

## **Solving aggregation problems of Greek cultural and educational repositories in the framework of Europeana**

### **Emmanouel Garoufallou**

Lecturer,

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

E-mail: [mgarou@libd.teithe.gr](mailto:mgarou@libd.teithe.gr)

### **Vangelis Banos**

IT Manager,

Hellenic Institute of Metrology, Intersection: NB6 and ΔA10 Industrial Area of Thessaloniki, Block 45, GR 57022 Sindos, Thessaloniki, Greece,

E-mail: [vbanos@gmail.com](mailto:vbanos@gmail.com)

### **Alexandros Koulouris**

Lecturer,

Department of Library Science and Information Systems, Technological Educational Institute of Athens, Agiou Spyridonos Str., 12210 Aigaleo, Greece,

E-mail: [akoul@teiath.gr](mailto:akoul@teiath.gr)

### **Biographical notes:**

**Emmanouel Garoufallou** is a Lecturer at the Department of Library Science and Information Systems at the Alexander Technological Educational Institute of Thessaloniki, Greece. He is member of the Europeana Network and the Director of Deltos Research Group. He is involved in various EC and National Projects such as the AccessIT Plus Project, the Light Programme, the Entitle Project, the EuropeanaLocal Project. Prior to that, he worked as Lecture and as Information Officer at the Department of Information and Communications at MMU, UK. He holds a PhD on the Impact of Digital Libraries on People from the Department of Information and Communications at Manchester Metropolitan University (MMU) and a MA in Information and Library Management from Northumbria University at Newcastle upon-Tyne, UK.

**Vangelis Banos** is an IT Manager in the Hellenic Institute of Metrology. He is also the Project manager of the BlogForever FP7 Project. In the past, he has participated in European programs such as EuropeanaLocal & AccessIT in which he collaborated with the Veria Central Public Library. He is a PhD candidate at the Computer Science Department of the Aristotle University of Thessaloniki. His research interests include digital preservation, repositories, online archives, open access, data mining & information retrieval. In the past, he has worked at the University of Macedonia Library. He holds a Bsc in Information & Communication Systems Engineering from

the University of the Aegean & an MSc in Information Systems from the Aristotle University of Thessaloniki.

**Alexandros Koulouris** is Lecturer in the Department of Library Science and Information Systems at the Technological Educational Institute (TEI) of Athens. He is member of the Europeana Network (formerly Council of Content Providers and Aggregators – CCPA) and of the Laboratory on Digital Libraries and Electronic Publishing at Ionian University and collaborator of the Veria Central Public Library. In the past, he has worked as a librarian for the National Technical University of Athens and for the National Documentation Centre of Greece. He holds a PhD and a BSc in Library and Information Science (LIS) from Ionian University and the TEI of Athens respectively, and a BSc in International and European Studies from Panteion University. More can be found on his website <http://users.teiath.gr/akoul/>.

**Corresponding author**

Emmanouel Garoufallou can be contacted at: [mgarou@libd.teithe.gr](mailto:mgarou@libd.teithe.gr)

**Abstract:** The paper describes a set of software tools and methodologies followed by the Hellenic Aggregator and Greek cultural, educational and scientific institutions in order to contribute their content to Europeana. The software tools are presented and their impact on the publishing, aggregation and dissemination process of Greek content to Europeana is described. The work of the Hellenic Aggregator, the methodologies developed and the new software tools, which were created improved the publishing process of Greek institutions and promoted their content to Europeana. During this process, difficult technical challenges were overcome and a clear path has been created for Greek content providers who would like to publish their content to Europeana. The present study outlines the work on supporting Europeana Semantic Elements (ESE) metadata publication. New metadata models such as the Europeana Data Model (EDM) are under development and future work should be committed to supporting them as well. The work presented in this publication is unique in covering Greek digital content providers, which publish their content to Europeana. Without the software and methodologies presented, harvesting Greek digital content from Europeana would not be possible.

**Keywords:** Europeana, OAI-PMH, Metadata harvesting, EDM, Institutional repositories, Digital Libraries, Aggregation, Learning resources, Educational resources, Cultural resources

## 1. Introduction

Repositories and digital libraries are widely distributed among European Countries. A broad range of standards including various formats, different content types and multiple metadata schemes are used. This knowledge, either in the cultural or in the scientific sectors, should be accessible to European citizens for awareness and dissemination as well as digital libraries and their digital working environment should be considered as a platform for sharing and disseminating knowledge (Garoufallou and Asderi, 2010; Garoufallou, Asderi and Koutsomiha, 2010)

. In order to cope with this need, various aggregation schemes have emerged. Europeana (the Digital Library, Archive and Museum of Europe) is an evolving service that tries to be a single access point for Europe's cultural heritage. According to recent surveys (IRN Research, 2011), Europeana is a service of vital importance for European cultural awareness. In this context, a set of special software tools and processes was developed in order to contribute Greek content to Europeana.

This article analyses the Greek paradigm, the procedures and the toolset developed to support Europeana data providers. This includes the Hellenic Aggregator (HA), the Open Archives Engine (OAE) software, DEiXTo web content extraction tool, and oai-pmh.com online OAI-PMH protocol validator and data extractor service.

The structure of the rest of this text is as follows. Section two (literature review), describes the way that the countries have solved aggregation issues through Europeana and what is the state-of-the-art in aggregation for Europeana. Section three presents the current status of Europeana content aggregation and heterogeneity and analyzes the example of EuropeanaLocal for two reasons. The reasons are:

a) this project contributed almost 30% of the available content in Europeana (Rowlatt, Davies, and Komen, 2011), and

b) many Greek cultural organizations build interoperable repositories and were helped by the network and the tools that implemented in the Europeana Local framework.

Section four analyzes the heterogeneity and the aggregation problems of the Greek cultural content. It describes an actual example of the heterogeneity of the content in Europeana. Section five offers practical information to Greek cultural organizations that are looking to connect to Europeana. Section six outlines the state of the Greek digital content and proceeds with a detailed presentation of the HA, showing some data related to the process that was deployed in the case of the HA. For example, man-hours that people worked on the metadata to publish the content, costs associated with technical work and theoretical/metadata related work. In addition, it offers practical steps for registering a repository to the HA (section 6.1) and including participating organizations, development and procedures. It also explains supports why the design decisions taken by the software engineers should be followed in similar cases. Section seven ~~is dedicated to~~ presenting the software tools developed in the context of this work, and how they were adopted by the Greek cultural organizations, ~~giving a practical example of the American Farm School repository (section eight)~~. Section eight refers to the American Farm School (AFS) repository implementation, gives a practical example of the costs associated with technical work and theoretical/metadata related work. Finally, section nine presents some conclusions and future work.

## 2. Literature review

This section contains a description of the state-of-the-art in aggregation for Europeana and explains references on how various countries, including Greece, have solved aggregation issues in the framework of EuropeanaLocal project and Europeana. The metadata development, mapping and aggregation are not an easy procedures. In addition building interoperable repositories is very crucial in the information integration era. Cultural organizations in Greece, in other European and non-European Countries have made efforts, either themselves (Olensky, Stiller, and Droge, 2012) or through the framework of Europeana and its supportive programs to tackle these issues. For example, EuropeanaLocal created a best practices network of interoperable repositories and helped small cultural institutions to provide content in Europeana (this issue is analyzed on the next section). Europeana also offered a great chance for cultural organizations to build repositories and to facilitate complex problem solving of metadata mapping and aggregation.

The Europeana Foundation centrally launched and implemented tools for helping aggregators and providers. In 2011, the Europeana Office launched the Europeana Professional website (2011d), where librarians, curators and archivists share digital expertise. The Europeana cultural organizations have implemented different types of aggregators. For example, country and projects aggregators or independent organizations that ingest cross-domain cultural content in Europeana (The Europeana

Office, 2011a). A technical working group was formed in the framework of Europeana Foundation. This working group provides useful information and tools for the [Europeana Semantic Elements \(ESE\)](#) and [Europeana Data Model \(EDM\)](#) implementation and other metadata mapping and aggregations issues, and helps the participants and prospective cultural organizations that need to participate in Europeana (The Europeana Office, 2011e).

Finally, successful case studies of EDM (The Europeana Office, 2011c) and ESE implementation has been made in many countries, except Greece. For example, in Poland the Poznan Supercomputing and Networking Center (PSNC, 2009a) Digital Libraries Team (PSNC 2009b), which acts as the hub of the Polish Digital Libraries Federation and aggregates metadata from Poland's regional and institutional repositories, has a successful ESE case story. The team describes the steps towards integrating the Polish content into the Europeana portal and the way they mapped and implemented ESE and how they helped by the Content Checker of Europeana (The Europeana Office, 2011b).

### 3. Europeana content aggregation and heterogeneity

The 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). Europeana is an evolving service, which will constitute an umbrella of European metadata from distributed cultural organizations. Europeana currently gives access to more than 23 million items representing all Member States including film materials, photos, paintings, sounds, maps, manuscripts, books, newspapers and archival papers. This process brings together and links up heterogeneously sourced content, which is complementary in terms of themes, location and time. In February 2013, Europeana's active partner network consists of 2,200 organizations from 33 countries.

In order to achieve these goals European Union (EU) launched various projects. One of the most fruitful was EuropeanaLocal (2008), which ran from June, 1<sup>st</sup> 2008 to 31<sup>st</sup> May 2011. This project was designed to involve and support local and regional libraries, museums, archives and audio-visual archives to:

- a) make the enormous amount of content that they hold available through Europeana, and
- b) deliver new services.

The project was funded under the eContentPlus Programme of the European Commission, ~~and concluded~~ [It resulted](#) in a Best Practice Network of distributed and interoperable repositories. EuropeanaLocal had 32 partners from 27 countries, 1031 plus person months and €4.3 million budget. ~~By Up-to-date~~ (February 2013), EuropeanaLocal partners had made available to [the](#) Europeana live service almost [six million](#) ~~6.000.000~~ items. Over 800 organizations that provided content mobilized across 27 countries, enabled and motivated local institutions and their staff to participate in Europeana by enhancing skills and expertise of key staff involved in the project.

EuropeanaLocal also had a great impact on Europeana strategy and awareness, documentation and guidelines, workflows and on tools and support. For example, EuropeanaLocal promoted aggregation, provided information systems and standards in use, helped in improving the Europeana Semantic Elements (ESE) scheme, evolved the Europeana Data Agreements, first tested and provided feedback in tools like the

ESE XML Schema validations. Additionally, EuropeanaLocal partners benefited by learning how to install OAI-PMH repositories, better understanding the importance of metadata and its impact on search results, networking themselves, tuning harvesting procedures, ~~etc.~~ (Rowlatt, Davies, and Komen, 2011). ~~Nonetheless,~~ ~~p~~Part of the content that provided to Europeana via EuropeanaLocal project came from the Greek cultural organizations that built interoperable repositories because of participating in this state of the art network.

In conclusion, the EuropeanaLocal project contributed 26% of total current Europeana content (February 2013). Technical and interoperability challenges were overcome, the network has made tremendous progress in content aggregation and the European aggregators' infrastructure was enhanced. However, long-term systemic problems such as financial problems and availability of qualified staff remained (Rowlatt, Davies, and Komen, 2011). It is worth noting that currently more content is delivered to Europeana service not only from completed projects like EuropeanaLocal but also from ongoing projects and initiatives.

#### 4. Greek cultural digital content: heterogeneity and aggregation problems

In the ~~last~~ decade, from 2000 until 2010, Greek cultural organizations (libraries, archives, museums) and academic institutions (e.g. Universities, Research Centers) digitized various types of content in the framework of projects funded by the European Union. The content was heterogeneous, because of the organizations' diversity. For example, libraries and archives mostly digitized texts and images, museums digitized objects such as paintings from exhibitions and statues, sound archives digitized audiovisual materials such as video and television recordings. This mosaic of digitized items presents the heterogeneity of the Greek content in Europeana.

Apart from the content diversity, the cultural organizations in Greece have different level of expertise and skills concerning the implementation of repositories and digital libraries. Most of them faced various problems such as a) limited number of digitized materials, b) absence of metadata schema, c) lack of OAI-PMH compatible repositories, d) lack of qualified staff, and e) lack of quality metadata descriptions. These organizations through EuropeanaLocal had an opportunity with the support from the network to build interoperable repositories, using open source software DSpace, and ~~to~~ tackle all these ~~above~~ issues by following metadata standards like DC and ESE as well as aggregation standards such as OAI-PMH.

The diversity of content, skills, repositories, and infrastructure implies practical aggregation problems. Academic libraries and the Hellenic National Documentation Centre (NDC, EKT in Greek) aggregate their repositories content through the openarchives.gr service –the Greek digital libraries search engine- (National Documentation Centre, 2006). The engine was developed ~~one of the authors of this paper (by V. Banos) (paper co-author)~~ as a free-lancer service and is now maintained by NDC. The openarchives.gr has 430,443 records from 64 repositories using mainly simple DC.

~~On the contrary,~~ Greek cultural organization did not make any aggregation progress until the EuropeanaLocal project (2008-2011). The usefulness of the software tools described below, such the HA ~~that~~ implemented in 2010 ~~by~~from the Veria Central Public Library (2010), the Open Archives Engine (Banos, 2009), the oai.pmh validator, ~~etc.~~ helped the content providers to build interoperable repositories and

allowed them to provide their content in Europeana. This help was both in technical issues and in tackling metadata compatibility problems. For example, most of the repositories, including those that participated in the openarchives.gr search engine, they have implemented simple DC. The ESE schema (The Europeana Office, 2012) needs more elements/fields like the type of content, divided into text, image and video, and other specific data elements/fields. The content providers were helped technically by the Greek EuropeanaLocal team in batch importing of the ESE metadata fields and values. For example, the DSpace plugin for ESE (Banos, 2010), developed by the Veria Central Public Library (VCPL) and the NDC (Houssos *et al.*, 2011), was a useful tool for batch importing.

Finally, if we take into account that according to the openarchives.gr the digital content in Greece that is provided mostly by the academic and research sector is 430,443 records, the 13603,223482 that is provided through the HA to Europeana is a significant amount of content. About The one-fifth (1/5) of the Greek digital content and almost all the cultural heritage Greek digital content is harvested by the HA.

## 5. Helping the Greek cultural organizations: practical steps for participating in Europeana

This section provides the information and describes the practical steps that EuropeanaLocal provided to Greek cultural organization in order to assist them in providing quality content to Europeana. Greece participated in EuropeanaLocal with the Veria Central Public Library (VCPL) that served both as partner and content provider and by providing the Hellenic Aggregator (HA). Since August 2011, fourteen content providers, of which seven use DSpace software, have followed closely the Europeana standards and implemented and fully supported the ESE; moreover their content harvested successfully by both the HA (Veria Central Public Library, 2010) and Europeana (Koulouris, Garoufallou, and Banos, 2010). By August 2011, the HA had provided 136,223 items to Europeana (see also section six).

The most important aspects in the process of creating a Europeana compliant digital repository can be summarized as follows:

1. First of all, support for ESE, which is virtually a new Dublin Core Profile, developed by Europeana is paramount in order to fulfill its operational requirements. Existing digital repository software in general does not support ESE by default, as it is the case with Dublin Core (DC). Nevertheless, the nature of the formats makes it feasible to alter existing software and data in order to add support for ESE. Specific information about the process can be found at the DSpace plugin for Europeana Semantic Elements webpage (Banos, 2010), developed by the VCPL and the NDC (Houssos *et al.*, 2011).
2. After becoming familiar with ESE, the next step in the process is to use the Europeana XML Namespace (The Europeana Office, 2010) in order to augment existing systems' configuration adding support to the additional ESE elements.
3. Finally, the repository must have to be populated with the appropriate metadata values. This task can be either performed either manually through the appropriate user interface of each digital library or automatically by using special software tools developed for this purpose. Given the widespread usage of the DSpace software internationally and including Greece, the focus has been the implementation of tools for this specific platform.

In addition to modern digital repository platforms, there are also numerous digital libraries built with older or closed source technologies or legacy software, which do not support OAI-PMH or any other form of automatic metadata exchange. In these cases, special techniques should be applied in order to extract metadata through plain HTTP requests and web content extraction using a [special](#) tool developed for this purpose called DEiXTo (or ΔEiXTo) (Ntonas and Kokkoras, 2007). DEiXto, Hellenic Aggregator, Open Archives Engine and oaipmh.com are tools created to help organizations participate in aggregation schemes such as Europeana and assisting in solving interoperability issues among distributed repositories. The following sections present these tools.

## 6. The Hellenic aggregator for Europeana

By June 2012, fourteen Greek cultural organizations were participating in EuropeanaLocal [thereby](#) ~~and were~~ making their metadata available to the Europeana service. The situation was different in previous years, for example in 2010 ((Koulouris, Garoufallou, and Banos, 2010) and in 2011 (Koulouris, Banos and Garoufallou, 2011) there were less cultural organizations providing less content. Table 1 presents the Greek cultural organizations that participate in Europeana as well as the number of records they are sharing (data comes from February 2013).

**Table 1.** Greek cultural organizations that participate in Europeana

### Figure 1. The Hellenic Aggregator architecture

The HA architecture is outlined in Figure 1. The main component of the aggregator is also the communication point with Europeana. Additionally, the HA ~~is~~ [employing](#) two different mechanisms to extract metadata from digital libraries depending on their level of compliance with the standards. If a digital library is supporting OAI-PMH and ESE, the communication with the aggregator is performed by using the OAI-PMH protocol. On the other hand, if the digital library is using legacy technologies, a special proxy is necessary. Deixto Web data extraction tool is used as an intermediate to crawl the digital library, extract metadata, generate ESE XML files and communicate with the HA using OAI-PMH to share metadata.

Essentially, becoming part of Europeana means that Europeana is able to retrieve specific information from a digital library. One way of [achieving this](#) ~~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 v3.4.1 (The Europeana Office, 2012),
- 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. Some examples to highlight this diversity are listed below:

- The digital repository of the Aristotle University of Thessaloniki (2010) is based on [the Invenio platform](#) (CERN, 2012) and ~~is using~~ MARC21 as its underlying metadata format.
- The digital library and institutional repository at the University of Macedonia maintained by the Library and Information Centre (2006) is based on an old version of the DSpace platform (MIT Libraries and Hewlett Packard, 2002) and ~~is using~~ Dublin Core as its underlying metadata format.
- The digital repository of the Veria Central Public Library is based on the newest version of DSpace and is capable of exporting metadata in RDF and METS (Library of Congress, 2012) besides Dublin Core.
- The Music Library “Lilian Voudouri” is based on a proprietary ASP platform and cannot provide any kind of XML encoded metadata.

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 HA (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 HA’s role within Europeana is not restricted 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.
- Generating discussions amongst partners and providing valuable feedback to Europeana.
- Promoting and implementing standards further along the content provision chain.
- Providing domain specific expertise and skills to institutions and Europeana.

An outline of the architecture of the HA is presented in Figure 1. The system is implemented in order to ~~be able to~~ cope with digital libraries supporting OAI-PMH as well as older systems without interoperability services. A detailed presentation of the software used in the HA is presented in section seven.

The HA was developed by the members of the EuropeanaLocal Group in Greece over a period of six months. Development involved initial discussions between the whole group ~~concerning for~~ the specific standards and technologies to be used. ~~The software development was performed by a single developer over a period of three who was fully committed for three person months, to the project.~~ The outcomes were continuously evaluated by the group. No further expenses were necessary as the group was using existing an server from Veria Central Public Library’s infrastructure and open source software tools.

### *6.1 Steps for registering a digital library to the Hellenic aggregator*

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

1. Initial contact ~~is has to be~~ made between the digital library administrator and a member of the EuropeanaLocal Group in Greece. The digital library web site is examined by an expert who ~~decides eoncludes~~ whether it contains content suitable for Europeana.
2. If the digital library supports OAI-PMH, an expert from the EuropeanaLocal Group conducts metadata tests of the digital library using the Europeana Content Checker (Europeana, 2010) and oaipmh.com validator. ~~Ifn-ease~~ there are problems and deviations from the protocols, the digital library's administrator is advised to make the necessary changes and the process is repeated.
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 complies with Europeana standards, it is registered to the HA and periodic harvesting is initiated.
5. The HA performs validations on the registered digital libraries at regular intervals and proposes solutions for ~~potential~~ problems. For example, if there is a problem with the OAI-PMH web service of a content provider, the administrator is notified and the library is removed temporarily from the main index until the problem is resolved.
6. Content Provider Agreement is signed by the digital library director.
7. The digital library content is published in Europeana.

During the process of registering the 14 participant institutions to the HA, a variety of issues occurred:

- The most common problem was the sub-optimal configuration of the OAI-PMH service of most digital libraries. Most of the times, OAI-PMH support was configured with the default settings and was not tuned to the exact needs of each digital library's ~~collections~~. Fortunately, the solution to this issue was easy as the EuropeanaLocal Group helped each institution to configure their systems correctly.
- Another important issue was the lack of Europeana Semantic Elements support in all libraries. This problem was resolved using the appropriate ESE plugin and configuration for each type of digital library software.
- Finally, the legacy digital libraries participating in the HA posed a serious challenge to data extraction. This was due to two reasons: a) the legacy html markup used, and b) the slow performance characteristics of the old systems. Deixto Web data extraction software operators had to focus on the intricacies of each library in order to be able to extract all the necessary information and generate ESE compliant metadata.

The data are related to the process that was deployed in the case of the Greek Aggregator. Man-hours that people worked on the metadata to publish the content, costs associated with technical work and theoretical/metadata related work, ~~ete~~ are presented.

## 6.2 Software tools

The software infrastructure of the HA 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 [the](#) Open Archives Engine while specific data extraction tasks are handled by DEiXTo software. Finally yet importantly, OAI-PMH protocol support and standards compliance for all partners is evaluated using oaipmh.com.

## 7. How the providers may use the software tools

This section presents the tools that the HA developed through the EuropeanaLocal project ~~in order~~ to meet the demands of both diversity of content and the technical infrastructure of the repositories and digital libraries of Greek cultural organizations. Some of the tools like DEiXTo have been developed by other researchers and adapted by the HA in order to target more effectively some cultural organizations and maximize its outcome.

### 7.1 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 in turn can be used not only to search and filter information but also to export information in a variety of formats such as OAI-PMH 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.

#### 7.1.1 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 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 library (Ragget, 2008) and a number of errors such as adding missing tags or removing and resolving inappropriate XML characters are resolved.
- Validation against ESE and DC XML Schemas (DCMI, 1995) 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 HA is ready to publish the aggregated data, either through the Web Service or through the Web Interface component.

#### 7.1.2 OAE web service

OAE web interface is responsible for creating a web portal from which users are able to search, browse and view metadata records and navigate to the original archives. It can also help for outputting metadata in a variety of formats such as DC and ESE, as well as OpenSearch (2011) and [Javascript Object Notation JSON](#) (1999). Currently, Europeana communicates with the HA using the OAI-PMH interface.

OAE web interface enables normal web users to browse and interact with the aggregated contents. Nevertheless, due to the nature of the Europeana project these features are currently disabled and only the OAI-PMH web service is available.

## 7.2 DEiXTO

The broad and diverse sets of digital libraries willing to contribute their content to Europeana have posed great technical obstacles, especially on the level of metadata harvesting. Digital libraries developed in the past did not support technologies such as OAI-PMH or any kind of metadata extraction using web services. As a result, the inclusion of such libraries in the HA and the Europeana would be impossible without the use of advanced data extraction tools such as DEiXTO.

DEiXTO (Ntonas and Kokkoras, 2007) 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 describe what pieces of data to scrape from a website. DEiXTO consists of three separate components:

- **DEiXTO GUI:** an MS Windows™ application implementing a friendly graphical user interface that is used to manage extraction rules (build, test, fine-tune, save and modify).
- **DEiXTOBot:** a Perl module implementing a flexible and efficient sleepy Mechanize agent (essentially a browser emulator) capable of extracting data of interest using GUI DEiXTO generated patterns. It contains best of breed Perl technology and allows extensive customization. Thus, it facilitates tailor-made solutions.
- **Command Line Executor:** a stand-alone, DEiXTOBot-based, cross-platform utility that can massively apply an extraction rule on multiple target HTML pages and produce structured output in a wide variety of formats.

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 such as XML, Excel, CSV, RSS, Text, OpenDocument Spreadsheet and HTML. In order to facilitate data extraction for the HA, a special output plugin has been developed for DEiXTO in order to generate ESE XML files.

## 7.3 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. A digital library willing to contribute its content to Europeana must be capable of communicating with the OAI-PMH. Additionally, the metadata that becomes available through OAI-PMH must be compliant with specific schemas, the *Dublin Core Metadata Initiative* (DCMI) and the ESE. Thus, extra ~~sanity~~ checks must be performed ~~to~~ the metadata provided in order to ensure their correctness. 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 client. Users can utilize the web site of oaipmh.com in order to issue commands to the application, which performs all the necessary actions in the background. The results appear on the user interface in real time. 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. After retrieving the URL, the server-side application issues all OAI-PMH commands to the digital library and evaluates the output according to the standards. After this process is completed, the user is presented with a checklist of validation checks, which have been performed, and their results. The validation checks that are performed, are based on the XML Schemas of DC and ESE as well as the Europeana Guidelines for digital libraries.
- **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 users to inspect and evaluate the metadata records.

~~These conclude,~~ Oaipmh.com has improved the process of validating new and existing OAI-PMH enabled libraries; anyone concerned can evaluate digital libraries using a quick and intuitive tool.

Μορφοποιήθηκε: Αγγλικά (H.B.)

Figure 2. Oaipmh.com validator

### 8. Successful Greek IR implementation in the framework of Europeana: the American Farm School case

The American Farm School Archives and Historical Collection (AFSAHC) came under the administration of the Dimitris and Aliko Perrotis Library in 2001. The AFSAHC contain precious documents relating to the foundation of the American Farm School (AFS) and its history such as correspondence, reports, photographs and artifacts.

In December 2007, the AFS ~~was been~~ invited to participate in the EuropeanaLocal project ~~that was being~~ led in Greece by a project team from the VCPL. The inclusion of collections of the AFS Archives into the Europeana service, provided further incentive and mobilized the transformation of core photography collections into digital format.

In order ~~for this~~ to ~~be~~ accomplished ~~this task and~~ along with the processing of existing unsorted records, simultaneous selection for preservation and digitization of 1000 core photographs from three major AFS collections ~~s~~ for the EuropeanaLocal was carried out. Scanning and selection procedures commenced in January 2008 and ended in spring 2009. ~~These~~ collections are *Miscellaneous AFS photos*, *Card Postal collection* ~~and well as~~ documents and scanned artifacts from the *Girls School*.

DSpace software was selected as a platform in order to build the digital archive and to organize AFS's collections. Configuration, parametrisation and installation were outsourced. Setting up servers and a staff-training seminar took place in July 2009. Metadata description for 1000 digital objects using the DC schema started in July 2009 and ended in October 2010 (American Farm School, 2010).

### 8.1 Budget considerations

The cost of salary for outsourcing configuration, parametrisation, installation of DSpace as well as training and maintenance costs was 2,000€. The chosen equipment and servers that selected were HP Proliant ML115 (server) with extra RAM at a net-cost of 1,500€.

### 8.2 Outcomes

Digitizing a significant part of AFS collections made accessible information that was previously only available to a selected group of researchers who were able to undertake on-site research. Now users can search the collections rapidly and comprehensively from anywhere at any time and they can find what they are looking for quickly and independently, thus increasing use of collections.

Research questions increased and almost half of the total reference questions provided are aimed at AFSAHC. The user base ranges from university students who seek information for their dissertations, to publishers who request permission to use photos from AFSAHC in their publications, to alumni who request photos of the times they were students or who want to donate part of the pictures archives and AFS staff.

Digital provision of archives facilitates learning and scholarship. A collaboration with the Department of American Literature and Culture, of the English Literature School of Aristotle University of Thessaloniki commenced in order to initiate an internship program under the title: *Greek-American Cultural Exchanges: Archival Research at the AFS*. The internship program aims at fourth-year or MA students with an interest in exploring America's global role and studying the perceptions and impact of US in Greece. Students learn to perform research in an original archive setting. They learn how to search in the AFS digital collections environment and they contribute an article to the AFS newsletter and the American Studies Resource Portal (ASRP). Thus, they organize literary group activities (literary evenings, film-screenings, etc.) in collaboration with the ASRP team and the AFS (email from D. Koutsomiha, AFS, 15 February 2013).

The AFS case study, shows practically how cultural organization in Greece helped by the EuropeanaLocal and the tools that it has developed may implement interoperable repositories. This is the main reason of presenting extensive numbers and data given for EuropeanaLocal project. Details of costs in monetary and manpower terms. Man-hours that people of worked on the metadata to publish the content, costs and associated with technical work and some comment on the impact of theoretical/metadata related work as well as the impact that of this digital archive has on the community is presented in this section to highlight the impact of EuropeanaLocal in small cultural institution.

## 9. Conclusions and prospective

Greek Cultural Heritage Institutions implemented and continue to implement rich repositories, digitizing and preserving valuable historic and cultural content. Many of these cultural repositories were built through the EuropeanaLocal project, which gave the opportunity to small organizations, like the AFS, to implement interoperable repositories, to promote and preserve their content. The Europeana Network Greek team, which was formed through the EuropeanaLocal project, is dedicated to support GCHIs effort to publish and enrich their content, disseminate and preserving it through Europeana. In order to support their cause, the Europeana Network Greek

team has developed a set of software tools as well as best practices to implement interoperable digital collections, thus reducing the cost and the human effort. The AFS is a successful example of good practice by the Hellenic aggregator concerning developing a digital repository, digitizing content, applying metadata, train staff and bring value to the organization.

The GCHIs have to exploit the technical and cognitive capital created by EuropeanaLocal to implement digital repositories and to contribute more content to Europeana. These synergetic schemes, especially in today's economic crisis, enhance the viability of the GCHIs and their content.

In the context of economic crisis, the EU funds gave to the cultural organizations the opportunity to build infrastructures and to digitize content. Greek academic libraries are still enhancing their digital content through EC structural funds. This will increase the Greek digital (cultural) content, which may be available through Europeana.

While [the](#) Greek economy [is](#) shrinking rapidly, cultural industries and tourism as well as the multidisciplinary cooperation provides a great chance for Greek culture and economy to use [Europeana and the EU funds](#) -it as a development aid. Beyond the economic crisis, the information-integrated environment demands synergies and cooperation between the multiple disciplines of information providers (archives, libraries, museums, private sector companies, etc.). For example the Stavros Niarchos Foundation Cultural Centre (SNF, 2011) is funding the new buildings of the National Library and Opera House of Greece, supports nonprofit organizations such as the Future Library, public libraries across Greece as well as funding research in the cultural sector. It is widely believed that this could be a good example to follow in order to launch a new era in information integration and enhancement of information services in Greece. In this context, an official national aggregator service has to emerge from Hellenic Aggregator, utilizing the existing knowledge, software tools and infrastructure built through the EuropeanaLocal project in order to support all Greek institutions in this difficult financial crisis.

**Acknowledgements:** Our thanks goes to EuropeanaLocal, the Greek team of EDLocal and VCPL, the Greek content providers and their staff and especially to Damiana Koutsomiha for the information ~~that~~ provided.

## References

- American Farm School (2010) *DSpace Repository at American Farm School of Thessaloniki*. [online] <http://ouranos.afs.edu.gr/dspace> (accessed 19 February 2013).
- Aristotle University of Thessaloniki (2010) *Aristotle University of Thessaloniki – Psifiothiki*. [online] <http://invenio.lib.auth.gr/> (accessed 13 February 2013).
- Banos, V. (2009) *Open Archives Engine software*. [online] <http://openarchivesengine.com> (accessed 13 February 2013).
- Banos, V. (2010) *DSpace plugin for Europeana Semantic Elements (ESE)*. [online] <http://vbanos.gr?p=189> (accessed 23 August 2011).
- Banos, V. (2011) *Open archives initiative protocol for metadata harvesting validation and data extraction tool*. <http://oaipmh.com> (accessed 13 February 2013).
- CERN (2012) *Invenio*. [online] <http://invenio-software.org/> (accessed 13 February 2013).
- CulturaItalia*. [online] <http://www.culturaitalia.it/Language/LanguageGateway?lang=en&T=1312795936520> (accessed 13 February 2013).

DCMI (1995) *XML schemas to support the guidelines for implementing Dublin core in XML*. [online] <http://www.dublincore.org/schemas/xmls/> (accessed 13 February 2013).

Europeana (2012) *Europeana content checker user guide*. [online] <http://pro.europeana.eu/documents/900548/ae5e78e8-ce78-424d-b360-5c01eddb3564> (accessed 13 February 2013).

EuropeanaLocal (2008) *EuropeanaLocal*. [online] <http://www.europeanalocal.eu/> (accessed 13 February 2013).

Garoufallou, Emmanouel and Asderi, Stella (2010), 'Digital Libraries and the digital working environment: what is their effect on library staff for sharing their knowledge?', in A. Katsirikou and C. Skiadas (Eds) *New Trends in Qualitative and Quantitative Methods in Libraries, 2<sup>nd</sup> Qualitative and Quantitative Methods in Libraries. Proceedings of the International Conference on QQML 2010*, Chania, Greece, 2010, World Scientific, pp. 359-365.

Garoufallou Emmanouel, Asderi Stella and Damiana Koutsomiha (2010), 'Digital Libraries as Knowledge Management Systems', in *International Scientific Conference, eRA 5: The SynEnergy Forum: The Conference for International Synergy in Energy, Environment, Tourism and contribution of Information Technology in Science, Economy, Society and Education*, T.E.I. of Piraeus, Greece, 15-18 September 2010.

Houssos, N., Stamatis, K., Banos, V., Kapidakis, S., Garoufallou, E. and Koulouris, A. (2011), 'Implementing enhanced OAI-PMH requirements for Europeana' in *TPDL 2011: Proceedings of the International Conference on Theory and Practice of Digital Libraries, Lectures Notes in Computer Science (LNCS)*, Berlin, Germany, pp. 296-407.

IRN Research (2011) *Europeana – online visitor survey: research report*. [online]. [http://pro.europeana.eu/c/document\\_library/get\\_file?uuid=334beac7-7fc2-4a4e-ba23-4dcc1450382d&groupId=10602](http://pro.europeana.eu/c/document_library/get_file?uuid=334beac7-7fc2-4a4e-ba23-4dcc1450382d&groupId=10602) (accessed 13 February 2013).

*Javascript Object Notation*. [online] <http://www.json.org/> (accessed 13 February 2013).

Koninklijke Bibliotheek (2009) *Europeana*. [online] <http://www.europeana.eu> (accessed 13 February 2013).

Koulouris, A., Banos, V., & Garoufallou, E. (2011), 'Aggregating metadata for Europeana: the Greek paradigm' in G. Giannakopoulos, & D. Sakas (Eds.), *Proceedings of the International Conference on Integrated Information (IC-ININFO 2011)*, Kos Island, Greece, September 29-October 3, 2011, pp. 198-201. Piraeus: IDAS.

Koulouris, A., Garoufallou, E. and Banos, E. (2010), 'Automated metadata harvesting among Greek repositories in the framework of EuropeanaLocal: dealing with interoperability', in A. Katsirikou and C. Skiadas (Eds) *New Trends in Qualitative and Quantitative Methods in Libraries, 2<sup>nd</sup> Qualitative and Quantitative Methods in Libraries. Proceedings of the International Conference on QQML 2010*, Chania, Greece, 2010, World Scientific, pp. 331-337.

Library of Congress (2012) Metadata Encoding and Transmission Standard (METS). [online] <http://www.loc.gov/standards/mets/> (accessed 13 February 2013).

MIT Libraries and Hewlett Packard (2002) *DSpace*. [online] <http://www.dspace.org/> (accessed 13 February 2013).

National Documentation Centre (2006) *Greek digital libraries search engine*. [online] <http://openarchives.gr/> (accessed 13 February 2013).

Ntonas, K. and Kokkoras, F. (2007) *DEiXTo*. [online] <http://www.deixto.com> (accessed 13 February 2013).

Olensky, M., Stiller, J. and Droge, E. (2012) 'Poisonous India or the Importance of a Semantic and Multilingual Enrichment Strategy', in J.M. Dodero, M. Palomo-Duarte, & P. Karampiperis (Eds.), *Metadata and Semantics Research (MTR 2012), Communications in Computer and Information Science (CCIS)*, vol. 343, pp. 252–263. Berlin: Springer. [online] doi: [http://dx.doi.org/10.1007/978-3-642-35233-1\\_25](http://dx.doi.org/10.1007/978-3-642-35233-1_25) (accessed 22 February 2013).

*OpenSearch Protocol*. [online] <http://www.opensearch.org/> (accessed 13 February 2013).

PSNC (2009a) *Poznan Supercomputing and Networking Centre*. [online] <http://www.man.poznan.pl/online/en/> (accessed 22 February 2013).

PSNC (2009b) *PSNC Digital Libraries Team*. [online] <http://dl.psnc.pl/> (accessed 22 February 2013).

Ragget, D. (2008) *HTML tidy library project*. <http://tidy.sourceforge.net> (accessed 23 August 2011).

Rowlatt, M., Davies, R. and Komen, L (2011) *EuropeanaLocal: it's objectives, activities and impact. Project presentation: results D1.11*. [online] <http://www.europeanalocal.eu/eng/Document-Library/Project-Deliverables> (accessed 13 February 2013).

SNF (2011) *Stavros Niarchos Foundation Cultural Centre*. [online] [http://www.snf.org/snfcc/snfMain\\_en.html](http://www.snf.org/snfcc/snfMain_en.html) (accessed 22 February 2013).

The Europeana Office (2010) *Europeana: Europeana Semantic Elements (ESE)*. [online] <http://www.europeana.eu/schemas/ese/> (accessed 13 February 2013).

The Europeana Office (2011a) *Aggregators and providers*. [online] <http://pro.europeana.eu/web/guest/aggregators-and-providers> (accessed 22 February 2013).

The Europeana Office (2011b) *ESE case studies*. [online] <http://pro.europeana.eu/web/guest/case-study> (accessed 22 February 2013).

The Europeana Office (2011c) *Europeana Data Model (EDM) case studies*. [online] <http://pro.europeana.eu/web/guest/case-studies-edm> (accessed 22 February 2013).

The Europeana Office (2011d) *Europeana Professional*. [online] <http://pro.europeana.eu/> (accessed 22 February 2013).

The Europeana Office (2011e) *Technical requirements*. [online] <http://pro.europeana.eu/> (accessed 22 February 2013).

The Europeana Office (2012) *Europeana semantic elements specifications v3.4.1*. [online] <http://pro.europeana.eu/documents/900548/dc80802e-6efb-4127-a98e-c27c95396d57> (accessed 13 February 2013).

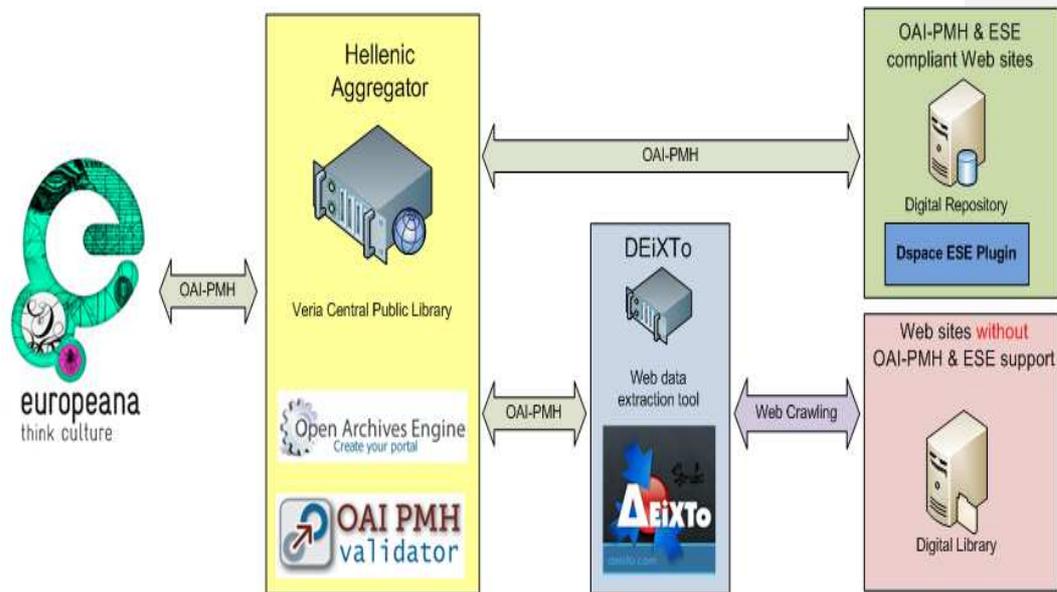
University of Macedonia, Library and Information Centre (2006) *Psepheda: digital library and institutional repository*. [online] <http://dspace.lib.uom.gr/> (accessed 13 February 2013).

Veria Central Public Library (2010) *Europeana Local Aggregator*. [online] <http://aggregator.libver.gr> (accessed 13 February 2013).

W3C (2005) *Document Object Model (DOM)*. [online] <http://www.w3.org/DOM/> (accessed 13 February 2013).

#	Organization	Records
1	Pandektis - National Documentation Center of Greece ( <a href="http://pandektis.ekt.gr/">http://pandektis.ekt.gr/</a> )	22.485
2	Medusa - Veria Central Public Library ( <a href="http://medusa.libver.gr/">http://medusa.libver.gr/</a> )	2.188
3	The Historical Archives of the American Farm School of Thessaloniki ( <a href="http://ouranos.afs.edu.gr/dspace">http://ouranos.afs.edu.gr/dspace</a> )	1.007
4	Technical Chamber of Greece Regional Department of Corfu ( <a href="http://lib.teeker.gr/">http://lib.teeker.gr/</a> )	216
5	Central Library of NTUA ( <a href="http://dspace.lib.ntua.gr/">http://dspace.lib.ntua.gr/</a> )	3.769
6	Music Library - Lilian Voudouri ( <a href="http://digma.mmb.org.gr/">http://digma.mmb.org.gr/</a> )	7.053
7	Corgialenios Digital Library ( <a href="http://www.corgialenios.gr/library/">http://www.corgialenios.gr/library/</a> )	7.381
8	University of Athens – Pergamos digital library ( <a href="http://pergamos.lib.uoa.gr/">http://pergamos.lib.uoa.gr/</a> )	24.663
9	Hellenic Ministry of Education – Educational Television ( <a href="http://www.edutv.gr/">http://www.edutv.gr/</a> )	661
10	Anatolia College - Digital Archives & Special Collections ( <a href="http://www.anatolia.edu.gr/digitalarchives">http://www.anatolia.edu.gr/digitalarchives</a> )	447
11	Technical Chamber of Greece Library ( <a href="http://library.tee.gr">http://library.tee.gr</a> )	5.783
12	Serres Central Public Library ( <a href="http://ebooks.serrelib.gr">http://ebooks.serrelib.gr</a> )	464
13	Levadia Central Public Library ( <a href="http://ebooks.liblivadia.gr">http://ebooks.liblivadia.gr</a> )	142
14	Athos Memory ( <a href="http://www.athosmemory.com/">http://www.athosmemory.com/</a> )	27.223
<b>Total</b>		<b>103.482</b>

**Table 1.** Greek cultural organizations that participate in Europeana



**Figure 1.** The Hellenic Aggregator architecture



**Figure 2.** Oaipmh.com validator