infeurope		EUROPEAN COMMISSION DIRECTORATE-GENERAL INFORMATICS Directorate B - Digital Business Solutions (DBS) **DIGIT B2 - Corporate Knowledge and Decision making Solutions (CKD)**

Framework:	EPRS/ARCH/SER/16/013 - Lot 1 ‘Analysis and Studies’: ISA² Programme: Interoperability Solutions for Public Administrations. **2017.01 ISA² Action: Standard-based archival data management, exchange and publication** **Phase 2: Work Package 4 – Link Open Data**
Version:	Final version 1.0
Date:	11/11/2019

Assignment

Date:	May 2019 – November 2019
Reference:	EPRS/ARCH/SER/16/013 - Lot 1 ‘Analysis and Studies’
Body:	EUROPEAN COMMISSION - DIRECTORATE-GENERAL INFORMATICS Directorate B - Digital Business Solutions DIGIT B2 – Solutions for Legislation, Policy & HR
Title:	ISA² Programme - Interoperability Solutions for Public Administrations ISA² Actions – Standard-based archival data management, exchange and publication Phase 2 – Proof of Concept Work Package 4: Link Open Data
Follow-up by:	Head of Sector DIGIT.B.2.004: Mr Béla Harsányi Project manager DIGIT.B.2.004: Mr Sorin Bobeică Action owner: OIB.OS.1.002 (HAS) Mr Sven Carnel Archivist specialists: OIB.OS.1.002 (HAS) Ms Julie Urbain, Mr Lieven Baert
Contract:	Infeurope[1]
Contact:	Mariana Damova[1], Bernhard Kerpen
	[1] Infeurope S.A. has subcontracted Mozaika Ltd. to execute on the LOD Prototype and deliverable of ISA² Action 2017.01 Phase II: Work package 4

Document

Version date:	11/11/2019
Status:	WP4 - Final version
Version:	1.00
Author:	Mariana Damova
Reviewed by:	Lieven Baert (OIB.HAS), Emmanuel Dervaux (OIB.HAS), Theodor-Bogdan Dimofte (DIGIT), Sorin Bobeică (DIGIT), Antonio Palma-Gómez (SG), Julie Urbain (OIB.HAS), Annemieke Vanlaer (SG).
Approved by:	Lieven Baert (OIB.HAS), Annemieke Vanlaer (SG), Béla Harsányi (DIGIT)

Version Releases

Versions	Date	Description
0.1	11/08/2019	Initial version - draft
0.2	06/09/2019	Revised version
0.3	24/10/2019	Version with comments included
0.4	17/11/2019	Final version with comments included
1.0	27/11/2019	Final reviewed version

Abstract

This document presents the LOD prototype that is part of Work package 4 of PHASE II of ISA² Action 2017.01. It outlines the scope and the approach to expose archival data as linked open data (LOD) and describes the implementation of the prototype as a SPARQL endpoint, being the reference architecture of a LOD component as part of a larger archival system, and the design principles of the LOD approach. This includes a proposed data model, the HAS ontology, the reused datasets and the LOD prototype knowledge base along with examples of queries and query results showing inference and enrichment on the implemented SPARQL end point. Finally, the integration of the LOD prototype with traditional AMS systems is presented.

Table of Contents

List of Figures

List of Tables

1 Introduction

ISA² is the programme of the European Commission that supports cross-border and cross-sectorial public services for public administrations, businesses and citizens in Europe. Its goal is to develop, maintain and promote an integrated approach to interoperability in the EU and to contribute to the development of reusable IT solutions at European, national, regional and local levels of public administration.

In this context, the ISA² action 2017.01 "*Standard-based archival data management, exchange and publication*" was launched by DG DIGIT (DIGIT.B.2), being the DG for Informatics of the European Commission and OIB (OIB.OS.1.002), being the Historical Archives Service of the European Commission (HAS). The first phase of this action had the following objectives:

1. On the one hand, to provide an overview of the current landscape, as far as the business processes and the use of data standards and IT tools implemented by archival institutions are concerned. This information can be used by the HAS for the selection of their future archives management system, and can serve as a reference and inspiration for other organisations and institutions in Europe that deal with archives management.

2. On the other hand, to analyse the latest trends and options for the publication of relevant archival metadata as (Linked) Open Data. In the analysis, elements that will streamline interoperability, such as the use of shared reference data or authority lists will get special attention.

During the second phase of the action, the "proof of concept phase", prototypes of specific archives management systems were delivered in order to examine the interoperability with existing architecture components (like the digital preservation and the digitisation systems). Secondly, a specific prototype was set up in context of linked open data publication. The current document reports the activities carried out in the context of linked open data prototype.

The purpose of the LOD prototype is to show how the archival information can be exposed in an optimal way for user consumption, opening the archival content for viewing and consultation providing access to the documental history of a nation, institution or a subject domain.

Archives have in general two major kinds of users: professional archivists and the general public. Archivists use archives management systems (AMS) to support the processing of archives in various ways: acquisition, appraisal, archival description and publication just to name some. As far as the general public is concerned, since the rise of the WWW, consistent efforts have been made to make information from the archives available for the public to facilitate access to the wealth of archival information for research, consultation or educational purposes. Public websites with search capabilities have been provided by most archives institutions, including the HAS. Archives can also be considered cultural heritage and can link with information from museums, libraries and other heritage institutions (GLAM[4] sector). For this reason, it becomes ever more important to make it possible to connect the information from the different institutions in harmonised knowledge bases, storing linked metadata and descriptions and providing operational interoperability between the heritage information available in various institutions.

The current technological progress offers continued improved mechanisms to search and retrieve relevant information. Looking at Google alone, it has not gone unnoticed that search results are being presented to the user in a more semantically structured way. This is made possible by the introduction of semantic or alias linked open data technologies within the Google search component.

Semantic Technologies[1], or Linked Open Data technologies, have been introduced to meet the information management needs of 21st century. They are based on standards for data modeling and representation, and rely on data storage software, called triple stores or semantic repositories. Semantic technologies allow for an unprecedented ease of integration of heterogeneous data sources, c.f. Figure 1. When comparing with the RDBMs paradigm, there is nothing that can be expressed with semantic technologies that cannot be expressed in relational models. However, the difference between these two technologies is in the cost of production, subsequent maintenance and efficiency of hardware utilization.

Advantages of semantic technologies:

- RDF[2] – the basic data format of semantic technologies, is schema agnostic having both data and schema represented in the same RDF format. It is as easy to add data as to change their schema by adding RDF triples into the semantic repository. This makes the cost of maintenance of semantic repositories over time much lower than the cost of maintenance of RDBMs.
- Only the available information is being stored in the triple stores, e.g. there is no waste of valuable hardware space because of the need to comply with the specification of the relational tables by leaving empty cells.
- Inference, e.g. the number of the explicitly introduced facts in the database can be up to one third or even more than the actual available facts for querying, as new knowledge gets created based on the models that allow for the generation of new implicit facts out of the explicit ones.
- Easy interlinking between data from different sources, as they are based on open standards. This enables retrieval of information from different datasets with a single query.
- Easy combination with natural language processing tools to make possible to register the occurrence of conceptual entities in texts.

[1] Semantic technologies are being used here as a synonym of Linked data technologies.

[2] http://www.w3.org/RDF/

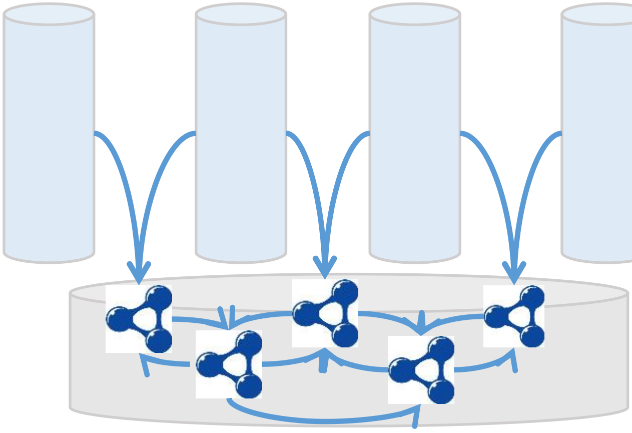


Figure 1 Semantic data integration breaking data silos

The information infrastructures based on these technologies allow for flexible querying, based not only on keywords, but also on semantic concepts. Thus, we can imagine queries, retrieving structured information about the searched objects and structured information about the documents they are associated with.

The LOD prototype is an implementation of such a semantic infrastructure that encompasses the HAS information that is of interest to the general public. By general public we envisage the typical users of this open access information to be, e.g. researchers – historians, political scientists, politicians, students, scholars or citizens, jurists, journalists, public bodies, etc. While each may have different end goals, these categories of users would be all interested in finding documents based on their title, topic, temporal coverage, time of creation, the people, the organisations and the locations that occur in the documents. Therefore, this information, available in the content level description of archives management system, is selected, modelled, represented and made available for querying in the LOD prototype.

The LOD prototype is designed to show the advantages for the different categories of users, e.g. citizens, researchers, historians, public bodies, but also archive professionals from other archival organisations, to have access to the archival data, managed by the HAS when they are represented as linked open data. That is why the most important feature of the LOD prototype has been to showcase the possibilities and the effects of linking, enriching and re-using existing resources for the end users, but also to the creators of authority vocabularies and respectively of archival records.

2 LOD Prototype

Archives and the archiving process encompass complex business processes that are suitable to be handled as linked data. The archiving requires the processing and the management of sometimes dislocated and dispersed data, data exchange and publication. LOD are most well known for their ability to connect objects and provide content enrichment. The main goal of the LOD prototype is to show how LOD technologies can be applied in the context of the HAS and how relevant they might be for the general public but also for reuse by other organisations. For this reason, the scope of the LOD prototype has been set up to cover end user facing features allowing to search and obtain information, and to demonstrate semantic integration of data from the HAS with datasets from the Publications Office, re-using the already existing resources of the European Commission, and with datasets containing general purpose information such as DBPedia and Wikipedia to show how data enrichment takes place. These general purpose datasets have been used as the European Commission's Publications Office does not provide with such a pool of structured linked data about the politicians and the organisations that are related to the records in the HAS.

To achieve this demonstration in the LOD prototype, it is necessary to answer the following questions:

1. What part of the information available at the HAS is relevant in this context?

2. How the information is exposed to the general public or other organizations?

3. How the information to be exposed is retrieved from the HAS?

To answer the first question, we have to analyse the archives management process and the information available for the archived entities. For the second question, we have to analyse the way archival information is being exposed by other archival institutions and what access and publications channels are already available within the European Commission. To answer the third question, we have to analyse the way the relevant information is organised, stored and accessible.

This document and the LOD prototype give an answer to these three questions. We deliver a site, a SPARQL end point that allows users, knowledgeable in the LOD query language SPARQL to formulate queries and to inspect and navigate through the query results experiencing the effects of them being in LOD format. The SPARQL[3] language is machine readable. Thus, the SPARQL end point is also the interface between the semantic linked open data, available in the semantic infrastructure and other components of a bigger system, such as a friendly graphical user interface, via a standard access protocol (REST API[4]).

[3] https://www.w3.org/TR/rdf-sparql-query/

[4] https://searchapparchitecture.techtarget.com/definition/RESTful-API

3 Archival Business Processes

The Historical Archives Service of the European Commission (HAS) archives physical documents (non digital) and digital documents that come from digitisation processes, or are only available in digital format. Submitted archives to the archive follow a transfer process. Documents are aggregated to files (record set) and metadata describe the relationships.. The records sets and documents contain (amongst others) descriptive metadata. To manage the archival context of the archives, ISAD(G) and ISAAR based encoding is applied. It comprises the archiving process level and the content level descriptions of the archived units, presented in detail in tables 1-28 in Appendix 2.

4 LOD approach

Technically, the linked open data are defined by the creator of these technologies, as part of the evolving WWW, Sir Tim Berners Lee, in “Architectural and philosophical points”[5]. Sir Tim Berners Lee describes the 5-Star linked data in the following manner. They should be:

★ Available on the web (whatever format) *but with an open license, to be Open Data*

★★ Available as machine-readable structured data (e.g. excel instead of image scan of a table)

★★★ as (2) plus non-proprietary format (e.g. CSV instead of excel)

★★★★ All the above, plus: Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff

★★★★★ All the above, plus: Link your data to other people’s data to provide context

Methodologically, the process of linked data creation consists in several steps. They can be summarised in the following way:

1. Selection or creation of a model, ontology, to describe the data conceptually
2. Creation of datasets based on the model
3. Linking the data by identifying items that describe one and the same object
4. Enriching the data by identifying LOD objects that can be associated with the data item and attaching them to it

The best practice for the realisation of point 1) above is identifying and re-using existing available resources, e.g. ontologies when possible. The second point (2) above usually involves the transformation of one data source into a semantic LOD format. The third and the fourth points (3) and (4) are realized by automatic, semi-automatic or manual processes of creating linked and enriched datasets.

There are different ways to approach and execute these four steps. They depend on the subject matter domain, on the available models, on the source datasets.

[5] https://www.w3.org/DesignIssues/

The adopted approach of building the ontological model of the archives management domain – the HAS ontology – is conceptual analysis, and consequently OntoClean[6]. OntoClean is a method for defining the ontology elements following the rule to distinguish between objects and roles. The analysis of the archival management domain and of the scope and requirements of the LOD prototype and the HAS sample datasets showed that there will be several conceptual domains to be covered in the HAS ontology and in the HAS knowledge base. These different conceptual domains have to be represented by proper conceptual models. Figure 2 shows how ontological models are built from different conceptual domains and respectively source datasets or data bases and then merged into a single ontological model.

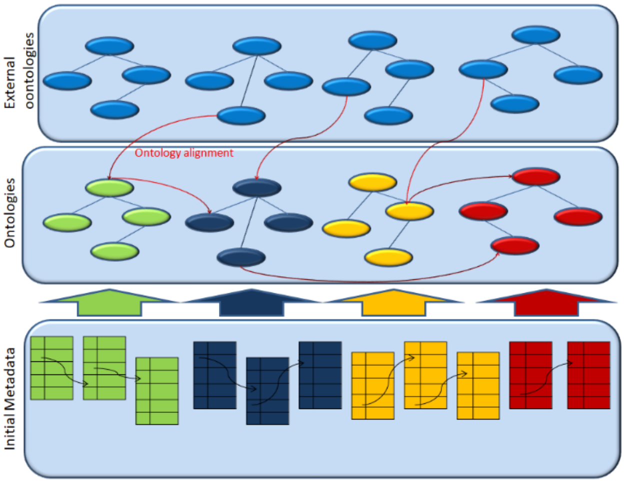


Figure 2 Ontology modeling approach

Following the best practices of linked open data models design, the LOD prototype ontology re-uses concepts and relations from popular and standard vocabularies, such as Dublin Core[7], SKOS[8], DBpedia ontology[9], PROtoN[10] that model general domain knowledge and will be outlined in greater detail in section Ontologies and Schemata below, and also a legal domain ontology to be also described in the same section.

In addition, the HAS ontology defines original concepts and relations based on the analysis of the sample data, provided by the HAS of the European Commission, c.f. Section HAS Archive Datasets below, and of several archive management models, e.g. RIC-CM[11], EDM[12], ISAD(G)[13], ISAAR(CPF)[14], RICO[15]. The conceptual model of HAS ontology (HASO) adopts the definitions of concepts from the standards ISAD(G), ISAAR and RIC-CM models. However, the adoption of concepts of these standards is incorporated into an original ontological model

[6] https://www.l2f.inesc-id.pt/~joana/prc/artigos/07a%20An%20overview%20of%20OntoClean%20-%20Guarino,%20Welty%20-%202004.pdf

[7] https://www.dublincore.org/schemas/rdfs/

[8] https://www.w3.org/2004/02/skos/vocabs

[9] https://wiki.dbpedia.org/services-resources/ontology

[10] https://ontotext.com/documents/proton/Proton-Ver3.0B.pdf

[11] https://www.ica.org/en/egad-ric-conceptual-model-ric-cm-01pdf

[12] https://pro.europeana.eu/resources/standardization-tools/edm-documentation

[13] https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

[14] https://www.ica.org/en/isaar-cpf-international-standard-archival-authority-record-corporate-bodies-persons-and-families-2nd

[15] https://www.ica.org/en/ric-o-extended-call-for-reviewers

that is not hierarchical, but a conceptual graph structure with named concepts and named relations between them.

The source concept in HASO is ISAD(G)'s “Unit of description”. In HASO the “Unit of description” instantiates different possible ISAD(G)'s “Levels of description”, the concept that gathers all manifestations of archival content organisation – Fonds, Series, File or Record Set, Item.

5 LOD Prototype Architecture

The high-level architecture of the system consists of three layers (c.f. Figure 3). The first layer, the top level on Figure 3, deals with the creation of the semantic metadata converting the metadata from the archives management system[16] (AMS) into RDF and injecting them into the second layer. Here the component “Metadata collector” harvests the data from the AMS, the component “Metadata normalize” prepares the harvested data to be converted into RDF by unifying their format. Next, the component “Metadata converter” transforms the harvested data into RDF, the “Metadata generator” component is where the metadata are enriched by attaching additional relevant information for them, and finally the “Metadata injector” sends the prepared RDF data into the semantic repository in the second layer, the middle level on Figure 3.

The second layer holds the semantic repository where the semantic metadata are stored, e.g. the component “Metadata repository”. We have provided in this layer also components for the version control (“Version control” component) and for backup (“Metadata backup” component) of the semantic data, incoming periodically from the first layer. The third layer is the one creating the interface between the users and the semantic linked open data. Note, that as the LOD prototype is implemented as a SPARQL endpoint (c.f. Section SPARQL Endpoint below), this layer is exemplified in the architecture on Figure 3, as part of the reference architecture on Figure 3 below, but not implemented in the LOD prototype that is the subject of the current deliverable. The component “Query generator” is supposed to take input from the graphical user interface, pictured on Figure 3 as the fourth layer “User interface”. The component “Query templates” is supposed to hold query patterns that can be sent to query the semantic linked open data repository as shortcuts. The function of these two components is performed in the LOD prototype by the end users, as the SPARQL end point exposes the data at the level of the semantic repository. As stated above the querying of the data on the SPARQL end point can be performed with SPARQL queries, the programming query language for semantic data. The SPARQL endpoint can be queried by users, knowledgeable in the SPARQL language, but also by algorithms, as they give access to linked machine readable semantic content. The component “Metadata analytics” can derive analytics about the metadata or other necessary intelligent operations.

The fourth layer of the architecture of Figure 3 below is the User interface layer. It refers to the proper graphical user interface, allowing end users to interact with the data in an easy and pleasing manner. This layer is also out of the scope of this LOD prototype, aiming to only showcase semantic data integration, metadata enrichment and exposure of archival data.

[16] Note. The figure shows one image of archives management system (AMS), implying connection with a single archives management system (AMS). It is important to keep in mind that the presented architecture can consume data for several archives management systems, not just one.

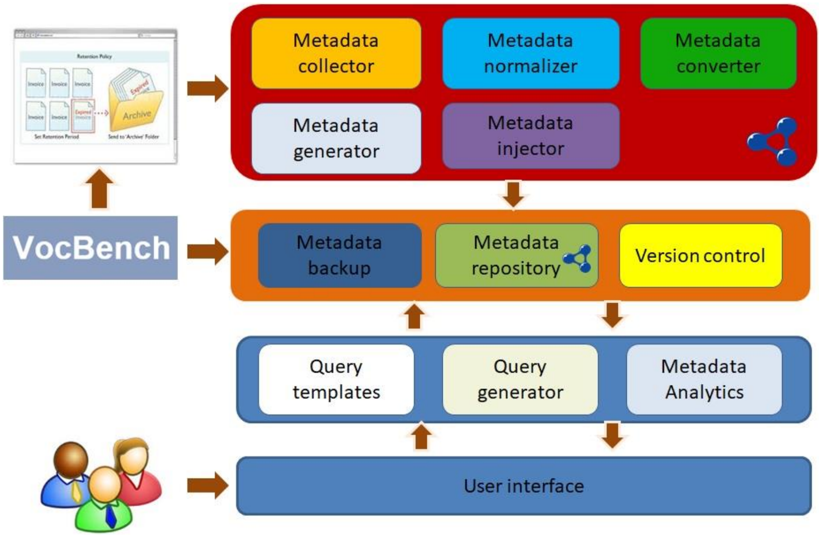


Figure 3 LOD Prototype Architecture

VocBench[17] is a collaborative, web-based, multilingual tool used by the European Commission to create authority vocabularies in RDF[18] and the format of the semantic linked open data. It also maintains thesauri, code lists and authority resources, provides advanced collaboration features such as history, validates and publishes workflows, and manages multiple users with *role-based access control"*[19]. That is why it is important to provide a vision of integration of VocBench with the semantic linked data infrastructure, the presumed LOD component of a larger operational system integrating AMS with LOD and VocBench (c.f. Figure 3) without implementing this in the LOD prototype. The integration is on Figure 3. Here we envisage two scenarios: 1) The RDF data generated in VocBench are introduced into the Archive management system as mere data records, and pulled from there by the Metadata collector of the first layer of the LOD component architecture on Figure 3 to be made available for linking, metadata enrichment and injected into the metadata repository; 2) The RDF data generated in VocBench get directly inserted into the Metadata repository in the second layer of the architecture on Figure 3.

[17] http://vocbench.uniroma2.it/

[18] https://eur-lex.europa.eu/browse/eurovoc.html?locale=en

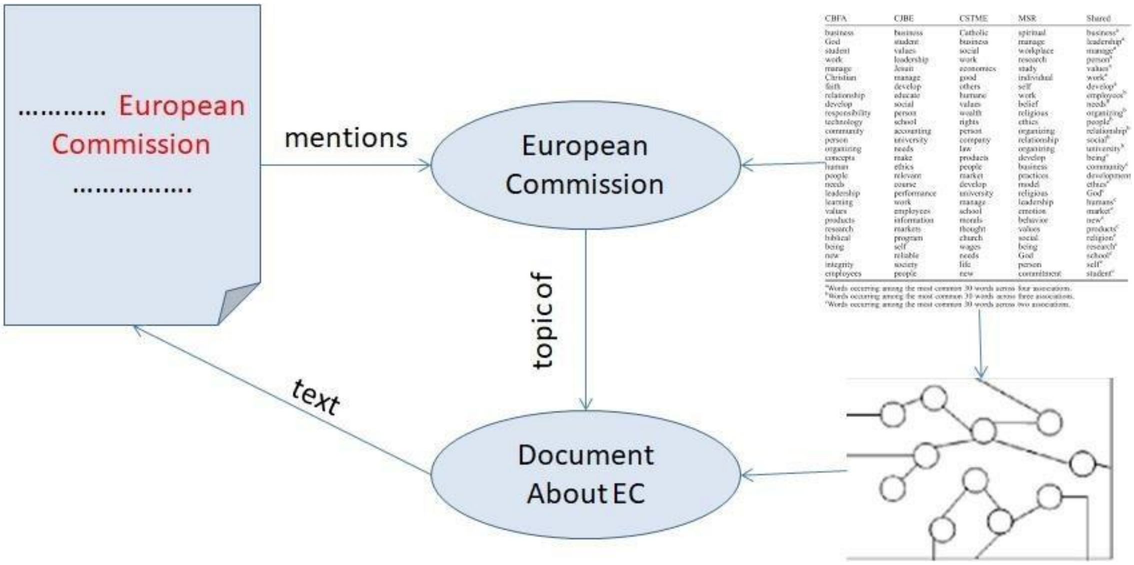
[19] http://vocbench.uniroma2.it/

6 Metadata enrichment

The metadata enrichment process is located, in the system architecture, under the Metadata generator component, in the first layer of the architecture on Figure 3. Because it involves complex procedures of identifying, extracting and linking of information, we dedicate a special section to it.

The metadata enrichment is a process of detecting objects in text and identifying and making available contextual information about these objects. Figure 4 shows an example of this process. It shows how a portion of a written text (e.g. “European Commission”) is recognised as describing a given object via the controlled vocabularies, or authority lists which is assigned as a topic to the object referring to the document holding the text. These controlled vocabularies are the lists of keywords, synonyms, topics, positions/roles, organisations, organisational functions, persons, etc. available in the HAS, but also provided by the Publications Office of the European Commission and generated with VocBench. VocBench enables the production of authority lists from the Publications Office and EuroVoc in RDF.

In other words, Figure 4 shows an example of metadata enrichment where it is identified that a conceptual object, an entity is mentioned in a text by making the connection between the string, e.g. the sequence of symbols (letters or words), in the text with the object from the authority list. The identified object is labelled as the topic of the document containing the text where the conceptual object has been mentioned. Furthermore, the authority lists and the description of the documents are part of the graph in the semantic linked open data repository, or the Metadata repository component of the architecture on Figure 3. The graph structure makes the objects and the relations between them explicit and accessible, thus linking all information in the graph.



CBFA	CJBE	CSTME	MSR	Shared
business	business	Catholic	spiritual	business[a]
God	student	business	manage	leadership[a]
student	values	social	workplace	manage[a]
work	leadership	work	research	person[a]
manage	Jesuit	economics	study	values[a]
Christian	manage	good	individual	work[a]
faith	develop	others	self	develop[a]
relationship	educate	humane	work	employees[b]
develop	social	values	belief	needs[b]
responsibility	person	wealth	religious	organizing[b]
technology	school	rights	ethics	people[b]
community	accounting	person	organizing	relationship[b]
person	university	company	relationship	social[b]
organizing	needs	law	organizing	university[b]
concepts	make	products	develop	being[c]
human	ethics	people	business	community[c]
people	relevant	market	practices	development[c]
needs	course	develop	model	ethics[c]
leadership	performance	university	religious	God[c]
learning	work	manage	leadership	humans[c]
values	employees	school	emotion	market[c]
products	information	morals	behavior	new[c]
research	markets	thought	values	products[c]
biblical	program	church	social	religion[c]
being	self	wages	being	research[c]
new	reliable	needs	God	school[c]
integrity	society	life	person	self[c]
employees	people	new	commitment	student[c]

[a] Words occurring among the most common 30 words across four associations.
[b] Words occurring among the most common 30 words across three associations.
[c] Words occurring among the most common 30 words across two associations.

Figure 4 Metadata generation

7 Archives Description Standards

The creation of the LOD prototype HAS ontology (HASO) is based on the archival description standards. The following sections give an overview of the standards that we took into consideration while modeling HASO, as most commonly used and fundamental standards of archival description.

7.1 ISAD(G)[20]

ISAD(G), General International Standard Archival Description, addresses the contextual information accompanying the description of archives. ISAD(G) also acknowledges linked contextual information to the combination of other elements used together to describe archives and records.

ISAD(G) standard provides general guidance for the preparation of archival descriptions by defining twenty-six elements that may be combined into the description of an archival entity. They are meant to be applicable to archival descriptions regardless of the nature or extent of the unit of description. The standard does not define output formats, or how these elements are presented – in inventories, catalogues, lists, etc.

The 26 elements are distributed into seven categories. They are:

- Identity statement area – elements identifying the record, such as reference code and title, etc.
- Context area – elements describing creators, history, source of acquisition or transfer, etc.
- Content and structure area – elements describing the scope (such as, time periods, geography) and content, (such as documentary forms, subject matter, administrative processes), accruals, appraisal, etc.
- Conditions of access and use area – elements about the conditions governing access, reproduction, physical characteristics, etc.
- Allied materials area – elements about the existence and location of originals, copies, related units of description, etc.
- Notes area – information that cannot be represented by other elements
- Description control area – archivist's notes, rules or conventions, etc.

There are levels of description, with varying degrees of detail, appropriate to each level of arrangement. The fonds form the broadest level of description; the parts form subsequent levels, whose description is often only meaningful when seen in the context of the fonds. Therefore, there may be a fonds-level description, a series-level description, a file-level description and/or an item-level description. Intermediate levels, such as a sub-

[20] https://www.ica.org/sites/default/files/CBPS_2000_Guidelines_ISAD%28G%29_Second-edition_EN.pdf

fonds or sub-series, may be expected. Each of these levels may be further segmented according to the complexity of the structure and/or functions of the organisation which generated the archival material and the organisation of the material.

ISAD(G) model is hierarchical. Figure 4 shows the hierarchical relationship between the different types of fonds in the archives.

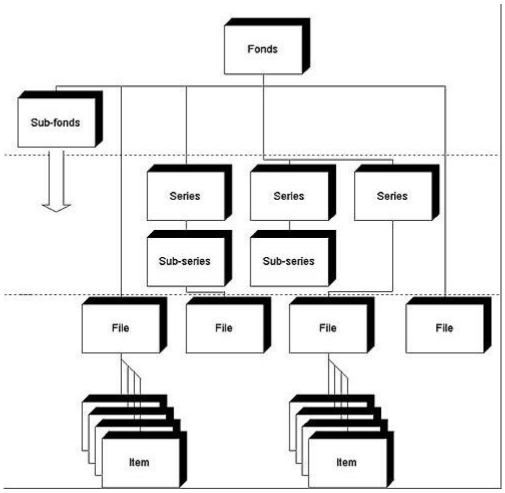


Figure 5 ISAD(G) Model of the levels of arrangement of a fonds

Very few elements are considered obligatory for international exchange of descriptive information. They are:

1. reference code
2. title
3. creator
4. date(s)
5. extent of the unit of description
6. level of description

The extent to which a given archival description will contain more than these elements depends on the nature of the unit of description.

7.2 ISAAR(CPF)[21]

This standard provides guidance for archival authority records which describe entities (corporate bodies, persons and families) associated with the creation and maintenance of archives. Archival authority records are used to describe a corporate body, person, or family as units within an archival descriptive system; to control the creation and use of access points in archival descriptions; to document relationships between different records creators and between those entities and the records created by them and/or other resources about or by them.

Description of records creators requires full documentation and constant maintenance of the context of records creation and usage, most importantly the provenance of information.

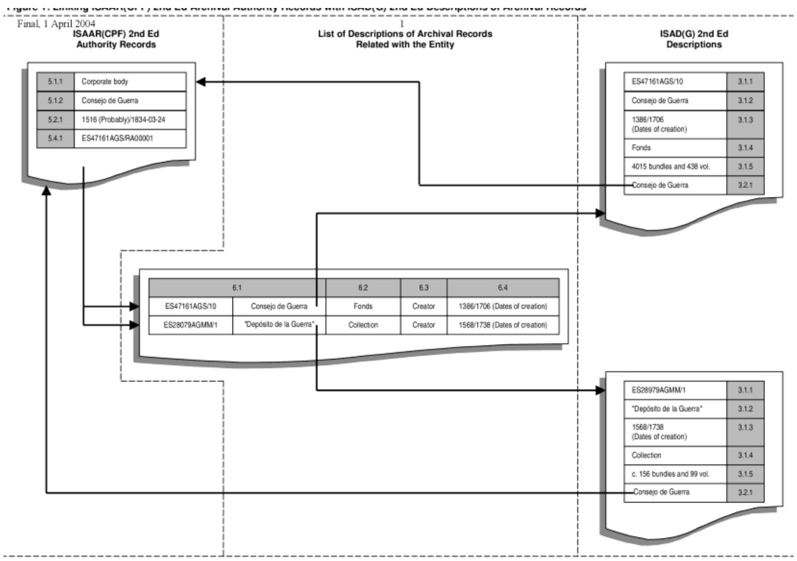
As a standalone description, ISAAR(CPF) allows linking of record creator descriptions and contextual information to record descriptions from the same creator(s) that could be found in more than one repository and to descriptions of other related resources – library and museum materials.

ISAAR(CPF) provides general rules for the standardisation of archival descriptions of records creators and the context of records creation, enabling access to archives and records based on the provision of these descriptions that are linked to descriptions of the frequently diverse and physically dispersed records themselves; understanding of the context underlying the creation and use of archives and records so that users can better interpret their meaning and importance; accurate identification of records creators incorporating descriptions of relationships between different entities, particularly documentation of administrative change within corporate bodies or change of circumstances in individuals and families; and the exchange of these descriptions between institutions, systems and/or networks.

Archival authority records are comparable to library authority records because both need to support the creation of standardised access points in descriptions. Nevertheless, the archival authority records go much further and usually contain much more information than library authority records.

Additionally, ISAAR(CPF) model has a direct link to the ISAD(G) model. Figure 5 shows the relationship between these two standards. This allows the creation of rich authority record description with equally rich contextual record description.

[21] https://www.ica.org/sites/default/files/CBPS_Guidelines_ISAAR_Second-edition_EN.pdf



Source: Subdirección General de los Archivos Estatales, Archivo General de Simancas. Spain 2003.

Figure 6 ISAAR(CPF)[22] based creator representation

7.3 RiC-CM[23]

RiC-CM – Records in Context conceptual model – has been created in 2016 by an Expert group (EGAD) within the International Council on Archives (ICA) as a standard for the description of records based on archival principles. It reconciles, integrates, and builds on the four existing standards: General International Standard Archival Description (ISAD(G)); International Standard Archival Authority Records—Corporate Bodies, Persons, and Families (ISAAR(CPF))[24]; International Standard for Describing Functions (ISDF)[25]; and International Standard for Describing Institutions with Archival Holdings (ISDIAH)4[26]. RiC-CM's primary usage is by the archival community, followed by the records management community, and the allied cultural heritage communities.

RiC-CM defines central descriptive entities, their properties or attributes, and important relations among them without further specifying and defining the relations among the entities in a formal way. It does not cover a model for the role and the activities of the archivist in the process of formulating a record and of maintaining it.

RiC-CM is driven by two methodological rules of archiving reflecting its ground principle: "the Principle of Provenance":

1. Respect of fonds – emphasizes and privileges the person or group that has accumulated a body of records

[22] https://www.ica.org/sites/default/files/CBPS_Guidelines_ISAAR_Second-edition_EN.pdf

[23] https://www.ica.org/sites/default/files/RiC-CM-0.1.pdf

[24] http://www.ica.org/en/isaar-cpf-international-standard-archival-authority-recordcorporate-bodies-persons-and-families-2nd

[25] http://www.ica.org/en/isdf-international-standard-describing-functions

[26] http://www.ica.org/en/isdiah-international-standard-describing-institutions-archivalholdings

2. Respect for original order – the intellectual order and the physical order of the records.

RiC-CM is intended to enable archivists to improve archival description. These descriptions help the management of records, their preservation, and their usage. For digital records, file systems facilitate the storage of files, most commonly using a hierarchical directory structure that is an equivalent of the hierarchical storage of physical files, including the use of directory, folder labels and other metadata. As the goal is to establish intellectual control over records and locate, identify, and get access to them, it is essential to supplement the rudimentary metadata with additional description of contexts. Records exist in shifting environments, and documenting as fully as possible both the complex origins and ongoing history is important for evaluating the evidentiary quality of records, and for understanding them.

The primary description entities of RiC-CM are:

- Record
- Record Component
- Record Set
- Agent
- Occupation
- Position
- Function
- Function (Abstract)
- Activity
- Mandate
- Documentary Form
- Date
- Place
- Concept/Thing

RiC-CM is a multidimensional description including the single, fonds-based, multilevel description modelled in ISAD(G), but also addressing the more extensive understanding of provenance. The Records and Sets of Records, their interrelations with one another, their interrelations with Agents, Functions, Activities, Mandates, etc., and each of these with one another, are represented as a network within individual fonds (see Figure 7).

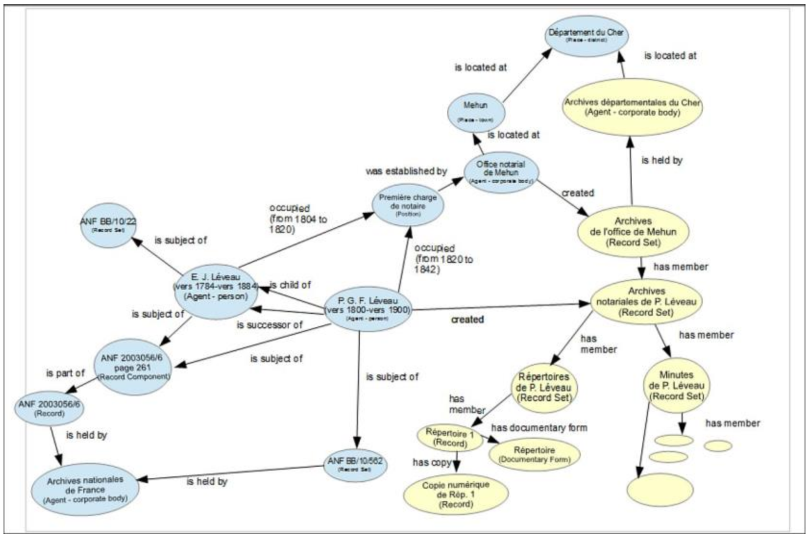


Figure 7 RIC-CM Model

RiC-CM is intended to provide the semantic and structural basis for developing record description systems or description modules in the context of records management systems. It defines 14 entities – RiC-E, 69 properties – RiC-P, and 792 relations – RiC-R.

Being a conceptual model RiC-CM is a guideline to be followed, but it is not a proper formal model. A closer look at the names and the definitions of entities, properties and especially relations points to a certain ambiguity, having, for example, the relation "associated with" used more than 100 times in different contexts, or the property "type" mentioned more than 10 times in different contexts. That is why the HAS model adopts RiC-CM not as a guideline, but as a formal model that ensures machine-readable format, machine-interpretable semantics and automated reasoning had to be built.

7.4 RIC-O[27]

RIC-O is an initiative to formalise the RIC-CM model into an ontology. It has been developed for several years now and the official release is due in the fall of 2020. As the HAS ontology (HASO), (c.f. section HAS ontology) is also based on RIC-CM model, it defines several concepts that can be found in RIC-O as well. However, the application domain of HASO – exposing of archival data in linked open data format to the general public, has a lot of concepts and relations that are not part of RIC-O, and can be adopted or re-used by RIC-O. HASO provides a comprehensive model of both ISAD(G) and ISAAR(CPF) standards reflecting the functional requirements of presenting archival information to the general public.

[27] http://piaaf.demo.logilab.fr/

8 Datasets of the LOD Prototype

The LOD prototype includes and re-uses several datasets from the Publications Office and General purpose datasets that are instrumental for the demonstration of the metadata enrichment. These datasets are described in the following sections.

8.1 Publications Office Datasets

The Publications Office of the European Commission develops, publishes, maintains and manages datasets of linked open data, produced by the EU institutions and bodies that are available for reuse – i.e. controlled vocabularies, schemas, ontologies, data models, authority lists. The authority lists are vocabularies describing different categories of conceptual objects, e.g. entities, according to SKOS[28] schema, and provide the identifier of the conceptual object, its category and the language expression that describes the entity in at least all official languages of the European Union. This makes it possible to identify the conceptual object, e.g. the entity, from documents in different languages. Publications Office- This is the reason why the LOD prototype includes general purpose datasets – to show the effects of metadata enrichment.

The following sections present the datasets that have been included into the LOD prototype.

8.1.1 Place

Two datasets fall into the category of "places." One contains objects describing places[29] (cf. example 1) and the other describes countries[30] (cf. example 2). No statistics about the size of the datasets is available, but the LOD prototype reuses the identifiers for cities and countries.

Example 1

```
<rdf:Description rdf:about="http://publications.europa.eu/resource/authority/place/1A0_PRN">
<skos:inScheme rdf:resource="http://publications.europa.eu/resource/authority/place"/>
<skos:topConceptOf rdf:resource="http://publications.europa.eu/resource/authority/place"/>
</rdf:Description>
```

[28] http://www.w3.org/2008/05/skos-xl#

[29] http://publications.europa.eu/resource/authority/place/

[30] http://publications.europa.eu/resource/authority/country/

Example 2

```
<rdf:Description rdf:about="http://publications.europa.eu/resource/authority/country/FRA">
<skos:inScheme rdf:resource="http://publications.europa.eu/resource/authority/country"/>
<skos:topConceptOf rdf:resource="http://publications.europa.eu/resource/authority/country"/>
</rdf:Description>
```

8.1.2 Corporate body

One dataset contains authority list of entities describing corporate bodies[31] (cf. example 3). The corporate bodies authority lists contain identifiers of European institutions and organisations, like the one in the example "Representation in France". There are no statistics about the number of entries of this authority list.

Example 3

```
<rdf:Description rdf:about="http://publications.europa.eu/resource/authority/corporate-body/REPRES_FRA">
<skos:inScheme rdf:resource="http://publications.europa.eu/resource/authority/corporate-body"/>
<skos:topConceptOf rdf:resource="http://publications.europa.eu/resource/authority/corporate-body"/>
</rdf:Description>
```

8.1.3 Person

One dataset contains authority list of entities describing people[32] (cf. example 4). The people represented in this authority list are European politicians and historical personalities. There are no statistics about the number of entries in this authority list.

Example 4

```
<rdf:Description rdf:about="http://publications.europa.eu/resource/authority/fd_014/SAVINI">
<skos:inScheme rdf:resource="http://publications.europa.eu/resource/authority/fd_014"/>
</rdf:Description>
```

8.1.4 Legal Datasets

The Publications Office has published lists describing legal entities[33]. The legal entities represented in these authority lists are legal concepts, such as the entry for APPEAL in example 5. They are also described with the identifier and the dataset they belong to. There are no statistics about the number of entries in this authority list.

[31] http://publications.europa.eu/resource/authority/corporate-body/

[32] http://publications.europa.eu/resource/authority/fd_014/

[33] http://publications.europa.eu/resource/authority/procjur/

Example 5

```
<rdf:Description rdf:about="http://publications.europa.eu/resource/authority/procjur/APPEAL">
<skos:inScheme rdf:resource="http://publications.europa.eu/resource/authority/procjur"/>
<skos:topConceptOf rdf:resource="http://publications.europa.eu/resource/authority/procjur"/>
</rdf:Description>
```

8.2 EuroVoc[34]

EuroVoc is a multilingual, multidisciplinary thesaurus describing the activities of the EU with terms describing concepts and entities from the government and from the public sector domains in 27 EU languages[35]. The structure of the vocabulary entry is shown in example 6. It contains the entity, its status, the time of creation and the language description of the entity in the 27 languages. The identifiers in EuroVoc authority lists are sequences of digits. Example 6 shows only the language description of the entity in French.

Example 6

```
<rdf:Description rdf:about="http://eurovoc.europa.eu/230549">
    <rdf:type rdf:resource="http://www.w3.org/2008/05/skos-xl#Label"/>
    <rdf:type rdf:resource="http://eurovoc.europa.eu/schema#SimpleNonPreferredTerm"/>
    <euvoc:startDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1952-06-16</euvoc:startDate>
    <euvoc:status rdf:resource="http://publications.europa.eu/resource/authority/concept-status/CURRENT"/>
    <dct:created rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1995-10-02</dct:created>
    <owl:versionInfo rdf:datatype="http://www.w3.org/2001/XMLSchema#string">n/a</owl:versionInfo>
    <dct:type rdf:resource="http://publications.europa.eu/resource/authority/label-type/ALTERNATIVELABEL"/>
    <skosxl:literalForm xml:lang="fr">aperçu historique</skosxl:literalForm>
</rdf:Description>
```

EuroVoc has also authority lists, describing places, countries, people, corporate bodies, EU institutions and bodies with identifiers different from the ones in the Publications Office datasets presented in the previous section.

The LOD prototype uses both the Publications Office and the EuroVoc identifiers of people, countries, places, corporate bodies, EU institutions and bodies, depending on their availability in the corresponding datasets. The LOD prototype also demonstrates linking between identifiers from datasets from the Publications Office and identifiers from datasets from EuroVoc.

[34] http://eurovoc.europa.eu/

[35] Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, Gaeilge, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish and Swedish , plus in three languages of countries which are candidate for EU accession: македонски (mk), shqip (sq) and српски (sr).

8.3 External Datasets

In addition to the Publications Office datasets, the LOD prototype includes entries from external datasets to demonstrate the metadata enrichment capabilities of linked open data architectures. We have used DBpedia and Wikipedia as the most popular and rich open data resources.

8.3.1 DBpedia[36]

DBpedia is an open online database that collects structured machine-readable content extracted from the information provided by various Wikimedia projects. The structured content resembles an open knowledge graph, which makes it possible for information to be organised, searched and utilised. DBpedia data is served as Linked Data, which means that anyone can navigate it with standard web browsers, automated crawlers or use queries with SQL-like languages. DBpedia is available in 125 languages.

The English version of DBpedia contains more than 4.5 million articles, 4.2 million of those are classified in an ontology. It provides information about 1.45 million persons, 735,000 places, 411,000 creative works (including 123,000 music albums, 87,000 films and 19,000 video games), 241,000 organisations (including 58,000 companies and 49,000 educational institutions), 251,000 species and 6,000 diseases.

8.3.2 Wikipedia

Wikipedia is a free online encyclopedia. It was launched on January 15 2001 and hosted by the Wikimedia Foundation. Wikipedia consists of articles that provide links to related pages with additional information and sources for all facts stated. The articles are created by volunteers and anyone can contribute, either anonymously, or under their real name. The Wikipedia community has developed many policies and guidelines related to content creation and editing, in order to address credibility of information and to improve the service.

Wikipedia is widely used and very popular reference website. It attracts more than 374 million unique visitors monthly. The online encyclopedia provides articles written in 307 languages (297 active). There are about 315,000 active users working on more than 51 million articles. As of today, there are almost 6 million articles in English.

8.4 HAS Sample Datasets

The most important part of the LOD prototype are the HAS sample datasets that have been included in it. The sample data sets that have been selected for the prototype are:

[36] https://wiki.dbpedia.org/about

- The collection of official COM documents (1958-1967). These documents play a role in the EU's decision process and concern autonomous acts, proposals and other communications (including preparatory documents) from the Commission to the Council and/or other institutions.
- The archives of the President Roy Jenkins's Cabinet (1977-1981), when Sir Roy Harris Jenkins was the president of the European Commission.

Both datasets are available in several files comprising:

- the description of the schema of the dataset,
- the Datasets (csv),
  - Main descriptive metadata;
  - the creators and
  - samples of keywords pertaining to the entries of the datasets.

These separate files, which correspond to relational database tables, have been represented in a single semantic knowledge graph represented in the LOD prototype data model and knowledge base (c.f. section Data model and section HAS Knowledge Base). Figure 8 shows the chain of relations between the different tables, e.g. files, describing the data. The small squares point to the fields, that ensure the link between two tables, e.g. files, represented with the bigger squares. We have tables with keywords (MCL) that are linked to DOSSIERS-RUBRIQUES via the field dos_obj_id, DOSSIERS_RUBRIQUES linked to RUBRIQUES_INV via the field Rub_code, PRODUCTEUR_RUBRIQUES linked to RUBRIQUES_INV via the field Rub_code and finally PRODUCTEURS_RUBRIQUES linked to creators_ISAAR(CPF) via the field not_code. It will be explained in the data model and the HAS Knowledge Base how this structure has been represented semantically.

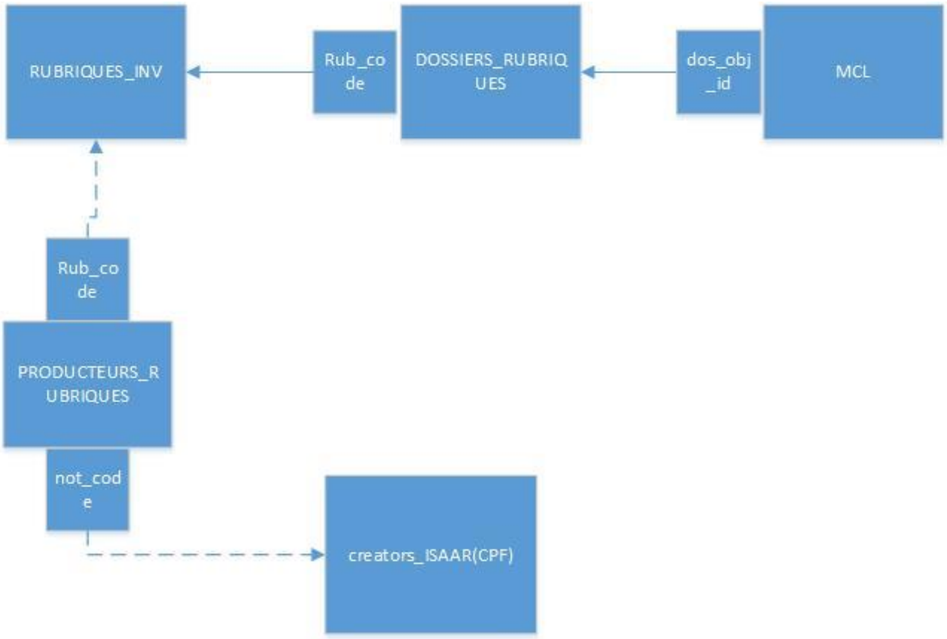


Figure 8 Relational model of HAS datasets

8.4.1 COM documents

The COM dataset concerns the collection of official COM documents (1958-1967). These documents play a role in the EU’s decision process and concern autonomous acts, proposals and other communications (including preparatory documents) from the Commission to the Council and/or other institutions. The collection has been created by the General secretariat of the European Commission and more precisely the Office of the registrar (1958 – 1967). Some of the descriptions are associated to internal controlled vocabularies such as topics places, people, or European institutions and bodies. Table 1 shows an example of the data from the COM dataset. These are the first two records, describing record sets, from year 1958.

A closer look at the scope_content column of Table 1 below shows that the cells contain text, describing two volumes of the record set described on the line, the one in French and the other in German. The text also contains further information about the two separate volumes, e.g. the Volume ID, its indicative data and its title. There is a separate column Languages making explicit in what languages the record sets are available. And finally, the columns links_display and links_uri show the description of the link to access and show the corresponding volumes. It will be shown below how these correlations have been semantically represented in the LOD prototype.

original_ref_code	Title	dates	description_level	extent_medium_units_of_description	scope_content	Languages	links_display	links_uri
COM(1958)1	COM(1958)1 – Questions concernant le personnel de la Communauté	1958	Dossier	2 volume(s) papier	Volume 1958/0001 Date indicative: 15/01/1958 Questions concernant le personnel de la Communauté (FRA) Volume 1958/0013 Date indicative: 15/01/1958 Fragen des Personals der Gemeinschaft (DEU)	DE, FR	Volume_1958/0001.pdf, Volume_1958/0013.pdf	http://publications.europa.eu/resource/cellar/3eddaa07-fead-4c40-9b83-496107a287b9.0001.01/DOC_1, http://publications.europa.eu/resource/cellar/b621c996-3f0a-46b3-9c90-7bfbbe23c807.0001.01/DOC_1
COM(1958)2	COM(1958)2 – Financement des premières dépenses de la Communauté	1958	Dossier	2 volume(s) papier	Volume 1958/0001 Date indicative: 15/01/1958 Financement des premières dépenses de la Communauté (FRA) Volume 1958/0013 Date indicative: 15/01/1958 Finanzierung der ersten Ausgaben der Gemeinschaft (DEU)	DE, FR	Volume_1958/0001.pdf, Volume_1958/0013.pdf	http://publications.europa.eu/resource/cellar/2ca82156-ae98-4417-a6ce-952d5af7e62b.0001.01/DOC_1, http://publications.europa.eu/resource/cellar/5ded1764-8189-4a2c-ade2-aa38c761c2cc.0001.01/DOC_1

Table 1 Data from COM dataset for 1958.

8.4.2 Archives of the Cabinet Roy Harris JENKINS (1977 – 1981)

The archives of the Cabinet Roy Harris JENKINS (1977 – 1981) are considered as a sub-fonds in the filing framework developed by the HAS. It gathers the archives of the Cabinet of Sir Roy Harris JENKINS, who was the president of the European Commission from 1977 to 1981.

It contains various documents that are useful to understand the overall functioning of the Commission and Communities and the evolution of the European integration, but also the economic, social and geopolitical context for the years 1977 to 1981 in Europe. The sub fonds includes miscellaneous types of documents, such as reports, minutes of meetings, administrative documents, correspondence, speeches transcripts, interviews, as well as notes and reports of diplomatic visits, conferences and ceremonies.

Table 2 shows a sample of two lines of the dataset, describing the record sets from the Archives of the Cabinet Roy Harris Jenkins.

It will be shown below how these correlations have been semantically represented in the LOD prototype.

rub_code	tfd_code	fnd_annee	fnd_num	dos_num	dos_obj_id	description_level	dos_per_deb	dos_per_fin	mdo_titre	mdo_analyse
11593	INV	2018	1	1	2509062	Dossier	1978	1979	Discours et allocutions de personnalités politiques et de commissaires. – 1978-1979.	Dont Sir D. Ezra, M. Burke, H. Schmidt, M. Thatcher, O. Lambsdorff, B. Ecevit, M. Ohira, D. O'Malley, J. François-Poncet, G. Richardson, E. Davignon, H.D. Genscher, H. Vredeling, C. Tugendhat, F.O. Gundelach.
11593	INV	2018	1	2	2509063	Dossier	1980	1980	Discours et allocutions de personnalités politiques et de commissaires. – 1980.	Dont A. Jørgensen, H. Schmidt, E. Davignon, R. Burke, H. Vredeling, F. O. Gundelach, L. Natali, C. Tugendhat, H. D. Genscher

Table 2 Data from Jenkins dataset

9 Ontologies and Schemata

Adopting the best practices of the linked open data design, the LOD prototype employs an especially designed ontology – the HAS ontology (HASO), mentioned earlier in this document, and reuses several ontologies. They are outlined in the following sections starting with the Semantic Web Technology Stack enabling Linked Open Data description, publishing and deployment.

9.1 Semantic Web Technology Stack

As the LOD prototype implements linked open data architecture, the first models to be presented are the technologies of the Semantic Web technology stack – RDF[37], RDFs[38], OWL[39], SPARQL[40]. These are a series of languages, semantic data exchange format, modeling formalisms and a query language, created by working groups of the World Wide Web Consortium (W3C)[41], and based on the vision of Sir Tim Berners Lee in the beginning of 21st century, they allow the modeling and implementing of linked open data. Figure 8 shows the visual representation of the layered structure of the Semantic Web technology stack. The modeling formalisms of this technology stack – RDFs and OWL are based on formal logic and enable inference according to the principles of formal logic rules for automatic derivation of conclusions, e.g. reasoning or inference.

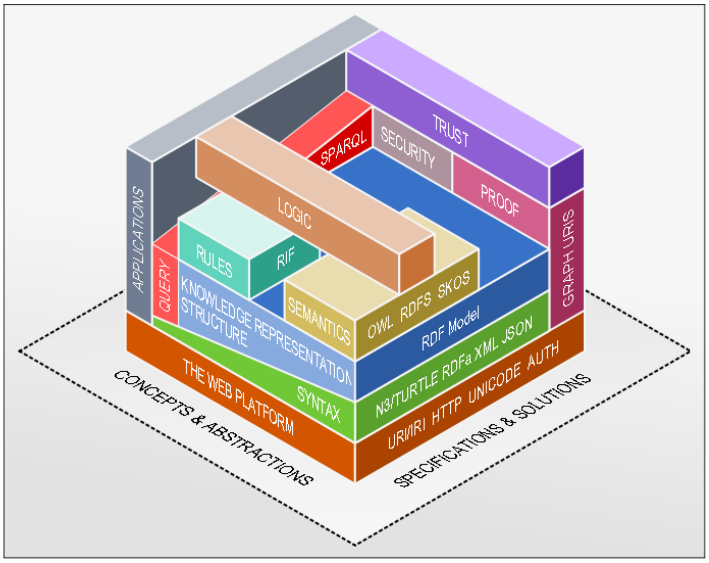


Figure 9 Semantic Web Technology Stack

[37] http://www.w3.org/1999/02/22-rdf-syntax-ns#

[38] http://www.w3.org/2000/01/rdf-schema#

[39] http://www.w3.org/2002/07/owl#

[40] https://www.w3.org/TR/rdf-sparql-query/

[41] http://w3c.org

The HAS ontology (HASO), designed and developed for the HAS prototype, is an OWL ontology, allowing the automatic generation of new facts according to formal logic rules.

9.2 XSD schema

The XSD schema[42] that implements the admissible datatypes in RDF is included in the HAS prototype. The XSD schema is published by the World Wide Web Consortium[43] (W3C), specifies how to formally describe the elements in an Extensible Markup Language[44] (XML) document in RDF.

9.3 SKOS

Simple knowledge organization system (SKOS) is a language used to develop specifications and standards to describe knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading system and taxonomies within the framework of the Semantic Web. SKOS allows a description of concepts hierarchically. The Publications Office of the European Commission make extensive use of SKOS to model and publish in open access vocabularies of conceptual objects, e.g. entities, as discussed earlier in the deliverable.

9.4 Dublin Core

The Dublin Core Metadata Element Set[45] is a vocabulary of fifteen properties that is used to describe metadata about documents and authorship of documents, like creator and identifier. Dublin core has an RDF version[46].The relation for identifier (dc:identifier) of Dublin Core has been adopted in the LOD prototype.

9.5 Upper-level ontologies

Upper-level ontologies describes very general concepts that are common across all domains. An important function of upper-level ontologies is to underlie domain-specific ontologies by providing them with a starting point for the formulation of the domain specific definitions. The best practices of ontology design reuse terms from the upper-level ontologies when developing domain models. For example, concepts in the domain ontology are ranked under the terms in the upper ontology, e.g., the upper ontology classes are superclasses or supersets of all the classes in the domain ontologies.

[42] http://www.w3.org/2001/XMLSchema#

[43] https://www.w3.org/

[44] https://www.w3.org/XML/

[45] https://www.dublincore.org/specifications/dublin-core/dces/

[46] http://purl.org/dc/terms/

The LOD prototype has followed these best practices and adopted concepts and relations from several upper-level ontologies – DBPedia and Proton, the former – as the most common, the later – as the most concise upper-level ontology. They are described in the following sections.

9.5.1 DBPedia Ontology

The DBpedia Ontology[47] is a shallow, upper level ontology, covering the information from the infoboxes of Wikipedia[48]. The ontology currently covers 685 classes that form a class hierarchy and are interconnected with 2795 different properties.

9.5.2 Proton

The PROTON[49] (PROTo ONtology) ontology is a lightweight upper-level ontology, serving as a seed for ontology generation (new ontologies constructed by extending PROTON); it has been used for automatic entity recognition and for metadata generation. PROTON ontology contains a class hierarchy of about 500 classes and 150 properties, providing coverage of the general concepts necessary for a wide range of tasks.

LOD prototype makes use of the top layer of PROTON ontology – PROTON TOP[50] that includes all general concepts like Agent, Location, Person, Organization put in use in the LOD prototype.

9.6 Legal ontology

The legal ontology[51] that was adopted for the LOD prototype is produced by CNR in Italy. This ontology models legal concepts and relations. It consists of 68 top classes and 77 properties. Examples of legal concepts are Legal parameter, Legal concept, Legal role, Legal task.

The review of ELI[52] ontology identified that the Legal parameter and the Legal concept from the CNR legal ontology adopted by the LOD prototype have equivalent concepts in ELI, so they were mapped accordingly in the LOD prototype, e.g. Legal concept from CNR ontology, as equivalent to Legal recourse from ELI, and Legal parameter from CNR, as equivalent to Legal expression from ELI. Thus, one can access the legal data in the LOD prototype by using CNR ontology or ELI ontology concepts.

However, it is important to point out that these ontologies do not provide models for legal document parts, such as article, paragraph. Therefore, these concepts have been introduced and defined in the especially crafted for the LOD prototype HAS ontology.

[47] http://dbpedia.org/ontology/

[48] https://en.wikipedia.org/wiki/Main_Page

[49] https://ontotext.com/documents/proton/Proton-Ver3.0B.pdf

[50] http://www.ontotext.com/protontop#

[51] http://www.loa-cnr.it/ontologies/IOLite.owl#

[52] https://op.europa.eu/en/web/eu-vocabularies/model/-/resource/dataset/eli

9.7 ELI (European Legislation Identifier) ontology

The European Legislation Identifier (ELI) is a representation standard to make legislation metadata available online. The ELI ontology[53] provides the model for structuring metadata of legislative resources and publishing them as linked data. It captures the relationships between national and European legislative resources. ELI is based on the conceptual FRBR[54] model (Functional Requirements for Bibliographic Resources) and describes each conceptual level of a legislative resource.

9.8 HAS ontology

The HAS ontology[55] has been especially designed for the LOD prototype. It is a domain model for the archival domain in the specific context of the HAS and reuses concepts and relations from all ontologies and schemata outlined in the previous sections. HAS ontology implements the data model of the LOD prototype described in the next section.

[53] https://eur-lex.europa.eu/eli-register/news_item_23.html

[54] https://www.oclc.org/research/activities/frbr.html

[55] http://ec.europa.eu/isa/ontology#

10 Data model

The archive records contain information related to the content of the Unit of description on the one hand and to the record creation, maintenance and management on the other. The LOD prototype data model covers two facets, c.f. Figure 10:

1. LOD representation of the archives metadata management,

2. LOD representation of the content of the units of description enriched by using controlled vocabularies.

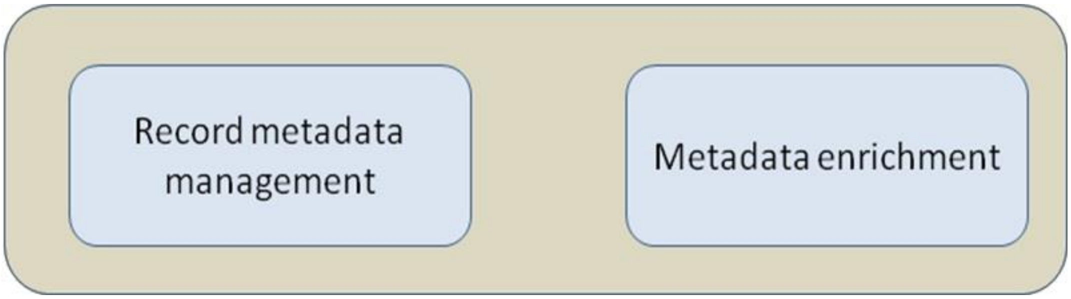


Figure 10 Archives representation with LOD

The analysis of the data samples described in Section “Datasets” above, and of the archives description models, e.g. RIC-CM[25], ISAD(G)[27], ISAAR(CPF)[28] helped to build a conceptual model that re-uses concepts from ISAD(G) and ISAAR standards. The core concept in the HAS ontology is ISAD(G)’s “Unit of description”. The “Unit of description” instantiates different possible ISAD(G)’s “Levels of description”, the concept that gathers all manifestations of archival content organization – Fonds, Series, File or Record Set, Item, but also includes the concepts of Digital Object, Physical object, Record, or Transfer that the HAS data model views as SubClasses of the “Levels of description”. Figure 11 shows all these archival concepts defined as subClasses of the concept “Level of description”.

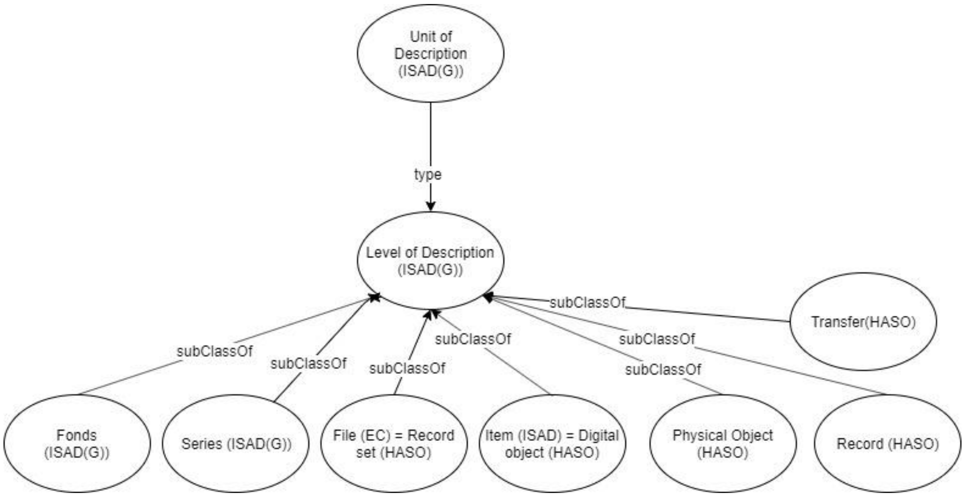


Figure 11 Model of Level of description class hierarchy

Based on the insight from the archives business processes section above, the digital objects are described as being stored in records and the physical objects have their metadata directly attached to the digital object itself. The model on Figure 12 has been designed for the representation between File (Records set), Digital object and Record. Furthermore, the model captures the files or the record sets, that are collections of digital objects, with other digital objects as attachments. They will also capture these sets as physical objects in cases when there is a paper copy of the described digital object. The fact that the metadata of the digital object are described as records is represented by the direct link between the Digital object and the Record – hasRecord. Finally, the fact that the Record set contains the Record if this Record describes a Digital object from the Record set is described by the implicit relation containsRecord, that is automatically generated by a logical implication rule, of the form shown in Example 7 below.

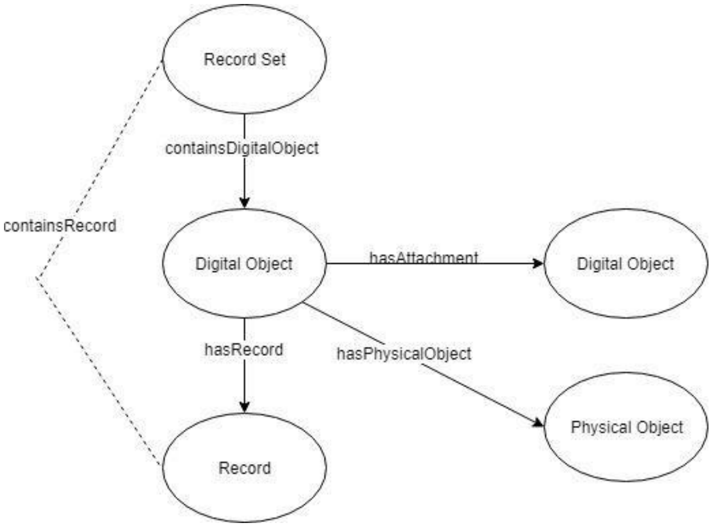


Figure 12 Model of Record set, Digital object, Record

The content of the units of description is represented in the LOD prototype model by relations around the concept of Level of Description, as shown on Figure 13. The model of the Level of description provides information about the metadata within the archives – identifier, title, startDate, endDate, clearance for access, clearance for reproduction, accrual, place of creation, place of holding, medium, but also information about the content of the level of description – title, extent, description, language. To allow for linked data enrichment, the relations “extent”, “title” and “description” have both an attribute that points to plain text and an attribute that points to objects that give access to linked information. The specification of the concepts on Figure 12 takes into consideration the definitions in ISAD(G) and the description of the business concepts identified in the study released in the context of the 1st phase of the ISA²Action 2017.01 (published on July 4, 2019).

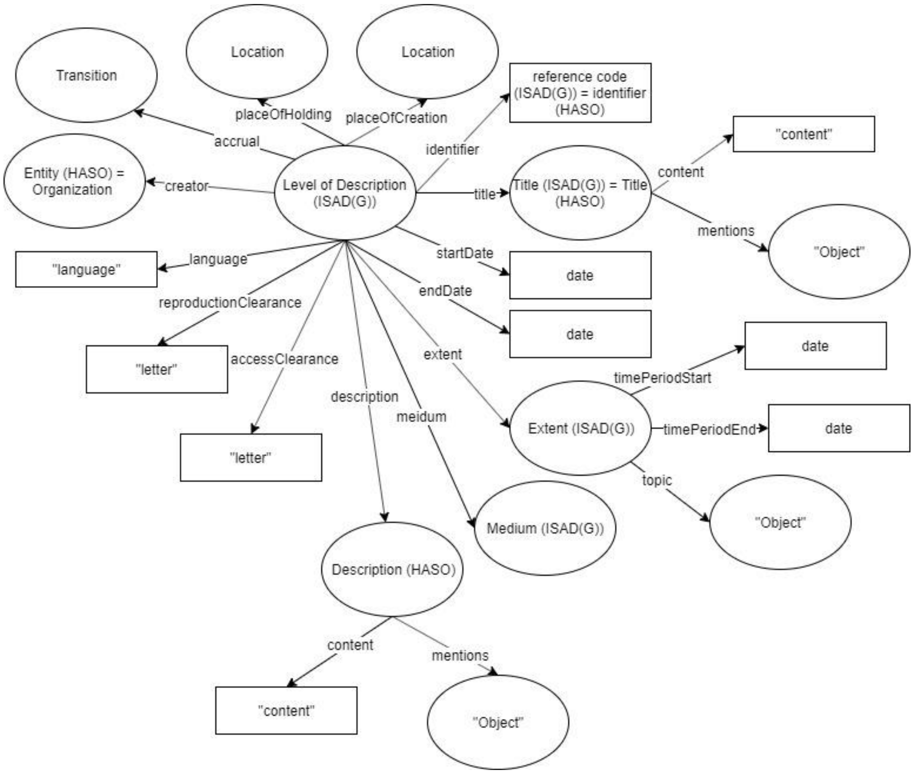


Figure 13 Model of Level of Description

The creators of the Levels of description are Entities, as per the definition of ISAAR(CPF) model and as per the business concepts description of the HAS. The Entity stands for an Agent that can be an Organisation, a Family or a Person. The model on Figure 14 mirrors to a great extent the ISAAR(CPF) model, and covers a variety of aspects needed in order to describe the Entity, like identifier, name, function, location of operation, location of foundation, date of establishment, date of dissolution, address, email, related entities, record.

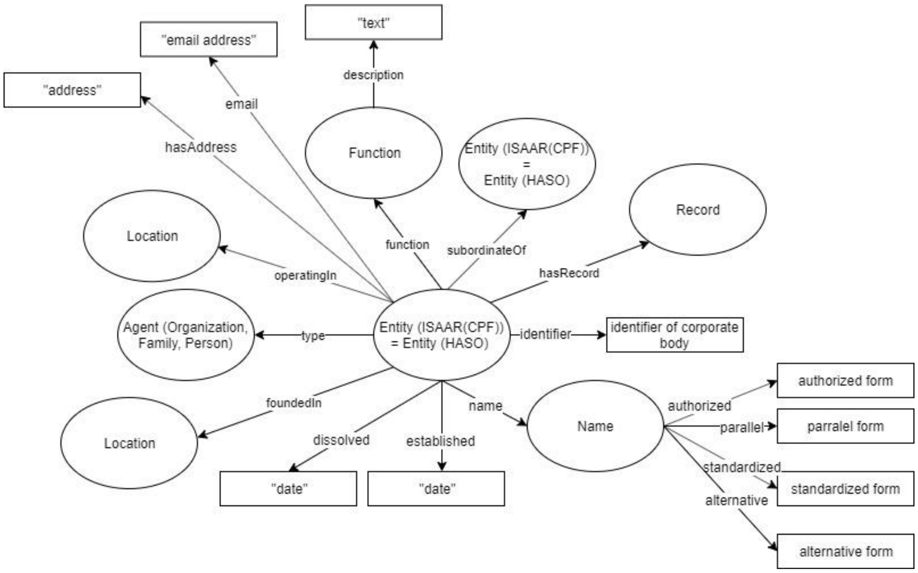


Figure 14 Model of Entity, e.g. organisation

The content descriptions of the Level of Description and the Entity on Figures 12 and 13 make explicit what kind of information will be exposed for querying by the user, and how the queries, listed in section “Use cases” will be solved by navigating the graph of the data model.

People, creators of records are modelled according to PROTON model, e.g. a Person has a Role and a Position at a Department that is part of an Organization (cf. Figure 15).

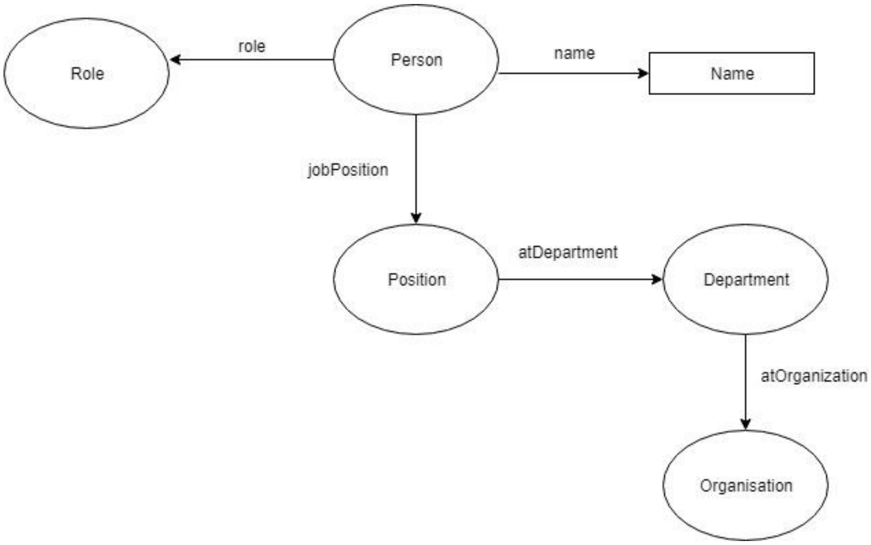


Figure 15 Model of Person

As mentioned earlier the legal references are modelled by reusing the legal ontology of CNR that is partially mapped to ELI. HAS ontology defines the concepts Article and Paragraph as subclasses of the concept Legal Parameter from the legal ontology of CNR. This is shown in Figure 16.

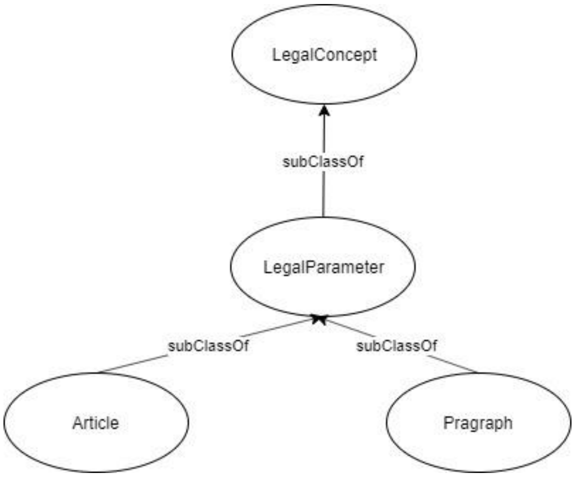


Figure 16 Model of Legal concept

Legal references appear in the Archive datasets included in the prototype as keywords. The HAS model represents legal references semantically with their structure. Figures 17 and 18 show two ways of describing them. In the first model, c.f. Figure 16, Article and Paragraph are linked to the Legal concept that refers to the legal object that the article and the paragraph are from.

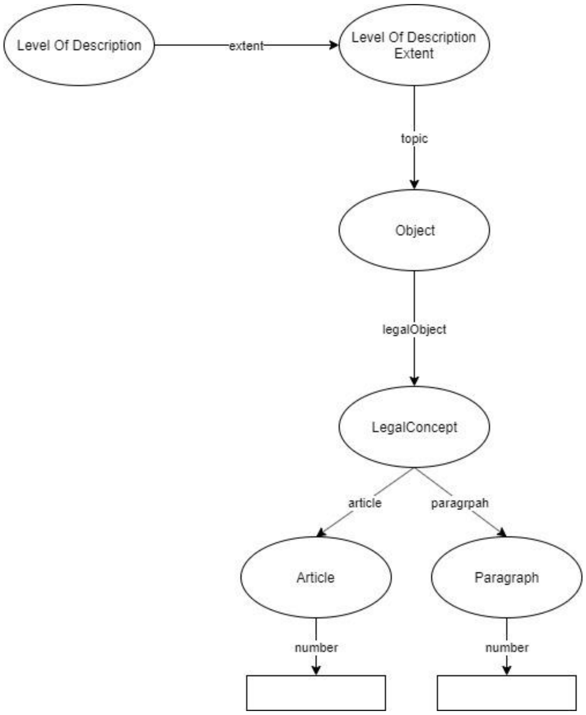


Figure 17 Model of Legal References – version 1

The drawback of this model is that it fails to represent the fact that the article and the paragraph are related. The second model, c.f. figure 18 makes the link between the legal object, the article and the paragraph explicit, by representing the three of them attached to the object representing the topic of the document, referring to the key word, describing the legal reference with a string listing these semantic objects as one whole.

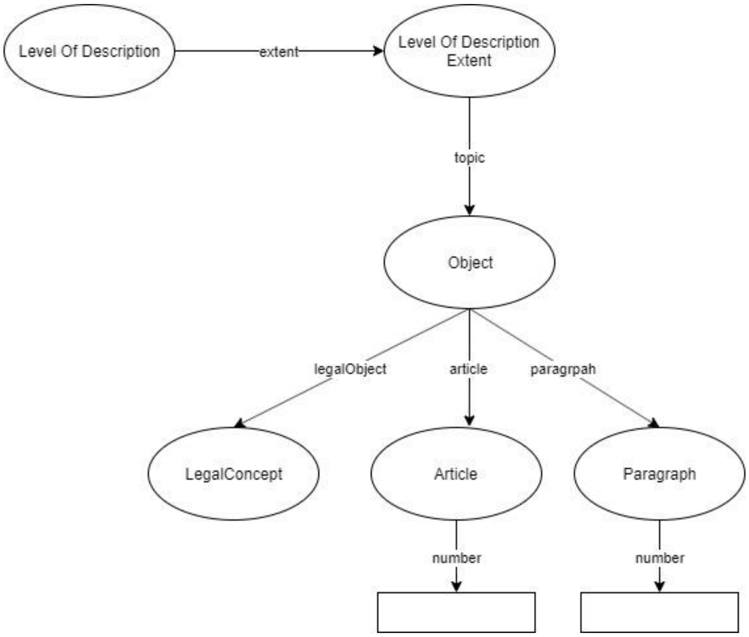


Figure 18 Model of Legal references – version 2

11 HAS Ontology

The HAS ontology can be seen as Archival domain ontology. It is an OWL ontology comprising 26 classes and 44 properties, and reuses concepts and properties from the ontologies and schemas presented in Section Ontologies and Schemata. Figure 18 shows the graph that represents the HAS ontology. As shown in the data model, the most referred to concept in the graph is “Level of Description”. The concept Entity also has many relations that describe it in a more precise way. It implements the data model as shown on Figure 19 below.

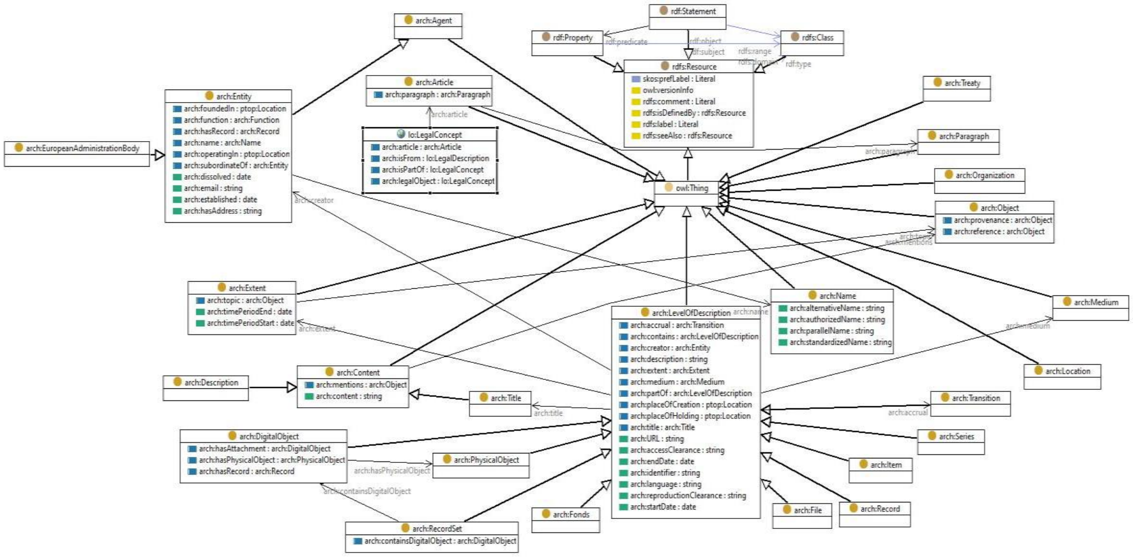

arch:Agent
rdf:Statement
rdf:Property
rdfs:Class
rdf:predicate
rdf:object
rdf:subject
rdfs:range
rdfs:domain
rdf:type
rdfs:Resource
skos:prefLabel : Literal
owl:versionInfo
rdfs:comment : Literal
rdfs:isDefinedBy : rdfs:Resource
rdfs:label : Literal
rdfs:seeAlso : rdfs:Resource
arch:Article
arch:paragraph : arch:Paragraph
arch:article
arch:Entity
arch:foundedIn : ptop:Location
arch:function : arch:Function
arch:hasRecord : arch:Record
arch:name : arch:Name
arch:operatingIn : ptop:Location
arch:subordinateOf : arch:Entity
arch:dissolved : date
arch:email : string
arch:established : date
arch:hasAddress : string
arch:EuropeanAdministrationBody
lo:LegalConcept
arch:article : arch:Article
arch:isFrom : lo:LegalDescription
arch:isPartOf : lo:LegalConcept
arch:legalObject : lo:LegalConcept
arch:creator
arch:Treaty
arch:Paragraph
arch:paragraph
arch:Organization
owl:Thing
arch:Object
arch:provenance : arch:Object
arch:reference : arch:Object
arch:topic
arch:mentions
arch:Extent
arch:topic : arch:Object
arch:timePeriodEnd : date
arch:timePeriodStart : date
arch:extent
arch:Name
arch:alternativeName : string
arch:authorizedName : string
arch:parallelName : string
arch:standardizedName : string
arch:name
arch:Medium
arch:medium
arch:LevelOfDescription
arch:accrual : arch:Transition
arch:contains : arch:LevelOfDescription
arch:creator : arch:Entity
arch:description : string
arch:extent : arch:Extent
arch:medium : arch:Medium
arch:partOf : arch:LevelOfDescription
arch:placeOfCreation : ptop:Location
arch:placeOfHolding : ptop:Location
arch:title : arch:Title
arch:URL : string
arch:accessClearance : string
arch:endDate : date
arch:identifier : string
arch:language : string
arch:reproductionClearance : string
arch:startDate : date
arch:Location
arch:Description
arch:Content
arch:mentions : arch:Object
arch:content : string
arch:Title
arch:title
arch:Transition
arch:accrual
arch:Series
arch:DigitalObject
arch:hasAttachment : arch:DigitalObject
arch:hasPhysicalObject : arch:PhysicalObject
arch:hasRecord : arch:Record
arch:PhysicalObject
arch:hasPhysicalObject
arch:containsDigitalObject
arch:Item
arch:Fonds
arch:File
arch:Record
arch:RecordSet
arch:containsDigitalObject : arch:DigitalObject

Figure 19 HAS Ontology

12 HAS Knowledge Base

The HAS knowledge base is constituted by converting the HAS archive datasets into an RDF graph following the ontology described above in Figure 18. As stated earlier the tables and the keys linking them have been removed and replaced with semantic relations. Thus, in the case of keywords, the key referring to the keyword is present in the table where the keywords are described and also in the file where the fonds items are described. It plays the role of connector between the keyword and the fonds it refers to. In the HAS knowledge base this relation is made explicit semantically by providing the keyword with the object it describes, e.g. country, person, European body, legal reference, and asserting this object to be the topic of the Record set it is attached to via the key in the relational table model, shown and explained on Figure 7. Further, the description of the COM records sets, for example, contains in one cell a general description about how many volumes there are in the given record set, and then the content of all volumes is also described in a single cell. The conversion of these data into RDF follows semantic principles as well. Thus, text mining has been performed inside the cells, describing the volumes (Table 1, column scope_content) to represent semantically each volume of the record set with its characteristics and then each semantically represented volume has been attached to the semantic representation of the Record set it belongs to. Figure 20 shows the graph describing COM 1958(1) records set of the COM dataset, described above. The single volumes are represented as digital objects, e.g. COM19581_DO1 and COM19581_DO2, and described consequently as the language they are in, the title and the date they are produced.

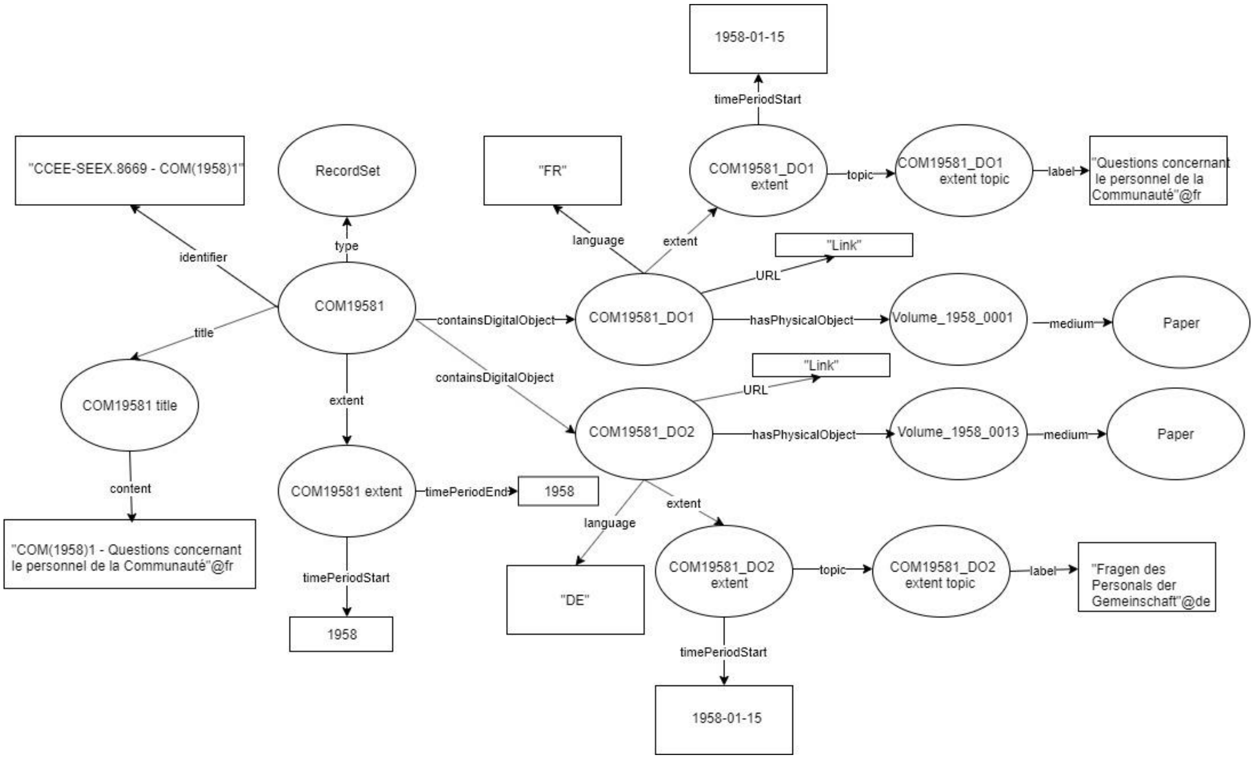


Figure 20 Semantic graph for one example of COM dataset

The graph on Figure 19 shows the instantiation of the model on Figure 11 above describes the relation between record set and digital object. Figure 19 interprets the volume as a physical object of the digital object. However, there is a different way of representing this relation showing that the record set has volumes in paper format on the one hand and digital objects for them on the other. This interpretation is provided on Figure 21.

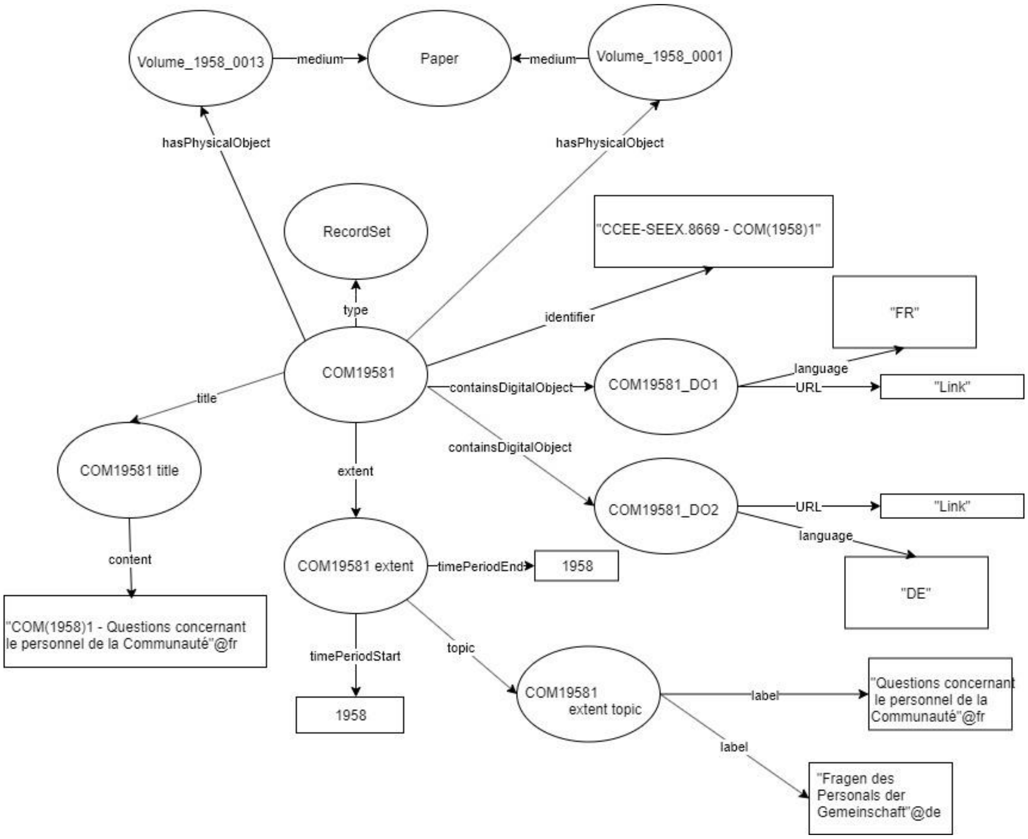


Figure 21 Semantic graph for one example of COM dataset – alternative model

Figure 22 shows a graph of the converted data from the Jenkins dataset. Here too the keys between the relational tables have been removed and captured semantically. Furthermore, the titles have been text mined, and the recognised entities linked and enriched. For example, the topic of Record set 2538968 is the object, describing the person Roy Jenkins. This object will be subsequently enriched with additional information.

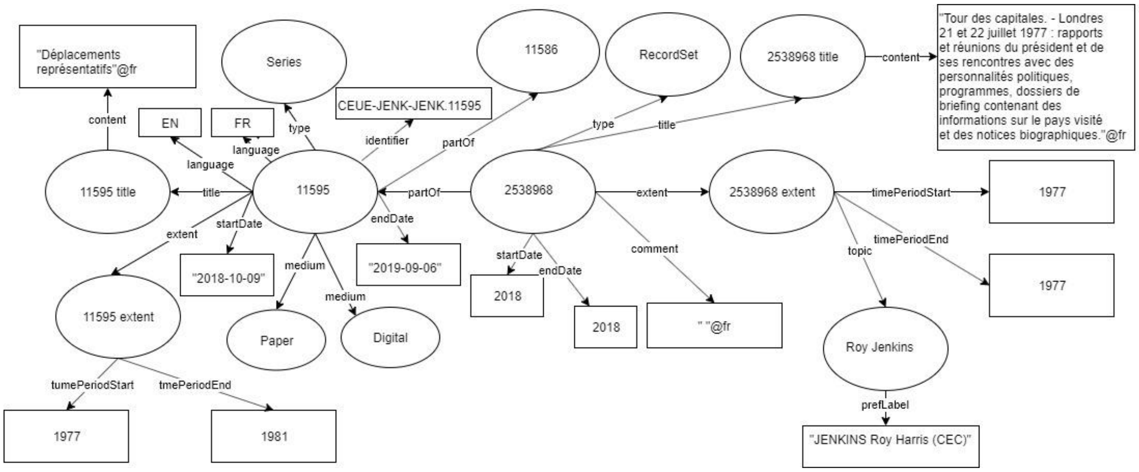


Figure 22 Semantic graph for one example of Jenkins dataset including all dataset files

HAS data model presented two interpretations of legal references. Figures 23 and 24 show these two interpretations. The one on figure 24 seemingly links correctly the article to the legal object on the one side and to the paragraph on the other side, but with this representation the link that the EURATOM Treaty with exactly this article and exactly this paragraph are the topic of the record set 2538978 gets lost.

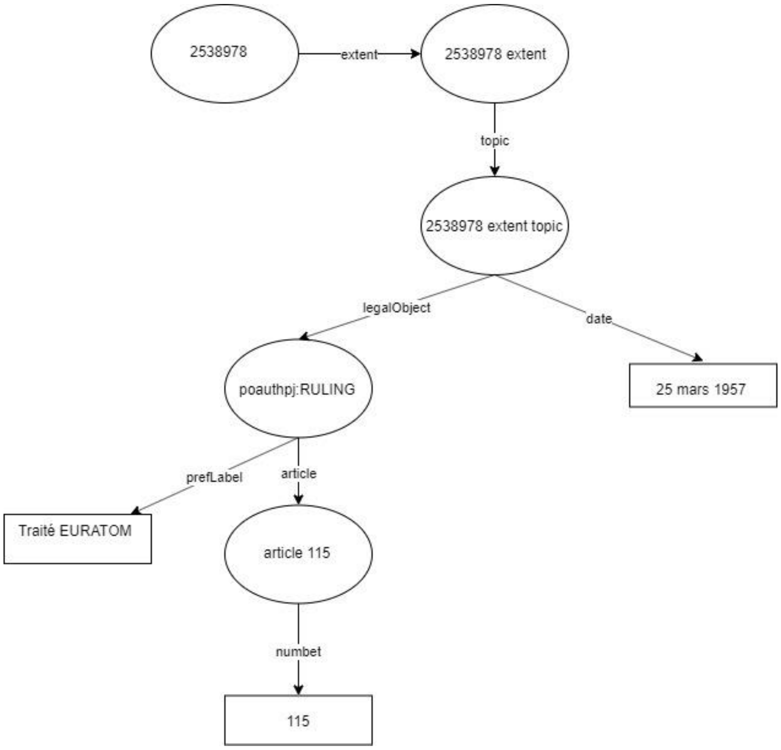


Figure 23 Semantic graph for legal references of Jenkins dataset – version 1

That is why an alternative representation that keeps this information has been adopted to preserve the information that the EURATOM Treaty, article 115 from 25.03.1957 is actually the topic of the record set in question, as shown on Figure 24.

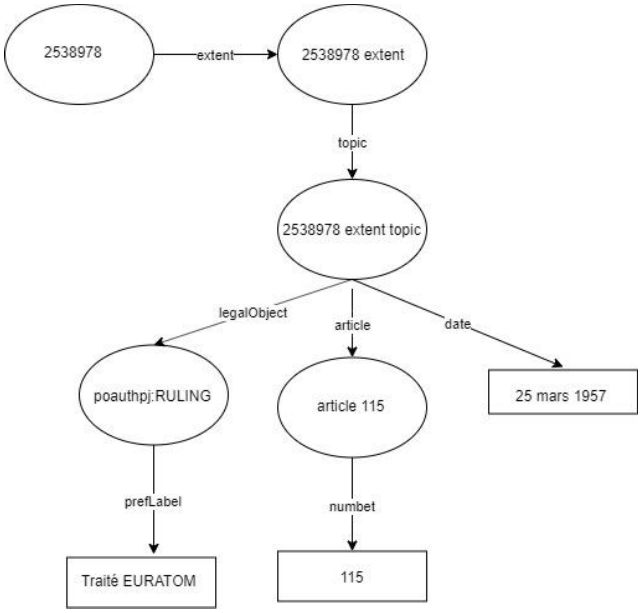


Figure 24 Semantic graph for legal references of Jenkins dataset – version 2

The HAS knowledge base has the following size, c.f. Table 3:

Statements	Number of statements
Total statements	49,019 768
Explicit statements	9,716 038
Implicit statements	39,264 276
Expansion ratio	5.06

Table 3 Size of HAS knowledge base

The explicit statements are the ones generated by the ETL[56] process, i.e. the data conversion procedures, from the COM and Jenkins datasets and physically inserted into the semantic repository. The implicit statements are those generated automatically in the process of loading the explicit data as a result of the application of OWL reasoning, i.e. the inference rules that are elicited from the ontologies via formal logic. Except for the ontologies, inference rules extending OWL reasoning rules have been added, that also generate implicit statements. Example 7 below shows such an inference rule implementing the graph of Figure 12 above.

Example 7

```
Id: containsDO
        a arch:containsDigitalObject b
        b arch:hasRecord c
        --------------------------------
        a arch:containsRecord c
```

[56] Extraction, transformation, loading.

And the ontologies and datasets described in sections Datasets and Ontologies and Schemata above are also inserted into the HAS knowledge base. They are repeated here below with their namespaces[57]:

- owl: http://www.w3.org/2002/07/owl#
- rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
- xsd: http://www.w3.org/2001/XMLSchema#
- rdfs: http://www.w3.org/2000/01/rdf-schema#
- dbo: http://dbpedia.org/ontology/
- dbr: http://dbpedia.org/resource/
- ptop: http://www.ontotext.com/protontop#
- dbp-ont: http://dbpedia.org/ontology/
- dc: http://purl.org/dc/terms/
- arch: http://ec.europa.eu/isa/ontology#
- archr: http://ec.europa.eu/isa/resource#
- poloc: http://publications.europa.eu/resource/authority/place/
- pocoun: http://publications.europa.eu/resource/authority/country/
- eurovoc: http://eurovoc.europa.eu/
- dbpedia: htt://dbpedia.org/resource/
- skos: http://www.w3.org/2008/05/skos-xl#
- pocb: https://publications.europa.eu/en/web/eu-vocabularies/at-concept/-/resource/authority/corporate-body/
- poauthcb: http://publications.europa.eu/resource/authority/corporate-body/
- popers: http://publications.europa.eu/resource/authority/fd_014/
- poauthpj: http://publications.europa.eu/resource/authority/procjur/
- lo: http://www.loa-cnr.it/ontologies/IOLite.owl#

[57] A namespace is a group of related elements that each have a unique name or identifier. https://techterms.com/definition/namespace

13 SPARQL Endpoint

The SPARQL endpoint is implemented in graphDB[58] semantic repository. It is accessible at:

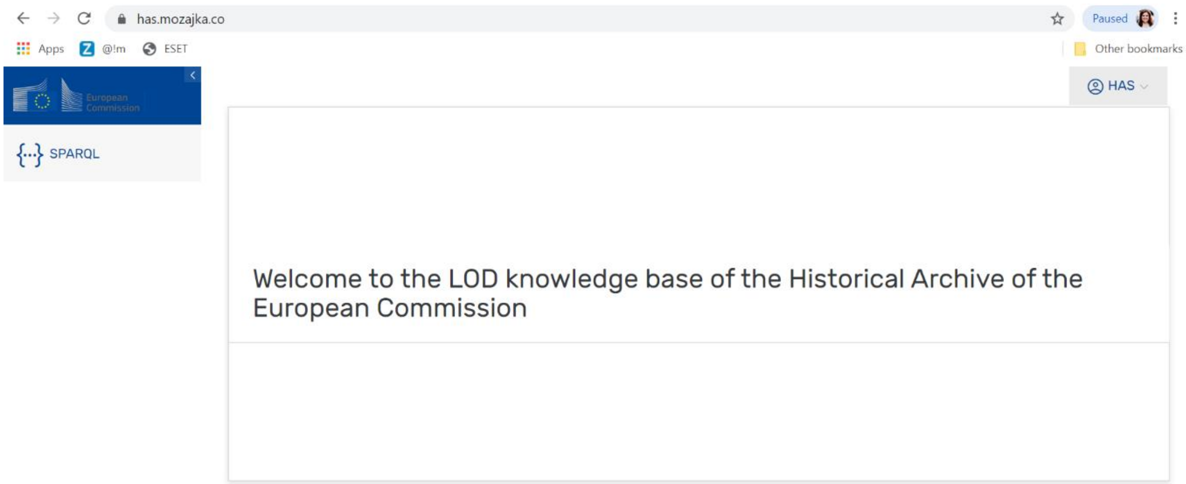
http://has.mozajka.co

Figure 25 shows the screen of loading the SPARQL endpoint.

HAS LOD knowledge base is loading...

Figure 25 SPARQL endpoint loading

Figure 26 shows the initial page of the SPARQL endpoint.



Welcome to the LOD knowledge base of the Historical Archive of the European Commission

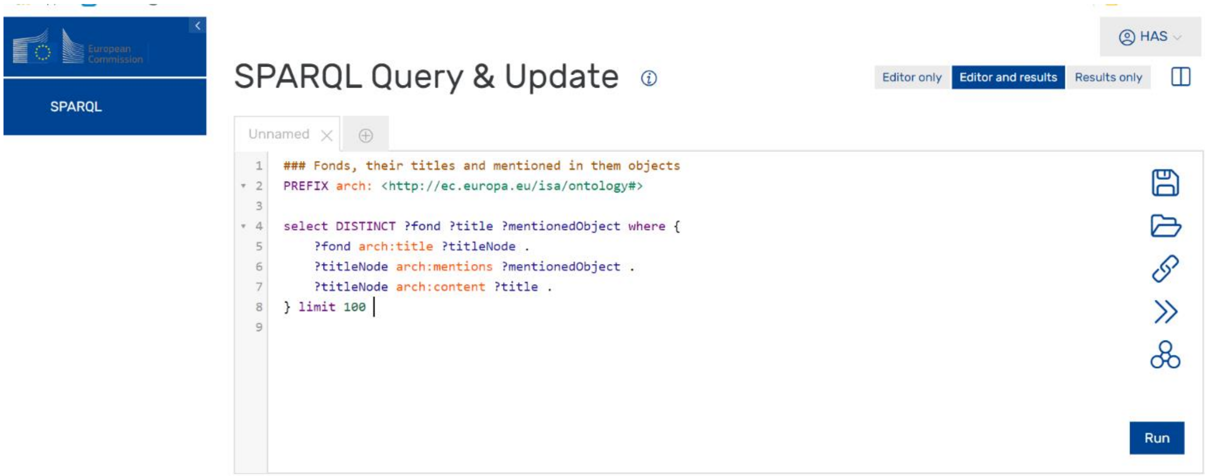
Figure 26 SPARQL endpoint initial page

[58] https://www.ontotext.com/products/graphdb/

The SPARQL query: “Fonds, their titles and mentioned in them objects”

```
PREFIX arch: <http://ec.europa.eu/isa/ontology#>
select DISTINCT ?fond ?title ?mentionedObject where {
?fond arch:title ?titleNode .
            ?titleNode arch:mentions ?mentionedObject .
?titleNode arch:content ?title .
} limit 100
```

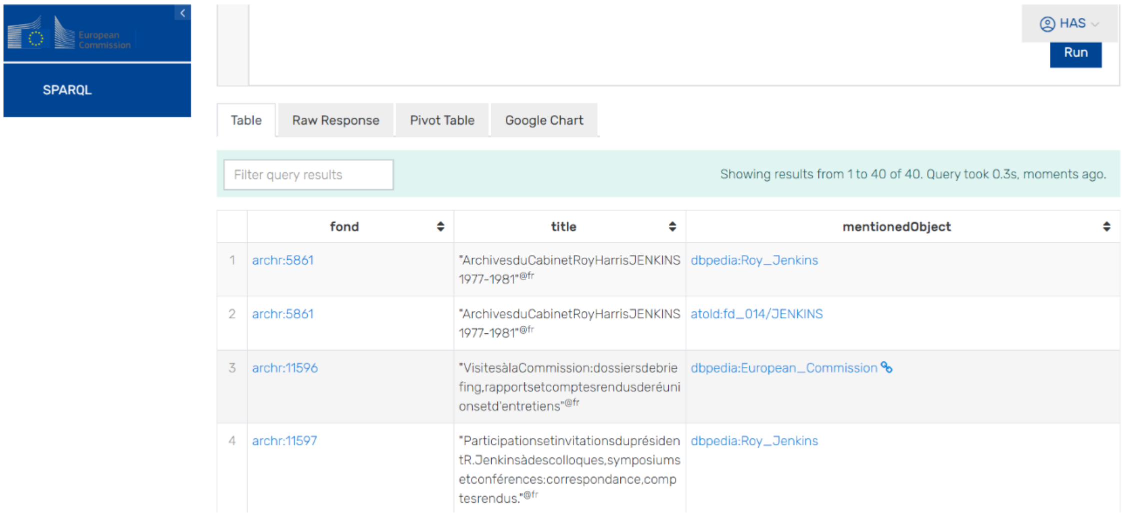
is shown on Figure 27 along with the SPARQL endpoint.



```
### Fonds, their titles and mentioned in them objects
PREFIX arch: <http://ec.europa.eu/isa/ontology#>

select DISTINCT ?fond ?title ?mentionedObject where {
    ?fond arch:title ?titleNode .
    ?titleNode arch:mentions ?mentionedObject .
    ?titleNode arch:content ?title .
} limit 100
```

Figure 27 SPARQL endpoint

The results of this query are shown on Figure 28. Here one notices that the entries under fond and the entries under mentionedObject are links that will lead to further information about the entities they describe. This is shown on figures 28 – 30 in the section Explicit and Implicit Information below.



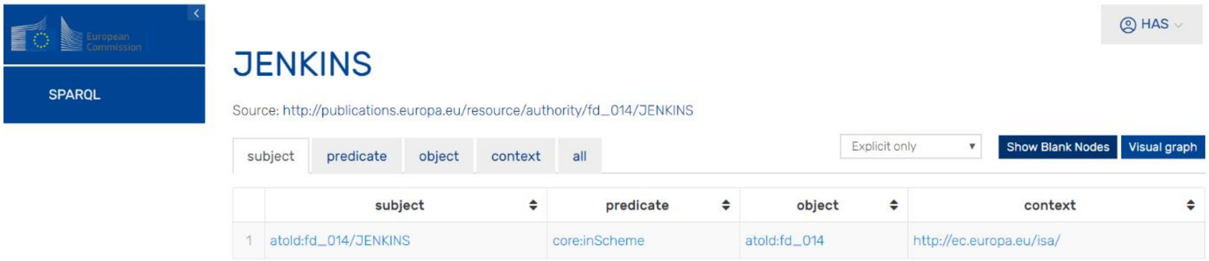
Showing results from 1 to 40 of 40. Query took 0.3s, moments ago.

	fond	title	mentionedObject
1	archr:5861	"ArchivesduCabinetRoyHarrisJENKINS 1977-1981"$^{@fr}$	dbpedia:Roy_Jenkins
2	archr:5861	"ArchivesduCabinetRoyHarrisJENKINS 1977-1981"$^{@fr}$	atold:fd_014/JENKINS
3	archr:11596	"VisitesàlaCommission:dossiersdebriefing,rapportsetcomptesrendusderéunionsetd'entretiens"$^{@fr}$	dbpedia:European_Commission
4	archr:11597	"ParticipationsetinvitationsduprésidentR.Jenkinsàdescolloques,symposiumsetconférences:correspondance,comptesrendus."$^{@fr}$	dbpedia:Roy_Jenkins

Figure 28 Query results for query “fonds, their titles and mentioned Objects in the titles”

13.1 Explicit and Implicit Information

Selecting the second mentionedObject from the results table of Figure 28, we obtain the Publications Office description of Jenkins that is shown in its full form in Figure 29, following the link, e.g. the URI of Jenkins, of the Publications Office.



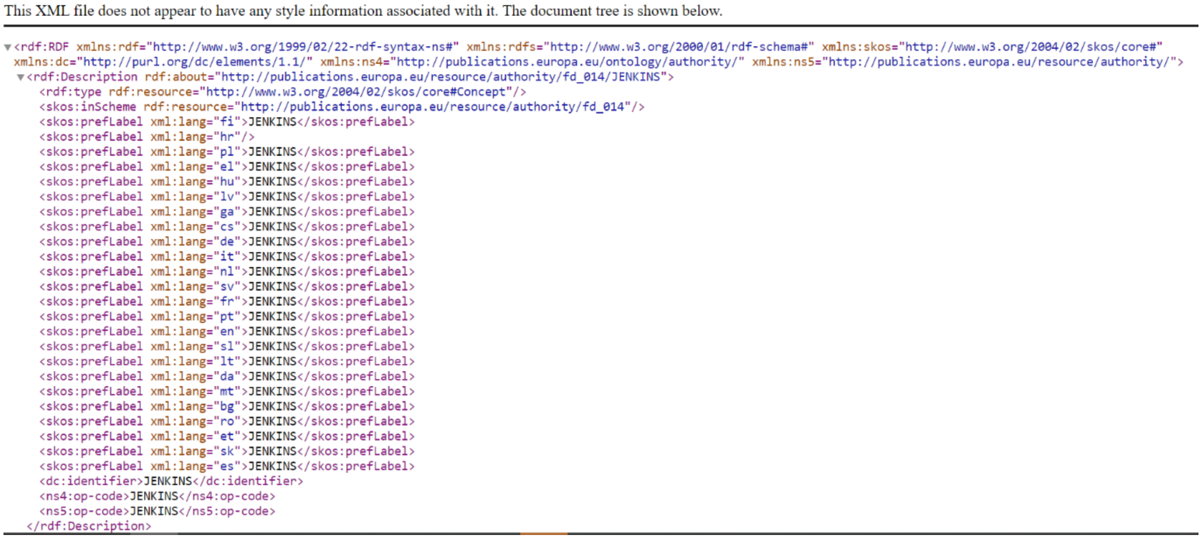
JENKINS

Source: http://publications.europa.eu/resource/authority/fd_014/JENKINS

	subject	predicate	object	context
1	atold:fd_014/JENKINS	core:inScheme	atold:fd_014	http://ec.europa.eu/isa/

Figure 29 Linked entry of Roy Jenkins from Publications Office authority list explicit statement

Figure 30 shows the authority list entry of Jenkins in the Publications Office where it becomes visible what information about Roy Jenkins is available in the Publications Office.

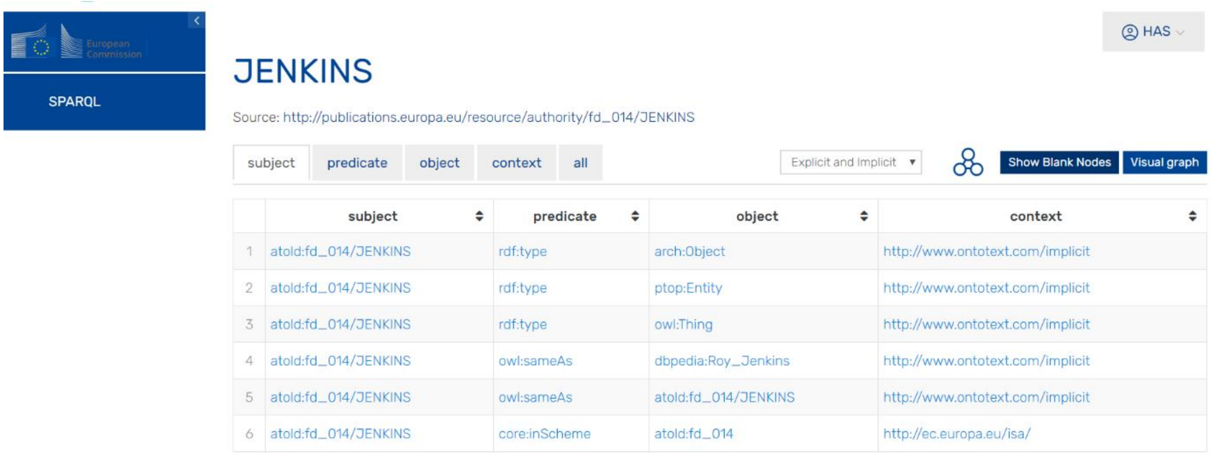


This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xmlns:skos="http://www.w3.org/2004/02/skos/core#"
 xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:ns4="http://publications.europa.eu/ontology/authority/" xmlns:ns5="http://publications.europa.eu/resource/authority/">
  ▼<rdf:Description rdf:about="http://publications.europa.eu/resource/authority/fd_014/JENKINS">
     <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
     <skos:inScheme rdf:resource="http://publications.europa.eu/resource/authority/fd_014"/>
     <skos:prefLabel xml:lang="fi">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="hr"/>
     <skos:prefLabel xml:lang="pl">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="el">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="hu">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="lv">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="ga">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="cs">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="de">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="it">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="nl">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="sv">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="fr">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="pt">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="en">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="sl">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="lt">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="da">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="mt">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="bg">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="ro">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="et">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="sk">JENKINS</skos:prefLabel>
     <skos:prefLabel xml:lang="es">JENKINS</skos:prefLabel>
     <dc:identifier>JENKINS</dc:identifier>
     <ns4:op-code>JENKINS</ns4:op-code>
     <ns5:op-code>JENKINS</ns5:op-code>
   </rdf:Description>
```

Figure 30 Linked data entry of Roy Jenkins from Publications Office authority list

So far, all results about Jenkins for the query mentioned above are explicit facts, i.e. statements inserted into the knowledge base from the converted data. LOD technologies, however, have the property to be able to apply formal logic to generate automatically new facts based on the ontologies and the semantic web technology stack introduced into the knowledge base. Thus, Figure 31 shows additional facts about Roy Jenkins that have been generated automatically. It is interesting to point out that only one result is obtained through the explicit relationship (c.f. figure 28), while six "facts" appear thanks to implicit and automatically generated relationships (cf. figure 31). Moreover, this example highlights the benefit that can be made of the relationships existing between the Publications Office data concerning Roy Jenkins and DBpedia, c.f. line number 4.



JENKINS

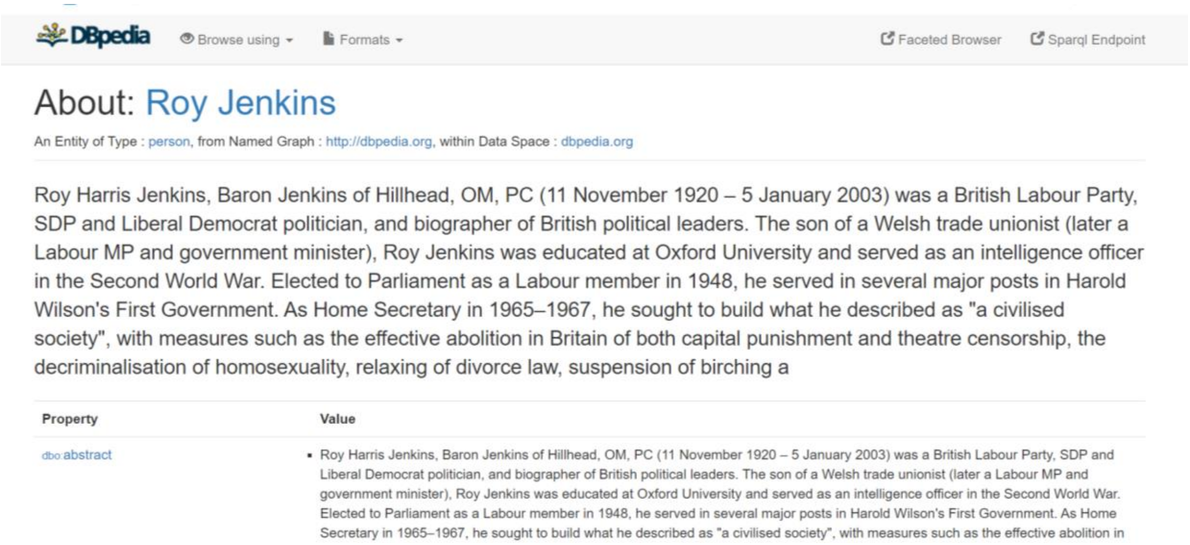
Source: http://publications.europa.eu/resource/authority/fd_014/JENKINS

	subject	predicate	object	context
1	atold:fd_014/JENKINS	rdf:type	arch:Object	http://www.ontotext.com/implicit
2	atold:fd_014/JENKINS	rdf:type	ptop:Entity	http://www.ontotext.com/implicit
3	atold:fd_014/JENKINS	rdf:type	owl:Thing	http://www.ontotext.com/implicit
4	atold:fd_014/JENKINS	owl:sameAs	dbpedia:Roy_Jenkins	http://www.ontotext.com/implicit
5	atold:fd_014/JENKINS	owl:sameAs	atold:fd_014/JENKINS	http://www.ontotext.com/implicit
6	atold:fd_014/JENKINS	core:inScheme	atold:fd_014	http://ec.europa.eu/isa/

Figure 31 Explicit and implicit statements about Roy Jenkins

Linked entry of Roy Jenkins from Publications Office authority list explicit statement and implicit data

This makes possible to obtain more information about Jenkins by referring to its DBpedia entry. Figure 32 below shows the initial description of the DBpedia entry of Jenkins. Thus, the HAS knowledge base entry for Jenkins has been enriched with the information from the Publications Office available about it, and with the information from DBpedia available about it. If there has been a similar to DBPedia knowledge base of data from the European Commission and the HAS, the enrichment of Roy Jenkins entry would have been enriched with information from the European Commission as well.



About: Roy Jenkins

An Entity of Type : person, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Roy Harris Jenkins, Baron Jenkins of Hillhead, OM, PC (11 November 1920 – 5 January 2003) was a British Labour Party, SDP and Liberal Democrat politician, and biographer of British political leaders. The son of a Welsh trade unionist (later a Labour MP and government minister), Roy Jenkins was educated at Oxford University and served as an intelligence officer in the Second World War. Elected to Parliament as a Labour member in 1948, he served in several major posts in Harold Wilson's First Government. As Home Secretary in 1965–1967, he sought to build what he described as "a civilised society", with measures such as the effective abolition in Britain of both capital punishment and theatre censorship, the decriminalisation of homosexuality, relaxing of divorce law, suspension of birching a

Property	Value
dbo:abstract	• Roy Harris Jenkins, Baron Jenkins of Hillhead, OM, PC (11 November 1920 – 5 January 2003) was a British Labour Party, SDP and Liberal Democrat politician, and biographer of British political leaders. The son of a Welsh trade unionist (later a Labour MP and government minister), Roy Jenkins was educated at Oxford University and served as an intelligence officer in the Second World War. Elected to Parliament as a Labour member in 1948, he served in several major posts in Harold Wilson's First Government. As Home Secretary in 1965–1967, he sought to build what he described as "a civilised society", with measures such as the effective abolition in

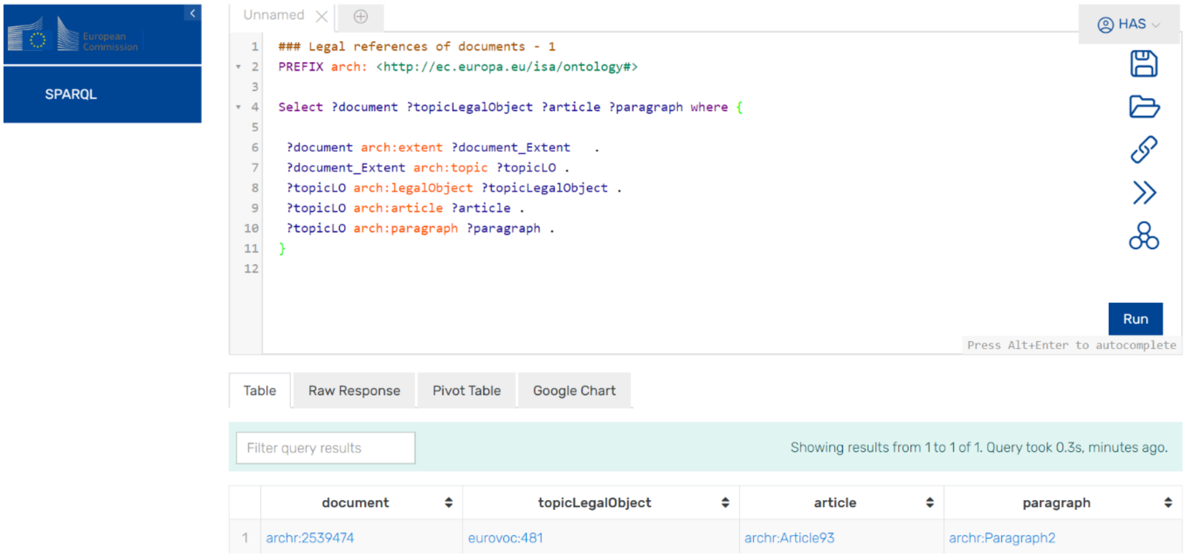
Figure 32 Metadata enrichment about Roy Jenkins

Linked data entry of Roy Jenkins from DBPedia, accessed via HAS Prototype SPARQL endpoint

Legal references have been discussed earlier in sections Data model and Knowledge base. Here is a SPARQL query, exemplifying the model and the data in action. The query "Legal references with articles and paragraphs of record sets"

```
PREFIX arch: <http://ec.europa.eu/isa/ontology#>
Select ?document ?topicLegalObject ?article ?paragraph where {
 ?document arch:extent ?document_Extent .
 ?document_Extent arch:topic ?topicLO .
 ?topicLO arch:legalObject ?topicLegalObject .
 ?topicLO arch:article ?article .
 ?topicLO arch:paragraph ?paragraph .
}
```

returns the result of figure 33 Here the topical legal object, the article and the paragraph are links that will lead the user to further information about these entities.



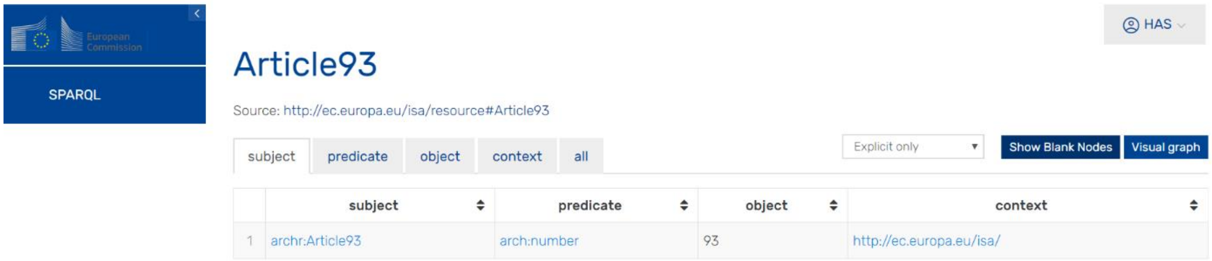
```
### Legal references of documents - 1
PREFIX arch: <http://ec.europa.eu/isa/ontology#>

Select ?document ?topicLegalObject ?article ?paragraph where {

 ?document arch:extent ?document_Extent   .
 ?document_Extent arch:topic ?topicLO .
 ?topicLO arch:legalObject ?topicLegalObject .
 ?topicLO arch:article ?article .
 ?topicLO arch:paragraph ?paragraph .
}
```

Showing results from 1 to 1 of 1. Query took 0.3s, minutes ago.

	document	topicLegalObject	article	paragraph
1	archr:2539474	eurovoc:481	archr:Article93	archr:Paragraph2

Figure 33 Query result with Legal concepts

Figure 34 shows the expansion of the link Article93. The explicit statement shows that it has the number 93.



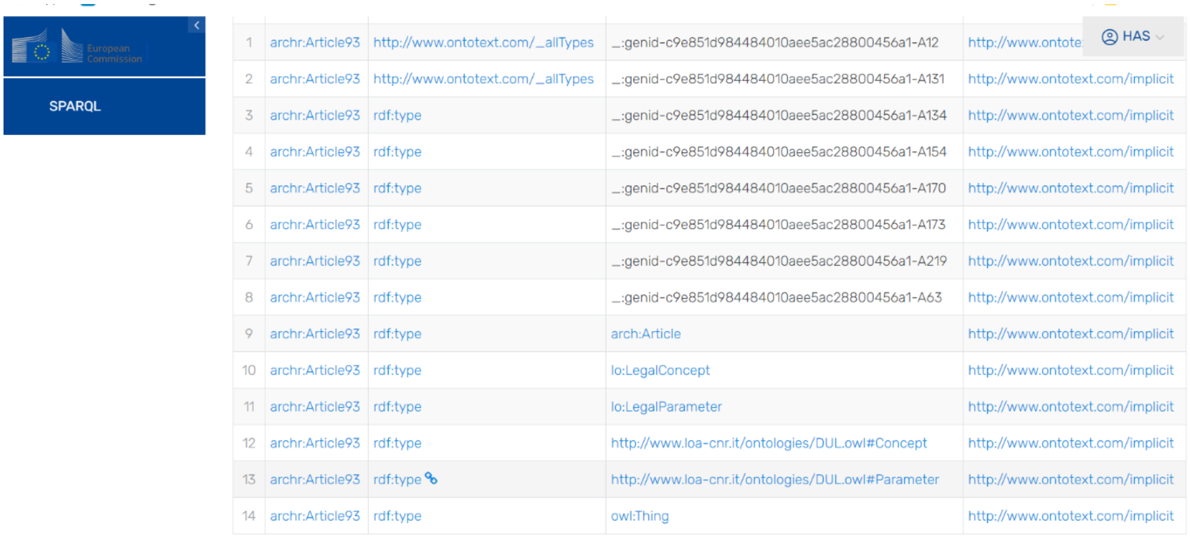
Article93

Source: http://ec.europa.eu/isa/resource#Article93

	subject	predicate	object	context
1	archr:Article93	arch:number	93	http://ec.europa.eu/isa/

Figure 34 Expansion of entity Article 93

The implicit statements about the entry Article93 show that it is an Article, a legal parameter, and a legal concept can be seen in Figure 35.



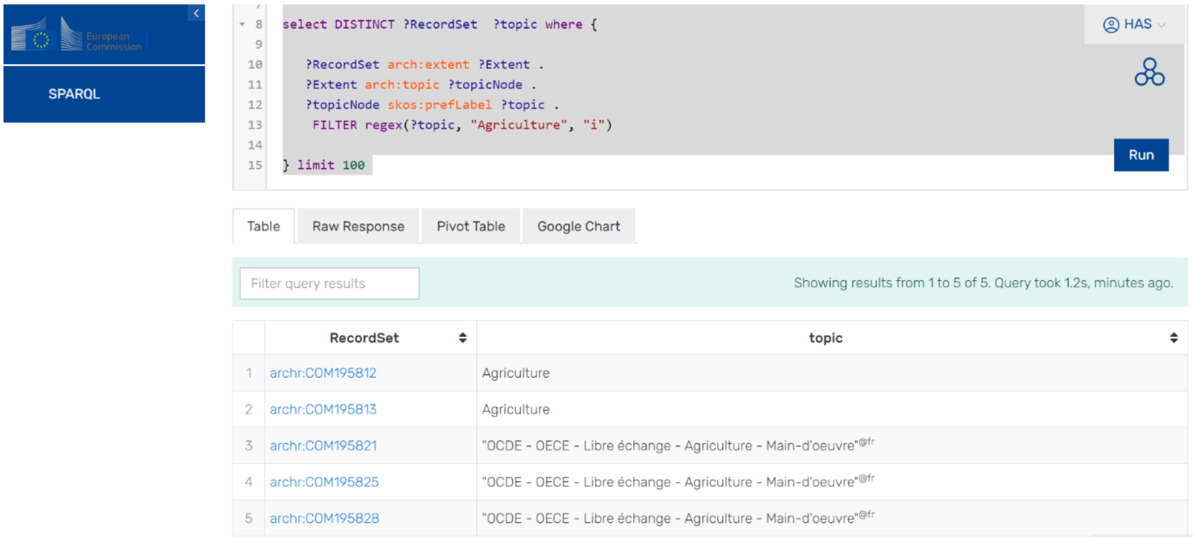
1	archr:Article93	http://www.ontotext.com/_allTypes	_:genid-c9e851d984484010aee5ac28800456a1-A12	http://www.ontote
2	archr:Article93	http://www.ontotext.com/_allTypes	_:genid-c9e851d984484010aee5ac28800456a1-A131	http://www.ontotext.com/implicit
3	archr:Article93	rdf:type	_:genid-c9e851d984484010aee5ac28800456a1-A134	http://www.ontotext.com/implicit
4	archr:Article93	rdf:type	_:genid-c9e851d984484010aee5ac28800456a1-A154	http://www.ontotext.com/implicit
5	archr:Article93	rdf:type	_:genid-c9e851d984484010aee5ac28800456a1-A170	http://www.ontotext.com/implicit
6	archr:Article93	rdf:type	_:genid-c9e851d984484010aee5ac28800456a1-A173	http://www.ontotext.com/implicit
7	archr:Article93	rdf:type	_:genid-c9e851d984484010aee5ac28800456a1-A219	http://www.ontotext.com/implicit
8	archr:Article93	rdf:type	_:genid-c9e851d984484010aee5ac28800456a1-A63	http://www.ontotext.com/implicit
9	archr:Article93	rdf:type	arch:Article	http://www.ontotext.com/implicit
10	archr:Article93	rdf:type	lo:LegalConcept	http://www.ontotext.com/implicit
11	archr:Article93	rdf:type	lo:LegalParameter	http://www.ontotext.com/implicit
12	archr:Article93	rdf:type	http://www.loa-cnr.it/ontologies/DUL.owl#Concept	http://www.ontotext.com/implicit
13	archr:Article93	rdf:type	http://www.loa-cnr.it/ontologies/DUL.owl#Parameter	http://www.ontotext.com/implicit
14	archr:Article93	rdf:type	owl:Thing	http://www.ontotext.com/implicit

Figure 35 Implicit statements about Article 93

The description of the representation of the multiple digital objects of a single record sets is illustrated in the following example. The SPARQL query “Record Sets about Agriculture”, c.f. Figure 36,

```
PREFIX arch: <http://ec.europa.eu/isa/ontology#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX skos: <http://www.w3.org/2008/05/skos-xl#>
select DISTINCT ?RecordSet  ?topic where {
  ?RecordSet arch:extent ?Extent .
  ?Extent arch:topic ?topicNode .
  ?topicNode skos:prefLabel ?topic .
   FILTER regex(?topic, "Agriculture", "i")
} limit 100
```

Present as results several record sets.



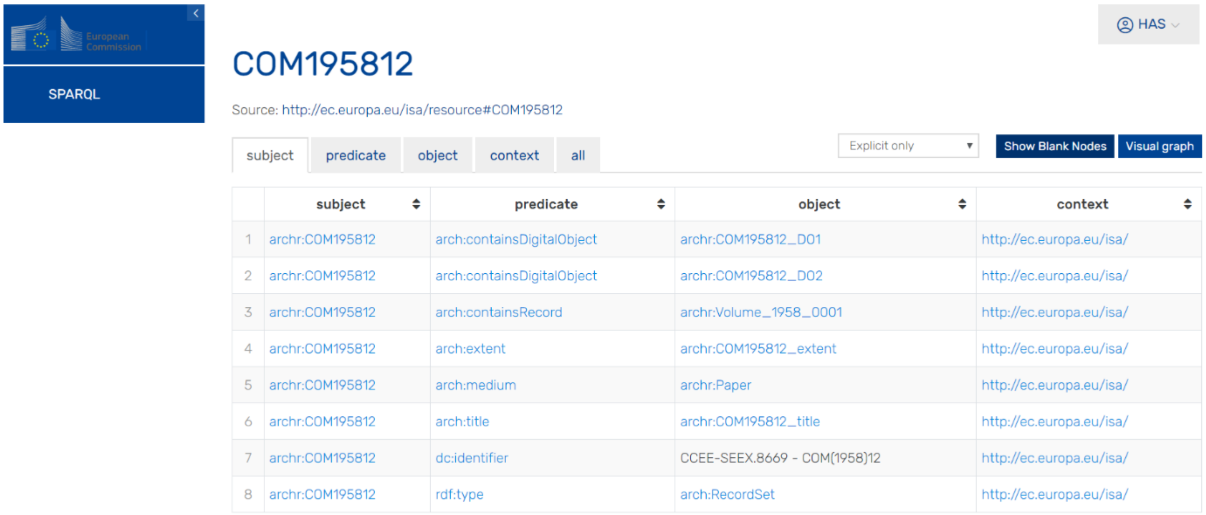
```
select DISTINCT ?RecordSet  ?topic where {

   ?RecordSet arch:extent ?Extent .
   ?Extent arch:topic ?topicNode .
   ?topicNode skos:prefLabel ?topic .
    FILTER regex(?topic, "Agriculture", "i")

} limit 100
```

Showing results from 1 to 5 of 5. Query took 1.2s, minutes ago.

	RecordSet	topic
1	archr:COM195812	Agriculture
2	archr:COM195813	Agriculture
3	archr:COM195821	"OCDE - OECE - Libre échange - Agriculture - Main-d'oeuvre"@fr
4	archr:COM195825	"OCDE - OECE - Libre échange - Agriculture - Main-d'oeuvre"@fr
5	archr:COM195828	"OCDE - OECE - Libre échange - Agriculture - Main-d'oeuvre"@fr

Figure 36 Query result for query about topic "Agriculture"

Selecting the first one, e.g. COM195812, shows two relations containsDigitalObject, e.g. COM195812_DO1 and COM195812_DO2. Figure 37 shows the explicit information available for COM195812_DO1.



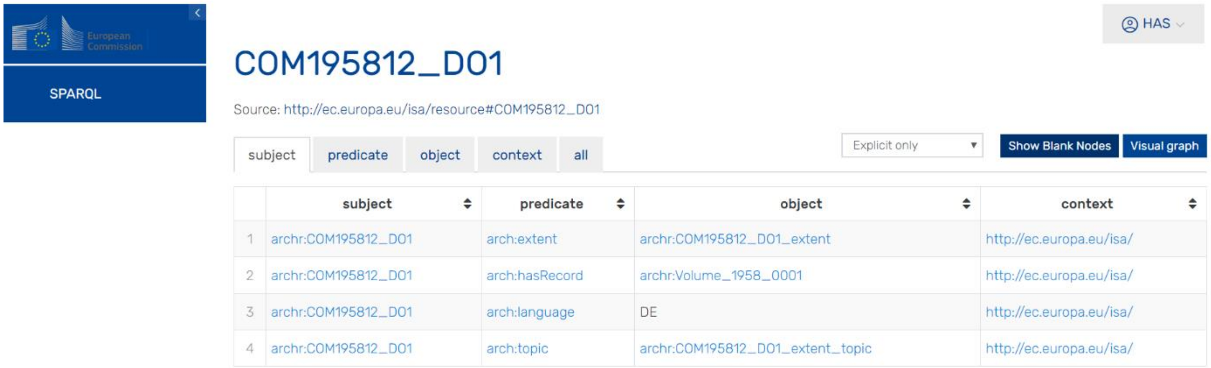
COM195812

Source: http://ec.europa.eu/isa/resource#COM195812

	subject	predicate	object	context
1	archr:COM195812	arch:containsDigitalObject	archr:COM195812_DO1	http://ec.europa.eu/isa/
2	archr:COM195812	arch:containsDigitalObject	archr:COM195812_DO2	http://ec.europa.eu/isa/
3	archr:COM195812	arch:containsRecord	archr:Volume_1958_0001	http://ec.europa.eu/isa/
4	archr:COM195812	arch:extent	archr:COM195812_extent	http://ec.europa.eu/isa/
5	archr:COM195812	arch:medium	archr:Paper	http://ec.europa.eu/isa/
6	archr:COM195812	arch:title	archr:COM195812_title	http://ec.europa.eu/isa/
7	archr:COM195812	dc:identifier	CCEE-SEEX.8669 - COM(1958)12	http://ec.europa.eu/isa/
8	archr:COM195812	rdf:type	arch:RecordSet	http://ec.europa.eu/isa/

Figure 37 Query results about two volumes of one record set

This is that it has Record Volume_1958_0001, extent and topic and is in German.



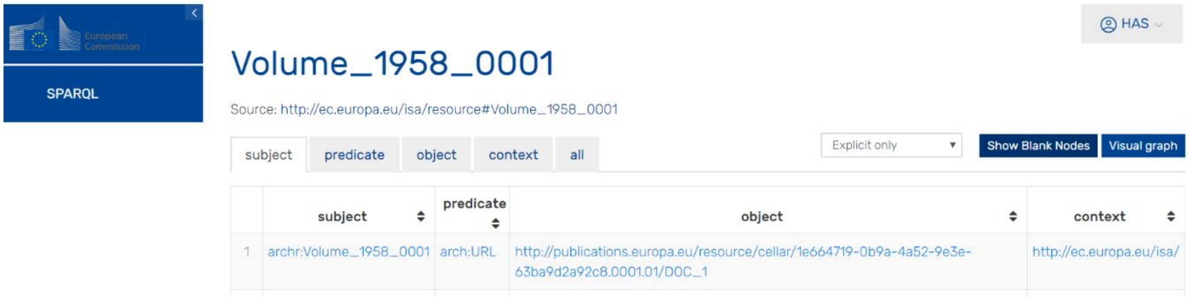
COM195812_D01

Source: http://ec.europa.eu/isa/resource#COM195812_D01

	subject	predicate	object	context
1	archr:COM195812_D01	arch:extent	archr:COM195812_D01_extent	http://ec.europa.eu/isa/
2	archr:COM195812_D01	arch:hasRecord	archr:Volume_1958_0001	http://ec.europa.eu/isa/
3	archr:COM195812_D01	arch:language	DE	http://ec.europa.eu/isa/
4	archr:COM195812_D01	arch:topic	archr:COM195812_D01_extent_topic	http://ec.europa.eu/isa/

Figure 38 Expansion of the entity describing one volume of one record set

If we continue following the links, we will find the link to the PDF file of this volume, c.f. figure 39, and



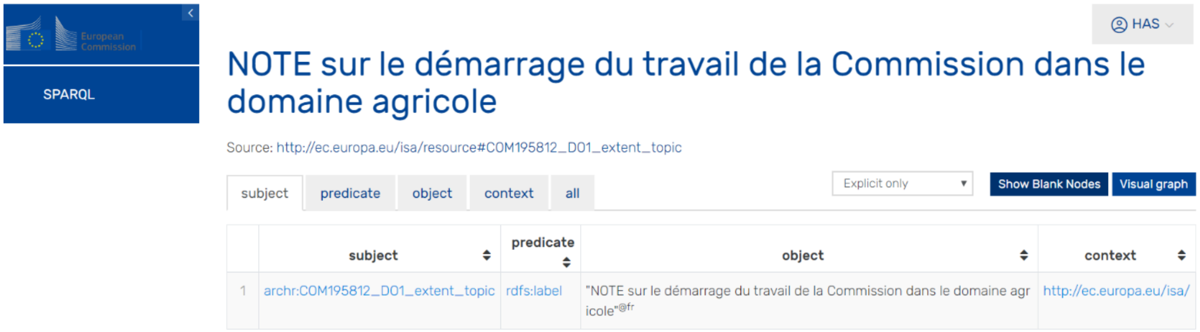
Volume_1958_0001

Source: http://ec.europa.eu/isa/resource#Volume_1958_0001

	subject	predicate	object	context
1	archr:Volume_1958_0001	arch:URL	http://publications.europa.eu/resource/cellar/1e664719-0b9a-4a52-9e3e-63ba9d2a92c8.0001.01/DOC_1	http://ec.europa.eu/isa/

Figure 39 expansion to the link to the text of the volume

The topic of the record set, this time not pointing to an object that can be enriched, but to a plain text description.



NOTE sur le démarrage du travail de la Commission dans le domaine agricole

Source: http://ec.europa.eu/isa/resource#COM195812_D01_extent_topic

	subject	predicate	object	context
1	archr:COM195812_D01_extent_topic	rdfs:label	"NOTE sur le démarrage du travail de la Commission dans le domaine agricole"@fr	http://ec.europa.eu/isa/

Figure 40 Navigation to the topic of one record set

Continuing to query and explore HAS knowledge base it will become clear that some entries are enriched with data from Wikipedia, others are from EuroVoc, third ones from the European Commission or from the Publications Office, cities, countries, corporate bodies assigned to be topics of given Record sets, bringing all the information available about themselves to the user.

13.2 SPARQL queries

This section presents a list of exemplary SPARQL queries that can be formulated and executed on HAS knowledge base. These are:

- Fonds, their titles and mentioned in them objects
- Fonds, created by the EC
- The type of the Fond whose title mentions the High Authority of the European Coal and Steel Community
- The temporal extent and the type of the Fonds with extent starting in 1958
- The French titles of fonds and their date of creation
- Available Record Sets with the Digital Objects and their URLs
- The identifiers and the titles of the record sets
- Record Sets about Agriculture
- What fonds are about legal references
- Entities
- Titles of Fonds and their topics
- The temporal extent of Fonds about Roy Jenkins
- Fonds about “Treaty CEE”

14 AMS integration

The LOD prototype is conceived as an example of a LOD component, being part of a larger Archives management solution that includes a standard AMS and other components, as shown on Figure 3, describing in detail the LOD prototype architecture and connecting to AMS and to VocBench.

The integration of the LOD component, exposing archival data to the general public by means of search capabilities, is represented in Figure 3 – LOD prototype architecture – above. Figure 3 shows how the AMS system connects with the LOD component to exchange data.

This section describes the features of two AMS that allow technical integration for data exchange with the LOD component – Ligeo Archive[59] and Archeevo[60]. Both AMS provide data exchange via OAI-PMH[61] standard connector and support XML format. Both vendors require further developments to fit into HAS data model. However, the conversion rules on the LOD component side can make this task easier and less resource demanding. With this respect both Ligeo Archive and Archeevo have the same capability. The following sections present both AMS in further detail.

14.1 Ligeo

Ligeo is a premier AMS from a French vendor. It has deployed close to 90 archives management applications.

Ligeo software has two components implemented as two portals:

- Ligeo Dissemination (Diffusion) – portal for publishing archival data as a web site.

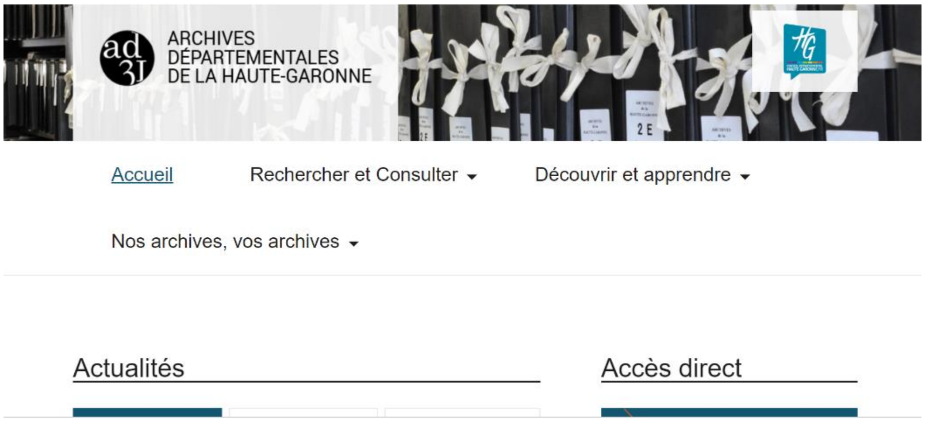


Figure 41 Example of Ligeo dissemination portal

[59] https://www.ligeo-archives.com

[60] https://www.keep.pt/en/produts/archeevo-archival-management-software/

[61] https://www.openarchives.org/pmh/

- Ligeo Management – portal for records management. Several modules are provided to ensure the following functionality: collection and classification of concepts, producers[62], transfer[63], pre-transfer, indexing, search tools, conservation. These functionalities are supplied with graphical user interfaces to allow archivists to insert and manage archival entries in the different stages of their workflow.

Ligeo supports Dublin Core, EAD XML[64], and SKOS. This makes them compatible with VocBench and the Publications Office description standards. It tentatively also fits the data model presented in this deliverable, as the data model is based on the standards ISAD(G) and ISAAR(CPF) that are covered by EAD and EAC XML. If there are concepts and relations that go beyond these standards in the HAS ontological model, this information will be obtained by means of transformation rules or by means of extension of the data standards that Ligeo Archive has been supporting.

Regarding the semantic data, Ligeo does not envisage further development in order to allow to store RDF data in Ligeo storage facilities. This is an argument for the implementation of a HAS solution including a LOD component, where the open for the public data will be exported, transformed, enriched, stored and exposed to the public. This solution will implement the architecture of Figure 3.

14.2 Archeevo

Archeevo is an AMS that aims to support all the functional areas of an archival institution. Amongst its users are the Presidency of the Portuguese Republic, the Navy, the Army, the Armed Forces and several Ministries including the ministries of Internal Administration, Economy, Public Works, Education and Sciences. The Archeevo AMS is designed to support the archivist, whereas the front end, facing the user, uses API to access the data.

The data standards that Archeevo supports are ISAD(G), ISAAR(CPF), ODA[65], EAD, BagIt[66], OAI-PMH. They ensure interoperability with archives aggregation portals such as Archives Portal Europe and Europeana. With respect to data description, Archeevo supports Dublin Core, SKOS, EAD XML. The system has a component that harvests the tagged as public data from the AMS and exposes them for publication by other components. This capability is well suited for the integration between Archeevo and the LOD component, subject of the current deliverable. It allows the implementation of the architecture from Figure 3 above.

Archeevo is prepared to implement capabilities that will allow to store all EuroVoc and necessary Publications Office vocabularies in the AMS itself. This will allow easy integration with VocBench and the vocabularies generated in it. This will provide single access point of the LOD component to all necessary data. All vocabularies will be pulled from the Publications Office and EuroVoc and loaded into the LOD component of the larger archival solution by the AMS. The integration between the AMS and VocBench will enable easy maintenance of the vocabularies. Further, Archeevo is prepared to implement capabilities that will store the URIs[67] of the HAS archive data in the AMS as authority lists. The URIs can be maintained within VocBench,

[62] https://www.ligeo-archives.com/n/producteurs/n:123

[63] https://www.ligeo-archives.com/n/services-versants-et-prearchivage/n:124

[64] https://www.loc.gov/ead/

[65] https://www.springer.com/gp/book/9783642769221

[66] http://www.dcc.ac.uk/resources/external/bagit-library

[67] URI – Uniform resource identifier of RDF.

and then exported either into the AMS or directly into the LOD component. If the URIs are loaded into the AMS as authority lists, they have to be subsequently injected into the Metadata repository of the LOD component of the solution. If the URIs are loaded directly into the LOD component of the solution, the version management of the URIs will be maintained by the LOD component. During the loading of the data into the Metadata repository of the LOD component inference will be performed generating new data and enrichment and consequently will expose the linked open archive data to the public.

15 Conclusion

This document presented the LOD prototype of the PHASE II: Work Package 4 of ISA² Action 2017.01. The scope and the LOD approach have been outlined followed by discussion of archive representation standards, datasets, ontologies and schemas that have been included in the prototype. Thorough description of the HAS data model has been provided, and the especially designed archive domain ontology has been presented. Furthermore, the knowledge base and the prototype has been presented and exemplified in a way to explain, on the one hand, how it works, and on the other hand, to serve as a user guiding tool to showcase and to emphasize the advantages of the linked open data approach for exposing archival information to the general public.

The LOD prototype built on the documented specifications successfully shows the potential of the linked open data technologies for providing archival information to the public. It is a solid starting point for the full scale implementation of LOD in the European Commission, bringing all advantages of the semantic web architectures, search, linking, enrichment, and easy maintenance. The HAS ontology and the HAS knowledge base reuse many available models and resources. Most importantly, the Publications Office datasets and EuroVoc have been put into action by using entities from them for metadata generation and enrichment. The LOD prototype showed how these resources are intertwined within the produced semantic facts. It also showed the advantages of inference, providing more useful information about the users by the application of logical reasoning over the ontologies and over the inference rules. The representation of the archival content is semantic, and hence all keys from the relational tables, where it is currently stored, have been removed. This makes the representation more compact and senseful. The LOD showed the flexibility of the semantic representation to analyse text and structure the information of the text in a way that it becomes queryable and accessible. Further to this, it has been determined that the LOD prototype can easily fit into a larger software architecture including a standard AMS and collaborative vocabulary creation tools, such as VocBench. The AMS providers identified in the previous phases of the ISA2 initiative are also suitable for the successful realisation of an integrated solution, exposing archival content to the general public as linked open data, and organising the internal archives management and vocabularies creation with standard AMS and other tools. Finally, the LOD prototype showed the great variety of navigation paths available through the linked open data that can lead the users to exciting journeys of knowledge discovery.

Adopting a LOD approach in order to expose archival data to the general public brings a number of advantages. Starting from the user experience and the capacity to navigate through archival data, but also the ability to reach contextual information that is outside of the archives is a key element for a fundamentally new way of interacting with archival content. The archival content is better represented when it is semantically represented, as this provides access and the ability to query much more precise and detailed structured information, answering exactly the question asked instead of returning paragraphs of texts. The maintenance of such a LOD model and knowledge base, once in place, are much cheaper to maintain and upgrade, as they do not require reengineering, but just extending the models and adding new facts to the knowledge base. The LOD approach to the HAS general public facing component will also make use of resources of the European Commission, like the Publications Office vocabularies that have not been used so far in HAS. It will also contribute to the augmentation of the vision about these resources of the European Commission by demonstrating how they can be used, what they bring and how they can be extended and improved.

Standard websites displaying archival content show strictly specified limited content as it is the case in the current standard archives portals, except for few adopters of linked open data architecture like The National Archives of the UK[68]. They look rigid, not so interesting, and not providing the way to a wealth of data that can be uncovered with a LOD based solution, as shown in the examples.

LOD technologies have matured and have been becoming mainstream. There are plenty of resources available for reuse and for linking. LOD technologies are a powerful engine that generates a completely different way of experiencing the interaction with data, content and information. This is what the archives are about to provide; access to the documental history of the world to different kinds of users. Linked open data are the perfect instrument to help this to be realised in an optimal manner.

[68] https://www.nationalarchives.gov.uk/

16 Annex 1. HAS data model equivalents

ISAD	ISAAR	ISA business concepts	HASO concepts and properties	RiC-CM
Concepts				
Unit of description			Unit of description	
Level of description			Level of description	RiC-P23 Type
Fonds			Fonds	
Series			Series	
			Record set	RiC-E3 Record set
Item			Digital object	
			Physical object	
			Record	RiC-E1 Record
			Transition	
Title		Title	Title	
Description			Description	
	Entity	Entity	Entity	RiC-E4 Agent
Extent			Extent	RiC-P7 Content extent
Medium			Medium	RiC-P14 Medium
			Description	
	Organization		Organization	
	Family		Family	
	Person		Person	
			Location	RiC-E13 Place
Properties				

ISAD	ISAAR	ISA business concepts	HASO concepts and properties	RiC-CM
			containsDigitalObject	
			hasRecord	
			hasAttachment	
			hasPhysicalObject	
Place of holding			placeOfHolding	
Place of creation			placeOfCreation	
Accrual			accrual	
Creator			creator	
Language			language	RiC-P11 Language information
Reproduction right			reproductionClearance	RiC-P19 Conditions of use
Access right			accessClearance	RiC-P18 Conditions of access
Description			description	RiC-P4 General note
Medium			medium	RiC-P12 Media Type
			Topic	
			timePeriodStart	
			timePeriodEnd	
			startDate	RiC-E12 Date
			endDate	RiC-E12 Date
Title			Title	RiC-P3 Name
			content	RiC-P6 Content Type
Reference code		Identifier	identifier	RiC-P2 Local identifier
			mentions	

ISAD	ISAAR	ISA business concepts	HASO concepts and properties	RiC-CM
	email		Email	RiC-P39 Contact information
	Address		hasAddress	RiC-P39 Contact information
	Location of operation		operatingIn	
	Location of foundation		foundedIn	
	Dissolved on		dissolved	
	Established on		established	
	Name		Name	
	Authorized name		authorized	
	Parallel name		parrallel	
	Standardized name		standardised	
	Alternative name		alternative	
	Subordinate of		subordinateOf	
	Function		function	RiC-E7 Function

Table 4 HAS data model with equivalents from other standards