

Aggregating metadata for Europeana: the Greek paradigm

Alexandros Koulouris¹, Vangelis Banos² and Emmanouel Garoufallou³

¹ *Technological Educational Institute of Athens. Department of Library Science and Information Systems, 12210, Egaleo, Greece.*

akoul (at) teiath.gr

² *Aristotle University of Thessaloniki. Department of Informatics, 54124, Thessaloniki, Greece.*

vbanos (at) gmail.com

³ *Technological Educational Institute of Thessaloniki, Department of Library Science and Information Systems, P.O. BOX 141, 57400, Thessaloniki, Greece.*

mgarou (at) libd.teithe.gr

Abstract: *Europeana is an evolving service that tries to be a single access point for Europe's cultural heritage. The metadata aggregation for Europeana is a procedure that needs interoperability and software tools for transforming metadata to Europeana Semantic Elements (ESE). This aggregation imposes requirements difficult to be implemented in many small institutions. To accommodate these needs and help the Greek institutions contribute their content to Europeana, we developed a series of tools. The Hellenic Aggregator that has established a single communication point with Europeana. The Open Archives Engine (OAE) software for creating digital library aggregators. The oaipmh.com tool for validating OAI-PMH enabled repositories. Other wrappers that enable repeatable generation and harvesting of ESE-compatible metadata via OAI-PMH. We present these tools and the way that helped the Greek cultural institutions in contributing content to Europeana.*

Keywords: *Europeana, EuropeanaLocal, OAI-PMH, Metadata harvesting, Europeana Semantic Elements.*

DOI: It would be provided by publication house

I. INTRODUCTION

Repositories and digital libraries are distributed among European Countries. Various formats, different content types, multiple metadata schemes are used. This knowledge either in cultural or in science sector should be accessible to Europeana citizens for awareness and dissemination. From this need various aggregation schemes have derived. *Europeana* (the European Digital Library) is an evolving service that tries to be a single access point for Europe's cultural heritage. According to recent researches, *Europeana* seems to be service of a vital importance for European cultural awareness. This may be certified by recent surveys (IRN Research, 2011). In this context and in order these content to be ingested various tools have been developed.

This paper analyses the Greek paradigm and the toolset that has been developed for data providers.

II. EUROPEANA AND EUROPEANALOCAL: WHERE WE STAND TODAY

Europeana service (Koninklijke Bibliotheek, 2009) is designed to increase access to digital content across Europe's cultural organizations (i.e. libraries, museums, archives and audio/visual archives). In order to achieve these goals *European Union* (EU) launched various projects. One of the most fruitful was *EuropeanaLocal*, which ran from 1 June 2008 to 31 May 2011. It had 32 partners from 27 countries, 1031 plus person months and €4.3 million budget. Up to 21 July 2011, *EuropeanaLocal* partners made available to *Europeana* live service 4.984.952 items, and 160.000 items from Region Marche ingested via *CulturaItalia* (2006), make a total of 5.144.952 items. More than 20 million items, held across 27 countries. Finally, another 1.075.791 items are on the way for harvesting.

In conclusion, *EuropeanaLocal* project, till the 4th of July 2011, contributed 26% of total current *Europeana* content. It had a great impact on *Europeana* strategy and awareness, documentation and guidelines, workflows and on tools and support. Technical and interoperability challenges were solved and the European aggregations infrastructure was enhanced. However, long term systemic problems (e.g. finance, staff availability that is qualified), remained. Finally, more work has to be done in ingesting more local archival content from municipal/regional archives, church councils, local history associations, etc (Rowlatt *et al.*, 2011).

III. THE GREEK CASE: SOLVING TECHNICAL ISSUES AND DEVELOPING SOFTWARE TOOLS

Greece is participating in *EuropeanaLocal* with content providers and the Hellenic Aggregator created and supported by the Veria Central Public Library (VCPL). One of the most important aspects in the process of creating a *Europeana* Compliant digital repository is the support for ESE. Existing digital repository software in general does not support ESE by default. The DSpace plugin for *Europeana* Semantic Elements webpage (Banos, 2010), developed by the VCPL and the Hellenic National Documentation Centre (EKT), provides specific information about the process (Houssos *et al.*, 2011).

DEiXto, Hellenic Aggregator, Open Archives Engine and oaipmh.com are tools created to help organizations participate in aggregation schemes such as Europeana, even if they do not use open digital repository platforms, but closed source technologies or legacy software, which do not support OAI-PMH or any other form of automatic metadata exchange.

IV. THE HELLENIC AGGREGATOR FOR EUROPEANA

Until August 2011, thirteen Greek cultural organizations are already participating in EuropeanaLocal (Koulouris *et al.*, 2010) and are making their metadata available to the Europeana service. These organizations are:

#	Organization	Records
1	Pandektis - National Documentation Center of Greece (http://pandektis.ekt.gr/)	38.880
2	Medusa - Veria Central Public Library (http://medusa.libver.gr/)	1.547
3	The Historical Archives of the American Farm School of Thessaloniki (http://ouranos.afs.edu.gr/dspace)	712
4	Technical Chamber of Greece Regional Department of Corfu (http://lib.teeker.gr/)	141
5	Central Library of NTUA (http://dspace.lib.ntua.gr/)	3.103
6	Music Library - Lilian Voudouri (http://digma.mmb.org.gr/)	2.796
7	Corgialenios Digital Library (http://www.corgialenios.gr/library/)	7.053
8	University of Athens – Pergamos digital library (http://pergamos.lib.uoa.gr/)	74.494
9	Hellenic Ministry of Education – Educational Television (http://www.edutv.gr/)	661
10	Anatolia College - Digital Archives & Special Collections (http://www.anatolia.edu.gr/digitalarchives)	447
11	Technical Chamber of Greece Library (http://library.tee.gr)	5.783
12	Serres Central Public Library (http://ebooks.serrelib.gr)	464
13	Levadia Central Public Library (http://ebooks.liblivadia.gr)	142
	Total	136.223

Table 1. Greek cultural organizations that participate in Europeana.

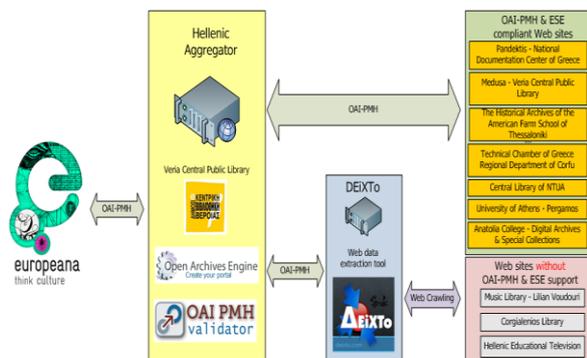


Figure 1. The Hellenic Aggregator Architecture.

Essentially, becoming part of Europeana means that Europeana is able to retrieve specific information from a digital library. One way of doing such a thing would be to connect with each digital library using the appropriate network communication protocol and retrieve the following data:

- Metadata (descriptive, administrative) describing a digital object. The metadata must be mapped to the ESE v3.4 (Europeana v1.0, 2011),
- 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.

Obviously, the amount and type of content, the technical infrastructure, the output formats and the documentation available can vary significantly among all these content providers. It is, therefore, practically impossible for Europeana to work individually with every content provider due to the enormous amount of work that the harmonization and normalization of metadata would require. As a result, Europeana works with an intermediate layer of content providers, the aggregators.

The Hellenic Aggregator (Veria Central Public Library, 2010) functions as an intermediary on a national level, retrieving data from each participating organization and creating a single communication point with Europeana. What is more, the Hellenic Aggregator's role within Europeana is not confined to submitting metadata. Aggregators also play a key role in other fields:

- 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.

A. Registering the Hellenic Aggregator

The process of registering a new digital library to the Hellenic Aggregator is described below:

1. Initially, the digital library web site is examined by an expert who concludes whether it contains content suitable for Europeana.
2. If the digital library supports OAI-PMH, an expert from the EuropeanaLocal Group conducts metadata tests, identifies potential problems and suggests possible solutions.
3. If the digital library does not support OAI-PMH, DEiXto software is used to harvest the required metadata from the target HTML pages.
4. As soon as the digital library's metadata comply with the Europeana standards, it is registered in the Hellenic Aggregator.
5. Content Provider Agreement is signed by the digital library director.

6. The digital library content is published in Europeana.

B. Software Tools

The software infrastructure of the Hellenic Aggregator consists of three different tools which function complementarily in order to implement the full lifecycle of digital library validation, data extraction, storage and communication with Europeana. The core of metadata harvesting, storage and communication with Europeana is implemented by Open Archives Engine while specific data extraction tasks are handled by DEiXTo software. Last but not least, OAI-PMH protocol support and standards compliance for all partners is evaluated using OAIPMH.com.

V. OPEN ARCHIVES ENGINE

Open Archives Engine (OAE) (Banos, 2009) software can be used to create a metadata aggregator and search portal using OAI-PMH enabled, web accessible digital repositories. OAE utilizes the OAI-PMH protocol in order to retrieve metadata from multiple digital libraries and create an index, which then can be used not only to search and filter information but also to export information in a variety of formats such as OAI-PMH Dublin Core (DC) and ESE. Additionally, OAE leverages the technology of DEiXTO in order to extract metadata from legacy digital libraries.

The main components of OAE are the metadata harvester and the web interface.

A. OAE Metadata Harvester

The OAE Metadata Harvester is responsible for connecting with digital libraries and extracting their metadata. After retrieving the metadata from a content provider the one way or the other, the OAE software applies filtering and normalization techniques in order to prevent errors and increase the quality of the metadata.

- XML document encoding and structure is checked using the HTML Tidy (Ragget, 2008) library and a number of errors such as adding missing tags or removing and resolving inappropriate XML characters are resolved.
- Validation against ESE & Dublin Core XML Schemas (DCMI, 2011) is performed.
- Validation for invalid metadata values such as invalid URLs, dates or missing fields is performed.
- Special library-specific bug fixes are applied.

Finally, system indexes are updated and the Hellenic aggregator is ready to publish the aggregated data.

B. OAE Web Service

OAE web interface is responsible not only for creating a web portal from which users are able to search, browse and view metadata records and navigate to the original archives, but also for outputting metadata in a variety of formats such as Dublin Core and ESE, as well as OpenSearch (A9.com, Inc, 2011) and JSON (1999). Cur-

rently, Europeana communicates with the Hellenic Aggregator using the OAI-PMH interface.

VI. DEIXTO

Digital libraries developed in the past did not support up to date technologies such as OAI-PMH or any kind of metadata extraction using web services. As a result, the inclusion of such libraries in the Hellenic Aggregator and the Europeana would be impossible without the use of advanced data extraction tools such as DEiXTo.

DEiXTo (Donas, 2010) is a powerful web data extraction tool that is based on the W3C Document Object Model (DOM) (W3C, 2005). It allows users to create highly accurate "extraction rules" (wrappers) that scribe what pieces of data to scrape from a website. DEiXTo can contend with a wide range of websites with high precision and recall. It provides the user with an arsenal of features aiming at the construction of well-engineered extraction rules. Wrappers built with GUI DEiXTo can be scheduled to run automatically providing automated access to resources of interest and saving users a lot of time, energy and repetitive effort.

DEiXTo extracts data and stores it in various formats. In order to facilitate data extraction for the Hellenic Aggregator, a special output plugin has been developed for DEiXTo in order to generate ESE XML files.

VII. OAIPMH.COM

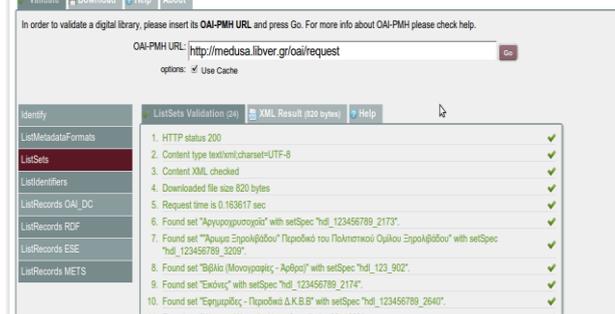
The process of validating an OAI-PMH enabled digital library is quite complex and may become tedious when dealing with a large number of digital libraries. In order to alleviate the task of validating a large number of libraries contributing content to Europeana, a novel online tool has been developed. OAIPMH.com is a web application capable of performing all the necessary checks required to ensure that an OAI-PMH enabled digital library is ready to be part of Europeana.

OAIPMH.com (Banos, 2011) consists of a server-side application running in the background and a modern web interface running on the web. Users can either validate a digital library or download all the records from one or more digital libraries in their computers.

- **Validation:** The validation of an OAI-PMH enabled digital library requires only the submission of the OAI-PMH web service URL. The application issues all OAI-PMH commands to the digital library and evaluates the output according to XML Schemas of DC and ESE as well as the Europeana Guidelines.
- **Metadata extraction:** Users can input a list of OAI-PMH URLs and retrieve all the metadata records which are available from them in parallel. Using this feature, users can retrieve a large number of metadata records from multiple libraries rapidly and easily, thus enabling them to inspect them and evaluate them.

OAIPMH.com has improved the process of validating new and existing OAI-PMH enabled libraries. Any-

one concerned can evaluate digital libraries using a quick and intuitive tool.



VIII. CONCLUSIONS AND FUTURE DEVELOPMENTS

The Greek Cultural Heritage Institutions implemented interoperable digital libraries, digitized valuable historic content, populate, and preserve it through Europeana service. They enriched their metadata quality, familiarized their staff with digital skills and developed new services for their users. On the other hand, the EuropeanaLocal Greek team developed software tools, which facilitated the Geek CHI to implement interoperable digital collections by reducing the cost and the human effort. Automated harvesting procedures were established.

The Geek Cultural Heritage Institutions have to take advantage of the knowledge that achieved during EuropeanaLocal, to implement repositories and to contribute more content to Europeana service. These synergistic schemes, especially in this economic crisis, enhance the viability of the Greek Cultural Heritage Institutions and content. In this context, European Commission should support similar projects, because this is of great importance for the European cultural heritage.

Finally, as the most important future obstacle and because VCPL has the aggregation obligation only through EuropeanaLocal, a national Hellenic aggregator that will continuously support the content providers, share knowledge and expertise between partners should be established.

REFERENCES

- A9.com, Inc., *OpenSearch Protocol*, <http://www.opensearch.org/> (2011).
- Banos, V., *DSpace plugin for Europeana Semantic Elements (ESE)*, <http://vbanos.gr?p=189> (2010).
- Banos, V., *Open Archives Engine Software*, <http://openarchivesengine.com> (2009).
- Banos, V., *Open Archives Initiative Protocol for Metadata Harvesting Validation & Data Extraction Tool*, <http://oaipmh.com> (2011).
- CulturaItalia, *CulturaItalia*, <http://www.culturaitalia.it/Language/LanguageGateway?lang=en&T=1312795936520> (2006).
- DCMI, *XML Schemas to Support the Guidelines for Implementing Dublin Core in XML*, <http://www.dublincore.org/schemas/xmls/> (2011).
- Donas, K., *DEiXTo*, <http://www.deixto.com> (2010).
- Europeana v1.0, *Europeana Semantic Elements specifications*. Version 3.4, 11/08/2011, http://www.version1.europeana.eu/c/document_library/get_file?uuid=77376831-67cf-4cff-a7a2-7718388eec1d&groupId=10128 (2011).
- Europeana, *Europeana Content Checker User Guide*, http://version1.europeana.eu/c/document_library/get_file?uuid=d9ca0106-affb-4a38-83e9-e886289dd0d9&groupId=10602 (2010).
- EuropeanaLocal, *EuropeanaLocal*, <http://www.europeanalocal.eu/> (2008).
- Houssos, N., Stamatis, K., Banos, V., Kapidakis, S., Garoufallou, E. and Koulouris, A., "Implementing enhanced OAI-PMH requirements for Europeana", *Proc. International Conference on Theory and Practice of Digital Libraries (TPDL 2011)*, Berlin, Germany, September 25-29, 2011, and *Lectures Notes in Computer Science (LNCS)* (2011) (in press).
- IRN Research, *Europeana – Online Visitor Survey: Research Report*, version 3, 23rd June 2011, https://version1.europeana.eu/c/document_library/get_file?uuid=334beac7-7fc2-4a4e-ba23-4dcc1450382d&groupId=10602 (2011).
- Javascript Object Notation (JSON)*, <http://www.json.org/> (1999).
- Koninklijke Bibliotheek, *Europeana*, <http://www.europeana.eu> (2009).
- Koulouris, A., Garoufallou, E. and Banos, E., "Automated metadata harvesting among Greek repositories in the framework of EuropeanaLocal: dealing with interoperability", *Proc. 2nd Qualitative and Quantitative Methods in Libraries International Conference (QQML2010)*, Chania, Greece (2010) (in press).
- Ragget, D., *HTML Tidy Library Project*, <http://tidy.sourceforge.net> (2008).
- Rowlatt, M., Davies, R. and Komen, L., *EuropeanaLocal: it's objectives, activities and impact. Project presentation: results D1.11*, <http://www.europeanalocal.eu/eng/Document-Library/Project-Deliverables> (2011).
- Veria Central Public Library, *Europeana Local Aggregator*, <http://aggregator.libver.gr> (2010).
- W3C, *Document Object Model (DOM)*, <http://www.w3.org/DOM/> (2005).