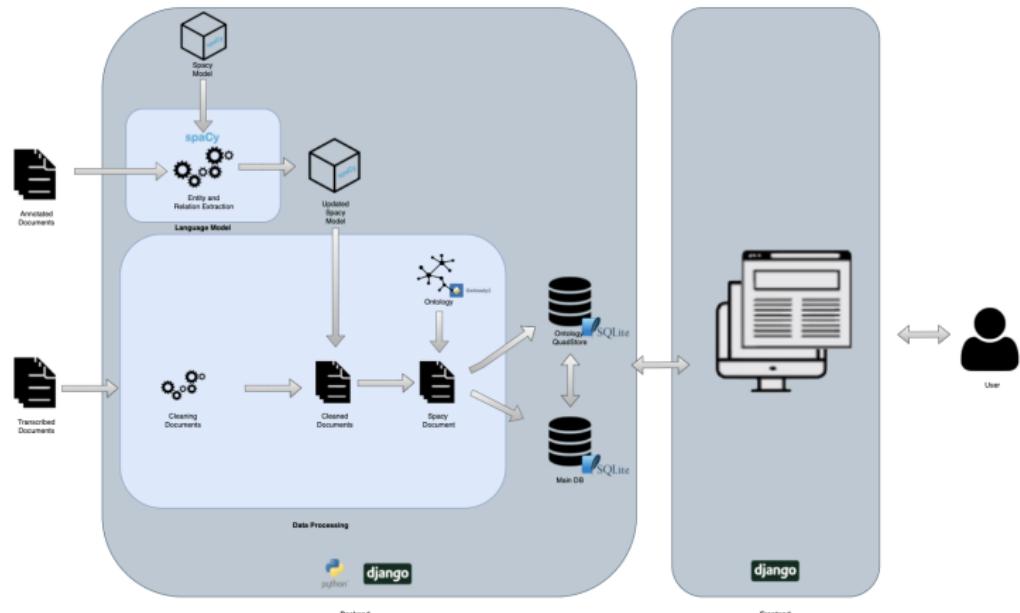


Figure: General Architecture

Architecture

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

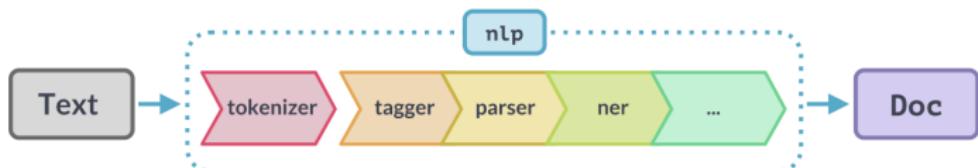
Relation Extraction

Pipeline Integration

Example

Conclusion

- Adaptable Natural Language Model. These linguistic models perform standard NLP tasks.
- Automate the annotation process, by processing a document through a pipeline.



Credit: *Spacy Documentation*

Figure: Spacy Processing Pipeline

Architecture

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

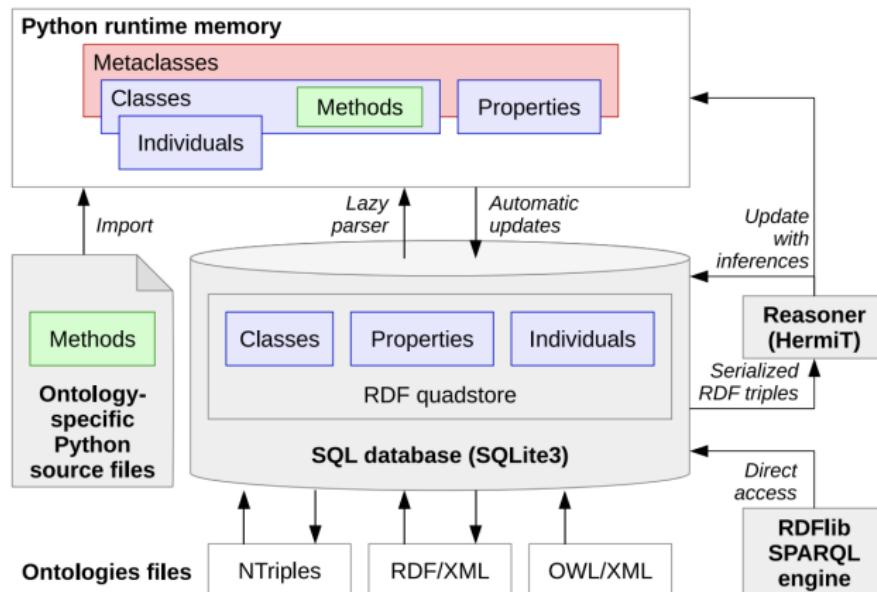


Figure: OWLReady Architecture

Architecture

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

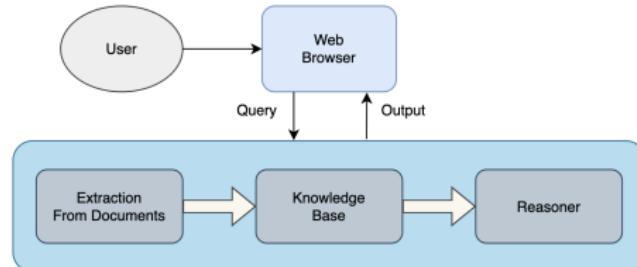


Figure: Repository Interaction

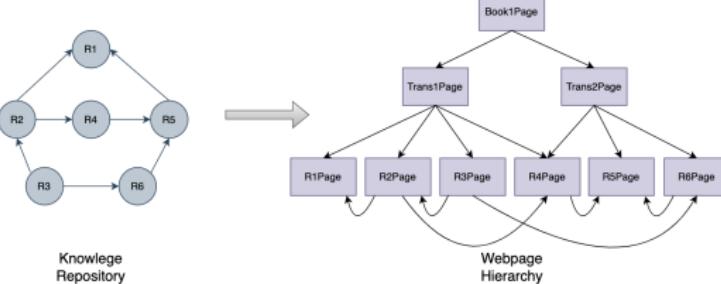


Figure: Repository Rendering

Models and Persistence

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

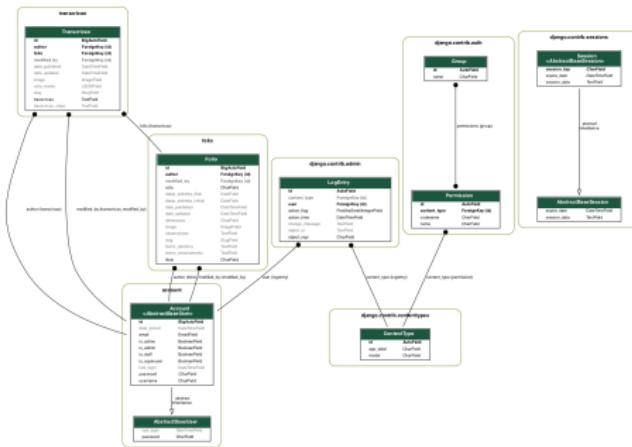


Figure: Database ERD



Figure: Ontology Hierarchy

Entity Extraction

Contextualization
Motivation
Objectives
Documents
Architecture
Implementation
Models and Persistence
Entity Extraction
Relation Extraction
Pipeline Integration
Example
Conclusion

- Spacy provide a NER trainable component, responsible for labelling sequences of text.
- The goal is to extract from each document a set of named entities, accordingly to the ontology concepts and attributes.
- The standard Portuguese model is retrained and permanently stored.



Figure: NER Identified by the Model

Relation Extraction

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion



Figure: NER Identified by the Model

- The goal is to extract meaningful relations between named entities.

Relation Extraction

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion



Figure: Entities Identified

```
pessoa_local = [  
    {"ENT_TYPE" : 'pessoa'},  
    {"POS": "PUNCT", "OP": "?"},  
    {"POS": "ADP", "OP": "?"},  
    {"ENT_TYPE" : 'Local'}  
]
```

Figure: Matcher Pattern

Relation Extraction

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

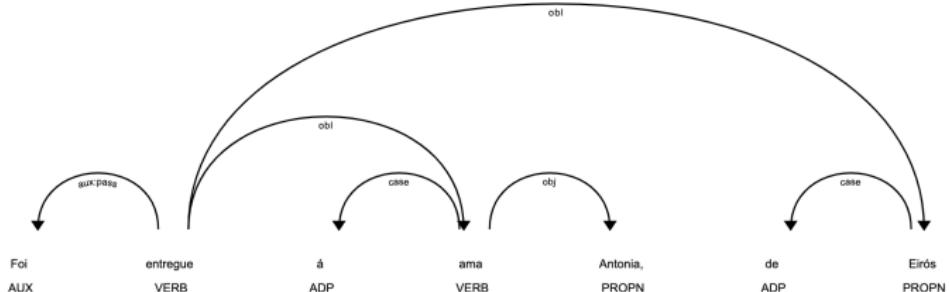


Figure: Dependency Example

```
evento_pessoa = [
    {
        "RIGHT_ID": "evento",
        "RIGHT_ATTRS": {"ENT_TYPE": "Evento"}
    },
    {
        "LEFT_ID": "evento",
        "REL_OP": ">>",
        "RIGHT_ID": "participa",
        "RIGHT_ATTRS": {"ENT_TYPE": "pessoa"}
    }
]
```

Figure: Dependency Matcher Pattern

Pipeline Integration

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

Transcription processing pipeline, responsible for establishing the connection between the independent repositories.

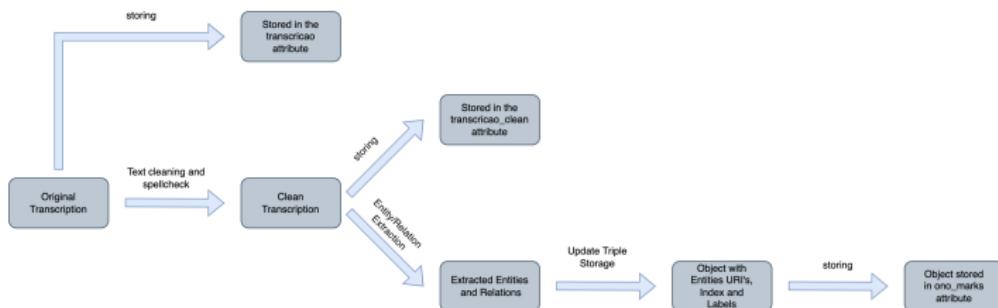


Figure: Transcription Processing Pipeline

Navigation Example

Contextualization
Motivation
Objectives
Documents
Architecture
Implementation
Models and Persistence
Entity Extraction
Relation Extraction
Pipeline Integration
Example
Conclusion

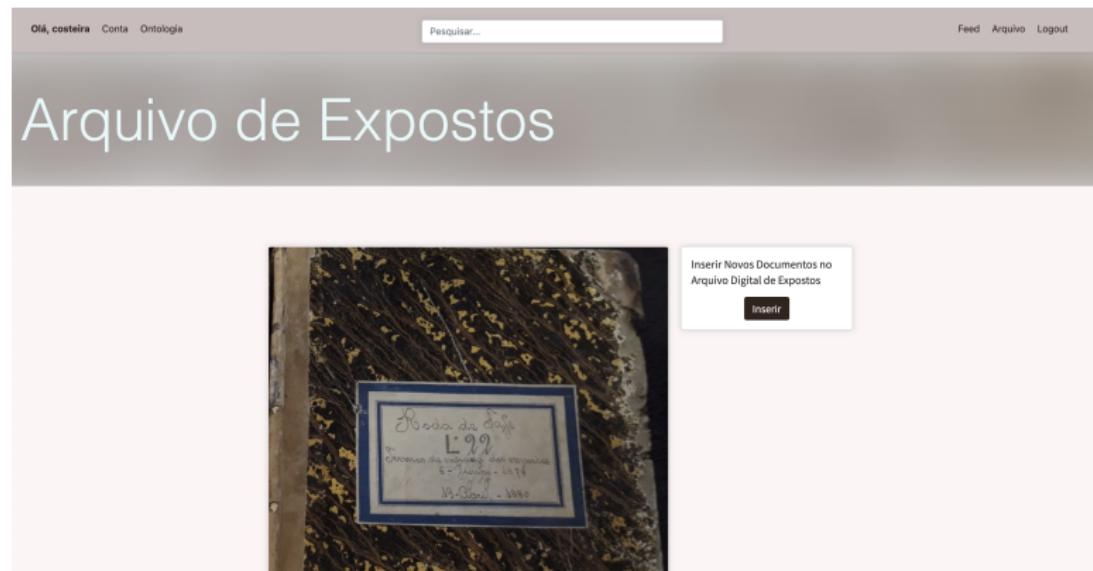


Figure: Feed Page

Navigation Example

- Contextualization
- Motivation
- Objectives
- Documents
- Architecture
- Implementation
 - Models and Persistence
 - Entity Extraction
 - Relation Extraction
 - Pipeline Integration
- Example
- Conclusion

The screenshot shows a web application interface. At the top, there is a navigation bar with links for 'Olá, costeira', 'Conta', 'Ontologia', a search bar labeled 'Pesquisar...', and buttons for 'Feed', 'Arquivo', and 'Logout'. The main title 'Arquivo de Expostos' is displayed prominently. Below the title, a green success message box says 'Folio eliminado com sucesso' with a close button 'X'. The main content area is a table titled 'Documentos Disponíveis' (Available Documents). The table has columns for 'Título' (Title), 'Cota' (Cote), and an 'Inserir' (Insert) button. The data in the table is as follows:

Título	Cota	Inserir
Livro de Termos de entrada de Expostos - Cabeceiras de Basto – Entrada dos Expostos	1-1-27-48	
Nº9 - Entradas dos Expostos na Roda de Celorico de Basto.	1-1-24-26	
Nº8 - Entradas dos Expostos na Roda de Celorico de Basto.	1-1-24-25	
Nº8 Roda de Celorico de Basto – Matricula dos Expostos.	1-1-24-37	
Nº2 Celorico - Matricula.	1-1-24-31	

Figure: Book Table

Navigation Example

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

The screenshot shows a web-based transcription interface. At the top, there is a navigation bar with links for 'Olá, costeira' (Hello, costeira), 'Conta', 'Ontologia', a search bar labeled 'Pesquisar...', and user options like 'Feed', 'Arquivo', and 'Logout'. Below the navigation bar, the breadcrumb trail indicates the current location: 'Livros / Livro / Transcrição'. The main content area is titled 'Transcrição processada' (Processed transcription). The text of the transcription is as follows:

Nº126 de 1862. Termo de **entrada EVENTOS** do Exposto **Guergório PESSOAS**. Aos vinte e seis de Novembro de mil oitocentos sessenta e dois, neste **Concelho de Cabeceiras de Basto LOCALIS** e Paços do mesmo, perante mim Escrivão compareceu a rodeira, e por ela foi dito que pelas nove horas da noite antecedente apareceu lançado na **roda INSTITUIÇÃO** o Exposto **Guergório PESSOAS** com os seguintes enxovalis uma camisa de pano cru folhos de morim nova, outra dita de morim folhos do mesmo, outra dita de pano cru folhos de tremola velha, dois coeiros de linho velhos, dois ditos de chita preta, um dito saiote velho, e um vestido de baetilha velha, um cinto de seragoça velho, um lenço de três pontas risca amarela, uma fita de lã vermelha velha, atada no pulso esquerdo, mostra ter de idade três dias e não tem sinais particulares. Foi **baptizado EVENTOS** na **Igreja de Refojos LOCALIS** foram Padrinhos **Policarpo Pereira PESSOAS** e **Rita Joaquina PESSOAS**. E para constar fiz este termo, que assina o marido da rodeira que assina comigo. Manoel Leite Araújo. Francisco José Alves Pacheco. **Faleceu EVENTOS** em 29 de Agosto de 1863.

Figure: Transcription Page

Navigation Example

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

Informação Pessoal

Nome: Guergório ♂
Género: Masculino ♂
Função Pessoa: Exposto ⓘ
Identificador: 126 , do ano 1862 ⓘ
Participa: Entrada
Possui o Seguinte Enxoaval:

dois ditos de chita preta
morim folhos
saiote velho
cinto de seragoça velho
lenço de três pontas risca amarela
dois coeiros de linho velhos
vestido de baetilha velha
pano cru folhos de tremola velha
fita de lã vermelha velha
camisa de pano cru folhos de morim nova

Figure: Personal Information

Informação de Evento

Tipo de Evento: Entrada
Data do Evento: 1862-11-26 ⓘ
Ocorreu: Cabeceiras de Basto ⓘ
Comparece: Guergório ♂

Figure: Event Card

Navigation Example

Motivation

Documents

Architecture

Example

Termo de Encerramento	
Tem este livro cento e dezassete meias folhas que ficão numeradas e rubricadas com a minha rubrica de A. Rodrigues- que uso e costumo.	
Observações	
Contém documentos soltos	
Datas Extrema Inicial	
14/11/1862 <input type="checkbox"/>	
Datas Extrema Final	
12/05/1869 <input type="checkbox"/>	
Dimensões	
29,5cm x20,3cm+2cm	
 <input type="button" value="Mudar Capa"/>	

Figure: Update Book Form

Conclusion

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

- Research Hypothesis verified, by the development of a knowledge repository that supports a digital platform.
- The proposed solution automates the extraction of knowledge from documents, resorting to the integration of Natural Language components.

Conclusion

Contextualization

Motivation

Objectives

Documents

Architecture

Implementation

Models and Persistence

Entity Extraction

Relation Extraction

Pipeline Integration

Example

Conclusion

As future Work, a set of improvements were identified:

- Expand the knowledge repository,
- Incorporate an annotation tool in the application,
- SPARQL engine for an advanced user base,
- Integration with similar projects

Contextualization
Motivation
Objectives
Documents
Architecture
Implementation
Models and Persistence
Entity Extraction
Relation Extraction
Pipeline Integration
Example
Conclusion

Exploration of documents concerning Foundlings in Fafe along XIX Century

João Costeira Faria Gomes

Universidade do Minho

2022

