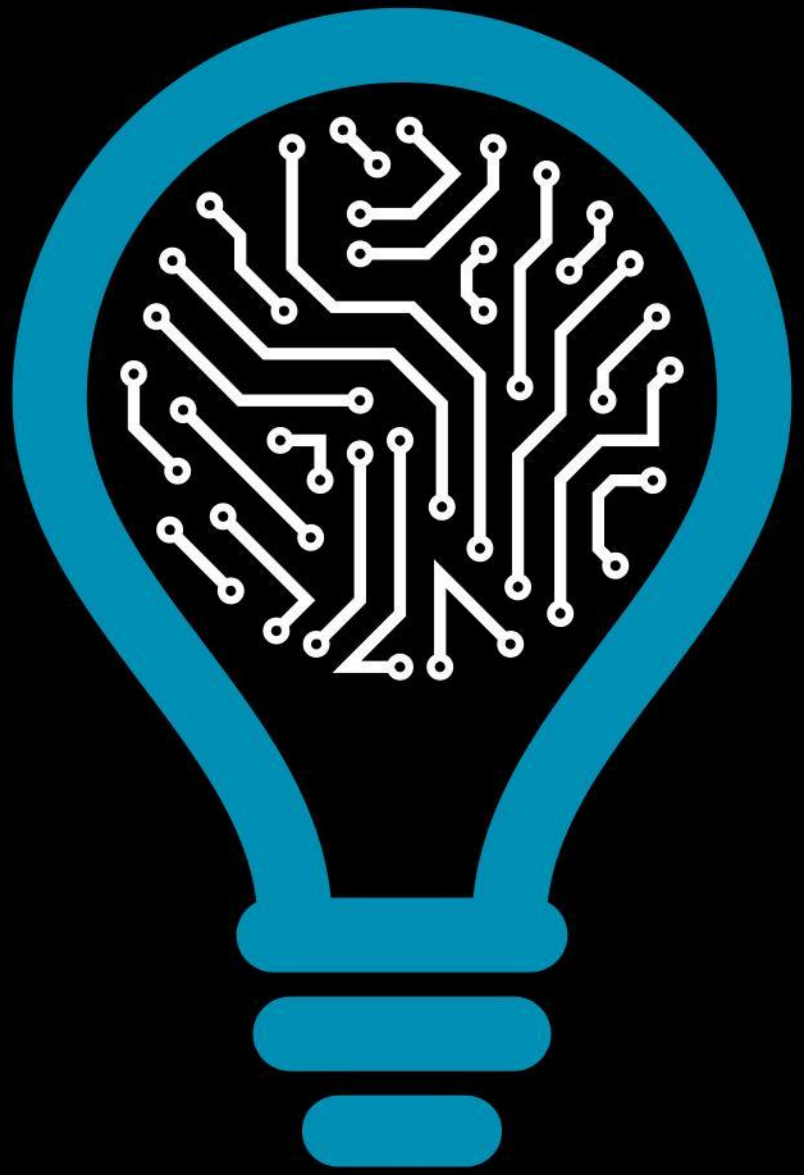
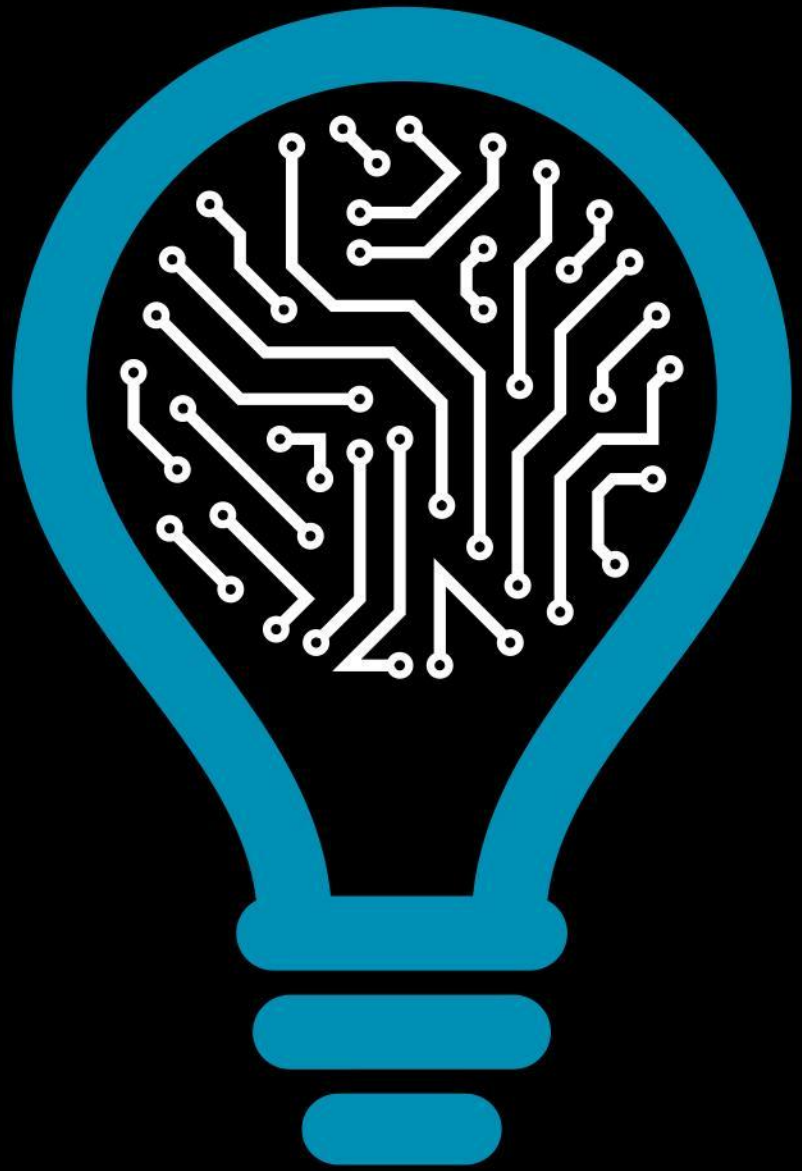
from knowledge production to science-based innovation



INESCTEC

INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE

Linked Data for the Archives: Models, Technologies and Applications



DCMI Virtual 2021
October 7, 2021

Carla Teixeira Lopes, Cristina Ribeiro
INESC TEC, University of Porto

INESCTEC
INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE

INESCTEC

Outline

EPISA: the view

EPISA: the project

Linked Data: distributed wisdom

Ontologies: where it all starts

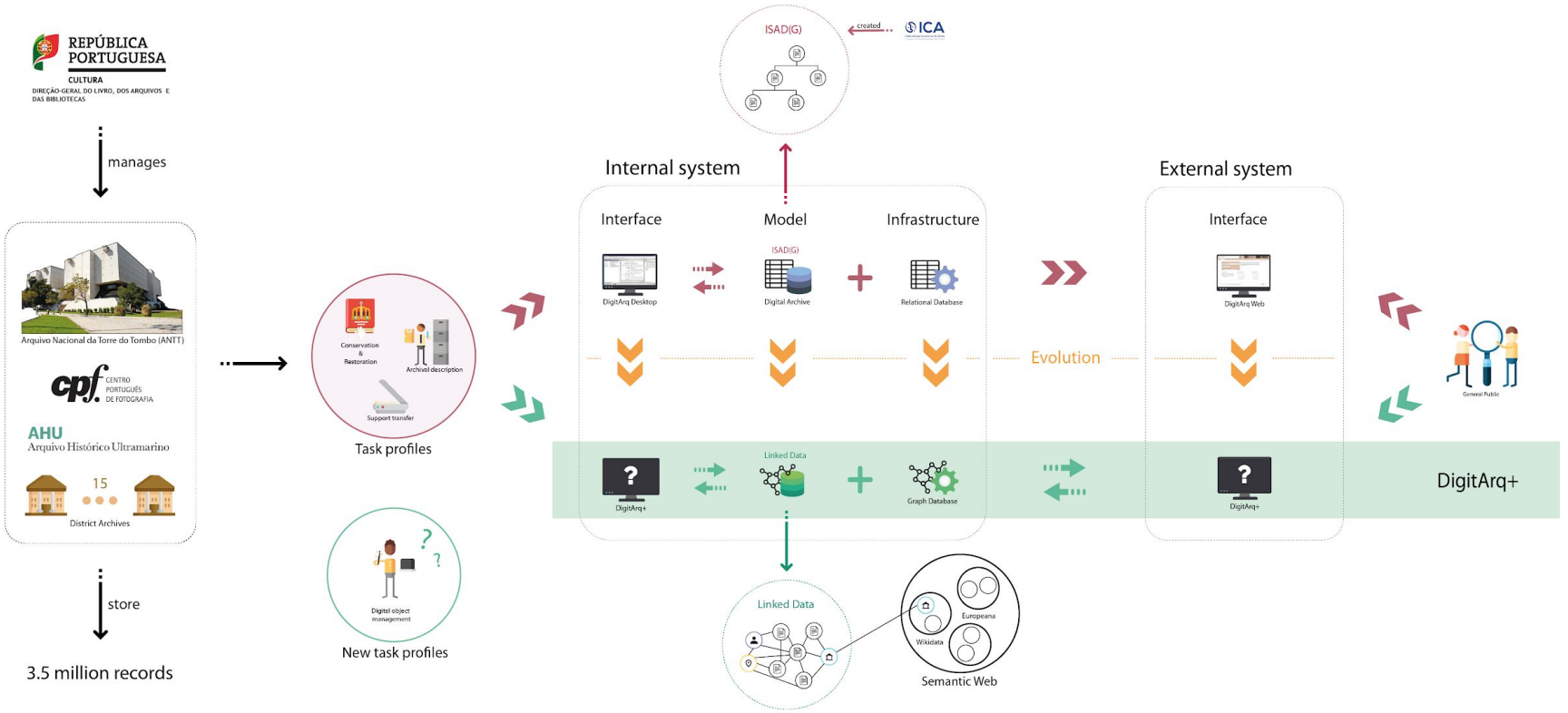
Technologies: how it comes true

Tradeoffs and challenges

INESCTEC

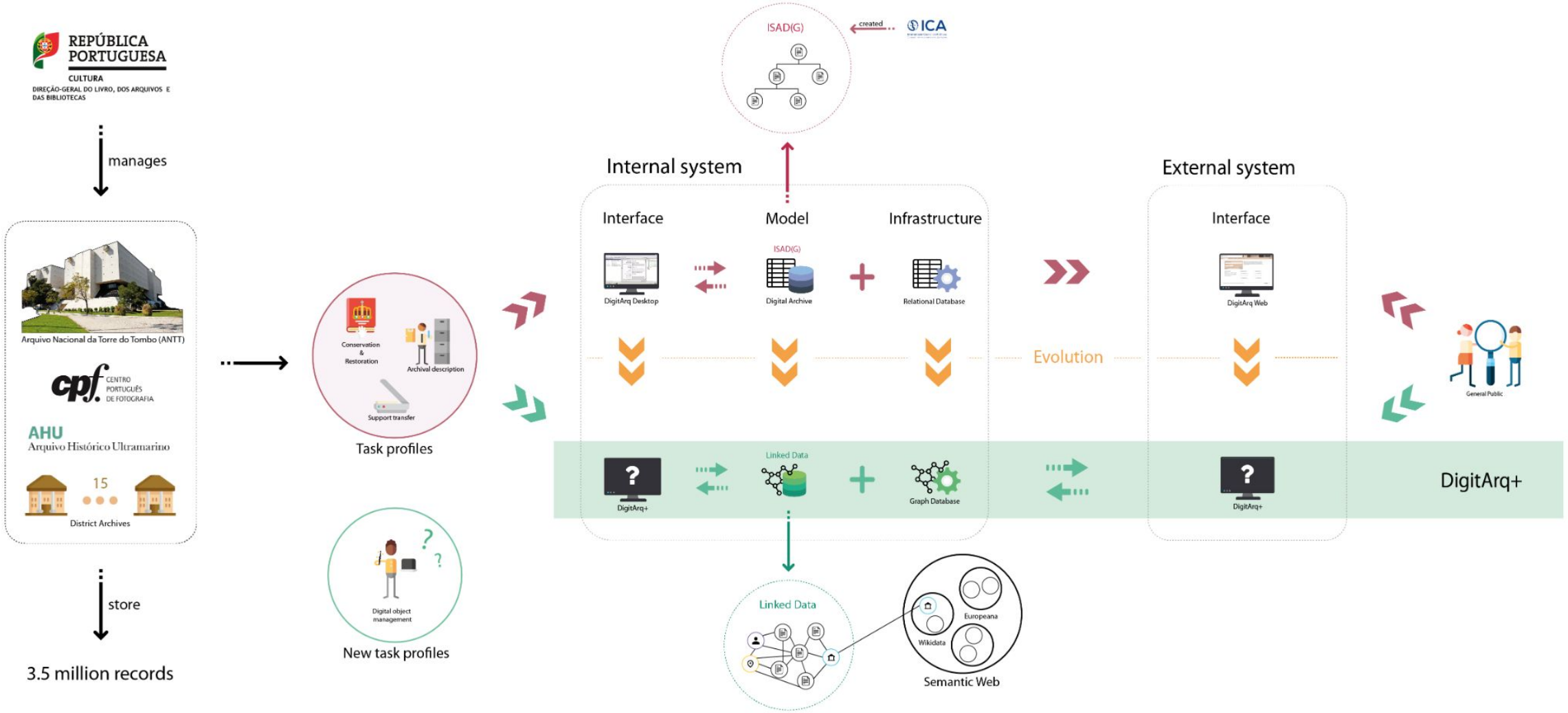
Evolving Archival Systems

Portuguese National Archives and Digitarq

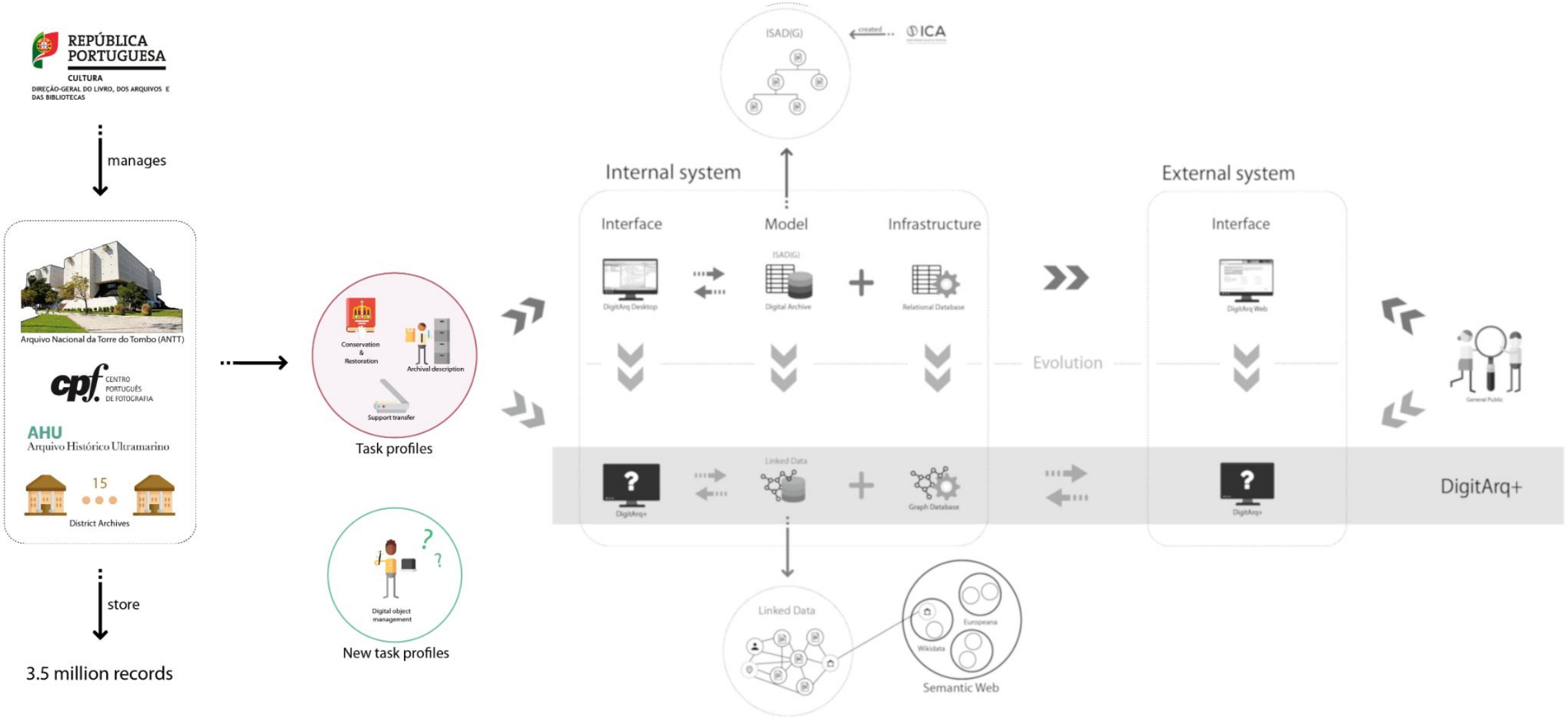


EPISA Project			
Author: Cláudia Guedes	Content: From DigitArq to DigitArq+	Version: 1	Data: 11/03/2020
Led by: INESCTEC	Partners: DGLAB	Financed by: FCT	

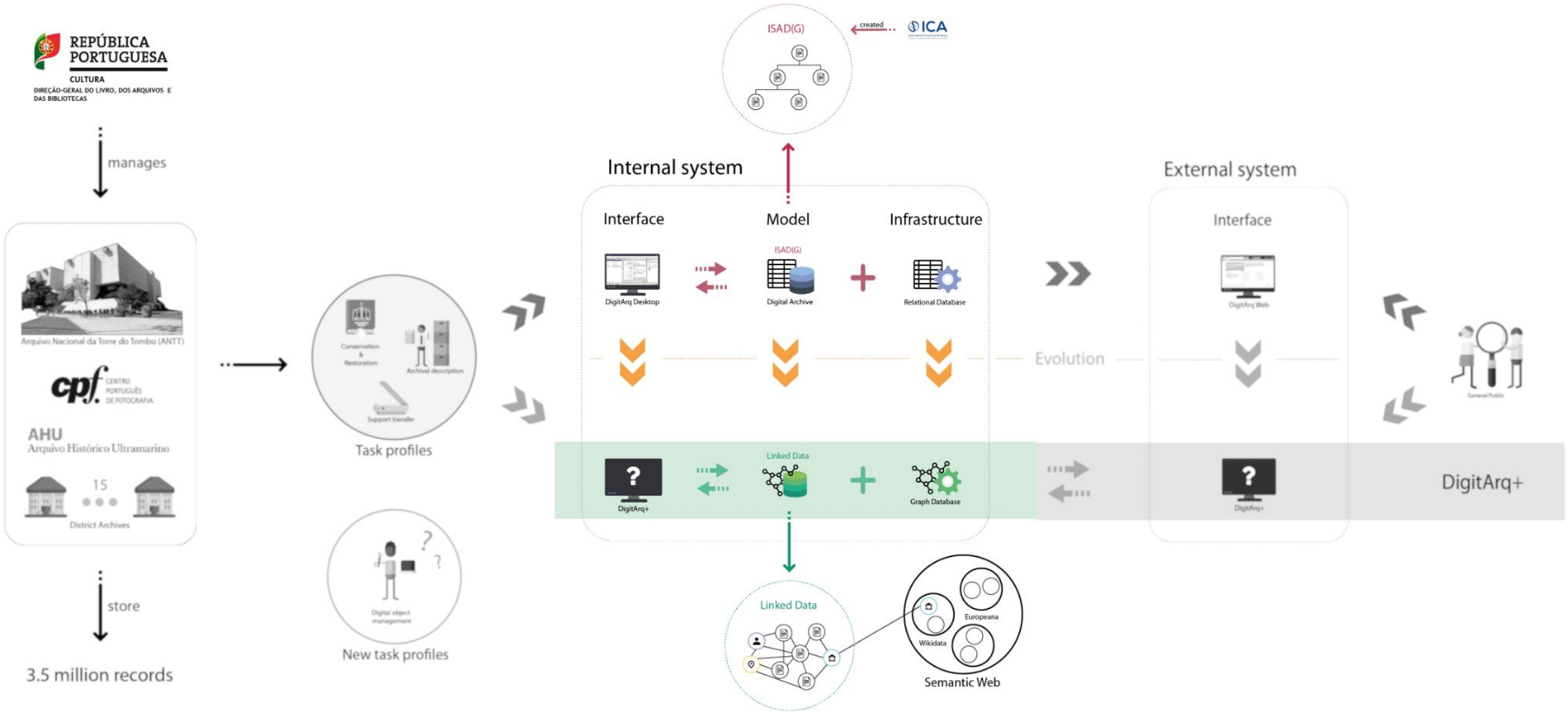
Portuguese National Archives and Digitarq



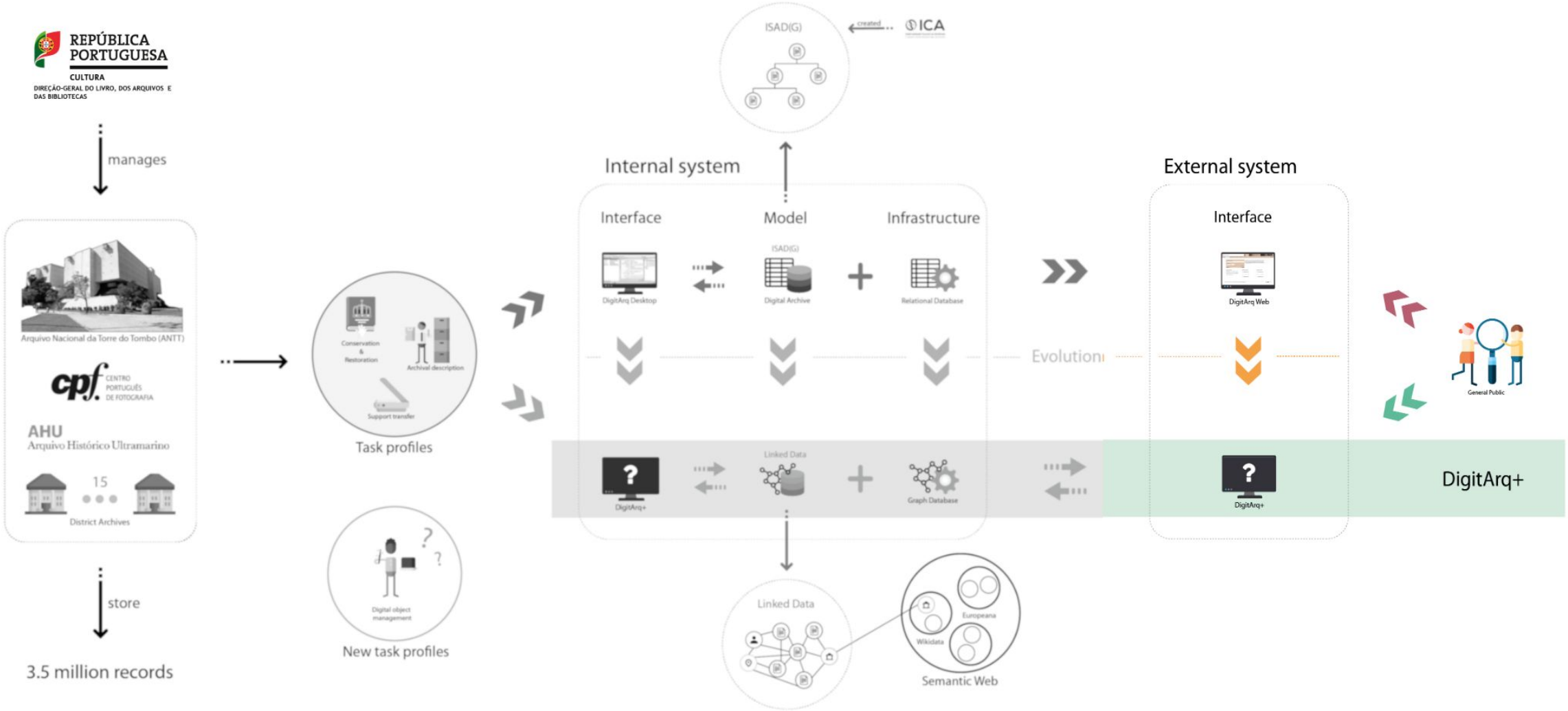
The tasks at the Archives



Digitarq and Digitarq+



External interface



EPISA partners and resources

Scholarships

3 graduates
2 Master
1 postdoc

Students

Graduate projects
Master dissertations

Academic

6 part-time
(~2 full-time equivalent)

Specialists

3 part-time
(~1 full-time equivalent)

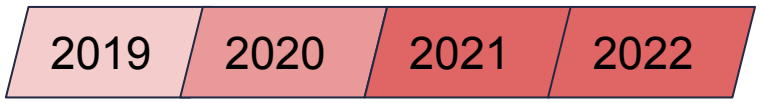
Partners

DGLAB
INESC TEC
Univ. Évora

Budget

~300 k€
Scholarships + equipment

Plan



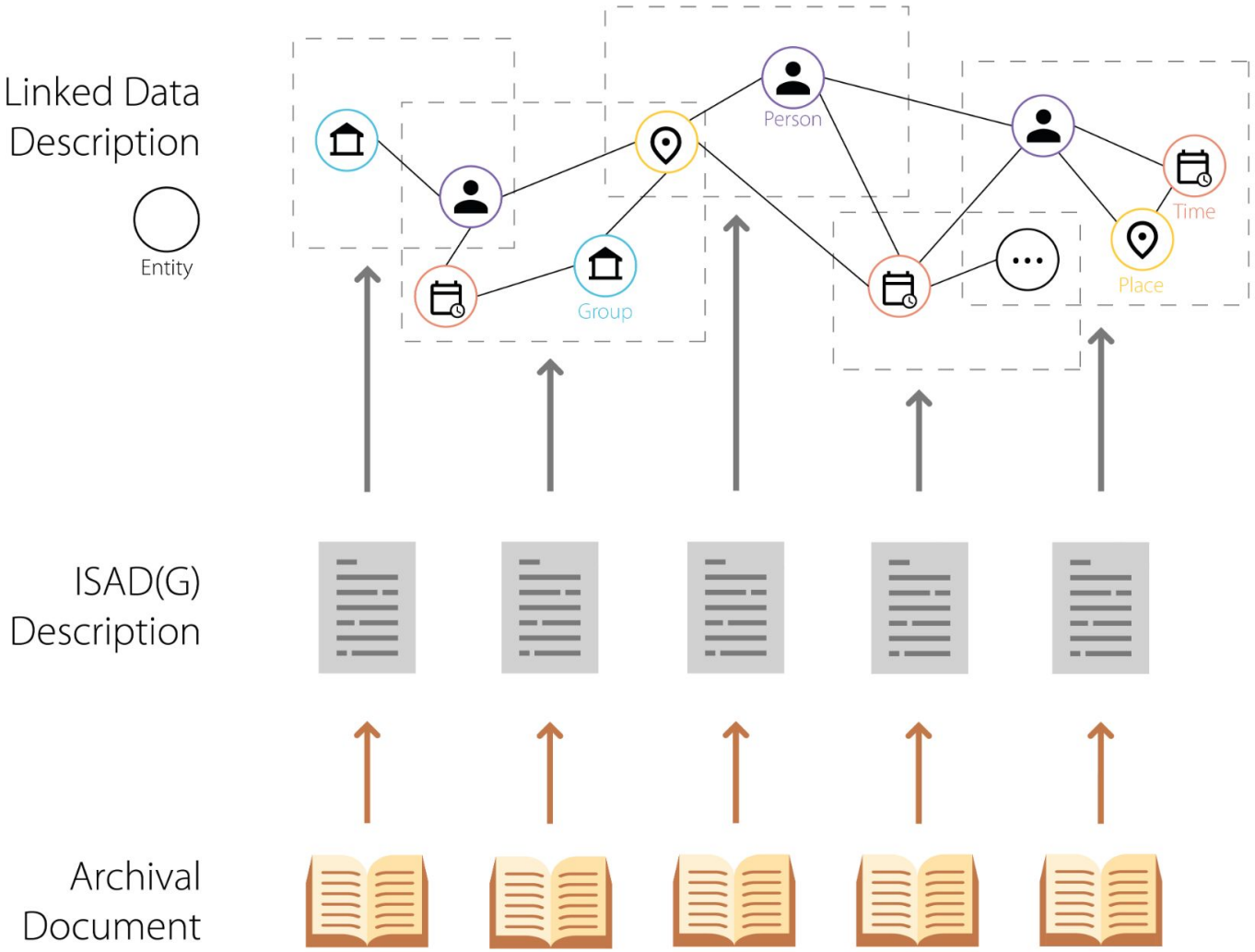
Deliverables

Papers
Theses
Prototypes

INESCTEC

Linked Data in the Archives

Descriptive standards vs Linked Data



Semantic view of the archival record

ISAD(G) Record

PROCESSO DE INVENTÁRIO ORFANOLÓGICO POR ÓBITO DE MARIA HENRIQUETA FRAGOSO BARAHONA CARVALHO E MIRA

DESCRIPTION LEVEL
File

REFERENCE CODE
PT/ADEVR/JUD/TCEVR/A/001/02503

TITLE TYPE
Atribuído

DATE RANGE
1885-06-20 ✔ a 1885-06-20 ?

DIMENSION AND SUPPORT
1 proc; papel

EXTENTS
717 Folhas

GEOGRAPHIC NAME
Évora

SCOPE AND CONTENT
Inventariada: Maria Henriqueta Fragoso Barahona Carvalho e Mira, casada com José Paulo de Barahona Carvalho e Mira, residentes na cidade de Évora.

A inventariada não fez testamento.

Data de óbito: 1885-05-28

Inventariante: José Paulo de Barahona Carvalho e Mira

Filhos:

1-José Barahona Fragoso Mira, de 10 anos.

2-Maria do Anjo Barahona Fragoso Mira, de 8 anos.

3-Francisco Barahona Fragoso Mira, de 6 anos.

ARRANGEMENT
Ordenação cronológica, do documento mais antigo no tempo para o mais recente.

ACCESS RESTRICTIONS
Ordenação cronológica, do documento mais recuado no tempo para o mais recente.

CONDITIONS GOVERNING USE
Sujeita à tabela de preços emolumentares em vigor

PHYSICAL LOCATION
Tribunal da Comarca de Évora; Secção: Cível; Série:001-Inventários obrigatórios; Cx:249; Dc.2503

PREVIOUS LOCATION
Maço: 204; doc.3112

LANGUAGE OF THE MATERIAL
Português

PHYSICAL CHARACTERISTICS AND TECHNICAL REQUIREMENTS
Bom estado de conservação

NOTES
A inventariante fez contrato antenupcial.

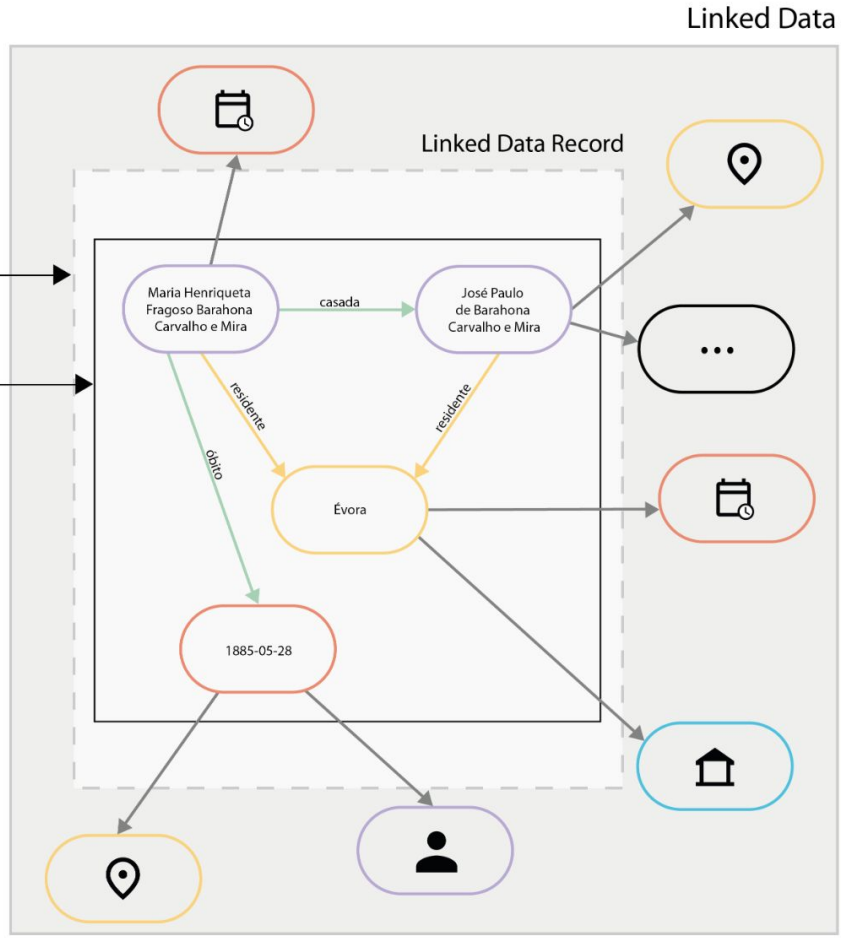
CREATION DATE
28/07/2017 10:25:40

LAST MODIFICATION
19/02/2019 16:20:30

Record not reviewed.

Atomization

Linked Data

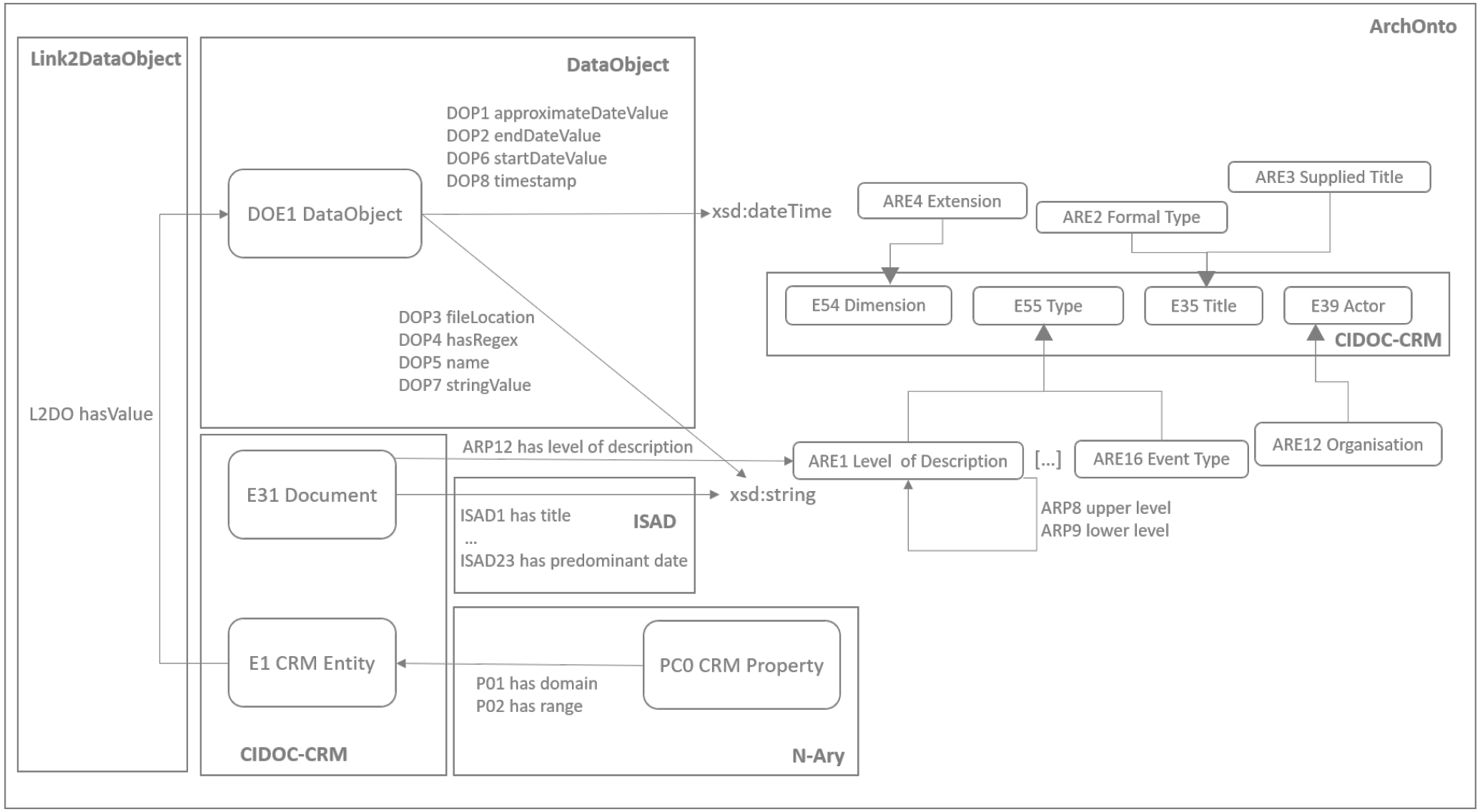


EPISA Project
Author: Claudia Guedes
Led by: INESCTEC
Content: Atomization of an ISAD(G) Record in Linked Data
Version: 1
Data: 11/03/2020
Partners:
Financed by: FCT

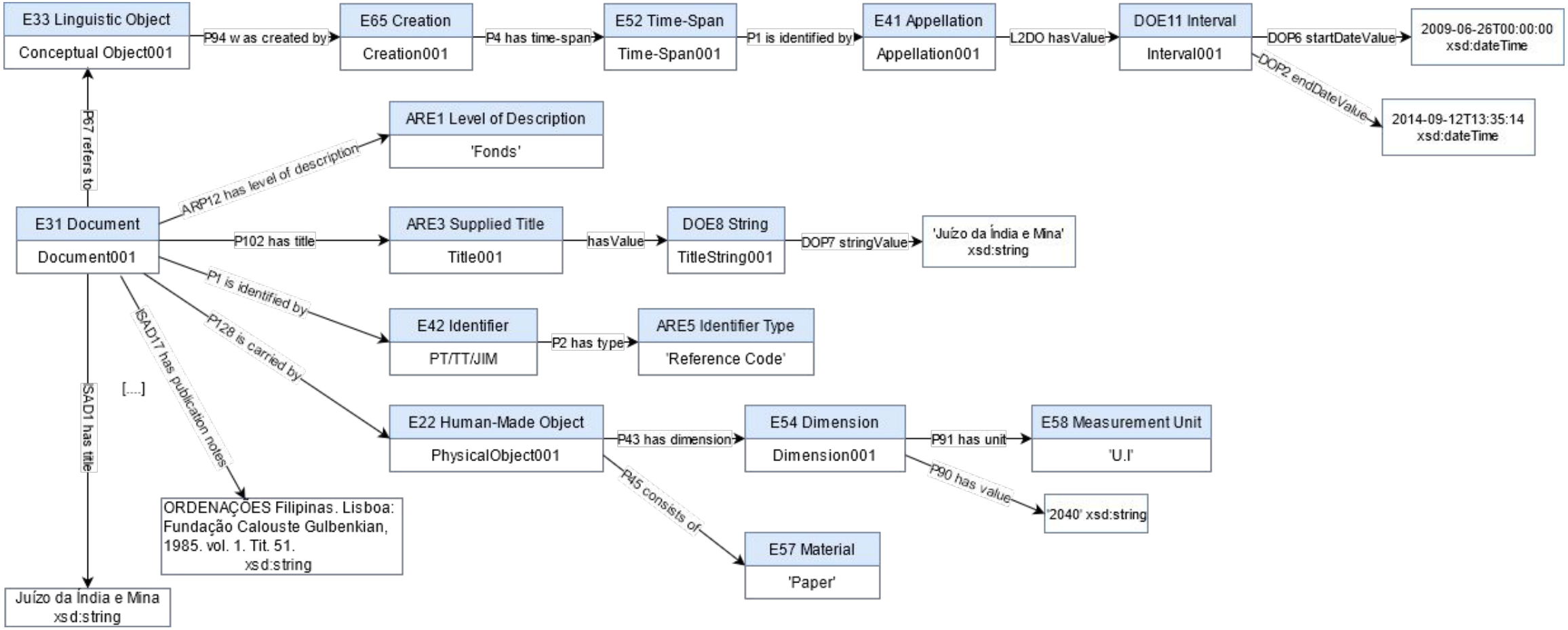
INESCTEC

Ontologies in Archives

ArchOnto

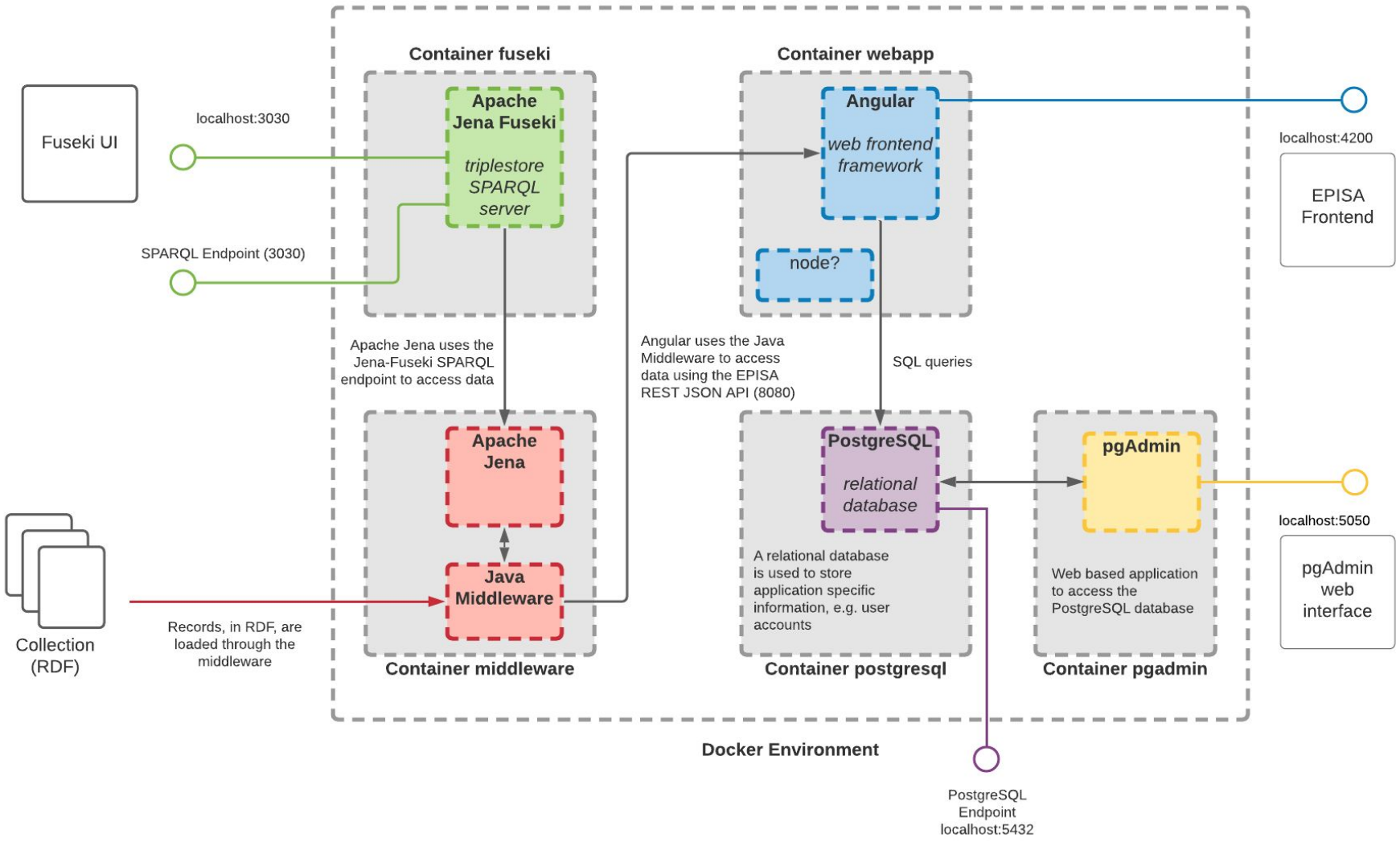

ArchOnto
Link2DataObject
DataObject
DOP1 approximateDateValue
DOP2 endDateValue
DOP6 startDateValue
DOP8 timestamp
DOE1 DataObject
xsd:dateTime
DOP3 fileLocation
DOP4 hasRegex
DOP5 name
DOP7 stringValue
L2DO hasValue
ARE4 Extension
ARE3 Supplied Title
ARE2 Formal Type
E54 Dimension
E55 Type
E35 Title
E39 Actor
CIDOC-CRM
ARP12 has level of description
ARE1 Level of Description
[...]
ARE16 Event Type
ARE12 Organisation
E31 Document
xsd:string
ISAD1 has title
...
ISAD23 has predominant date
ISAD
ARP8 upper level
ARP9 lower level
E1 CRM Entity
PC0 CRM Property
P01 has domain
P02 has range
CIDOC-CRM
N-Ary

ArchOnto example

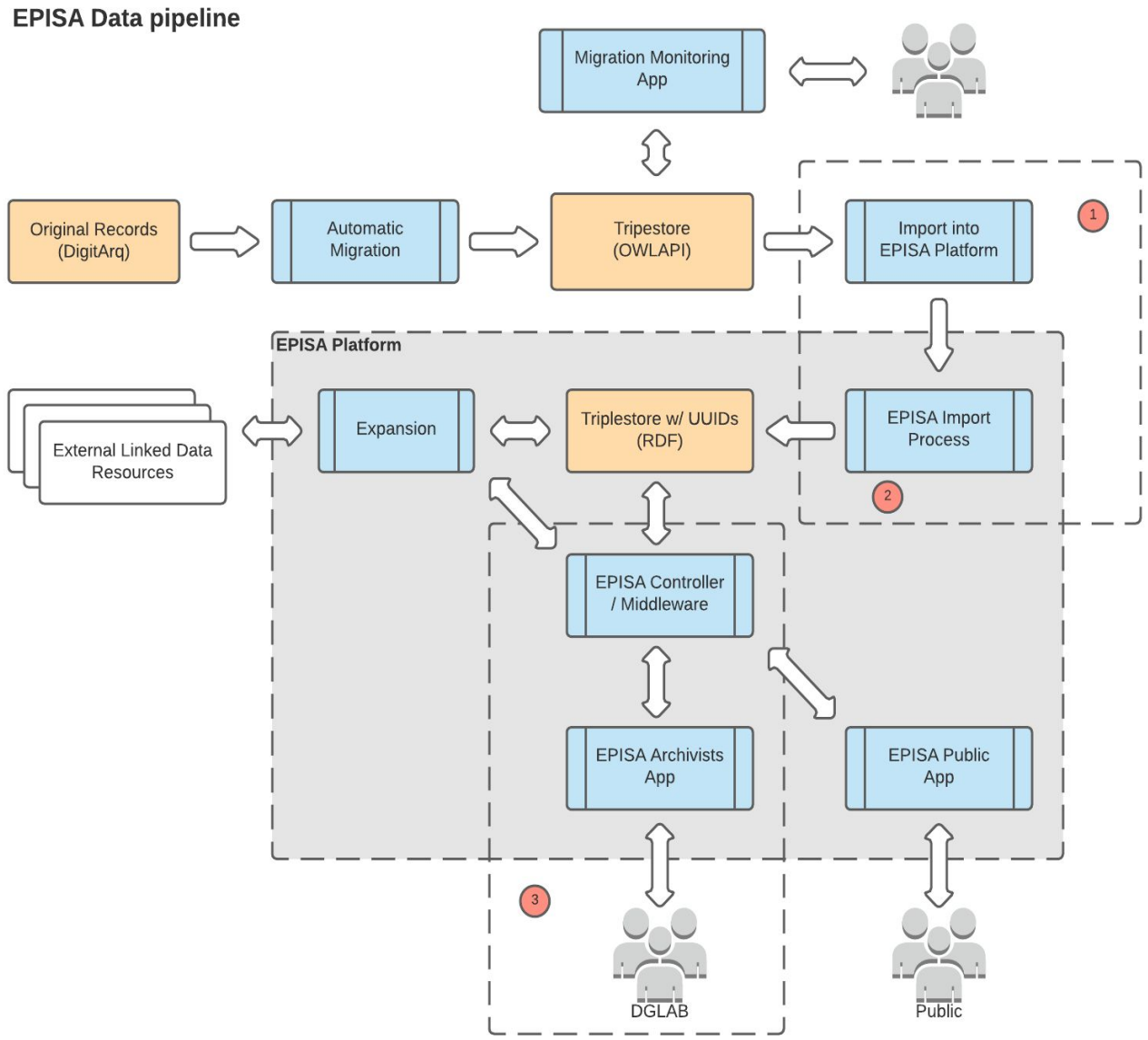


INESCTEC

Technologies for Knowledge Graphs


Container fuseki
Container webapp
Fuseki UI
localhost:3030
Apache Jena Fuseki
triplestore SPARQL server
Angular
web frontend framework
localhost:4200
EPISA Frontend
SPARQL Endpoint (3030)
node?
Apache Jena uses the Jena-Fuseki SPARQL endpoint to access data
Angular uses the Java Middleware to access data using the EPISA REST JSON API (8080)
SQL queries
Apache Jena
PostgreSQL
relational database
pgAdmin
localhost:5050
pgAdmin web interface
Java Middleware
A relational database is used to store application specific information, e.g. user accounts
Web based application to access the PostgreSQL database
Collection (RDF)
Records, in RDF, are loaded through the middleware
Container middleware
Container postgresql
Container pgadmin
Docker Environment
PostgreSQL Endpoint localhost:5432

Data Flow



1 May be integrated in a single step. Shown as two separate processes to reflect the fact that the development is distributed by two teams.

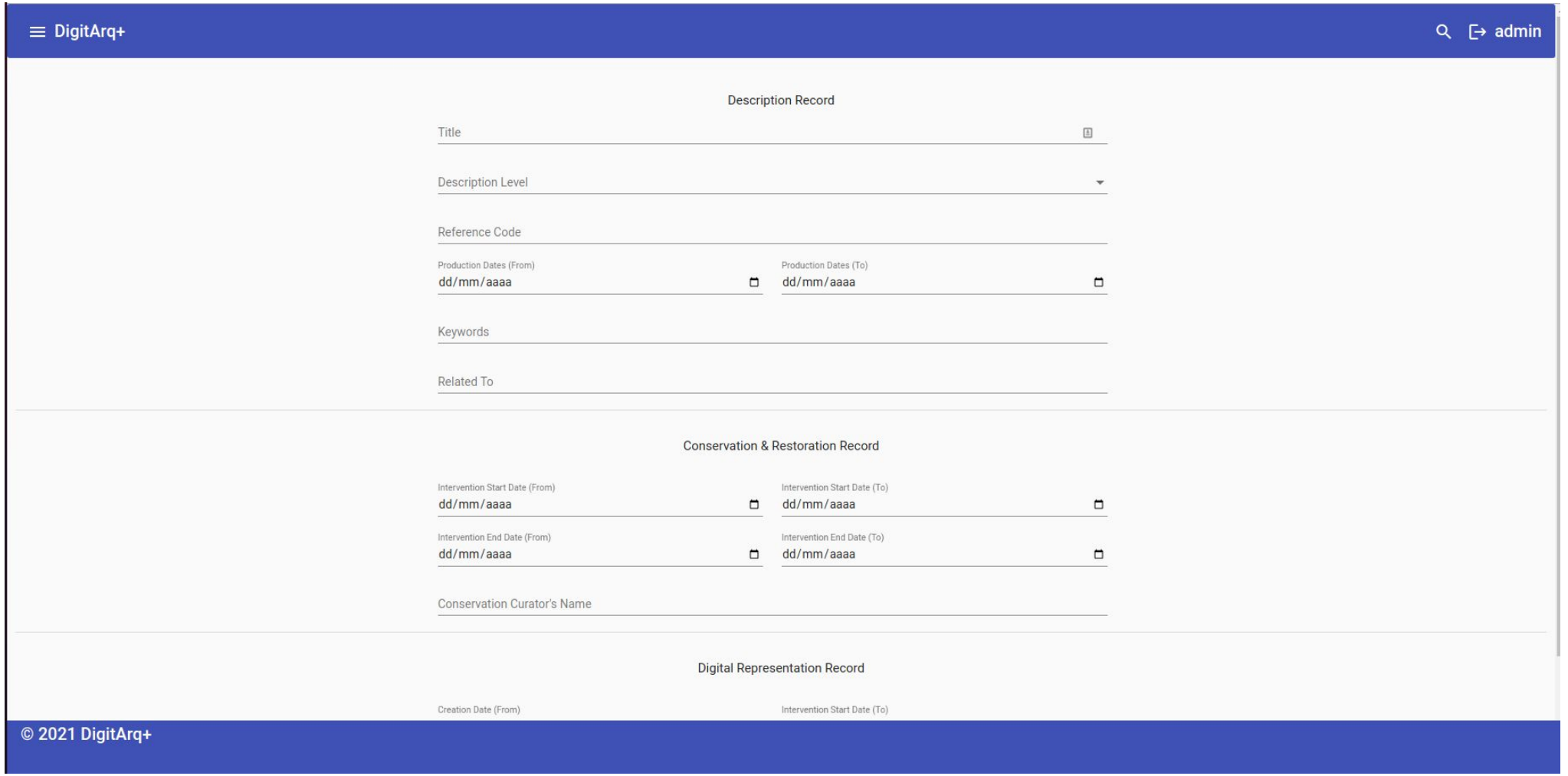
2 The EPISA 'import process' needs to:
- Import the OWL to the triplestore;
- Attribute a UUID to each object;
- Log the import process;

3 This process needs to be detailed to include additional steps, namely:

- Incorporate "migration" functions to help parse texts;

- Incorporate the functions to parse digital representations;

Prototype

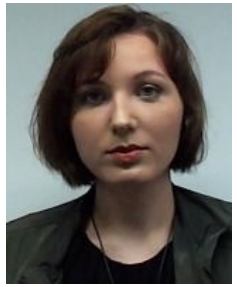

DigitArq+
admin
Description Record
Title
Description Level
Reference Code
Production Dates (From)
dd/mm/aaaa
Production Dates (To)
dd/mm/aaaa
Keywords
Related To
Conservation & Restoration Record
Intervention Start Date (From)
dd/mm/aaaa
Intervention Start Date (To)
dd/mm/aaaa
Intervention End Date (From)
dd/mm/aaaa
Intervention End Date (To)
dd/mm/aaaa
Conservation Curator's Name
Digital Representation Record
Creation Date (From)
Intervention Start Date (To)
© 2021 DigitArq+

INESCTEC

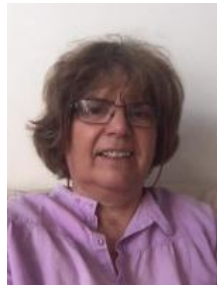
Team

Research Team

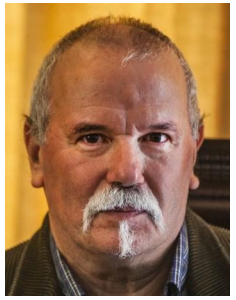
INESC TEC



Univ. Évora



DGLAB



Recent Publications

- Davide Varagnolo, Dora Melo, Irene Pimenta Rodrigues. **A Tool to Explore the Population of a CIDOC-CRM Ontology**. In: 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. 2021.

- Lázaro Costa, Nuno Freitas, João Rocha da Silva. **An evaluation of Graph Databases and Object-Graph Mappers in CIDOC CRM-compliant digital archives**. Special Issue on "Computational Archival Science" of the ACM Journal on Computing and Cultural Heritage (JOCCH). [Pending publication]

- Cláudia Guedes, Bruno Giesteira, Sérgio Nunes. **Designing User Interaction with Linked Data in Historical Archives**. Special Issue on "Computational Archival Science" of the ACM Journal on Computing and Cultural Heritage (JOCCH). [Pending publication]

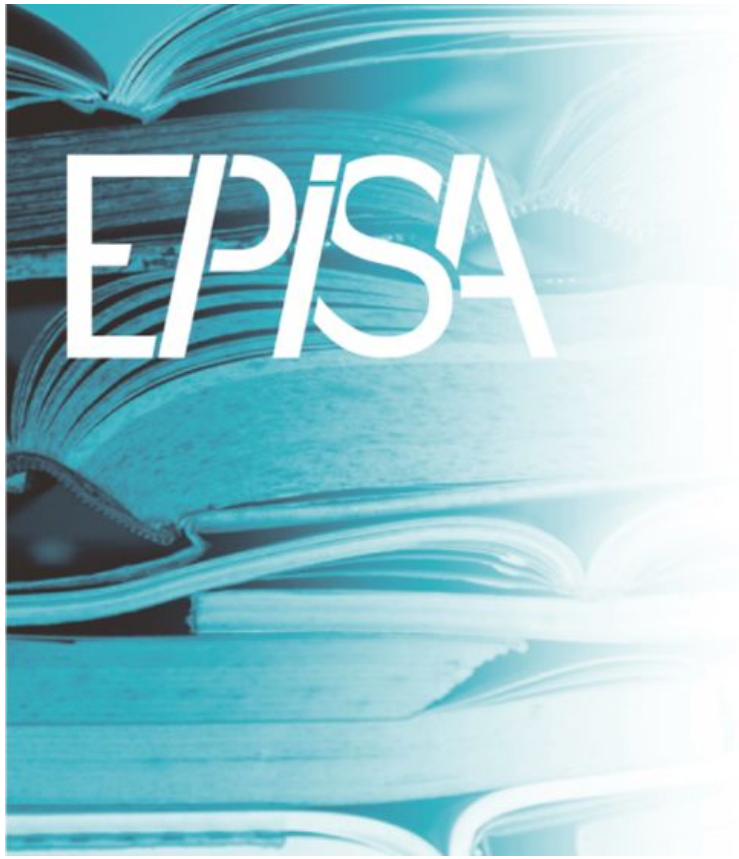
- Inês Koch, Carla Teixeira Lopes, Cristina Ribeiro. **Moving from ISAD(G) to a CIDOC-CRM-based Linked Data Model in the Portuguese Archives**. ACM Journal on Computing and Cultural Heritage (JOCCH). [Pending publication]

- Marta Araújo, Carla Teixeira Lopes. **How Can an Archive Be Characterized?** TPDL 2021.

Workshop, September: LinkedArchives@TDPL 2021

https://linkedarchives.inesctec.pt/

EPISA website



SOBRE EQUIPA PARCERIAS PUBLICAÇÕES NOTÍCIAS CONTACTOS

Visão Geral

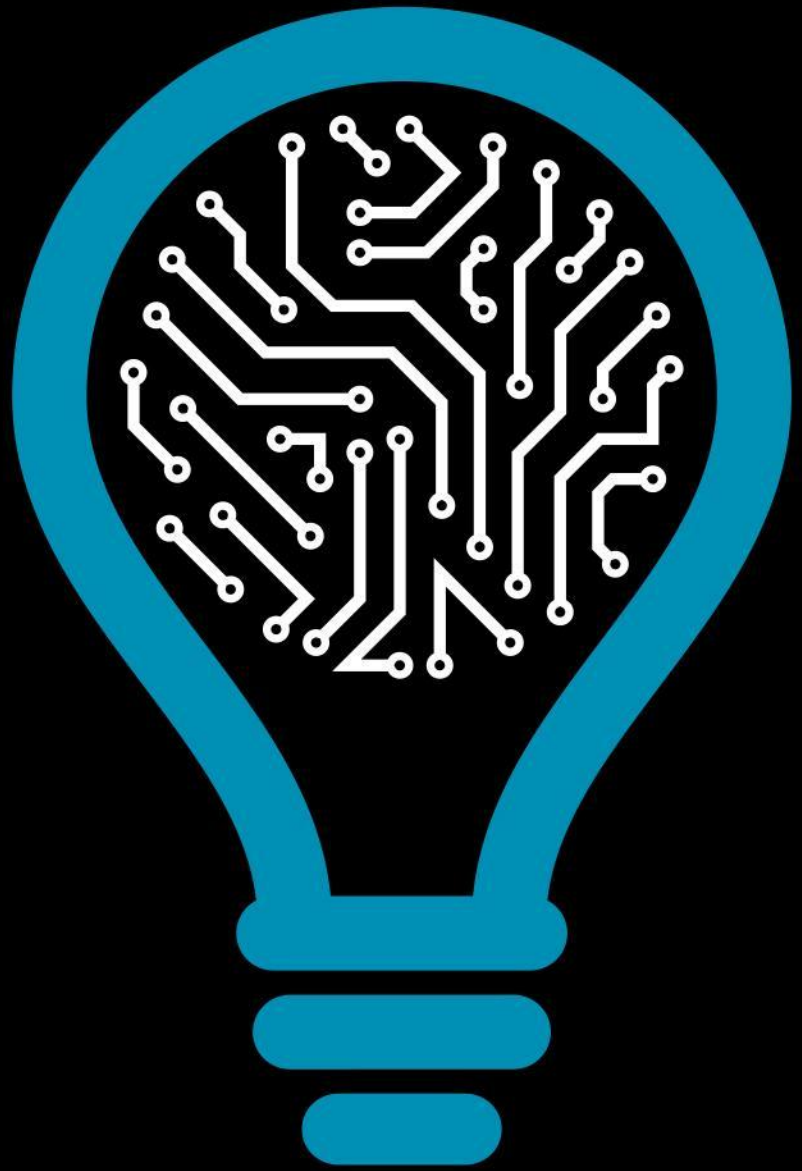
O projeto EPISA – Inferência de Entidades e Propriedades para Arquivos Semânticos – trilha um caminho importante no panorama nacional rumo à acessibilidade do património cultural português e da informação em geral.

Propõe-se a criar um novo modelo de descrição para os arquivos e a promover a criação semiautomática de metadados. O objetivo principal consiste em incorporar os arquivos nacionais na rede global dos dados ligados semânticos.

O projeto está a ser desenvolvido pelo INESC TEC (Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência), em parceria com a Universidade de Évora e a DGLAB (Direção-Geral do Livro, dos Arquivos e das Bibliotecas), sob coordenação das Professoras Carla Lopes e Cristina Ribeiro.

https://episa.inesctec.pt/

Linked Data for the Archives: Models, Technologies and Applications



DCMI Virtual 2021
October 7, 2021

Carla Teixeira Lopes, Cristina Ribeiro
INESC TEC, University of Porto

INESCTEC
INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE