METADATA AND TOOLS FOR INTEGRATION AND PRESERVATION OF CULTURAL HERITAGE 3D INFORMATION

Achille FELICETTI[1], Matteo LORENZINI[2]

[1]PIN, Università degli Studi di Firenze
Piazza Ciardi 25, 59100 - Prato, Italy

achille.felicetti@pin.unifi.it

[2]ICCU, Ministero per i Beni e le Attività Culturali
Viale Castro Pretorio 105, 00185 - Roma, Italy

matteo.lorenzini@beniculturali.it

Keywords 3D, Digital Repositories, Metadata, Open Source, Ontologies, CIDOC-CRM

Abstract:

In this paper we investigate many of the various storage, portability and interoperability issues arising among archaeologists and cultural heritage people when dealing with 3D technologies. On the one side, the available digital repositories look often unable to guarantee affordable features in the management of 3D models and their metadata; on the other side the nature of most of the available data format for 3D encoding seem to be not satisfactory for the necessary portability required nowadays by 3D information across different systems. We propose a set of possible solutions to show how integration can be achieved through the use of well known and wide accepted standards for data encoding and data storage. Using a set of 3D models acquired during various archaeological campaigns and a number of open source tools, we have implemented a straightforward encoding process to generate meaningful semantic data and metadata. We will also present the interoperability process carried out to integrate the encoded 3D models and the geographic features produced by the archaeologists. Finally we will report the preliminary (rather encouraging) development of a semantic enabled and persistent digital repository, where 3D models (but also any kind of digital data and metadata) can easily be stored, retrieved and shared with the content of other digital archives.

1. INTRODUCTION

1.1 Cultural heritage and 3D technologies

3D modeling has become widespread in many areas of archaeological research and nowadays comprises a wide range of applications to cover every aspect of the archaeological work (e.g. landscape analysis, excavation area documentation, creation of digital representations of monuments and artifacts). After some initial skepticism, 3D seems to have finally seduced the cultural heritage experts and, especially in archaeology, 3D technologies are increasingly used not only for the typical operations of reconstruction and presentation to the public, but also for restoration and preservation purposes. The use of information technology to capture or represent the data studied by archaeologists, art historians and architects falls now under the name of *Virtual Heritage*, a brand new and fascinating branch of knowledge [1]. A lot of work has been carried out during the last decade and a huge amount of digital information has been produced. But at the same time, this rapid and uncontrolled growth has given rise to brand new issues, which must now be faced.

One of the most common concerns, not only in the cultural heritage field, is interoperability. Portability and integration of 3D information across different systems and over the Web is very often impeded by the 3D

acquisition and processing tools, mainly because of the proprietary (i.e. "closed") data formats used to encode information, but also for the non-standard way in which these tools capture the *provenance* information of each digital model. Other relevant obstacles regard the long and often too diversified pipeline necessary for the creation of the final 3D model: the whole encoding process is usually split into a series of complex operations (from digital acquisition to the final creation of the model, through the various processing steps) which are very often performed by different people at different locations and times. Interoperability requires a solution for these kinds of problems and all our efforts are conveyed towards this purpose.

2. INTEROPERABILITY, OPEN SOURCE AND OPEN STANDARDS

2.1 3D Content and metadata formats

The history of standardization of 3D data formats is longer and perhaps more complex than that of any other digital resource. One of the reasons behind this is that the 3D formats have been (and are even today) strictly bound to the various tools and software used in turn to acquire and process 3D information. Most of these tools and software still provide proprietary file formats to encode the resulting 3D objects. As a result this practice makes 3D content very difficult to share and exchange. Proprietary data formats have always created barriers to the integration among 3D data and the other 2D digital information and have for a long time impeded the creation of a suitable open and standard format able to encompass all the interoperability issues. Popular 3D formats like DWG and 3DShape are still considered de facto standards, although they frustrate any possibility of efficient data sharing and exchange due to their "closed" nature.

One attempt to create an open format was performed by Adobe, who invented a way to embed 3D content into PDF documents. 3D PDF is now an ISO standard (ISO 32000-1:2008) enabling users to create their own 3D PDF library and related software [2]. But its range of use is very narrow due to its absolute lack of flexibility and inadequacy in facing the huge demand of interoperability required by modern information technology when dealing with 3D. Even if many attempts have been made to implement a conversion framework between (at least) the most popular formats, many issues of data loss still remain while converting from one 3D format to another.

As mentioned above, a serious matter of non-integration affects the metadata generated by the 3D model creation pipeline (provenance data). In a very common scenario, the format of these metadata remains strictly related to the acquisition/processing tools. This information, which is usually very detailed and comprises in some cases of hundreds of data fields, is often encoded following *ad hoc* and non-standard schemas, making it impossible to compare it to and integrate with other similar information coming from different sources.

2.2 Open formats and open standards for 3D encoding

Most of the possible solutions to all of the format problems mentioned above rely on the adoption of *open* standards and on their ability to guarantee the necessary portability and cross-compatibility of digital information. Open standards can be used throughout the whole pipeline, from the acquisition/creation of the 3D model and the related metadata, to the processing and presentation operations, until its storage in digital repositories. Many formats already exist and provide a good degree of standardization: the COLLADA, for instance, was created to represent 3D models with a standard syntax [3] and many important applications, like Google SketchUp, natively support it; taxonomies like CityGML [4] can describe specific elements of a 3D scene and their mutual relationships, and make them interact with the spatial elements of a geographic context; the new HTML5 specification will hopefully simplify the visualization of X3D-encoded models on the Web [5].

2.3 Metadata formats

Standardization is not only a matter of data formats, but it is something that can also be achieved on other levels, for example by providing 3D information with valuable sets of metadata. Many improvements have been achieved in this field, particularly after the publication of various guidelines and recommendations addressing the importance of having good quality and standardized metadata for documenting 3D content, especially in the cultural heritage field. The contribution of the London Charter for computer-based

visualization of cultural heritage has been one of the most essential to overcome this issue [6]. Today many ontologies and schemas are available which produce standard sets of metadata. Most provide rich collections of classes and properties to capture every degree of granularity required for the description of 3D models and for their enrichment with annotations and other similar techniques. In particular CIDOC-CRM and its derivatives are among the preferred ontologies for the representation of cultural heritage descriptive metadata [7], even if the difficulties in using these schemas impose on the developers a high degree of automation and the creation of extremely user-friendly interfaces.

The world of digital libraries has always made use of various metadata schemas to describe the archived objects and if the Dublin Core model has always been preferred for its simplicity in encoding basic description of objects, nowadays the METS is becoming more increasingly used for the so called *structural metadata,* which describes the logical or physical relationships between the various parts of a compound object [8].

One of the biggest requirements in 3D documentation is the so called *digital provenance*, which puts the creation pipelines in relation with the digital objects, just as the *physical provenance* puts physical places in relation with physical artifacts. Digital provenance is gaining increasing importance in cultural heritage research and practice since it deals with the uninterrupted chain linking the original to the processed outcome. The detailed documentation of this chain provides the necessary transparency and, from the scientific point of view, the repeatability and verifiability of the whole process. However this can occur only when the documentation is properly acquired. The provenance recording process should also be dense enough to document every step of a digital object's life in order to build a complete *fingerprint* for the preservation of the necessary referential integrity of the metadata.

Many new standards are also appearing on the stage for the encoding of provenance information. The CRMdig seems to be one of the most promising among them [9]. CRMdig is an extension of the CIDOC-CRM ontology and was developed for the documentation purposes of the 3D-COFORM project [10]. It provides an event-centric model to capture the *technical metadata* typical of the data acquisition. CRMdig provides a superclass "Data Acquisition Event" and many subclasses to describe the various related sub-events and the entities involved in the process, e.g. the information concerning the acquisition tools (calibration, data formats), the actors participating in the acquisition event and the places where the acquisition event was performed. The model is intended to provide a flexible infrastructure to build provenance information in a very precise way.

3. DIGITAL CONTAINERS FOR 3D INFORMATION

3.1 The quest for a 3D digital repository

Along with the issues of data formats and metadata creation, there is another challenge to face in order to achieve real interoperability, a flexible exchange and a safe preservation of the 3D digital content. It concerns the nature of the available digital repositories, their inadequacy in dealing with 3D models and in guaranteeing affordable storage features, reliable and meaningful data retrieval and long-term preservation of digital artifacts and metadata.

Good results have recently been achieved by the open source community on the "repository" side: if the storage model based on the MPEG-21 "containers" for data and metadata recommended by the EPOCH project was lacking in flexibility, recent developments of new semantic-oriented paradigms make the storage/retrieval operations much more effective. We have surveyed many existing content and media management repositories to find a flexible and adaptive technology. At the end of our investigation we focused our research on the most interesting ones as of today: the 3D-COFORM Repository Infrastructure and the digital repository provided by Fedora Commons.

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. Even if Fedora can be used as a standalone repository service, its power is in its flexibility which 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) [11].

The 3D-COFORM Distributed Object Repository is a digital repository developed to provide cultural heritage experts and practitioners with a working platform to access, use, share and modify digital content. It comes as an integrated repository to ingest, store and manage complex digital objects together with the related metadata, to enable efficient access and to export the information for reuse in other contexts. The 3D-COFORM repository also provides features to manage the digital object provenance information, descriptions and semantic classification of the modeled objects, including their physical location, their history, sources and expert annotations about modeling and related historical data [12].

We have chosen to base the core of our experimental system on the Fedora architectural model instead of using the 3D-COFORM infrastructure, mainly because the latter, notwithstanding the good transparency kept towards the external services interacting with the core, implements an internal separation and decentralization between the object archive and the metadata archive in the core itself [13]. This architectural approach very often causes a lack of performances at the level of the core itself, i.e. in the very place where high performances are constantly required. The modular approach provided by Fedora Commons, with a very rigid but perfectly integrated "dual" core of data and metadata management, assisted by distributed services, offers a wide range of possibilities and can guarantee fast and affordable interaction between the digital objects and their metadata.

We will make the connected services, developed on top of the Fedora core, compatible with the 3D-COFORM Repository Infrastructure as soon as its architecture will become stable and the final version of the necessary 3DC communication APIs will be released. For this purpose we are already implementing the same data models used by the 3D-COFORM project, in particular the CIDOC-CRM ontology for structural and descriptive metadata and the CRMdig model (among the others) for the encoding of provenance metadata, in our system.

3.2 The Fedora-based 3D repository

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

- **A digital object repository** to ingest, store, aggregate manage and extract digital objects coming from different institutions in different formats (images, videos, documents and other relevant files).

- **A semantic resource index** that provides the infrastructure for indexing the complex network of information regarding relationships between objects and components. We have extended this module to support the storage of the RDF representation of all the metadata used within the system (including CIDOC-CRM, METS and CRMdig).

On top of the core we have implemented a set of services, both by extending some already provided by Fedora itself and by developing ex novo the services required and not available in the Fedora framework, using open source technology (Figure 1). In particular, one of the main issues we've had to face concerned the search/retrieval service: Fedora actually provides only a basic query/retrieval mechanism based on textual search to look for digital objects and a very trivial bunch of functions to query metadata. For this reason we bound a SOLR framework to the Fedora core. SOLR is a scalable, totally open source and extremely powerful enterprise search platform which provides, among other features, a dynamic geospatial search, a strong integration framework and one of the most advanced faceted searches available today [14]. For this and many other reasons, SOLR has been chosen for the construction of the *query framework.* The other available services provided by the system are:

- **A set of ingestion tools and technologies,** for the preparation of SIPs, the standard packages for the ingestion of digital objects and metadata and for the appropriate archiving of all the digital and semantic information available. These tools also provide user interfaces to reduce user interaction and to guide the user through all the different phases of the metadata and URI creation to guarantee full internal compliance with the digital archive.

- **A conversion framework,** to create descriptions of the 3D models in COLLADA, X3D, CityGML (and if required, other open formats) and to make them available for online visualization and download.

- **An enrichment mechanism,** to combine existing metadata with new information regarding the same objects (e.g. geographic information, data coming from different thesauri, annotations and so on). The enrichment framework is also able to create "aggregated objects" by adding semantic and geographic information directly into the COLLADA and X3D code.

- **A content versioning mechanism**, to track when a change is made on a certain object and by whom. Every time a change occurs, a new version of the modified data is added to the object's metadata. This allows users to retrieve older versions of a data object by performing a "date and time" search, or to retrieve the "current version", i.e. the one that is most up-to-date.

- **A query framework** able to interact with all kinds of metadata and the semantic relations between them for retrieval, conversion and redistribution of the 3D models and the related metadata information. As explained above, this framework is implemented by using the SOLR technology. Fedora also provides a SPARQL endpoint which allows to query the semantic resource index directly.

- **A set of plugins** for Blender and QuantumGIS which allows them to directly interact with the repository to download and re-ingest the digital content and to perform annotation and geographic data enrichment of the 3D models.

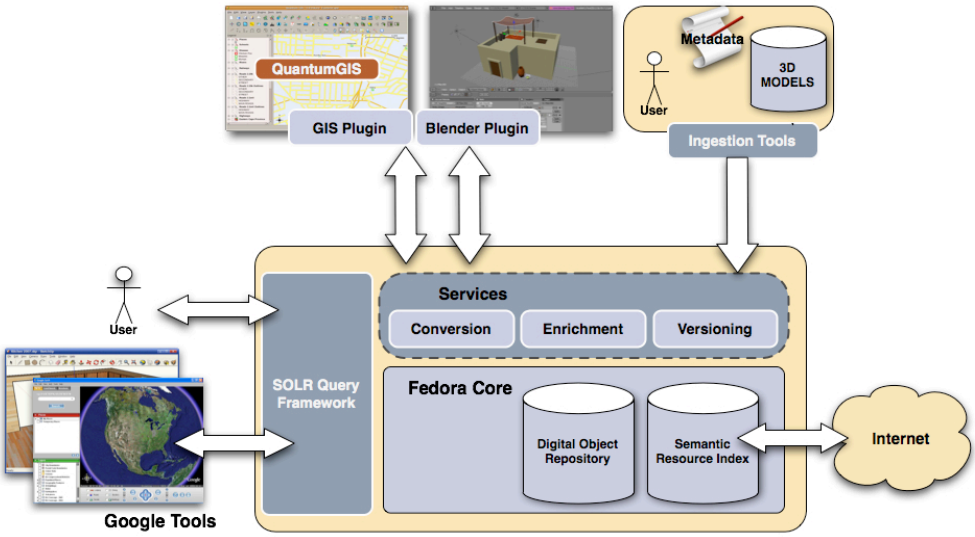

Figure 1: The structure of our 3D digital repository

3.3 Ingesting operations

The operations required by the system to archive, retrieve and manage the 3D digital content are straightforward. The preliminary operation to execute before storing 3D content is to collect all the available information in order to create rich sets of metadata. In an ideal scenario, the acquisition tools (i.e. laser scanner, digital cameras etc.) would be able to provide the necessary provenance metadata together with the 3D models they produce in a standard encoded format. A lot of effort has been put into resolving this issue within the 3D-COFORM project; but the direct production of standard metadata during the acquisition phase 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 multiples and different proprietary formats used by each acquisition tool. When both the digital and the information content is ready, a package is created to ingest it all into the repository.

The ingestion stage is the most delicate of the entire process. Normally we don't encounter particular problems when uploading the 3D model(s), 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. If metadata pass all the necessary validity tests, a SIP containing

both the digital data and the metadata can be created and sent to the repository to be stored. At ingestion time, a set of new compound objects are created in accordance with the physical objects or monuments digitized with the acquisition operations; after ingestion, each 3D model and metadata set referring to a particular object or monument becomes a *datastream* of the related compound object.

The system is able to generate on the fly the Dublin Core and CIDOC-CRM description of each compound object 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 new objects. The new information, once uploaded, will be immediately made available to all the other services.

3.4 Search and retrieval

There will be different ways to query the digital repository, in particular it will be possible to retrieve 3D models of a particular object or monument by querying the descriptive metadata or the information in the semantic network using the SPARQL endpoint provided by the system. Refinements of the queries and advanced search criteria can be specified from the SOLR-based query interface or by using the faceted browsing facilities offered by the SOLR framework. As a result of the query operations, a set of 3D models (with the related thumbnail for preview) will be returned. The user can then perform the following operations:

- Visualize all the available versions and the available formats of a given 3D object.
- Load a chosen version of the object in a browser. This operation will be performed by the *conversion framework* which will create (where possible) an X3D representation of the selected model and will publish it through an HTML5 page created on the fly.
- Download the original object or a standard encoded version of it (in COLLADA, X3D or CityGML) together with the related metadata, for personal use or for further processing and enrichment.
- Get a Google-compatible COLLADA+KML representation of the object (if geographic information is available), suitable to be used in different scenarios (e.g. loaded in Google SketchUp, Google Earth/Maps and other similar applications).

3.5 3D Content enrichment

The standard encoded versions of all the 3D models stored in our archive are ready to be enriched with geographic and *structural* information using the QuantumGIS plugin and the Blender CityGML markup plugins. The metadata provided by the system give the possibility to immediately upload the processed 3D models as an extension of the already existing compound objects without the need to recreate metadata from scratch. The new processing information will also enrich the provenance metadata of the existing objects to extend the internal semantic network. The processed 3D models, once re-ingested, will become new datastreams of the original compound object. The original 3D models will never be affected by any conversion or processing operation: they will continue to exist in their original format until a delete operation will be explicitly invoked.

4. TESTING THE SYSTEM

4.1 The Uchi Maius dataset

The digital content we have used to test our repository comes from the digital archive of Uchi Maius, an archaeological excavation site located about 100 km south of Tunis which includes Roman and Islamic remains. The archive comprises of the 3D model of the whole excavation area and many 3D models of single monuments. It was created by the University of Pisa and the University of Sassari (both in Italy), which surveyed the area in 2002 by using a total station (Leica TCR 307) and a digital camera with calibrated parameters (Canon EOS 400) [15]. We used all the information coming from these tools for creating the provenance metadata for our digital content. A good set of information regarding the acquisition process (calibration, resolution of the tools etc.) is also available. Additional surveys and measurements were carried

out to take notes of the most articulated details of the buildings and their apparatus and to acquire a very detailed set of spatial data describing the whole area in geographic terms; most of this information was used to test the data enrichment framework.

We have prepared various SIPs, each one containing the digital content (i.e. the 3D models in .dxf format), the METS description of compound objects' structure to be recreated and the provenance metadata encoded using CRMdig. Afterward we validated and ingested the whole content into the repository. The SIPs creation was carried out by using a preliminary version of one of the ingestion tools that we are developing and that will provide (in its final version) all the basic functions for metadata creation and validation, for SIPs aggregation and upload into the central repository. The tool is currently in a very early stage of development and only a basic metadata creation mechanism (based on templates) and the upload service are fully working. The metadata validation has been performed manually by using the various XML plugins provided by the JEdit editor.

After the ingestion of 3D models and metadata, we tested the query/retrieval system in order to verify the performances of the SOLR framework and of the SPARQL endpoint. In both cases we got meaningful information on our digital objects and on the internal relationships between them.

4.2 Data enrichment of the Uchi Maius data

In a previous work we already described in technical terms the process of building a CityGML representation of a 3D models and how the archaeologists can enrich the CityGML code with CIDOC-CRM entities to insert semantic information into the model itself (e.g. historical information like the year of foundation or destruction of the city and so on) [16]. We have the process more efficient and the interaction of the Blender plugin with the system is now straightforward. The new Blender plugin is now able to download the CityGML encoded version of one of our 3D models (generated directly by the repository), to import it into Blender, to enrich it with CIDOC-CRM code and to re-ingest it back into its original context (Figure 2):

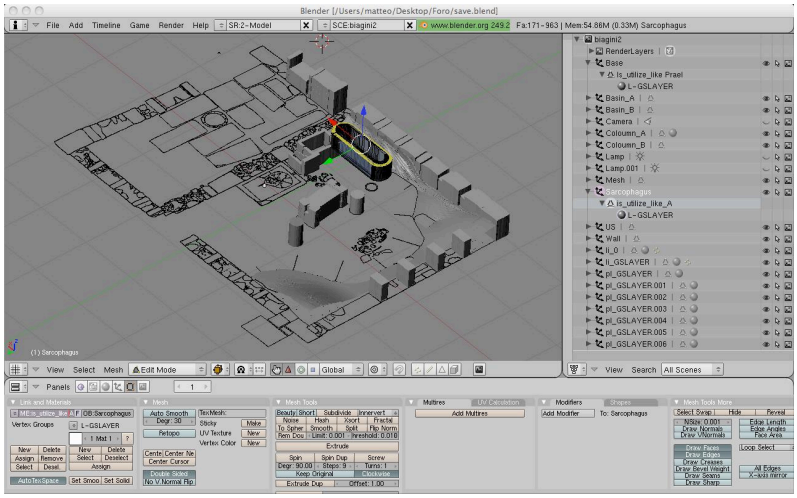

Figure 2: The Blender plugin in action on the Uchi Maius 3D Data

The same operations were performed with the QuantumGIS plugin, which is now able to work in a similar way on the same CityGML code to create geographic information for a given 3D model and ingest them into the repository. The rich set of spatial information provided by the archaeologists for the Uchi Maius excavation site was used to enrich the COLLADA representation of the 3D models with a lot of spatial data which made them suitable to be exported and used in a Google Earth/Google Maps context.

5. CONCLUSIONS AND FURTHER WORK

This activities described in this paper are the prosecution of a very fruitful collaboration between PIN, University of Florence and the Italian Ministry for Cultural Heritage started in 2010 and using the Uchi Maius dataset kindly provided by the archaeologists of the Department of History of the University of Sassari [16,17]. The final goal of our effort is the definition of an open repository of 3D cultural heritage models, providing standard mechanisms for preservation, updating, and dissemination. Even if we are still at the very beginning of the development activity, the preliminary results seems very encouraging mainly

because of the maturity of the open source technology we are using and the affordability of the available standards, able today to provide a true and solid integration environment, which was impossible to imagine in the past.

Notwithstanding this optimistic view, a lot of work remains to be done and most of it concerns the adaptation of the sophisticated technology we are dealing with to the every day needs of the cultural heritage people, usually very sensible to the interoperability issues, but rarely willing to sacrifice the simplicity and usability of the tools for their achievement. A lot of development is still needed also at the repository level: even if most of the integration problems can be solved today with the tools we have already implemented, other issues still remain: detailed and high quality metadata are required to achieve the scientific authentication of the 3D artifacts and various issues of digital preservation remain open. Fedora and other similar digital object repositories have always kept an eye on preservation, but a more active management of the life cycle of digital resources, from data creation and management to data use and rights management, needs to be implemented in order to reduce the risk of loosing control over the digital content and compromise their survivability.

In the future we will focus our effort on the development of more clear and user-friendly interfaces and on good quality documentation on how to use them. The stability of the core and the query/retrieval service make us free to concentrate our work on the improvement of the other services, in particular on the conversion and enrichment mechanisms. The system is already able to host standard-encoded thesauri (in SKOS) and gazetteers. Future development will extend the enrichment framework to take advantage of this kind of knowledge as well. We are also planning to implement an OAI-PMH repository to publish information concerning the 3D digital objects stored in our archive. The improvement of the existing plugins and the development of new ones to extend the functions of our system will be among the priorities of our future development activity.

6. REFERENCES

[1] Koller, D., Frischer, B., Humphreys, G.: *Research challenges for digital archives of 3D cultural heritage models*, ACM J. Comput. Cult. Herit. 2, 3, December 2009, Article 7.

[2] *3D PDF Technology,* http://www.adobe.com/manufacturing/solutions/3d_solutions/.

[3] *COLLADA: Digital Asset and FX Exchange Schema,* http://www.collada.org.

[4] *CityGML: a common information model for the representation of 3D urban objects,* www.citygml.org.

[5] *X3D: Open Standards for Real-Time 3D Communication*. http://www.web3d.org/x3d/.

[6] *The London Charter Initiative*, www.londoncharter.org, June 2006.

[7] Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M.: *Definition of the CIDOC Conceptual Reference* Model. Tech. rep., http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.1_Mar09.pdf, March 2009.

[8] *Metadata Encoding and Transmission Standard (METS)*: http://www.loc.gov/standards/mets/.

[9] Pitzalis, D., Niccolucci, F., Theodoridou, M., Doerr, M.: *LIDO and CRMdig from a 3D Cultural Heritage Documentation Perspective*, Proceedings of VAST 2010, Paris, September 2010, 87-95.

[10] *3D-COFORM: Tools and expertise for 3D collection formation*, http://www.3d-coform.eu.

[11] *Fedora Repository Project: General purpose, open source digital object repository system*, www.fedora-commons.org.

[12] Doerr, M., Tzompanaki, K., Theodoridou, M., Georgis, C., Axaridou, A., Havemann, S.: *A repository for 3D model production and interpretation in culture and beyond,* Proceedings of VAST 2010, Paris, September 2010, 97-104.

[13] Pan, X., Beckmann, Ph., Havemann, S., Tzompanaki, K., Doerr, M., Fellner, D.W.: *A Distributed Object Repository for Cultural Heritage,* Proceedings of VAST 2010, Paris, September 2010, 105-114.

[14] *SOLR - Open Source Enterprise Search Platform From The Apache Lucene Project,* http://lucene.apache.org/solr/.

[15] Lorenzini, M.: *Dati e conoscenza archeologica: il CityGML per il 3D del foro di Uchi Maius in Tunisia,* Graduation Thesis, Università degli Studi di Pisa, 2009.

[16] Felicetti, A., Lorenzini, M., Niccolucci, F.: *Semantic Enrichment of Geographic Data and 3D Models for the Management of Archaeological Features,* Proceedings of VAST 2010, Paris, September 2010, 115-122.

[17] Felicetti, A., Lorenzini, M.: *Open Source and Open Standards for using integrated geographic data on the Web,* Proceedings of VAST 2007, Brighton, November 2007, 63-70.