Preservation Aspects of a Curation-Oriented Thematic Aggregator

Dimitris Gavrilis
Digital Curation Unit
Athena Research Center
Artemidos 6 & Epidavrou
Maroussi, Greece
+30 2106875425, GR
d.gavrilis@dcu.gr

Stavros Angelis
Digital Curation Unit
Athena Research Center
Artemidos 6 & Epidavrou
Maroussi, Greece
+30 2106875425, GR
s.angelis@dcu.gr

Christos Papatheodorou
Dept. of Archives and Library Science,
Ionian University, Corfu, Greece and
Digital Curation Unit
Athena Research Center
Artemidos 6 & Epidavrou
Maroussi, Greece
+30 2106875425, GR
papatheodor@ionio.gr

Costis Dallas
Digital Curation Unit
Athena Research Center
Artemidos 6 & Epidavrou
Maroussi, Greece
+30 2106875425, GR
c.dallas@dcu.gr

Panos Constantopoulos
Digital Curation Unit
Athena Research Center
Artemidos 6 & Epidavrou
Maroussi, Greece
+30 2106875425, GR
p.constantopoulos@dcu.gr

ABSTRACT
The emergence of the European Digital Library (Europeana) foregrounds the need for aggregating content using smarter and more efficient ways taking into account its context and production circumstances. This paper presents the main functionalities of MoRe, a curation oriented aggregator that addresses digital preservation issues. MoRe combines aggregation, digital curation and preservation capabilities in a package that shields content providers from changes, and that ensures efficient, high volume metadata processing. It aggregates data from a wide community of archaeological content providers and integrates them to a common metadata schema. The system provides added-value digital curation services for metadata quality monitoring and enrichment so that to ensure metadata reliability. Furthermore it provides preservation workflows which guarantee effective record keeping of all transactions and the current status of the repository.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *system issues*

D.2.12 [**Software Engineering**]: Interoperability

General Terms
Documentation, Performance, Design, Reliability, Standardization.

Keywords
Digital curation, aggregators, Europeana, CARARE, workflow, metadata enrichment, digital preservation, micro services.

1. INTRODUCTION
The emergence of the European Digital Library (Europeana) presents the need for aggregating content from multiple content providers and delivering this content to Europeana in a single metadata schema and in a uniform way. The CARARE project (Connecting Archaeology and Architecture in Europeana – http://www.carare.eu/) has delivered successfully over 2 million records (about 10% of Europeana's total content) from over 22 different content providers. The cultural assets made available are very diverse, from prehistoric and Iron Age archaeological survey results to complex Mediterranean archaeological sites and historic buildings. The digital resources representing such assets are also heterogeneous, ranging from paintings and prints to photographs, archaeological and architectural plans, sections and drawings, and, increasingly, digital 3D models.

The challenge faced by CARARE was that each content provider had their information in their native schema, using different ways to describe heterogeneous objects. Heritage assets are associated with geographic information, both in the form of geographic coordinates according to some grid standard, and in the form of named geographic entities such as historical place and area names; as expected, contentproviders used different coordinates systems, place names etc. Moreover archaeological sites are characterized by a nested mereological structure, being composed of buildings, each of which is also composed of particular architectural elements. Thus the main requirement for the descriptive metadata of such resources is to represent architectural and archaeological assets at quite different levels of complexity.

CARARE was the first of Europeana's projects to employ operationally the recently defined Europeana Data Model (EDM) [3]. EDM is a semantic graph schema that allows for a rich representation of a digital record. However it is still under development, so CARARE was facing the challenge of accommodating continuous changes in its delivery metadata schema. Additionally, even when EDM reaches a stable status, a part of the partners' metadata information content might be lost in the process of mapping to EDM, which is a generic schema and not especially suited to capture archaeological monument

documentation. An important challenge was, therefore, how to guarantee the preservation of the integrity of original archaeological information supplied by content providers.

This paper presents Monument Repository (MoRe), a repository system which addresses these issues by operating as an information broker between content providers and Europeana, offering value added curation services.

2. BACKGROUND

The traditional approach to aggregating metadata and links to digital resources into Europeana involves an aggregator [2] [11], which implements a crosswalk to transform original metadata records to records following a common output schema such as Europeana Semantic Elements (ESE) [4] or EDM. The crosswalks are based on a set of rules that map a source schema to the target schema (ESE or EDM).

CARARE represents a significant departure from this architecture. It introduces the notion of an information broker – an intermediate repository acting as a mediator – intended to ensure the integrity, authenticity and content enrichment of metadata provided to Europeana by heterogeneous collections. The overall architecture is shown in Figure 1. The content supplied by providers comprises administrative/scientific national registries of sites and monuments, archaeological museum collections, collections of 3D models describing any of these types of objects, as well as digital historical document collections such as the Visual Fortune of Pompeii archive. The metadata of all of these sources are transformed to a rich, thematic (in our case: sites and monuments) schema, the CARARE schema [9], and are stored in the CARARE repository, implemented using the Monument Repository system (MoRe) The CARARE schema is an application profile drawing on MIDAS Heritage, LIDO and the CIDOC CRM. The CARARE metadata are aggregated and delivered into the common format now used by Europeana to describe its content, the Europeana Data Model (EDM).

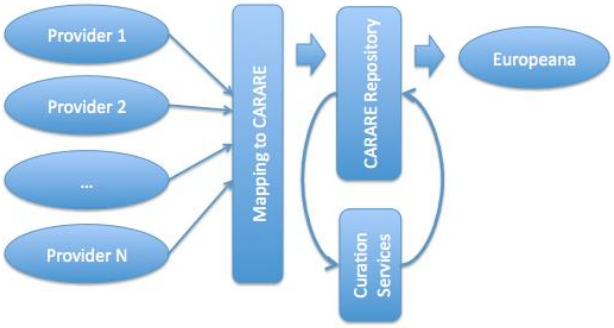

Figure 1. Overall architecture of CARARE

Compared with Europeana's approach for content aggregation, the repository approach followed by the CARARE project has the advantage of having the entire content available thus allowing to perform tasks at repository level, collection level, and content provider level, according to need. The architecture is based on a trusted repository approach [7] [8] and it aims to provide "reliable, long-term access to managed digital resources to its designated community, now and in the future" [12]. Hence MoRe demonstrates an organizational system that curates the archaeological information in accordance with commonly accepted standards and conventions.

This architecture matured alongside with the progress of the CARARE project, as one goal was for the repository to be flexible enough to tackle possible challenges that may appear. This allowed for the introduction of added value features, such as new services along the way of metadata harvesting. For instance, the usage of edm:Place element was introduced at a time when CARARE was already delivering content to Europeana. This element presented the need for visualization of information objects over a map. MoRe had to incorporate this element and provide the appropriate information to Europeana without necessitating a change in the original metadata or extra effort by content providers.

3. MONUMENT REPOSITORY (MoRe)

The mission of MoRe is to support the effective management of supplied information with minimal content providers' involvement. To this end it provides:

- versioning support for subsequent ingests of the same digital objects
- preservation services
- curation services

MoRe was built on top of Mopseus [5] [6], a Fedora-commons based digital repository developed by the Digital Curation Unit - Athena Research Centre.

3.1 Repository architecture

The repository architecture (Figure 2) consists of a core layer of services that receive information packages, pre-process them and store the metadata (datastreams) in a Fedora-Commons installation. The indexes of those datastreams are stored in a MySQL database. The metadata supplied by the content providers is transformed to the CARARE schema, stored, preserved, curated, and then made ready for publication.

MoRe functionalities are based on the implementation of micro-services, i.e., "small and well-defined procedures/functions that perform a certain task" [1] [13]. Chains of micro-services implement a larger macro-level functionality, which are called actions. Micro-services offer modularity in the construction of the MoRe workflows as a feature, and in tandem provide system administrators with full control of what happens in a workflow.

The core services of the repository, along with a set of curation services, have been developed in Java and run on an Apache Tomcat server. All services are orchestrated by a workflow engine and mainly operate at datastream level, although there are services that use only the MySQL index database. For instance, the clustering of records according to geographical proximity needs only to access the relevant indexes.

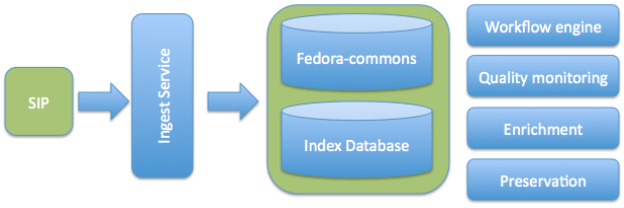

Figure 2. Monument Repository (MoRe) architecture

MoRe is fully OAIS compliant and handles three distinct types of information packages – Submission, Archival and Dissemination (SIP, AIP, DIP) - following certain specifications. Submission packages are created on ingestion and include the native (content provider's) metadata, the XSLT document that transforms the source metadata to the CARARE schema, as well as the corresponding CARARE metadata. All this information is accompanied by a technical metadata XML file (Figure 3) and ensures the tracing of provenance.

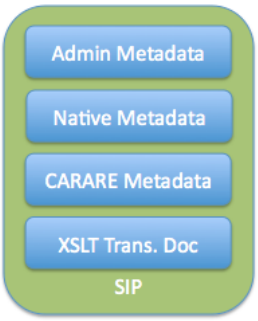

Figure 3. Submission Information Package

Each information object becoming part of the repository is wrapped into an Archival Information Package (AIP) which includes the SIP datastreams, as well as a PREMIS [10] datastream that contains a log with information about the ingestion of the object and relevant events (such as the datastream generation events, curation events, etc.), so that the object can be assigned preservation metadata. Finally, the Dissemination Package (DIP) includes both the CARARE and EDM datastreams. The EDM metadata (datastream) are harvested by Europeana, while the CARARE metadata are provided through the MoRe interface. The EDM datastream is created by the mapping service after the SIP package is ingested.

The services offered by MoRe are distinguished into core and curation services: the first are necessary to carry out the functions of content aggregation and delivery, while the second are intended for improving quality and adding value.

3.2 Core Services

Core services is the minimum set of services required in order to receive, transform and deliver content from the content providers to Europeana. These include:

3.2.1 Ingest

The ingest service is responsible for receiving submission information packages, performing various integrity checks and ingesting them into the repository. As SIP packages are received by the respective web service, they are stored in a temporary space awaiting to be verified. The verification process includes integrity checks on the SIP package in order to make sure that:

a) it contains the necessary information (e.g. package level admin and technical metadata);

b) all items in the package contain all the necessary XML datastreams (e.g. native metadata, CARARE metadata, admin metadata, XSLT transformation);

c) all XML documents are well formed;

d) each item contains valid item and provider identifiers.

After the verification step, the ingest service ingests the datastreams into Fedora-commons following the process below:

a) If the item (based on its provider identifier / native identifier) is new, a new digital object is created;

b) if the item exists, the existing identifier is retrieved;

c) all datastreams are ingested along with the corresponding PREMIS events;

d) the index service is triggered.

3.2.2 Indexing

Indexing is a fundamental service in all repositories. Mopseus comes with its own indexing mechanism which uses a descriptive XML document to define not only which parts of the metadata will be indexed, but also the structure of the SQL database that they will be indexed to. This approach simplifies and in part automates the work of other services such as the quality monitoring service. The repository manager is able to create the indexes and map them to any SQL schema. This approach allows to easily plug in services as they usually require specific table structures in the SQL database. For example, a service that discovers records that are in close proximity to each other, needs acces to a table that contains record identifiers and lat/lon coordinates.

3.2.3 Mapping

The CARARE Schema has been designed so as to capture the complexity of information represented within the CARARE aggregator, namely: collections, heritage objects, digital resources and activities. Thus, Archival Information Packages in the CARARE aggregator consist, in practice, of heterogeneous information, which needs to be re-expressed through mapping in order to allow harvesting and use by Europeana. All content ingested in MoRe is described in the CARARE Schema. Extracting these information objects to Europeana requires a mapping between CARARE Schema and EDM. This transformation is implemented through use of XSLT stylesheets. Depending on the native records, the transformation takes place on ingestion, or at a second step, if a particular set of data needs to be firstly de-duplicated (see Section 3.3.1). The mapping to EDM has been revised many times throughout the CARARE project, as the EDM Schema is still under development. Each time a new element was introduced or altered in EDM, the mapping had to be updated and the transformed objects reproduced and republished to Europeana. All this happens without the need of any effort on the part of content providers.

3.2.4 Delivery

The delivery service is responsible for delivering content through the OAI provider subsystem. The content to be delivered can be grouped per provider, per collection, per package (received package), and of course it has to take versioning into account (always the latest version is sent).

3.2.5 Repository Manager

The repository manager holds a key role in MoRe and in the CARARE project in general. The repository manager is in charge of executing second level checks on the data, making decisions about their overall quality and coordinating the proper operational scheme of the repository.

3.2.6 Quality monitoring

Quality monitoring is an essential part of an aggregator, as it informs content owners about the status of their information. It is based on policies, practices and performance that can be audited and measured in several ways, summarized per collection or even per submission package, as it is not feasible to inspect each information item separately. Some of the quality criteria are:

- Metadata completeness
- Unity of reference to information objects
- Element – Attribute completion
- Accuracy of spatial information

For example, metadata completeness measures whether the information captured per CARARE record meets the project's minimum acceptable standards. Although this task seems trivial at first glance, in a schema like CARARE it becomes somewhat more complicated. Consider the following examples:

a) Information completeness may vary among the top level elements of a CARARE object, and one could get a record with rich information in the heritageAsset and digitalResources elements, but minimal in the Activity elements. In this case, the overall quality is higher than estimated because the Activity element can be discarded during the mapping (to EDM) process.

b) Information can be captured in various ways. For example a spatial object may contain x/y coordinates without specifying the coordinate reference system, and these coordinates are not represented using WGS84. In this case, the actual quality is very low.

3.3 Curation Services

Curation services is a set of services running on MoRe, monitoring information objects and performing actions (curation actions) that aim to provide higher quality content. These are categorized in the table below and have various effects on the resulting metadata records.

Table 1. Curation actions

Action	Effect
Element & attribute cleaning	Homogeneity
De-duplication	Unity of reference, identification
Element & attribute fill	Improved completeness
Relation add	Additional information
Spatial transform	Homogeneity

For example, setting the language attributes (e.g. el, gre, GR to el) provides homogeneity to the resulting records and allows for building better services for end users. Spatial information is often encoded using different coordinate reference systems and has to be transformed to enable unified processing (for instance to the WGS84 system). In other instances, further information needs to be added to records, i.e.: a) a relation that denotes rights usage, b) a language attribute, or, c) the format type of a record.

The workflow used to execute curation services is especially important. Services that perform cleaning and simple element filling (with little built-in logic) are executed first. Following those are the more complex services which have more intelligent logic built into them, such as adding relations, performing de-duplication of records, etc. This sequence helps increase the information available to the more complex micro-services, thus yielding better results.

Below we discuss in some more detail two specific curation services, namely, De-Duplication and Geo-Spatial.

3.3.1 The De-Duplication (De-Dup) curation service

Each CARARE record may consist of four top-level elements: heritageAssets, digitalResources, collectionInformation, activities. The de-duplication service is responsible for:

- identifying duplicate top level elements among CARARE records of the same content provider / collection;
- removing the duplicates and replacing them with relations.

An illustration of how this service works is given in figure 4, where a set of 3 CARARE records are received (top row), and are transformed by the De-Duplication service (bottom row). Top level elements, such as the digitalResource of CARARE Record 2 and CARARE Record 3 (with id: D-1), are replaced by a relation (with id: H1) that points to the same element in CARARE Record 1. This process ensures unity of reference and identification, resulting in more robust sets of records.

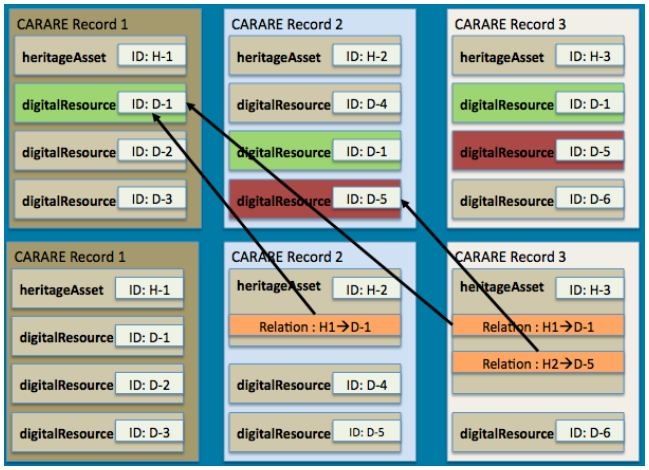

Figure 4. De-duplication example

3.3.2 The Geo-Spatial curation service

CARARE records descibe monuments. As a result, most of them contain spatial information in various forms: latitude/longitude coordinates (including coordinate system); historic place names address, and; country.

All the above information is encapsulated in a Spatial Element Block in CARARE schema [9], and usually not all information is provided. The Geo-Spatial curation service mainly performs three operations:

a) It checks coordinates (if provided) by verifying the correct coordinate reference system (Europeana only accepts WGS84), addressing errors in the provided x/y coordinates.

b) It checks the provided address, place name and other textual information and compiles them into one string (used in the target prefLabel element of the EDM set).

c) It geo-parses place names (if provided) using various openly available geo-parsers, and returns the place names they were mapped to the user. This feature is only provided to the end user through the UI.

An illustration of how the Geo-Spatial service works is given in Figure 5, where the spatial block from a CARARE record is displayed in the box on the left. This block of information contains a compilation of real use cases related to the geo-spatial information that had to be handled. Firstly, due to a mapping error, the x/y coordinates were concatenated in the x element. These are split (the parsing algorithm can detect and handle several cases). After the x/y coordinates are extracted, the coordinate reference system is checked. If it is different from WGS84, the coordinates are converted. If it is not provided, the x/y are checked to verify if they fall into the proper range. In the third step, the x/y are mapped to lat/lon, or vice versa (they must fall within the respective country). Finally, in the fourth step, the lat/lon are placed on a map and the country is checked out.

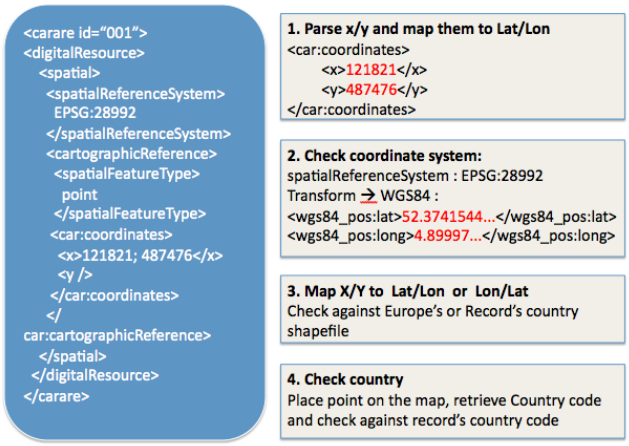

Figure 5. Geo-Spatial service example

3.3.3 Service Orchestration

When running a set of services in a streamline mode, the execution workflow is important especially with regard to preservation aspects. For example, in the execution workflow example presented in 3.4:

- In order to ensure integrity, the transform service must be executed after the Geo-Spatial service (because Geo-Spatial operates only on CARARE streams)
- In order to reduce processing resources required, the De-Dup service must precede all other services (because it results in fewer top level element sets requiring little or no processing).

Proper orchestration of these services helps reduce the amount of resources required, ensures integrity and helps formalize the overall ingest process.

3.4 Preservation Service

The preservation service is responsible for maintaining the metadata of the records provided to the repository, enabling their revision, versioning and validation, as well as maintaining the bond among various forms, and thus preserving provenance information. Each curation action that generates new content, or in any way modifies existing information, produces a new datastream version that is stored in Fedora-commons along with its PREMIS event log. A PREMIS [10] event log is maintained across the entire collection [5].

The Submission Information Package specification requires that each CARARE item is accompanied by its native record, the XSLT document that was used to transform between them, and the administrative metadata associated with the record. All these data are ingested as separate datastreams under the same item in MoRe, along with the appropriate PREMIS event (which is generated during ingest).

From a preservation point of view, the services layer of MoRe handles all preservation tasks. For example, consider a simple ingest of the 3 records shown in Figure 4, and assume that the De-Dup, Geo-Spatial and Mapping services are executed:

- **Ingest**. Each CARARE record is ingested
  - The Native datastream is added
  - The CARARE datastream is added
  - The XSLT (native→carare) is added
  - A PREMIS event record is generated and is added to the object
- **De-Dup**. For each record the De-Dup service processes
  - If the CARARE datastream is updated, a new datastream is added
  - A PREMIS event record is generated and is added to the object
- **Geo-Spatial**. For each record the Geo-Spatial service processes
  - If the CARARE datastream contains geo-information that needs to be updated, a new datastream is added
  - A PREMIS event record is generated and is added to the object
- **Mapping**. For each record the Mapping service processes
  - The CARARE record is transformed into EDM and the EDM datastream is added
  - A PREMIS event record is generated and is added to the object

This approach allows to track all the changes to the objects and to roll-back these changes if needed. Furthermore, the PREMIS records contain references to the services that operated on the datastreams, timestamps, user identifiers that possibly triggered the events, etc.

4. APPLICATION EXPERIENCE

During the 3-year CARARE project, over two million digital records were ingested, curated and delivered to Europeana using the system presented in this paper. From the 307 SIP packages that were received, 212 were ingested (the rest were discarded for not conforming to standards). These 212 packages contained approximately 3.6 million records from which only 2.6 million records were delivered to Europeana. The rest were discarded due to quality reasons, or were duplicates, a fact that demonstrates the importance of the De-Duplication service. The scale of digital records, as well as the number of the content providers accessing and making requests to MoRe, raised significant performance issues that were addressed successfully. For example, some typical big packages contained: 748.651, 487.882, 288.634 records. These had to be processed in short timeframes in order to

meet the strict deadlines that were laid out by the project. Using MoRe we were able to cope with the continuous changes in the EDM schema without having to burden content providers in order to re-harvest data. We also managed to provide clean, enriched records with the help of the curation services. Minimum amount of effort was required by content providers, as they had to provide their data once and all other processes were handled by the repository. This approach allows for future use of the same data without the need for further effort by content providers, as this data can be manipulated in the repository. Communication with content providers, including monitoring of original data quality and notification about issues with the data, was an important issue that was also successfully addressed.

5. CONCLUSIONS

This paper presented the added-value features of MoRe, a system that aims to aggregate, curate, preserve and make available quality metadata for archaeological monuments. MoRe aggregates information from a wide community of institutions and homogenizes it, obtaining interoperability between the diverse metadata schemas they use, on the basis of common well-documented policies and a common schema for metadata submission. In addition, MoRe provides procedures for access control and user authentication.

MoRe supports workflows for the effective record keeping of all transactions, as well as micro-services for the assessment of the completeness of the submitted metadata, combined with digital curation micro-services for the enrichment of aggregated metadata, and for increasing their quality and reliability. MoRe enables content curators and adminstrators to define workflows which implement policies for specifying how and at what level digital information is preserved, and how access is provided to users. It employs a functional de-duplication service, and ensures transformation to standardized geographic co-ordinates, both important features for accessing location-based, unique cultural heritage assets through an online user interface.

In conclusion, MoRe implements services that combine constitutive traits of both aggregators and trusted repositories. It offers a carefully prioritized workflow of services, optimized for high volume, industrial grade processing of complex metadata. It integrates curation services on top of established digital preservation standards, such as conformance with the OAIS model, and PREMIS metadata audit. It shields content providers from potential updates to the delivery schema. However, its most significant contribution is in empowering content providers to adopt good practices for the creation of digital materials, and to ensure the generation of clear, meaningful and homogeneous metadata for aggregation and online access.

6. REFERENCES

[1] Abrams, S., Kunze, J., Loy, D. 2010. An Emergent Micro-Services Approach to Digital Curation Infrastructure. *International Journal of Digital Curation*. 5, 1, 172 - 186. DOI= http://dx.doi.org/ 10.2218/ijdc.v5i1.151.

[2] Drosopoulos, N., Tzouvaras, V., Simou, N., Christaki, A., Stabenau, A., Pardalis, K., Xenikoudakis, F., Kollias, S. 2012. A Metadata Interoperability Platform. In *Proceedings of the Museums and the Web 2012* (San Diego, USA, April 11-14, 2012). MW2012. http://www.museumsandtheweb.com/mw2012/programs/a_metadata_interoperability_platform

[3] Europeana. 2012. Definition of the Europeana Data Model elements, version 5.2.3. http://pro.europeana.eu/documents/900548/bb6b51df-ad11-4a78-8d8a-44cc41810f22

[4] Europeana. 2012. Europeana Semantic Elements Specifications, version 3.4.1. http://pro.europeana.eu/documents/900548/dc80802e-6efb-4127-a98e-c27c95396d57

[5] Gavrilis, D., Angelis, S., Papatheodorou, C. 2010. Mopseus – A Digital Repository System with Semantically Enhanced Preservation Services. In *Proceedings of the 7th International Conference on Preservation of Digital Objects* (Vienna, Austria, September 2010). iPRES2010. 135-143.

[6] Gavrilis, D., Papatheodorou, C., Constantopoulos, P., Angelis, S. 2010. Mopseus – A Digital Library Management System Focused on Preservation. In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries* (Glasgow, UK, September 6-10, 2010). ECDL2010. Springer-Verlag, Berlin, LNCS 6273, 445-448.

[7] Jantz, R. 2005. Digital Preservation: Architecture and Technology for Trusted Digital Repositories. *D-Lib Magazine*, 11, 6 (June 2005). http://www.dlib.org/dlib/june05/jantz/06jantz.html

[8] Moore, R., Rajasekar, A., Marciano, R. 2007. Implementing Trusted Digital Repositories. In *Proceedings of the DigCCurr2007 International Symposium in Digital Curation* (Chapel Hill, North Carolina, USA, April, 2007). https://www.irods.org/pubs/DICE_DigcCur-Trusted-Rep-07.pdf

[9] Papatheodorou, C., Dallas, C., Ertmann-Christiansen, C., Fernie, K., Gavrilis, D., Masci, M.E., Constantopoulos, P., Angelis, S. A New Architecture and Approach to Asset Representation for Europeana Aggregation: The CARARE Way. 2011. In *Proceedings of the 5th International Conference on Metadata and Semantic Research* (Izmir, Turkey, October 12-14, 2011) MTSR 2011. Springer-Verlag, Berlin, CCIS 240, 412-423.

[10] PREMIS Presevation Metadata. http://www.loc.gov/standards/premis/

[11] Reis, D., Freire, N., Manguinhas, H., Pedrosa, G. 2009. REPOX – A Framework for Metadata Interchange. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries* (Corfu, Greece, September 27- October 2, 2009). ECDL2009. Springer-Verlag, Berlin, LNCS 5714, 479-480.

[12] RLG - OCLC. 2002. *Trusted Digital Repositories: Attributes and Responsibilities*. Technical Report. RLG. http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf

[13] Ward, J.H., Wan, M., Schroeder, W., Rajasekar, A., de Torcy, A., Russell, T., Xu, H., Moore, R.W. 2011. *The integrated Rule-Oriented Data System (iRODS) Micro-service Workbook*, CreateSpace Independent Publishing Platform