

Ontotext CH/DH Projects

(A lot more details in [LOD for CH webinar](#), 2016-09, 130 slides)

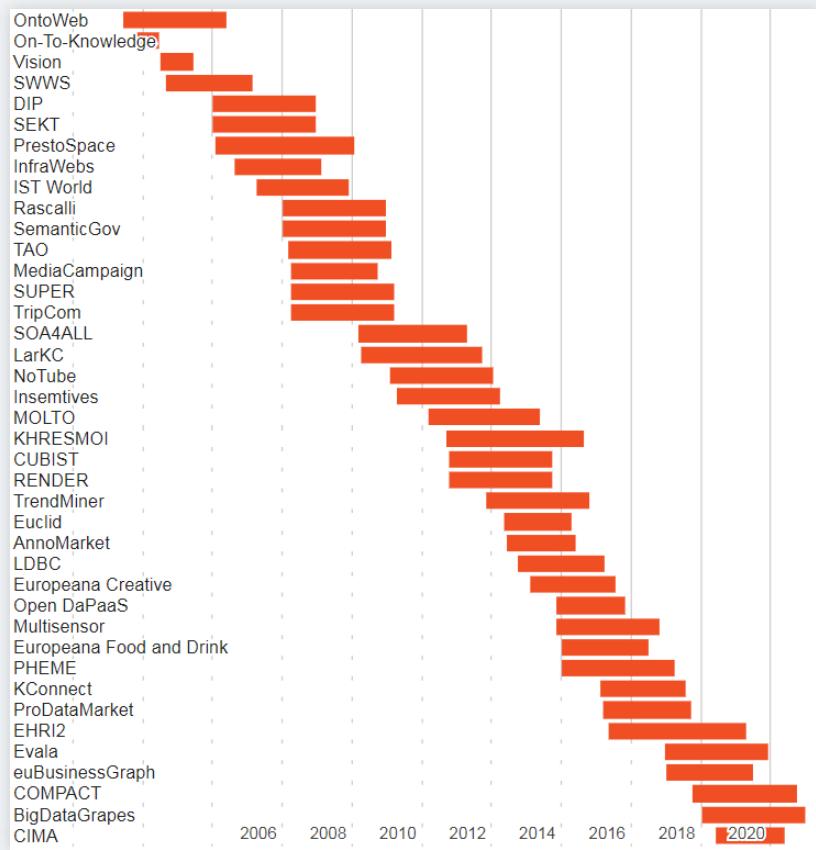
Vladimir Alexiev, PhD, PMP

2018-06-13, CLADA-BG Meeting, Sofia

About Ontotext

- ▶ Founded 2000, part of [Sirma Group](#) (400 people, [BSE:SKK](#), part of SOFIX), venture funding 2008
- ▶ 65 people: 7 PhD, 30 MS, 20 BS, 6 university lecturers. Offices in Sofia, Varna, London
- ▶ Core part of [Sirma Strategy 2022](#) with focus on cognitive computing
- ▶ Working on: semantic technologies, semantic repositories, semantic text analysis, machine learning
- ▶ Semantic Graph Database: [Ontotext GraphDB](#)
- ▶ Semantic data integration and building of Knowledge Graphs
- ▶ Semantic text analysis: entity, concept, relation extraction, document classification
- ▶ Recommendations, sentiment analysis
- ▶ Machine learning: entity disambiguation, deep learning in graphs, etc

Research Projects



Current Projects:

- ▶ EHRI2: European Holocaust Research Infrastructure (H2020 RI): **CH**
- ▶ Evala: Cognitive and Semantic Links Analysis and Media Evaluation Platform (EuroStars)
- ▶ euBusinessGraph: Innovative Data Products and Services for Company Data (H2020 BigData Experimentation)
- ▶ COMPACT: From Research Through Policy on Social Media and Convergence (H2020 CSA)
- ▶ BigDataGrapes: BigData to Enable Global Disruption of the Grapevine-Powered Industries (H2020 BigData Research)
- ▶ CIMA: Company Intelligent Matching and Linking (BG OPC ISIS)

Research and Innovation Awards

Arguably, Ontotext is the most innovative Bulgarian software company.

- ▶ [Innovative Enterprise of the Year 2017](#)
- ▶ [EU Innovation Radar Prize 2016 nomination](#)
- ▶ [BAIT Business Innovation Award 2014](#)
- ▶ [Innovative Enterprise of the Year 2014](#)
- ▶ [Washington Post “Destination Innovation” Competition 2014 Award](#)
- ▶ [Pythagoras Award 2010](#) for most successful company in EU FP6 projects

We have more EU research projects than some universities combined

Industries and Clients

80% of our sales are in the UK and US

- ▶ Media: BBC, UK Press Association, NL Press Association (NDP)...
- ▶ Financial Info: S&P Global Platts, Euromoney, Financial Times, Nikkei...
- ▶ STEM Publishing: IET, Oxford University Press, Wiley, Elsevier, Springer Nature...
- ▶ Life Science: AstraZeneca, Novartis...
- ▶ Government: UK Parliament, The National Archives, Natural Resources Canada...
- ▶ Cultural Heritage and Digital Humanities (see next)

CH/DH Projects

- ▶ ResearchSpace: British Museum, Yale Center for British Art. Largest museum collection, CIDOC CRM, semantic search...
- ▶ (with Sirma Enterprise) [ConservationSpace](#), [Sirma MuseumSpace](#)
- ▶ [Medieval Cultures and Technological Resources](#) (VCMS) COST action
- ▶ Europeana: [Creative](#), [Food and Drink \(sem app\)](#), [OAI PMH](#), [SPARQL](#), [members council](#), 5 work groups, [Data Quality Committee](#)
- ▶ [Bulgariana](#) national aggregator: initiator
- ▶ [Getty Research Institute](#): [vocabularies LOD](#) and helping on Getty Museum LOD
- ▶ [Carnegie Hall LOD](#)
- ▶ [American Art Collaborative](#) consulting: 14 US museums integrating data using CIDOC CRM
- ▶ [European Holocaust Research Infrastructure](#): semantic archive integration. 4+4 years, heading towards ERIC
- ▶ [Canadian Heritage Information Network](#) consulting (national aggregator moving to LOD)
- ▶ Wikidata: frequent contributions (authority control)
- ▶ DBpedia: contributions, association member, data quality and ontology committee
- ▶ CLADA BG: key participant in both CLARIN (NLP) and DARIAH (CH/DH)

Knowledge Graphs

Knowledge Graph	Year	M obj	B triples
British Museum	2013	2	0.92
Polish Digital Library	2013	3.1	0.53
Europeana	2014	20.3	3.8
FactForge	2006-now	~14	3.2
LinkedLifeData	2008-now	~12	10.2
Company Graph	2017-now	6	3
Dun & Bradstreet	2017	210	30

Details about the first 5 are in V.Alexiev et al, [Large-scale Reasoning with a Complex Cultural Heritage Ontology \(CIDOC CRM\)](#), Workshop Practical Experiences with CIDOC CRM and its Extensions (CRMEX), TPDL 2013, slide 17

ResearchSpace

- Semantic integration based on CIDOC CRM, search (first implementation of Fundamental Relations search), data & image annotation, data basket, etc

Find all objects with images created/modified by Rembrandt

and is/has/about drawing and is/has/about mammal +

Search Add To Data Basket Export Print

13 Results 1

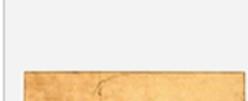
List Thumbnails Timeline

Object Type
1 album
13 drawing

Creator
1 Anonymous
13 Dutch
2 Italian
2 Jan Baptist Weenix
1 Jan Lievens
12 Rembrandt

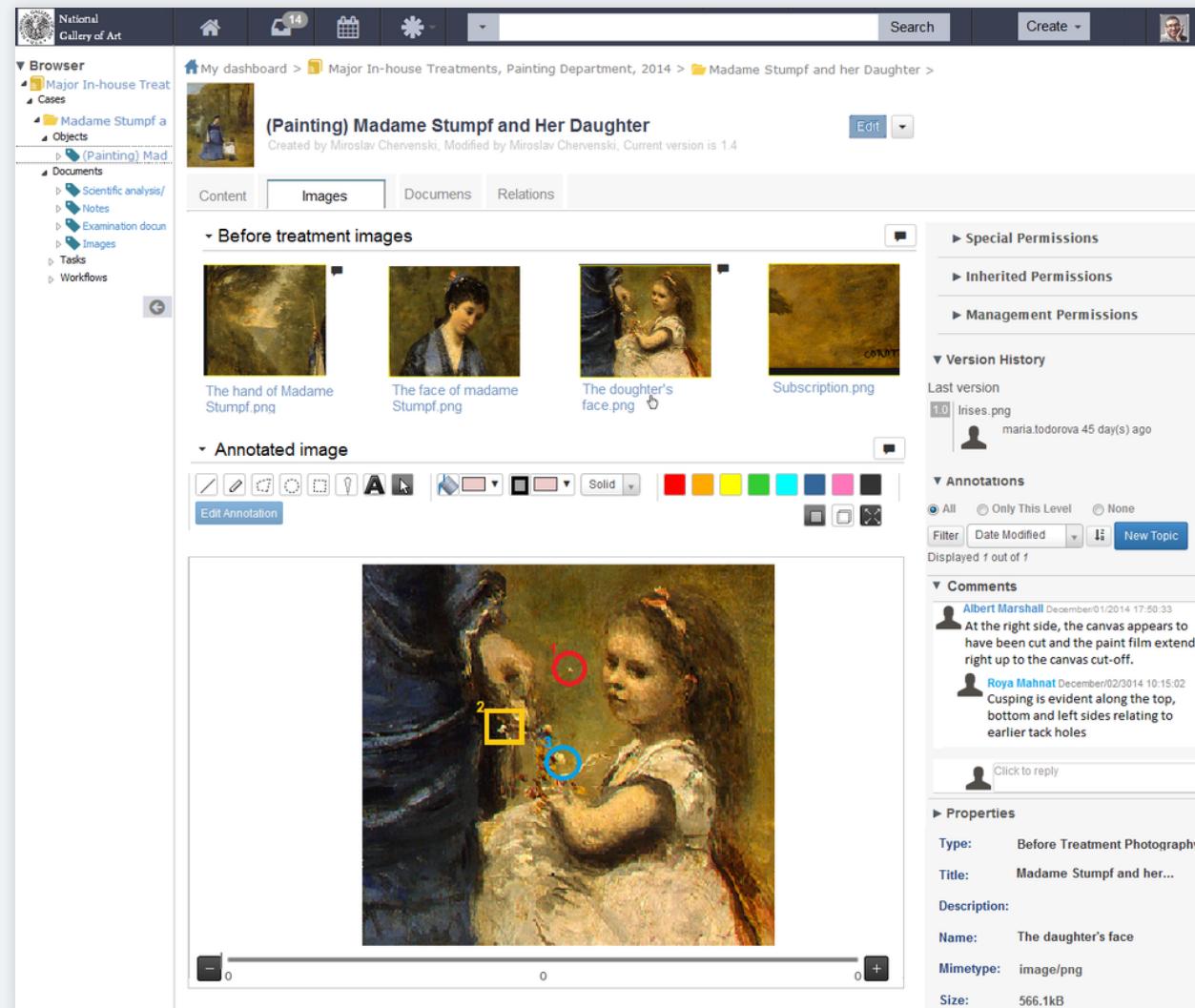
Places
13 (others)

sorted by: Title; then by...

			
PDO13612 A horse lying down; with head to right. ... by Jan Lievens, Anonymous, Dutch, and Rembrandt	PDO13924 Study of a pig, facing left. c.1638-1639... by Dutch and Rembrandt	PDO13925 A tethered pig, facing right. c.1638-1639... by Dutch and Rembrandt	PDO13926 A lion drinking from a pail; crouching on... by Dutch and Rembrandt
			

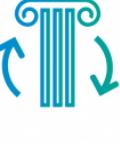
ConservationSpace

- Line-of-business application for conservation specialists. International consortium (US NGA, DK SMK, UK Courtauld etc). Based on the [Sirma Enterprise Platform](#) and [Ontotext GraphDB](#), Ontotext helped with the ontologies.



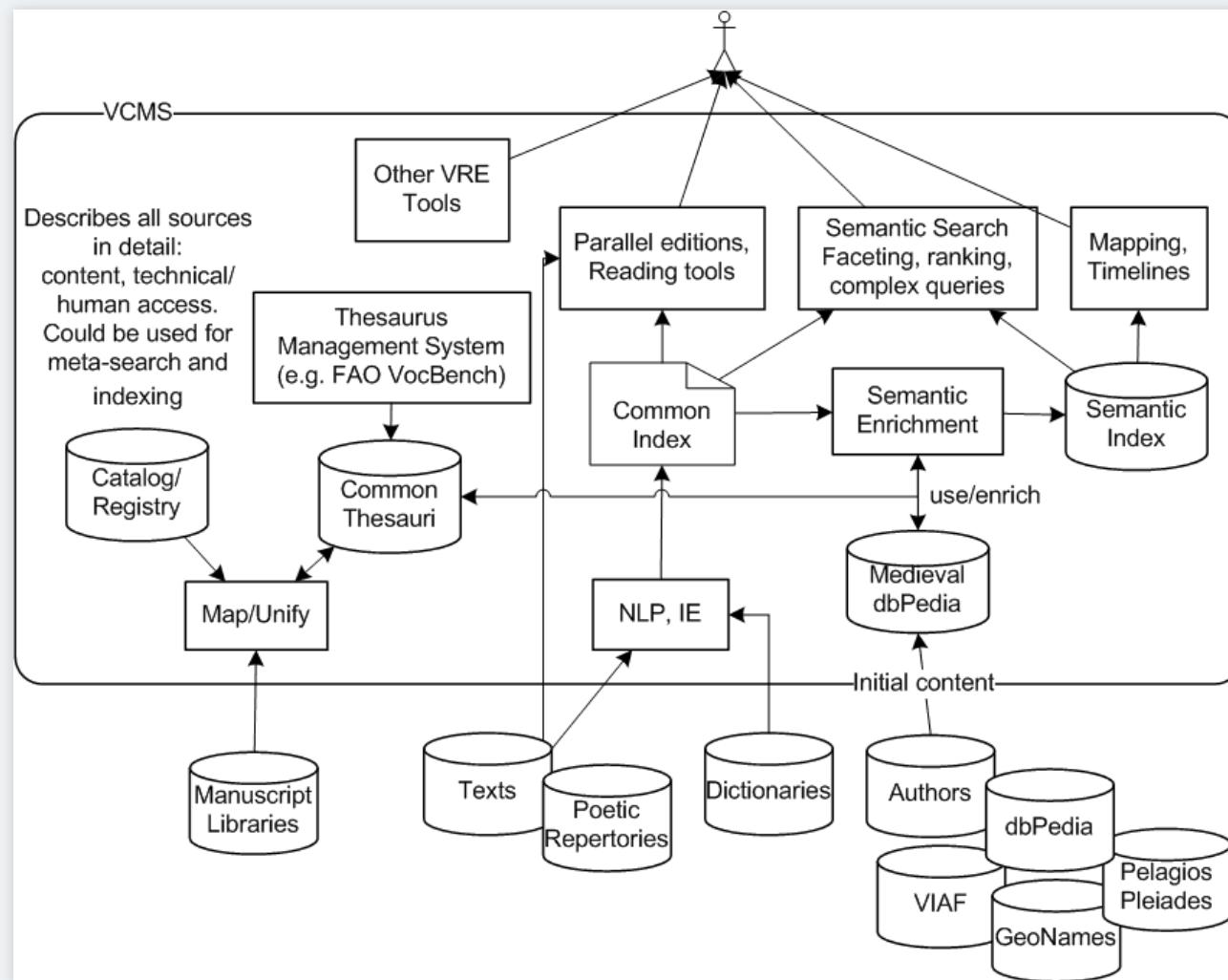
MuseumSpace

► Based on the Sirma Enterprise Platform and ConservationSpace experience. Collections, exhibitions, curation...

			
<p>Conservation Management is the commercially supported and cloud-based version of the open-source Conservation Space project of the National Gallery of Art (USA) with the support of The Andrew W. Mellon Foundation....</p>	<p>Collection Management simplifies your work, such as creating libraries with cultural objects, organized by different criteria. Perfect for small and medium organizations with basic collection management requirements. All cultural objects are...</p>	<p>Image Annotation feature, uses Mirador's image tool. Mirador is an open-source, configurable, extensible, and easy-to-integrate image viewer, which enables image annotation and comparison of images from repositories dispersed around the world....</p>	<p>Exhibition management helps organizations to prepare for an upcoming exhibition. This feature allows you to manage all budgeting, marketing activity, floor plans, contracting and other processes related to exhibitions process. [gallery...</p>
			
<p>Loans Management allows institutions to manage all incoming and outgoing loans regarding all cultural objects within the institution. Incoming/Outgoing Loans Prepare contract agreements or Document special requirements Schedule/Prepare for Log... Arrival/Approval...</p>	<p>Location and Movements feature, allows institutions to monitor the location of each cultural object and the status of every room/building within the facility. Manage Locations Record information for all buildings/rooms,</p>	<p>Acquisition Management allows institutions to track all current, pending, and historic acquisitions of the institution, creating an audit trail of each cultural object that enters the institution. Case for New Acquisition...</p>	<p>Deaccessions Management gives managers a bird's eye view of all deaccessions, past and present to easily manage and maintain records of deaccessions for the institution. Donor Records Log information relevant to...</p>

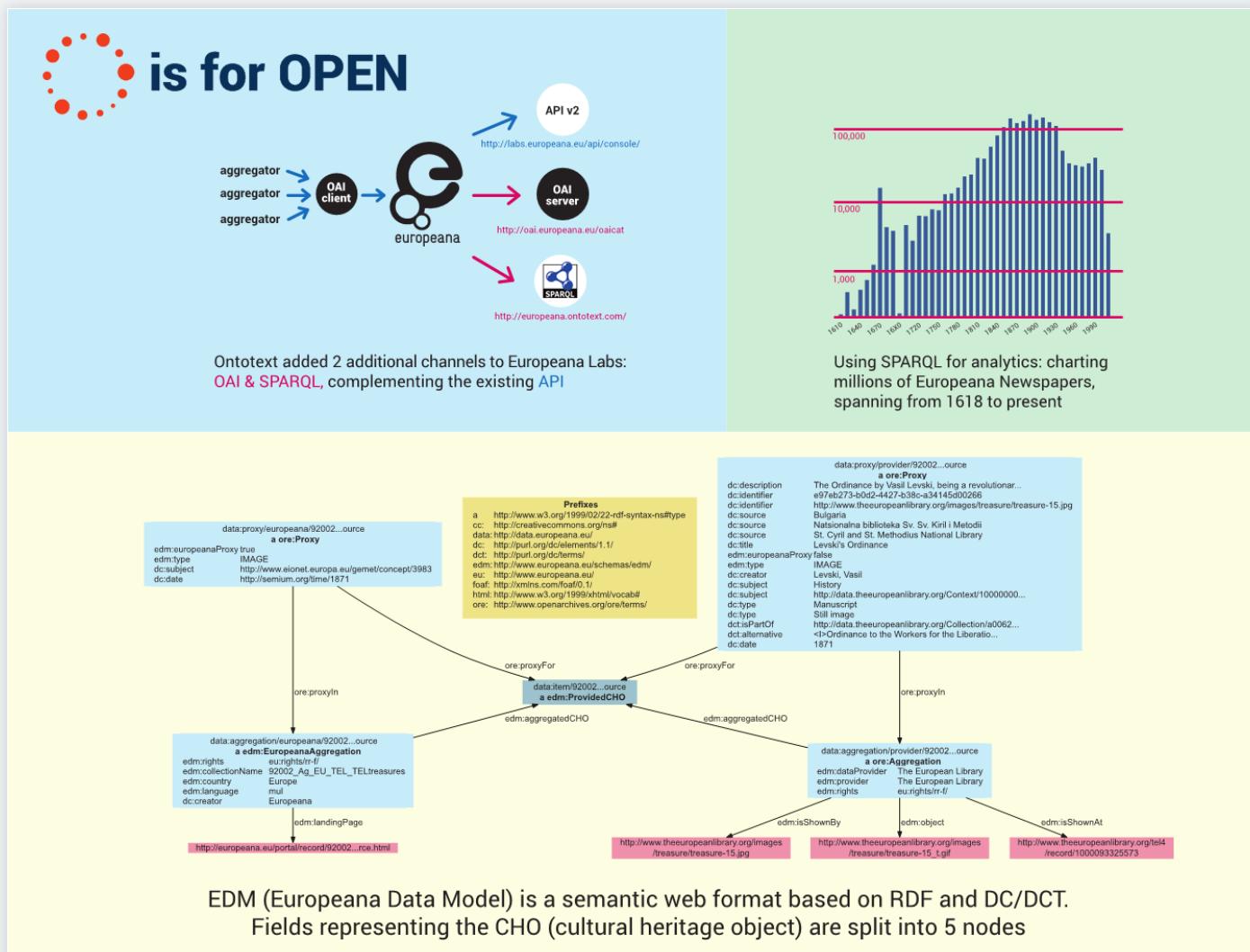
Virtual Center for Medieval Studies

- [Medieval Cultures and Technological Resources](#) (VCMS) COST action. FET proposals for medieval lexicography, historic research, Virtual Research Environments



Europeana Creative

OAI PMH, and SPARQL servers for Europeana (part of Europeana Labs).



Europeana Food and Drink

► sem app: enrichment (EN, FR, manual BG), hierarchical semantic facets, contributed BG recipes



Europeana Food & Drink

The Semantic Demonstrator shows the use of semantic technologies for classification and discovery of Europeana objects related to Food and Drink. Detailed description, data, SPARQL endpoint.

Selected filters: FD: Beer ✕ Data provider: Bulgariana ✕

Food and Drink

- + Agriculture 148
- + Beverages 149
- + Cuisine 149
- + Eating behaviors 149
- + Food and drink by country 108
- + Food and drink preparation 149
- + Food and drink terminology 138
- + Food culture 149
- + Food decorations 30
- + Food industry 148
- + Food politics 143

Places

- + Europe 149

Type (resource)

Language

Data provider

- Bulgariana 149

Results per page: 24 ▾

Results 49 - 72 of 149

◀ Page 3 of 7 ▶



Бирени палачинки
Разбийте яйцата, прибавете постепенно брашното и бирата до получаване на гладка смес. Продължете да чупите и добавете



Панирано пиле с бира
Пилето се сварява предварително. Приготвя се специална паста за паниране. За целта се отделят белтъците от



Пилешки сърца с бира на фурна
Смесете бирата, меда, пресования чесън, зехтина, малко сол и черен пипер на вкус и розмарина и

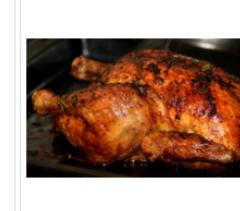


Свински бут във фолио на фурна
Разбийте всички продукти за маринадата, добрее облейте месото, след като сте го пробили на няколко



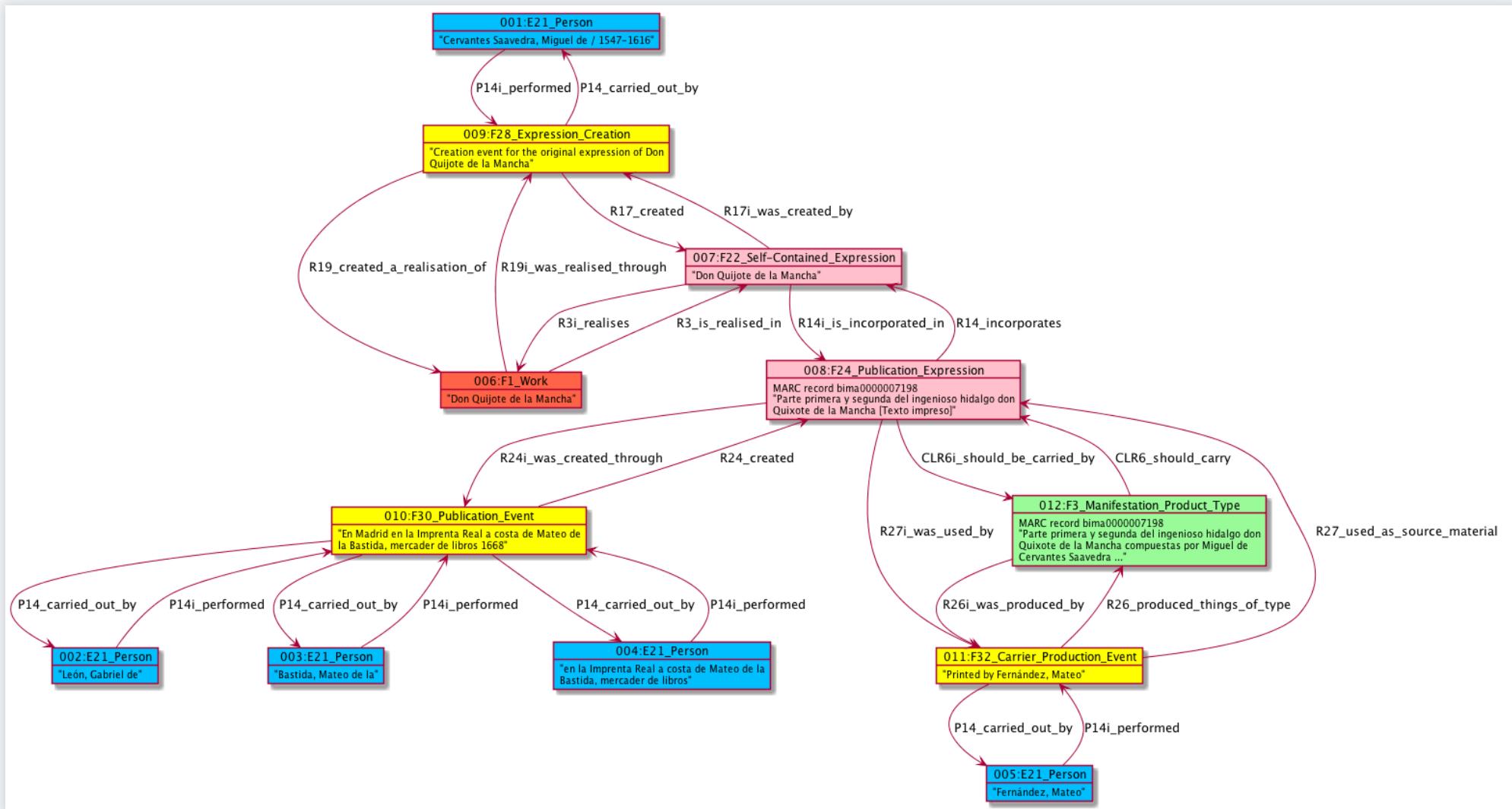






Europeana

- Members council, 5 task forces (e.g. below: working on FRBRoo-EDM profile), Data Quality Committee



Bulgariana

- National aggregator: initiator

[Bulgariana](#) - An aggregator that contributes Bulgarian cultural heritage content to Europeana

[Bulgariana Collections Published in Europeana](#) [Collections](#) ▾ [Digital Repository](#) ▾ [News, Events](#) ▾ [Related Materials](#) ▾

Pra-historic and Thracian Civilizations

Unpublished Thracian archaeological objects collected by Prof. Valeria Fol, Center of Thracology at the Institute for Balkan Studies

Click on the image to view the collection.



Getty Research Institute: Vocabularies LOD

► Complete services: GraphDB, ontology design, mapping, documentation, support... Helping on Getty Museum LOD

The screenshot shows the Getty Vocabularies LOD interface. The top navigation bar includes the logo, "Getty Vocabularies: LOD", "SPARQL", "Queries", search fields, and a "BETA" indicator. The left sidebar lists categories and numbered items:

- Family
 - 5.11 ULAN Subjects Linked to LCNAF
 - 5.12 German, Dutch, Flemish printmakers, listed with their teachers
 - 5.13 Artists Whose Identity May be Associated or Confused With Another
 - 5.14 Ordered Hierarchy of Given Subject
 - 5.15 Ancient Artists or Groups by Nationality
 - 5.16 Art Repositories in the USA by State
 - 5.17 Popes and Their Reigns** (4)
 - 5.18 Pope Reign Durations
- 6 Language Queries
 - 6.1 Scientific Names by Language
 - 6.2 Scientific Names not in English and Latin
 - 6.3 Find Terms by Language Tag
 - 6.4 Languages and ISO Codes
 - 6.5 Language URLs
 - 6.6 Custom Language Tags
 - 6.7 Chart AAT Languages with VISU
 - 6.8 Chart TGN Languages with VISU
- 7 Counting and Descriptive Info
 - 7.1 Descriptive Info from VOID
 - 7.2 Number of Entities from VOID
 - 7.3 Number of Local Sources (Dynamic)
 - 7.4 Number of Global Sources (Dynamic)
 - 7.5 Associative Relations Count
 - 7.6 Number of AAT Revision Actions
 - 7.7 ULAN Facet Counts
 - 7.8 ULAN Agents by Type
 - 7.9 ULAN Agents by Nationality
 - 7.10 ULAN Events by Type
- 8 Explore the Ontology
 - 8.1 Ontology Classes and Properties
 - 8.2 Ontology Values

The main area contains a SPARQL query editor with numbered callouts:

- SPARQL button
- Search field
- Search button
- Results table header
- Query code (for 5.17):

```
1 select ?x ?name ?bio ?start ?end {  
2   ?x gvp:agentTypePreferred [rdfs:label "popes"@en];  
3     gvp:prefLabelGVP [xl:literalForm ?name];  
4     foaf:focus [  
5       bio:event [dct:type [rdfs:label "reign"@en]; gvp:estStart ?start; gvp:estEnd ?end];  
6         gvp:biographyPreferred [schema:description ?bio]]  
7 } order by ?start
```

- Checkboxes for "Include inferred" and "Expand results over equivalent URIs"
- Submit button
- Section title: 5.17 Popes and Their Reigns
- Query code (for 5.17):

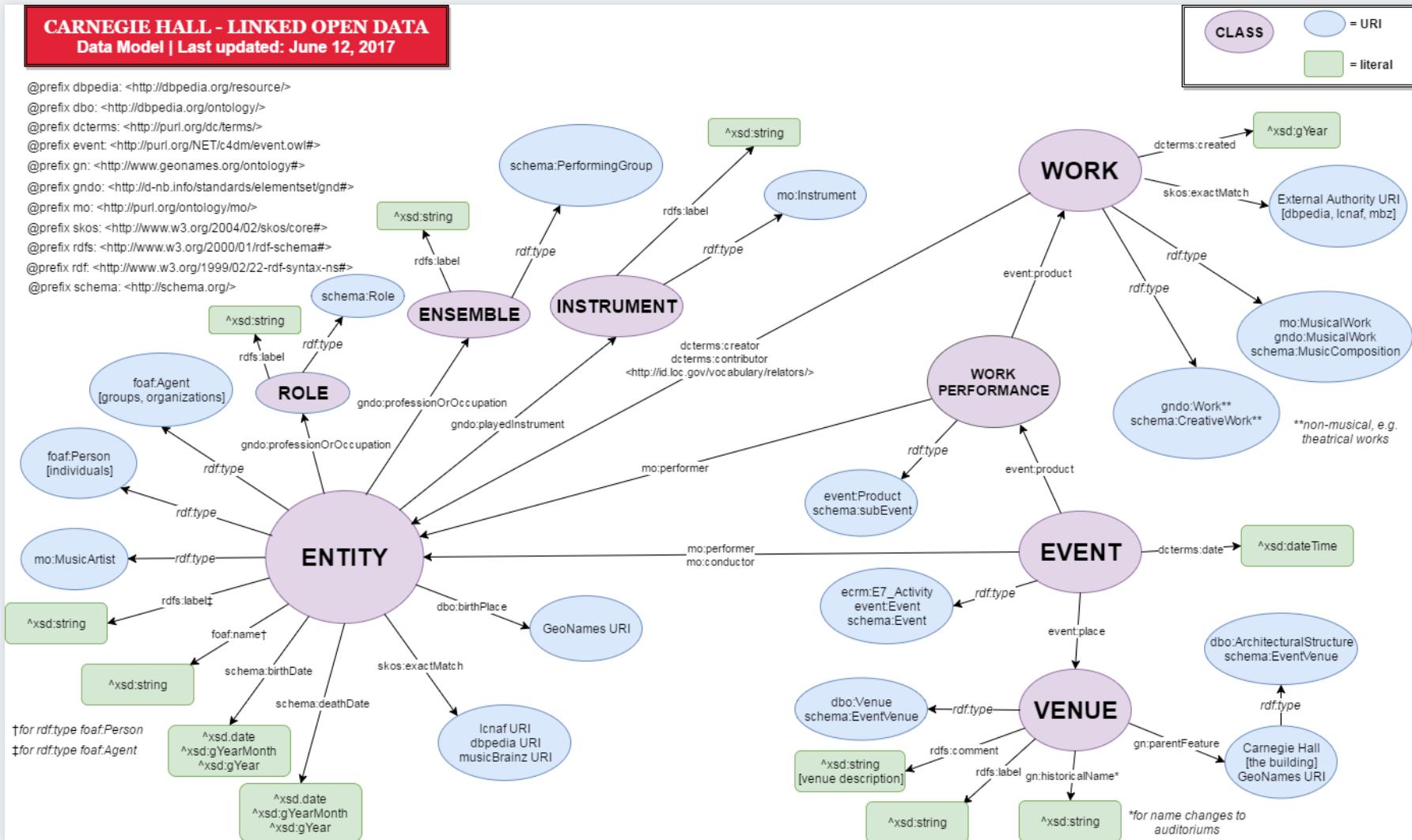
```
select ?x ?name ?bio ?start ?end {  
?x gvp:agentTypePreferred [rdfs:label "popes"@en];  
gvp:prefLabelGVP [xl:literalForm ?name];  
foaf:focus [  
  bio:event [dct:type [rdfs:label "reign"@en]; gvp:estStart ?start; gvp:estEnd ?end];  
  gvp:biographyPreferred [schema:description ?bio]]  
} order by ?start
```

Returns 127 popes. There is one ([ulan:500324155](#)) Pius VI for which the reign is not recorded.

- Section title: 5.18 Pope Reign Durations
- Text: Let's chart the durations of Popes' reigns.
- Query code (for 5.18):

```
select ?dur (count(*) as ?c) {  
?x gvp:agentTypePreferred [rdfs:label "popes"@en];  
  foaf:focus [bio:event [dct:type [rdfs:label "reign"@en]; gvp:estStart ?start; gvp:estEnd ?end]].  
  bind(xsd:integer(str(?end))-xsd:integer(str(?start)) as ?dur)
```

Carnegie Hall LOD



American Art Collaborative

Consulting: 14 US museums integrating data using CIDOC CRM. [Mapping Review app](#)

Classification

The type of object the work is.

Style

The style or period of the work.

Subject

What the work depicts.

Concept

What the work is about.

Technique

How the object was made.

Materials

What the work is made of.

Medium Text

A medium descriptive text.

Physical Object

Current Owner ↗

Who owns the work.

Current Location

Where the work is located.

Responsible Department

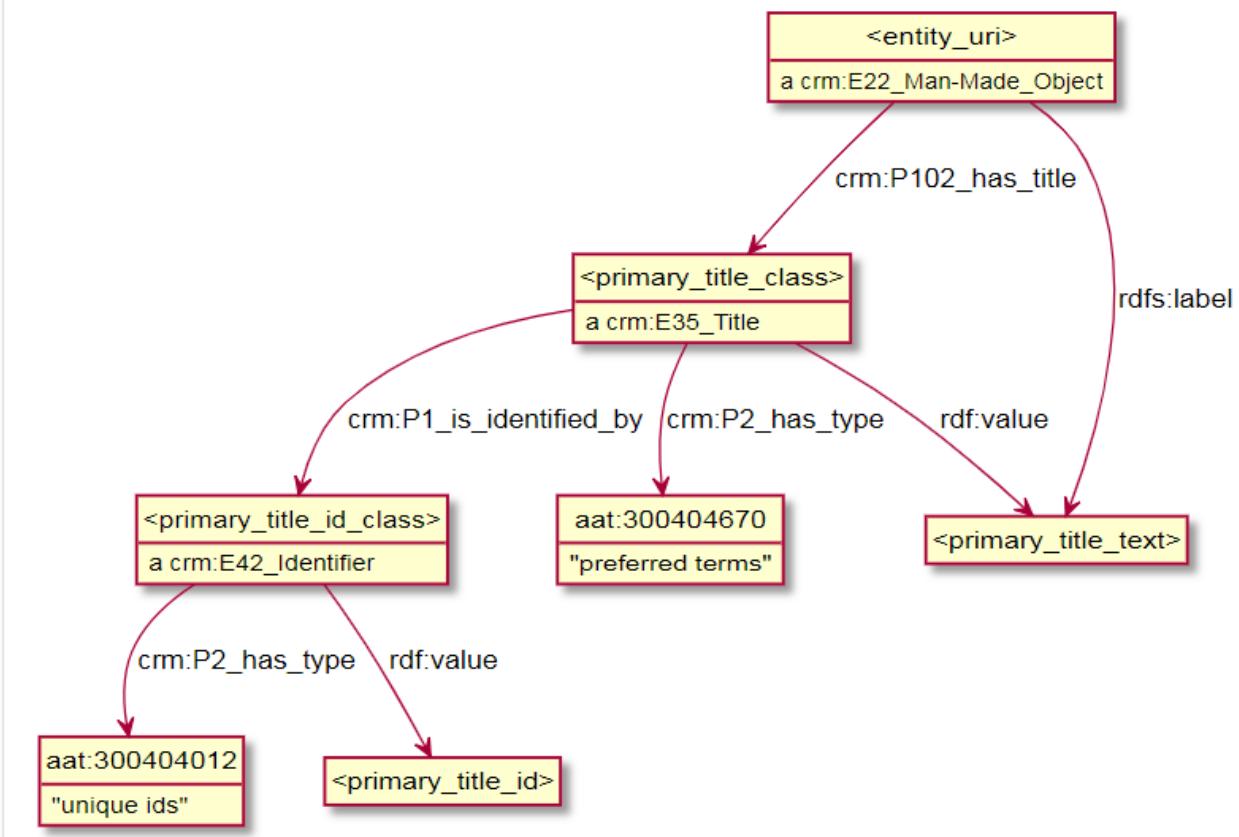
Institution Department responsible for the work.

Dimensions

Physical size of the work.

Dimensions (Part)

AAC Target Mapping For Primary Title



European Holocaust Research Infrastructure

- Semantic archive integration. 4+4 years, heading towards ERIC. EAD conversion

The image shows a composite screenshot. At the top left is a black and white photograph of a long, low concrete wall with a tall, textured chimney-like structure in the background. To the right of the wall is the EHRI logo, which consists of a purple square containing a white stylized 'E' and 'H' above the letters 'RI'. Below the logo, the text 'EUROPEAN HOLOCAUST RESEARCH INFRASTRUCTURE' is written in a smaller, white, sans-serif font.

Below this header, the word 'EAD CONVERSION TOOL' is centered in a bold, dark purple font. Underneath it is a horizontal navigation bar with six circular icons numbered 1 through 6. To the right of the bar is a link labeled 'DOCUMENTATION' and below it is a link labeled 'VIEW GOOGLE SPREADSHEET'.

The main area features a Google Sheets interface titled 'US-USHMM-mapping-config'. The sheet has four columns: 'target-path' (A), 'target-node' (B), 'source-node' (C), and 'value' (D). The data is as follows:

target-path	target-node	source-node	value
/	ead	//doc	
/ead/	eadheader	.	
/ead/eadheader/	profiledesc	.	
/ead/eadheader/profiledesc/	creation	/str[@name="datetimemodified"]	"This EAD is created by EHRI on ", <date>
/ead/	archdesc	.	
/ead/archdesc/	did	.	
/ead/archdesc/did/	unitid	if (/str[@name="accession_number"]/text() != /str[@name="id"]/text()) then /str[@name="id" text()]	attribute label ("accession_number"), &
/ead/archdesc/did/	unitid	/str[@name="accession_number"]	attribute label ("former_accession_number")
/ead/archdesc/did/	unitid	/arr[@name="accession_number_add"]/str	attribute label ("recordgroup_number")
/ead/archdesc/did/	unitid	/arr[@name="rg_number"]/str	attribute label ("subtitle"), text()
/ead/archdesc/did/	unittitle	/arr[@name="subtitle"]/str	attribute label ("alternative"), text()
/ead/archdesc/did/	unittitle	/arr[@name="title_alternate"]/str	

At the bottom of the sheet, there are buttons for '+', 'Sheet1', and a green 'Разглеждане' (View) button.

Below the sheet, there are two dropdown menus: 'Select Local Mapping' and 'Select local mapping file'. At the very bottom are 'PREVIOUS STEP' and 'NEXT STEP' buttons.

EHRI EAD Validation

EAD validation, HTML preview and integrated error display; publishing/transport and ingest to Portal

[Profile Description](#)

[Archival Description](#)

[Date of the Unit](#)

[Date of the Unit](#)

[Conditions Governing Access](#)

[Acquisition Information](#)

[Arrangement](#)

[Biography or History](#)

[Biography or History](#)

[Scope and Content](#)

[Conditions Governing Use](#)

Profile Description

[ERROR]
element "profiledesc" not allowed yet; missing required element "eadid"

Creation

This EAD is created by EHRI on 2017-02-02+02:00 based on the JSON file provided by USHMM on TODO: find out where to get this . This JSON file is constructed on a Catalog Record that was last modified on 2016-11-17 11:12:18 .

Archival Description

[ERROR]
element "archdesc" missing required attribute "level"

Descriptive Identification

ID of the Unit

irn515021

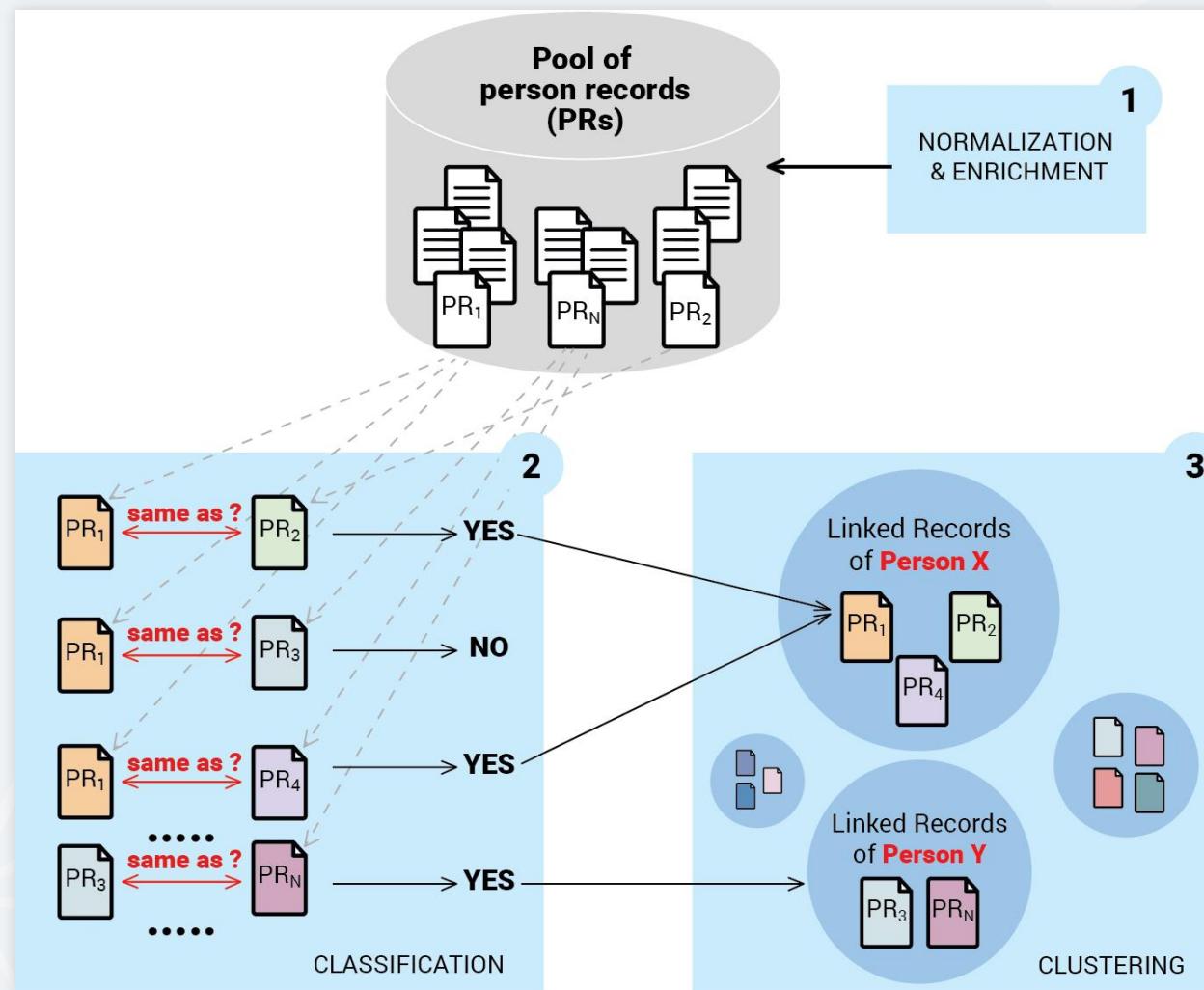
ID of the Unit

Label accession_number

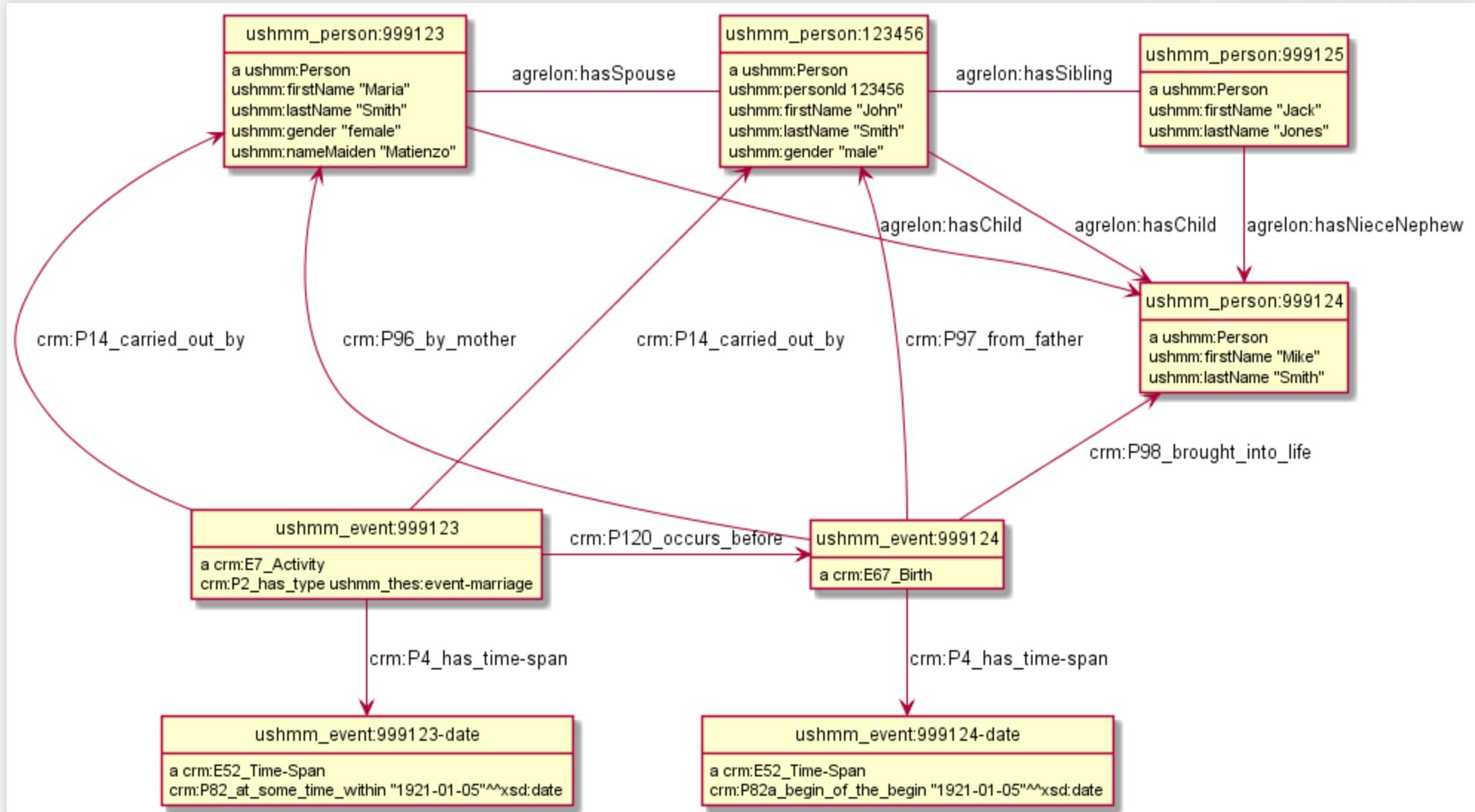
2004.273.1

EHRI Semantic Services

- Geonames Coreferencing service (from EAD access points etc), EHRI Thesaurus (VocBench), [Person Deduplication](#) (record linking)

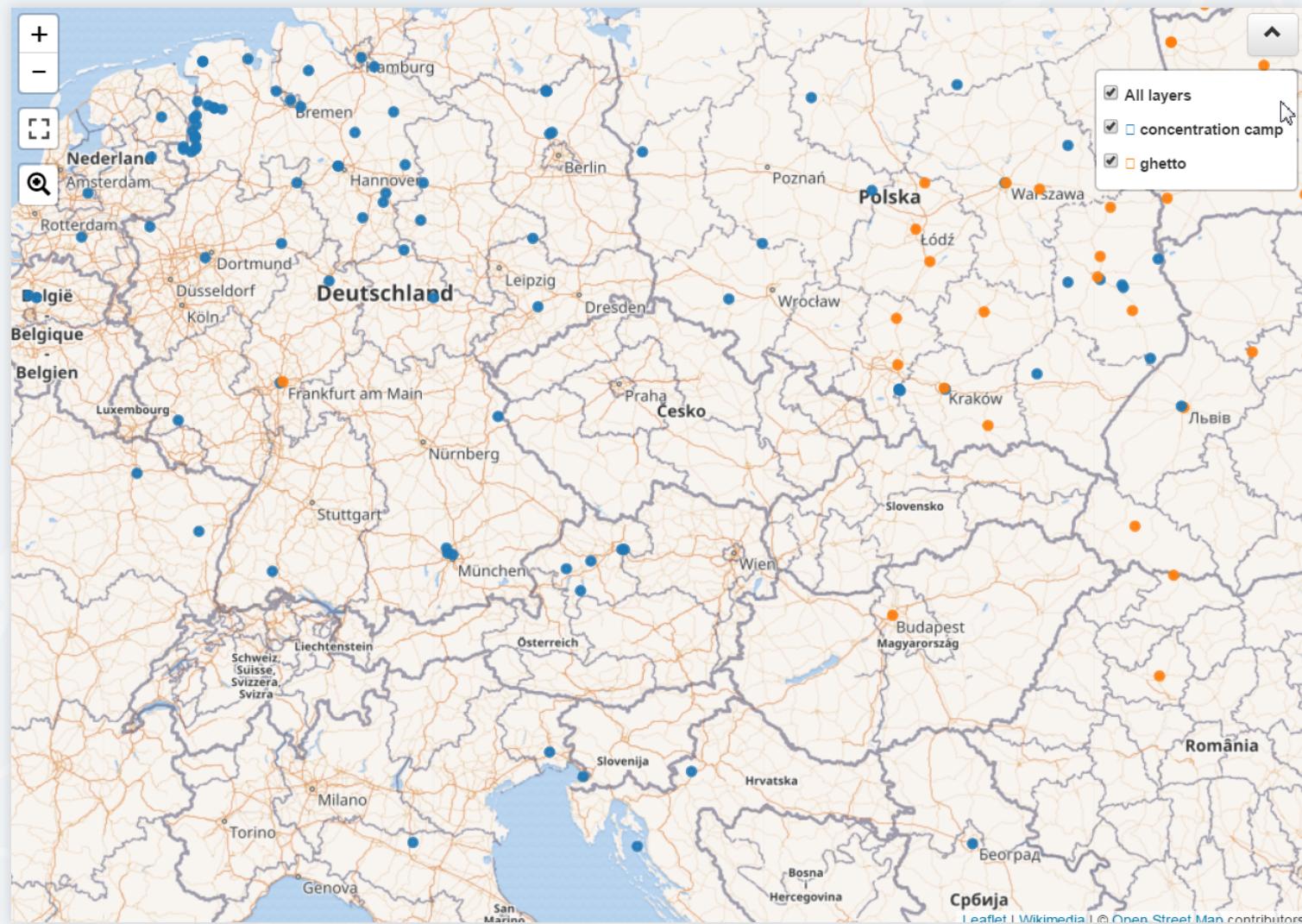


EHRI Person networks



EHRI Camps and Ghettos

► Integrating Camps and Ghettos info between EHRI and Wikidata



Canadian Heritage Information Network

- ▶ CA national aggregator is transitioning to LOD. 4 Consulting projects: environment scan, strategy, Artefacts Canada data analysis, national authorities



Artefacts Canada Data Analysis

Authors: Nikola Tulechki, Laura Tolosi, Vladimir Alexiev. Ontotext Corp

Version 1.2, Last Edited 27 Mar 2018

1 Contents

[1 Contents](#)

[2 Introduction](#)

[2.1 Executive Summary](#)

[2.1.1 How Much to Clean-Up](#)

[2.1.2 Another LOD Pilot?](#)

[2.1.3 Future Efforts](#)

[2.2 Glossary and Links](#)

[2.3 Revisions](#)

[2.4 Deliverables](#)

[2.4.1 Exports](#)

Wikidata Authority Control

[1-50](#) | [51-100](#) | Show unmatched | Show auto-matched | Show user-matched | Show NoWD | Show N/A | [Site stats](#)

Title/Q	Description	Actions
pyrolusite	mineral, inorganic material, <materials by composition>, materials (matter), Materials (Hierarchy Name), Materials Facet	Matched by Vladimir Alexiev
pyrolusite  Q413293	Rutile mineral group, named after fire and washing; oxide mineral	Remove
sodium chlorite	sodium compounds, sodium, inorganic material, <materials by composition>, materials (matter), Materials (Hierarchy Name), Materials Facet	Matched by Vladimir Alexiev
sodium chlorite  Q411294	Chemical compound; chemical compound	Remove
silanol	compounds (materials), <materials by chemical form>, <materials by form>, materials (matter), Materials (Hierarchy Name), Materials Facet	Matched by Vladimir Alexiev
Silanol  Q420482	Chemical substance	Remove
gellan gum	gel, colloid (particulate material), <materials by physical form>, <materials by form>, materials (matter), Materials (Hierarchy Name), Materials Facet	Matched by Vladimir Alexiev
Gellan gum  Q416694	Chemical compound, thickening agent, and polysaccharide; chemical compound	Remove
stonemasonry	<processes and techniques by material>, <processes and techniques by specific type>, <processes and techniques (processes and techniques)>, Processes and Techniques (Hierarchy Name), Activities Facet	<i>Not matched</i>
Search Wikidata Search en.wikipedia Google-search Wikipedia Google-search Wikidata Create Wikidata item		Set Q New item N/A

Wikidata/DBpedia vs VIAF/GND; and Europeana

Name Data Sources for Semantic Enrichment (Europeana Creative D2.4). [Wikidata, a Target for Europeana's Semantic Strategy](#). (GlamWiki 2015)

VIAF	id in VIAF	Wikidata	id in Wikidata
viafID	49268177	VIAF	49268177
BAV	ADV10197613		
BNC	.a10853637		
BNE	XX907273		
BNF	cb12176451h	BNF	12176451h
DNB	118522582	GND	118522582
ISNI	00000000121319721	ISNI	0000 0001 2131 9721
JPG	500115364	ULAN	500115364
LC	n50020861	LCCN	n50020861
LNB	LNC10-000002573		
NDL	00436834		
NKC	jn20000700335		
NLA	000035031951		
NLI	000035532,001445575,001448179		
NLP	a16828161		
NTA	068435312	NTA PPN	068435312
NUKAT	vts000190728		
SELIBR	182422		
SUDOC	028710010		
WKP	Lucas_Cranach_the_Elder	Many Wikipedias	
IMAGINE	T7238,T267474	Cantic	a10853637
		Commons Creator	Lucas Cranach (I)
		Commons category	Lucas Cranach d. Ä.
		Freebase	/m/0kqp0
		RKDartists	18978
		SIMBAD	CRANACH, Lucas the Elder
		Your Paintings	lucas-the-elder-cranach

→ Can be leveraged to fill the gaps, e.g. bring RKDartists into VIAF

