Europeana Semantic Publishing: Followup Report

Author: Vladimir Alexiev, PhD, PMP

Updated: 6/10/2013 6:43 PM.

1 Contents

Europeana Semantic Publishing: Followup Report 1
1 Contents 1
2 Project Data 1
2.1 Project Description 1
2.1.1 Main project goals 1
2.1.2 Secondary project goals 1
2.1.3 Project components 2
2.1.4 Achieved results 2
2.2 Project Partners 2
2.3 Specificatons, Reviews, Reports 2
3 News and Events 3
3.1 Bulgariana.eu News 3
3.2 Ontotext News 3
3.3 Other News 4
4 Sustainability After the End of the Project 4
5 Attachments 5

2 Project Data

- Project Name: **Semantic Publishing of Bulgarian Cultural Heritage in Europeana**
- Project start: **18 Oct 2011**
- Project end: **21 Feb 2012**

2.1 Project Description

Bulgaria has very rich and important cultural heritage, but little of it is digitized and publicized widely.

Europeana is Europe's virtual Gallery, Library, Archive and Museum (GLAM) that aggregates over 20M cultural objects from 1500 institutions (and growing).

The aim of this project is to showcase some key Bulgarian collections in Europeana.

2.1.1 Main project goals

- Digitize data about selected Bulgarian cultural heritage collections
- Convert the data to semantic format
- Publish the data in Europeana

2.1.2 Secondary project goals

- Establish a technical aggregator to Europeana to publish Bulgarian cultural heritage (www.bulgariana.eu)
- Establish a semantic publication process that's easy to use by other cultural heritage institutions
- Establish a Special Interest Group (SIG) for this purpose and coordinate with other like-minded people and institutions

2.1.3 Project components

This is a relatively complex project with several components that were executed by different organizations (see next section about the project partners)

- Create specifications
- Provide collection content
- Establish digital repository for storing the collections
- Convert data to semantic format, establish Europeana aggregator, publish the data to Europeana, networking

2.1.4 Achieved results

All project goals were achieved; in some areas with much greater impact than expected (see News and Events below)

- Established a repository for Bulgarian cultural heritage, based on D-Space. This will allow the preservation and presentation of Bulgarian cultural heritage by smaller institutions that don't have their own repository. This was done by project partner Sirma Media
- Created the necessary metadata models and added two collections to repository (with their respective sub-collections per object type):
  - Pra-historic and Thracian Civilizations: ancient golden treasures, tombs, etc. High-quality photos and metadata descriptions. This work was done by project partner Center for Thracology, Balkan Studies Institute
  - Golden Pages of the Bulgarian Renaissance: a small collection of hand-written songs. This work was done by Ontotext

  We were also hoping to include already digitized psalters and manuscripts by the Cyrillo-Methodian Research Institute, but this part was canceled by ITCC for lack of funding
- Established an appropriate mapping and converted the data to semantic format
- Established a technical aggregator to Europeana to publish Bulgarian cultural heritage (http://bulgariana.eu). Created an infrastructure that enables organizations to publish their digital collections online and then make publish them for viewing via Europeana
- Published the data to Europeana
- Formed a community of experts, officials and activists who work actively for the digitalization, presentation and preservation of Bulgarian cultural heritage.
  - Established a Google group as communication channel
  - Established http://bulgariana.eu as a source of information on digitization, publishing to Europeana and events on these topics. All papers and presentations from events organized by Ontotext are freely available at this site. We firmly believe in the open exchange of information
  - Organized a number of networking events and conferences (see below)

2.2 Project Partners

- Sirma Media Corp
  An IT company with a long history in cultural heritage and e-learning.
  They established a repository for storing cultural heritage collections.
- Center for Thracology, Balkan Studies Institute, Bulgarian Academy of Sciences.
  Described and contributed the currently largest collection published by this project.

2.3 Specificatons, Reviews, Reports

- Project specifications: a number of detailed specifications were developed in the summer of 2011. Attached
- Mid-term review: 21 Nov 2011: presentation given to Korean representative. Attached
- Progress report: 18 Jan 2012. Submitted to ITCC
- Final report: 22 Feb 2012. Submitted to ITCC and Attached.
- Independent Review: Mar 2012 by Dr Radko Pavlov, Digitization Lab, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences. Submitted to ITCC

3 News and Events

The underlined links below are active in the electronic version and lead to more information.

- News specifically about Bulgariana are marked with this special bullet

3.1 Bulgariana.eu News

https://confluence.ontotext.com/display/Bulgariana/News%2C+Events

- 20120130 Европеана - Мисията възможна
  Conference "Europeana: Mission Possible" organized by Ontotext in Sofia. 75 participants.
- 20120208 bulgariana.eu представена по БНР
  Bulgariana.eu presented in an interview on Bulgarian National Radio
- 20120306 First Bulgariana Collections Published in Europeana
  Successful publication of collections in Europeana, i.e. de-facto completion of the ITCC-funded project
- 20120319 Семинар по Европеана (В. Търново)
  A conference co-organized by Ontotext in Veliko Tarnovo. 47 participants. Ontotext paper:
  Mariana Damova: Bulgariana Collections in Europeana
- 20120404 Bulgariana and Europeana discussed on National Radio
  Second interview about Bulgariana and Europeana on Bulgarian National Radio
- 20120503 Multilingual Digital Repositories and Services
  A conference in Sofia supported by Ontotext
- 20120613 Europeana Plenary 2012
  Ontotext participated in the Europeana Plenary in Leuven, Belgium
- 20120918 DiPP 2012 Conference. Ontotext participated with two papers:
  Vladimir Alexiev: Types and Annotations for CIDOC CRM Properties
  Mariana Damova: Bulgariana: A Bulgarian Aggregator to Europeana
- 20121017 Българското участие в Европеана. Сътрудничество и развитие
  Conference "Bulgarian Participation in Europeana. Cooperation and Development" co-organized by Ontotext in Varna. Ontotext paper:
  Mariana Damova: Europeana Semantic Data in OWLIM
- 20121109 Round Table at the Bulgarian Ministry of Culture with Participation of Europeana
  Organized by Ontotext in Sofia. 37 participants, including Annette Friberg and Francesca Morselli, Europeana Foundation, and Arch. Stoev, Bulgarian Deputy Minister of Culture
- 20121121 Coverage of the Round Table at the Bulgarian Ministry of Culture with Participation of Europeana on Bulgarian National Radio
  A third interview on Bulgarian National Radio
- 20121127 Ontotext Featured at Europeana AGM
  Ontotext participated in the Europeana Annual General Meeting in Berlin. Mariana Damova of Ontotext, was named an "Unsung Hero" in the "Aggregate" strategic track, for "the high quality of images and metadata that Ontotext provides and for its efforts to establish a Bulgarian national aggregator".
- 20130321 Национален форум ПРЕОБРАЖЕНИЯТА НА Б-ПРОСТРАНСТВОТО
  A conference in Varna. Ontotext participated with a presentation:
  Mariana Damova, The Road to Creating a Bulgarian National Aggregator for Europeana

3.2 Ontotext News

http://www.ontotext.com/news

- November 17, 2011: Ontotext wins landmark project with The British Museum
  ITCC's Bulgariana project helped Ontotext define a focus in Cultural Heritage, and win this important international project partly due to this factor
- February 16, 2012: Ontotext launches bulgariana.eu
  The main internal goal of the project is achieved
- March 15, 2012: Bulgariana collections published in Europeana
  The main external goal of the project is achieved

- May 9, 2012: The British Museum and Yale Center for British Art use OWLIM
  The Yale Center for British Art museum at Yale University, USA joins the British Museum in using Ontotext's semantic database.
- June 12, 2012: Cultural Heritage Aggregation for Dutch Public Library
- June 21, 2012: ResearchSpace to be Presented at OpenCulture 2012
- December 21, 2012: Ontotext releases Europeana SPARQL end-point
  Ontotext loaded Europeana's EDM data in OWLIM and provided semantic query (SPARQL) over it. This comprises 20M cultural objects, making about 950M facts (triples)
- January 15, 2013: europeana.ontotext.com named: "The first large public SPARQL 1.1 endpoint"
  The Europeana EDM endpoint received good acceptance from the semantic web community
- March 7, 2013: ConservationSpace project
  ConservationSpace is the second project funded by the Andrew Mellon foundation which Ontotext wins (in consortium with our sibling company Sirma ITT)
- March 7, 2013: Europeana Creative Kick-off
  Partly as a result of our work on the Europeana EDM endpoint, we were invited in this exciting Europeana-related project to make Europeana content more available to the Creative industries. This will open the way to future profitable collaborations
- April 5, 2013: ResearchSpace featured by Collections Trust
  Collections Trust is an influential UK organization in the museum sector.
- April 19, 2013: Interview with Europeana CEO in major Bulgarian daily
- May 23, 2013: Ontotext organizes workshop Practical Experiences with CIDOC CRM at TPDL 2013
  Ontotext organizes the workshop Practical Experiences with CIDOC CRM and its Extensions (CRMEX 2013), affiliated with the 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013). The workshop will be held on 26 September 2013 in Valetta, Malta. This is a high-profile and popular scientific conference.
- June 5, 2013: Bulgariana Collection Featured in Europeana's First Free iPad App
  An engaging iPad app featuring 350,000 images in categories: Maps and Plans, Treasures of Art, Treasures of the Past, Treasures of Nature and Images of the Past. The Bulgariana collection Bulgarian Heritage: Pra-historic and Thracian Civilizations" is featured. It is an open source app, so other developers can reuse, develop and extend it further. This app shows the power of Europeana's open cultural content approach.

3.3 Other News

News that are not published on our sites (view them in chronological order with the ones above)

- 19 May 2011: Participation in KR Global Smart SOC Initiative, Seoul, Korea. Presentation "Semtech for Cultural Heritage", attached
- Apr 2013: Ontotext added an extra collection Sacred Places with a 3D scan of one of Bulgaria's most important temples, the St. Alexander Nevsky cathedral
- 15 May 2013: Ontotext joined project proposals for 2 more Europeana-related projects (decision expected Aug 2013)
- Jun 2013: Ontotext created a company brochure dedicated to Cultural Heritage (Attached)
- Jul 2013: Ontotext is hiring a part-time person to work with Bulgarian museums to start deploying collection management and publishing through Bulgariana
- before year-end 2013: Ontotext hopes to establish Bulgariana as a Non Government Organization under Bulgarian law

4 Sustainability After the End of the Project

An ongoing concern of Korean ITCC initiatives worldwide is the application of project results after the end of the project (actual applications or potential applicability), i.e. the sustainability of project results.

We hope it should be obvious from the information above that this project has excellent sustainability. We can summarize them as follows:

- The project established Bulgariana as an Europeana aggregator, a digital repository, and a platform that cultural heritage institutions can use to store and publish their collections

- The project helped to promulgate the Europeana idea in Bulgaria. News, meetings and conferences continue and the cultural heritage community is prepared for larger-scale digitization and publication
- The project and additional Ontotext efforts helped to raise visibility for Bulgarian cultural heritage and efforts in Europeana
- Ontotext established itself as a strong vendor of semantic technologies in the cultural heritage domain, with key clients such as the British Museum, Andrew Mellon Foundation, etc
- Ontotext became visible on the global Europeana landscape, by hosting all Europeana semantic data on its repository (EDM SPARQL endpoint over the OWLIM repository)
- Ontotext joined its first EC-funded Europeana-related project, Europeana Creative, and is partner in 2 more proposals

What is still lacking in Bulgaria is national funding for digitization, preservation and publication of cultural heritage. We hope this will change with the new Bulgarian Reference Framework and the respective Operating Programs to come into force in 2014.

What we also hope may come in the future is **interest from Korean institutions and companies**:

- Unfortunately we were not able to secure Korean partners for this project. We established a semantic technology partner in Korea (**Saltlux**), but they were not particularly interested in cultural heritage.
- We firmly believe that cultural heritage preservation and presentation is a very important topic, and semantic technologies can play a very big role in it. We presented these ideas at the KR Global Smart SOC Initiative, Seoul, Korea (19 May 2011), including the idea of **Asiana** by analogy with Europeana
- **Japan** is already moving in this direction, see the http://lod.ac/ site by the National Institute of Informatics. The LOD.AC Museum tackles several challenges including:
  - Aggregate cultural heritage and museum information from all museums in Japan
  - Present it as a website for researchers and audience
  - Integrate natural language search and provide reliable SPARQL endpoint

  LOD.AC uses Ontotext's semantic repository OWLIM. See details at http://www.ontotext.com/case/Lodac

5 Attachments

This file includes the following documents in the indicated order. Report dates are also presented:

- Bulgariana Followup Report (20130610): this document

Specifications:

- Europeana- general (in Bulgarian): general project description
- Europeana P1 Cyril-Methodius (in Bulgarian): component for Cyrillo-Methodian collection
- Europeana P2 Thracian (in Bulgarian): component for Thracian treasures collection
- Europeana P3 Digital Repository (in Bulgarian): component for establishing a digital repository
- Europeana P3 Appendix-Survey-Repository-Tools (in English): survey of appropriate repository tools
- Europeana P4 Semantic Publishing (in Bulgarian): component for semantic publishing
- Europeana P4 Appendix- Europeana HowTo (in English): background on semantic publishing in Europeana

Presentations:

- Semtech for Cultural Heritage (20110519), KR Global Smart SOC Initiative, Seoul, Korea
- Mid-term review (20111121): presentation given to Korean representative.
- Europeana Publishing Final Report (20120222): Submitted to ITCC.
- Ontotext Cultural Heritage (20130601): company brochure

Софийски Университет
Българо-Корейски Център за ИТ кооперация
(ITCC)

Публикуване на Българско културно наследство в семантичен формат в Европеана

Създаден: 15 юни 2011
Редактиран: 22 авг 2011

1 Съдържание

Публикуване на Българско културно наследство в семантичен формат в Европеана 1
1 Съдържание 1
2 Въведение 2
2.1 Културно наследство 2
2.2 Цифровизация 2
2.3 Разпространение и Европеана 2
2.4 Принципи на агрегация 3
2.5 Бъдеще на Европеана 3
2.6 Принос на България в Европеана 3
2.7 Семантичен уеб 4
2.8 Linked Open Data 5
2.9 LOD за културно наследство 6
2.10 Семантичен уеб за културно наследство 6
2.11 Семантични данни за културно наследство 7
2.12 Пример за семантично публикуване 9
2.13 Примери за семантично анотиране 10
3 Цел на проекта 12
3.1 Разделение на проекта 12
3.2 П0: Подготовка на проекта 12

2 Въведение

2.1 Културно наследство

Културното наследство представлява голяма, хетерогенна и мулти-дисциплинарна област на човешката дейност. То включва:

- **Области**: културно, научно, инженерно, индустриално, архитектура, библиографско, етнографско, природно...
- **Артефакти** (богатства): исторически, археологически, религиозни, писмени, картини, музика, филми...
- **Институции**: музеи, библиотеки, галерии, архиви, аудио-визуални колекции, университети, научни институти...
- **Дейности**: проучване, намиране, придобиване, консервация , реставрация, научни изследвания, цифровизация, каталогизация, разпространение...

Българското културно наследство представлява значително национално богатство, което трябва да се запази за поколенията. Освен това то трябва и да се разпространява, както в България така и в обединена Европа. Това отговаря както на националните интереси на България, така и на целите и идеалите на Европа свързани с мултикултурализъм.

2.2 Цифровизация

Най-добрият и модерен начин за запазване на културното наследство е чрез цифровизация (дигитализация). Това включва:

- снимане или сканиране на културните богатства,
- запазването им в цифрово хранилище,
- добавяне на метаданни (вкл. библиографски, реставраторски, исторически, археологически, научни и пр),

2.3 Разпространение и Европеана

Разпространението на културното наследство често става чрез публикуването му на уеб сайт, със съответни инструменти за търсене.

- Един от най-добрите начини за разпространение е посредством включването (публикуването) в Европеана
- **Европеана** (www.europeana.eu) може да се нарече пан-Европейски цифров музей
- Мисията на Европеана е да позволи на хората да се насладят на цифровото наследство на европейските музеи, библиотеки, архиви, аудио-визуални колекции... И да инспирира хората чрез богатството на европейското културно и научно наследство
- Европеана спомага за научни изследвания и съвместна работа в областта на културното наследство, в една много-езична среда.
- Към лятото на 2011г, над 1600 културни институции са допринесли в колекцията на Европеана, която наброява 18 милиона експоната.

europeana
think culture

2.4 Принципи на агрегация

Европеана **централизира метаданните** (описанията) на всеки предоставен обект, както и едно малко иозображение (thumbnail). Така може да се осъществи търсене в цялата глобална колекция, и в бъдеще семантично свързване на обектите.

Самите цифрови обекти остават на уебсайта на доставчика. В този смисъл Европеана работи на принципа на **федерирано съдържание** (цифрови обекти).

За да се справи с логистичните проблеми при голям брой доставчици, Европеана работи с по-малък кръг **агрегатори**, които събират данни от доставчиците, помагат им да конвертират данните в нужния формат и комуникират с Европеана. Агрегаторите биват:

- регионални (напр. culture.fr обслужва Франция и е най-големият агрегатор),
- тематични (напр. Europeana Judaica за еврейско наследство),
- секторни (напр. European Library от европейски национални библиотеки)

Желан ефект от този проект е да установи технически агрегатор за България, който да може да помогне на всяка културна институция да публикува в Европеана.

2.5 Бъдеще на Европеана

Европейската Комисия е възприела политика за активно насърчаване на работата по Европеана. Докладът "Новият ренесанс"[1] на т.нар. "Комитет на мъдреците" ясно показва, че Комисията ще финансира проекти в тази насока с нарастващо темпо; и очертава важността на национални политики за принос към Европеана.

Желан страничен ефект от този проект е да създаде работна група и мрежа от професиналисти, които да работят заедно за дефинирането и осъществяването на такива проекти с международно финансиране.

2.6 Принос на България в Европеана

Приносът на България в Европеана до момента (лятото на 2011) е ограничен. България е допринесла 23 хиляди от 18 милиона обекта, или 0.13%:

- Варненска регионална библиотека[2]: 22,821
  - Текстове (20,618): стари вестници, стихосбирки, бюджетни документи...
  - Изображения (2,203): картини, стари снимки, стари пощенски картички
  - Доставчик: проектът Europeana Local
  - Много положителна черта е, че са включени материали не само на Варна, но и на други библиотеки и музеи в региона. Т.е. Варна се явява един вид агрегатор за тези по-малки институции
  - Изображенията са с добро качество.
  - Метаданните са с много добро качество
  - Някои от текстовете от по-малките библиотеки липсват, напр. от Старо-Загорската библиотека Родина[3] (сайтът не отговаря)
    Този и много други примери показват необходимостта от **централизирано съдържание** за някои по-малки доставчици

[1] http://ec.europa.eu/information_society/activities/digital_libraries/doc/executivesummery/bg.pdf

[2] http://catalog.libvar.bg/

[3] http://catalog.rodina-bg.org

- Национална библиотека: 196
  - Доставчик: The European Library
  - Включва някои ръкописи, стари вестници, картини.
  - Има множество проблеми:
  - Изображенията са с много ниско качество и не може да се различи нищо
  - Използваните URL[4] трудно могат да се нарекат "персистентни", което е основно изискване на Европеана
  - Сървърът връща следната грешка: 404 - File or directory not found. The resource you are looking for might have been removed, had its name changed, or is temporarily unavailable.
  - Метаданите са с много ниско качество, напр.
    - Description: "Александър Мутафов - Пред буря (1939)": правилно описание на обекта
    - Date: 1960 - не съответства на истинската дата (1939)
    - Title: "Национална художествена галерия" - това е хранилището, а не името на картината
    - Creator: "Въведение от Николай Шмиргела" - истинският автор е Александър Мутафов
- Централна библиотека на БАН: 144
  - Снимки на руски паметници в България, събрани от проекта "България и славяните 2 (Руснаци в България)"[5]
  - Метаданните са добри, но липсват географски координати
  - Снимките не са с много добра разделителна способност: надписите не се виждат.

Директната цел на този проект е да увеличи българския принос в Европеана

- Има и поне още 3 текущи европейски проекта с българско участие, които се надяваме да допринесат към Европеана.
- Евродепутатът Емил Стоянов е организирал 3 конференции за ролята на България в Европеана и е много активен в тази област
- Този проект трябва да осъществи контакт с Емил Стоянов и да се опита да установи съвместна работа постигане на общата цел

2.7 Семантичен уеб

Семантичният уеб е иновативна ИТ област:

- Предложена е от Тим Бернерс-Лий (изобретателят на уеба) около 1998
- Стандартизира се от WWW Consortium (W3C)
- Нарича се още "уеб 3.0", където уеб 1.0 е традиционният уеб, и уеб 2.0 е социалният уеб (интерактивни уеб приложения, Facebook, Twitter и пр)
- Ако уеб 1.0 може да се оприличи на един огромен хиперсвързан **документ**, то уеб 3.0 може да се оприличи на огромна хиперсвързана **база данни**

---

[4] Напр http://193.200.14.178/SampleFC/ALBYMIKA.000001/000/000/000/0000000001/f23.jpg

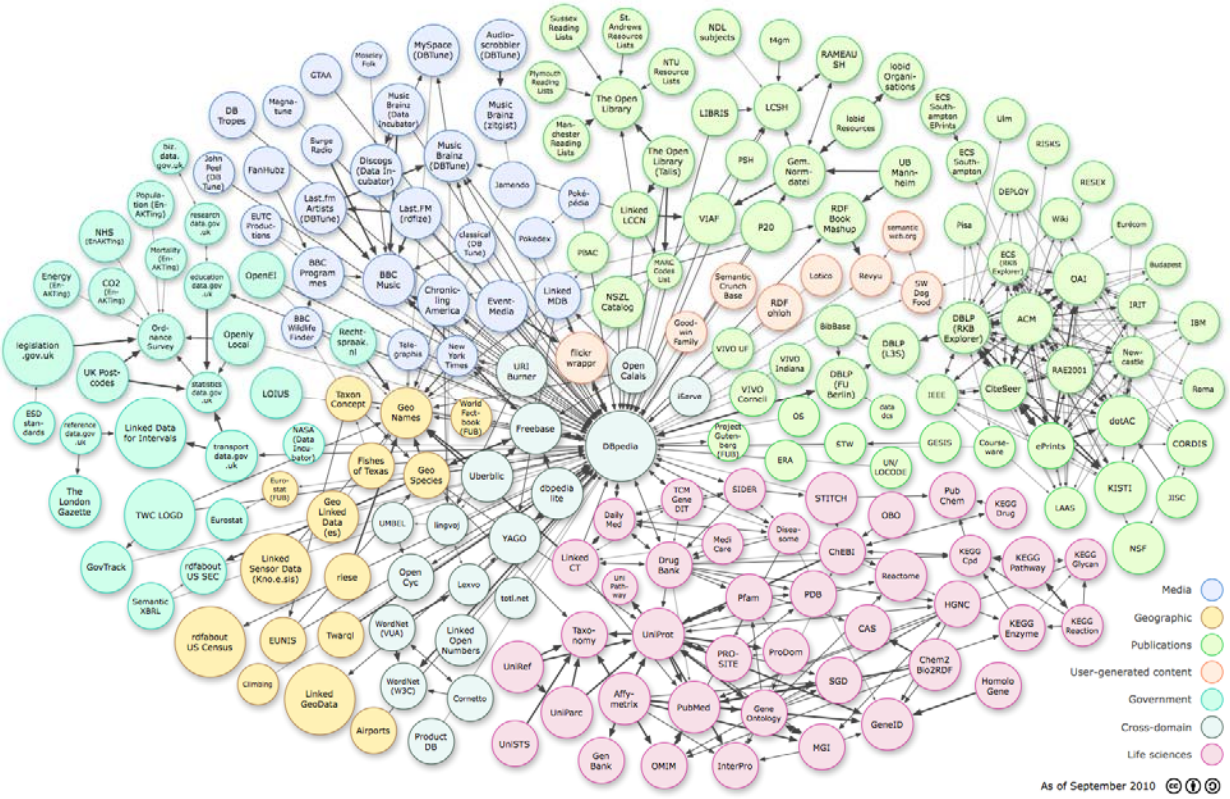
[5] http://rusmir.cl.bas.bg

През последните 5г семантичният уеб се очерта като най-перспективният подход за интеграция на данни:

- Използва максимално просто и унифицирано представяне на данните: RDF факти (тройки) "субект - предикат - обект"
- Метаданните (схемата) се представя по същия начин както данните, което позволява максимална гъвкавост на представяне и произволно много нива на метаданни
- Обектите и релациите (предикатите) получават глобално уникални идентификатори (URI) което позволява интеграция на данни чрез простото им поставяне на едно място
- Базира се на W3C стандарти (RDF, RDFS, OWL2, RIF) които позволяват логически извод (разсъждения) с различни нива на сложност, съобразени с конкретното приложение

2.8 Linked Open Data

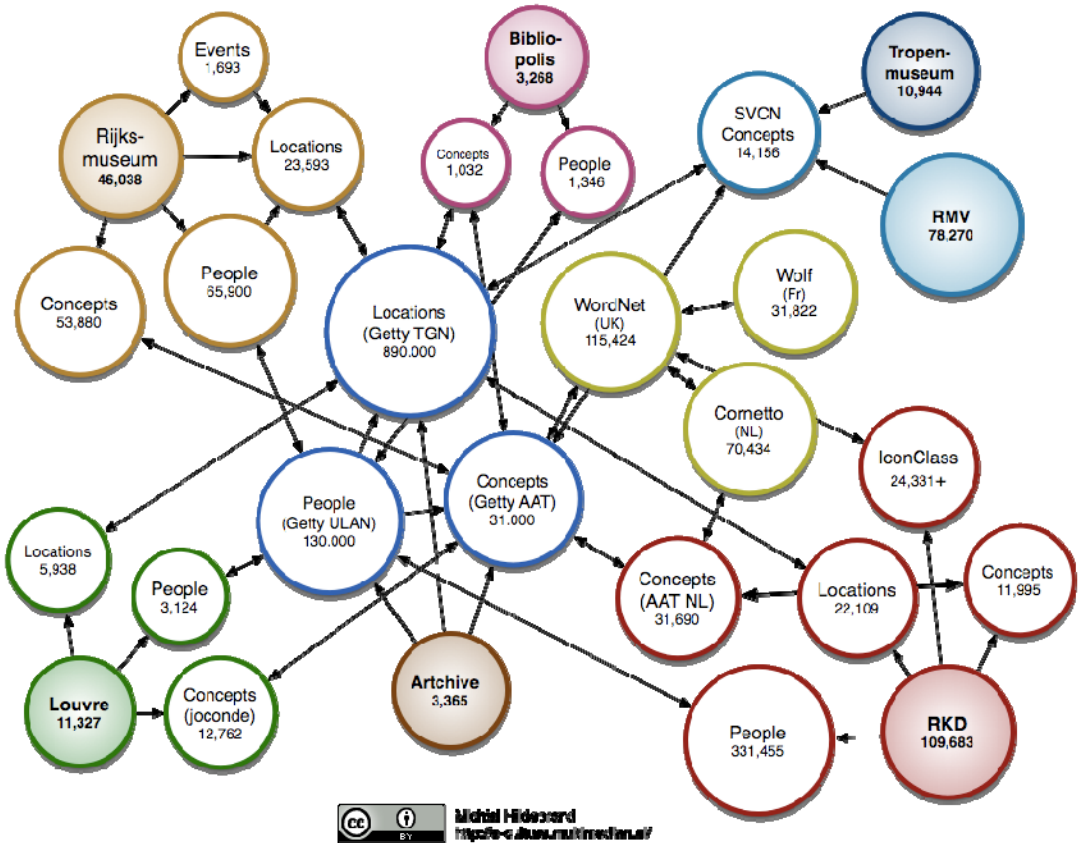
Linked Open Data (LOD) е множество от свободно достъпни множества от данни, които са представени в семантичен формат и са свързани помежду си. По-долу са показани различните множества, оцветени тематично:



- LOD расте експоненциално още от създаването си през 2008
- Към края на 2010 съдържа 207 множества данни и 28 **милиарда** факта в разнородни области: енциклопедични (dbpedia), география, политически, статистически, библиографски, музика, биохимични, генетични и пр.
- Съдържа много полезна фонова информация. Всяка нова информация може да се свърже с LOD, за да се обогати с такова контекстно знание

2.9 LOD за културно наследство

LOD за културно наследство се развива бурно в последната година. Един от основните фокуси до момента е свързване на терминологии (хора, места, концепции, материали, техники) разработени от различни организации, така че да може да се ползват глобално. Напр. по-долу са показани терминологии от фондация Getty, Rijksmuseum, RKD, Лувъра, WordNet и др.



2.10 Семантичен уеб за културно наследство

Много професионални области имат сходни проблеми с фрагментирани и трудни за управление данни. Областта Културно наследство е такава, тъй като данни за един и същ артефакт е разпиляна в множество фрагментирани бази данни:

- Каталожни данни
- Данни за придобиване
- Данни за реставрация и съхранение
- Научни данни (датиране, научни изследвания и пр)

По този начин се спъват изследвания, които биха могли да обхванат няколко институции или държави. Възможността свободно да се свързват артефакти, които се намират в различни музейни колекции, позволява принципно нови методи на изследване, и се цени много високо. Някои примери за такива връзки (виж илюстрации в секция 2.13):

- Търсене по състава на стоманата на определено оръжие (определен чрез научен анализ) и предполагаемата дата на производство, за да се намерят други оръжия от същата ковачница.

- Географско и времево картиране на местата, където са намерени такива оръжия
- Анализ на пътищата за разпространение и местата за използване, допринасящ към военната история.

Институции като Европеана и Британският музей твърдо се ориентират към семантичния уеб и LOD.

- Европеана възприема LOD за всички бъдещи разработки
- EDM е базиран на принципите на семантичния уеб, което го прави лесно разширяем
- Вече допринесените артефакти ще се конвертират към семантичен уеб. Първият Europeana LOD пилот е налице: въведение[6], стандартно описание[7], данни[8]

Първата концептуална статия на Европеана (White Paper) озаглавена **Knowledge = Information in Context** обяснява ключовата роля, която LOD ще играе в бъдещото развитие на Европеана.

- *"Свързаните данни дават на компютрите възможността да правят асоциации и да поставят ключови думи за търсене в техния контекст. Без свързани данни, Европеана ще е просто още една колекция от цифрови обекти. Със свързани данни, потенциалните приложения са много повече".*

Фондацията Andrew Mellon през 2009:

- *"Препоръката на фондацията е проектите да изберат RDF, като подход който увеличава шансовете, че учените ще могат да правят изследвания в каквито и нови посоки да се появят.*
- *Семантичните бази данни (RDF "triple stores") не поставят твърди ограничения върху моделите данни и позволяват моделите да бъдат променяни или дори подменяни с малки усилия, които възможности са важни поради отворения характер на научните изследвания"*

2.11 Семантични данни за културно наследство

Някои модели за представяне на културни данни включват следните:

- Dublin Core (DC)
  - Прости метаданни за описание на ресурси: автор, дата на създаване, актуализация и пр.
  - Стандарт ISO 15836
- Europeana Semantic Elements (ESE)
  - Структурира данните, които Европеана събира, управлява и публикува в момента
  - Сравнително беден модел, който представлява "най-малко общо кратно" за между множество съществуващи стандарти, бази данни за музейни колекции, институции
  - Моделира информационни обекти (метаданни, снимки, текстове) и културни обекти (напр. статуи, картини)
  - Не е базиран на семантичния уеб

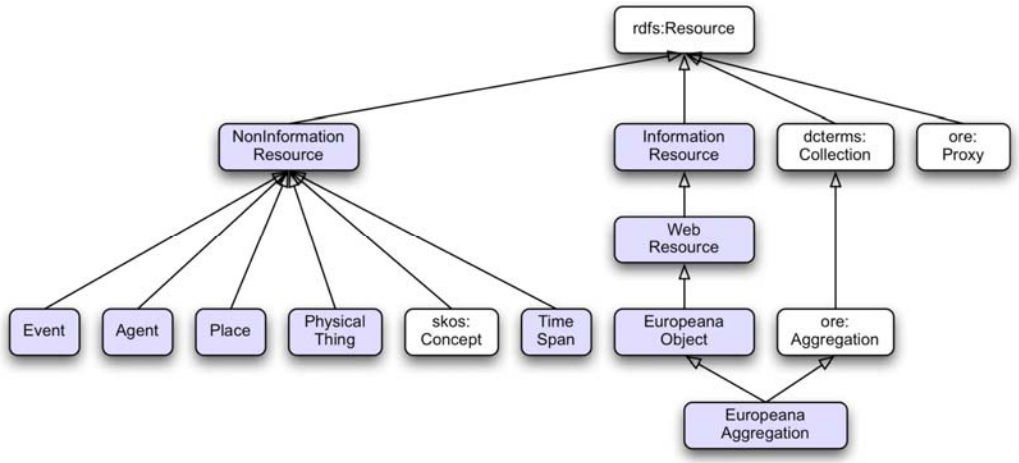
[6] http://version1.europeana.eu/web/lod/

[7] http://ckan.net/package/europeana-lod

[8] http://data.europeana.eu

    - "Плосък" модел без възможности за вложена информация и свързване между обектите
- Europeana Data Model (EDM)
    - Структурира данните, които Европеана ще управлява в бъдеще
    - Разработван до края на 2010, ще се внедрява до края на 2011
    - Отворена рамка, която се придържа към принципите за моделиране установени от семантичния уеб
    - Моделира **агрегации** следвайки модела OAI ORE. Агрегациите се използват, за да структурират обектите в контекст (от гледната точка на определена институция)
    - Освен метаданни и информационни обекти, моделира и "неинформационни обекти" (хора, места, периоди, събития и пр).
    - Позволява свързване между обекти
    - Покрива множество области и е предвиден да покрие вече установени стандарти като LIDO, EAD, METS
- Conceptual Reference Model (CRM)
    - Създаден от International Committee on Documentation (CIDOC) на International Council of Museums (ICOM), международна организация на музеи и музейни работници
    - Съдържа основни концепции от високо ниво, които са общи за различни културни организации
    - Например: данни за изработката на артефакта: място, дата и автор (ако е известен), от коя школа или последовател на кой майстор е автора (дори авторът да е неизвестен); данни за придобиването и собствеността; техники на изработка и мн.др
    - Стандарт ISO 21127 (2006) "Information and documentation — A reference ontology for the interchange of cultural heritage information"

По-долу са дадени класовете на EDM. Класовете в бяло са преизползвани от вече съществуващи онтологии, което е в пълно съгласие с принципите на семантичния уеб



Ето пример за EDM представянето на картината Мона Лиза от Леонардо, първо като обекти и после като събития

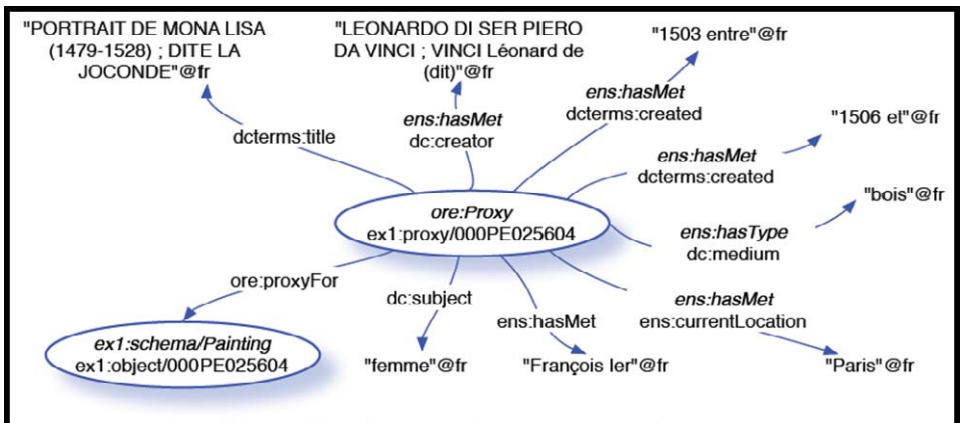


Fig. 11. Mona Lisa – an object-centric description

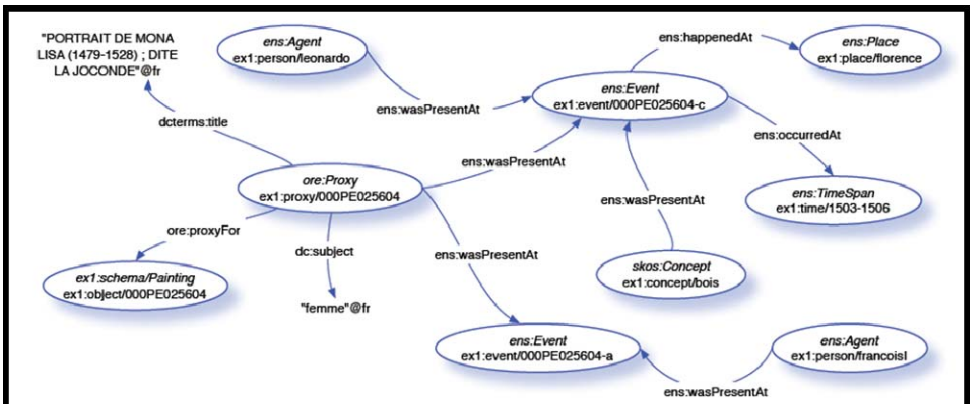


Fig. 12. Mona Lisa – an event-centric description

- EDM е обобщен стандарт, който представлява "най-малкото общо кратно"
- CRM е много по-богат като възможности
- В близко бъдеще е вероятно EDM да включи CRM, който вече е приет като " EDM профил на приложение" за следващото издание на Европеана.

2.12 Пример за семантично публикуване

(Линковете по-долу са активни)

LOD Experiment in Europeana Thought Lab:

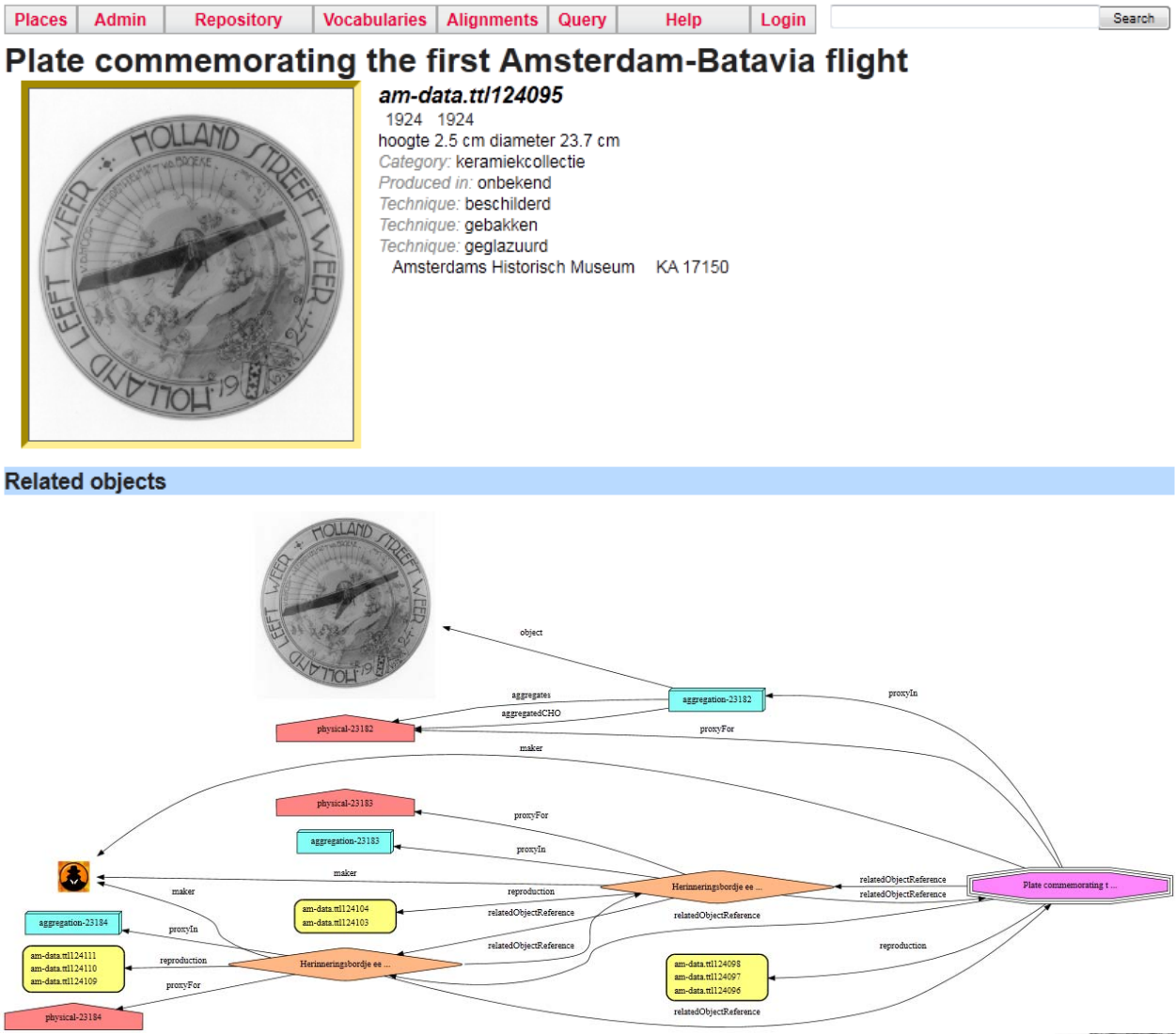
- Взети са данни и метаданни от Амстердамския музей, които описват 70 хиляди обекта, свързани с града
- Данните са извлечени чрез XML API от музейната колекция (от база данни на AdLib UK)
- Превърнати в семантичен формат: Converted to LOD compliant with EDM от Свободния университет на Амстердам (Vrije Universiteit, VU)
  - 5 милиона факта, т.е. средно по 70 факта за всеки обект
  - Софтуерът за конверсия е наличен като отворен код
- Данните са свързани със съществуващи отворени множества данни:
  - Холандски тезаурус на изкуствата и архитектурата: Dutch Art and Architecture Thesaurus (AATNed)

- Обединен списък на имена на творци на галериите Гети: Getty's Union List of Artists Names (ULAN)
- LOD версията на онлайн енциклопедията Wikipedia: DBpedia
- Географски наименования и тяхната йерархия: GeoNames

- Публикуван като източник в Comprehensive Knowledge Archives Network (CKAN), за да може да се открие от всеки заинтересован

Това е отличен пример за LOD от типа, който Европеана ще създава и промотира.

Пример за изглед на един музеен обект, както и свързани с него данни:

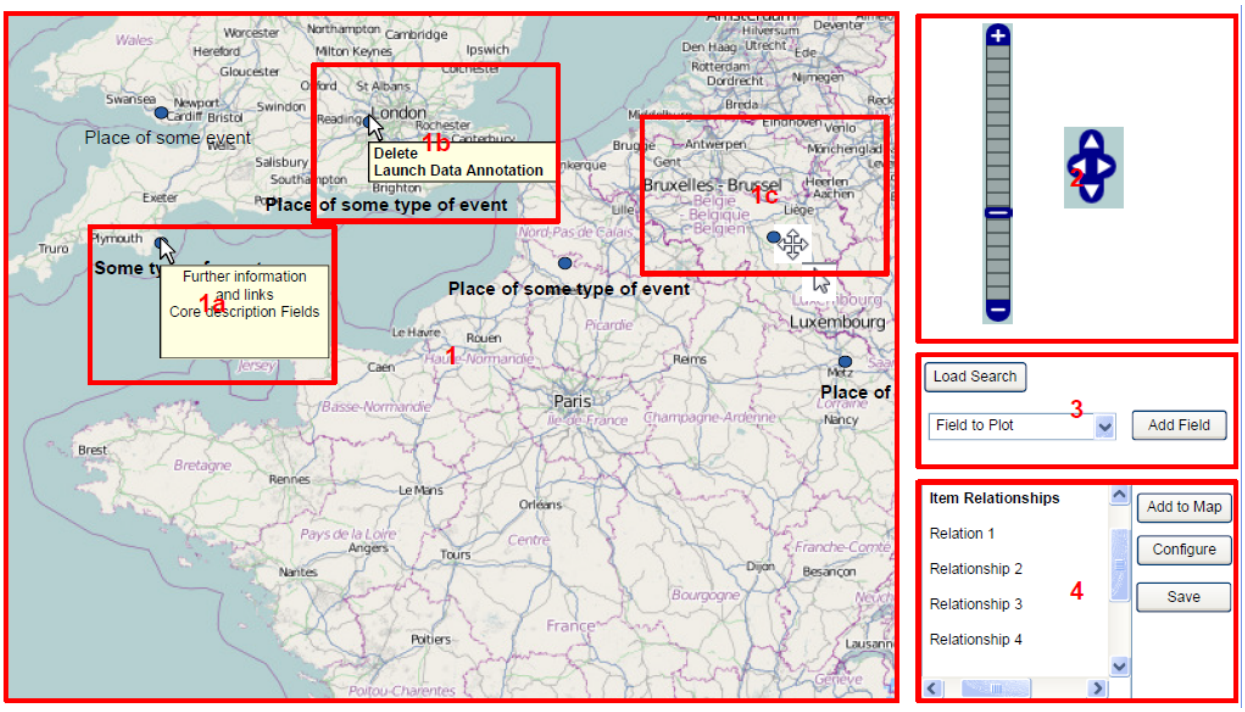


Изборът на произволен възел показва неговите данни и връзки, напр.

- Този плакет е изработен от "Heytze, J.C."
- Свързан е с друг плакет (показан вдясно), и пр.

2.13 Примери за семантично анотиране

Концептуален пример за географско картиране и анотация (от проекта www.ResearchSpace.org на Британския музей):



Пример за географска анотация на старинна карта (от проекта EuropeanaConnect Annotation, http://dme.arcs.ac.at/annotation).

- Добавянето на старинно име към картата (Стълбовете на Херкулес) намира съвременни свързани имена от GeoNames (част от LOD) като Гибралтар, Испания, Мароко, Средиземно море и позволява свързването им към анотираното понятие.
- Освен това се позволява свързване по името (Херкулес) към други артефакти

3 Цел на проекта

Основната цел на проекта е да цифровизира данни за избрано българско културно наследство, да ги конвертира в семантичен вид и да ги публикува в Европеана.

Спираме се на две от най-разпознаваемите български наследства:

- Кирило-Методиевски ръкописи, псалтири и пр
- Тракийски златни съкровища, гробници, светилища

Допълнителни цели:

- Установяване на процес за публикуване, който да е лесен за използване от други културни институции
- Създаване на група (Special Interest Group, SIG) за публикуване в Европеана, координация с други съмишленици
- Установяване на технически агрегатор за България (www.bulgariana.eu)

3.1 Разделение на проекта

Целият проект ще бъде разделен на следните координирани под-проекти, които ще се реализират с отделни договори.

- П0: Подготовка
- П1: Цифровизирано Кирило-Методиевско наследство
- П2: Цифровизирано Тракийско наследство
- П3: Хранилище на цифрови обекти, миграция на данни към него, публикуване в уеб
- П4: Конверсия, семантични връзки, публикуване в Европеана. Цялостна координация. Реализация на допълнителните цели.

Малка сума ще бъде заделена и за хардуер и хостинг на хранилището.

П0 е описан в този документ, а останалите под-проекти в отделни документи.

3.2 П0: Подготовка на проекта

Подготовката на проекта е извършена и включва:

- Разговори с доставчици на съдържание, създаване концепция на проекта
- Предпроектно проучване:
  - Начини на публикуване в Европеана, семантични формати данни, ESE спрямо EDM
  - Налични цифрови хранилища с отворен код, критерии за избор
  - Модели на избраните библиотеки
  - Начини на представяне на информацията (визуализация)
  - Възможности за свързване с LOD и терминологии
- Разработка на задание
- Създаване на предварителен проектен план (разбивка на работата)
- Оценяване на необходимите ресурси (труд, хардуер, софтуер)
- Подготовка на тръжна документация

Европеана П1: Цифровизиране на Кирило-Методиевско наследство

Създаден: 15 юни 2011
Редактиран: 06 окт 2011

1 Съдържание

Европеана П1: Цифровизирано Кирило-Методиевско наследство ..... **Error! Bookmark not defined.**
1 Съдържание 1
2 Въведение 1
2.1 Контрагент 1
2.2 Дейности на контрагента 1
2.3 КМНЦ Административна Система 2
3 Описание на работата 4
3.1 КМНЦ Метаданни 4
3.2 КМНЦ обекти за публикуване в Европеана 6
3.3 Собствен принос на КМНЦ 6
3.4 Нива на достъп до КМНЦ материали 7
3.5 Финансова калкулация 7
4 Уточняване на работната програма 8

2 Въведение

2.1 Контрагент

Кирило-Методиевски Научен Център (КМНЦ) при БАН

ул. Московска №13, етаж 2

Директор: д-р Светлина Николова

Емайл: nikolova@bas.bg, svetlina.nikolova@gmail.com

Телефон: 0886765084

2.2 Дейности на контрагента

КМНЦ има за задача да изследва и запази културното наследство на Солунските братя и техните ученици и последователи.

- КМНЦ е цифровизирал около 1052 документа, основно древни ръкописни книги, например псалтири.

- Сканирани са общо около 100 хиляди страници (63 хиляди "кадъра" т.е. двойни страници).
- Попълнени са около 15 полета метаданни за всеки артефакт
- Освен това са създадени около 8000 библиографски записа за артефакти, които се намират извън България (напр. в Русия)

Библиотеката все още не е публикувана в уеб поради липса на средства. Затова публикацията й в уеб и Европеана ще има значителен ефект.

2.3 КМНЦ Административна Система

Създадена е административна система:

- Разработена е на MS.NET и използва база данни Mysql
- Функциите включват въвеждане на данни, търсене, разглеждане на изображенията
- Също включва контрол и статистика на въвеждането

Статистика (само една част от системата)

Статистика

Общо документи		
Брой документи	1052	
Брой кадри	62774	
Попълнени документи	291	27,66%
Непопълнени документи	761	72,34%
Документи БАН		
Брой документи	690	
Брой кадри	26006	
Попълнени документи	113	16,38%
Непопълнени документи	577	83,62%
Документи КМНЦ		
Брой документи	356	
Брой кадри	36419	
Попълнени документи	177	49,72%
Непопълнени документи	179	50,28%

Списък ръкописи:

Статистика | Научен архив на БАН | Кирило-Методиевски научен център | Кодов | Библиография | Потребители

Документи от Кирило-Методиевски научен център

Документи КМНЦ		
Брой документи	356	
Брой кадри	36419	
Попълнени документи	177	49,72%
Непопълнени документи	179	50,28%

		Документ	Кадри	Тип на документа	Сигнатура	Действия
1	✓	KIvanova_M_CodSlav32	99	Копие на ръкопис		
2	✓	KIvanova_M_Dr700	409	Копие на ръкопис		
3	✓	KIvanova_M_Dr706	482	Копие на ръкопис		
4	✓	KIvanova_M_Dr724	301	Копие на ръкопис		
5	✓	KIvanova_M_Dr739	298	Копие на ръкопис		
6	✓	KIvanova_M_Ehrhard	371	Печатно издание (Книга)		
7	✓	KIvanova_M_H382	266	Копие на ръкопис		

Разглеждане на един ръкопис:

Документи от научния архив на БАН > NABAN_10_1

Метаданни за документ

В тази страница вие редактирате метаданните за избрания документ. Първите стъпка е попълване на общите метаданни.Тези метаданни ще бъдат използвани по подразбиране за всяка една страница от документа.Като следваща стъпка ще имате възможност да редактирате метаданните за всяка отделна страница от документа.

Име	NABAN_10_1
Тип	Ръкопис
Сигнатура	BAN_SB10_MANUSCRIPT

Документ | Страници на документа (1-3 от 110) | 1 | Отиди

NABAN_10_1 Текущ | стр. 1 Избери | стр. 2 Избери | стр. 3 Избери

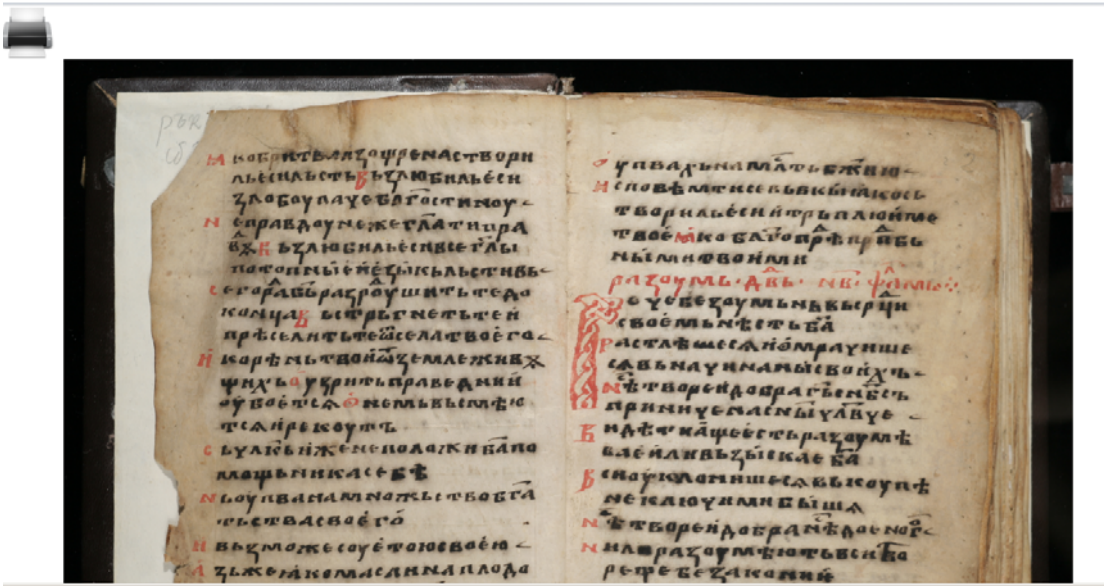
Редакция на метаданни за документ

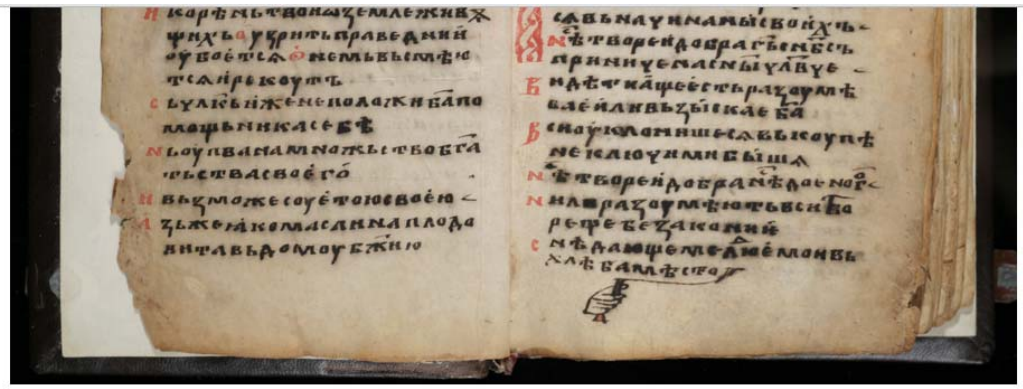
Колекция:
Научен архив на БАН

Сигнатура:
№ 1

Название на ръкописа:

Разглеждане на една страница:





Колекция:

Научен архив на БАН

3 Описание на работата

3.1 КМНЦ Метаданни

Колекцията от цифровизирани средновековни писмени паметници на КМНЦ съдържа три типа документи: ръкопис, копие на ръкопис и архив. Всеки тип документ се описва с определени метаданни, изброени по-долу. (Метаданните ще бъдат уточнени след мапинг към EDM).

Ръкопис

- Име
- ID на документа
- Тип: Ръкопис
- Сигнатура
- Кадри (брой снимки)
- Документ
- Страници от документа

- Колекция
- Сигнатура
- Название на ръкописа
- Датировка
- Материал
- Обем
- Подвързия
- Тип писмо
- Украса
- Бележки
- Съдържание
- Правопис
- Придобиване на ръкописа
- Пагинация
- Опис
- Фрагмент (да/не)
- Връзки с други материали на електронното хранилище
- Допълнителни бележки

Копие на ръкопис

- Име
- ID на документа
- Тип: Копие на ръкопис
- Сигнатура
- Документ
- Страници от документа
- Колекция
- Инвентарен №
- Название на ръкописа
- Датировка
- Обем
- Тип писмо
- Украса
- Бележки
- Съдържание
- Правопис
- Придобиване на ръкописа
- Пагинация
- Библиография
- Фрагмент (да/не)
- Връзки с други материали на електронното хранилище
- Книгохранилище на оригинала
- Оригинална сигнатура
- Допълнителни бележки

Архив

- Име
- ID на документа
- Тип: Архив
- Сигнатура
- Документ
- Страници на документа
- Фонд
- Опис
- Архивна единица
- Лист
- Автор
- Вид на документа
- Тип на документа
- Датировка
- Публикуван (да/не)
- Съдържание
- Допълнителни бележки (незадължително поле)

3.2 КМНЦ обекти за публикуване в Европеана

КМНЦ ще предостави следните материали за публикуване в Европеана:

- Ръкописна сбирка от 118 средновековни български ръкописи от XI до XVIII в., които вече са дигитализирани. Общият брой на заснетите кадри е 18000.
- Дигитализирани печатни издания на КМНЦ:
  - Списание Palaeobulgarica. Направени са дигитални копия на 104 книжки (от кн. 1 на 1977 г. до кн. 4 на 2002 г.). Общият брой на дигитализираните кадри е над 5200 кадъра;
  - Поредицата „Кирило-Методиевски студии“. Направени са дигитални копия на том 1–15 с общ обем на дигитализираните кадри 2245;
  - В/ Отделни книги, издадени от КМНЦ. Направени са дигитални копия на 10 книги. Общият брой на дигитализираните кадри е 1100.

  Общ брой на дигитализираните кадри за печатните издания: около 8545.
- Дигитализиране на нови документи и въвеждане на метаданните за тях: архивни материали от XIX и XX в., посветени на кирило-методиевското културно наследство. Предвижда се допълнително да се дигитализират 350 документа с общ обем на страниците 7000 стр., т.е. 3500 кадъра.

3.3 Собствен принос на КМНЦ

Дигитализирани 118 ръкописа и попълнени метаданни за тях. Заснети са 18000 кадъра, за което КМНЦ е заплатил 18000 лв.

За дигитализация на печатните издания на КМНЦ - заснети са над 8500 кадъра, за което КМНЦ е заплатил 8500 лв.

За попълване на метаданни (за труд) - около 3500 лв.

3.4 Нива на достъп до КМНЦ материали

Преценката на КМНЦ е да предостави частичен достъп до материалите, тъй като някои от тях са все още в печат и в продажба. В бъдещ проект ще се търси възможност за монетизация на цифровизираните материали (т.е. разплащателна система и система за достъп).

- **Ръкописи** – частичен достъп до всичките 118 ръкописа – да се показват по 5–6 кадъра от всеки ръкопис.
- **Печатни издания**:
  - **Palaeobulgarica** – достъп до титулните страници и съдържанието на всичките 104 книжки, както и до първата страница от всяка статия в тях. Което означава частичен достъп до около 1300 статии (документа), публикувани в тези 104 книжки.
  - **Кирило-Методиевски студии –** достъп до съдържанието на всеки от 15-те дигитализирани тома на поредицата и до началната страница на всеки от документите в тях (монографии, студии, статии). Това означава частичен достъп до над 200 документа.
- **Отделни книги -** достъп до съдържанието и до началната страница.

3.5 Финансова калкулация

Разходи за дигитализация на новите архивни документи (XIX и XX в.), посветени на кирило-методиевското културно наследство:

- За дигитализиране на 350 документа с общ обем 7000 страници, 3500 кадъра – 3500 лв.
- Разходи за труд за попълване на метаданни:
  - Въвеждане на метаданни за един материал: 10 часа (много от материалите съдържат по 10-15 статии)
  - Контрол на качеството на метаданните за един материал: 5 часа
  - За попълване на метаданни на дигитализираните нови архивни документи: Назначаване на двама млади учени (? на граждански договор) на 6-часов работен ден за 4 месеца по 450 лв. месечно възнаграждение = 3600 лв.
  - За попълване на метаданни на вече дигитализираните материали:
    - За печатните издания: списание Palaeobulgarica (попълване на метаданни за 20 книжки, 1500 кадъра); поредицата Кирило-Методиевски студии (попълване на метаданни за 10 тома, 1500 кадъра); отделни книги, издания на КМНЦ (попълване на метаданни за 10 отделни книги, 1100 кадъра). 3 сътрудника на 4-часов работен ден за 3.5 месеца по 200 лева месечно брутно възнаграждение на човек = 2100 лв.
    - За попълване на метаданни на средновековни български ръкописи от XI до XVIII в. – за 20 документа (1000 кадъра) – 1 сътрудник на 4-часов работен ден за 3.5 месеца по 200 лв. = 700 лв.
- Материали и консумативи - 2000 лв.
- Отчисления за базовата организация
  - Управление на проекта: 1000 лв
  - Инструктаж, помощ и координация на работата на временно наетите сътрудници: 1000 лв

    - Управление и контрол на качеството: 1000лв
- Общо: 14 900 лв.

4 Уточняване на работната програма

Следните детайли ще бъдат уточнени по време на проекта:

- Оптимизация на графика според времевите ограничения на проекта и възможностите за привличане на допълнителни сътрудници
- Оптимизиране на набора материали за публикация
- Нива на достъп: ако може древните ръкописи да са достъпни изцяло

Софийски Университет
Българо-Корейски Център за ИТ кооперация
(ITCC)

Европеана П2: Цифровизиране на Тракийско наследство

Създаден: 20 August 2011
Редактиран: 31 Oct 2011

1 Съдържание

Европеана П2: Цифровизиране на Тракийско наследство ... 1
1 Съдържание ... 1
2 Въведение ... 2
2.1 Контрагент ... 2
2.2 Дейност и собствен принос на Центъра по Тракология ... 2
3 Задачи ... 5
3.1 Метаданни за тракийско наследство ... 5
3.2 Траколожки обекти за публикуване в Европеана ... 5
3.2.1 Сложни предмети и обекти ... 6
3.2.2 Обикновени предмети ... 6
3.2.3 Допълнителни източници ... 6
3.3 Разбивка на работата ... 7
4 Уточняване на работната програма ... 7

2 Въведение

Българските земи са богати със съкровища, надгробни могили и религиозни археологически обекти, останали от древни времена: от халколитната Култура Варна (най-старото технологично обработено злато в света) до тракийските златни съкровища.

2.1 Контрагент

Център по Тракология "Проф. Александър Фол"

Институт по Балканистика на БАН

София 1000, ул. Московска 45

Директор: доц. д-р Александър Костов, kostov_al@yahoo.com, 0887 520 662, 0886 567 879

Сътрудник: проф. д-р Валерия Фол, valeriafol@gmail.com, 0888873443

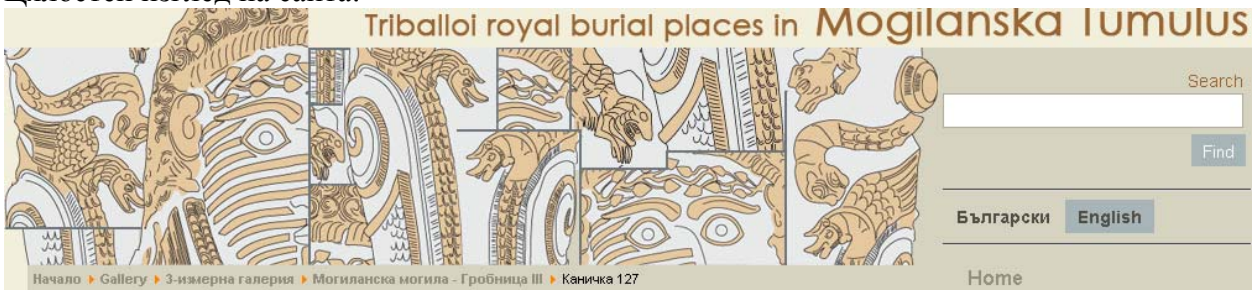
2.2 Дейност и собствен принос на Центъра по Тракология

Центърът по Тракология изследва, дигитализира и документира българските халколитни и тракийски съкровища. Центърът и Фондация Трацика са разработили няколко уебсайта с цифровизирано наследство, които ще бъдат използвани:

- Енциклопедия Древна Тракия (http://www.thracians.net)
- Трибалски царски гробници – Могиланска могила (http://tribaloi.thracians.net)
- Проф.д-р. Валерия Фол притежава множество високо-качествени снимки, които е направила сама и част от които ще бъдат използвани.

Следват няколко илюстрации от сайта за Могиланската могила:

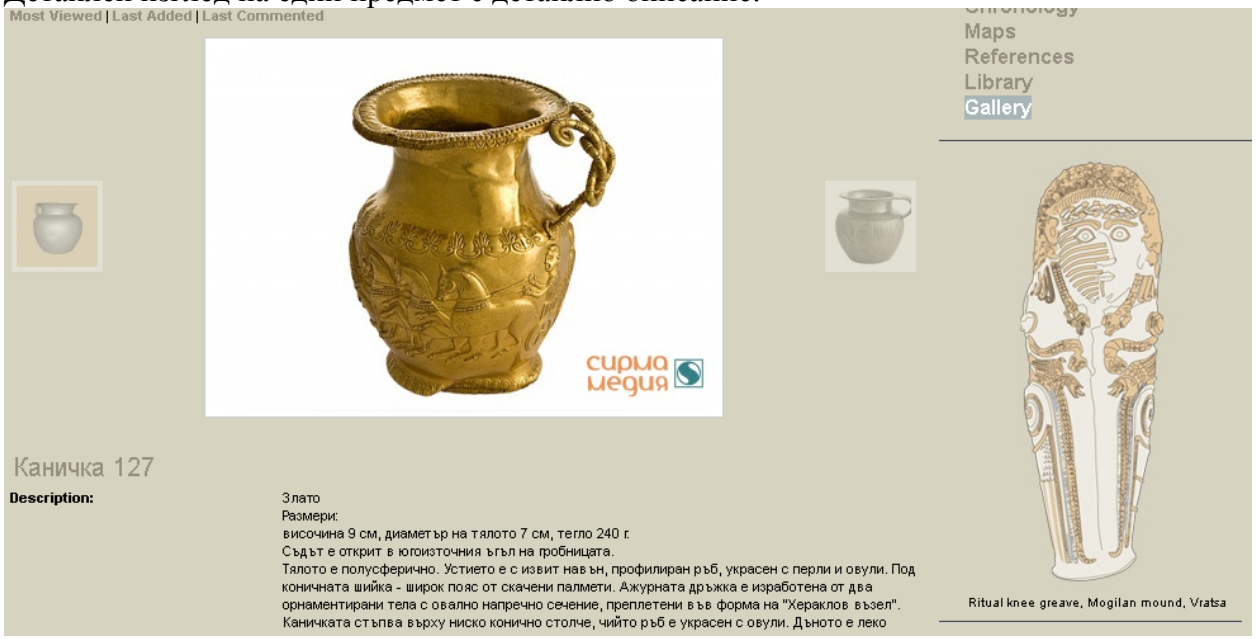
- Цялостен изглед на сайта:



- Мултимедиен модел на могилата. Цялостен изглед отгоре:

- Увеличаване на изгледа към Гробница 2 и единия предмет намерен в нея (наколенник)

- Детайлен изглед на един предмет с детайлно описание:



- Двете енциклопедии включват и тримерни модели на някои предмети, включително и горната каничка[1]

[1] http://www.thracians.net/index.php?option=com_datsogallery&Itemid=76&func=detail&id=49

3 Задачи

3.1 Метаданни за тракийско наследство

Предлага се да се въвеждат следните метаданни. (О) маркира опционални метаданни (въвеждат се когато са налични и приложими за обекта), всички останали метаданни са задължителни.

- За предмети:
  - Наименование на предмета;
  - Материал;
  - Размери;
  - (О) Тегло (за предмети от благороден метал)
  - Датировка;
  - Местонамиране;
  - Описание;
  - Съхранение – музей, инв. №;
  - Снимки и други цифровизирани материали; Брой
  - Публикация/и (Библиография; да/не)
  - Авторски права
  - (О) Връзка с други материали на електронното хранилище, напр. Предмет е част от Съкровище (обект) (когато е приложимо)
  - (О) Връзки със световни източници на отворени данни (Linked Open Data, LOD), напр. места, хора (когато е приложимо)
- За археологически обекти:
  - Наименование на обекта
  - Местонахождение (държава, село, град, район; GPS координати)
  - Материал (скала, камък, кирпич, дървена конструкция, пръст и пр.)
  - (О) Размери (отнася се само за някои обекти)
  - Описание
  - Брой снимки
  - Публикация/и (Библиография; да/не)
  - Авторски права
  - (О) Връзка с други материали на електронното хранилище, напр. Гробница 2 е част от Могиланската могила (когато е приложимо)
  - (О) Връзки с LOD (когато е приложимо)

Метаданните ще бъдат допълнени след:

- Определяне на данните необходими за Европеана
- Определяне на LOD, които могат да се използват: например GeoNames или Getty Thesaurus of Geographic Names за географски места

3.2 Траколожки обекти за публикуване в Европеана

Предлагаме да се въведат 300 единици (предмети и обекти), които имат две степени на сложност

- "Сложни" наричаме предмети и обекти, които:

- Позволяват семантични връзки с:
  - хора (богове, исторически личности, херои),
  - животни (които са митологически персонажи или зооморфни образи на богове, елемент на жертвоприносителна обредност),
  - архитектура,
  - природа,
  - атрибути (въоръжение, обредни предмети и др.),
- Предмети, които са част от съкровище
- Предмети с надписи;
- Обекти, които са част от комплекси и/или с богата украса
- Изискват множество снимки за адекватно представяне

- "Обикновени" са предмети и обекти с просто описание, за които е достатъчна една снимка

3.2.1 Сложни предмети и обекти

- Вълчитрънско съкровище – 20 сннмки
- Панагюрско съкровище - 30 снимки
- Боровско съкровище – 10 снимки
- Съкровище от Летница – 25 снимки
- Варненски некропол – 20 предмета, които позволяват семантически връзки; 20 снимки
- Култова сцена от Овчарово – 25 снимки
- Казанлъшка гробница (паметник на ЮНЕСКО) – 15 снимки
- Гробницата в Свещари (Паметник на ЮНЕСКО) – 10 снимки
- Гробница в Александрово – 15 снимки
- 5 светилища и мегалита – 20 снимки
- 3 гробници-хероони – 10 снимки

Общо – 200 снимки

3.2.2 Обикновени предмети

- 300 предмета от халколита до IV век сл. Хр. – 300 снимки

Всички снимки на Сложни и Обикновени предмети ще са с професионално качество. Разделителната способност и форматът на цифровите изображения ще са съобразени със следните фактори:

- Форматът трябва да е подходящ, за да не пострада качеството
- Достатъчна разделителната способност, за да не пострада качеството
- Големината на файловете не трябва да е прекалено голяма, защото предназначението е публикация в уеб сайт

Разделителната способност ще е поне 2048x1024 пиксела

3.2.3 Допълнителни източници

Ще се проучи възможността да се използват обекти от следните сайтове:

- http://ancient-treasure.info.
- http://thracian.info (може би е същото съдържание)

3.3 Разбивка на работата

Разходите за дигитализация и въвеждане на метаданни за обектите възлизат на общо: 14960 лв и включват:

- Авторски права за фотографиите
- Въвеждане на метаданни и авторски права за описанията
  - 200 изображения с описание от категория сложни
  - 295 изображения с описание от категория обикновени
- Материали и консумативи
- Отчисления за базовата организация
  - Управление на проекта, административно ръководство, сключване на граждански договори, счетоводно обслужване
  - Методологическо ръководство, инструктаж, помощ и координация на работата на временно наетите сътрудници
  - Управление и контрол на данните, редакция
- .

4 Уточняване на работната програма

Следните детайли ще бъдат уточнени по време на проекта:

- Оптимизация на графика според времевите ограничения на проекта и възможностите за привличане на допълнителни сътрудници
- Оптимизиране на набора материали за публикация

Софийски Университет
Българо-Корейски Център за ИТ кооперация
(ITCC)

Европеана П3: Хранилище на цифрови обекти

Създаден: 15 юни 2011
Редактиран: 06 окт 2011

1 Съдържание

Европеана П3: Хранилище на цифрови обекти ... 1
1 Съдържание ... 1
2 Въведение ... 2
2.1 Контрагент ... 2
2.2 Дейности на контрагента ... 2
2.3 Принципи на агрегация (централизиране и федериране) ... 2
2.4 Цифрово хранилище ... 3
2.5 Технически агрегатор за България ... 3
3 Задачи ... 4
3.1 Финансова калкулация ... 4
4 Сравнителен анализ ... 4
4.1 DSpace ... 5
4.2 CDS Invenio ... 6
4.3 Nuxeo DAM/DM ... 7
4.4 Alfresco ... 8
4.5 REPOX ... 9
4.6 Изводи ... 9

2 Въведение

2.1 Контрагент

Сирма Медия АД, http://www.sirmamedia.com/

1784 София, бул. "Цариградско шосе" № 135, ет. 3

Изпълнителен директор: Илиян Узунов, ilian@sirma.bg

Тел.: (02) 976 8304, (0888) 772 248

2.2 Дейности на контрагента

Сирма Медия е първото българско издателство на мултимедийни електронни знания – енциклопедии, речници, образователни игри, уеб базирани бази знания, инструменти за работа на български език, дигитални хранилища, системи за управление на процеса на учене и т.н.

Дейността на Сирма Медия започва през 2002 г. с издаването на първата електронна Българска енциклопедия „А-Я”, съвместно с Българска академия на науките и Книгоиздателска къща „Труд”. Издаването на мултимедийни знания продължава с нарастващ обхват и дълбочина.

Сирма Медия се развива в две основни бизнес направления

- ОБРАЗОВАНИЕ: мултимедийни образователни продукти; специализирани уеб приложения – уеб портали, семантични хранилища, системи за управление на образователния процес; собствени методики за преподаване и курсове за обучение и квалификация на учители; специализирани образователни хардуерни решения
- КУЛТУРНО-ИСТОРИЧЕСКО НАСЛЕДСТВО: дигитализация и съвременно представяне на културно-историческо наследство, дигитални хранилища на артефакти, уеб базирани информационни системи за управление, каталогизиране и представяне на културно-историческо наследство.

Мултимедийните енциклопедични продукти на Сирма Медия (образователен пакет „Продукти на знанието”), заедно с познавателния и образователен уеб портал www.znam.bg, както и семантичното хранилище на образователни ресурси са достъпни във всички български училища и се използват активно за нуждите на образователния процес.

Сирма Медия е партньор с големи световни брандове, притежаващи силни позиции в сферата на образованието като: Intel, National Geographic, Young Digital Planet, Luidia и много други.

Амбицията на Сирма Медия е да бъде определяща, движеща сила в създаването на система от принципи, стандарти и сътрудничещи си организации за постепенното изграждане на национална мрежа от знания като част от развитието ни към „общество на знанието”.

2.3 Принципи на агрегация (централизиране и федериране)

Европеана **централизира метаданните** (описанията) на всеки предоставен обект, както и едно малко изображение (thumbnail). Така може да се осъществи търсене в цялата глобална колекция, и в бъдеще семантично свързване на обектите.

Самите цифрови обекти се намират на уебсайта на доставчика. В този смисъл Европеана работи на принципа на **федерирано съдържание** (цифрови обекти).

2.4 Цифрово хранилище

Целта на проекта е да създаде цифрово хранилище за културни институции, които искат да публикуват Българско културно наследство. Тази система трябва да поддържа:

- Множество потребители и организации със съответните права за достъп
- Възможност за създаване на типове обекти и настройване на метаданните за всеки тип.
- Лесно попълване на метаданни от музейни работници, които не са ИТ специалисти
- Поддръжка на колекции от обекти
- Публикуване на цифровите обекти в уеб
- Търсене на обекти
- Визуализация на обекти по разнообразни начини, включително преглед на книги (напр. чрез PDF), прелистване на изображения, тримерни изображения и пр.
- Лесно създаване на "изложби

2.5 Технически агрегатор за България

Нашите наблюдения показват, че множество от по-малките културни институции в България нямат установени цифрови хранилища:

- Някои нямат никаква система за пазене на цифрово съдържание
- Други имат просто уебсайт, с който е трудно и времеемко да се поддържа значителна колекция
- Хранилищата или уебсайтовете на трети не се поддържат регулярно и не дават гаранции за работоспособност и достъпност. Това включва проблеми както със софтуера, така и с хардуера или инфраструктурата (електрическо захранване, свързаност и пр)

Примери за такива проблеми се срещат дори за обекти, публикувани в Европеана:

- Текстове от Старо-Загорската библиотека Родина[1] включени чрез доставчика Варненска регионална библиотека чрез проекта Europeana Local, но сайтът не отговаря
- Обекти на Националната библиотека включени чрез European Library не се намират. Тук проблемът е софтуерен, тъй като цифровата библиотека на Националната библиотека работи.
    - Използваните URL[2] трудно могат да се нарекат "персистентни", което е основно изискване на Европеана
    - Сървърът връща следната грешка: 404 - File or directory not found. The resource you are looking for might have been removed, had its name changed, or is temporarily unavailable.

Този и много други примери показват необходимостта от **централизиране на съдържанието** за по-малките доставчици.

Желан ефект от този проект е да установи технически агрегатор за България, който да може да помогне на всяка културна институция да публикува в Европеана. Създаването и

[1] http://catalog.rodina-bg.org

[2] Напр http://193.200.14.178/SampleFC/ALBYMIKA.000001/000/000/000/0000000001/f23.jpg

поддръжката на цифрово хранилище, което може да се използва от множество малки доставчици е съществена стъпка в тази насока.

3 Задачи

Задачите на този под-проект включват:

- Избор на цифрово хранилище (виж следващата секция)
- Инсталация и настройване на хранилището
- Създаване на акаунти за различните организации и потребители, настройване на правата за достъп
- Създаване на необходимите типове обекти, напр. ръкопис, статия, списание; исторически предмет, археологически обект
- Настройване на метаданни за различните обекти. Метаданните трябва да са координирани с изискванията на конкретната институция и данни, необходими за Европеана
- Събиране на номенклатурни списъци (терминологичен речник). Свързване на избраните термини с глобални идентификатори (URI), за да може да се търси между данни публикувани от различни институции и дори държави
- Миграция на съществуващи данни от избрани културни институции към хранилището.
- Курс за обучение на сътрудници (щатни или извън-щатни) от избрани културни институции, които ще въвеждат данни в хранилището
- Публикуване на колекциите в уеб, включително настройване и реализиране на търсене

3.1 Финансова калкулация

- Инсталация и настройване на хранилището: 1000 лв
- Създаване на акаунти, настройване на права за достъп: 1000 лв
- Създаване на необходимите типове обекти: 1000 лв
- Настройване на метаданни за различните обекти: 2000 лв
- Събиране на номенклатурни списъци: 1000 лв
- Миграция на съществуващи данни: 3000 лв
- Курс за обучение на избрани институции: 950 лв
- Публикуване на колекциите в уеб: 5000 лв

Общо: 14 950 лв

4 Сравнителен анализ

Тази секция представя сравнителен анализ на софтуерни средства с отворен код, предназначени за създаване на цифрови хранилища. По-детайлни сведения за особеностите на разгледаните софтуерни инструменти, които аргументират направените по-долу изводи, се съдържат в приложението Europeana- P3- Appendix-Survey-Repository-Tools-EN.doc.

Настоящото проучване има за цел извършване на сравнителен анализ на някои популярни софтуерни пакети с отворен код, предназначени за създаване на цифрови хранилища и осигуряване на достъп до тяхното съдържание, от гледна точка на приложимостта на тези софтуерни пакети за публикуване в уеб и в последствие в Европеана на избрани колекции от цифровизирани обекти на българско културно наследство.

Сравнението е извършено на базата на следните характеристики:

- наличие на удобни авторски средства (средства за въвеждане на данни и метаданни);
- поддръжка на изображения с висока разрешаваща способност;
- мащабиране на изображения;
- наличие на библиотеки за обработка на изображения;
- наличие на удобни средства за визуализация/разглеждане на обекти/данни/метаданни;
- функционални възможности на вградената машина за търсене;
- изисквания към формата на метаданните;
- възможности за работа с файлове от различни типове;
- използвана база от данни;
- публичен достъп до съдържанието;
- управление на правата на потребителите;
- възможности за публикуване в семантичен формат (опционално)
- съвместимост с протокола OAI-PMH който се използва от Европеана (опционално);

4.1 DSpace

http://www.dspace.org

Dspace е софтуерно средство с отворен код, предназначено за създаване на цифрови хранилища, което предоставя пълна гама от инструменти за управление и съхранение на цифрово съдържание. DSpace е най-широко използваната платформа за създаване и поддържане на цифрови хранилища, която има повече от 1000 инсталации по света и бързо нарастващо общество на активни потребители.

Обобщени технически характеристики

Авторски средства	Да (удобни за използване от библиотечни и архивни специалисти)
Поддръжка на изображения с висока разрешаваща способност	Да
Мащабиране на изображения	Не се поддържа
Средства за разглеждане на обекти/данни	Да
Вградена машина за търсене	Да (пълнотекстово и фасетно търсене; позволява използване на логически връзки между различни полета)
Формат на метаданните	Qualified Dublin Core е схемата на метаданни, валидна по подразбиране; може също да работи с метаданни във формат MARC/MODS
Допустими типове на файловете	Разпознава файлове с разнообразни формати (например TXT, DOC, PDF, JPEG, MPEG, TIFF и др.). Може да работи с файлове от произволен тип
База от данни	Използва PostgreSQL
Съвместимост с OAI-PMH	Да

Публичен достъп до съдържанието	Предвидени са възможности за достъп от страна на нерегистрирани потребители
Управление на правата на потребителите	Предоставя средства за управление на правата на потребители и групи по отношение на достъпа им до колекции и отделни обекти
Библиотеки за обработка на изображения	Да
Възможности за публикуване в семантичен формат	Може да се експортират метаданни в N3 и RDF/XML формат

В българските академични институции има натрупан опит в използването на DSpace, например научният портал на СУ „Св. Кл. Охридски" и цифровата библиотека на ИМИ-БАН са базирани на DSpace.

4.2 CDS Invenio

http://invenio-software.org

Invenio е софтуерно средство с отворен код, предназначено за създаване на уеб достъпни цифрови библиотеки и цифрови хранилища, което е разработено в CERN. Използва се в CERN като документен сървър, управляващ повече от 1 000 000 библиографски записа (представляващи статии, книги, списания, снимки, видео и др.) в областа на физиката на високите енергии, както и в множество други институции. Характеризира се с гъвкавост и бързодействие.

Обобщени технически характеристики

Авторски средства	Да
Поддръжка на изображения с висока разрешаваща способност	Да
Мащабиране на изображения	Не е намерена информация
Средства за разглеждане на обекти/данни	Да
Вградена машина за търсене	Да (поддържа интерфейси за различни типове търсене, включително комбинации от пълнотекстово търсене и търсене по ограничения върху стойностите на определени от потребителя метаданни)
Формат на метаданните	Използва MARC 21 като основен библиографски формат; поддържа описания на статии, книги, дисертации, изображения, видео, музейни обекти и др.
Допустими типове на файловете	Разпознава файлове с разнообразни формати (например TXT, DOC, PDF, JPEG, MPEG, TIFF и др.). Може да работи с файлове от произволен тип

База от данни	Използва MySQL
Съвместимост с OAI-PMH	Да
Публичен достъп до съдържанието	Предоставя възможности за свободен достъп до публикуваните колекции
Управление на правата на потребителите	Предоставя средства за управление на правата на отделни потребители и групи от потребители
Библиотеки за обработка на изображения	Не е намерена информация
Възможности за публикуване в семантичен формат	Възможно е новите етикети и конвенции на HTML5 да бъдат използвани по отношение на изхода на Invenio

В българските академични институции има натрупан опит в използването на Invenio: цифровата библиотека на СУ „Св. Кл. Охридски" в областта на хуманитарните науки "SUDigital" е базирана на Invenio.

4.3 Nuxeo DAM/DM

http://community.nuxeo.com, http://nuxeo.org

Nuxeo DAM е система за управление на богато цифрово съдържание (изображения, анимация, видео, музика и др.). Nuxeo DM е среда за управление на документи. Двете системи могат да бъдат интегрирани, като интерфейсът на Nuxeo DAM работи с файлове, съхранявани и управлявани от Nuxeo DM.

Обобщени технически характеристики

Авторски средства	Да (може да се анотират цели изображения или избрани техни области)
Поддръжка на изображения с висока разрешаваща способност	Да
Мащабиране на изображения	Изображенията могат да се намаляват и увеличават, както и да се показват самостоятелно или в мрежа
Средства за разглеждане на обекти/данни	Да
Вградена машина за търсене	Да (поддържа се търсене по ключови думи и фасетно търсене по зададени ограничения върху стойностите на избрани метаданни)
Формат на метаданните	Не е намерена достоверна информация
Допустими типове на файловете	Nuxeo DAM поддържа работа с файлове от произволен тип, в частност може да показва с ограничени подробности (preview) изображения, видео и аудио. Потребителят може да избира и

	коментира анотирани области от изображения. Автоматично се генерират поредици от умалени изображения (thumbnails) на кадри от видео файлове. Могат да бъдат изпълнявани видео и аудио файлове.
База от данни	PostgreSQL, MySQL, Oracle, Microsoft или H2
Съвместимост с OAI-PMH	Не е намерена достоверна информация
Публичен достъп до съдържанието	Предвидени са възможности за свободен достъп до публикуваното съдържание
Управление на правата на потребителите	Предоставя средства за управление на правата на отделни потребители и групи от потребители
Библиотеки за обработка на изображения	Да
Възможности за публикуване в семантичен формат	Да (с помощта на специални скриптове могат да се импортират и експортират RDF данни/метаданни)

BBC базира своя мобилен портал на Nuxeo DAM и публикува повече от 100 милиона информационни единици месечно.

4.4 Alfresco

http://www.alfresco.com

Alfresco е софтуерна система с отворен код, предназначена за управление на съдържание. Разработени са множество допълнителни модули (plugins), които позволяват на Alfresco да поддържа типичните функционални характеристики на система за управление на богато цифрово съдържание.

Обобщени технически характеристики

Авторски средства	Да (може да бъдат добавяни коментари към всяко изображение
Поддръжка на изображения с висока разрешаваща способност	Да
Мащабиране на изображения	Според системната документазия е възможно интегриране с ImageMagick с цел мащабиране на изображения
Средства за разглеждане на обекти/данни	Да
Вградена машина за търсене	Да (използва машината за търсене Lucene)

Формат на метаданните	Не е намерена достоверна информация
Допустими типове на файловете	Работи с произволен тип файлове – документи, изображения, видео и аудио
База от данни	PostgreSQL или MySQL
Съвместимост с OAI-PMH	Не е намерена достоверна информация
Публичен достъп до съдържанието	Предвидени са възможности за свободен достъп до публикуваното съдържание
Управление на правата на потребителите	Предоставя средства за управление на правата на отделни потребители и групи от потребители
Библиотеки за обработка на изображения	Да
Възможности за публикуване в семантичен формат	Alfresco управлява метаданните, използвайки т. нар. „аспекти”. Аспектът е колекция от стойности на отделни метаданни; с всеки документ може да се асоциира множество от аспекти. Към настоящия момент не се поддържа експортиране на метаданни в семантичен формат

Alfresco предоставя богат набор от средства за съвместна работа. Съществува подробна документация, която позволява лесно да бъдат разработвани допълнителни модули за Alfresco.

4.5 REPOX

http://repox.ist.utl.pt

REPOX заема особено място sред анализираните софтуерни инструменти за създаване на цифрови хранилища и предоставяне на достъп в рамките на Европеана до тяхното съдържание.

REPOX е самостоятелен OAI-PMH сървър, разработен в рамките на един от проектите – предшественици на Европеана, който поддържа добре документирана технология за подготовка на данни и трансформиране на метаданни във формата на Европейската цифрова библиотека.

4.6 Изводи

От разгледаните софтуерни пакети Alfresco предоставя най-развита среда за съвместна работа. Nuxeo DAM/DM предлага най-силна поддръжка на достъпа до публикувано богато цифрово съдържание (включително и до отделните му компоненти), както на асоциирането на семантични метаданни със съответното цифрово съдържание.

От гледна точка на ресурсите, необходими за уеб публикуване (и опционално публикуване в Европеана) на цифрови колекции от обекти на българско културно наследство, Invenio (евентуално в комбинция с REPOX) изглежда най-подходящото решение. DSpace е разумна алтернатива от гледна точка на опита, натрупан в някои водещи национални академични институции.

Приложение R1

Survey of Open Source Software Tools for Creating Digital Repositories

Introduction: Europeana

Europeana (http://www.europeana.eu/portal/) was launched in 2008, with the goal of making Europe's cultural and scientific heritage accessible to the public. The project is funded by the European Commission. It is based in the National Library of the Netherlands, the Koninklijke Bibliotheek. Europeana builds on the experience of The European Library (TEL, http://www.theeuropeanlibrary.org/), which is a service of the Conference of European National Librarians.

Europeana has a recognisable brand identity and worldwide visibility which each content provider benefits from because Europeana drives traffic to their site. Videoactive's traffic statistics for October 2009, for example, showed that 60% of the total traffic to the site was coming from Europeana. By harvesting and exposing the metadata to search engines Europeana enhances the routes users follow to reach content.

Europeana adds value to the content by juxtaposing related images, texts, videos and audio items, therefore, 'repatriating' content that is geographically dispersed into a single, coherent and contextual virtual space. As a result, Europeana enriches users' experience, giving them the opportunity to study related content, which comes from different collections in different countries and exists in different formats, in a single virtual space. The multilingual interface makes it possible for users to search in their native language and retrieve objects that are otherwise inaccessible to them.

In the wide public Europeana is primarily perceived as a Portal exposing increasingly impressive amounts of cultural heritage from various sources to Europe's citizens. On a very abstract level Europeana can be seen as a large collection of surrogate objects representing born digital or digitized cultural heritage objects which themselves remain outside the Europeana data space (they need to be accessed by Europeana once, however, for processing with the aim of producing the surrogate representations).

As indicated in Figure 1 [1], Europeana has not only an API for end user functionality, but also an I/O-API enabling data flow from and to the content providers – the latter creates the option of re-integrating enriched content in the remote applications of the data providers.

Europeana only harvests the metadata describing a digital object and provides a thumbnail with a link to the digital object's location on the provider's or aggregator's system. A user is always directed to the provider's or aggregator's site for viewing the digital object in detail. That enables the content holder to apply their own terms to use to their content. Conditions and formats under which the provider makes the content available vary across providers (paid or free, upon registration, upon order, for viewing via a standard browser or commercial and downloadable plug-in, etc.).

National institutions/portals representing several sectors are the preferred first contact point for Europeana. A new content provider is, therefore, first directed to the national aggregator, if this one is established. However, in some countries these national institutions/portals are not yet established, and another routing will then be suggested. If an aggregator is proposed, the organisation should contact the aggregator directly for information about how to submit data. The Europeana Office may decide to take the content directly rather than via an aggregator.

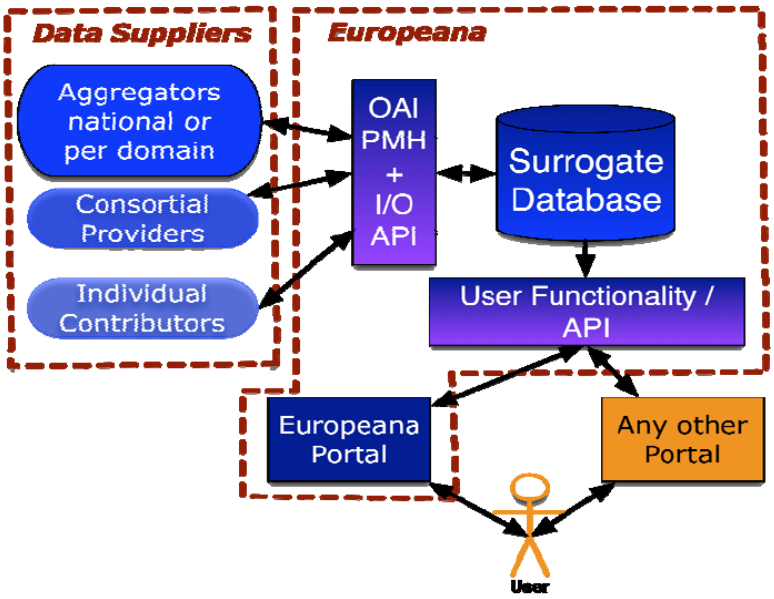


Figure 1 [1]. Europeana APIs

Every content provider and aggregator needs to comply with Europeana's technical requirements when submitting data. Europeana provides a common access point to digital cultural heritage objects across different cultural domains. It complements but does not duplicate the source environment of the digital objects. To achieve this, Europeana uses specialized portals. The object is linked to Europeana and is shown in a neutral environment, while at the same time still being available in the domain-specific portal, which may provide greater contextualization.

Europeana aims to provide access to digital objects at the lowest possible level of granularity. This implies giving direct access to the digitised object itself, that is, with a minimum click distance between the description and the object. The minimum granularity can vary, and it is in the discretion of the content provider to decide this. A broadcast provider, for example, might decide to cut down a news programme made of individual sections, and make each one available as a separate digital object. On other occasions, the complete program is of value due to the context it adds to the individual fragments. Europeana asks that content providers keep the users in mind when deciding on the granularity of their data.

Europeana's data model enables search and discovery of digital objects. Europeana maintains a common central index of the objects' metadata. It has, therefore, an object-centric rather than a collection-centric approach.

Europeana stores *representations* of digital objects and not digital objects themselves. Europeana generates a description and a preview of digital objects with the help of the metadata and thumbnails or previews of the digital objects. This requires that on the content provider's/aggregator's side there is a repository, where the digital objects are stored and can be linked to. It also requires a native website that can be used to view, play and reuse the objects.

A digital object in Europeana is a unique single entity, which can be viewed/played by users (e.g., mpeg movie, mp3 audio, jpeg photo, PDF text, etc.) on their computers. A digital object is the digitised version of a physical/analogue cultural item/artefact. Europeana does not accept descriptions, even extensive ones, that do not correspond to a digital object. A scanned catalogue card is, therefore, not an object Europeana would accept.

Europeana harvests, stores and indexes the metadata in a central index. This is important to enable integrated results display and processing of the data to provide various functionalities for the users (i.e. integrated and faceted search). This is also essential for the enrichment of the data. The preferred method for harvesting is the OAI-PMH protocol (http://www.openarchives.org/pmh/) and partners need to set up an OAI-PMH repository comprising their data mapped to the Europeana Semantic Elements (ESE) standard.

What data are submitted?

A content provider or an aggregator is responsible for making available to Europeana the following data [2]:

- Metadata (descriptive, administrative) describing a digital object. The metadata must be mapped to the ESE v3.2.2 (https://version1.europeana.eu/c/document_library/get_file?uuid=c56f82a4-8191-42fa-9379-4d5ff8c4ff75&groupId=10602). This is the Europeana current data model which consists of the Dublin Core (DC) metadata elements, a subset of the DC terms and a set of twelve elements which were created to meet Europeana's functionality needs [3],
- A preview or thumbnail of the described object,
- Persistent identifiers - active and stable links to the described digital object on the provider's site or the portal's site.

Persistent Identifiers are mandatory when submitting data to Europeana, because of the role they play in preventing duplication. Europeana has a wide network of aggregators and content providers, and the possibility of data being ingested more than once must be avoided. The development and attribution of Persistent Identifiers will help deduplicate content and provide greater control over the data.

Harvesting in Europeana should follow the implementation guidelines OAI-PMH v2 for repository implementers [4]. Repository implementers should consider exporting DC the first and most important step toward OAI-PMH interoperability. The addition of facilities to export other formats may then be added at a later date.

Migration from ESE to EDM

The Europeana Data Model (EDM) is a new approach towards structuring and representing data delivered to Europeana by the various contributing cultural heritage institutions. The model aims at greater expressivity and flexibility in comparison to the current ESE, which it is destined to replace [5]. The design principles underlying the EDM are based on the core principles and best practices of the Semantic Web and Linked Data efforts to which Europeana wants to contribute. The model

itself builds upon established standards like RDF(S), OAI-ORE, SKOS, and Dublin Core. It acts as a common top-level ontology which retains original data models and information perspectives while at the same time enabling interoperability.

EDM will re-use some of the reference ontologies already available, such as the W3Csponsored Simple Knowledge Organization System (SKOS). SKOS defines a model to represent the elements of Knowledge Organization System such as thesauri, classification schemes, and their likes.

Other ontologies that are of specific interest for EDM include Dublin Core (DC) and Friendof-a-Friend (FOAF). Dublin Core gives a compact vocabulary to describe the core features of culture objects (creators, relations to other resources, subject indexing, etc.) in a Semantic Web-enabled fashion that fits a very wide range of needs. DC is used as the basis for ESE: keeping it as a part of the model thus gives direct compatibility with legacy Europeana data. It also enables these providers that do not want to provide richer descriptions to keep to a simple vehicle for the data they submit. It finally makes EDM data better fit to sharing and re-use: DC is used by many applications, which could be very easily adopted to consume EDM data.

Typical object representations in Europeana mostly will be compound entities consisting of several parts, such as for instance metadata attributes, a thumbnail picture and a static html landing page (to give just a very simple example). For this and other reasons the OAI Object Reuse & Exchange (OAI-ORE) specifications (http://www.openarchives.org/ore/1.0/to) were chosen as the structural modelling framework for the EDM ontology.

This survey is aimed at an analysis of some widely used open source software tools for creating digital repositories with respect to their applicability for web publishing of digitized Bulgarian cultural heritage collections and making them accessible via Europeana.

DSpace (http://www.dspace.org)

DSpace is an open source repository software tool for creating repositories focused on delivering digital content to end users, and providing a full set of tools for managing and preserving content within the application. DSpace is the most widely used repository software platform, with over 1000 installations worldwide representing a growing and active user community.

DSpace was developed by the MIT Libraries and the Hewlett-Packard Company. It may be characterized as a set of cooperating Java applications and utility programs that maintain an asset store and an associated metadata store. The web applications provide interfaces for administration, deposit, ingest, search, and access. The asset store is maintained on a file system, or similar storage system. The metadata, including access and configuration information, is stored in a relational database.

DSpace is written in Java. It requires the use of either a PostgreSQL or Oracle database. PostgreSQL is a free to use and is open source. It also requires a Java Servlet Container, which can be Apache Tomcat, Jetty or Caucho Resin.

Technical information (summary)

Authoring/Annotation Tool(s)	Yes (familiar to librarians and archivists)
Supports High Resolution Images	Yes
Image Scaling	No support
Browsing Tool(s)	Yes

Built-in Search Engine	Yes (full-text search for end users; faceted search; can use logical conjunctives across different fields – see Figure 2 [8])
Metadata	Qualified Dublin Core (default metadata schema); can also translate metadata from MARC/MODS and other metadata schemas
File Types	Recognizes files of any common format (e.g. TXT, DOC, PDF, JPEG, MPEG, TIFF). Any file type can be uploaded to DSpace
Suitable Database	Uses PostgreSQL
OAI-PMH Compliance	Yes
Public Access to Content	DSpace has a user group called "Anonymous" which can have read access given to it. Thus, public and private access is managed using the same interfaces and rules as repository users
User Management	DSpace manages users and groups, which can have permissions over communities, collections and items
Image Manipulation Libraries	Yes
Semantic Support	Through the DSpace History System, metadata can be exported on the command line into N3 and RDF/XML

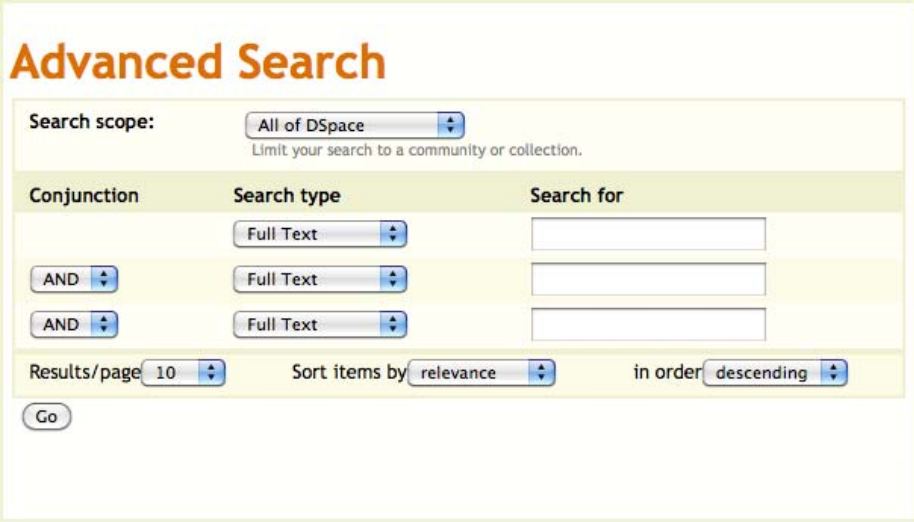


Figure 2 [8]. DSpace's advanced search interface

DSpace is easy-to-install and running quickly. It does not have much built in support for operations over specific file types such as images and videos. For example, preview thumbnails are not generated for images or videos.

DSpace is popular in Bulgarian academic institutions – e.g., the research portal of Sofia University (see Figure 3) and the digital repository of IMI–BAS run on it.

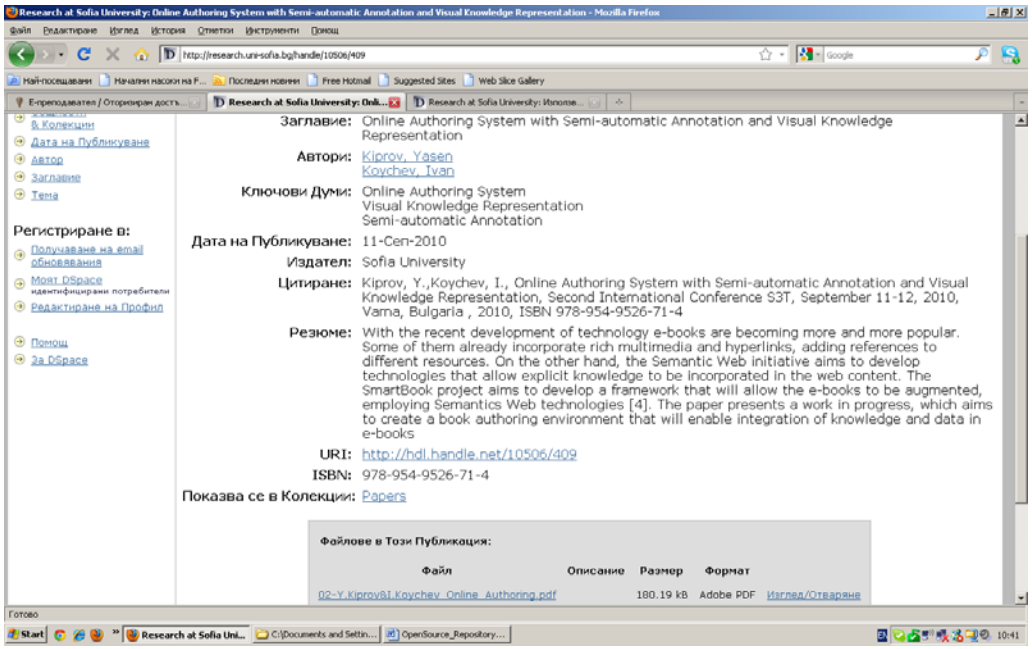


Figure 3. Browsing in the Research Portal of SU

Invenio (http://invenio-software.org/)

Invenio is a free software suite enabling one to run his/her own digital library or document repository on the web. The technology offered by the software covers all aspects of digital library management from document ingestion through classification, indexing, and curation to dissemination. Invenio complies with standards such as OAI-PMH and uses MARC 21 as its underlying bibliographic format. The flexibility and performance of Invenio make it a comprehensive solution for management of document repositories of moderate to large sizes (several millions of records).

Invenio has been originally developed at CERN to run the CERN document server, managing over 1 000 000 bibliographic records in high-energy physics since 2002, covering articles, books, journals, photos, videos, and more. Invenio is being co-developed by an international collaboration and is being used by about thirty scientific institutions worldwide

Technical information (summary)

Authoring/Annotation Tool(s)	Yes
Supports High Resolution Images	Yes
Image Scaling	No information found
Browsing Tool(s)	Yes
Built-in Search Engine	Yes (customizable simple and advanced search interfaces; combined metadata, full

	text and citation search in one go – see Figure 4)
Metadata	Uses MARC 21 as its underlying bibliographic format; handles articles, books, theses, photos, videos, museum objects and more
File Types	Recognizes files of any common format (e.g. TXT, DOC, PDF, JPEG, MPEG, TIFF). Any file type can be uploaded
Suitable Database	Uses MySQL
OAI-PMH Compliance	Yes
Public Access to Content	Ensures open access to the published collections
User Management	Invenio manages users and user groups (provides construction of user-defined document baskets and basket-sharing within user groups)
Image Manipulation Libraries	No information found
Semantic Support	Some (based on the employment of new HTML5 tags and conventions to Invenio output)

Invenio is popular in Bulgarian academic institutions – e.g., the digital library in the field of humanitarian studies at Sofia University SUDigital (see Figure 4) runs on it.

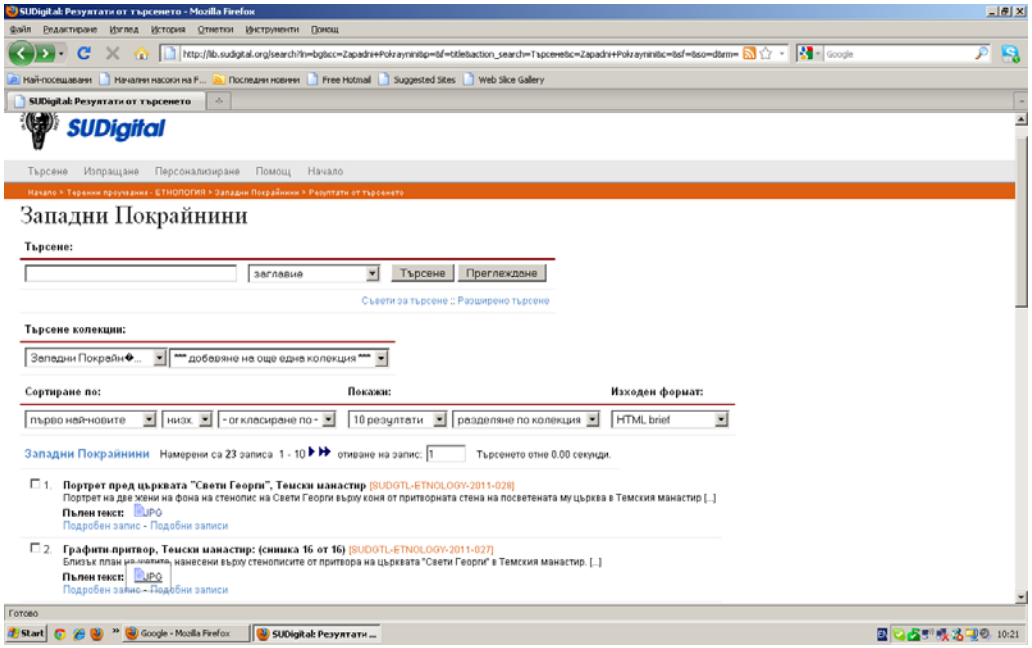


Figure 4. Search results in SUDigital

Nuxeo DAM/DM (http://community.nuxeo.com/)

Nuxeo DAM is a Digital Asset Management system, and Nuxeo DM is a Document Management system. Nuxeo DAM has a rich set of features for working with digital media assets, such as images and videos, while Nuxeo DM is a more mature product for dealing with the management of any files, without specific features for dealing with file types (like images and videos).

It is possible to integrate the two systems together, so that the Nuxeo DAM interface runs on media files that are stored and managed from Nuxeo DM.

The BBC has adopted Nuxeo DAM for their mobile portal, which receives 100 million hits per month. Nuxeo DAM is written in Java and uses JDBC to interface with databases. This allows users to use a PostgreSQL, MySQL, Oracle, Microsoft, or H2 database.

Technical information (summary)

Authoring/Annotation Tool(s)	Yes (annotations can be added to a whole image or selected areas of the image)
Supports High Resolution Images	Yes
Image Scaling	Can zoom in and out of images. Shows images individually in an inspector as well as in a grid
Browsing Tool(s)	Yes
Built-in Search Engine	Yes (searching the DAM is performed through a faceted filtering interface, where images can be filtered by: folder, category, file type, topic, geographic coverage and authoring date. This is in addition to keyword filtering)
Metadata	No reliable information about the used metadata format(s)
File Types	Nuxeo DAM supports any file type, and can preview images, videos and audio specifically. For images, annotated areas can be selected and commented on. For videos, a storyboard of thumbnails is automatically generated. Videos and audio files can be played inline
Suitable Database	PostgreSQL, MySQL, Oracle, Microsoft or H2
OAI-PMH Compliance	No reliable information found
Public Access to Content	Items can be exported, which provides their public URL. Through Nuxeo DM, workspaces can be explicitly exported as web pages
User Management	Manages users and groups, with access control per user and group
Image Manipulation Libraries	Yes
Semantic Support	Nuxeo has a built-in Relational engine that

	allows predicates to be assigned to items. Vocabularies are managed centrally, where URIs for predicates can be added, with a human readable label. Nuxeo uses Jena internally to model its RDF data. Through the Nuxeo import and export scripts, the RDF data can be exported

Nuxeo DAM offers file downloads for different sizes (original, medium and thumbnails) – see Figure 5 [8]. Through Nuxeo DAM users can comment on items, and annotate them. Through Nuxeo DM users can create collaborative workspaces, to collaborate on projects with other users; users can also engage in discussion forums.

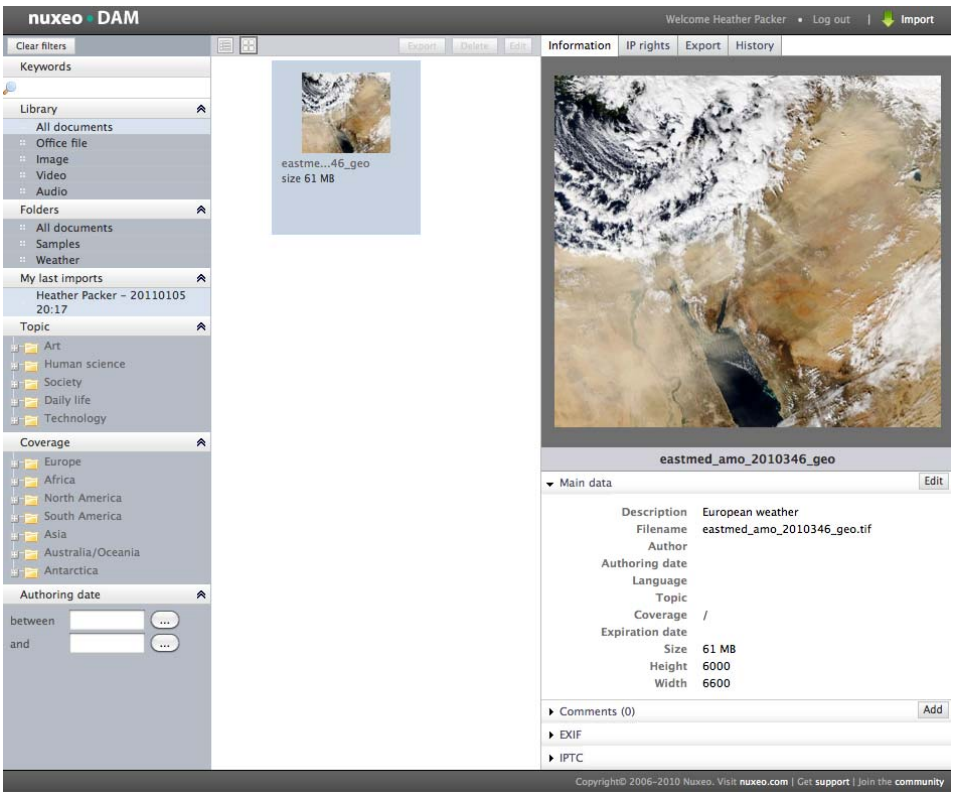


Figure 5 [8]. Nuxeo: A screenshot of the summary and thumbnail of a large image

Alfresco (http://www.alfresco.com/)

Alfresco is an open source enterprise content management system. It is not a dedicated Digital Asset Management (DAM) solution, however, there is support through plugins and configuration options that allow it to support the typical features of a DAM.

Alfresco is written in Java and can connect to either a MySQL or PostgreSQL database. It also uses Spring for its architecture, Hibernate for writing to the database, Lucene for searching, and iPhone apps so that it can be accessed from iPhones.

Technical information (summary)

Authoring/Annotation Tool(s)	Yes (comments can be added to each image)
Supports High Resolution Images	Yes
Image Scaling	It is possible to integrate with ImageMagick for transformations inline, but this is not available in their demo
Browsing Tool(s)	Yes
Built-in Search Engine	Yes (uses a Lucene based search with customisable support for field based searches)
Metadata	No reliable information about the used metadata format(s)
File Types	Manages any file type – documents, images, video and audio
Suitable Database	PostgreSQL or MySQL
OAI-PMH Compliance	No reliable information found
Public Access to Content	Each asset lists its public URL. Also, each asset can be checked out to Google Docs
User Management	Users and groups, with access control per user and group
Image Manipulation Libraries	Yes
Semantic Support	Alfresco manages metadata using "aspects". An aspect is simply a collection of metadata fields, and an item in the document library can have a number of aspects assigned to it, which increases the number of metadata fields the item has. Semantic export is not currently supported

Alfresco is divided into Sites. Each site has users that are members, and groups can also be added to sites. Each site has its own wiki, blog and discussion board, which can be used to make rich notes, make dated updates, and to have discussions about items, respectively.

Alfresco looks just like SharePoint to Microsoft Office, allowing users to upload, check-in, check-out and modify content right from Office. There is a lot of documentation on how to develop plugins for Alfresco. It lacks RDF export, but does have a sophisticated metadata system. The FISE/Stanbol project promises RDF integration, otherwise it would be possible to implement a plugin that can export the metadata.

REPOX (http://repox.ist.utl.pt/)

REPOX is a framework to manage metadata spaces. Some authors characterize it as a Data Aggregation and Interoperability Manager. It comprises several channels to import metadata from data providers, services to transform metadata between schemas according to user's specified rules, and services to expose the results to the exterior. REPOX aims to provide to all the TEL and

Europeana partners a simple solution to import, convert and expose their bibliographic data via OAI-PMH, by the following means:

- *Cross platform*
  It is developed in Java, so it can be deployed in any operating system that has an available Java virtual machine.
- *Easy deployment*
  It is available with an easy installer, which includes all the required software.
- *Support for several metadata formats and encodings*
  It supports UNIMARC and MARC21 schemas, and encodings in ISO 2709 (including several variants), MarcXchange or MARCXML. During the course of the TELplus project, support will be added for other possible encodings required by the partners.
- *Metadata crosswalks*
  It offers crosswalks for converting UNIMARC and MARC21 records to simple Dublin Core as also to TEL-AP (TEL Application Profile). A simple user interface makes it possible to customize these crosswalks, and create new ones for other formats.

REPOX is a standalone OAI-PMH server developed within the TELplus project. It aims to provide all The European Library partners with a simplified and flexible solution for exposing their data via OAI-PMH. The technology for creating a data provider and data source as well as the one for mapping metadata to TEL format are described in [6] and [7].

Conclusions

According to [8], Alfresco provides the richest collaboration environment of any of the evaluated systems, with easy to use wiki, blog and discussion areas for each sub-project on the site. Nuxeo DAM/DM provides the best tools for access to published rich digital content (and to its particular components) and allows the richest support for semantic metadata to be associated with digital media assets.

From the point of view of the resources needed for web publishing of existing digitized Bulgarian cultural heritage collections and making them accessible via Europeana, Invenio (possibly in combination with REPOX) seems to be the most adequate solution. DSpace is a reasonable alternative having in mind the experience gathered in national academic institutions.

References

[1] C. Concordia et al., Not (just) a Repository, nor (just) a Digital Library, nor (just) a Portal: A Portrait of Europeana as an API. Available at http://www.ifla.org/files/hq/papers/ifla75/193-concordia-en.pdf

[2] The Europeana Team, Europeana Aggregators' Handbook. Available at http://version1.europeana.eu/c/document_library/get_file?uuid=94bcddbf-3625-4e6d-8135-c7375d6bbc62&groupId=10602

[3] A. Friberg, Europeana. National Conference on Library Digitalisation (Budapest, March 17, 2010). Available at http://dtt.ogyk.hu/letoltesek/Friberg_Annette.pdf

[4] C. Lagoze, H. Sompel, M. Nelson, S. Warner (Eds.), Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: Guidelines for Repository Implementers. Available at http://www.openarchives.org/OAI/2.0/guidelines-repository.htm

[5] M. Doerr et al., The Europeana Data Model (EDM). Available at http://www.ifla.org/files/hq/papers/ifla76/149-doerr-en.pdf

[6] REPOX Documentation. Available at http://repox.ist.utl.pt/doc/index.html

[7] D5.3.1 – Europeana OAI-PMH Infrastructure – Documentation and final prototype, Appendix – REPOX User Manual. Available at http://www.europeanaconnect.eu/documents/01a_Europeana_OAI_PMH_APPENDIX_User%20Manual.pdf

[8] H. Packer, Comparison of Digital Asset Management Systems (DAMs) and Content Management Systems (CMSs). ResearchSpace Project, January 2011.

Европеана П4: Семантично публикуване

Създаден: 15 юни 2011
Редактиран: 30 авг 2011

1 Съдържание

Европеана П4: Семантично публикуване ... 1
1 Съдържание ... 1
2 Въведение ... 1
2.1 ESE спрямо EDM ... 1
3 Задачи ... 3
3.1 Основни задачи ... 3
3.1.1 Подготовка и Онтологична работа ... 3
3.1.2 Конверсия ... 5
3.1.3 Свързване на данните ... 6
3.1.4 Семантично запазване, визуализация ... 6
3.1.5 Триплификация, Публикуване в Европеана ... 7
3.1.6 Рационализиране (Streamline) ... 7
3.2 Допълнителни задачи ... 7

2 Въведение

Преди този документ, моля прочетете въведителния документ Europeana- general.doc, по-специално секции 2.11 и 2.12.

Целта на този под-проект е публикува обектите в Европеана, след като те бъдат въведени в цифровото хранилище.

2.1 ESE спрямо EDM

Една от задачите на П4 е да избере в какъв модел ще публикува данните: Europeana Semantic Elements (ESE), Europeana Data Model (EDM) или и двата. EDM е единен формат за представяне на данните от културното наследство, разработван за Европеана. Той е в процес на стабилизиране, което предполага кодиране да истински данни в този формат. EDM е създаден, за да измести настоящия модел за представяне на данни в Европеана (ESE). EDM предоставя набор от метаданни, които позволяват да се опишат обозначителни елементи от обектите на дадена колекция от културни обекти. Целта на този модел е да предостави унифициран формат за описание и достъп през единен портал до обекти на културно-историческото наследство.

По-детайлен обзор на ESE и EDM е даден във въведителния документ, сек. 2.11. Съображенията по този въпрос са:

- EDM е бъдещето на Европеана. Този модел ще измести ESE, защото той позволява по-богато представяне, разширяемост, "неинформационни обекти" (хора, места, периоди, събития и пр), връзки между обекти.
- Една от най-важните възможности за нас е да се използват глобални номенклатури (терминологични речници), така че публикуваните данни да се свързват с други световни данни за същите неинформационни обекти.
- Европеана закъснява спрямо предварително анонсираното внедряване на EDM, но по това внедряване се работи сериозно
- След постигане на техническа готовност, Европеана ще мигрира съществуващите данни от ESE към EDM (което вече е направено в Европеана LOD Pilot). Но по-богатите възможности на EDM няма да може да се използват при тази миграция
- Ако се реализира конверсия към EDM, то реализирането на конверсия към ESE ще е лесна задача, тъй като ESE е концептуално подмножество на EDM.

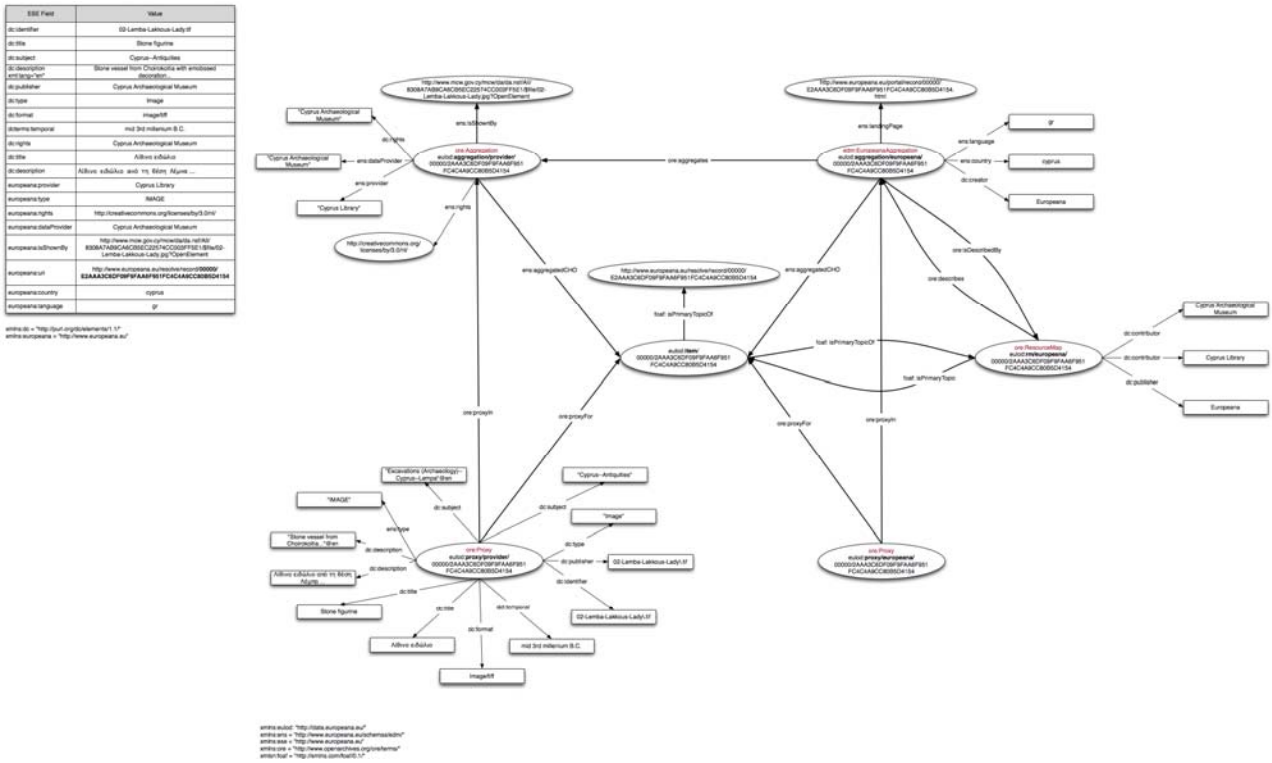
На 10.8.2010 осъществихме контакт с техническото ръководство на Европеана. Antoine Isaac, Scientific coordinator, Europeana ни писа следното:

- *Thanks a lot for your interest in Europeana and EDM! With my Semantic Web background I'd like very much to see you involved in our network.*
- *Regarding production-level ingestion of EDM into Europeana, unfortunately I can't really give you a precise timeline. I'm still hoping for something to happen this year, but don't have much visibility on what and when our technical team can deliver.*
- *However, I am very much interested in getting EDM data right now, on individual project basis. We may have to convert it to ESE ourselves, before making it accessible through the Europeana portal. But already having got it some way, will save us the pain of organizing two different data ingestion sprints with you (one ESE now, and one EDM later).*
- *Further, while we are still in a "prototyping" phase: getting the opportunity to experiment with EDM datasets will help us finalize our documentation, and maybe even adapt some details, depending on providers' feedback.*
- *Of course this would require you not to be afraid to undergo a less-tested process. We can guarantee that your data will end up on the Europeana portal ingested, but the process will be bumpier than with the ESE option. I however trust for people with your experience it should be quite ok.*

Тъй като смятаме, че е по-важно да се гледа към бъдещето, към момента решението ни е следното:

- Да започнем работа при допускането, че EDM ще влезе в реално използване скоро. Това включва набор метаданни, мапинг на метаданните, инструмент за и конверсия към EDM.
- Ако до ноември 2011 Европеана не покаже ясна готовност да внедри EDM или алтернативно да конвертира данните ни към ESE, то сами да направим такава конверсия

Фигурата по-долу представя модела за мапинг между ESE (отляво) и EDM и илюстрира както "плоската" структура на ESE (без вложени данни), така и относителната сложност на EDM.



3 Задачи

3.1 Основни задачи

Основни задачи на този под-проект:

3.1.1 Подготовка и Онтологична работа

Информацията, която описва обектите се определя от изследователите, които работят с тях. В много случаи тя варира в зависимост от типа на описваните обекти, тяхната датировка, предназначение и т.н.

Доставчиците на съдържание предоставят набор от метаданни, с които са описани обектите на собствената им колекция. Тези метаданни не съответстват задължително на набора от метаданни дефиниран в EDM. За да стане възможно публикуването на колекциите на доставчиците на данни в Europeana е нужно метаданните, които ги описват да бъдат хармонизирани с модела на Eropeana – EDM.

За целта е необходимо изготвяне на схема за съответствие между метаданните на доставчиците и метаданните в EDM. Тази схема трябва да включва:

(а) идентифициране на идентични метаданни в EDM и схемата на даден доставчик,

(б) идентифициране на метаданни без съответствие в EDM,

(в) идентифициране на метаданни на доставчиците, които нямат съответствия в EDM,

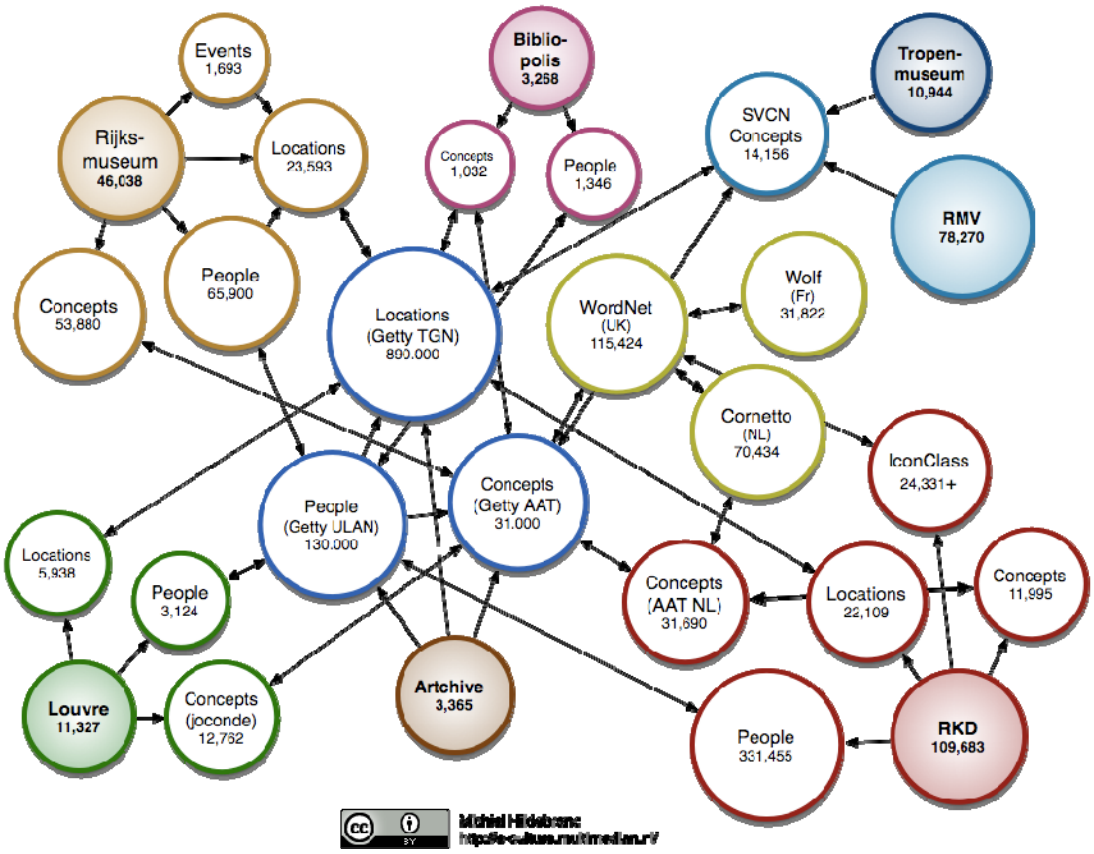
(г) осъществяване на мапинг между съответстващите си метаданни,

(д) определяне на метаданни от EDM, които липсват в схемата на доставчиците, за прибавяне към схемите на доставчиците.

Изграждането на такива съответствия ще даде възможност за публикуването на колекциите в Europeana като в същото време запази възможността на изследователите да

използват пълния набор от метаданни, необходими за адекватното експертно описание на обектите от техните колекции.

- Хармонизация на метаданните от доставчиците към тези на EDM (напр. промяна и добавяне на полета).
- Инструкции за конфигуриране на метаданни към под-проект П3.
- (Заедно с П3) Създаване на схеми на различните видове обекти в цифровото хранилище
- Създаване на съответствие между моделите на съдържанието на доставчиците и EDM (ontology mapping). Ще се получи модел подобен на фигурата по-горе
    - Най-добре е ако може съотвествието да се запише като атрибути на метаданните в хранилището
- Разработка на стратегия за присвояване на адреси (URL) и уникални идентификатори (URI), която да съответства на:
    - Изискванията на Европеана за перманентни URL на цифровите обекти
    - Рекомендациите за URI за свързани семантични данни[1]
- Проучване и избор на глобални номенклатури (терминологични речници) които са подходящи за избраните предметни области и доставчици. Например по-долу са показани няколко такива множества от данни.



[1] http://www.w3.org/DesignIssues/LinkedData.html

- Включени са хора, места, концепции и събития от няколко музея (Rijksmuseum и RKD в Холандия, Лувъра във Франция, Getty в САЩ), както и общ смислов речник на няколко езика (WordNet EN, Wolf FR, Cornetto NL)
- Горните множества далеч не са единствени, и освен това връзките между тях са непълни.
- Например Virtual International Authority File (VIAF)[2] включва данни за хора и организации, които са напълно корелирани между 20 национални библиотеки.
  - Напр. запис за Св.Методий[3] се намира в библиотеките на САЩ, Израел, Ватикана, Канада и Испания

- Друг пример е GeoNames,[4] което съдържа имена и географски координати на близо 8 милиона места[5]. Но за исторически места други тезауруси (специално Getty TGN) може да са по-подходящи.
- DBpedia[6] е семантичният аналог на Wikipedia и може да се използва като алтернативен източник на URI-та

Използването на глобални номенклатури ще позволи глобалното свързване на нашите данни.

3.1.2 Конверсия

- Проучване на съществуващ софтуер с отворен код
- Създаване или адаптация на софтуер за конверсия
- Запис на съответствието между данни в хранилището и семантичните данни като конфигурация на инструмента.
- Междинно извличане на данните от библиотеките в XML формат. (Ако е необходимо)
- Конверсия на колекциите от хранилището в семантичен формат
- Предоставяне на софтуера за преизползване от други институции и проекти

---

[2] http://viaf.org/

[3] http://viaf.org/viaf/20472719/

[4] http://www.geonames.org/

[5] http://www.geonames.org/statistics/

[6] DBpedia : http://dbpedia.org.

3.1.3 Свързване на данните

Семантичните технологии предоставят възможността за интегриране на данните от различни множества чрез тяхно свързване. Свързването на данните трябва да се осъществи с техники за идентифициране на обекти с идентично съдържание и обозначаване на тяхната идентичност или сходство.

Свързването на данните дава възможност за достъп до информация за даден обект, която се е намирала в различни бази данни или схеми на описание. Така например, ако в една колекция съществува описанието на размерите на един предмет, а в друга описанието на обстоятелствата на неговото намиране и идентификаторите на обекта в двете колекции са обозначени като отнасящи се до един и същ обект, то запитвания за обекта с неговото ID от едната колекция към семантичното хранилище ще получи като резултат информацията, както за размерите на обекта, така и за обстоятелствата на неговото намиране.

Освен свързване на обектите е възможно свързване на метаданните с LOD. Това свързване предполага обозначаване, че метаданни описващи една колекция се отнасят до метаданни, описващи друга колекция.

- Свързването на географски обекти с обекти от Geonames гарантира достъп до обширна географска информация за местоположението на даден обект, местности, с които то граничи, неговите географска ширина и дължина и т.н.
- Свързването на данните с DBpedia пък позволява достъп до информация за личности, създатели на дадени артефакти, за исторически епохи и събития, за истоията на географски обекти и т.н.
- Приложените методи трябва да позволяват свързване на данните с различни специализирани речници, специално изработени за целите на описанието и събирането на културното наследство. Напр:
    - VIAF вероятно има данни за повече исторически личности от DBpedia
    - Особено полезна за описанието на географските данни във времето е информацията за наименованията на дадени географски обекти във времето. Тази информация не се съдържа в Geonames. От друга страна Getty TGN съдържа и историческа географска информация.

Конкретни задачи:

- Свързване на данните към съществуващи отворени множества данни (LOD): места, древни и исторически личности, свързани събития. Предпочитаният начин е това да стане чрез мапинг на контролирани стойности към глобални терминологии
- Създаване на семантични връзки към други обекти в хранилището

3.1.4 Семантично запазване, визуализация

- Записване на получените RDF данни в семантична база (semantic repository), която да може да се ползва през SPARQL протокол
- Създаване на софтуер за уеб представяне (визуализация) на семантичните данни, който позволява:
    - Преглед, интегрирайки се към механизмите за визуализация на цифровото хранилище
    - Навигация към свързани обекти (виж примерна илюстрация във въведителния документ, сек. 2.12),

- (Заедно с П3) Изграждане на уеб сървър, който връща машинно-четимите семантични данни (RDF) или уеб представянето (HTML), като използва content negotiation
- Публикуване като източник в Comprehensive Knowledge Archives Network (CKAN)
- (Опционално) Статистики и логове за достъп

3.1.5 Триплификация, Публикуване в Европеана

- "Триплификация" на данните.
  Привеждането на данните на доставчиците в модела на EDM и създаването на модел за свързани данни за тяхното обогатяване е концептуалната страна на задачата за публикуване в Европеана. Работата по реализиране и популиране на данните се състои от:
  - създаване на RDF тройки на базата на създадените схеми
  - свързване на данните чрез правила за свързване до постигането на мрежа от свързани обекти и понятия,
  - зареждане на семантичното хранилище с новосъздадените данни
  - захранване на OAI-PMH сървър с генерираните данни. (Ако хранилището има такъв)
- Създаване на ESE (само ако Европеана няма готовност да използва или конвертира EDM)
- Координация на работата с Европеана
- Публикуване на данните в Европеана, вкл. необходимите административни процедури

3.1.6 Рационализиране (Streamline)

- (Ако има възможност) Разработка на процедура за инкрементално обновяване при добавяне
- Документиране на цялостния процес и "пакетиране" на разработения софтуер, за да може да се използва самостоятелно от библиотеките

3.2 Допълнителни задачи

- Цялостна координация на под-проектите
- Установяване и документиране на процес за публикуване, който да е лесен за използване от други културни институции
- Създаване на група (Special Interest Group, SIG) за публикуване в Европеана, координация с други съмишленици
- Установяване на технически агрегатор за България (www.bulgariana.eu)
  - Инсталиране и поддръжка на достатъчно мощен сървър, с осигурено работно време (Service Level Agreement)
  - Улесняване на други културни институции, които желаят да публикуват свои данни

How to contribute to EUROPEANA

(A Survey)

Author:	*Boyan Bontchev*
Creation Date:	*09.08.2011*
Last Update:	*23.08.2011*
Version:	*0.53*
Status:	*Draft for revisions*
Scope:	*Internal document*

Contents

1. Introduction to Europeana 3
   1.1 The Europeana foundation 3
   1.2 The Europeana service 4
2. Europeana aggregators 4
3. The EUROPEANA submitting process 5
   3.1 Routes and criteria for contributing to Europeana 5
   3.2 Steps for direct contributions to Europeana 7
   3.3 Indirect contributions to Europeana 8
   3.4 Europeana Local Technical Implementation Steps 9
      3.4.1 Constitution of digital collection(s) 11
      3.4.2 Repository installation 11
      3.4.3 Metadata Transformation 11
      3.4.4 Repository Population 12
      3.4.5 Metadata Harvesting 12
      3.4.6 Usage of aggregator repositories 12
      3.4.7 Starting end-user services 12
4. Europeana Semantic Elements 12
   4.1 ESE goals 13
   4.2 ESE structure and restrictions 13
   4.3 Content Ingestion 16
   4.4 Shortcomings of ESE 16
5. Europeana Data Model 17
   5.1 EDM goals 17
   5.2 EDM structure 18
   5.3 Shortcomings of EDM 20
6. Mapping ESE to EDM 20
7. Conclusions 21

References 21
   A. Principal references 21
   B. Other references 22

1. Introduction to Europeana

The Europeana project (http://version1.europeana.eu/web/europeana-project/) aims at realizing an integrated access to digital collections of European cultural heritage provided by various organizations of European countries. Such collections contain digital objects representing art works available at museums, libraries, or archives and, being a part of our multicultural and multilingual heritage, deserve to be discoverable within a unique on-line software environment.

According to a statement of European parliament, Europeana is "a digital library that is a single, direct and multilingual access point to the European cultural heritage". The Commission's objectives for Europeana could be summarized as follows (Meghini et al, 2010):

- To create a multilingual public-domain access point to Europe's cultural and scientific heritage
- To use digitized cultural and scientific heritage resources as input for a wide range of information products and services
- To play a key role in the future growth of sectors such as learning and tourism
- To inspire new creative enterprise and innovation
- To promote understanding of our common European background and the sense of a European identity

The name Europeana stands for two mayor things: EUROPEANA foundation and EUROPEANA service.

1.1 The Europeana foundation

The Europeana Foundation is the governing body of the Europeana service. Its members are the presidents and chairs of European associations for cultural heritage and information associations.

The Foundation promotes collaboration between museums, archives, audiovisual collections and libraries so that users can have integrated access to their content through Europeana and other services.

The Foundation is incorporated under Dutch law as Stichting Europeana and is housed within the Koninklijke Bibliotheek, the national library of the Netherlands. It provides a legal framework for the governance of Europeana, employing the staff, bidding for funding and enabling the sustainability of the service.

The Europeana Foundation has defined four main objectives to be pursued (Europeana foundation. 2007):

- to provide access to Europe's cultural and scientific heritage by way of a cross-domain portal
- to facilitate formal agreement across museums, archives, audiovisual archives and libraries on how to co-operate in the delivery and sustainability of a joint portal
- to stimulate and facilitate initiatives to bring together existing digital content
- to support and facilitate digitization of Europe's cultural and scientific heritage,

As presented in fig.1, the Europeana Foundation consists of the Council of Content Providers and Aggregators (CCPA), the Board of Participants and the Executive Committee (Europeana Team, 2010). The Council elects 6 members to the Foundation Board and plays an advisory role on strategy and policy from the content providers' point of view. Other members of the Board come from the major pan-European professional organisations representing museums, libraries, archives and audio-visual archives.

Every type of eligible organisation or project management body that provides or will provide content to Europeana is allowed to be a member of the Council. A provider or aggregator may choose more than one individual from its organisation to participate in the Council.

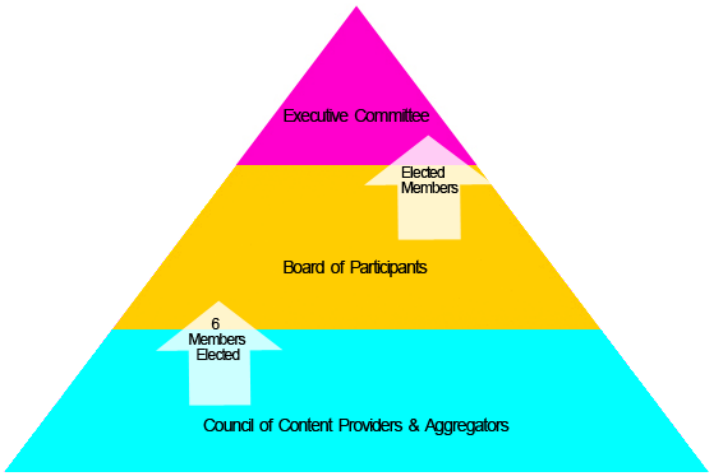


Fig. 1. Structure of the EUROPEANA foundation (Europeana Team, 2010)

1.2 THE EUROPEANA SERVICE

The Europeana service harvests only the metadata describing a digital object. It provides a thumbnail with a link to the digital object's location on the provider's or aggregator's system. Users are always directed to the provider's or aggregator's site for viewing the digital object in detail.

Until present, the European Commission has co-funded, mainly through its CIP ICT-PSP Programme (Competitiveness and Innovation Framework Programme), a several European projects forming the Europeana Group (http://group.europeana.eu/web/guest). They facilitate metadata standardisation, interoperability and IPR of various heritage sectors. As an outcome, online access to the cultural and scientific heritage is improved by integration of the aggregated content into Europeana. Many of these projects develop their own portals providing access to digital resources in a more specific context.

For the moment, organizations from *twenty-nine countries* from EU plus Norway and Switzerland contribute to Europeana. They provide metadata about content in *twenty-six languages*, with four types of materials: image, sound, video, text.

2. EUROPEANA AGGREGATORS

A content provider for Europeana may be any organisation (and individuals, as well – for the future) that provides digital content accessible via Europeana. Europeana stores *only the institution's metadata and indexes it*, however, digital objects remain storing at the content provider. The goal of Europeana is to provide access to digital content of cultural heritage of thousands of Europeen cultural institutions.The main obstacle for it is the huge variation of type of content, technical infrastructures and output formats available of these content providers. Thus, due to the enormous amount of work for metadata harmonisation and normalisation,

Europeana does not practically work individually with any content provider but rather works with an intermediate layer of *aggregators* of the content providers.

Definition: "An aggregator in the context of Europeana is an organisation that *collects metadata from a group of content providers and transmits them to Europeana*" (Europeana Team, 2010)

Functions: Aggregators are supposed to perform activities of several types, according to the Europeana guidelines and procedures, as follows:

A. Primary functions of aggregators:

- Gathering content from individual organisations;
- Standardizing file formats and metadata;
- Delivering metadata to Europeana;
- Supporting the content providers with administration, operations and training.

B. Secondary functions of aggregators (Europeana Team, 2010):

- "Disseminating the vision and objectives of Europeana to their network of institutions in order to increase support for and involvement with Europeana
- Providing valuable feedback about the issues and discussions from their field
- Promoting and implementing standards further along the content provision chain
- Providing domain specific expertise and skills to institutions and Europeana".

Types of aggregators:

- Public vs non-public - an aggregator may or may not have a portal where the content is made accessible to the public. If the aggregator's portal is not accessible to the public, it is referred to as a 'dark portal' or a 'dark aggregator.'
- Digital repository aggregators (act as a repository storing digital items) vs intermediary aggregators (only collect metadata with a link to digital items).

Benefits for aggregators:

- By harvesting and exposing the metadata to search engines Europeana enhances the access to digital cultural content
- Boost the traffic to the content providers
- Europeana groups geographically dispersed content into a single, coherent and contextual virtual space – it adds value to the content by juxtaposing related images, texts, videos and audio items

3. THE EUROPEANA SUBMITTING PROCESS

3.1 ROUTES AND CRITERIA FOR CONTRIBUTING TO EUROPEANA

Europeana works in partnerships with with aggregators (organisations, projects or portals) in order to harvest their metadata. Europeana frequently receives requests from local, regional or national institutions that wish to submit their data. These institutions are then guided by the Europeana Office to start contacting the appropriate aggregator for their content. However, if an institution cannot be redirected to an existing aggregator, the Europeana Office can also work directly with the individual institution, in situations in which there is no appropriate aggregator, as shown in fig. 2.

Organization have to select options to help Europeana to identify their route to providing content to Europeana. The route type for content providing by an organization may be one of the following:

- through a national aggregator,
- through a Europeana related project, or
- directly to Europeana.

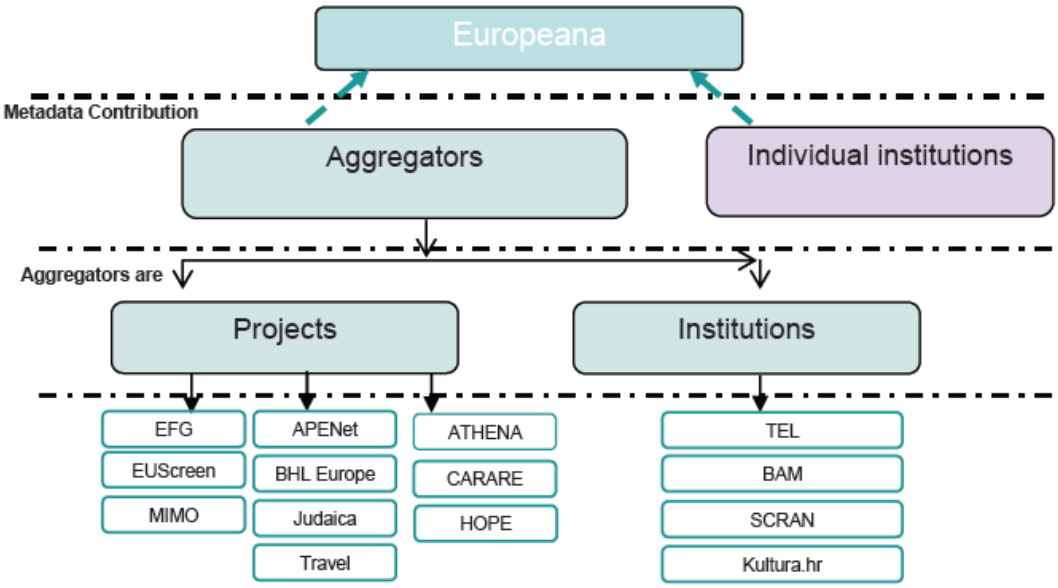


Fig. 2. Routes to deliver content to Europeana (Europeana Team, 2011)

The route to delivering content to Europeana is determined by four main criteria (Europeana Team, 2011):

a. **Country** - the national portals representing several sectors are the preferred first contact point for Europeana. Thus, a new content provider is first directed to the national aggregator, if such is already established; otherwise, another routing is to be suggested.

b. **Sector** - depends on the sector-specific activities, such as Museum/Gallery, Library, Archive, Audio Visual, Research & Education, Portal, Publisher, Private, Others.

c. **Scope** - relates to the scope of the material for which the organization will be collecting data, whether as an aggregator defined by three domains, or as an individual organization:

- Single aggregator: organisations representing and collecting data from one sector only, such as a museum, a library or an archive (e.g., a national/regional/local museum, national/regional/local library, national/regional archive),
- Cross-domain aggregator: organisations working across several sectors and collecting material from museums, libraries and archives (e.g., national portals and Europeana),
- Thematic aggregator: organisations bringing together content on a specific theme from one or several sectors (e.g., Judaica that collects material from multiple sectors about the Jewish urban culture),
- Individual organisation: the organisation is submitting data only from its own holdings and is not acting as an aggregator.

d. **Geographic level** - refers to the level at which an organisation operates in terms of the data it is collecting, such as regional, European or worldwide

The organization represent have to fulfill and to submit a Web form (see http://version1.europeana.eu/web/guest/providing-content/) selecting predefined values for the four main criteria stated over. Based on the information provided the Europeana will contact the organization within 5 working days to discuss the application. Next, Europeana will indicate whether the organization will contribute *directly* or *indirectly* to Europeana.

3.2 Steps for direct contributions to Europeana

Any direct data submission to Europeana implies that the organisation follows the steps described here (Europeana Team, 2011a). Therefore, this procedure also applies to the organisations (including aggregators) that have already submitted data to Europeana in the past.

Fig. 3 depicts the six steps for direct contributions to Europeana:

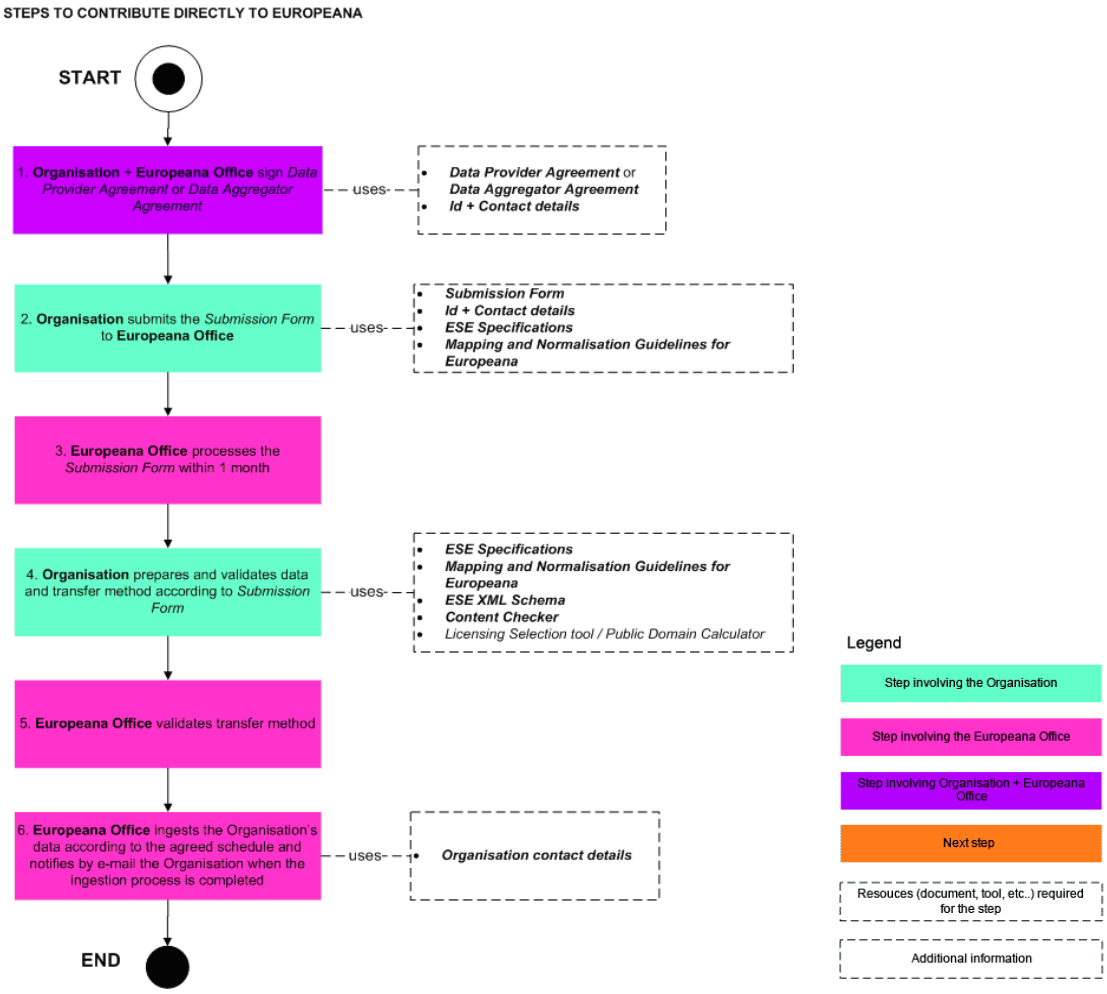


Fig. 3. Direct data submission workflow (Europeana team, 2010)

1. Step 1 - the aggregator and the Europeana Office formalize their collaboration by signing the Data Aggregator Agreement.
2. Step 2 - the *Europeana Content Ingestion team* sends the Data Submission Form to the aggregator, through which it provides a description of the submission it wants to make. This description includes type of submission (new/update), licensing and metadata information related to this particular submission, transfer mechanism that will be used, etc.
3. Step 3 - within one month (or, for urgent cases, even faster), the *Europeana Office* reviews the submission request, gives feedback to the aggregator and arranges the planning for the submission of the data.
4. Step 4 - in the next stage, the aggregator prepares the data sets to be submitted using the ESE (Europeana Semantic Elements) specifications (Europeana v1.0, 2010) - *for the current version of the services planned to be operational until the end of 2011*). The Mapping and Normalisation Guidelines (Europeana v1.0, 2010a) are available to support the mapping exercise. The XML Schema (Europeana v1.0, 2009) is available to validate the mapping to ESE (as explained in the next section). In addition to the previous resources, the aggregator can use the Content Checker (Europeana v1.0, 2010b) - a web tool allowing the validation of mapping against the XML Schema, and to simulate the ingestion operations so that the search, browse and display of the data in a copy of the Europeana portal can be tested. In parallel to the data preparation steps, the aggregator sets up and/or configures the data transfer infrastructure and test it. Most typically, Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is used (Reis and Freire, 2008)).
5. Step 5 - after the aggregator has tested the data sets using the Content Checker tool, the Europeana Office validates transfer methods with the aggregator.
6. Step 6 - the Europeana Office runs the operations aiming at ingesting the data into the Europeana production environment. When the operations are completed, the aggregator receives a notification from the Europeana Office. Final steps after the data have been published on the Europeana portal include the aggregator's checking of their data in production to ensure that the data are presented in the desired way.

During the submission procedure, the content provider is supposed to interacts with the Europeana representatives as follows:

- The *Europeana Business Development team* is responsible for initial assessment of the Partner Application Form, assistance to the new content provider as well as the signing of contractual agreements.
- The *Europeana Operations team* is responsible for the ingestion of the data sets according to the information provided to them on the submission forms. They provide support and guidance and do the harvesting and quality checking of the data provided. They also give aggregators feedback about the data.
- The *Europeana Office* has a liaison point for each project to facilitate the communication and coordination with Europeana. The Europeana Projects Liaison team is responsible for the alignment of planning, identification of contacts and the transmission of key information on both sides.

3.3 Indirect contributions to Europeana

The aggregator or project will provide the organization with the necessary requirements.

3.4 EUROPEANA LOCAL TECHNICAL IMPLEMENTATION STEPS

This section presents in brief technical implementation steps to be conducted by any content provider organization which wishes to contribute organization's content to Europeana Local throughout the project. You may find more information about them in (Bergheim S. et al, 2008).

The process includes several work tasks:

1. Constitution of digital collection(s)
2. Repository installation
3. Metadata Transformation – includes:
    a. Metadata Extraction
    b. Metadata Normalization
    c. Metadata Enrichment
4. Repository Population
5. Metadata Harvesting
6. Usage of aggregator repositories
7. Starting end-user services

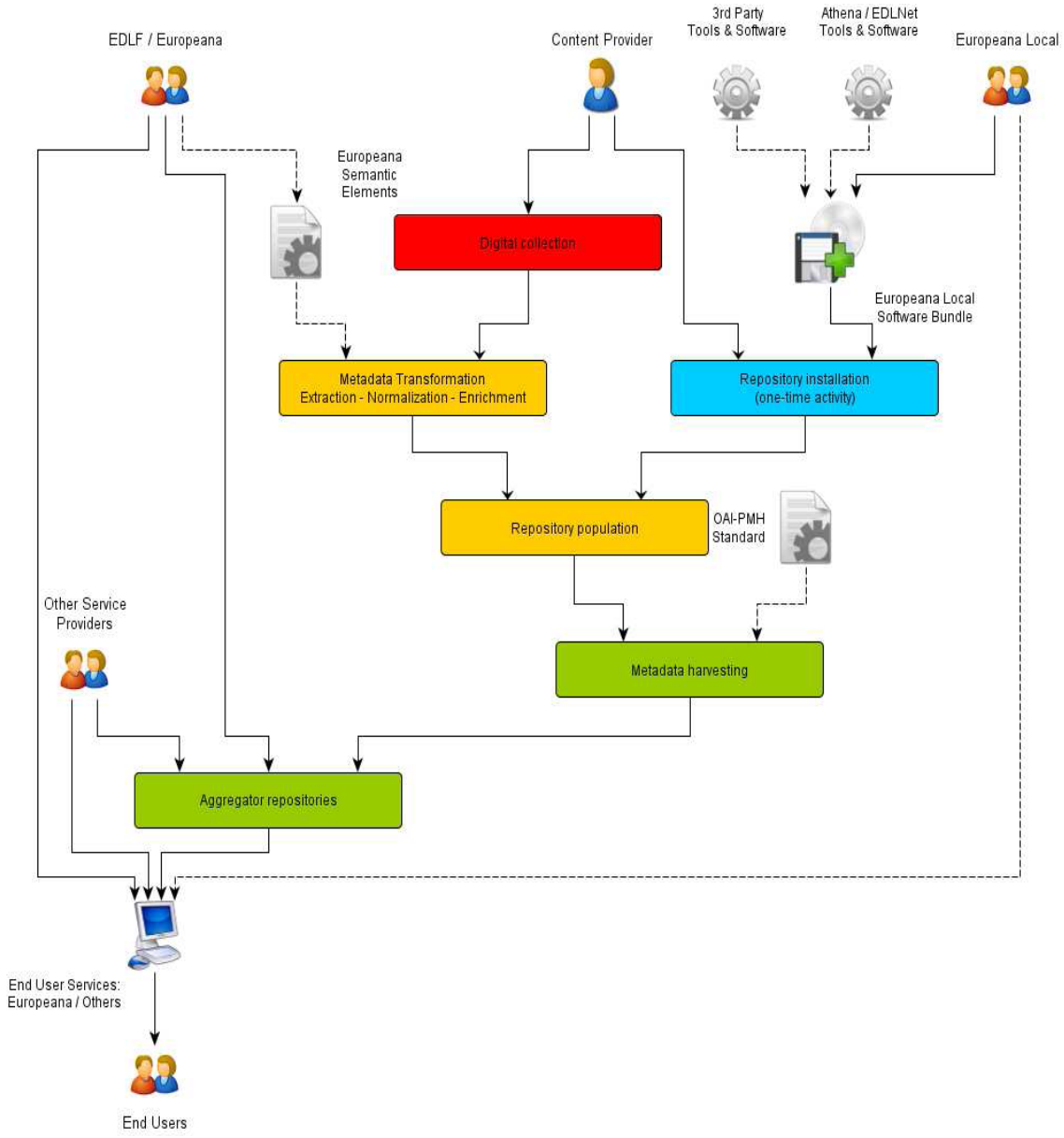


Fig. 4. Workflow of implementing OAI-PMH compliant repository solutions on top of digital collections held by local content providers (Bergheim S. et al, 2008)

The workflow including tasks listed over is presented in fig. 4. The red boxes show existing collections while the blue boxes present one-time activities such as repository installation. On other side, orange boxes depict tasks to be performed by content providers, while green boxes show tasks to be executed by Europeana. Finally, solid lines present direct participation and stippled lines signify indirect contribution such as supporting documents and standards.

It folllows a short description of each of these tasks. More information may be found in Europeana Local Implementation Manual – "a comprehensive document which forms the basis for technical workshops and which serves as a reference during implementation of OAI-PMH compliant repositories".

3.4.1 Constitution of digital collection(s)

Europeana Local uses the principally abstract term digital collection for content stored within the original project application. Practically, it may include also content from any new data sources throughout the project execution. From a technical point of view, a digital collection may be extracted from the same set of tables in one database without heed to thematic divisions based on content types. Note that *more than one collection may be added to a local OAI-PMH compliant repository through the same metadata extraction/repository population operation.*

3.4.2 Repository installation

The next step of the workflow for implanting a local collection into the Europeana infrastructure is to enable repository software locally – either to install it or to configure already existing repository software in the local collection management system. The task involves several steps:

- Choosing suitable repository software for the existing software and hardware infrastructure;
- Installing the repository software on local hardware;
- Configuring the repository to work with existing or embedded web server functionality and ensuring OAI-PMH web service protocol responds in a standards compliant manner to any requests from external harvesting repositories

3.4.3 Metadata Transformation

Metadata transformation is a complex process of "*moving data from its source collection management system, mapping to a target metadata profile, normalizing content attributes and enriching the metadata through various automatic and semi-automatic techniques*". Let describe each of these sub-tasks below.

Metadata extraction from local collection management system to local repositories - may require a lot of manual customization work and low level programming, depending on support of standards. A possible solution is to write simple SQL-scripts connecting to the data source (SQL-compliant RDBMS) and reading metadata directly from the physical data structure. Next, metadata will be transformed into an element set defined by the European Digital Library Foundation called Europeana Semantic Elements (ESE) - note that until the end of 2011 it should be substituted by Europeana Data Model (EDM) as described in next sections. ESE a simplified application profile based on qualified Dublin Core with the addition of a few custom elements supporting extended functionality of Europeana end-user service and is described in the next section.

Metadata normalization is a process of harmonizing attributes expressing the same information but in different formats. It is performed for ensuring the data required to support Europeana functionality is present in each record and, thus, to allow interoperability between content of heterogeneous sources.

Metadata enrichment is a process of improving the quality of metadata for using it out-of-context and together with metadata from other sources e.g. including multi-lingual content, spatial references, temporal references, or mapping to common vocabularies and identification of authorities. Here, various appropriate standards, tools, and technologies may be used such as:

- Dictionary description notations like SKOS
- Translation utilities
- Indexing and keyword extraction utilities
- Location extraction utilities
- Date extraction utilities

Practically, most metadata enrichment techniques involve some manual effort on the part of the content contributor because automated techniques may lead to some errors. Therefore, some quality procedures could be used.

3.4.4 Repository Population

Once metadata is exported from the content source system, the next step is to setup automatic loading of the metadata into the metadata repository, upon a schedule defined according the update frequency of the local digital collection. If the collection is finite (frozen), then loading of metadata can be done once.

3.4.5 Metadata Harvesting

After completion of all the previous steps, harvesting repositories can connect to your local repository, issue queries and download metadata content as authorized by the contributor to Europeana in the role of repository manager. An external web service connects to organization's local repository web service upon given schedule and downloads either all metadata or delta sets containing last changes such as inserts/edits/deletions since previous download. See (Reis and Freire, 2008) for details about OAI-PMH implementation and tools guidelines.

3.4.6 Usage of aggregator repositories

The aggregator repositories include the underlying database of the Europeana Service and some national portals which base their services on harvesting via OAI-PMH required for enabling local content for the Europeana infrastructure. Aggregator repositories can use the same technology as the source repositories installed in contributing organization but will contain also a web service requester for scanning a list of source repositories according to a preset schedule. Thus, Aggregator Repositories get their content from harvesting source repositories rather than from local collection management systems.

3.4.7 Starting end-user services

Europeana end-user services ensure cross-domain and cross-geography searches. Many local repositories benefits an increased quary traffic generated by the Europeana end-user services.

As readers may note metadata transformation is a rather complex process including extracting, mapping, normalizing and enriching metadata according either Europeana Semantic Elements (ESE) notation (the only one supported by Europeana until the end of 2011) or by the relatively new Europeana Data Model (EDM) being still under development. Next sections are devoted to the specifications of ESE and EDM and their benefits and shortcomings.

4. Europeana Semantic Elements

The present section of the document presents the Europeana Semantic Elements (ESE) according to their specification version V3.2.2 (Europeana v1.0, 2010). This metadata set was developed for the prototype version of Europeana which is operational since November 2008 until present moment. It is formed as an enriched

application profile of Dublin Core metadata[1] and provides a generic set of terms appropriate for to heterogeneous digital objects.

4.1 ESE GOALS

In order to provide common access to European collections of digital objects, Europeana harvests and indexes ONLY descriptive metadata associated with the digital objects. It is a matter of fact, it does not exist any universal metadata standard which could be applied across domains of culture and arts. That's why a set of metadata elements has been selected by the Europeana team in other to facilitate user functionality of an integrated world-wide search of digital objects.

Contributors to Europeana must provide metadata in the ESE format (before the new format of EDM enters in use as described in next chapter). For this purpose, contributors are supposed to map **elements from their own metadata format to the ESE metadata elements**. Next, a **normalization process** should be performed on some values to enable machine readability – two tasks which are transmitted from the Europeana Office to content providers or aggregators.

Thus, Europeana provides to aggregators an XML Schema as a tool for assisting providers in ensuring compliance with ESE (Aloia, 2009). There are developed Guidelines for the normative specification for ESE metadata found in ESE v3.2.2 (Europeana v1.0, 2010). The XML schema checks for basic conformance to this specification.

Note that it is an extension of the DC terms schema and accepts all DC terms. Providers should include in their metadata only the elements defined in the specification.

4.2 ESE STRUCTURE AND RESTRICTIONS

Europeana Semantic Elements (ESE) V3.2 was developed specially for the first prototype of Europeana in 2008. It is a Dublin core-based application profile and serves for cross-domain schema for heterogeneous data intended not to capture the full semantics of provider's data. ESE contains:

- 37 Dublin Core terms – used principally to describe the objects and incorporates elements from both the dc and dcterms namespacesplus some locally coined terms which have been added specifically to support functionality in Europeana. A full alphabetic declaration of these terms can be found in the ESE V3.2.2 Specification
- 12 Europeana coined terms - used to support portal functionality and working consistency

The list of elements is separated into four groups building the four columns in table1 (Europeana v1.0, 2010a).

[1] http://dublincore.org/documents/

Table 1. ESE elements

Strongly recommended	Recommended	Additional elements	Europeana elements
dc:title dcterms:alternative dc:creator dc:contributor dc:date dcterms:created dcterms:issued	dc:coverage dcterms:spatial dcterms:temporal dc:description dcterms:isPartOf dc:language dc:publisher dc:source dc:subject dc:type	dc:format dcterms:extent dcterms:medium dc:identifier dc:rights dcterms:provenance dc:relation dcterms:conformsTo dcterms:hasFormat dcterms:isFormatOf dcterms:hasVersion dcterms:isVersionOf dcterms:hasPart dcterms:isReferencedBy dcterms:references dcterms:isReplacedBy dcterms:replaces dcterms:isRequiredBy dcterms:requires dcterms:tableOfContents	europeana:country europeana:hasObject europeana:isShownAt europeana:isShownBy europeana:language europeana:object europeana:provider europeana:type europeana:unstored europeana:uri europeana:usertag europeana:year

Note: dc and europeana namespaces both have “Type” and “Language” elements, these have different uses

The ESE v3.2 Schema defines several restrictions (Aloia, 2009), as follows below. The ESE v3.2 Schema doesn't accept the following elements (Europeana office is responsible for providing all these elements):

- europeana:country
- europeana:hasObject
- europeana:language
- europeana:uri
- europeana:usertag
- europeana:year

The values for the following elements must be valid URI:

- europeana:isShownBy
- europeana:isShownAt
- europeana:object

The values for europeana:type element must be one of the following:

- TEXT
- IMAGE
- SOUND
- VIDEO

While DC elements can appear in a user preferred order, in the instances of documents, the Europena elements must appear after all DC elements and in the following order:

- europeana:unstored
- europeana:object
- europeana:provider (to be replaced by europeana:dataProvider)
- europeana:type
- europeana:isShownBy
- europeana:isShownAt

The europeana:isShownBy and europeana:isShownAt elements may appear in any order (one of the two is mandatory) after the europeana:type element.

ESE Field	Value
dc:identifier	02-Lemba-Lakkous-Lady.tif
dc:title	Stone figurine
dc:subject	Cyprus--Antiquities
dc:description xml:lang="en"	Stone vessel from Choirokoitia with emobssed decoration...
dc:publisher	Cyprus Archaeological Museum
dc:type	Image
dc:format	image/tiff
dcterms:temporal	mid 3rd millenium B.C.
dc:rights	Cyprus Archaeological Museum
dc:title	Λίθινο ειδώλιο
dc:description	Λίθινο ειδώλιο από τη θέση Λέμπα ...
europeana:provider	Cyprus Library
europeana:type	IMAGE
europeana:rights	http://creativecommons.org/licenses/by/3.0/nl/
europeana:dataProvider	Cyprus Archaeological Museum
europeana:isShownBy	http://www.mcw.gov.cy/mcw/da/da.nsf/All/8308A7AB9CA6CB5EC22574CC003FF5E1/$file/02-Lemba-Lakkous-Lady.jpg?OpenElement
europeana:uri	http://www.europeana.eu/resolve/record/**00000/E2AAA3C6DF09F9FAA6F951FC4C4A9CC80B5D4154**
europeana:country	cyprus
europeana:language	gr

xmlns:dc = "http://purl.org/dc/elements/1.1/"
xmlns:europeana = "http://www.europeana.eu"

Fig. 5. Sample ESE test record (Haslhofer and Isaac, 2011)

Extra note on DC terms elements and ESE conformance: The XML schema checks for basic conformance to the "Europeana Semantic Elements specifications", document available at http://version1.europeana.eu/web/europeana-project/provide-content.

Note however that it is an extension of the DC terms schema and will therefore accept all DC terms - even those that are not importable into the portal. Contributors have to include in their metadata only the elements defined in the specification.

The document (Europeana, 2010a) provides mapping guidelines, i.e. general principles of mapping practice for the Europeana Portal. It also offers some mapping examples and additional guidance for mapping information that is a difficult fit with ESE. The general rules are as follows:

1. Map as many as possible of the original source elements to the available ESE elements.
2. If it is not possible to map the source element to an appropriate ESE element then leave it unmapped or consider using <europeana:unstored>.
3. If possible use one of the more specific <dcterms> refinements (as indicated in the text) but only when the semantic of the source term clearly corresponds to the narrower term.
4. The persistent link to digital object and/or full information page should be given as a URL. These may need to be constructed from metadata values and information external to the metadata.
5. If it is difficult to decide which ESE element to map a source term to, consider how best to meet expectations of the user and the functionality of the system.
6. Where there are multiple values for the same element repeat the element for each instance of the value.
7. Consider how the data would perform in response to "who, what, where and when" queries. This therefore encompasses names, types, places and dates relevant to the object and what it depicts.
8. To ensure that your data will be meaningful when displayed in the new context consider adding a prefix or suffix. As a simple example, "100 x 200" could become "100 x 200cm". Such additions are especially important where data from several source elements have been aggregated into one ESE element.

4.3 Content Ingestion

Europeana has provided a Content Checker tool (Europeana Team, 2010a) which has two parts:

- The Content Ingestor - allows uploading of a data set (i.e., sets of records); as well, provides validation against the ESE V3.2 XML schema. Next, its imports the data into the database and does indexing of data plus caching of thumbnails
- The Test Portal - it is separated from the operational portal and enables data providers and aggregators to search and browse uploaded data by means of a visual interface

The Content Checker has been updated to accept ESE v3.3. The two new elements, europeana:dataProvider and europeana:rights can now be included in the metadata.

4.4 Shortcomings of ESE

ESE (Europeana Semantic Elements) is used since 2008 for the first version of Europeana. It represents lowest common denominator for object metadata, by converting datasets to Dublin-Core like standard. Thus, ESE forces interoperability but is very based solution for it and has several shortcomings. The major drawback here is the loss of original metadata. Other drawbacks of ESE (Haslhofer and Isaac, 2011) are as follows:

- ESE is a “flat” model that uses simple string values for the metadata fields and thus, it prevents properly linking an item ingested by Europeana to other objects (e.g., a series of portraits), or to contextual entities as represented by complex resources, e.g., a creator with many naming variations, or a concept that is more specific than another one, all of which could help improving access to Europeana items
- ESE aggregates in one and the same record metadata fields that can apply to different entities, breaking the “one-to-one” principle. Therefore, it causes great confusion: for example, some providers use rights- or daterelated fields to give information for the “real-world” object they hold, while others use the same fields for data on the digital representations of these items.

5. EUROPEANA DATA MODEL

5.1 EDM GOALS

In last four years, a new, RDF-based, Europeana Data Model (EDM) is yet under development aiming at replacement the old the ESE schema in order to solving the shortcomings explained over. Europeana team expects it will enable Europeana's transition (in Europeana Version 2.0) from a closed data repository to an open Web 3.0 information space of Linked Data (Haslhofer and Isaac, 2011).

There may be encountered several key goals of EDM design, namely:

- data integration
- preservation of original data while still allowing for interoperability
- semantic Web representation
- support rich functionality (e.g., semantic search)
- optimize the use of resources in time

There are several ways to reach these goals, as follows:

- ability to distinguish a “provided object” (being a painting, book, manuscript, etc.) from its digital representations
- possibility of several descriptions of a same item to co-exist in Europeana
- support for complex item representation, with sequence or partition information
- semantic linking between objects.

Moreover, EDM uses three fundamental design principles (Europeana v1.0, 2011):

- D1. EDM allows data integration in an *open* environment: it is impossible to anticipate all data contributed
- D2. EDM allows for rich functionality, possibly via extensions
- D3. EDM should re-use existing (standard) models as much as possible

Therefore, EDM is a new attempt for structuring the data that Europeana will be ingesting, managing and publishing. It follows a description of its structure.

5.2 EDM STRUCTURE

EDM sticks to the modelling principles supporting the Web approach Linked Data ("Semantic Web"). There is no fixed XML schema that restricts only one way to represent data. EDM can be seen rather as an interoperable mean for attaching various finer-grained models can be attached with preserwing original expressivity and richness of data. EDM is not built on any particular community standard but rather adopts an open, cross-domain Semantic Web-based framework that can accommodate particular community standards such as LIDO, EAD or METS.

There are defined *seven* base requirements about EDM (Europeana v1.0, 2011), namely:

- R1. distinction between "provided object" (painting, book, movie, archaeology site, archival file, etc.) and digital representation
- R2. distinction between object and metadata record describing an object
- R3. multiple records for the same object should be allowed, containing potentially contradictory statements about an object
- R4. support for objects that are composed of other objects
- R5. compatibility with different abstraction levels of description (e.g. if a provider wishes to submit descriptions that follow the distinctions introduced in FRBR Group 1
- R6. EDM provides a standard metadata format that can be specialized
- R7. EDM provides a standard vocabulary format that can be specialized

EDM relies on the following standards in order to fulfill these seven requirements:

- Open Archive Object Reuse and Exchange Model (OAI-ORE[2]), a reference model for the description and exchange of aggregations of Web resources for organization of metadata about an object – for Requirements 1-4 - specified with an RDF model, by four key notions (as RDF classes):
    - Object: the book/painting/program being described
    - Aggregation: organizes object information from a particular provider (museum, archive, library) – *EDM uses multiple aggregations for multiple providers*
    - Digital representation: some digital form of the object with a Web address
    - Proxy: the metadata record for the object – the ORE ***proxy*** mechanism allows several descriptions for a same provided item to co-exist.

- Dublin Core for metadata representation - for Requirement 5 - EDM uses the latest version of DCMI Metadata Terms[3], specified with an RDF model, with a specialization of 15 original DC elements such as dcterms:coverage, dcterms:spatial and dcterms:temporal

- W3C SKOS[4] model for Knowledge Organization Systems for vocabulary representation - for Requirement 6 - specified with an RDF model

- OAI ORE, Dublin Core and SKOS together fulfill Requirement 7

Thus, EDM allows flexible model with different semantic grains:

---

[2] http://www.openarchives.org/ore/

3 http://dublincore.org/documents/dcmi-terms/

4 http://www.w3.org/2004/02/skos/

- "preserve original data while still allowing for interoperability"
- Keep data expressed as close as possible to original model
- Using mappings to more interoperable level
- Preserving and exploiting original data also means being compatible with descriptions beyond simple object level - objects, events and the rest

Advanced modeling in EDM

- Relations between provided objects
- Part-whole links for complex (hierarchical) objects
- Derivation and versioning relations

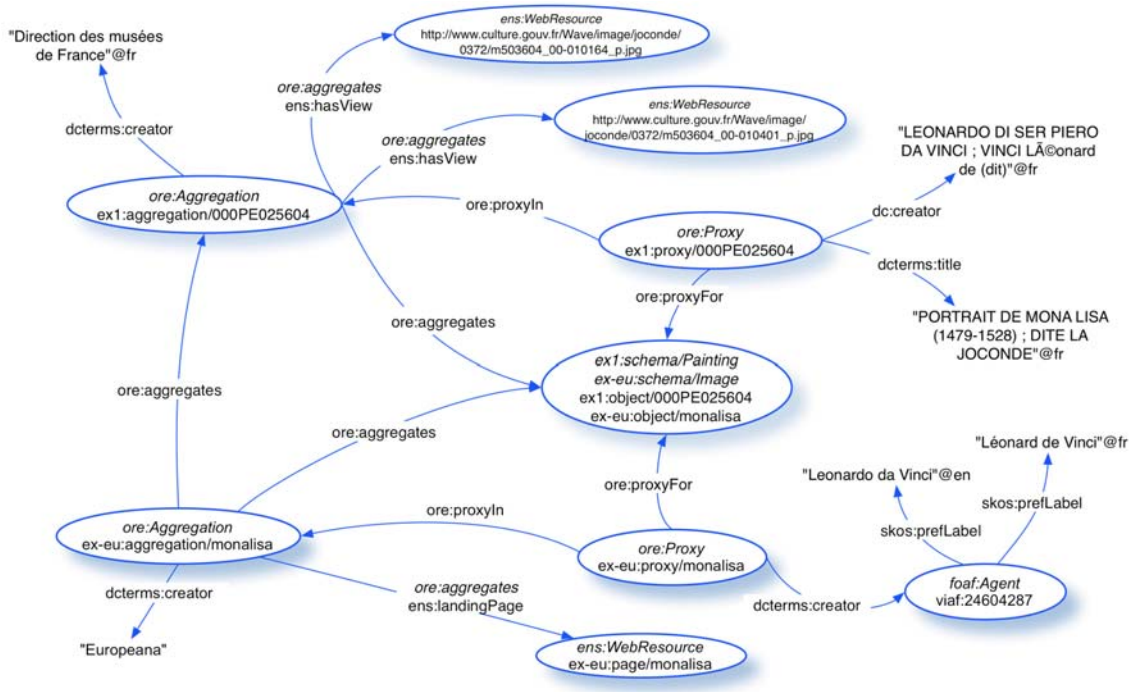


Fig. 6. EDM sample RDF

data.europeana.eu is the first Europeana EDM pilot system (Europeana Team (2011b). It serves several classes of resources (fig. 6):

- ***Item*** - such resources present an object (painting, book, manuscript, etc.) for which an institution provides a representation to be accessed through Europeana. Item URIs are the main entry points in data.europeana.eu. Descriptive metadata (creator, subject, etc.) is not directly attached to item URIs but instead is provided by the ***proxies*** that represent a view on the item. This view is made from a specific institution's perspective - from Europeana provider or from Europeana.
- ***Provider's proxy*** - such resources describe creator, subject, date of creation, etc for an ***item***, which are contributed by a Europeana provider. In the OAI-ORE model, proxies enable separating different views for a same resource, in the context of different aggregations. This allows us to *distinguish the original metadata for the object from the metadata that is created by Europeana*. Descriptive properties of these proxies usually are taken from the Dublin Core metadata standatd. Proxies connected to the item represent a facets – by means of the ore:proxyFor property. They are attached to the aggregation that contextualizes them, using ore:proxyIn. *Note: due to no support for the present moment for named graphs in the RDF standard, ORE introduces Proxies in order to support referencing resources in the context of a graph.*

- **Provider's aggregation** - these resources provide data related a Europeana provider's gathering digitized representations and descriptive metadata for an item. They are related to digital resources about the item - such as files directly representing it (ens:object and ens:isShownBy) or web pages showing the object in context (ens:isShownAt). They may also provide with controlled rights information applying to these resources (ens:rights).
- ***Europeana's proxy*** - the second type of proxies served at data.europeana.eu, Europeana proxies provide access to the metadata created by Europeana for a given item, distinct from the metadata provided by the provider. Here one can find ens:year statements, indicating a normalized date associated to the object. Finally, a proxy is connected to the item it represents a facet of, using the ore:proxyFor property, as well as to the aggregation that contextualizes it, using ore:proxyIn.
- ***Europeana's aggregation*** - a Europeana aggregation bundles together the result of all data creation and aggregation efforts for a given item. It aggregates the ***provider's aggregation*** (using ore:aggregates), which in turn will connect to the ***provider's proxy***. Next to the provider aggregation, one can find the digitized resources Europeana.eu serves for the item, i.e., an object page (ens:landingPage) and a thumbnail (using a combination of ens:hasView and foaf:thumbnail). The Europeana proxy is also connected to this aggregation, as mentioned above.
- ***Resource map*** - OAI-ORE Resource maps are constructs for indicating meta-level statements on the creation and publication of ORE data (ORE aggregations and their aggregated resources) - as a contextualization mechanism for the Europeana aggregation. Maps are connected to an item they are about using foaf:primaryTopic, and to its corresponding Europeana aggregation using ore:describes. They sum up the provenance of data using dc:creator and dc:contributor statements.

5.3 SHORTCOMINGS OF EDM

Shortcomings of EDM are to be identified by exploitation of EDM in the scope of Europeana 2.0.

6. MAPPING ESE TO EDM

There is defined a mapping between ESE and EDM that covers the creation of the EDM entities (items, aggregations, proxies), the assignment of dereferencable HTTP URI identifiers to these entities, as well as the attachment of metadata fields to these entities. Metadata values from legacy ESE records are distributed over the various EDM entities, and new ones are created.

The major difference between ESE and EDM is that ESE records are flat lists of property value pairs while EDM data are networked resources (Haslhofer and Isaac, 2011). Converting ESE data to EDM thus requires distributing ESE elements over EDM resources, each playing a different role in the EDM model. The major issue here was to identify the target EDM resource for a given ESE property since the one-to-one mapping principle does not apply. In many cases, indeed, providers have not strictly followed the ESE guidelines, and use values that apply to digital representations in fields that in theory would apply to the provided “real-world” item.

Another problem concerns the structure of EDM data itself. The requirement for distinguishing different data sources, and data that applies to different resources, leads to creating complex networks of aggregations, proxies and other resources. While this has many benefits, it also raises the barrier to data access and consumption.

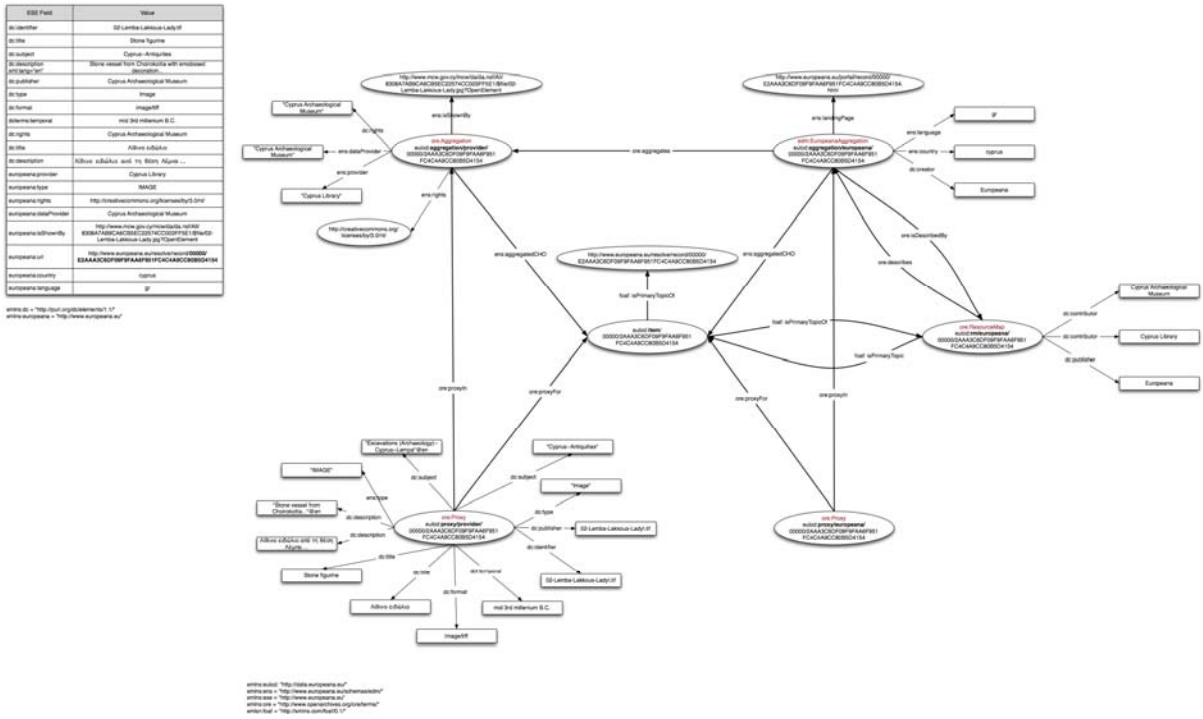


Fig. 7 ESE to EDM mapping

7. Conclusions

The launch of data.europeana.eu has set the first trial of making the Europeana metadata set available as Linked Open Data. data.europeana.eu includes data from content providers who wish their data to be published in the public domain. Next developments will emphasize even more on the EDM and on a smooth transition from ESE to EDM. Therefore, present contributors to Europeana who got used with the flat ESE metadata structure, should be prepared to EDM entrance until the end of the year.

References

A. Principal references

- Aloia N. et al (2009) Europeana Semantic Elements V3.2 XML Schema, Version: 3.2, 2009-08-07, http://www.europeana.eu/schemas/ese/ESE-V3.2.xsd
- Bergheim S. et al. (2008) EUROPEANALOCAL Technical Primer - Overview of Europeana Local Implementation Steps, Revised version 16.12.2008
- Carlo Meghini, Antoine Isaac, Stefan Gradmann, Guus Schreiber, et al. (2010) The Europeana Data Model: tackling interoperability via modelling, DL.org Autumn School, Athens, October 5, 2010
- Haslhofer, B., Isaac, A. (2011) data.europeana.eu - The Europeana Linked Open Data Pilot, Proc. Int'l Conf. on Dublin Core and Metadata Applications 2011 (to appears)

- Europeana foundation (2007) Articles of Association of the Europeana Foundation – "Stichting European Digital Libtrary", http://version1.europeana.eu/web/europeana-foundation/statutes.
- Europeana Team (2010) Europeana Aggregators' Handbook, Edition 1 May 2010.
- Europeana Team (2010a) Europeana Content Checker User Guide – revised for ESE v3.3, 1 September 2010
- Europeana Team (2011) Routes to deliver content to Europeana, http://version1.europeana.eu/web/guest/providing-content/
- Europeana Team (2011a) STEPS TO CONTRIBUTE DIRECTLY TO EUROPEANA, http://www.version1.europeana.eu/web/guest/stepstocontribute
- Europeana Team (2011b) Europeana LOD Intro.mht
- Europeana v1.0 (2009) Europeana Semantic Elements V3.2 XML Schema, Version: 3.2, 2009-08-07
- Europeana v1.0 (2010) Europeana Semantic Elements specifications, Version 3.2.2, 18/01/2010.
- Europeana v1.0 (2010a) Metadata Mapping & Normalisation Guidelines for the Europeana Prototype, Version 1.2.1, 18/01/2010.
- Europeana v1.0 (2010b) Europeana Content Checker User Guide – revised for ESE v3.3, 1 September 2010
- Europeana v1.0 (2010c) Metadata Mapping & Normalisation Guidelines for the Europeana Prototype, Version 1.2.1, 18/01/2010
- Europeana v1.0 (2011) Europeana Data Model Primer, 10/02/2011 – Editor's Draft
- Reis, D., Freire, N. (2008) OAI-PMH implementation and tools guidelines, ECP-2006-DILI-510003, TELplus D-2.1, 31 May 2008

B. OTHER REFERENCES

http://www.scribd.com/doc/32110457/Europeana-White-Paper-1
http://www.proxml.be/users/paul/weblog/473b6/Europeana_and_Linked_Data.html
http://www.openarchives.org/ore/1.0/primer
http://www.w3.org/TR/2004/REC-rdf-primer-20040210/
http://www.w3.org/TR/skos-primer/#secscheme
http://www.cidoc-crm.org/
http://dublincore.org/metadata-basics/
Gleaning **R**esource **D**escriptions from **D**ialects of **L**anguages - http://www.w3.org/2001/sw/grddl-wg/
RDFA-SYNTAX
RDFa in XHTML: Syntax and Processing (See http://www.w3.org/TR/rdfa-syntax.)
CC
Creative Commons (See http://creativecommons.org.)
DC
Dublin Core Metadata Initiative (See http://dublincore.org.)
FOAF
The Friend of a Friend (FOAF) Project (See http://www.foaf-project.org/.)
N3
Notation 3 (See http://www.w3.org/TeamSubmission/n3/.)

RDF
Resource Description Framework (See http://www.w3.org/RDF/.)
RDFHTML
RDF-in-HTML Task Force (See http://www.w3.org/2001/sw/BestPractices/HTML/.)
RDF-SCHEMA-PRIMER
RDF Primer - Section 5 on RDF Schema (See http://www.w3.org/TR/2004/REC-rdf-primer-20040210/#rdfschema.)
SWD-WG
Semantic Web Best Deployment Working Group (See http://www.w3.org/2006/07/SWD/.)
SWBPD-WG
Semantic Web Best Practices and Deployment Working Group
(See http://www.w3.org/2001/sw/BestPractices/.)
XHTML2-WG
XHTML2 Working Group (See http://www.w3.org/MarkUp/.)
http://www.w3.org/DesignIssues/LinkedData.html
http://linkeddata.org/guides-and-tutorials

Semantic Technologies for Cultural Heritage

Vladimir Alexiev, PhD, PMP
vladimir.alexiev@ontotext.com

Outline

- Semantic Web
  - Linked Open Data
  - Ontotext Corp
- Cultural Heritage
  - Applicability of Semantic Web
  - Data Standards
  - Europeana Data Model
- Related Projects
  - EU Projects
  - EU Projects with BG Participation
  - BG Projects
- ITCC Project

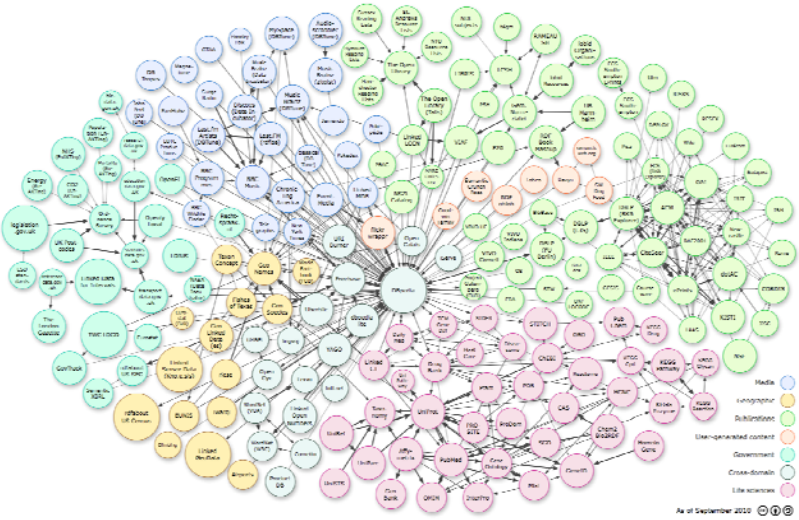
➢ 40 slides in 15 minutes???
  ➢ Half the slides are at the end (Backup section)
  ➢ You can read them offline

SEMANTIC WEB, LOD, ONTOTEXT

Semantic Web

- What is the semantic web?
  - "Artificial Intelligence meets the web": Recasting of AI developments, but with realistic goals and at global web scale
  - Proposed around 1998 by T.Berners-Lee (same person who proposed the WWW)
  - Standardized by the WWW Consortium (W3C)
  - If Web 1.0 is a giant hyperlinked **document**, then Web 3.0 is a giant linked **database**
- Emerged as the best Data Integration approach
  - Many professional fields have a data fusion problem
  - Semantic web becomes the "Rosetta Stone" linking various types of independent data sources, using a unified, globalized and abstracted representation (RDF, RDFS, OWL2, RIF)
  - Schema info (metadata) is represented the same way as data
- Intro videos (from the TED series)
  - Tim Berners-Lee: The next Web of open, linked data
  - Tim Berners-Lee: The year open data went worldwide
  - Hans Rosling: No more boring data

Linked Open Data



- Grown exponentially since 2008. Currently 207 datasets, 28 billion facts
- Provides useful background information (geographic, political, statistical)
- Any new data can be linked to LOD to benefit from background/context knowledge

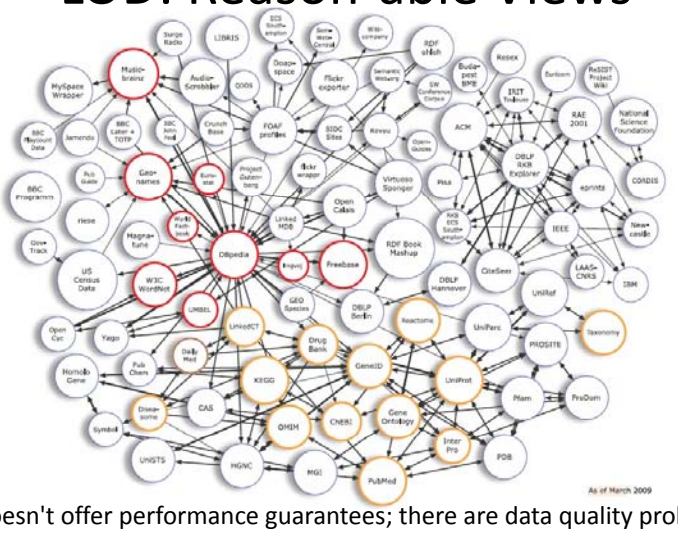
Ontotext Corp

- Who is Ontotext?
  - The leading semtech company in Eastern Europe and one of the leaders world-wide
  - 55 people: Bulgaria (Sofia, Varna), Austria, USA
  - Worked in this area since 2000
  - Venture funding and commercial clients since 2008
  - Bulgaria's most successful participant in EU FP5,6,7 research
  - www.ontotext.com
- 360-degree semantic technology:
  - Semantic Repository (OWLIM), ETL
  - Text Mining, Semantic Annotation and Search (KIM, U.Sheffield GATE, Teamware, MIMIR)
  - Web Mining and Crawling
  - Ontology Engineering and Exploitation
  - Master Data and Linked Data Management

Ontotext Clients (selected)

- British Broadcasting Corporation (BBC)
  - Runs its World Cup 2010 sites on top of OWLIM
  - Next is BBC Sports (2011) and the 2012 Olympics
- The National Archives (UK)
  - The UK Government's official archive contracted Ontotext to implement semantic search for the Government Web Archive
- AstraZeneca
  - Analysis and retrieval of clinical trial reports
  - Integration of biomedical databases for drug target identification
- KT
  - Applications of OWLIM in Interactive TV and IPTV
- All these clients need to:
  - integrate massive amounts of heterogeneous data
  - provide efficient search and navigation

LOD: Reason-able Views



- LOD doesn't offer performance guarantees; there are data quality problems
- Ontotext hosts 2 public services that aggregate, clean-up, and correlate large parts of LOD, and provide uninterrupted service:
  - Red: www.factforge.net (9 sets, 2.1B statements, 10B retrievable statements)
  - Orange: www.linkedlifedata.com (25 sets, 4.1B statements): use in biotech research

SemWeb Tipping Point

- Richard McManus (founder of ReadWriteWeb):
  - "…the tipping point for the Semantic Web may be when one can … deliver –using Linked Data– a comprehensive list of locations of original Modigliani art works …"
- Query http://factforge.net/sparql: uses data from 3 independent datasets (DBpedia, FreeBase, UMBEL)

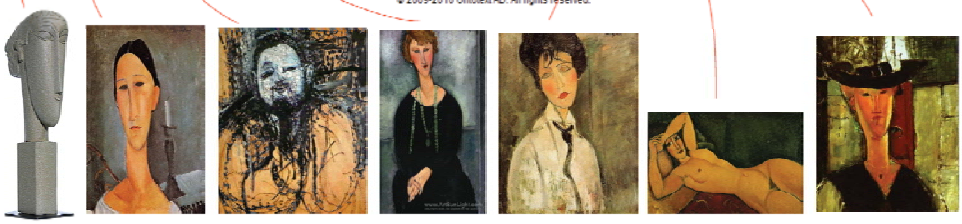
```
SELECT DISTINCT
  ?painting_l ?owner_l ?city_fb_con ?city_db_loc ?city_db_cit
WHERE {
  ?p fb:visual_art.artwork.artist dbpedia:Amedeo_Modigliani ;
     fb:visual_art.artwork.owners [
       fb:visual_art.artwork_owner_relationship.owner ?ow ] ;
     ff:preferredLabel ?painting_l.
   ?ow ff:preferredLabel ?owner_l .
    OPTIONAL { ?ow fb:location.location.containedby [
                  ff:preferredLabel ?city_fb_con ] } .
    OPTIONAL { ?ow dbp-prop:location ?loc.
               ?loc rdf:type umbel-sc:City ;
                    ff:preferredLabel ?city_db_loc }
    OPTIONAL { ?ow dbp-ont:city [ ff:preferredLabel ?city_db_cit ] }
}
```


Modigliani Test: FactForge Results

Results for PREFIX fb: <http://rdf.fr... (8) View as Exhibit Download in JSON | SPARQL Results in XML | SPARQL Results in JSON

painting_l	owner_l	city_fb_con	city_db_loc	city_db_cit
Head@en	Museum of Modern Art	New York City		
Anna Zborowska@en	Museum of Modern Art	New York City		
Portrait of Diego Rivera@en	The São Paulo Museum of Art@en		São Paulo	
Woman with a Necklace@en	School of the Art Institute of Chicago@en			Chicago
Portrait of a Woman@en	School of the Art Institute of Chicago@en			Chicago
Reclining Nude@en	Museum of Modern Art	New York City		
Madam Pompadour@en	School of the Art Institute of Chicago@en			Chicago
Jeanne Hébuterne@en	Barnes Foundation@en	Philadelphia		

© 2009-2010 Ontotext AD. All rights reserved.



- Reaching the tipping point of the semantic web
- Interestingly, with an application to cultural heritage

CULTURAL HERITAGE

Cultural Heritage

- Cultural Heritage (C.H.) represents
  - **artifacts**: historical, archeological, written, art, etc
  - **institutions**: museums, galleries, archives, audiovisual collections, libraries, trusts, universities, research institutes, …
  - **activities**: preservation, scientific research, cataloging, dissemination, digitalization, …
  - **fields**: artistic, scientific, industrial, bibliographic, ethnographic, nature …
  - A large, heterogeneous, cross-disciplinary field
- Social importance
  - The EU agenda includes multiculturalism and cultural preservation
  - Both BG and KR have rich history and heritage

Why Semweb for C.H.?

- Many C.H. institutions have a data integration problem, since data about the same artifacts is scattered in separate silos:
  - cataloging data
  - acquisition data
  - conservation data
  - scientific data, …
- Cross-institution and cross-country research is stuttered
  - Which BG archeological sites exhibit the same style as MK sites? Who knows…
- British Museum hosted a Cultural Heritage & SemWeb day (13.1.2011)
  - Examined use of the emerging web of data by c.h. and humanities organizations. See this interesting blog
  - Looked at practical applications of semweb technology relevant to the sector
  - Out of 8 presenters:
    - **John Sheridan** (Ontotext client): Government and National Archive initiatives (the 2nd most used web archive in the world is being converted to a Semantic Knowledge Base with 50M documents)
    - **Atanas Kiryakov** (Ontotext CEO): FactForge – the Fast Track to the Centre of the Data Web

PROJECTS RELATED TO CULTURAL HERITAGE & SEMTECH

Related Projects

- Numerous related projects are sponsored by the EU, in line with its multiculturalism agenda
  - Europeana
  - EU FP7 and DigiCult projects
- Some EU projects with BG participation
  - MOLTO
  - CULTURA
  - IMPACT
  - Linked Heritage
- BG Projects
  - SUdigital
  - Romulus, Theseus, Thelamon
  - Collaborative Portal

Europeana

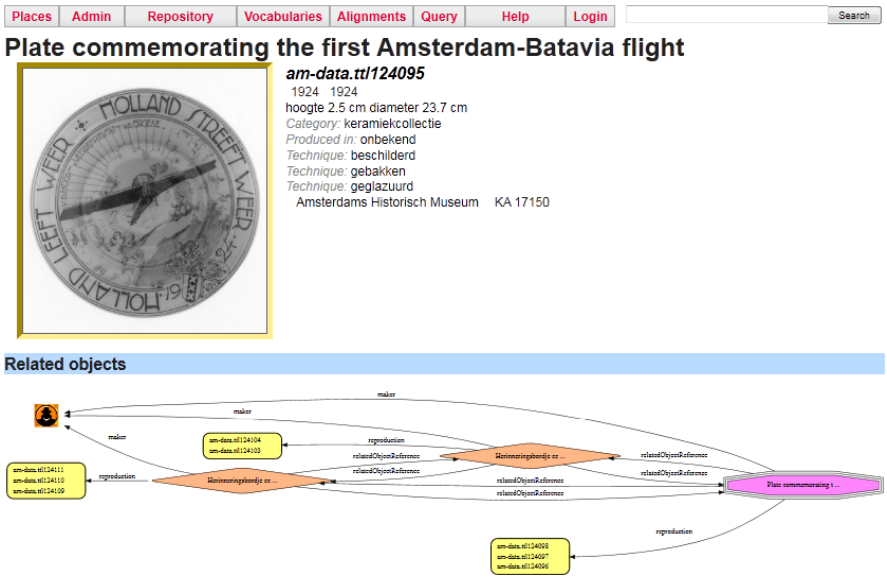
- http://europeana.eu
- Pan-European Digital Museum
- People can explore the digital resources of Europe's museums, libraries, archives, audio-visual collections
  - And be inspired by the rich diversity of Europe's cultural and scientific heritage
- Promotes discovery and networking opportunities in a multilingual space
- Around 1500 institutions contributed
  - Includes over 15 M items
  - Still none from BG, but that will change soon

europeana
think culture

Europeana and semweb

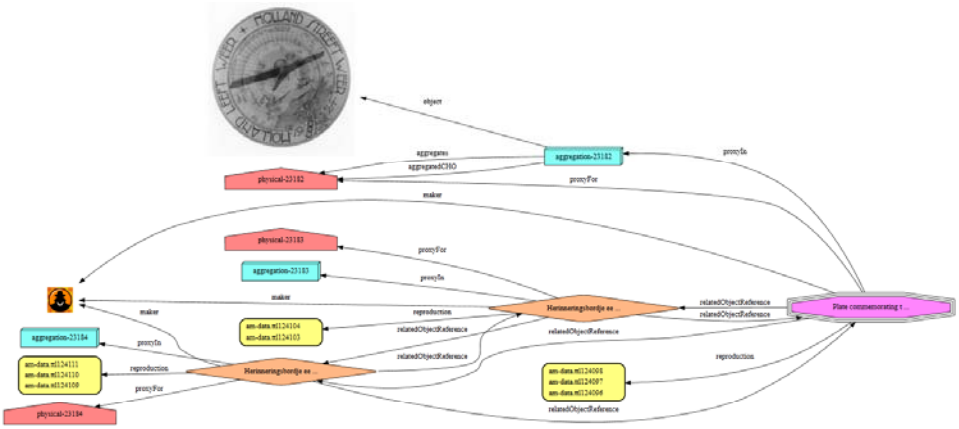
- Europeana adopts semtech for all future development
  - EDM is based on semweb, which makes it easily extensible
  - Current holdings will be converted to semweb format
- Europeana's first White Paper **Knowledge = Information in Context** looks at the key role LOD will play in Europeana's development
  - "Linked data gives machines the ability to make associations and put search terms into context. Without linked data, Europeana could be seen as a simple collection of digital objects. With linked data, the potential is far greater"
- LOD Experiment in Europeana Thought Lab:
  - Data and metadata from the Amsterdam Museum (AM)
    - 70k cultural heritage objects related to the city
  - Retrieved from an XML API of the museum's collection (database by AdLib UK)
  - Converted to LOD compliant with EDM by Vrije Universiteit (VU): 5M facts
    - Conversion software is available as open source
  - Linked to existing datasets
    - Dutch Art and Architecture Thesaurus (AATNed)
    - Getty's Union List of Artists Names (ULAN)
    - DBpedia (data version of Wikipedia)
    - GeoNames (structured geographical data)
  - Published to the Comprehensive Knowledge Archives Network (CKAN)
  - Prime example the type of LOD that Europeana will be producing

Amsterdam Museum object



- http://semanticweb.cs.vu.nl/europeana/browse/list_resource?r=http://purl.org/collections/nl/am/proxy-23182

Museum Object Relations



- Click on any node to see its object and relations, e.g.
  - The plate was made by "Heytze, J.C."
  - It is related to another plate (see on right), etc
- This is the essence of LOD!

EU FP7 and DigiCult

- DigiCult and TeLearn
  - **DigiCult (Cultural Heritage)**: leading-edge ICT for accessing, experiencing and preserving cultural and scientific resources
  - **TeLearn (Technology Enhanced Learning)**: how ICT can make learning more efficient
- FP6 Projects
  - 25 research projects were executed to mid 2010
- FP7 projects and calls:
  - 11 projects from the first two calls, started between Nov 2007 and Feb 2009
  - 15 projects from the latest call, started between Dec 2010 and Apr 2011
  - ICT SME DCL: open till Apr 2011, Digital Content and Languages, targets SMEs
    - Ontotext submitted a proposal with VU Amsterdam and AdLib (UK) similar to the Europeana pilot
  - ICT call 9: open Jan-Jul 2012, targets Digitization
- Budget (specifically for DigiCult and TeLearn, other calls are also be applicable)
  - WP 2007-2008: 102 MEUR DigiCult and TeLearn (Calls 1 and 3)
  - WP 2009-2010: 69 MEUR for "Digital Libraries and Digital Preservation"
  - WP 2011-2012: 110 MEUR (Call 9, Challenges 4 and 8) for "Digital Preservation" and "ICT for access to cultural resources"

BG RELATED PROJECTS

SUDigital

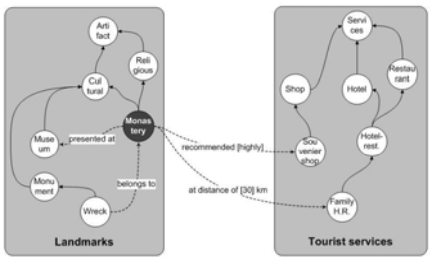
- Digital Library of Sofia University (sudigital.org)
- A common platform and presentation portal for digitization initiatives at SU (e.g. see next 3 such initiatives)
- Cross-disciplinary effort. Current participants:
  - University Library,
  - Faculty of History,
  - Faculty of Theology,
  - Faculty of Journalism,
  - Faculty of Philology (Slavic, Western, Eastern, Ancient languages)
- Built using the CDS Invenio digital repository by CERN (CH), currently 1k items
- Extensible variety of document types
  - Currently extended to epigraphics (stone inscriptions).
  - A universal system for describing and indexing c.h. items that is unique for BG
  - Adheres to the spirit of Europeana and hopefully will soon contribute to Europeana

Romulus, Theseus, Telamon

- Digitization projects of the Faculty of Philology
- **Romulus Bulgaricus** (romulus-bg.net): bilingual library of Latin texts and their Bulgarian translations
  - Uses XML tags and an automatic aligner between the two texts
  - Presentation by books and chapters, original text and/or one of the translations, parallel view
  - Electronic Encyclopedia of the Ancient world and an image gallery
- **Theseus Bulgaricus** (theseus.proclassics.org): bilingual library of Ancient Greek texts and their Bulgarian translations
- **Telamon** (telamon.proclassics.org): library of Ancient Greek inscriptions in BG. Epigraphic heritage is very rich in BG
  - Universal system for XML tagging of text and archeological data about epigraphics
  - Conforms to the tag set of Text Encoding Initiative (TEI)
- These projects are fully integrated into the SUDigital portal

SU Collaborative Portal

- Developed at Sofia University
- Collaborative creation of data about immovable monuments (monasteries, archeological sites, nearby tourist services, …)
- Creation of new content types (bottom). Inter and intra-ontology relations (right)
- Uses Google Maps for geo-referencing of sites
- Described in "Ontology-Based Content Development In Collaborative Environments with Semantic Services" (B.Bontchev, R.Jardim-Gonçalves)



GUI control	Content type	Mnd	No	ID	BG title	EN title	BG values	EN values
Label	String	Yes	1	type	Тип	Type	манастир	monastry
TextBox	String	Yes	2	title	Име	Title		
CheckBoxList	String	Yes	3	active	действащ	active	да,не	yes,no
TextBox	String	Yes	4	established	основан	established		
TextBox	String	Yes	5	renovated	обновен	renovated		
TextBox	Other	Yes	6	patronalFeast	храмов празник	patronal Feast		
DropDownList	String	Yes	7	femaleMale	тип	type	мъжки,женски	male,female
Multiline	Other	Yes	8	description	описание	description		
Image	Other	Yes	9	photo	снимка	photo	снимка на манастира	monastery photo

ITCC-SPONSORED PROJECT

Project Partners

- Sponsor: BG-KR ITCC (www.itcc-bg.com)
- BG Partners
  - Ontotext Corp
  - Sofia University
    - Faculty of Mathematics and Informatics
    - Faculty of History
    - Faculty of Classical and New Philology
- KR Potential Partners
  - Saltlux
  - Wonkwang University

Project Goals

- Semweb publishing for SUDigital
  - Map SUDigital content models to EDM (ontology mapping)
  - (optional) Enrich SUDigital with more item relations
  - Convert SUDigital holdings to LOD
  - Put the RDF data to a semantic repository
  - Publish the RDF data to Europeana
  - Use content negotiation to serve human-readable content to browser agents
    - Similar to Europeana pilot (Amsterdam Musem) and MOLTO (Gothenburg City Museum)
  - Streamline a semweb publishing pipeline for c.h. institutions
  - (optional) Contribute semweb publishing to Invenio by CERN (inventor of the web)

Project Goals (cont)

- Semweb publishing of data about immovable monuments
  - Map the Collaborative Portal data models to EDM
  - Enrich the Collaborative Portal with a semweb publishing pipeline
  - Publish data about immovable monuments to Europeana
- (optional) Extend and integrate
  - Integrate the Collaborative Portal to SUDigital to allow description of immovable monuments
  - Create in-depth Archeological ontology, in particular about Tombs and rich graves in Thracia and the Mediterranean area during Classical and Hellenistic ages

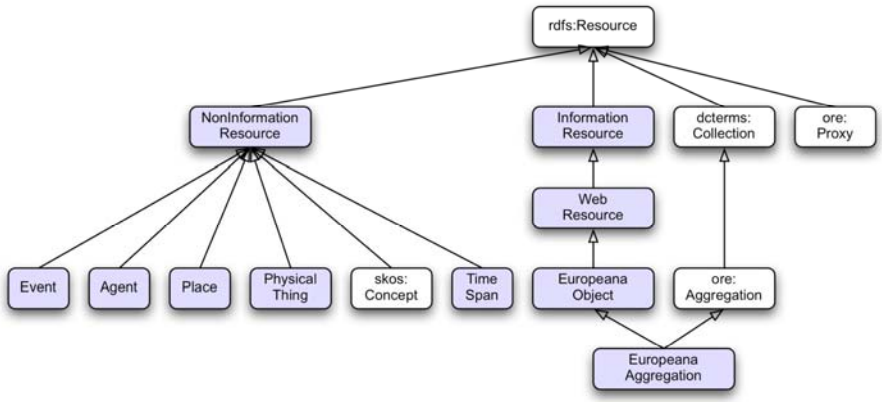
BACKUP SLIDES

CULTURAL HERITAGE DATA STANDARDS

C.H. Data Standards

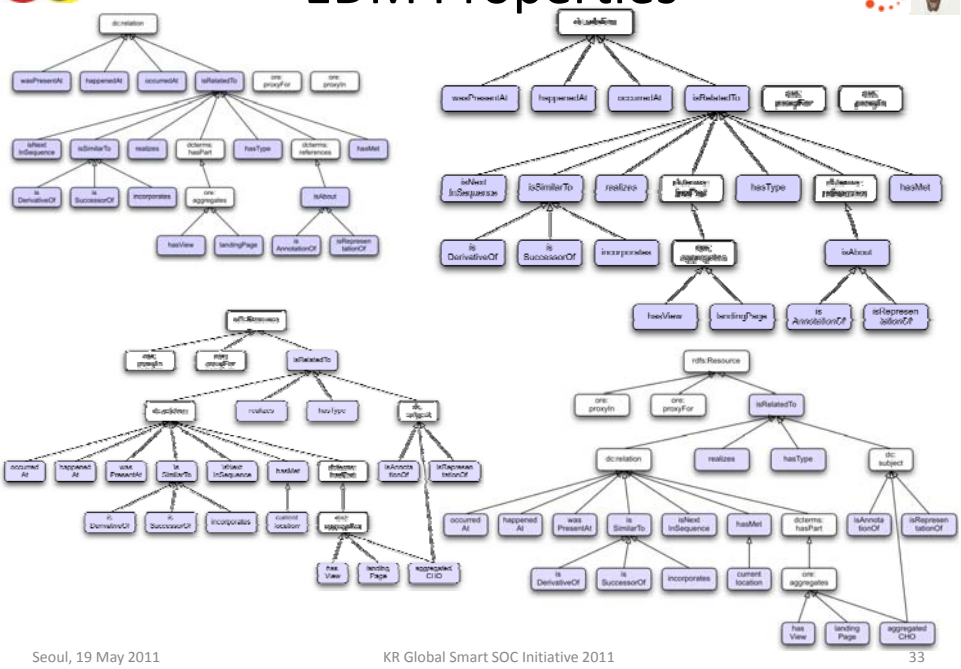
- Dublin Core (DC)
  - Simple resource descriptions: author, date created, etc
  - ISO 15836
- Conceptual Reference Model (CRM)
  - By International Committee on Documentation (CIDOC) of the International Council of Museums (ICOM), an organization devoted to conserve and communicate to society the world's natural and cultural heritage
  - Core upper-level concepts common across c.h. documentation
  - ISO 21127 (2006) "Information and documentation — A reference ontology for the interchange of cultural heritage information"
- Functional Requirements for Bibliographic Records (FRBR)
  - by International Federation of Library Associations and Institutions (IFLA)
  - Conceptual entity-relationship model
- Europeana Data Model (EDM)
  - Structuring the data that Europeana is ingesting, managing and publishing
  - Refined until the end of 2010, will be implemented during 2011
  - Open, cross-domain framework, adheres to semweb modeling principles
  - Intended to accommodate specific community standards such as LIDO, EAD, METS
  - Probably in the future will incorporate CRM

EDM Classes

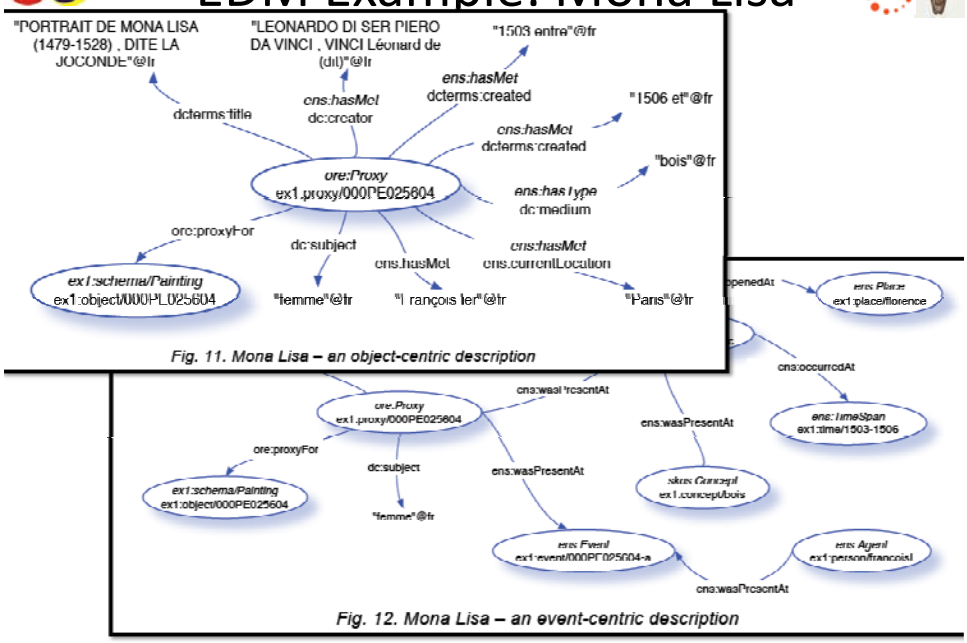


- Reuses 4 other ontologies (shown in white)
- This is the true spirit of semweb

EDM Properties

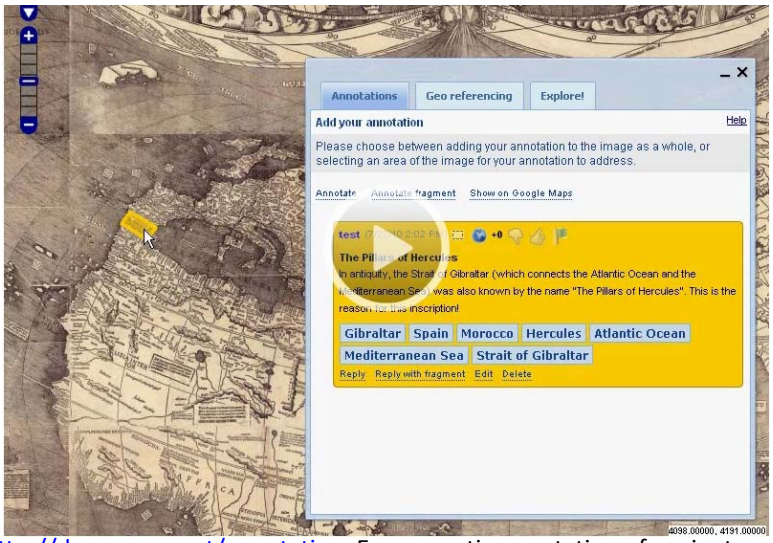


EDM Example: Mona Lisa



EU PROJECTS

EuropeanaConnect Annotation



- http://dme.arcs.ac.at/annotation. Eg: semantic annotation of ancient maps
- Adding an ancient name to a map (**Pillars of Hercules**) finds contemporary geonames (e.g. Gibraltar, Spain, Morocco) and offers them for association

DigiCult Semtech Projects

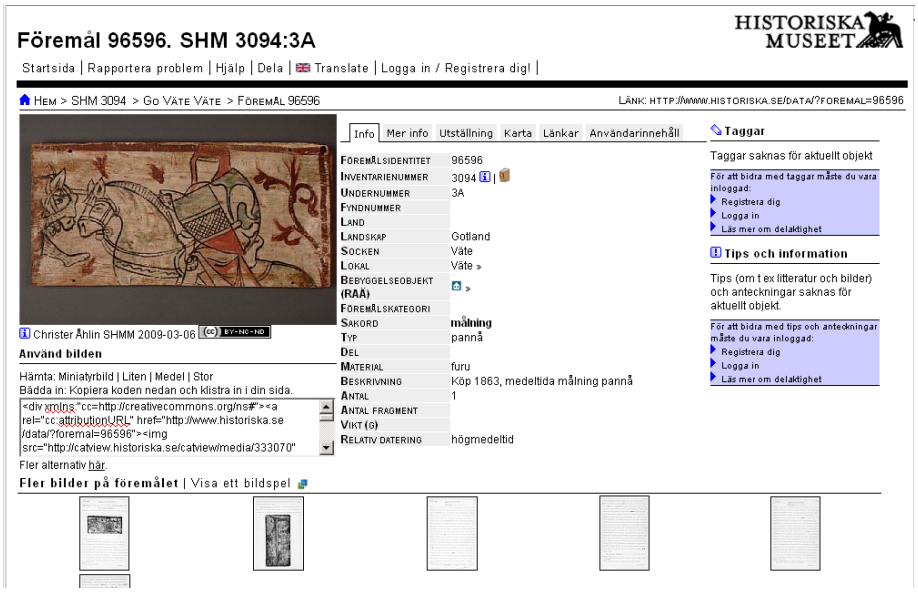
- Digital cultural experiences: exploration of leading edge technologies (e.g. knowledge management systems, semantic tools, graphics, interfaces, novel displays)
  - Empower applications that improve the meaningful use of cultural resources and the experiences users get from visits of museums and sites
  - ➢ 3D-COFORM, ARtSENSE, AXES, CHESS, CINeSPACE, CULTURA, DECIPHER, EPOCH, ISAAC, iTACITUS, MOSAICA, PATHS, P2P-FUSION, QVIZ, V-City, …
- Information retrieval and use: application areas such as digital libraries and archives, exploring methods and tools for (semi)automatic indexing, semantic annotation of non-textual objects (music, speech, images).
  - Aim at enhancing information and knowledge retrieval through efficient search engines able to deliver results from complex multimedia resources, from distributed collections, across languages.
  - ➢ CONTRAPUNCTUS, EASAIER, ethnoArc, IMAGINATION, MEMORIES, MultiMATCH, PAPYRUS, QVIZ, Treble-CLEF, …

EU PROJECTS WITH BG PARTICIPATION

MOLTO

- Multilingual Online Translation (MOLTO), www.molto-project.eu
  - develops knowledge infrastructure, provides interoperability between natural language and structured queries
- BG partner: Ontotext
- Objectives:
  - tool for creating domain-specific translation systems;
  - set of tools for translators and the general public to translate documents;
  - three extensive case studies: mathematical exercises, biomedical patents, museum objects
- Ontotext is in charge the Representation Infrastructure, including interoperability between NL P and SPARQL and case study 3
  - Goal: provide access to museum information, allow querying for museum items, discover information related to them from other sources
  - Data comes from Gothenburg City Museum
  - Uses LOD and Reason-able views to publish museum data
  - Integrates standard museum-specific ontologies (CRM) with upper-level ontologies, to permit linking museum data to existing LOD (e.g. from DBpedia)
  - Integrates c.h.-specific systems such as the Swedish **K-search** system

Example SE Museum Data



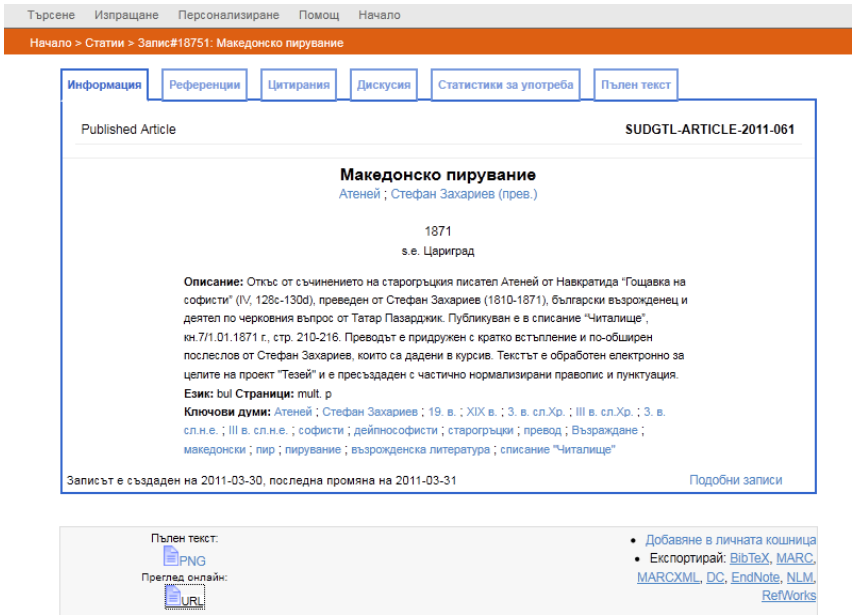
- CULTivating Understanding and Research through Adaptivity (CULTURA), www.cultura-strep.eu
- BG partners: Sofia University, ComMetric (BG branch of US company)
- Development of adaptive systems, new forms of multi-dimensional adaptivity:
  - personalized information retrieval and presentation: user and contextual intent models
  - community-aware adaptivity: responds to community activity, interest, contribution and experience
  - content-aware adaptivity: responds to entities and relationships automatically identified within the artifacts and across collections
  - personalized dynamic storylines: generated across individual artifacts and entire collections
- Advances and integrates:
  - Natural language processing: normalizes ambiguities in noisy historical texts
  - Entity and relationship extraction: highlights the key individuals, events, dates and other entities and relationships within unstructured text
  - Social network analysis of entities and relationships within the content, and also of the individuals and broader community of users engaging with the content
  - Multi-model adaptivity to support dynamic reconciliation of multiple dimensions of personalization

- Improving Access to Text (IMPACT), www.impact-project.eu
- Goal: improve access to historical texts
- Innovate OCR technology
  - explore challenges using different approaches, rather than just one sided approaches
  - develop cutting edge approaches such as collaborative correction
- Innovative language technologies to remove the historical language barrier
- Create center for access to digitized materials of all participating institutions
- Remove constraints to mass digitization by providing Best Practice guidance
- In 2010 the project was extended to include more countries and languages, totaling 25 beneficiaries all over Europe.
- BG participants: Bulgarian Academy of Sciences (ICT Institute), National Library
  - Digitize 2000 pages and create a dictionary of 10k old BG words

- Coordination of Standards and Technologies for the enrichment of Europeana, www.linkedheritage.org
- Consortium includes key stakeholder groups from 20 EU countries, Israel and Russia
  - Several partners participate in related Europeana ecosystem projects; guarantees alignment with Europeana's evolution
  - Organizations which have not in the past been involved will contribute for the first time to Europeana
- BG participant: Bulgarian Academy of Sciences (Central Library)
- Main objectives:
  - Contribute large quantities of new content to Europeana (from both public and private sectors). 3 M new items will be delivered to Europeana, covering a wide spectrum of types of cultural content
  - Demonstrate enhancement of quality of content, in terms of metadata richness, re-use potential and uniqueness
  - Enable improved search, retrieval and use of Europeana content
- Facilitate and deliver large-scale, long-term enhancement of Europeana and its services. Address problems associated with:
  - Non-standard descriptive terminologies
  - Preservation of complex metadata models within the Europeana metadata schema

BG PROJECTS

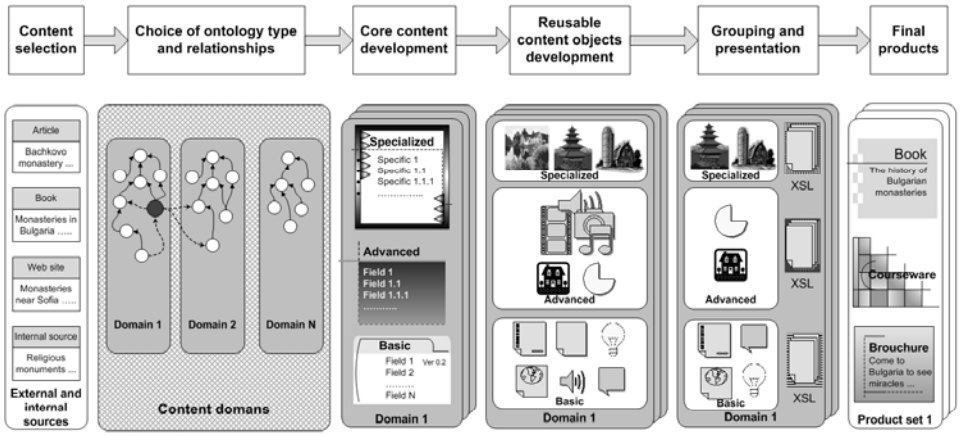
SUDigital Sample Record



Theseus Sample

- Sample text of the SUDigital record shown earlier

SU Collaborative Workflow



- Proposed collaborative content development workflow
- Could be extended to reuse existing artifacts (e.g. Wikipedia BG articles, Panoramio photos, etc)
- Potential to be used by the BG Ministry of Culture for creating an immovable c.h. catalog

Semantic Technologies for Cultural Heritage

Europeana Publishing of Bulgarian Cultural Heritage

P1.1 Dev Semantic web pilot system

Project Review, 21 Nov 2011

Mariana Damova, PhD
mariana.damova@ontotext.com

ontotext

Ontotext Corp

- Who is Ontotext?
  - The leading semtech company in Eastern Europe and one of the leaders world-wide
  - 55 people: Bulgaria (Sofia, Varna), Austria, USA
  - Worked in this area since 2000
  - Venture funding and commercial clients since 2008
  - Bulgaria's most successful participant in EU FP5,6,7 research
  - www.ontotext.com
- 360-degree semantic technology:
  - Semantic Repository (OWLIM), ETL
  - Text Mining, Semantic Annotation and Search (KIM, U.Sheffield GATE, Teamware, MIMIR)
  - Web Mining and Crawling
  - Ontology Engineering and Exploitation
  - Master Data and Linked Data Management

Ontotext Clients (selected)

- British Broadcasting Corporation (BBC)
  - Runs its World Cup 2010 sites on top of OWLIM
  - Next is BBC Sports (2011) and the 2012 Olympics
- The National Archives (UK)
  - The UK Government's official archive contracted Ontotext to implement semantic search for the Government Web Archive
- AstraZeneca
  - Analysis and retrieval of clinical trial reports
  - Integration of biomedical databases for drug target identification
- KT
  - Applications of OWLIM in Interactive TV and IPTV
- All these clients need to:
  - integrate massive amounts of heterogeneous data
  - provide efficient search and navigation

Outline

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

Europeana

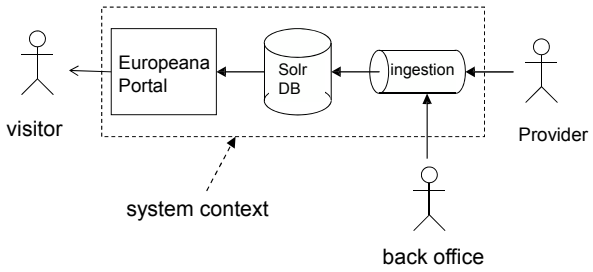
http://www.europeana.eu

- 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

europeana
think culture

Community Building

Google Group: Cultural Heritage Digitalisation
Meetings with
- professionals from the cultural heritage domain
  Archeologists, Folklorists, Epigraphists, Thracologists, Researchers on Medieval Bulgaria, Museum workers
- professionals from the digitalization domain
  IMI
- professionals who work on creating metadata for representing cultural heritage

Conferences
- Digital Preservation and Presentation of Cultural and Scientific Heritage (september 2011)

Identification of digital collections of Bulgarian Cultural Heritage
- F-museum
- Regional Museum of Kjustendil
- WebFolk

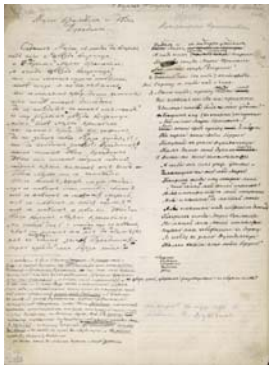
Connection with Europeana Officials and Technology Creators

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

Collections

WebFolk
Веда Словена
an online collection of Bulgarian folk songs gathered in 1997 by a team from the Bulgarian Academy of Sciences
Website funded by
"Open Society" Foundation , available in English and Bulgarian

Originally, a collection made by the Bosnian-Croatian Stefan Verkovitch, 1874
Entitled "Bulgarian Folk Songs from the Prehistoric and Prechristian Times"

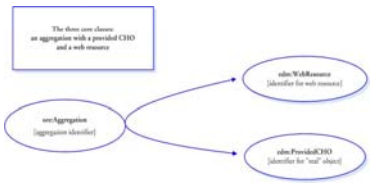
Europeana Data Standards

Europeana Data Standards

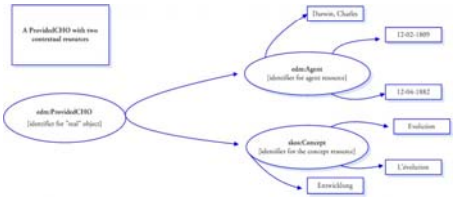
- Unified metadata
  - ESE – Europeana Semantic Elements
    - DublinCore & Europeana fields
    - 36 fields: flat, limited ability semantic links

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

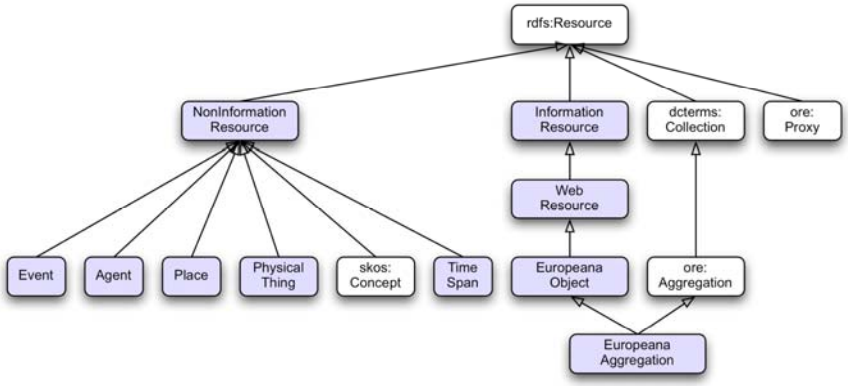


Basic data model



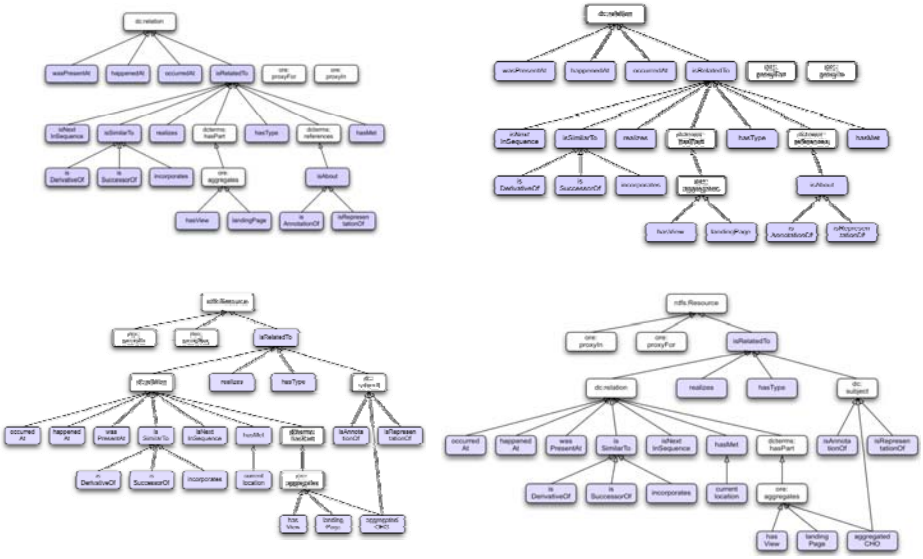
Two contextual classes

EDM Classes

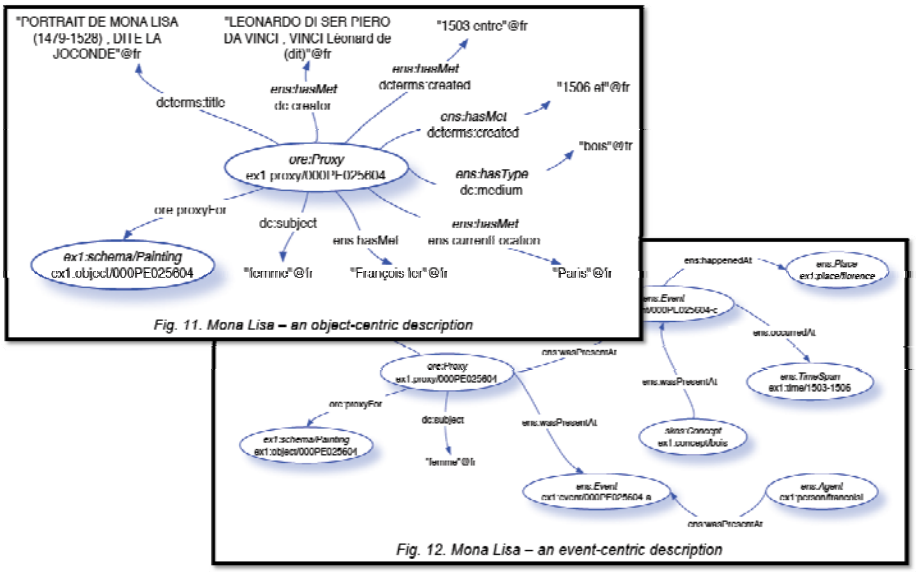


- Reuses 4 other ontologies (shown in white)
- This is the true spirit of semweb

EDM Properties



EDM Example: Mona Lisa



Metadata mapping, conversion and ingestion

Metadata mapping

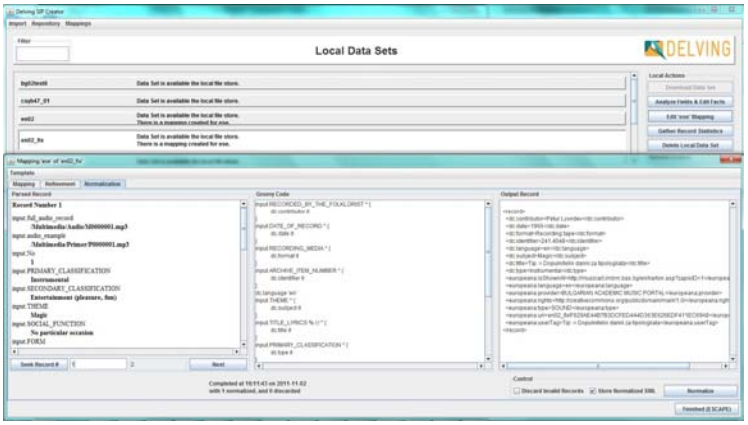
- WebFolk

	WebFolk		Europeana	
	Veda Slovena - EN	Veda Slovena - BG	Source	Element
1	TITLE / LYRICS	заглавие и текст	DC	title
2	THEME	тематика	DC	subject
3	RECORDED BY THE FOLKLORIST	записвач	DC	contributor
4	DATE OF RECORD	година на записа	DC	date
5	PRIMARY CLASSIFICATION	музикална класификация	DC	type*
6	RECORDING MEDIA	начин на записване	DC	format
7	ARCHIVE ITEM - NUMBER	архивна сигнатура	DC	identifier
8	FORM	форма	Bulgariana	FORM
9	METER	размер	Bulgariana	METER
10	TEXTURE	фактура	Bulgariana	TEXTURE
11	WAY OF PERFORMANCE	начин на изпълнение	Bulgariana	WAY OF PERFORMANCE
12	SECONDARY CLASSIFICATION	функционална класификация	Bulgariana	SECONDARY CLASSIFICATION
13	SOCIAL FUNCTION	битова функция	Bulgariana	SOCIAL FUNCTION
14	INFORMANT/S	изпълнява	Bulgariana	INFORMANT/S
15	ETHNIC DATA	етнос	Bulgariana	ETHNIC DATA
16	INFORMANT - PLACE OF BIRTH	родно място на изпълнителя	Bulgariana	INFORMANT - PLACE OF BIRTH
17	COUNTY	околия	Bulgariana	COUNTY
18	PLACE OF RECORD	място на записа	Bulgariana	PLACE OF RECORD
19	TYPE	артефакти/типология	Bulgariana	TYPE
20	ARCHIVE ELABORATION	архивна обработка	Bulgariana	ARCHIVE ELABORATION
21			Europeana	country
22			Europeana	language*
23			Europeana	uri
24			Europeana	userTag
25			Europeana	year

Metadata conversion

- WebFolk
  - crawling the website
  - processing HTML to XML
  - mapping schema

DELVING
tool used to convert 1400 entries into ESE



Metadata ingestion

- administrative steps
- ftp access provided to Europeana
- ingestion tests

The other Collections

Process
- Metadata mapping
- Conversion of all data in ESE format
  - manually
  - with DELVING (original Europeana tool) or MINT (used by FP7 Athena)
- Preparation for Ingestion

Pre-historic and Thracian Civilizations

- digitalization and metadata assignment
- storing in digital repository for cultural heritage
- preparing for Europeana publishing
- publishing in Europeana early 2012

Golden Pages from the Bulgarian Renaissance

- specifying metadata
- storing in digital repository for cultural heritage
- preparing for Europeana publishing
- publishing in Europeana early 2012

Digital Repository for Cultural Heritage

Digital Repository for Cultural Heritage

Sirma Media –
The mission of Sirma Media is to become a leading developer and supplier of interactive educational and cultural heritage environments enabling active acquisition of knowledge through direct experience.

- Selection of technical solution/platform to be used as digital repository for the needs of the project
- Setup and configuration of the digital repository – account management – organizations, users, roles
- Definition of the different types of entities – manuscripts, articles, artifacts, archeological sites, bibliographical records
- Elaboration of the metadata properties in accordance with the content providers requirements
- Migration of databases and digitalized artifacts from available online resources and cultural bodies collections
- Training of users to work with the admin panel of the digital repository – metadata input and editing, media files upload
- Publication of the digitalized collections on the web – UI layer enabling rich visualization, various search options, browse by thematic categories, etc…

Conclusion

Aggregator for Bulgarian Cultural Heritage to Europeana

Thank you for your attention!

Semantic Publishing of Bulgarian Cultural Heritage in Europeana

Start date: 18.10.2011
End date: 21.02.2012

The purpose of this project is threefold:

- to build a technical aggregator to Europeana to publish Bulgarian cultural heritage
- to create an infrastructure which will allow the preservation and the presentation of Bulgarian cultural heritage by enabling organizations to publish their digital collections online and then make them available for viewing via Europeana
- to form a community of experts, officials and activists who work actively for the digitalization, presentation and preservation of Bulgarian cultural heritage, and trigger the establishment of different instruments that help institutionalize, organize and finance these activities in Bulgaria.

Bulgariana (bulgariana.eu) is the name of the created technical aggregator to Europeana. A spaeically designed logo according to Europeana requirements has been approved by Europeana officials.

Two collections have been produced:

- Golden Pages of the Bulgarian Renaissance
- Pra-historic and Thracian Civilizations

The metadata have been converted to Europeana format (Europeana Semantic Elements, ESE), and the administrative procedure for becoming a technical aggregator to Europeana has been successfully completed.

The two collections were harvested by Europeana in February 2012 and appeared online on March 06.2012, where they are accessible since then via http://europeana.eu , using

- Search by provider "Ontotext" at the http://www.europeana.eu/ portal:
- Search by provider using direct link: http://www.europeana.eu/portal/search.html?query=*:*&qf=PROVIDER:Ontotext
- Search by collection:

Search	Collection	Link
europeana_collectionName:2021501*	Golden Pages from the Bulgarian Renaissance	http://europeana.eu/portal/search.html?query=europeana_collectionName:2021501*
europeana_collectionName:2021502*	Pra-historic and Thracian Civilizations	http://europeana.eu/portal/search.html?query=europeana_collectionName:2021502*

- Search by keyword in cyrillic letters

The metadata of the two collections were modelled in Europeana Data Model (EDM), which is a semantic representation and the basis of the next generation of Europeana infrastructure.

The collections have been made available in this format for evaluation by Europeana experts, and is available for viewing at http://britishmuseum.ontotext.com/sparql

A digital repository bulgarianheritage.bulgariana.eu serves as a hub for Bulgariana collections. The user interface is currently only in Bulgarian.

Community building activities have been initiated during the project. A Bulgarian wide conference including representatives of memory institutions, research institutions, business and the government gathered to share successes and discuss issues around the topic of digitalization, preservation and presentation and Bulgarian cultural heritage and Europeana. This meeting: "Europeana: Mission Possible" was a first of a series of meetings, co-organized by Ontotext, hosted by different organizations in Bulgaria. The community is counts a couple of hundred representatives and close to 100 organizations from all stakeholders of this cause. A Google group Cultural Heritage Digitalization was established which counts about 50 people. It serves as an instrument for easy information exchange.

Impact and Sustainability of the Project

- Two meetings about Europeana were organized as a consequence of the first one in Veliko Tyrnovo and in Varna (co-organized by the corresponding Regional Libraries)
- Ontotext is a partner of a ICT-PSP project "Europeana Creative" from February 2013 with Europeana.
- Ontotext hosts the entire collection of Europeana in the new semantic EDM format at http://europeana.ontotext.com since October 2012.
- Ontotext co-organizes with the Bulgarian Ministry of Culture a round table with Europeana officials in November 2012, where the Europeana vision for the local organization of Europeana aggregators will be presented and good practices of local aggregators established in other European countries will be discussed.
- D-r Mariana Damova was awarded with the Award “Unsung Hero” of Europeana during Europeana AGM 2012 in Berlin in the category “Aggregate” for her contribution in building a relationship with Europeana and efforts for establishing a Bulgarian National aggregator.
- A new collection has started Sacred places/Christian Orthodox Temples in February 2013.
- The collection “Pra-historic and Thracian Civilizations” is one of the 5 featured in the iPad application of Europeana for Open Culture in June 2013.
- Publication at DiPP’2012:
  Mariana Damova. **Bulgariana: A Bulgarian Aggregator to Europeana**. In: Proceedings of "Digital Presentation and Preservation of Cultural and Scientific Heritage" (DiPP2012) conference, Veliko Tarnovo, Bulgaria, September 2012.
- Bulgariana is going to become an NGO.