Europeana Semantic Data in OWLIM

Mariana Damova, PhD

The Bulgarian Participation in Europeana.
Cooperation and Development
Varna
October 2012

Ontotext

- Top-5 provider of core Semantic Technology
- Established in year 2000; offices in Bulgaria, UK, USA
- Active both in research and commercial projects (FP7 funding for 10 years)

- 360° semantic technology – unique portfolio:
  - **Semantic Databases**: high-performance RDF DBMS, scalable reasoning
  - **Semantic Search**: text-mining (IE), metadata generation, Information Retrieval (IR)
  - **Web Mining**: focused crawling, screen scraping, data fusion
  - **Linked Data Management and Data Integration**

Good recognition in the SemTech community

- Ontotext pages are ranked #1 for “semantic annotation” and “semantic repository” at GYM, #3 for “linked data management” at Google

Several joint ventures and subsidiaries

- Innovantage: leading online recruitment intelligence provider in UK

ontotext

Ontotext Clients (selected)

British Broadcasting Corporation (BBC)

- Run its World Cup 2010 sites on top of OWLIM
- Since Mar'12 BBC Sports
- 2012 Olympics sections are driven by OWLIM and a Concept Extraction service developed by Ontotext

Press Association (UK)

- Analysis of Sports news
- Concept extraction
- Linked data generation

Top-3 USA media (not allowed to name)

The National Archives (UK) contracted Ontotext to implement semantic KB and semantic search for the Government Web Archive

British Museum (UK) Ontotext leads the development of Phase 3 of ResearchSpace project on collaborative research in cultural heritage; British Museum's public SPARQL end-point is powered by OWLIM

de Bibliothek (Holland) aggregation of data from 150 library databases

ontotext

Outline

- Europeana
- bulgariana.eu
- Collections
- Europeana Data Standards
- Metadata mapping, conversion and ingestion
- Digital repository
- Conclusion

Europeana

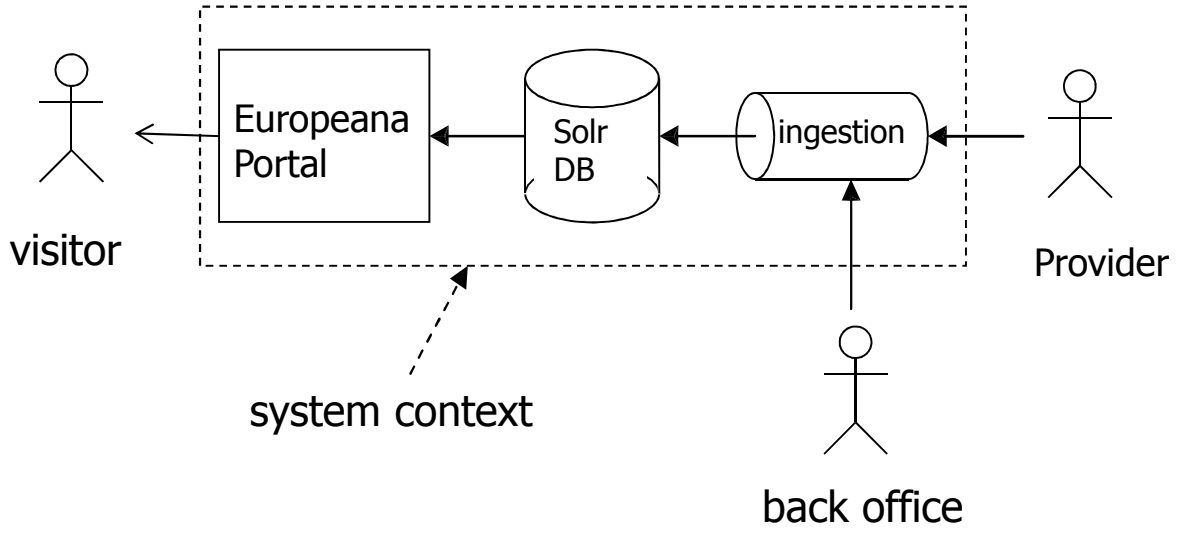
http://www.europeana.eu

- Launched in 2008
- Project funded by the European Commission
- Based in the National Library of the Netherlands, the Koninklijke Bibliotheek
- Goal to make Europe's cultural and scientific heritage accessible to the public
- Over 180 heritage and knowledge organzations and IT experts across Europe
- Europeana Collection: 5M objects in 2009, 10M in 2010, 20M at present
- Endorsed by the European parliament in 2010
- **2011** "Comité des Sages" makes recommendations about Europeana

  to put online the collections held by Europe's libraries, archives, museums and audiovisual archives – vast numbers of books and periodicals (there are some 2.5bn items in Europe's libraries alone), and millions of hours of film and video covering the whole of Europe's diverse history and culture.

Europeana

- Collection types: Image, Sound, Video, Text
- Present Europeana Architecture

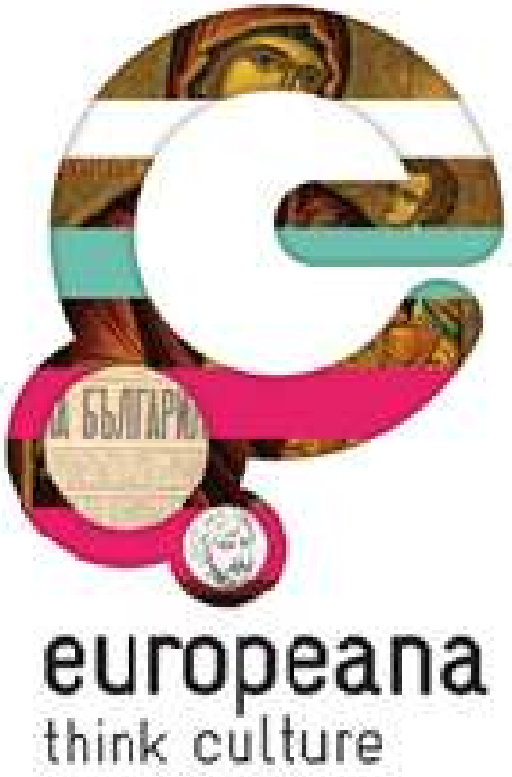


- Europeana data standards
- Europeana aggregators (by country or cultural heritage sector)
- Process of ingesting content (4-6 weeks)

bulgariana.eu

bulgariana.eu

- Main Purpose: BG aggregator for Europeana
- Secondary Purpose: networking and special interest group for BG Cultural Heritage



Collections

Collections

Golden Pages from the Bulgarian Renaissance

Златни страници от Българското Възраждане

unique manuscripts of Bulgarian folk songs collected in 19th century
by Miladinov Brothers, renowned Bulgarian Folklorists
published in 2008 by D-r Luchia Antonova,
Institute of Bulgarian Language, Bulgarian Academy of Sciences

МАРКО КРАЛЕВИКИ БОЛЕН СЕ КАИТ И СЕ ИСПОВЕДВИТ

Поболил се Марко Кралевике,
що си лежал токму три години,
от нищо се иляч (1) не на'ож'ал.
И му рече негва стара майќа:
"Ай ти, Марко, ай ти, синко милий;
не си болен, синко, от господа,
тук си болен, синко, от гре'о'и,
да ти викна попой (2), ду'овници,
лепо да се синко исповедиш,
да си кажиш твоите гре'о'и!"
....

Collections

Pra-historic and Thracian Civilizations

Праисторическа и Тракийска цивилизация

Unpublished Thracian archeological objects collected by Prof. Valeria Fol, Center of Thracology at the Institute for Balkan Studies at the Bulgarian Academy of Sciences

Links

- http://bulgariana.eu
- http://bulgarianheritage.bulgariana.eu
- http://www.europeana.eu
  - europeana_collectionName: 20215*
  - for the individual sets use europeana_collectionName: 2021501* (or 2021502*)
- http://britishmuseum.ontotext.com

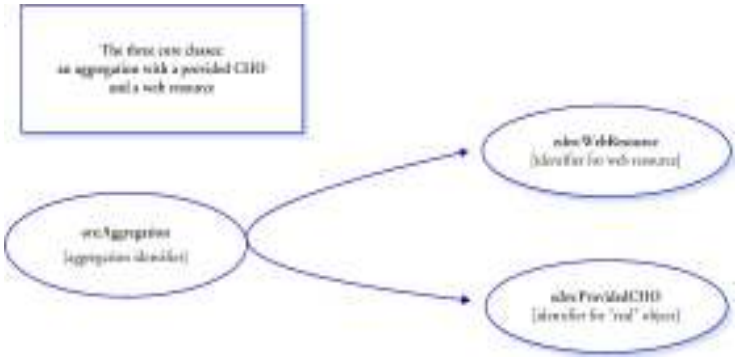
Europeana Data Standards

Europeana Data Standards

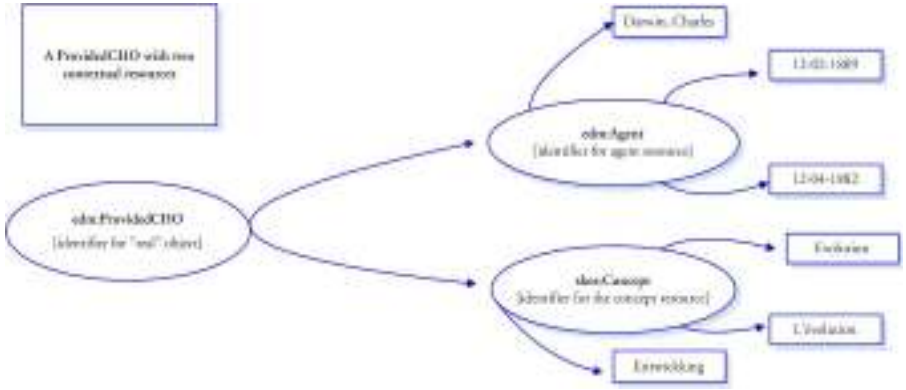
- Unified metadata
  - ESE – Europeana Semantic Elements
    - DublinCore & Europeana fields
    - 36 fields: flat, limited ability semantic links

dc:title	europeana:provider
dc:creator	europeana:dataProvider
dc:subject	europeana:rights
dc:description	europeana:type
dc:publisher	europeana:isShownBy and/or europeana:isShownAt
…	…

  - EDM - Europeana Data Model



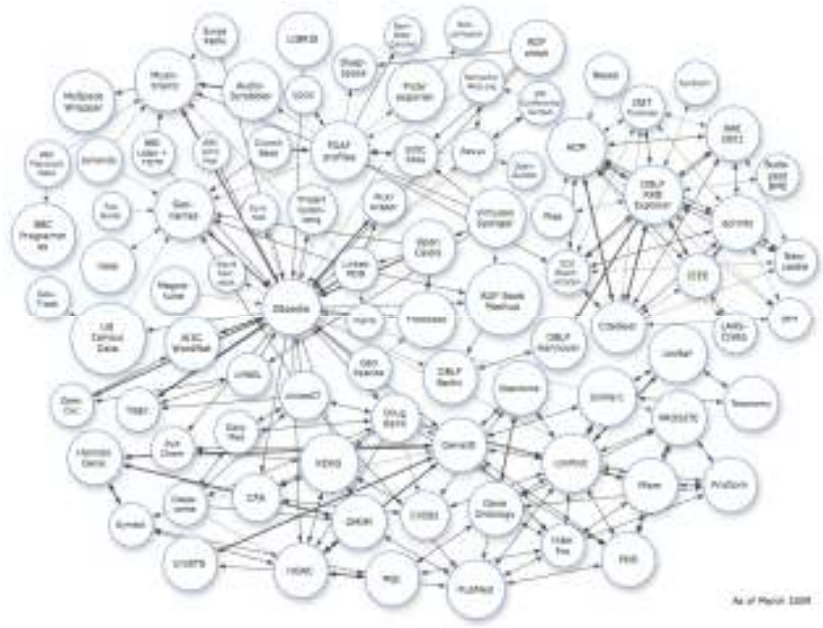
Basic data model



Two contextual classes

Linking Open Data

- Linking Open Data (LOD) W3C SWEO Community project
  http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData



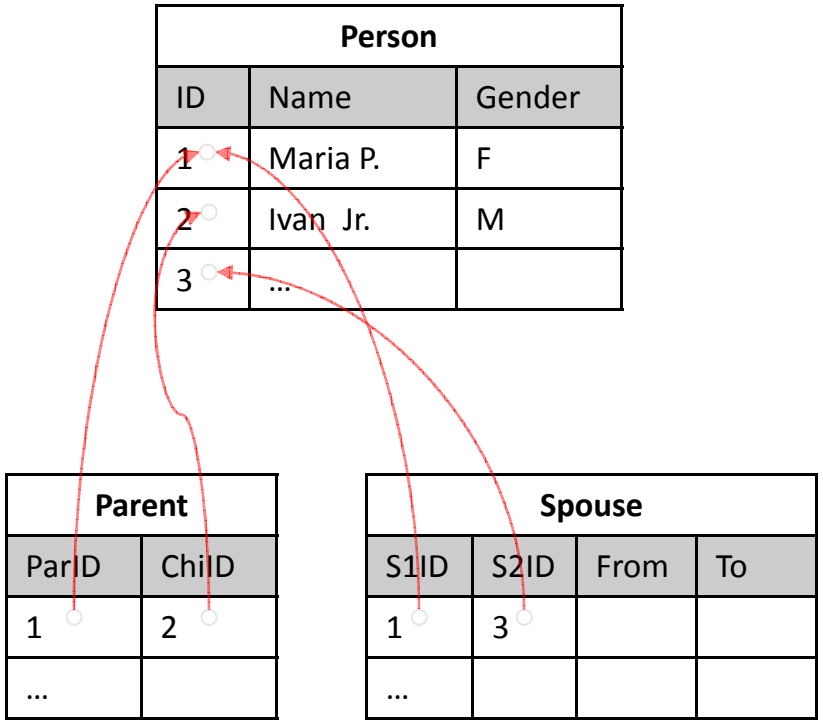
- Initiative for publishing “linked data” – a set of principles, which allows browsing of RDF data, spread across different servers, in the way HTML is browsed

ontotext

Semantic Repositories: Major Characteristics

- *Easy integration of multiple data-sources*
  - once the schemata of these sources is semantically aligned, the inference capabilities of the engine supports the interlinking and combination of the facts from the different sources;
- *Easy querying against rich or diverse data schemata*
  - inference is applied to match semantics of the query to the semantics of the data, regardless of the vocabulary and the data modeling patterns used for encoding of the data;

Physical data representation: RDF vs. RDBMS



Person		
ID	Name	Gender
1	Maria P.	F
2	Ivan Jr.	M
3	…	

Parent	
ParID	ChiID
1	2
…	

Spouse			
S1ID	S2ID	From	To
1	3		
…			

Statement		
Subject	Predicate	Object
myo:Person	rdf:type	rdfs:Class
myo:gender	rdfs:type	rdfs:Property
myo:parent	rdfs:range	myo:Person
myo:spouse	rdfs:range	myo:Person
myd:Maria	rdf:type	myo:Person
myd:Maria	rdf:label	“Maria P.”
myd:Maria	myo:gender	“F”
myd:Maria	rdf:label	“Ivan Jr.”
myd:Ivan	myo:gender	“M”
myd:Maria	myo:parent	Myd:Ivan
myd:Maria	myo:spouse	myd:John
…		

OWLIM

OWLIM is a family of semantic repositories, or RDF database management systems, with the following characteristics:

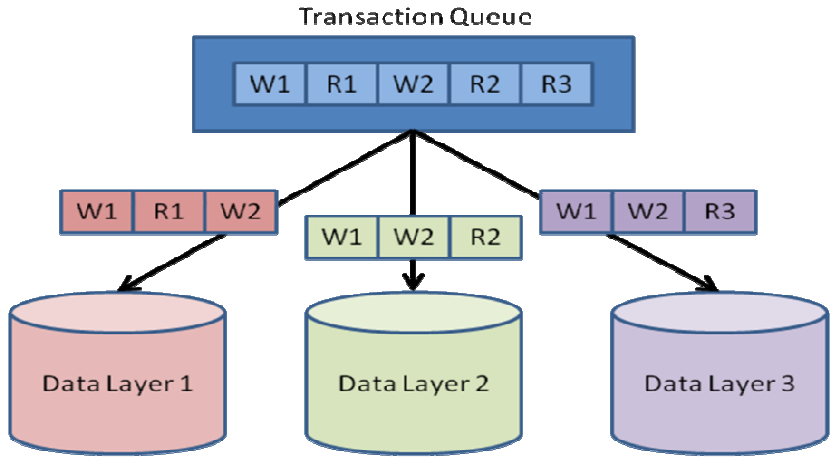
- native RDF engines, implemented in Java
- delivering full performance through both Sesame and Jena
- robust support for the semantics of RDFS, OWL 2 RL and OWL 2 QL
- best scalability, loading and query evaluation performance

OWLIM is used in a large number of research projects and software tools. Independent opinions justifying our bold claims are referred to here.

The presentation Lowering the Cost of Data and Content Integration and enabling Searching and Querying of Billions of Facts on the Web presents the key features of OWLIM alongside an introduction to the benefits of using RDF databases for data integration and a discussion on linked data management.

OWLIM Replication Cluster

- Distribution through data replication is used to ensure:
  - Better handling of concurrent user requests
  - Failover support
- How does it work?
  - Every user request is pushed in a transaction queue
  - Each data write request is are multiplexed to all repository instances
  - Each read request is dispatched to one of the instance only
  - To ensure load-balancing, each read requests is send to the instance with smallest execution queue at this point in time



ontotext

Europeana Data in EDM

- 268GB of data
- cultural objects data and linkages to other datasets

Dataset size:

NumberOfStatements=3,899,531,218
NumberOfExplicitStatements=993,332,911
NumberOfEntities=264,523,842

EDM model
SKOS

Prototype available at

http://europeana.ontotext.com

to become

-> http://data.europeana.eu

ontotext

Upcoming …

Europeana Creative - PSP project

lead by the Austrian National Library

26 partners

Objective: experimenting with re-use of cultural content for creativity

Project: Europeana re-use framework and 6 pilots in different domains such as education, tourism, etc.

Ontotext: participate in the infrastructure for re-use with the semantic repository OWLIM, and data integration

Upcoming …

Round table organized by the Ministry of Culture and Ontotext where Europeana officials will explain the organizational principles of Europeana data collection and aggregation and will share experience with setting up national aggregator's to be held in November 2012 in Sofia

Thank you for your attention!

mariana.damova@ontotext.com