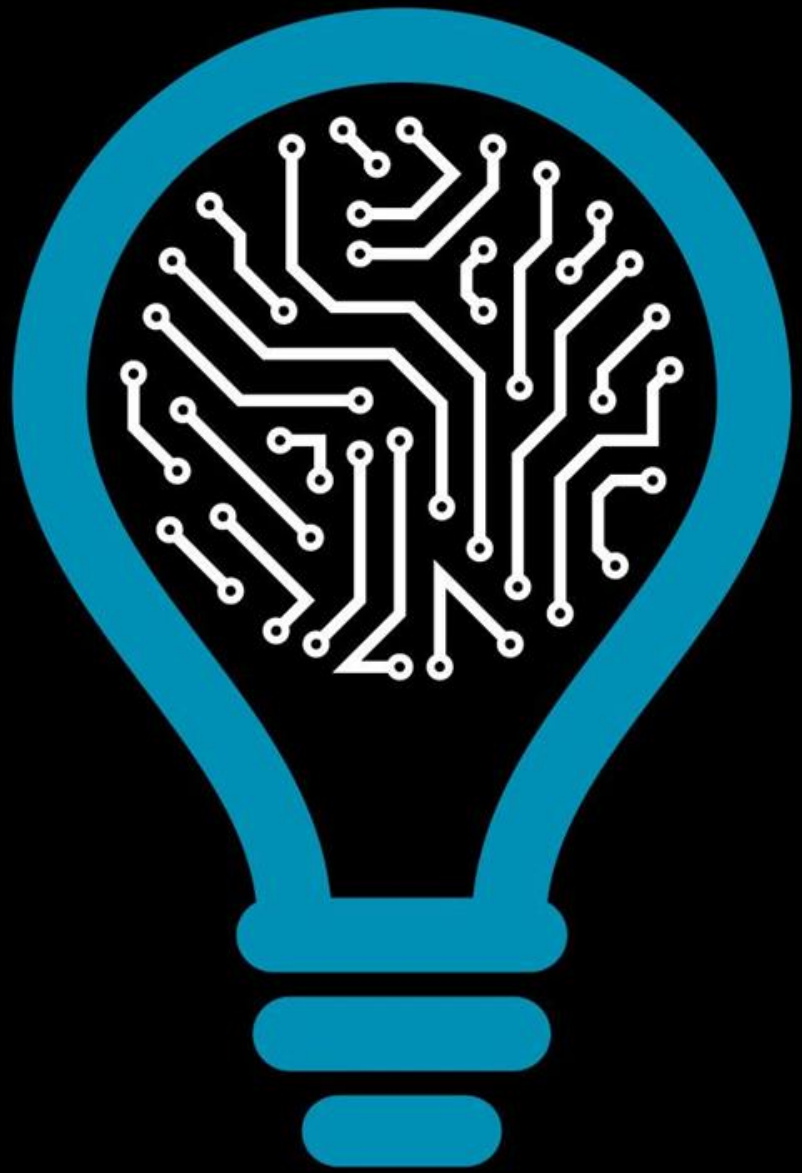
EPISA - ENTITY AND PROPERTY INFERENCE FOR SEMANTIC ARCHIVES

55th CIDOC CRM and 48th FRBR CRM

Inês Koch
Carla Teixeira Lopes
Cristina Ribeiro
Sérgio Nunes

This work is financed by National Funds through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia within project DSAIPA/DS/0023/2018. Inês Koch is also financed by FCT within the research grant 2020.08755.BD.



INESCTEC

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Inês Koch

Ph.D. candidate - Doctoral Programme in Digital Media

EPISA Project - Entity and Property Inference for Semantic Archives

Master in Information Science

Porto, Portugal

Outline

- EPISA Project
- ArchOnto
- EPISA Platform

INESCTEC

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

EPISA

Entity and Property Inference for Semantic Archives

EPISA Project

- Part of the ongoing renewal of the The Portuguese National Archives' existing data infrastructure

- Goals:
  - Develop a prototype for an open-source knowledge graph platform representing archival information on a linked data model.
  - Find ways to guarantee the migration of contents stored according to ICA (International Council on Archives) standards to an ontology-based model, the CIDOC CRM (Conceptual Reference Model).

EPISA team

INESC TEC: Cristina Ribeiro, Carla Teixeira Lopes, Gabriel David, Sérgio Nunes, Bruno Giesteira, Inês Koch, Mariana Dias, Cláudia Martins, Tiago Silva, Tiago Gomes, Catarina Pires, Camilla Oliveira, Pedro da Silva

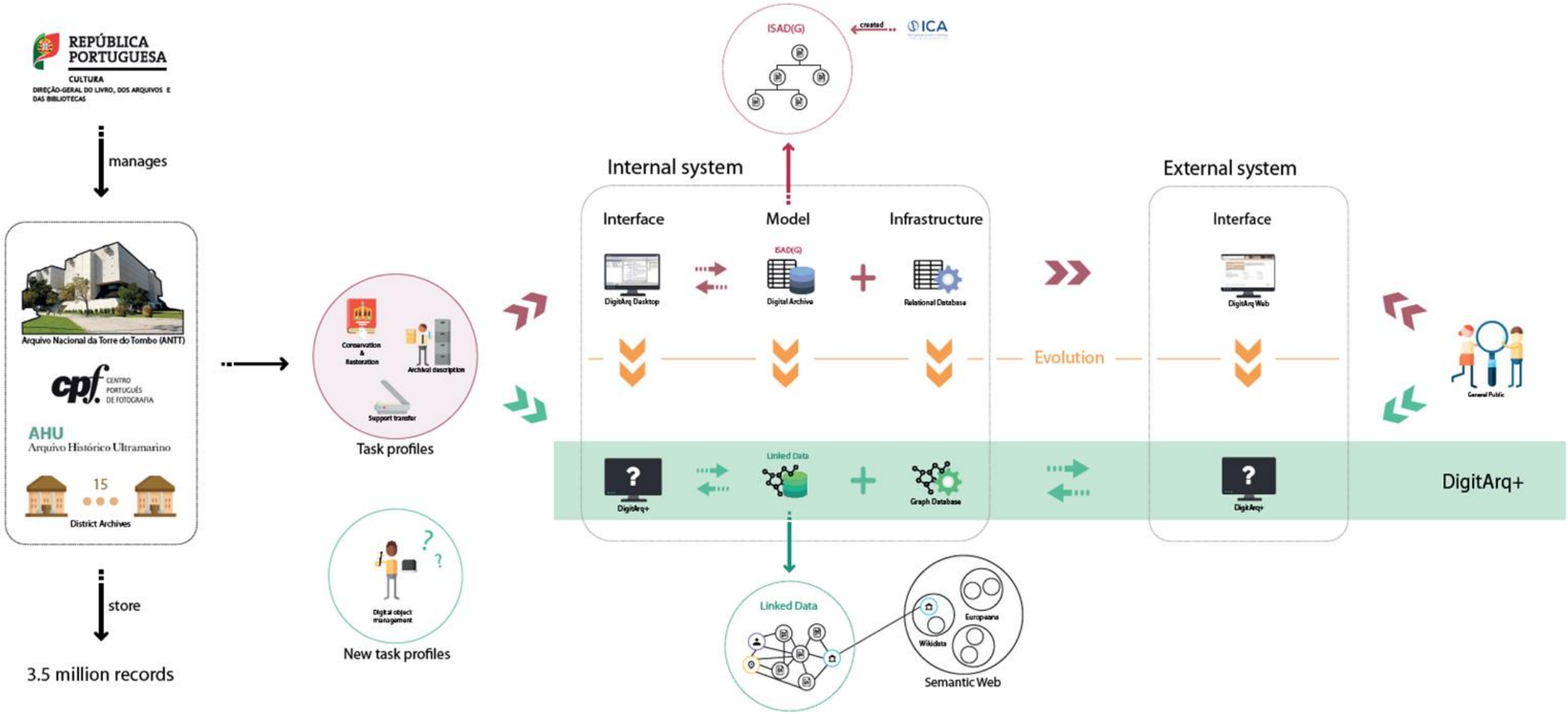
U.Évora: Irene Rodrigues, Dora Melo, Davide Varagnolo

DGLAB: Ana Maria Rodrigues, Lucília Runa, Maria José Almeida, Cristiane Casaca, Silvestre Lacerda

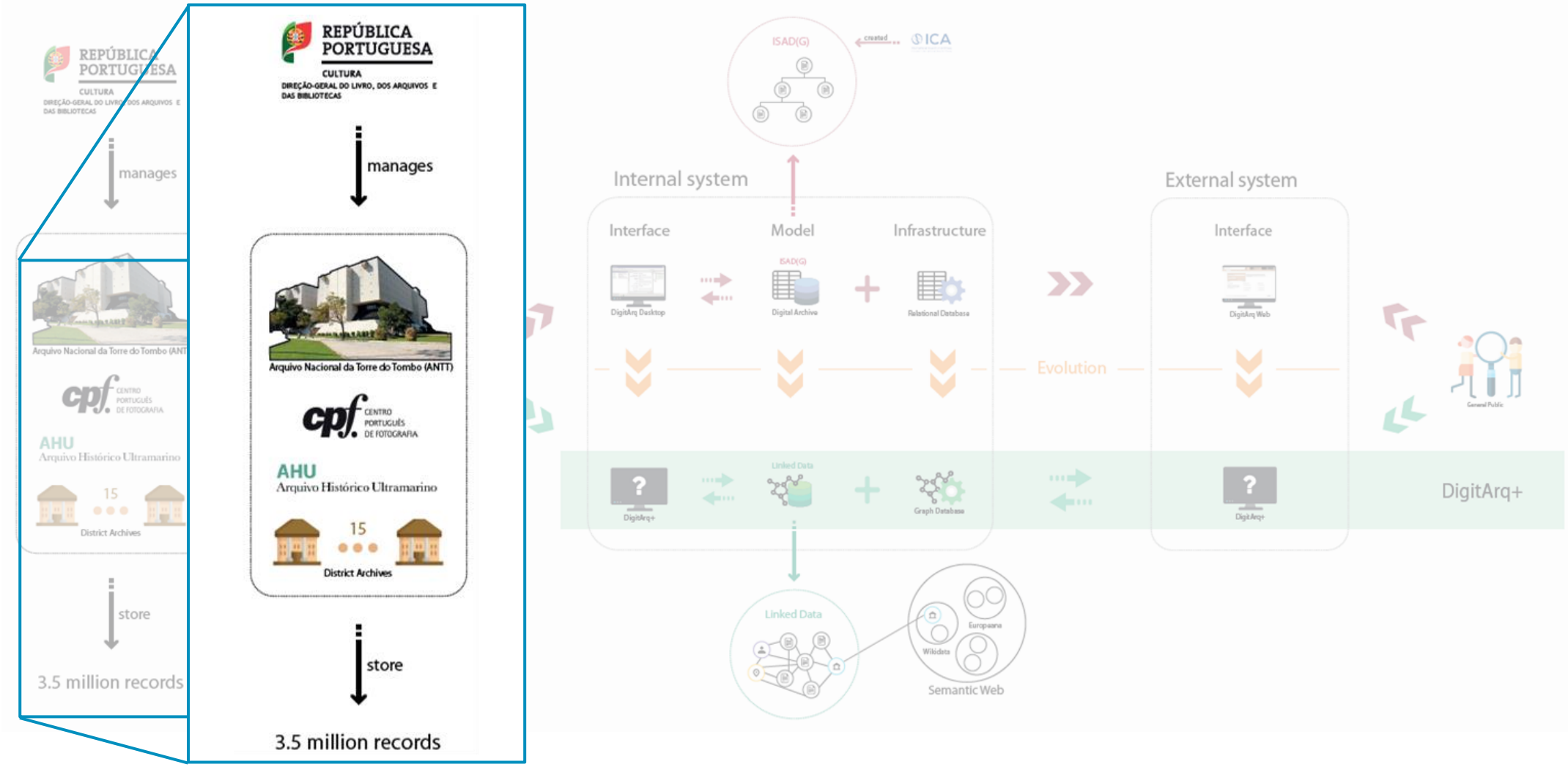
EPISA tasks

- T1: Project management and dissemination of results
- T2: Linked data model for cultural content
- T3: Semantic migration with relations inference
- T4: Knowledge propagation
- T5: Document mining for automatic metadata records
- T6: Knowledge graph management platform KGMP
- T7: Exploration and querying interface
- T8: User evaluation

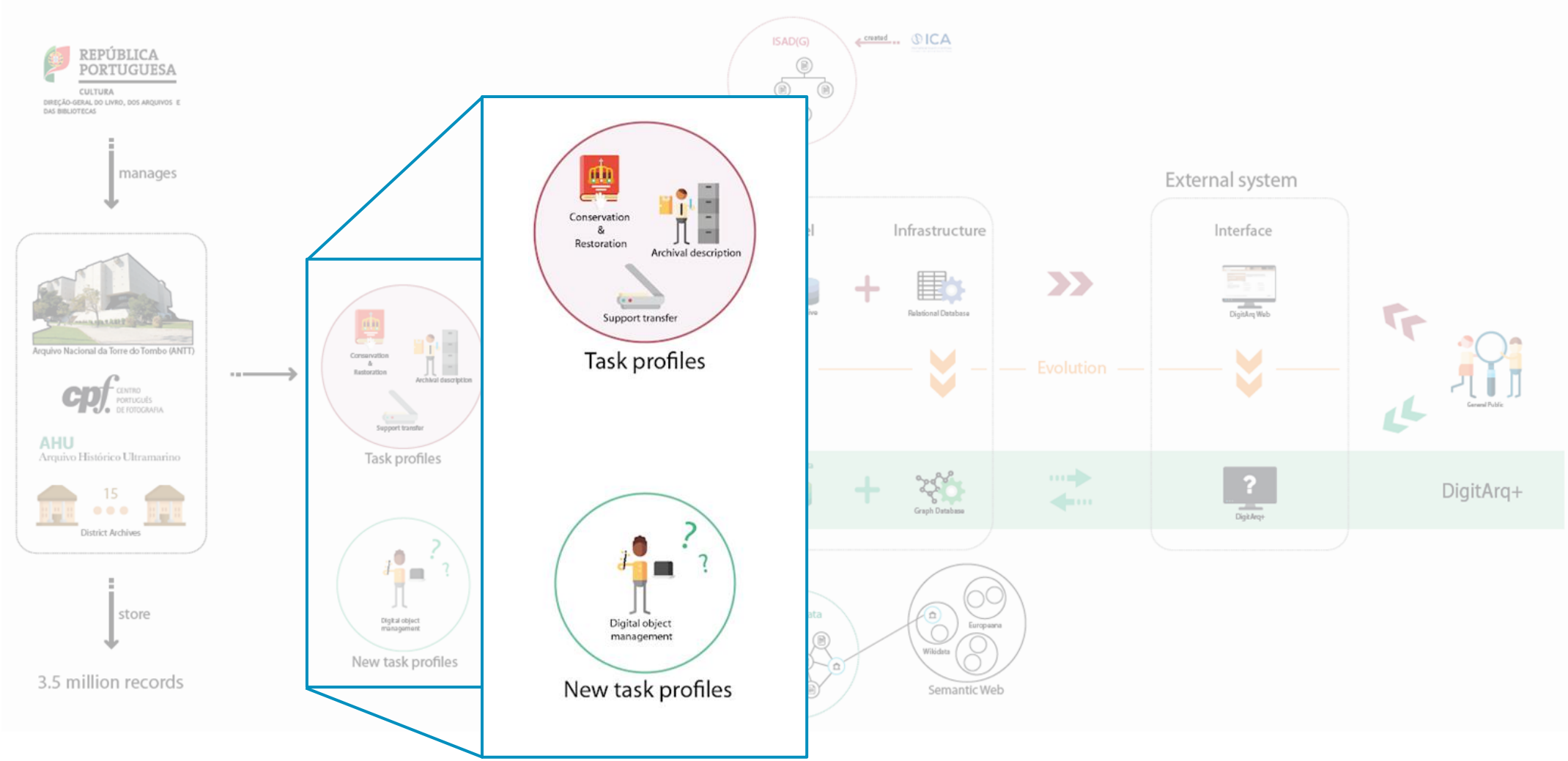
From Digitarq to Digitarq+



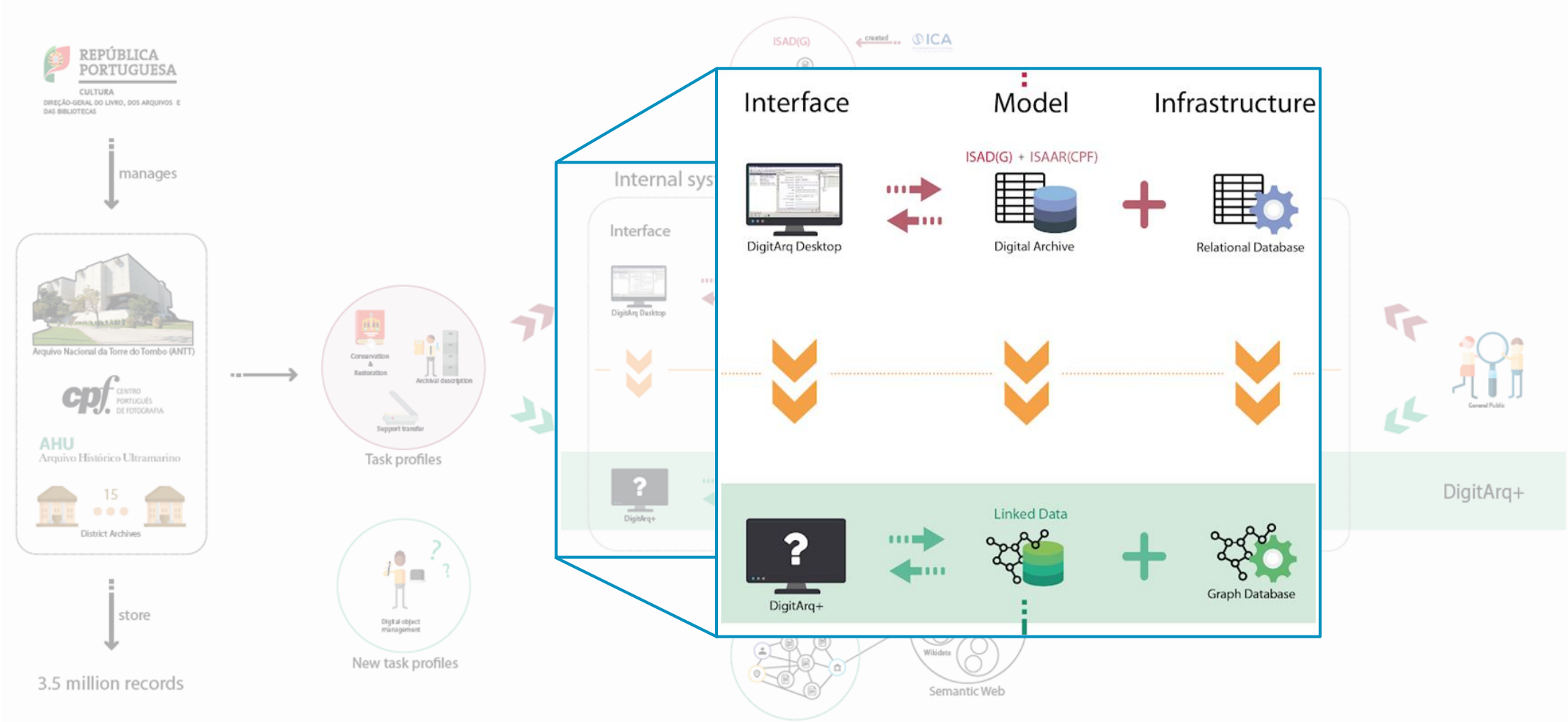
From Digitarq to Digitarq+



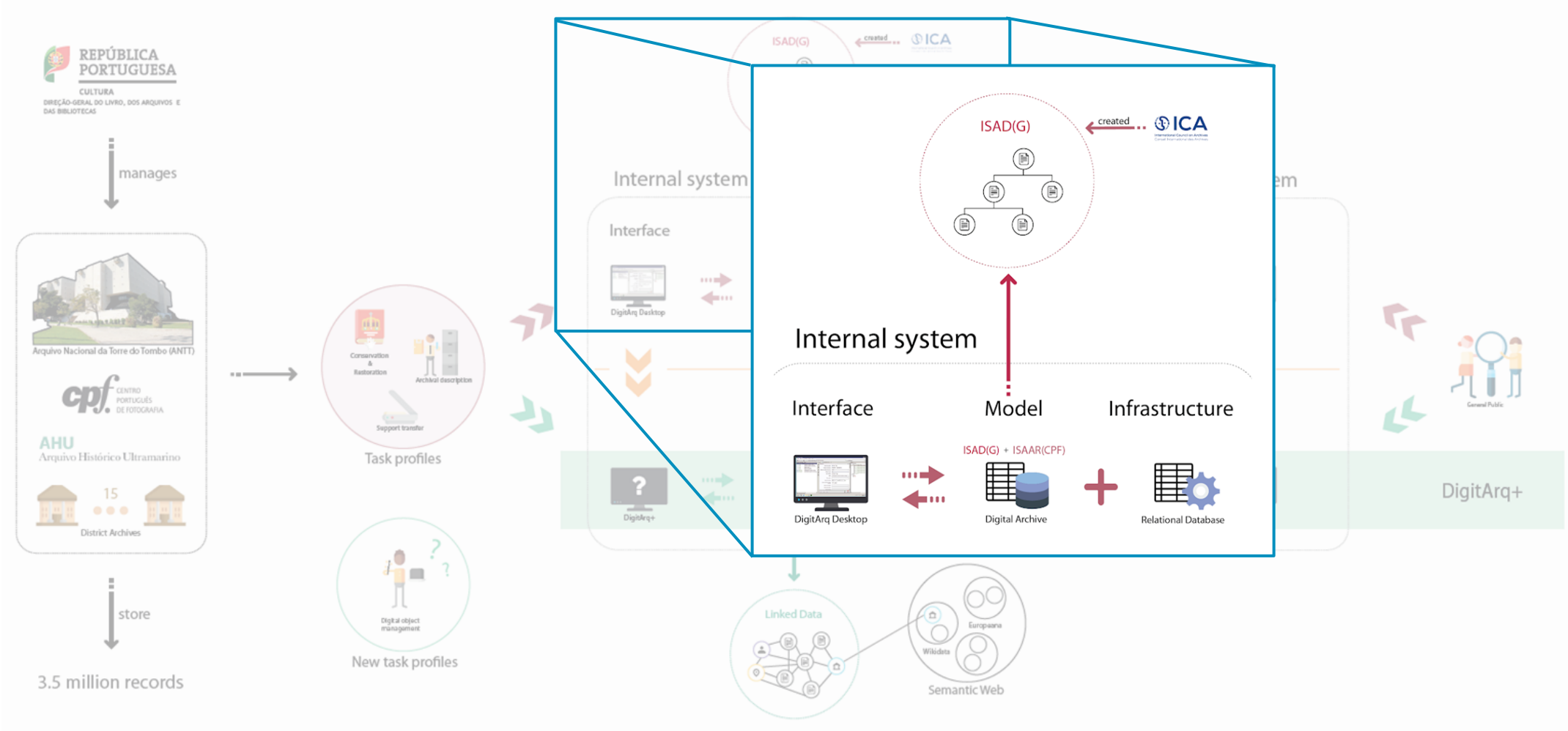
From Digitarq to Digitarq+



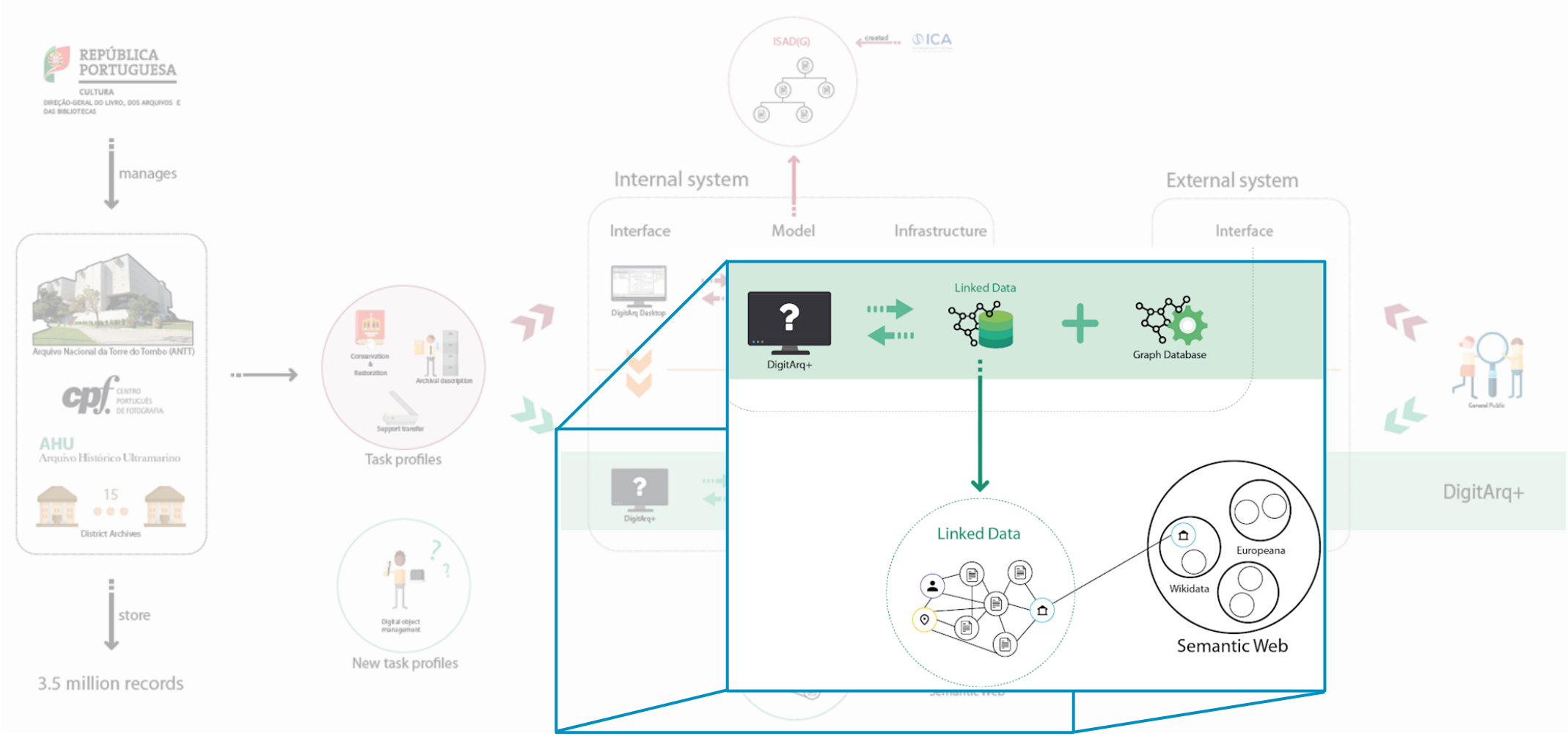
From Digitarq to Digitarq+



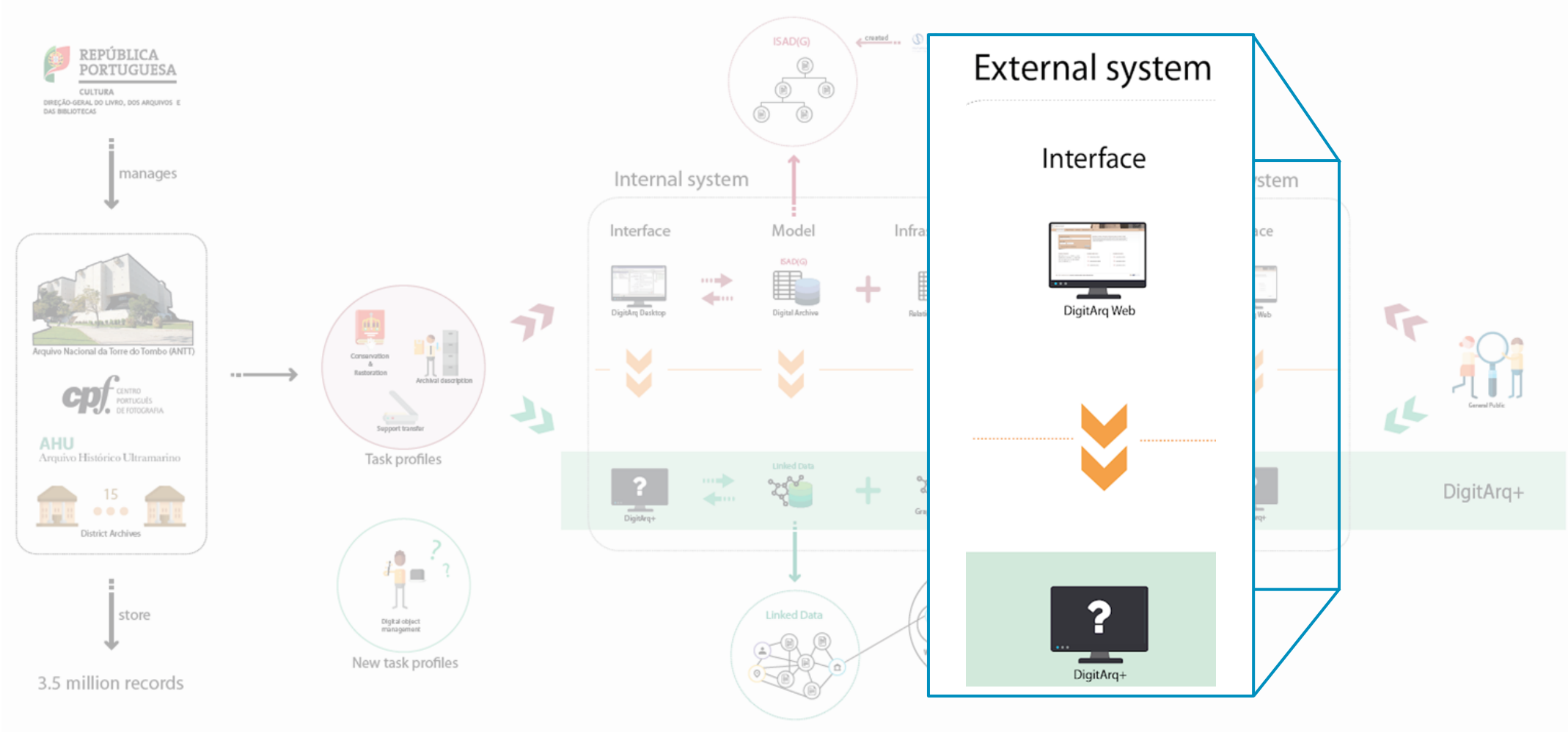
From Digitarq to Digitarq+



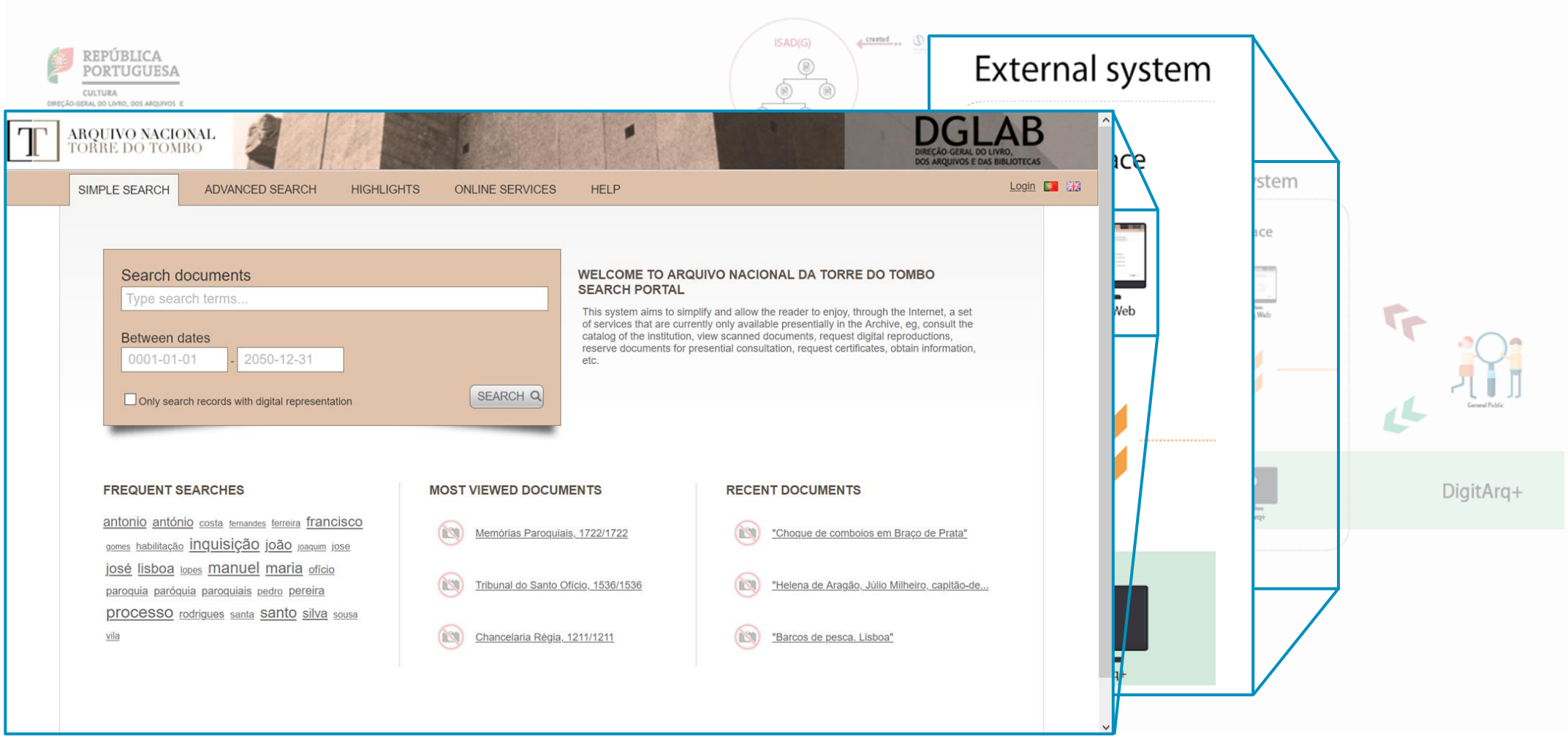
From Digitarq to Digitarq+



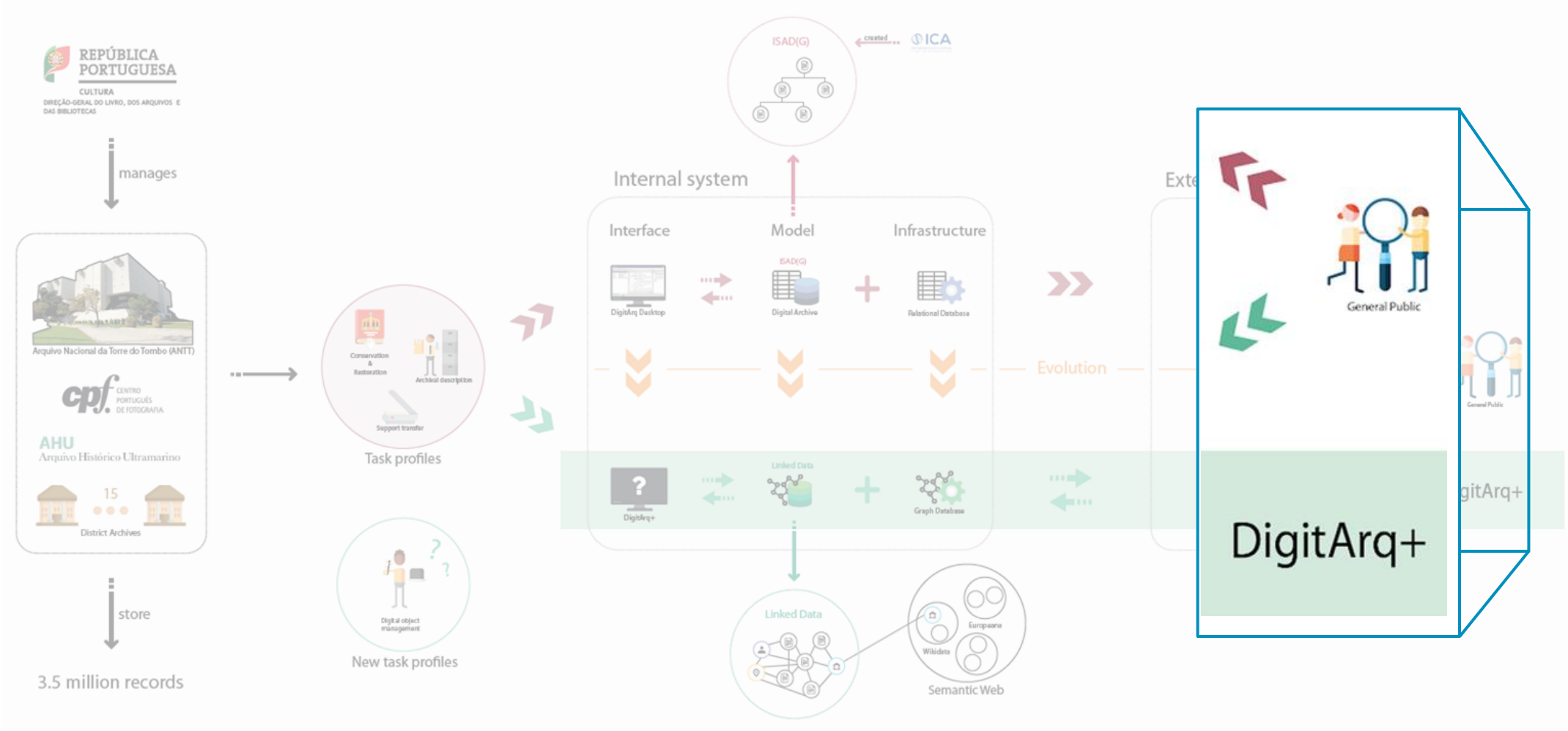
From Digitarq to Digitarq+



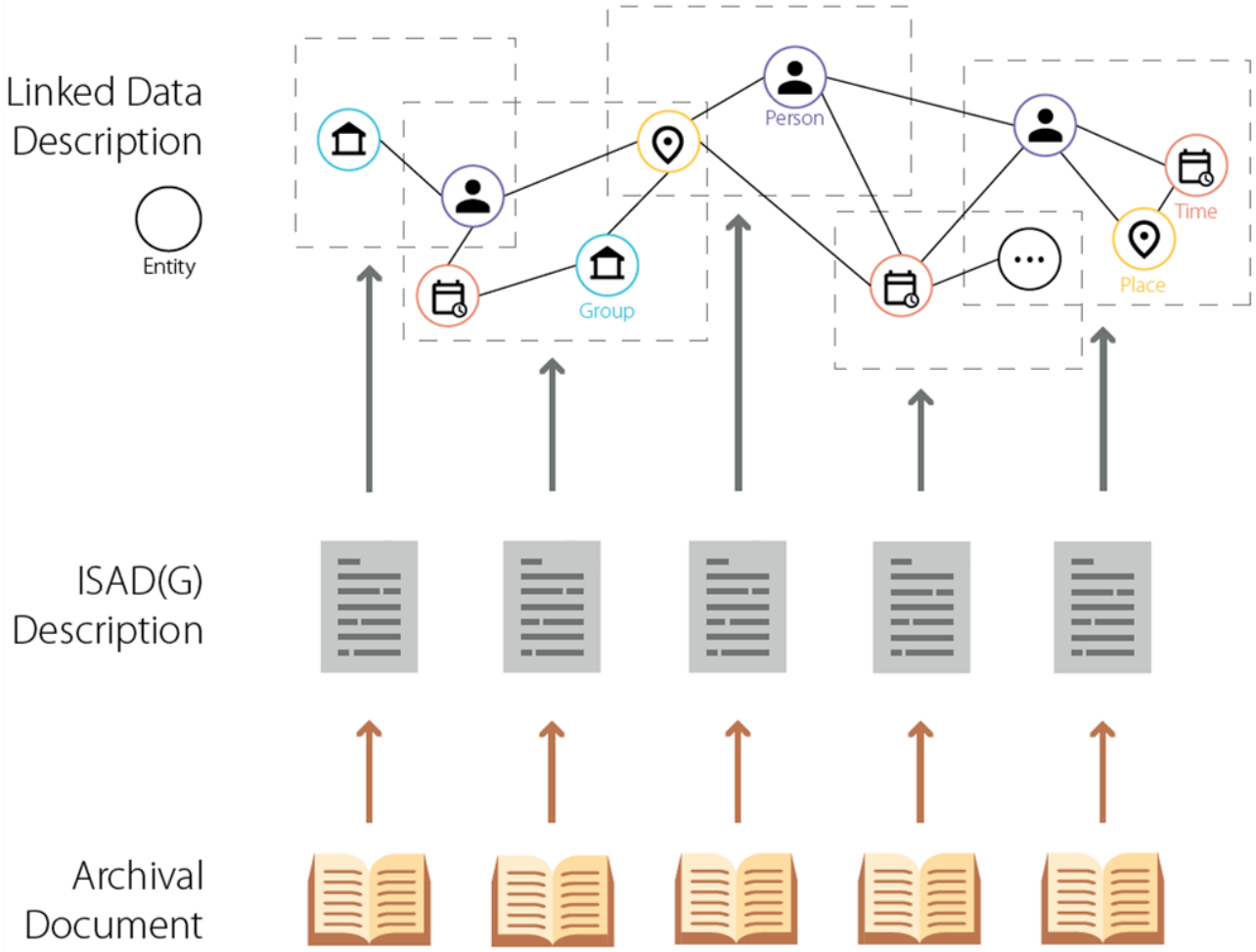
From Digitarq to Digitarq+



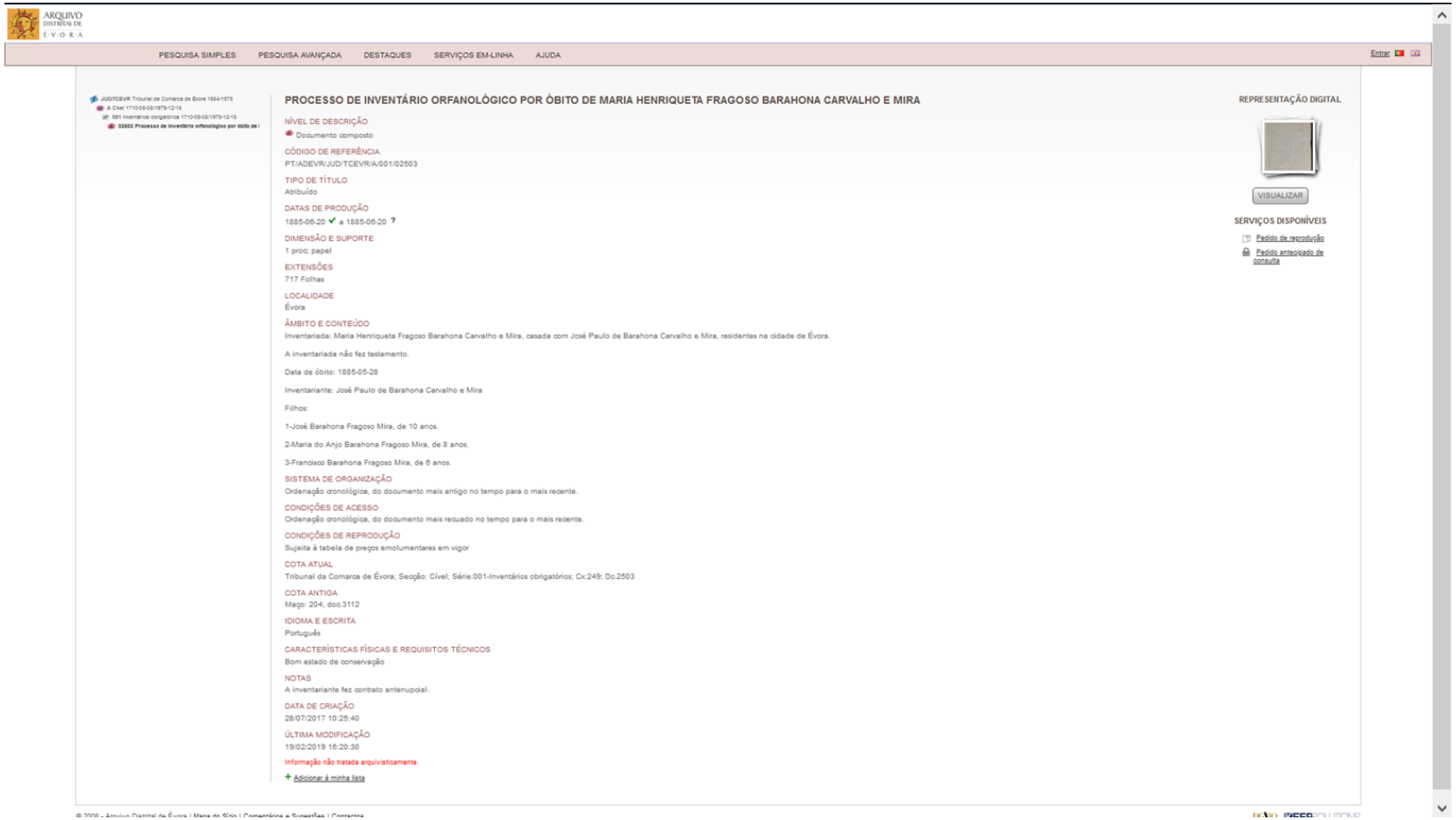
From Digitarq to Digitarq+



Proposed evolution of archival description

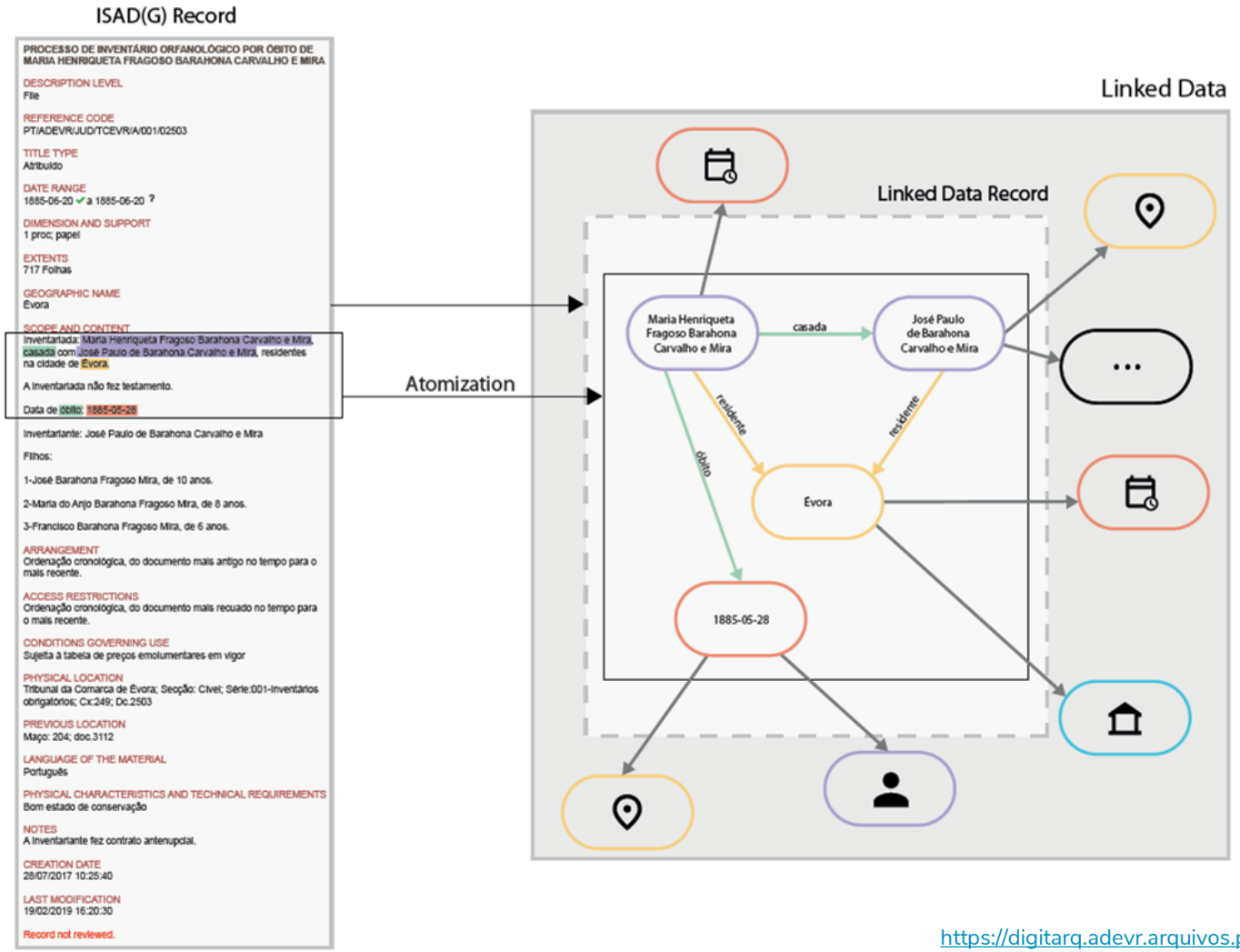


Atomization of an ISAD(G) record in linked data



https://digitarq.adevr.arquivos.pt/details?id=1174365

Atomization of an ISAD(G) record in linked data



https://digitarq.adevr.arquivos.pt/details?id=1174365

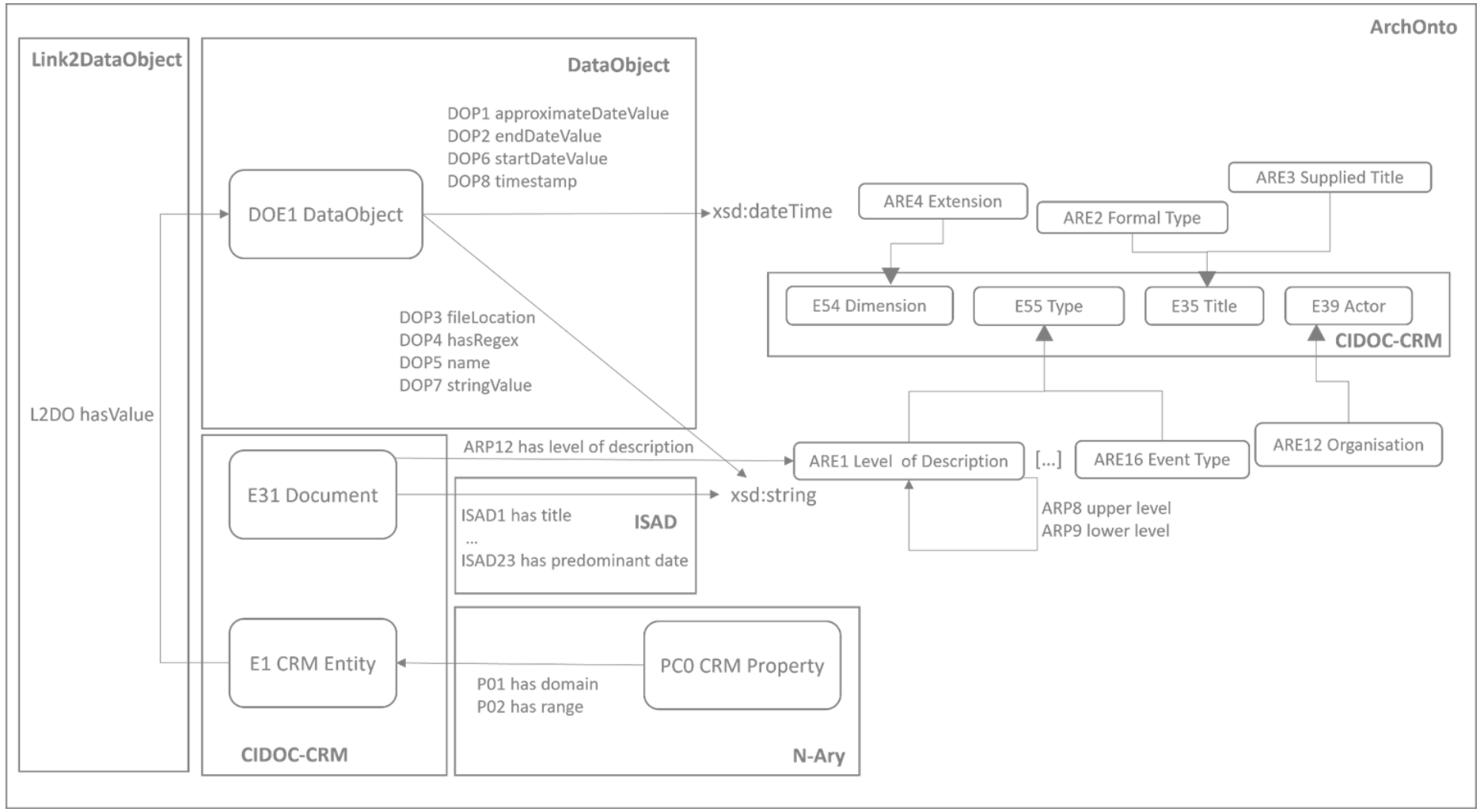
ArchOnto

An extension of CIDOC CRM for archives

ArchOnto

- Is composed by 5 different ontologies:
  - CIDOC CRM
  - DataObject
  - N-ary
  - ISAD Ontology
  - Link2DataObject
- Use the prefixes:
  - ARE – Archival Entity, ex.: ARE1 Level of Description.
  - ARP – Archival Property, ex.: APR12 has level of description.

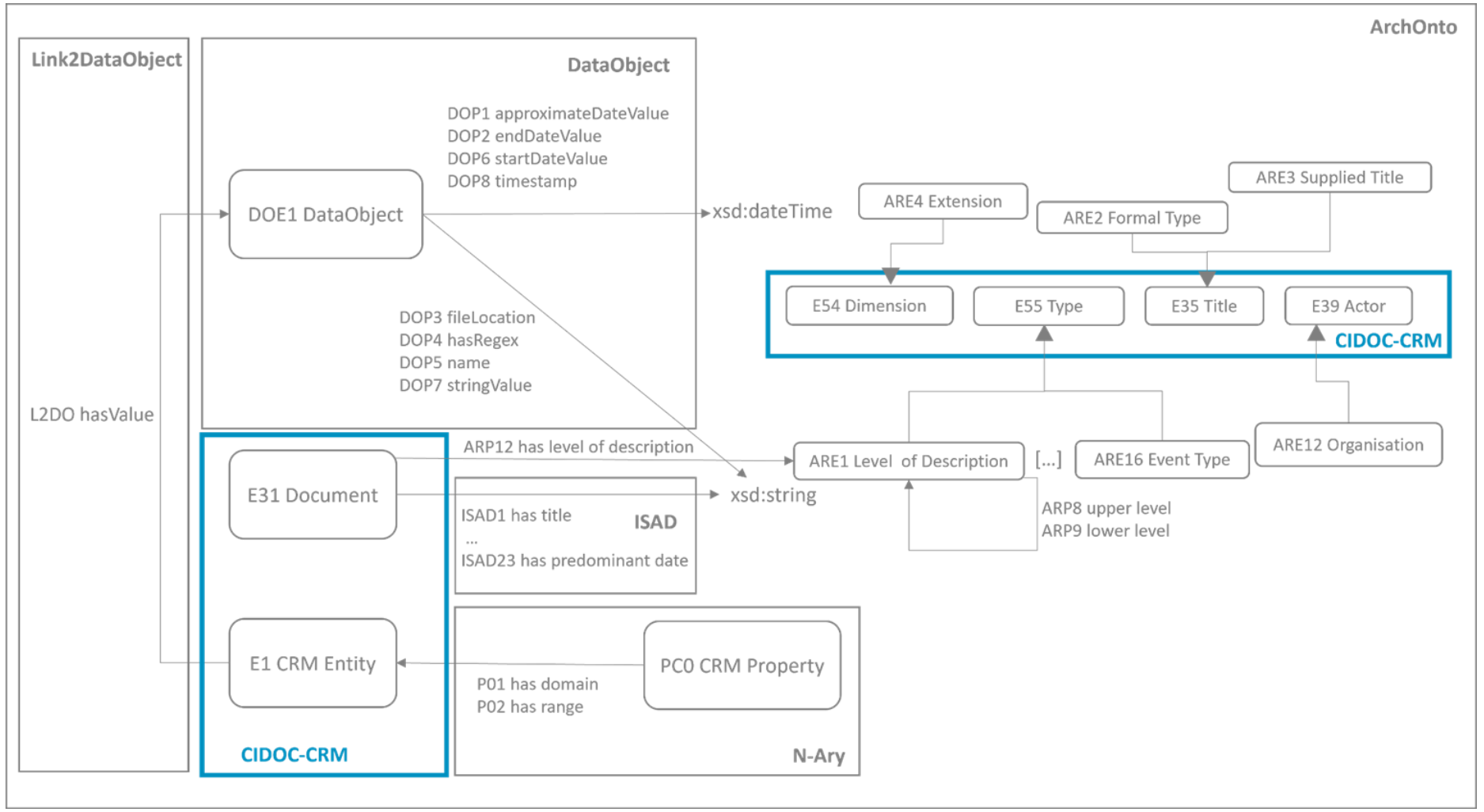
ArchOnto


ArchOnto
Link2DataObject
DataObject
DOP1 approximateDateValue
DOP2 endDateValue
DOP6 startDateValue
DOP8 timestamp
DOE1 DataObject
xsd:dateTime
ARE4 Extension
ARE3 Supplied Title
ARE2 Formal Type
E54 Dimension
E55 Type
E35 Title
E39 Actor
CIDOC-CRM
DOP3 fileLocation
DOP4 hasRegex
DOP5 name
DOP7 stringValue
L2DO hasValue
ARP12 has level of description
E31 Document
xsd:string
ARE1 Level of Description
[...]
ARE16 Event Type
ARE12 Organisation
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

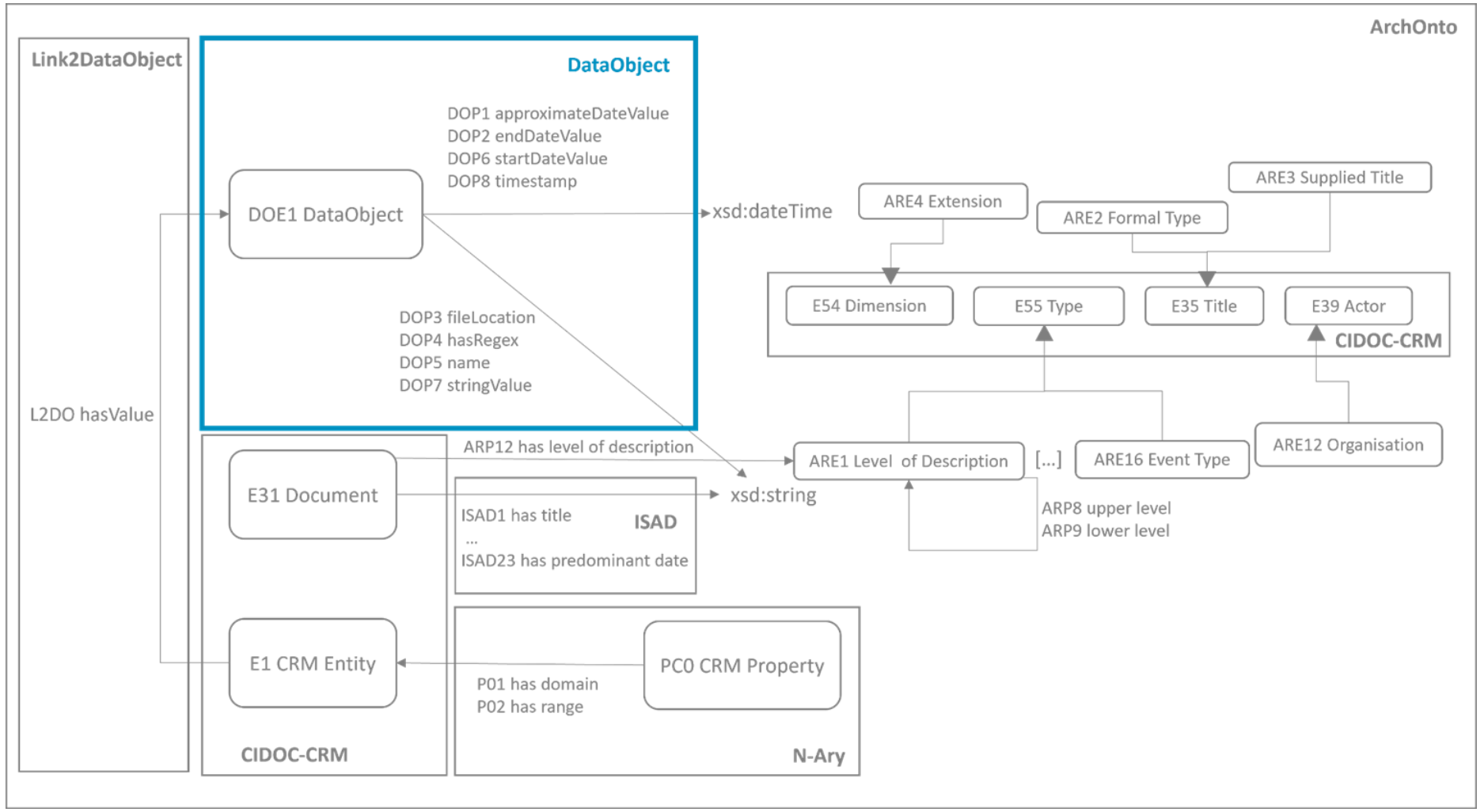
ArchOnto - Controlled vocabularies

Class	Vocabulary	Example Values
ARE1 Level of Description	Level of description	Fonds; Series; Section; File; Item
ARE2 Formal Title ARE3 Supplied Title	Title Type	Formal, Supplied
ARE5 Identifier Type	Identifier of collective person/group	PT; VCT; AGH01; 161016; ADLSB; 600084892; PT-LiBN
ARE6 Date Type ARE9 Date Certainty	Type of time period	Exact dates; Inferred dates; Predominant dates
ARE7 Name Type	Type of name of collective person/group	Authorized form of name; Another form of the name; Parallel name form
ARE8 Role Type	Role played	Producer; Material Author; Recipient
ARE11 Documentary Typology	Documentary Typology	Certificate; Income book; Patent
ARE13 Subject Type	Subject	Education; Science; Law; Management
ARE14 Place Type	Type of jurisdictional entity	Ocean; Archipelago; Mountain range; Country; District
ARE15 Acquisition Type	Transfer of Custody / Acquisition Identifier	Purchase; Giving; Donation; Deposit; Swap; Legacy; Reintegration; Transfer
ARE16 Event Type	Event Type	Evaluation; Expertise; Financial management
E56 Language	Language Identifier	Portuguese; Latin; French; Greek
E57 Material	Support	Paper; Parchment; Photosensitive film
E58 Measurement Unit	Measurement Unit	Centimeter; Gram; Byte; Minute; Pack
E98 Currency	Currency	Euro; Dollar; Kwanza

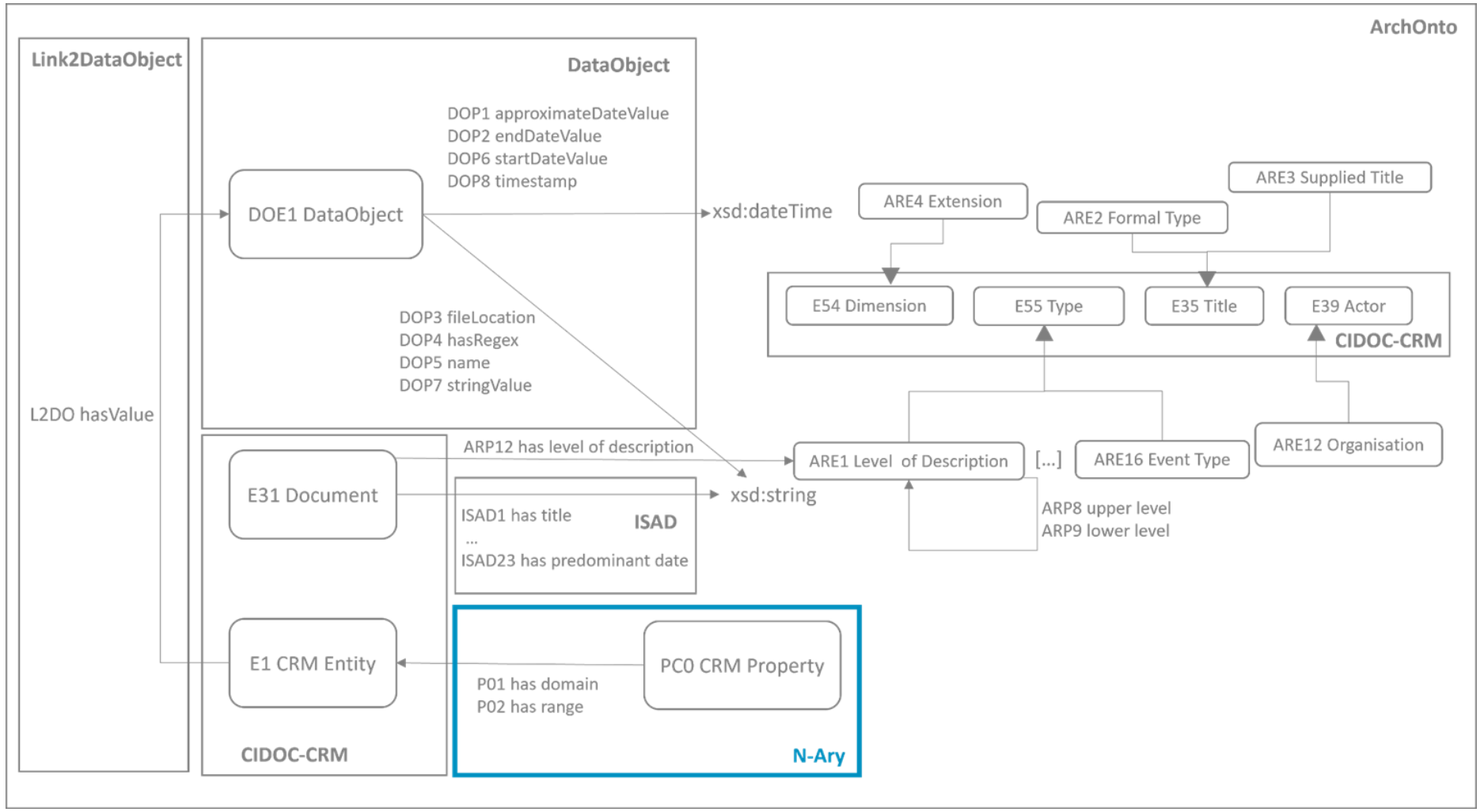
ArchOnto - CIDOC CRM



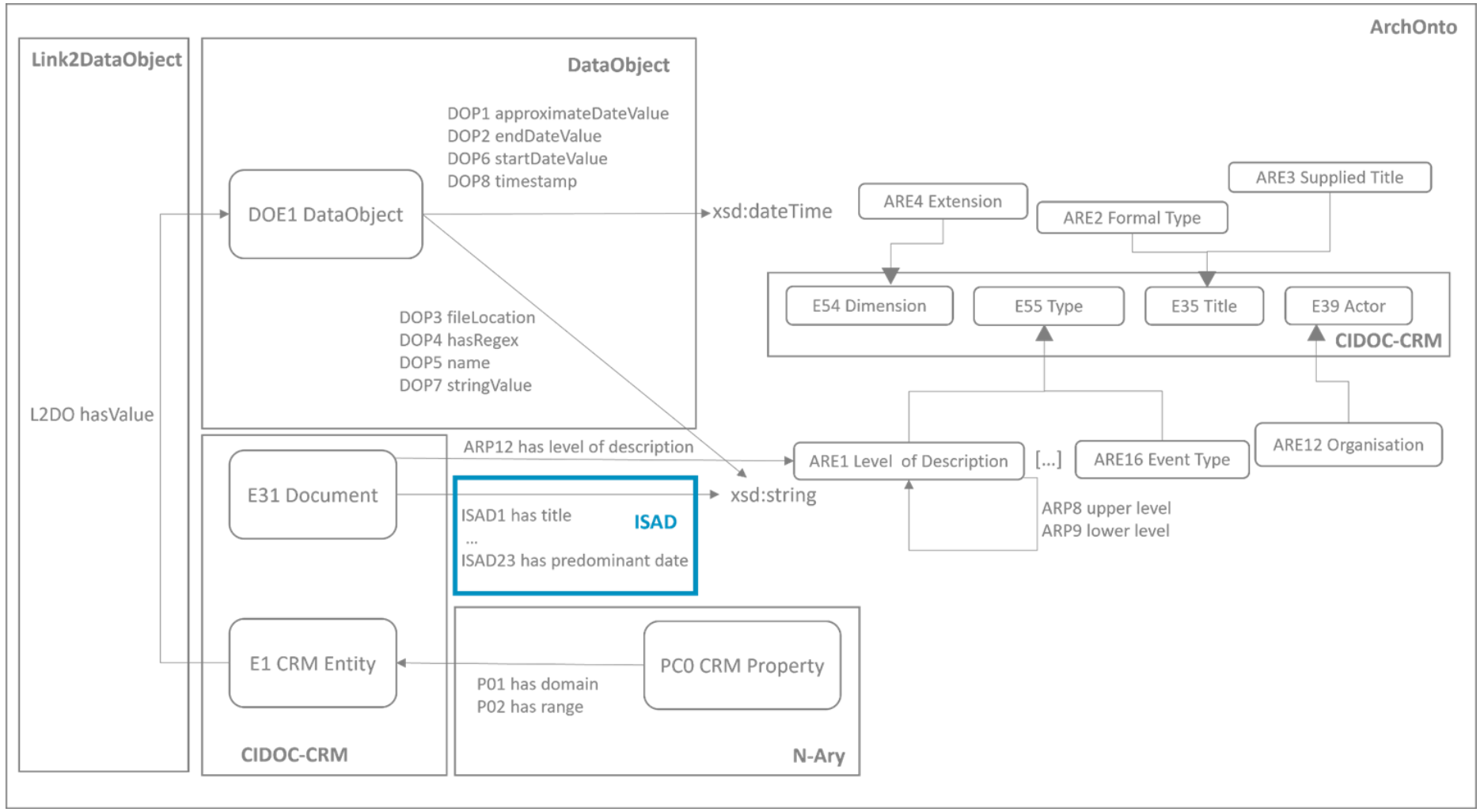
ArchOnto - DataObject



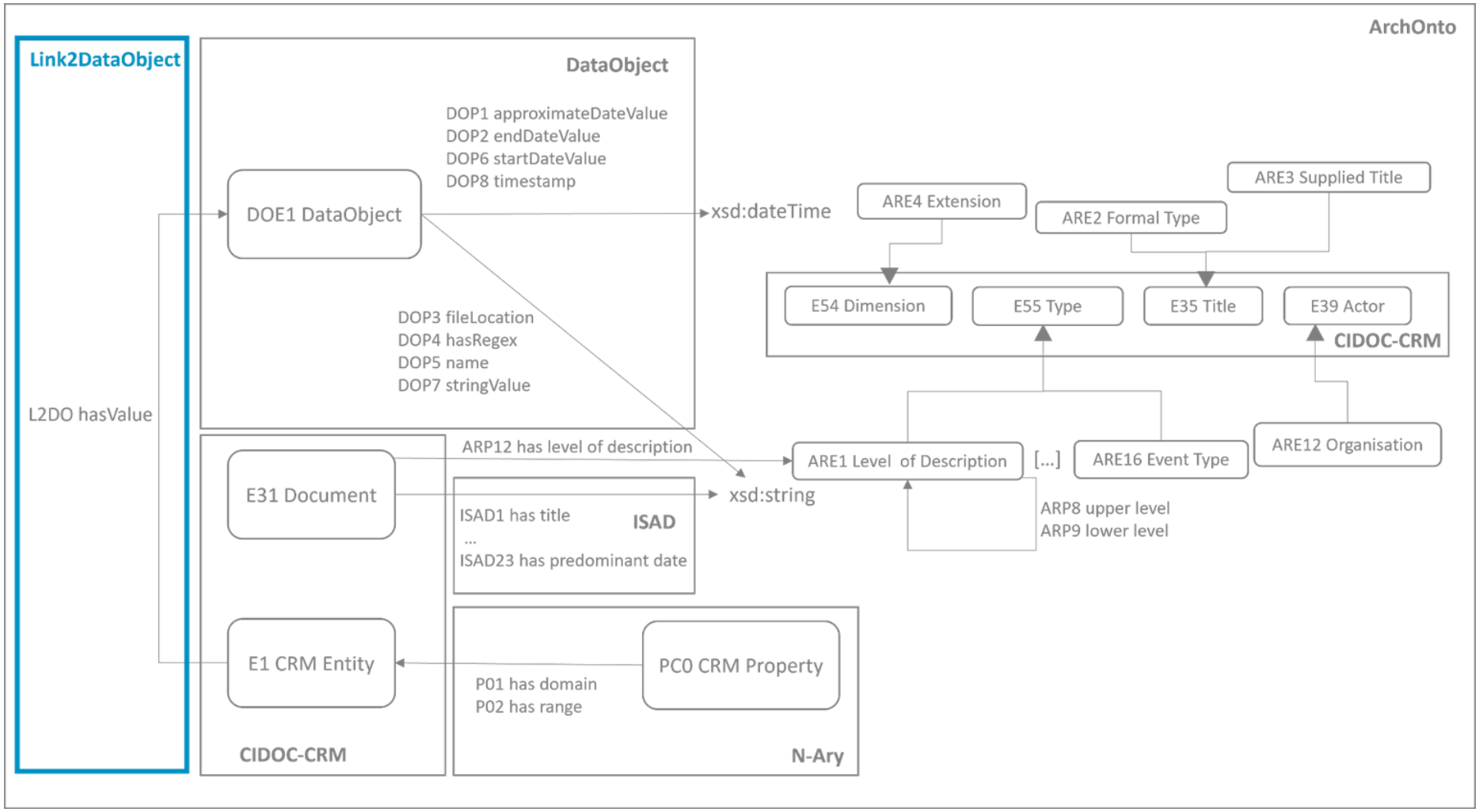
ArchOnto - N-ary



ArchOnto - ISAD Ontology



ArchOnto - Link2DataObject



EPISA Platform

An archival management system based on linked data technologies

EPISA Platform

The EPISA Platform is the software infrastructure developed to support archival records management based on linked data technologies.

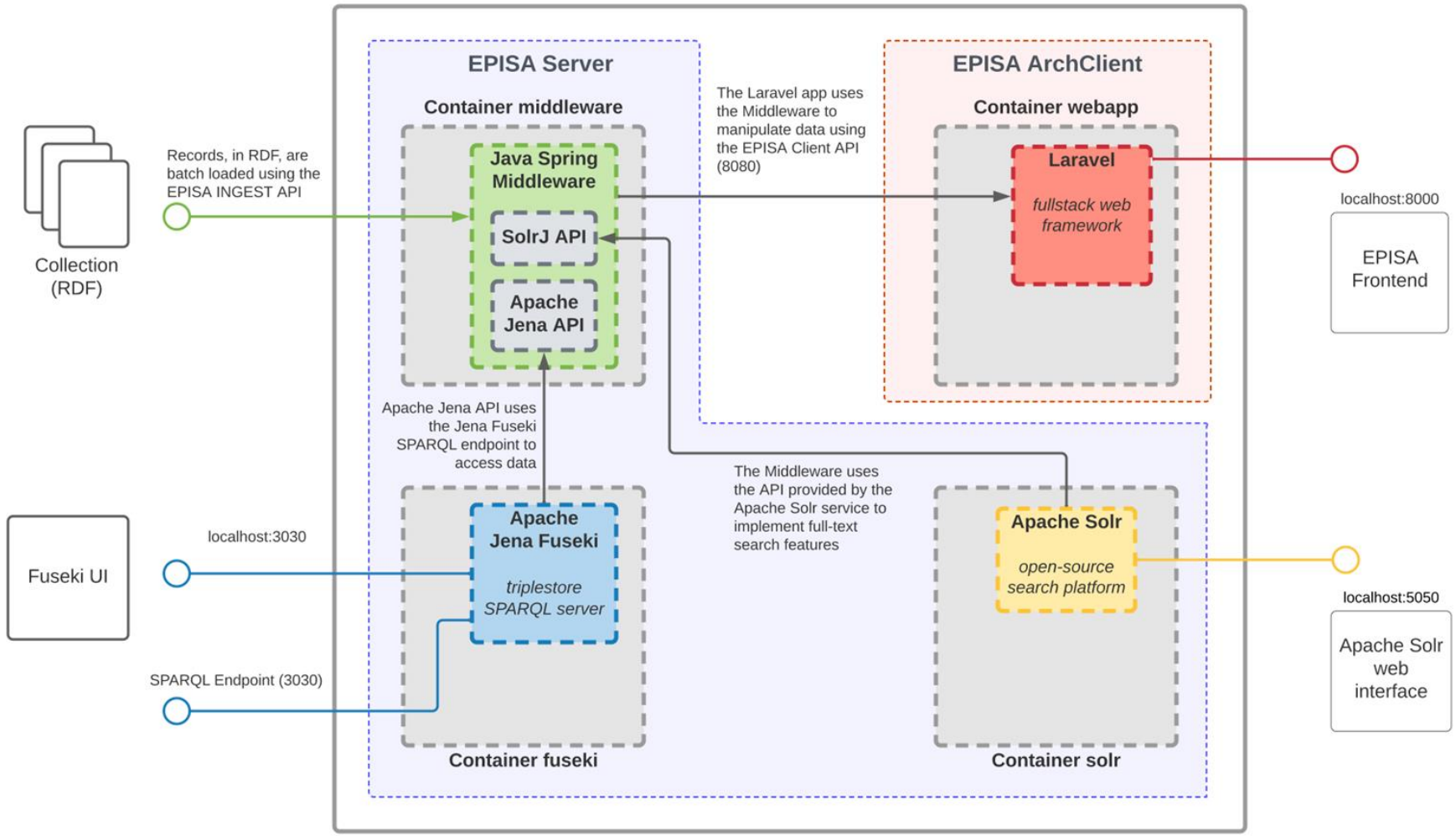
Supports storage, creation, search, access, and navigation over archival records and related entities.

Is based on open-source native linked data technologies.

Organized in two main components:

- **EPISA Server** is responsible for storing, reasoning, managing access, and providing an effective search mechanism over the archival data.
- **EPISA ArchClient** is a web application providing a graphical user interface for archivists to access, manage and describe collections of archival records.

EPISA Platform Docker Environment



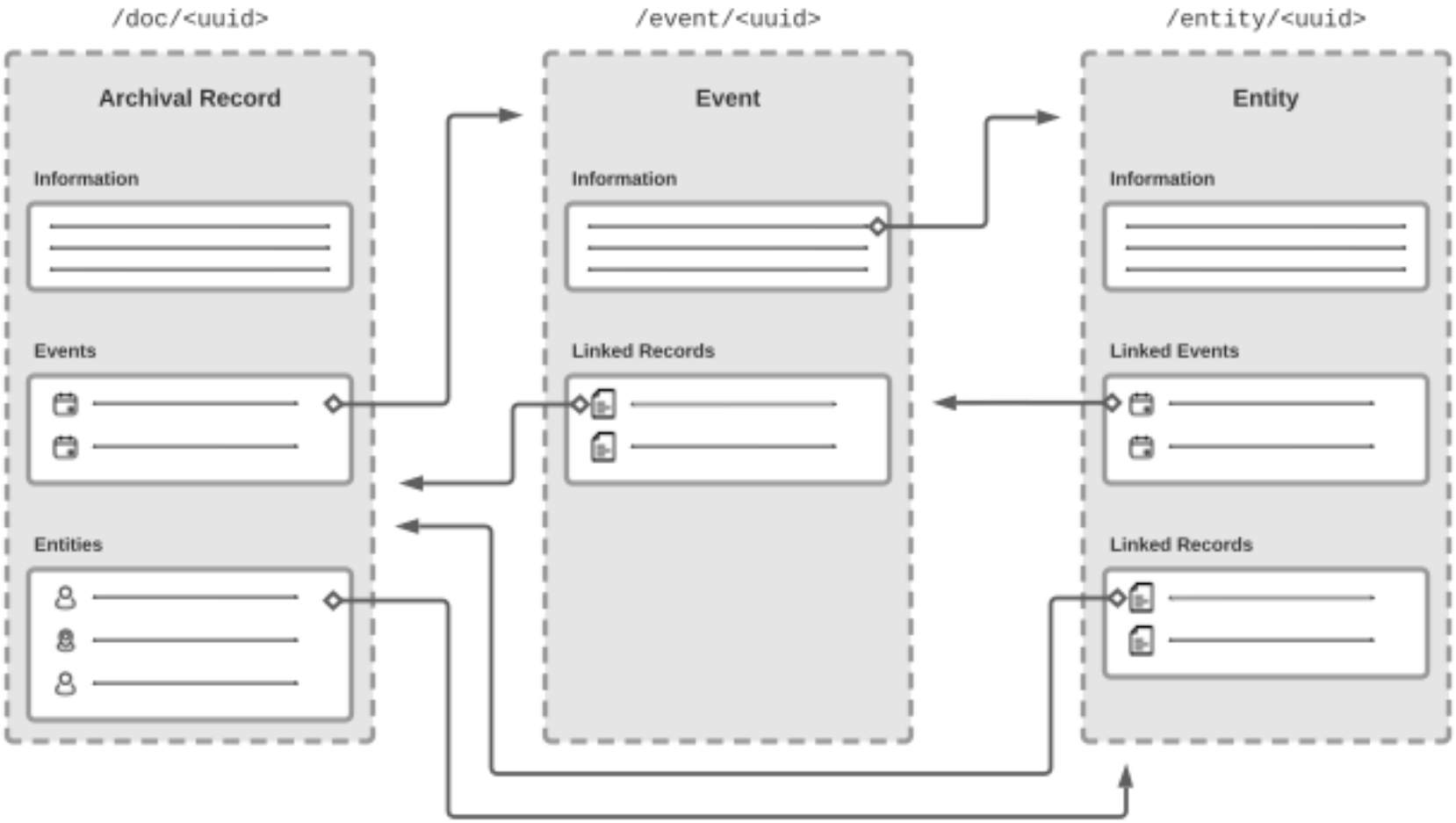
User Interface Abstractions

Higher-level abstractions need to be defined at the user interface level.

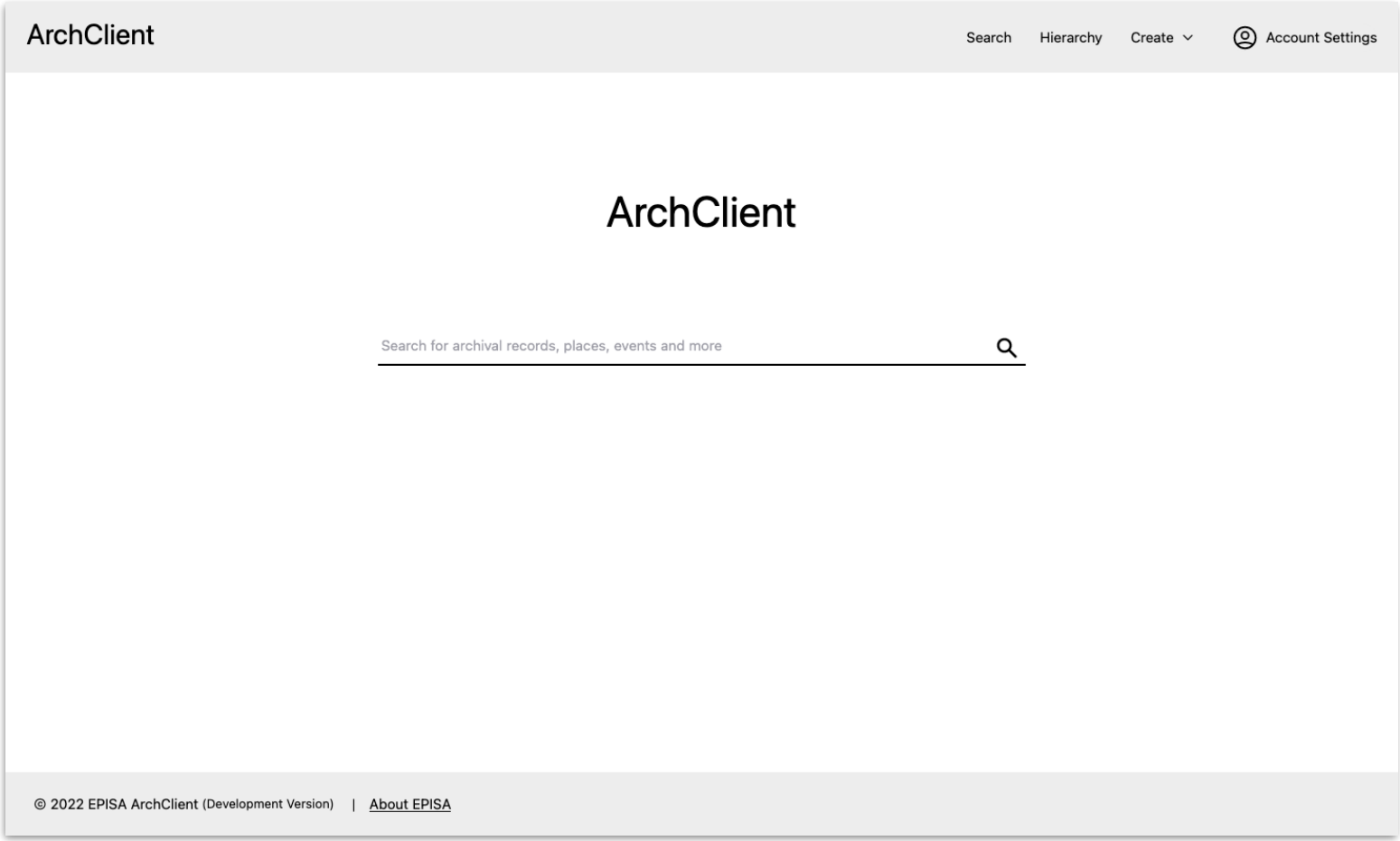
The abstractions defined are:

- **Records**, representing the collection of archival documents (e.g., fonds, collections, records).
- **Entities**, representing concepts (i.e., persons, places, organizations) mentioned in the records.
- **Events**, allowing the creation of complex structures linking records and entities — e.g., birth and death events, marriages, places of domicile.

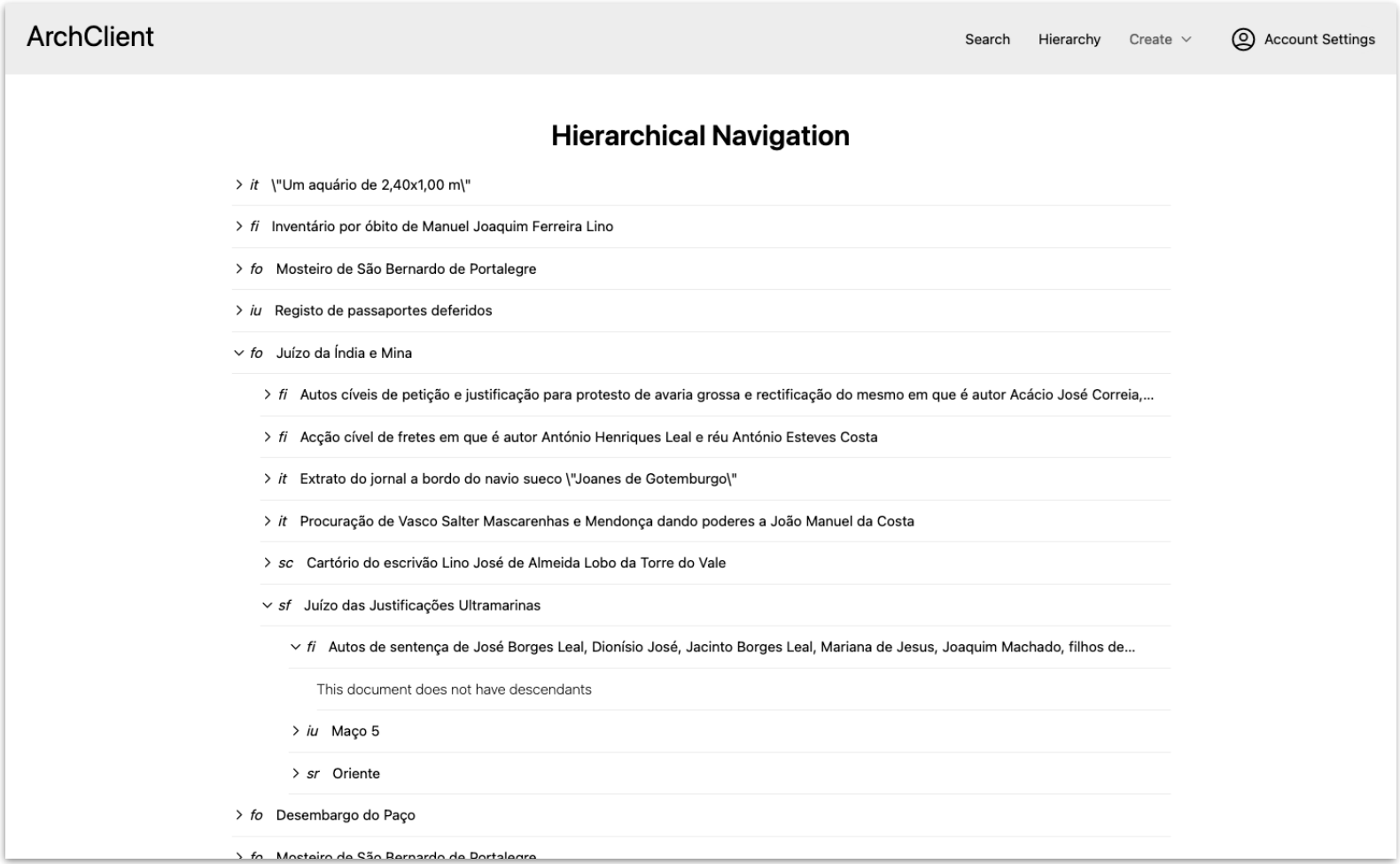
Central Information Concepts at the UI Level



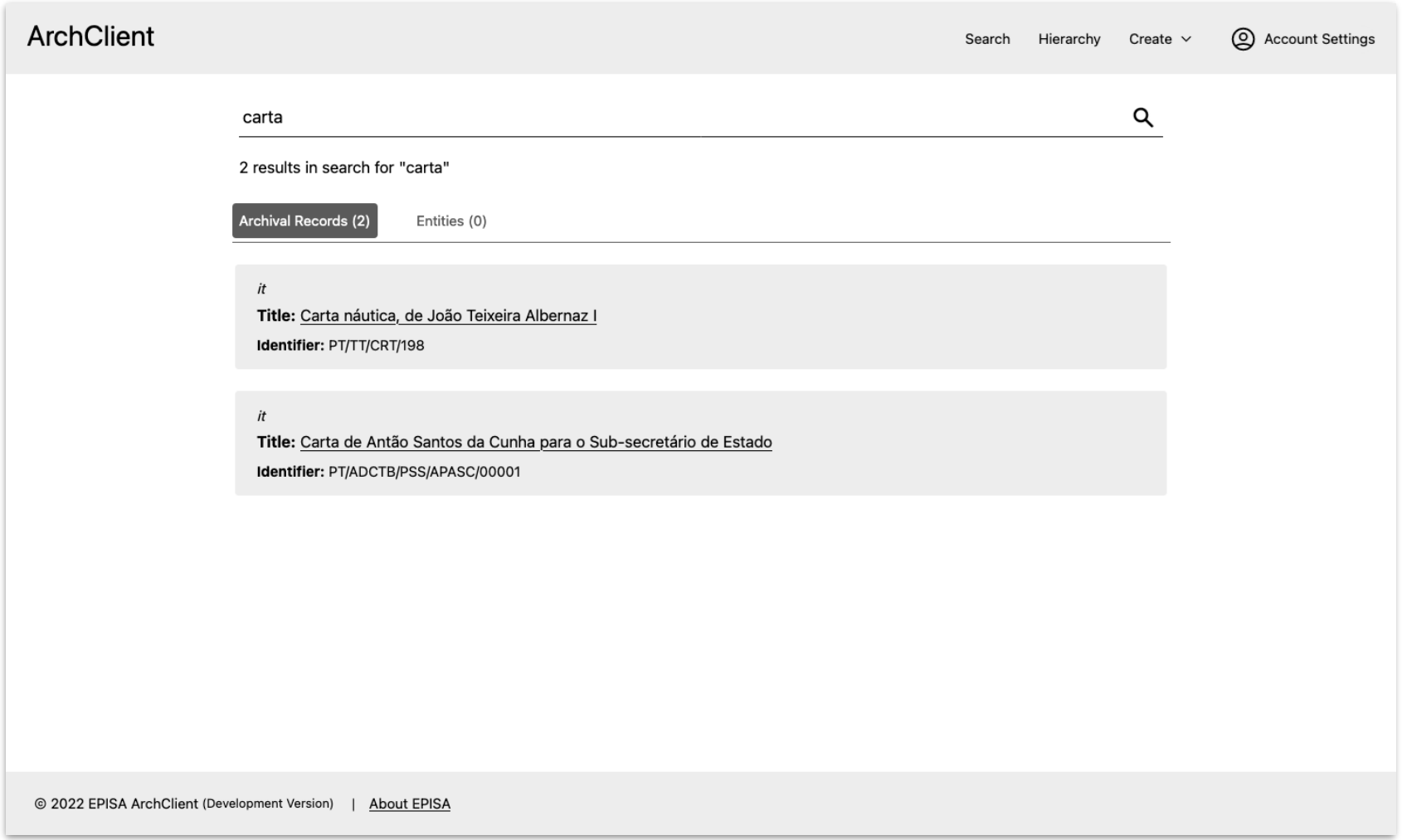
EPISA ArchClient Homepage



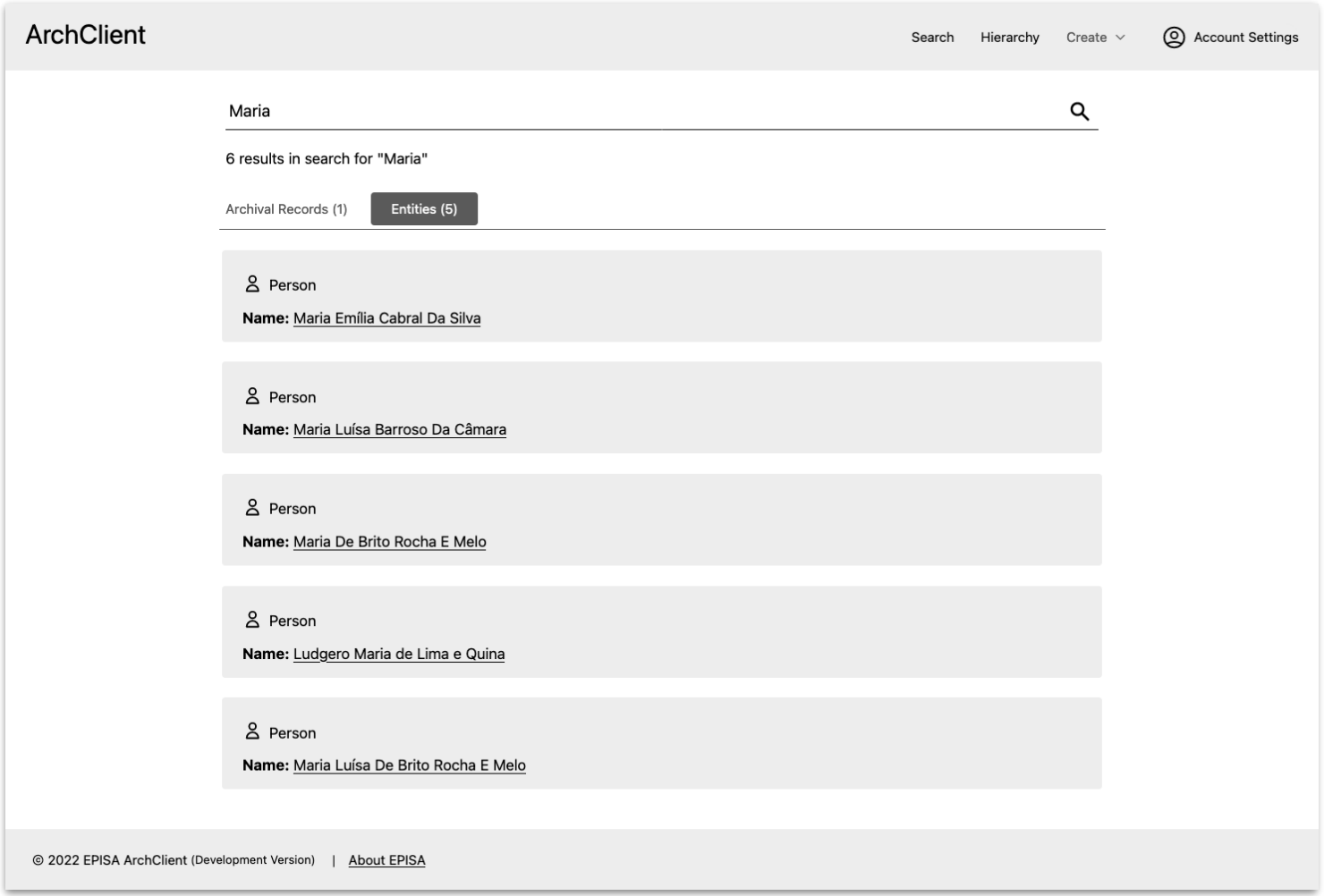
Hierarchical Navigation



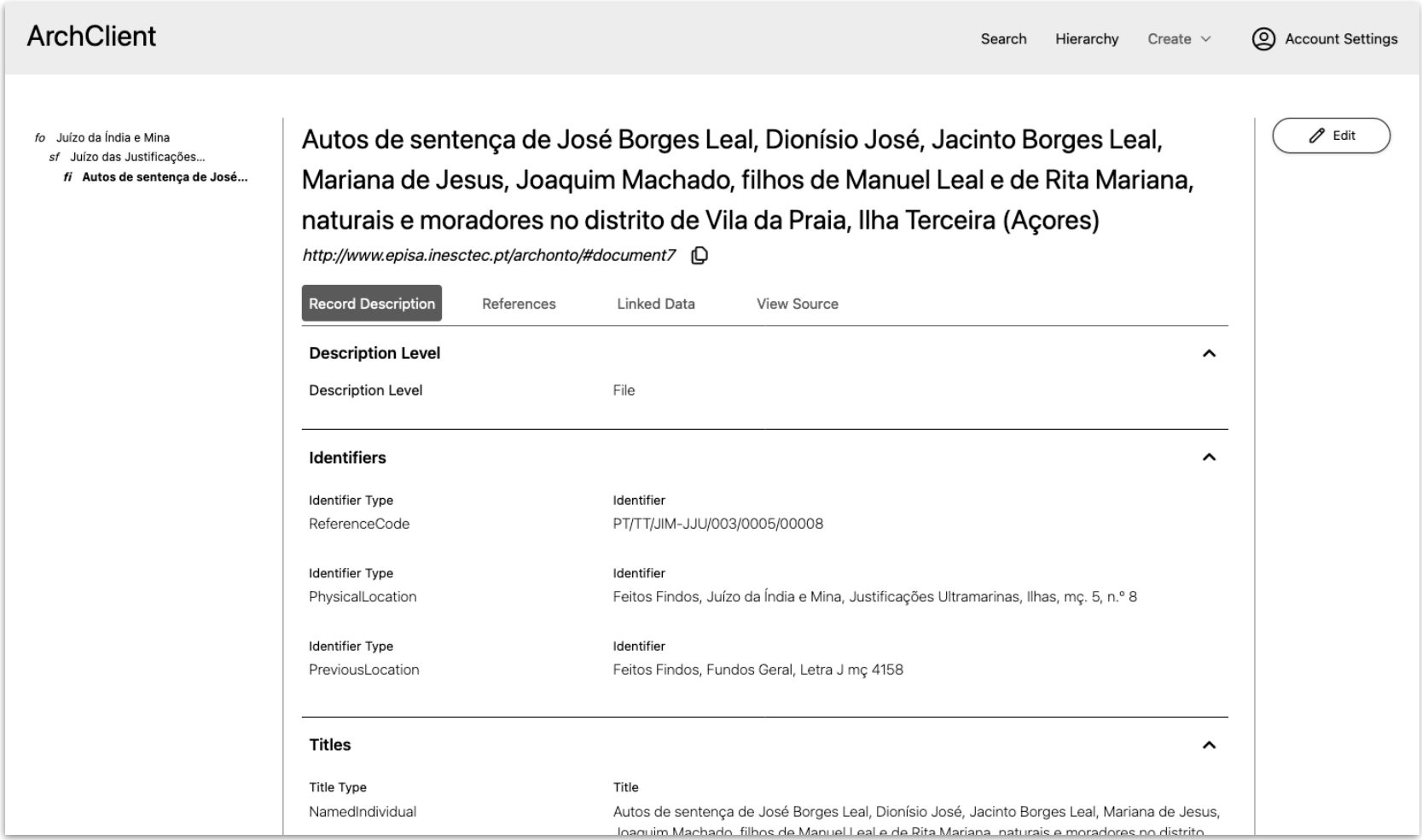
Search by Keywords (archival records results)



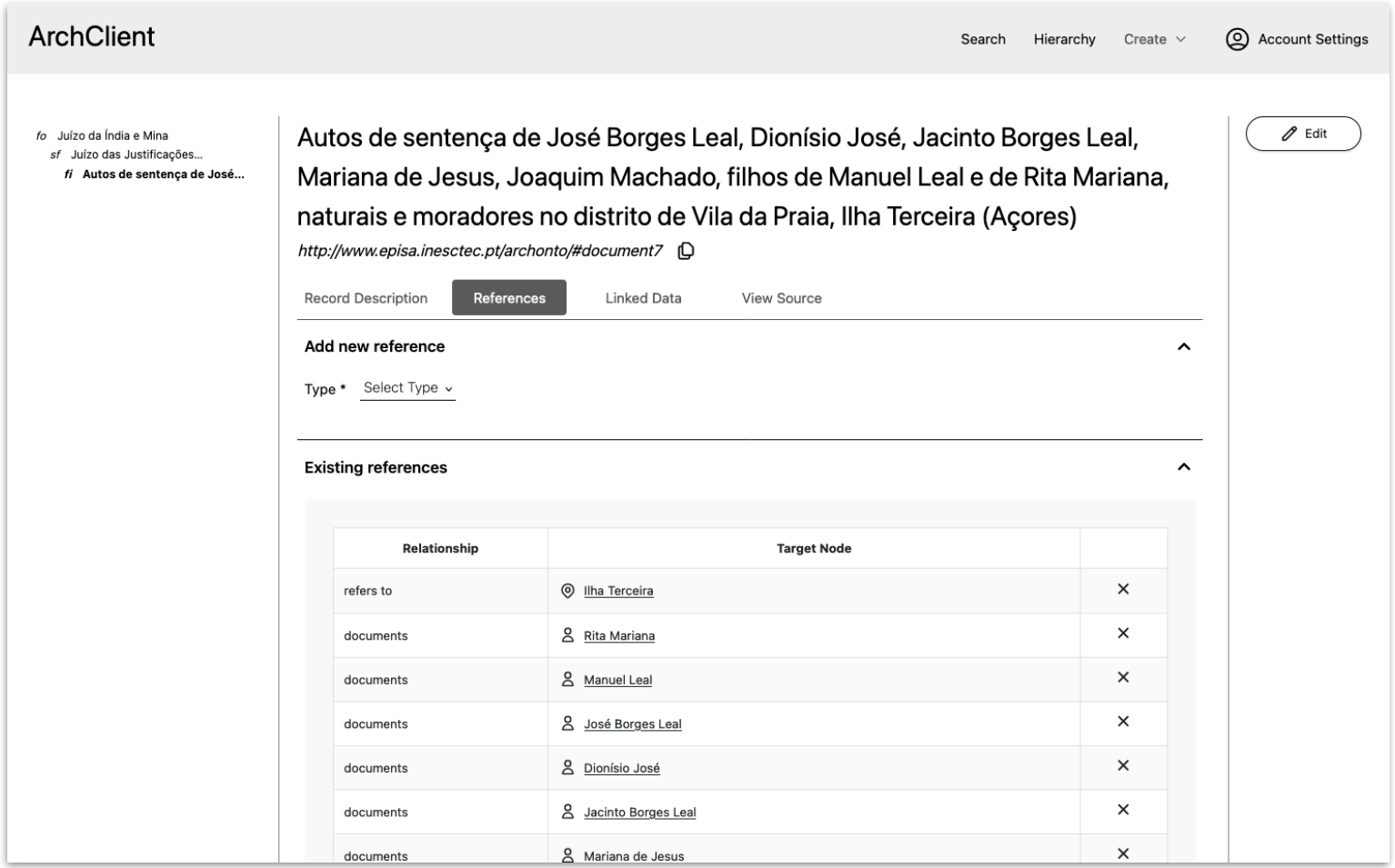
Search by Keywords (entities results)



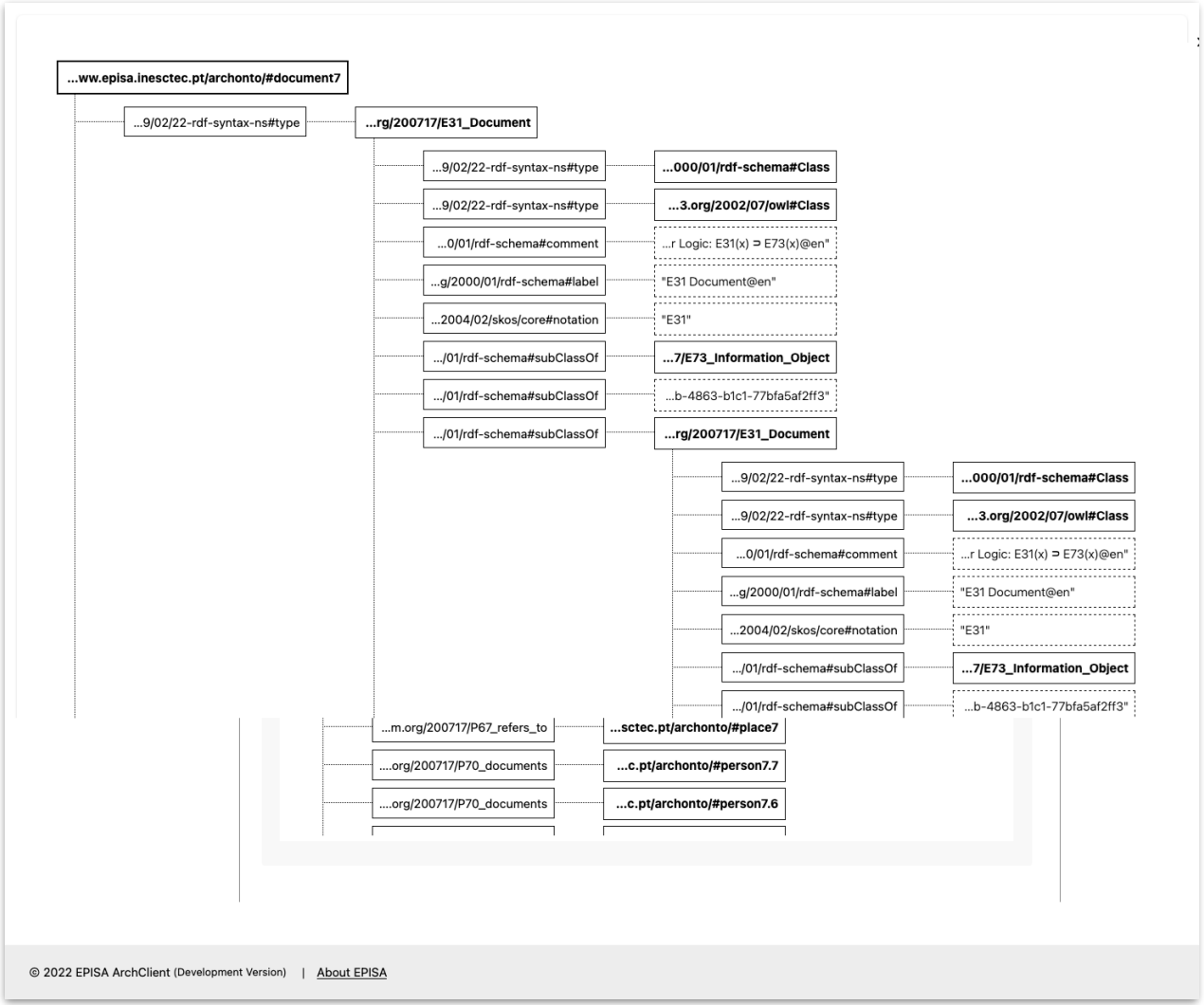
Single Record Description



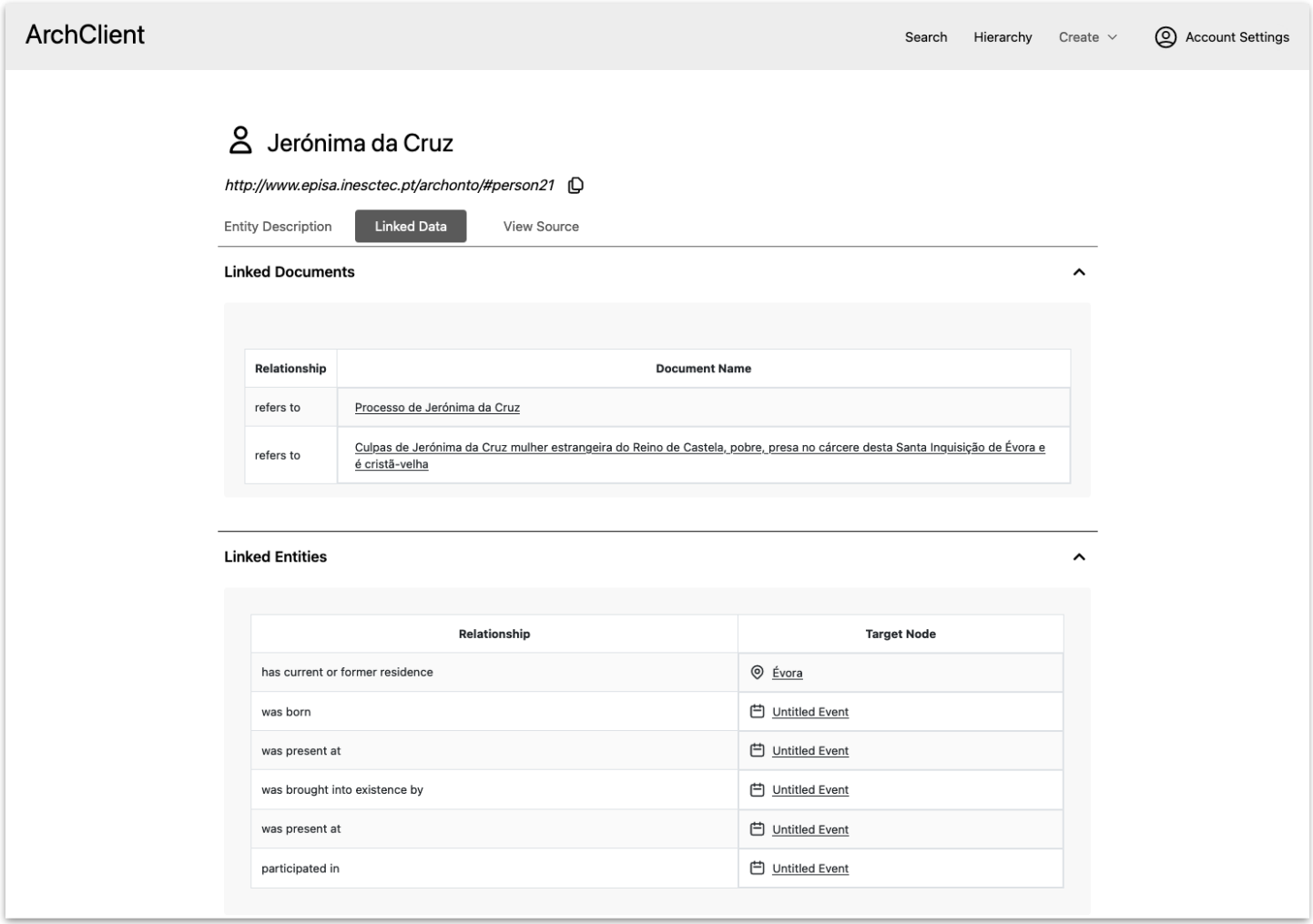
Single Record References



View Record ArchOnto Source



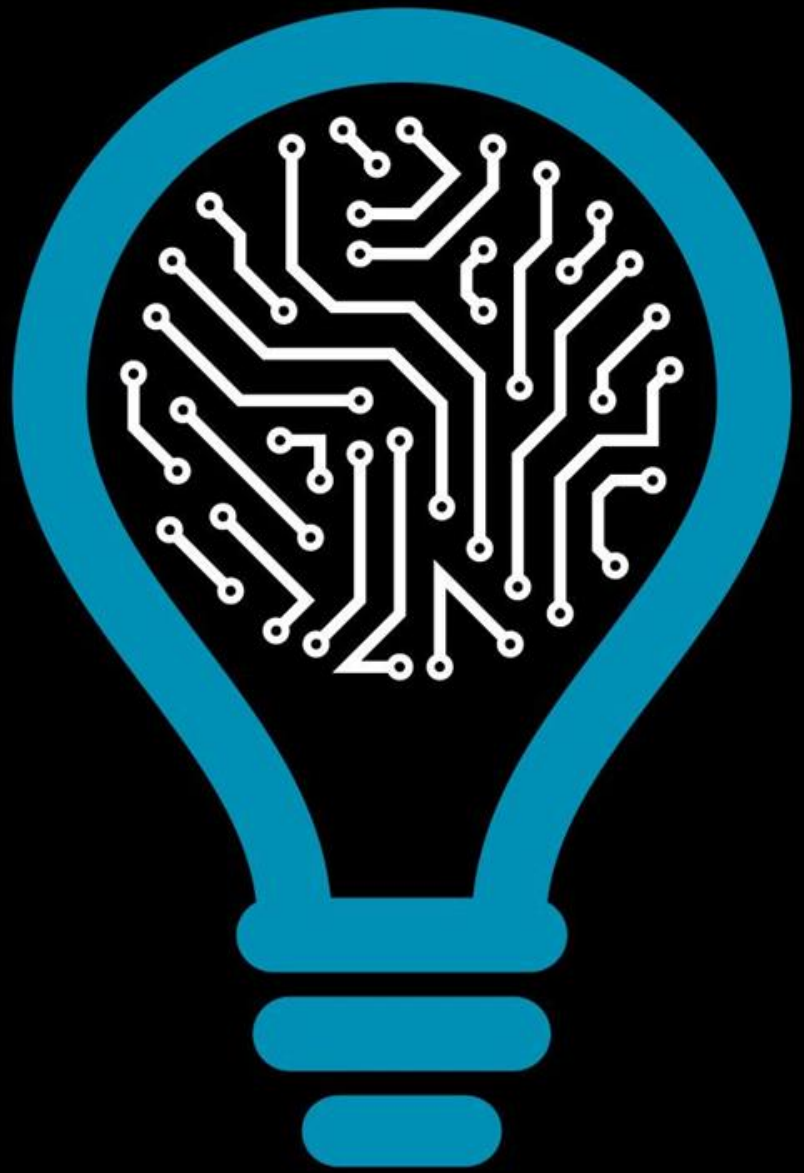
View Entity of Person Type



Development Challenges

- Limited maturity of linked data technologies, i.e. no standard off-the-shelf solutions.
- Lack of tools and resources focused on linked data authoring platforms.
- Lack of examples or guidelines for user interaction in authoring linked data.
- Complex underlying linked data model.

ENTITY AND PROPERTY INFERENCE FOR SEMANTIC ARCHIVES



55th CIDOC CRM and 48th FRBR CRM

Inês Koch
Carla Teixeira Lopes
Cristina Ribeiro
Sérgio Nunes

Follow us:

http://episa.inesctec.pt

https://purl.org/episa/archonto

INESCTEC

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO