The 12th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST (2011)
M. Dellepiane, F. Niccolucci, S. Pena Serna, H. Rushmeier, and L. Van Gool (Editors)

A Repository for Heterogeneous and Complex Digital Cultural Objects

A. Felicetti[1] and F. Niccolucci[2]

[1]PIN, Università degli Studi di Firenze, Italy
[2]STARC, The Cyprus Institute, Cyprus

Abstract

The paper proposes a solution for a repository of digital cultural objects, which can manage complex data as 3D objects, videos and more, together with the related metadata. The repository is built with open source components and may be easily installed and managed. Basing on an example, interfaces are shown for the most common operations. The system allows for text searches, semantic searches as well as facet refinements. The proposed system can support a full-featured digital library for its modularity and easy personalization.

Categories and Subject Descriptors (according to ACM CCS): Information Storage and Retrieval [H.3.5]: Web—based Services

1. Introduction

The increasing use of 3D technologies for Cultural Heritage is generating an increasing number of 3D replicas of cultural objects, as museum items or monuments. So far, most teams store these objects and the related information in the most diverse, and usually inefficient, ways, keeping the metadata (text information) in some database and storing the 3D digital objects by means of the file system [KFH10]. It must be noted that although often just one 3D model may be considered as "final" and used for the purpose required by the research, several intermediate models are also generated, increasing the complication of such naïve storage procedure. This was the case that generated the present paper.

One of the authors is involved with the CARARE project [CAR], which aims at providing 3D digital content to Europeana. In fact, the 3D models to be eventually produced will use a somehow simplified format - currently decided as 3D PDF - but to arrive to this result, for example starting from 3D scans, several intermediate digital objects are produced and stored in the CARARE partner's storage system. A "final" version will then be generated for ingestion by Europeana. There is the need of availing of a simple system to store all the metadata and the intermediate 3D data, and possibly the final ones as well.

For the purpose of ingestion, such a system does not need to have a sophisticated interface nor an autonomous search tool. However, since the 3D digital objects have an autonomous interest, it might be interesting to set up a system that besides harvesting also time enables autonomous access, search, retrieval and display. Our goal here is to present a system with the following characteristics:

- Storing and preserving digital assets acquired or created by the user.
- Granting access to designated users.
- Storing metadata (including provenance data, i.e. the documentation of the production process of the associated 3D object).
- Relying on a robust repository system.
- Enabling queries and possibly semantic reasoning.
- Visualizing the visual content, e.g. 3D models, possibly using special plug-ins for different formats.
- Be easy to install and maintain.

We are aware of the fact that a sophisticated repository system for 3D objects is under development within the 3D-COFORM project [3DC]. However, sometimes the richness of functions that a system provides, as the forthcoming 3D-COFOM one, goes together with the complexity of installation [PBH*10] [DTT*10].

There are also possible issues related to the physical location of the 3D assets: some cultural institutions may be reluctant to allow the circulation of reproductions of their objects because they feel unsafe (possibly with no reason)

DOI: 10.2312/VAST/VAST11/081-087

as regards the protection of their IPR. So we believe that a simple turnkey system may be of some usefulness and that is why we are proposing it now. We have tested it using a set of 3D objects with their metadata according to the CARARE metadata model.

Of course, with some easy personalization the system may accept any metadata schema, including CIDOC-CRM, and different kinds of complex objects such as audios, videos, virtual reality models and so on. However, we will fully develop the example related to the CARARE metadata schema and including 3D models as "complicated" digital objects.

2. The Overall Design

The proposed system is composed of three components, plus the interfaces towards the user: Fedora, Apache Solr and Islandora. We will briefly highlight the most important features of these three packages.

2.1. Fedora Digital Repository

The Fedora digital repository provides a flexible digital content repository, which can be adapted to a wide variety of scenarios and can store any kind of digital content including images, videos, datasets and so on, together with a complex network of relationships linking the digital objects to each other. Fedora is based on a very rigid but perfectly integrated "dual" core of data and metadata management, offering an affordable interaction between the digital objects and their metadata, and the possibility to attach a set of modules providing a plethora of distributed services for the various ingestion and retrieval operations [FED]. Even if Fedora can be used as a standalone repository service, its power resides in its flexibility that makes it easy to integrate into an application or system that provides additional functions to satisfy particular user needs (e.g. a robust triple store or a fast and reliable query/retrieval framework) [LPSW05].

The core of our repository is currently based on version 3.4 of the Fedora digital archive, which natively provides:

- A digital object repository to ingest, store, aggregate manage and extract digital objects (images, videos, documents and other relevant files).
- A semantic resource index (i.e. an instance of the Mulgara triple store) which provides the infrastructure for indexing the complex RDF network of metadata concerning the relationships between objects and components.

On top of the core, Fedora also provides a set of modules which implements the most typical operations of a digital repository, like for instance the content versioning module, capable to track when a change is made on a certain object and by whom, and to create a new version of the modified data every time a change occurs.

Another very interesting and important module is the OAI-PMH module, to implement an OAI-PMH repository for the standard publication of the information concerning the digital objects stored in our archive towards other institutional repositories (such as the Europeana one).

Internally Fedora uses Dublin Core and an internal object model (FOXML) to represent the stored digital objects and collections and their mutual relationships (such as the *isPartOf* and the *isMemberOfCollection* relations, used to describe the internal tree structure of the collections). The METS metadata schema is also available to be used for the same purposes. But, since the semantic Resource Index deals with XML information, any other schema can be used in addition to the default ones or in substitution of them. Thanks to this flexibility we were able to implement a fully featured CARARE metadata schema for every description of heritage assets and for their relations to digital resources, activities and to collection information.

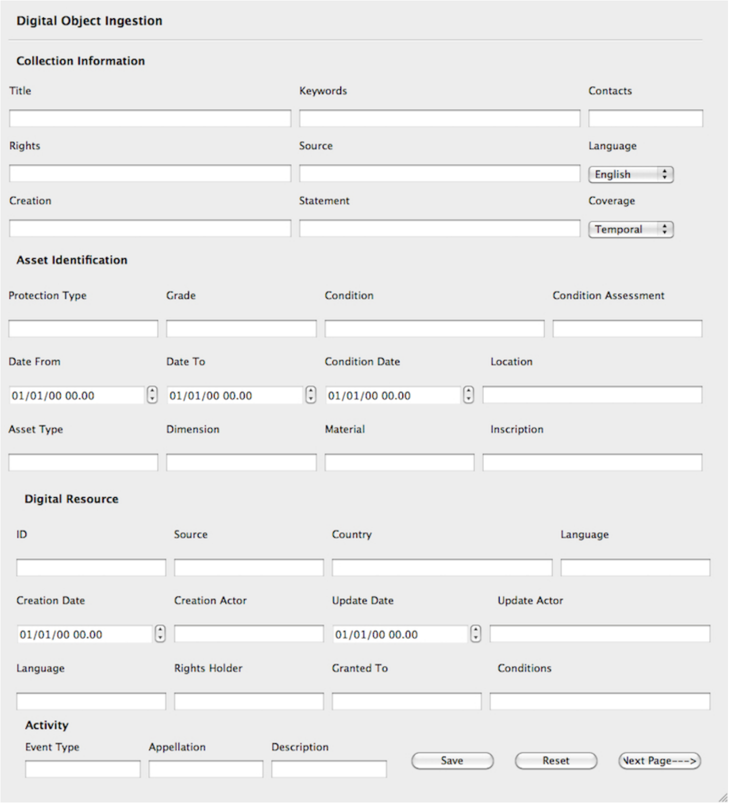
2.2. Islandora

One of the main issues every user or developer has to face to use Fedora for digital data management is that Fedora comes with very poor user interfaces and does not provide any standard entry point from which is possible to manage every aspect of the framework, like for instance the ones provided by EPrints or DSpace. This is of course justified, from the developers' point of view, with the flexibility of the modular approach of Fedora, which allows users to build up an *ad hoc* interface to suit their particular needs. But since the development process is usually a very tedious and time-consuming task, it would be useful to have, for instance, some middleware libraries or tools providing standard modules already connected with the Fedora core, which can afterwards be customized by the developers.

Recently a number of such middleware tools have been developed by third parties. As of today, one of the most promising is the Islandora module developed by the University of Prince Edward Island's Robertson Library, whose main aim is to provide standard and customizable interfaces to the most common functionalities required to manage a Fedora digital repository [ISL].

Basically Islandora is a PHP framework which uniquely combines and harnesses the power of the Drupal content management system [DRU] to create a robust digital asset management system based on Fedora. This means that users can discover, view, and manage Fedora objects through Drupal, to take advantage of the Drupal modularity and flexibility in providing user-friendly interfaces for content management.

Additionally, the Islandora framework already comes with a customizable instance of Apache Solr, to bring lightning-fast searching capabilities on the Fedora database, and the related interfaces to take advantage of its power.



Language English

Coverage Temporal

Date From 01/01/00 00.00

Date To 01/01/00 00.00

Condition Date 01/01/00 00.00

Creation Date 01/01/00 00.00

Update Date 01/01/00 00.00

Figure 1: *The First Page of the Interface for Manual Data Ingestion.*

2.3. Solr

Solr is a scalable, open source and extremely powerful search platform that provides, among other features, a dynamic geospatial search, a strong integration framework and one of the most advanced faceted searches available today [SOL]. It also perfectly integrates with FEDORA through a set of services making it capable of indexing the entire FEDORA Resource Index and to perform full-text searching of any attached documents [DPWG10].

3. Ingestion

The ingestion stage is the most delicate of the entire process. Usually there are no particular problems when uploading digital objects, even if encoded in proprietary formats, thanks to the abstraction of the container and to its capability to store any type of file it receives. But the metadata should be validated before ingestion, to debug the XML code from syntactic errors and to check the structural and semantic coherence towards the chosen schemas. Based on the different steps necessary to create the metadata, there are two main ways to ingest digital content into the repository: the first requires a preliminary operation to be executed before the storage itself, which consists in collecting all the available information in RDF to create the metadata layer describing the digital objects.

A typical scenario of this process can be observed during the process of creation of provenance metadata: when a digital object is created, the acquisition device (i.e. a scanner, a digital camera etc.) also creates the so-called provenance information, i.e. information about the acquisition conditions (e.g. type and model of acquisition device, environmental conditions, author, and so on) often encoded in a non-standard format. The standardization of provenance metadata still remains a big challenge to overcome. Anyway, it is always possible to retrieve and put this information into a standard format (i.e. CIDOC-CRM and CRMdig) through various mapping operations, even if the mapping process is often slowed down by the multiple and different proprietary formats used by each acquisition tool. It must be noted that since the digitization stage is beyond the scope of the CARARE project - which is the provision of 3D digital content to Europeana - the metadata schema does not incorporate any provenance information. However, there is work in progress to include also provenance information and enable adoption of the schema outside of the project context.

The actual ingestion is performed in a second stage, where

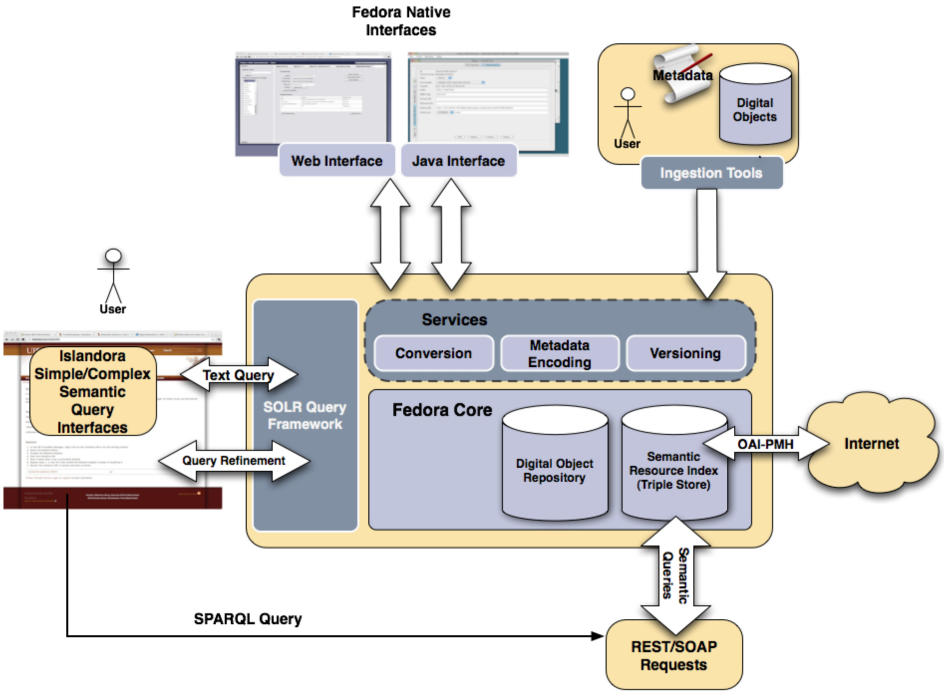


Figure 2: *Overall scheme of the system and details of the query system.*

a package containing both digital data and metadata (the submission information package, SIP) is created and sent directly to the repository. It is important to notice that with this technique it is possible to include many objects and the related metadata into the package (in order to create a compound object) and to invoke one big ingestion process for all of them.

The second way of data ingestion is straightforward: the user ingests the objects in sequence, one by one, by creating metadata on the fly to "describe" each digital object before sending it to the repository.

4. Ingestion Tools

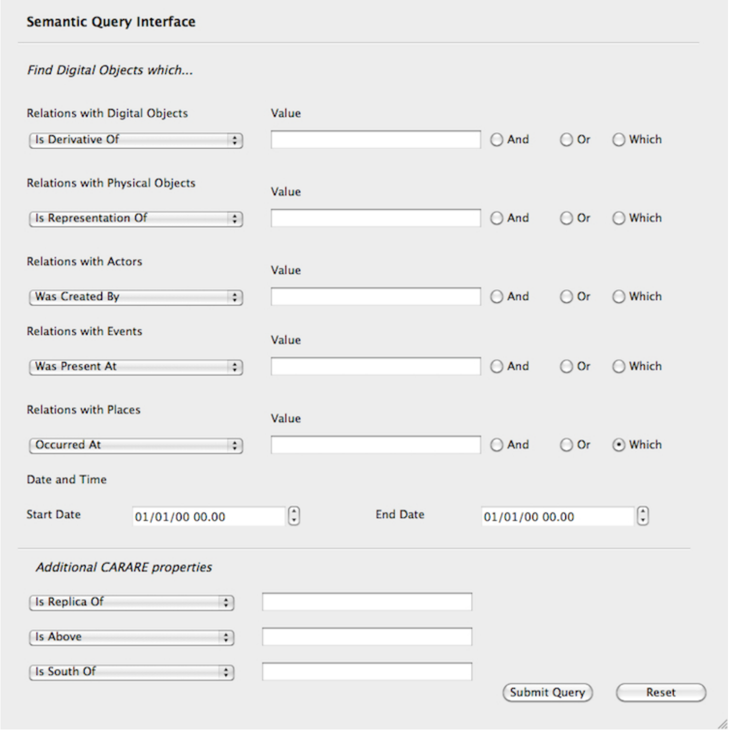
The two different ingestion models described above require different ingestion tools and interfaces. For the first case (automated ingestion) we have designed a tool that includes a validator to check the coherence and the validity of the metadata against the ontologies; a packager to combine digital data and metadata; a transfer framework to serialize and put the content of the package in the right place inside the repository.

Optionally the tool might include also a mapper to convert provenance metadata into valuable CIDOC-CRM/CRMdig or CARARE entities (RDF). In the second case (manual ingestion), a set of interfaces to input relevant information is required.

We have extended the Drupal interfaces provided by Islandora, which natively provides features to encode full instances of the Dublin Core metadata for the digital object to be ingested into Fedora, and added a set of new interfaces able to capture all the information required by the adopted metadata schema, in our case CARARE.

Such interfaces are also able to create RDF representation of the metadata inserted by the user on the fly, every time the ingestion process is invoked to transfer each object to the repository. Since the interfaces are designed ad hoc to collect the required information, the controls on the validity and coherence of the encoded metadata is easier in this case. The procedure has been tested on the CARARE metadata schema but it could be any other.

Whatever interface is used, according to the user's choice of ingestion mode, the FEDORA system is always able to generate on the fly the basic Dublin Core description of each compound or single objects upon creation; the internal structural relationships inside the compound object are provided through the SIPs as well and encoded using the METS format. Additionally, each SIP could contain the thumbnails of the 3D models, very useful when visualizing a preview of the object during the browsing and query/retrieval operations. All the metadata information will also be stored in the semantic Resource Index and used to extend the internal semantic network with the necessary descriptions of the



Semantic Query Interface

Find Digital Objects which...

Relations with Digital Objects — Is Derivative Of — Value: — ○ And ○ Or ○ Which

Relations with Physical Objects — Is Representation Of — Value: — ○ And ○ Or ○ Which

Relations with Actors — Was Created By — Value: — ○ And ○ Or ○ Which

Relations with Events — Was Present At — Value: — ○ And ○ Or ○ Which

Relations with Places — Occurred At — Value: — ○ And ○ Or ● Which

Date and Time

Start Date: 01/01/00 00.00 — End Date: 01/01/00 00.00

Additional CARARE properties

Is Replica Of

Is Above

Is South Of

Figure 3: *The Semantic Query Interface.*

new objects. The new information, once uploaded, will be immediately made available to all the other services.

5. Search and Retrieval Features

Taking advantage of the Islandora and Solr frameworks, there are many different ways to query the digital repository.

The simplest query mechanism consists of full text search on the metadata, looking for a term present in them. This extends also to the ingestion event, for example to search all objects ingested on 22/7/2011, because the ingestion (Dublin Core) metadata are automatically created and stored by the FEDORA system. The search may be refined adding more parameters.

Secondly, advanced search criteria and faceted browsing of the results are provided by Islandora through Solr as well.

The last query option is the possibility of using the SPARQL technology to perform semantic queries against the Resource Index of Fedora. This service is already included in the basic FEDORA distribution and it is reachable via REST and SOAP protocols. But the graphic interface offered by Fedora to the user is very minimal and the SPARQL query needs to be input by hand.

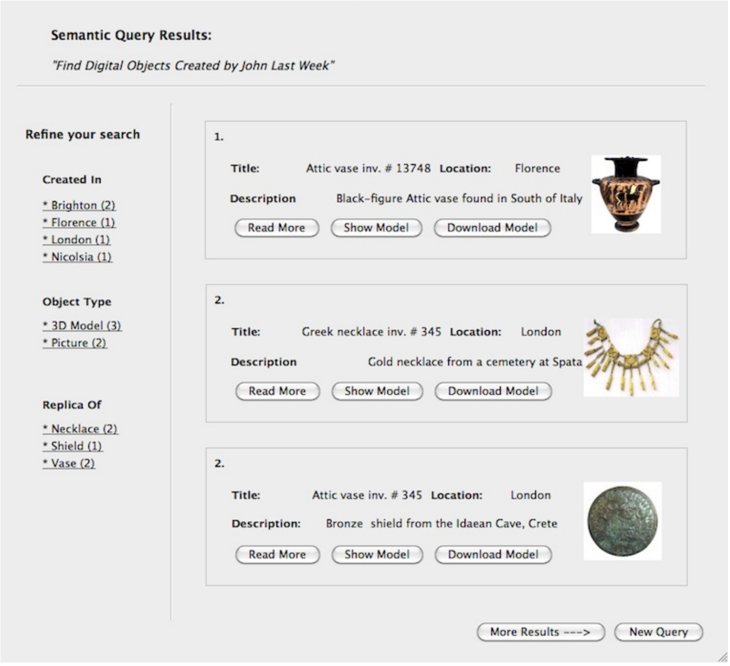
Again the Islandora flexibility has allowed to create a prototype of a new interface, which will receive the different query options specified by the user by using the entities and properties of the underlying metadata schema (CARARE in our case) and will translate them in SPARQL language to be executed by the Resource Index to perform queries like the following:

"Show me all the digital objects created by John during last week."

Then the translated query will be sent to the repository through SOAP or REST and the resulting information returned to the user via another Islandora instance which will show the list of requested digital objects (with the related thumbnail for preview). The faceted browsing functionalities of Solr can be used also here to refine the visualization of the results by slicing them according with other parameters [FAC].

After the result of the query operations is displayed, the user can perform any of the following operations:

- Browse the related collection.
- Visualize all the available versions and the available formats of a given digital object.
- Visualize (if possible) a chosen version of the object in a browser. This operation will be performed by the conversion framework, which will create (where possible) wa eb representation of the selected object and will publish on an Islandora/Drupal dynamic page.



Semantic Query Results:

"Find Digital Objects Created by John Last Week"

Refine your search

Created In

* Brighton (2)
* Florence (1)
* London (1)
* Nicolsia (1)

Object Type

* 3D Model (3)
* Picture (2)

Replica Of

* Necklace (2)
* Shield (1)
* Vase (2)

1\.

Title: Attic vase inv. # 13748 Location: Florence

Description Black-figure Attic vase found in South of Italy

2\.

Title: Greek necklace inv. # 345 Location: London

Description Gold necklace from a cemetery at Spata

2\.

Title: Attic vase inv. # 345 Location: London

Description: Bronze shield from the Idaean Cave, Crete

Figure 4: *Display of Query Results.*

- Download the original object together with the related metadata, for personal use or for further processing and enrichment.

6. Conclusions and Future Work

The system presented in this paper solves, in our opinion, a number of needs for CH practitioners and researchers. It provides a simple but effective way of managing digital assets created within research, and allows very flexible queries to be carried out on the repository. It can be installed and work very quickly, as it can be distributed together with all the necessary libraries and set-up parameters. At the same time, personalization is easy and extension to different metadata schemas as well as object types is straightforward, requiring only the preparation of appropriate interfaces.

Plans for future development include improving the interfaces that so far are just essential; testing the system on different datasets and metadata schemas; and extending the system to a distributed repository.

7. Acknowledgements

Authors have received partial support for this work by the European Project 3D-COFORM contract number 231809.

References

[3DC] Tools and expertise for 3D collection formation. http://www.3d-coform.eu/.

[CAR] Carare Project. http://www.carare.eu/.

[DPWG10] Devarakonda R., Palanisamy G., Wilson B. E., Green J. M.: Mercury: reusable metadata management, data discovery and access system. *Earth Science Informatics* (2010), 87–94.

[DRU] Drupal: Open Source CMS. http://drupal.org/.

[DTT*10] Doerr M., Tzompanaki K., Theodoridou M., Georgis C., Axaridou A., Havemann S.: A repository for 3d model production and interpretation in culture and beyond. In *VAST10: The 11th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage* (Paris, France, 09/2010 2010), vol. Full Papers, Eurographics Association, Eurographics Association, pp. 97–104.

[FAC] Faceted Search with Solr (by Yonik Seeley), January 2009. http://www.lucidimagination.com/Community/Hear-from-the-Experts/Articles/Faceted-Search-Solr.

[FED] Fedora Repository Project: General Purpose, Open Source Digital Object Repository System. http://www.fedora-commons.org.

[ISL] Islandora: building a rich digital repository ecosystem. http://islandora.ca.

[KFH10] Koller D., Frischer B., Humphreys G.: Research challenges for digital archives of 3d cultural heritage models. *J. Comput. Cult. Herit. 2* (January 2010), 7:1–7:17.

[LPSW05] Lagoze C., Payette S., Shin E., Wilper C.: Fedora: An architecture for complex objects and their relationships. *Complex Objects and Their Relationships, International Journal of Digital Libraries 6(2)* (2005).

[PBH*10] Pan X., Beckmann P., Havemann S., Tzompanaki K., Doerr M., Fellner D. W.: A distributed object repository for cultural heritage. In *VAST10: The 11th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage* (Paris, France, 09/2010 2010), vol. Full Papers, Eurographics Association, Eurographics Association, pp. 105–114.

[SOL] Open Source Enterprise Search Platform From The Apache Lucene Project. http://lucene.apache.org/solr/.