Information design for cultural documentation

Chryssoula Bekiari[1], Panos Constantopoulos[1,2], Martin Doerr[1]

[1] Institute of Computer Science, FORTH
[2] Athens University of Economics and Business
{bekiari,panos,martin}@ics.forth.gr

Abstract.

Digital cultural repositories contain digital surrogates of cultural objects, objects born digital, or documentation data. The latter comprise data (or 'metadata') about physical or informational objects, as well as about the various related processes, such as acquisition, recording, digitization, research, conservation, exhibition, publication, etc. Interoperability of the different, autonomous repositories is important for the joint utilization of their contents. This, in turn, implies compliance with certain syntactic and semantic representation conditions determined by relevant standards and recommendations. The elaboration of information structures and associated guidelines that can support the development of interoperable digital cultural repositories, in particular of the respective documentation parts, for application on a national scale is the object of a Greek Information Society project we have undertaken. Drawing on the recommendations of various established international and national bodies concerning archaeological, ethnological, museological, archival, geographical, terminological and digital preservation data, and adopting as general frameworks for syntactic and semantic interoperability the standards of W3C and CIDOC CRM (ISO/DIS 21127) respectively, we have defined a set of appropriate information patterns that can meet the requirements of a large variety of cultural material in an integrated manner while maintaining enough flexibility to accommodate further special needs. This paper presents the approach and the main design decisions taken in this large information design project.

Introduction

Holders of "cultural goods" increasingly practice the production of related digital material. This involves digitizing documents and images, digital photography, analog-to-digital conversion of audio or video recordings, or digital transcription of object information recorded in various manners. Adding to these digital surrogates inherently digital cultural products, as well as recordings of cultural information directly in digital form, one gets a grand total of digital material that might be called "digital cultural inventory". By virtue of its function as digital surrogate of physical

objects or machine-processible representation of information objects, this digital material has a value by itself, which is compounded by the capability of uniform access to independent collections of digital material, i.e. the potential for a unified digital space. As digital collections are created independently by autonomous organizations, the emergence of a unified digital space is not automatic nor easy. Besides legal and organizational aspects, it technically requires certain conditions for the interoperability of repositories.

The "Information Society" Operational Programme, currently deployed in Greece, provides, through its specific Action 1.3: *Documentation, exploitation and promotion of Greek culture*, a framework for the generation of a digital cultural inventory. The anticipated results include large digital repositories, as well as significant infrastructures and experiences, both technical and organizational, that should provide the basis for further development, preservation and exploitation of the inventory. Serious challenges arise as this endeavour involves the introduction of information technologies in a large number of organizations most of which made little, if any, use of it so far. The lack of experience and "IT culture" implies increased difficulty in ensuring the quality of the results. The *quality indices* considered include *validity, accuracy* and *completeness* of data, *ease of access*, *interoperability* of the various information repositories, and *preservability* of the inventory. There is an obvious need for referring to a common set of general guidelines for the design and implementation of digitization and documentation projects, and for promoting common practices.

Part of the Operational Programme aims indeed at formulating a set of guidelines for (a) digitization methods and procedures, (b) organization, integration and preservation of information, (c) Web design and educational applications, and (d) intellectual property rights management.

The work reported here concerns the above item (b), specifically the development of a guide for designing and applying information structures for cultural documentation and for supporting the preservation and interoperability of digital information [32]. Cultural documentation comprises a wide spectrum of information concerning the objects themselves, physical or informational, as well as the processes that take place in the course of diverse activities ranging from primary data acquisition to various scientific studies, conservation, exhibition design, publication, etc. These activities may be separately documented and multiple different relevant data sets may exist. We thus contend that the capability of accessing and associating relevant information from disparate sources is an important factor of the usage value of the information. Consequently, we primarily focus on addressing the interoperability issue. This complements previous works that have specified information categories or have developed systems for various application classes (e.g. [7,8,9,10,14,15,16,17,19, 25,27,28,29,30,33]) by providing a common ontological layer from which compatible views are derived. In addition, in the case of documenting moveable objects and site monuments in particular, we have integrated concepts from multiple relevant standards into a new, comprehensive, common XML DTD compatible with the ontology provided by the CIDOC CRM [6]. To the best of our knowledge, an information structure of such genericity has not appeared in international practice before.

Interoperability has a syntactic and a semantic aspect. Syntactic interoperability is achieved by conforming to standards for information encoding and exchange.

Semantic interoperability means the capability of different information systems to communicate information consistent with the intended meaning [6,31]. In practice, semantic interoperability aims at associating and aggregating knowledge dispersed in various carriers and forms. This requires conforming with standards for representing objects, functions and content during documentation, as well as during "productive" uses of digital information, so that common concepts and relations can be automatically identified. Our work goes beyond simply registering common and non-common metadata elements as in the SCHEMAS Project [34], because the compatibility with an ontology yields rich generalizations over more specific metadata elements that can be used for integrated information access. To a certain degree, we have also merged concepts from metadata elements from multiple sources in order to complement functionality similar to the ARCO Project [35].

Our work comprises three parts: normative framework, documentation and interoperability.

A. Normative framework. This provides *recommendations* and *suggestions*, the former being more compelling than the latter. It concerns conformance with standards for ontologies, term thesauri, formatting and data exchange in the Web, museological documentation, archaeological and ethnological documentation, geographical information, archival documentation, encoding and management of multilingual texts, multimedia content description, learning object description, and digital preservation.

B. Documentation. Information structures are specified to be used in registration, description and conservation of objects, digital preservation and publication of digital information. Specifically, a family of digital object record types is provided, that conform with the normative framework along with respective XML DTDs.

C. Interoperability. Guidelines are provided for applying technologies and standards for interoperability, information resource access, and terminology management. In particular, the conceptual reference model of the ICOM/CIDOC, CIDOC CRM, also ISO/DIS 21127, is recommended as the basis for semantic interoperability of cultural documentation systems. The first edition of CIDOC CRM in Greek is also included.

In this paper we give an overview of the approach followed in our project.

Documentation: from objects to data

Apart from digital cultural products themselves, the rest of the digital cultural inventory is the result of documenting cultural objects, be they *physical* or *conceptual* (e.g. a piece of music) which, of course, have physical *carriers* (e.g. a CD). In an abstract view, a documentation process generally records and describes selected objects. Traditional, non-digital forms of documentation produce archives of records, texts, photographs, designs, maps, audio recordings, video recordings, etc. These differ in nature from each other and from the objects they refer to. Digital documentation, on the other hand, produces archives that differ in terms of information content, but share digital kinship among them as well as with digital surrogates and inherently digital objects.

A general model of the digital documentation process comprises four basic branches (see fig. 1):

(1) Initial (often prior) non-digital documentation of non-digital objects.
(2) Documentation of non-digital objects directly in digital form.
(3) Creation of digital surrogates both of non-digital objects and of non-digital documentation.
(4) Digital documentation of digital surrogates and inherently digital objects.

The outcome of this process includes

- documentation data for physical and conceptual objects
- digital objects (surrogates or inherently digital)
- documentation data for digital objects (metadata).

The integration of these digital elements, enabled by fulfilling interoperability conditions, is expected to compound the aggregate value of the digital cultural inventory.

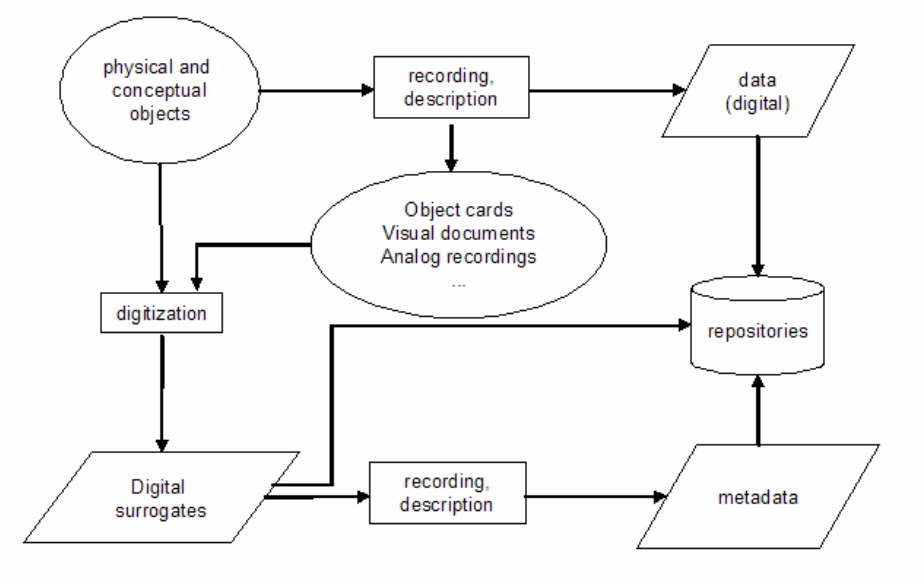

Fig. 1. Documentation process

Object records

General structure of an object record

The information generated by the documentation process is saved in a *record* and can be distinguished into five major classes, or sections, according to the purpose it serves: record identification, object identification, scientific documentation, administration, and references.

Record identification contains metadata concerning the record as a digital object in itself.

Object identification contains the minimum data necessary to identify the object and uniquely refer to it independently from any particular context.

Scientific documentation contains elements that concern

the **description** of the object as it is in our hands, such as various classifications, physical constituency and condition, symbolic content, etc.;

the **history** of the object as reported by witnesses or inferred from traces and evidence, including descriptions of events and activities, such as construction, use, discovery, conservation, etc., in which the object took part; and

associations of the object with other objects (e.g., similarity) and events.

Strictly speaking there is some overlap of the scope of scientific documentation with that of object identification: certain descriptive elements are deemed necessary for identifying the object, in which case they are placed in the identification section.

Administration contains data which pertain to the current handling of the object in a museum or collection, e.g. acquisition, location, exhibition, loan, etc., and which may later be regarded as relevant to the object history or not.

References, finally, contain metadata about sources of documentation and related bibliography.

Nature of documentation data

The elements of an object record generally describe entities (physical or conceptual), events and associations. *Physical entities* include the object being documented and, possibly, others related to it. *Conceptual entities* appear in the context of their relation to the object being documented. *Events* are determined by their kind, persons, organizations and objects involved in specific roles, their limits in place and time, and constituency from other sub-events. An important specialization of events are activities, which are further characterized by actor, purpose and technique. Events are only recorded in the context of their relationship to the object being documented. *Associations* may represent comparisons between objects (e.g. similarity) or cultural context (e.g. joint use of objects, depiction or copy making, witness).

Furthermore, data are distinguished with respect to their temporal validity into *permanent* (unlimited validity) and *volatile* (limited over a specific time interval). Volatile data should normally be tagged with their validity time.

The information contained in an object record can be considered as a set of *logical propositions* that may refer to

- *specific situations* or occurrences, e.g. the pen with which Eleftherios Venizelos signed the Protocol of the Sevres Convention, the necklace worn by Queen Amalia on her wedding, the inscription on Odysseus Androutsos'sword; or
- *categories*, e.g. wedding dress, flag carried in the battlefield, clay pot.

So, a set of propositions may convey

- part of the history of a particular object;
- a frame of hypotheses about part of the history of an object, which refers to categories of events and other entities; or
- categorical knowledge, i.e. knowledge about the kinds of objects and events, not about a particular object.

In the first case propositions are based on specific evidence. In the second, which is probably the most common in practice, the hypotheses expressed by the propositions are based on general knowledge and its connection to specific evidence represented by the documented item. The third case represents a useful by-product of the documentation process, in which the documented item plays an exemplary role.

Information patterns

Evidently certain parts of an object record may contain information of the same nature, e.g. time, place, object composition, event, etc., even though they refer to different subjects, and so may happen with different records even concerning objects of different kinds. By defining certain specializable types of information units, which we shall call *information patterns*, we can reduce the problem of designing an object record to one of designing a set of information patterns and a general, flexible record structure. As the information patterns are much fewer than the record fields, the design and the conformance with relevant standards are much better controlled. We thus effectively obtain a family of records, conformant at the information pattern level, rather than a single record, which facilitates both addressing different needs and adopting systematic data entry procedures. Examples of information patterns are given in table 1.

Data entry

Data entry into the object record naturally follows the sequence of data generation, which in turn follows the sequence of object handling acts, e.g. as specified in Spectrum [19]. However, the prescribed data may unavailable, incomplete, uncertain, or they simply require time for editing before official entry, so in practice a certain autonomy of the data entry process is desirable.

Each record *field* must obey a respective *data entry rule*, which determines how compelling the entry of a value in that field is. We distinguish three cases: *necessary* (value omission disallowed), *compulsory* (value must be entered if it exists and is known) and *optional*. The data entry rule facilitates, but does not impose, an organization of the documentation work favouring breadth of coverage over depth.

Such an organization policy aims at ensuring the documentation of *all* objects in a collection at a minimum acceptable level of detail in the least time possible. Of course, the level of detail, or completeness, remains a quality index of the record.

Date from until	Object composition number of parts part name kind code or cardinal number
Chronology within throughout cultural period social time justification	Place name code cadastral number kind geopolitical hierarchy address coordinates values reference point precision of measurement geodesic coordinate system link to design
Dating chronology time measurement value method laboratory	Event name kind chronology place description persons involved organizations involved objects involved comprises events
Person name biographical data communication data role / capacity / social group	Organization Title legal address communication data department role / capacity / social group

Table 1. Examples of information patterns

The value multiplicity of a field must be specified. Our default choice is to allow assigning *multiple values*, in order to support the recording of multiple opinions.

Correspondingly, whenever it makes sense, one should be allowed to record the *source* and *degree of certainty* of the value. These are metadata that can be assigned to a field whenever required. Single value assignment is only imposed in a few cases, such as certain id codes.

Data are often fuzzy or uncertain, in which case a "least binding" data entry policy might appeal as safe, e.g. date 'unknown' or 'before 1900 AD'. In terms of search effectiveness though, the opposite policy, i.e. entering the most precise values within the limits of the documenter's knowledge, would have been preferable. E.g., a personal computer of unknown production date could be safely dated 'after 1980 AD'.

Object record types

Although the general structure of the object record is quite flexible, it would be in practice desirable to have a controlled variety of records thus supporting uniform documentation practices. To this end we can define a set of record types on the basis of two criteria: the intended use of the records and the type of objects recorded. These criteria are listed in table 2.

The actual object record types we consider in our work along with the corresponding normative framework (recommendations and suggestions) are shown in table 3.

Intended use	Object type
registration	site monuments
description	moveable objects
administration	text documents
conservation	audio
digital preservation	pictures (still, moving)
publication of digital material	digital surrogates
	inherently digital objects

Table 2. Object record type generation criteria

Interoperability

It has already been noted that the digital cultural inventory should on one hand remain available and safe despite future failures of equipment or technological changes (*preservation objective*) and, on the other, support integrated access and use (*integration objective*). From an economic point of view attaining the first objective depends on the costs of recovery, re-creation and permanent loss of information, while attaining the second depends on the costs of access and re-use of distributed and

Record type	Recommendations	Suggestions
registration and description of moveable objects	CIDOC Conceptual Reference Model (CRM) [6] Object-ID (Getty) [10] User Manual for the Electronic Documentation of Monuments (National Monuments Record – Ministry of Culture) [14]	International guidelines for museum object information (CIDOC) [7] Metadata Standards for Museum Cataloguing (CHIN) [16] International Core Data Standard for Ethnology/Ethnography (CIDOC) [8]
registration and description of site monuments	CIDOC Conceptual Reference Model (CRM) [6] Object-ID (Getty) [10] User Manual for the Electronic Documentation of Monuments (National Monuments Record – Ministry of Culture) [14] OpenGIS Reference Model (Open GIS Consortium) [18]	Draft International Core Data Standard for Archaeological Sites and Monuments (CIDOC) [9] MIDAS (English Heritage) [17]
administration of museum collections		SPECTRUM (Museum Documentation Association) [19]
administration of site monuments		Established practice of Ministry of Culture – National Monuments record System [15,27,28]
documentation of digitized images and video	DIG35 (Digital Imaging Group) [12] MPEG7 (ISO) [11]	
documentation of digitized text	TEI (Text Encoding Initiative Consortium) [13]	
archive documentation	EAD (Encoded Archival Description – Library of Congress) [20]	
digital preservation	Digital Preservation Coalition [21] Online Computer Library Center/Research Library Group (OCLC/RLG) [22]	
publication of digital material for educational use	LOM (IEEE) [23] SCORM (US DoD Advanced Distributed Learning Initiative) [24]	
multilingual publications		TMX (Localisation Industry Standards Association) [36]

Table 3. Object record types and corresponding normative framework

heterogeneous information. From a technical point of view the decisive factors are portability across platforms, data and system interoperability among repositories, and Web access. A necessary condition for these is conformance with certain standards concerning *syntactic* and *semantic interoperability*.

Syntactic interoperability is achieved by using common means for data representation; more precisely, external data representation, while individual repositories maintain the freedom to use different encodings for internal representation and processing. It appears that XML [1] is a good choice for this purpose.

Semantic interoperability is achieved by employing a common conceptual model in formulating semantic descriptions of objects and digital resources to support uniform access to them. The ICOM/CIDOC Conceptual Reference Model (CIDOC CRM, also ISO/CD 21127) [6] emerges as the most suitable such model for the cultural field today. The CIDOC CRM is an ontology for the cultural domain, which formally describes the concepts and relations involved in cultural documentation. It provides a common base for the interpretation of various forms of documentation, but does *not* dictate the documentation elements. The ontology can be used both as a framework for designing information structures for documentation systems and as a communication medium, at the semantic level, between heterogeneous systems. Thus it plays an indispensable role in building an integrated digital cultural inventory.

purpose	**recommendations**	**suggestions**
syntactic interoperability	XML (W3C) [1]	
semantic interoperability	ICOM/CIDOC Conceptual Reference Model (ISO/CD 21127) [6]	
terminology management	ISO2788 [3] ISO5964 [4]	SKOS [5] RDF/ RDFS (W3C) [2]

Table 4. Interoperability normative framework

While CIDOC CRM is an ontology for cultural documentation in general, at the level of specific sub-domains it can be extended to produce specific application ontologies. These extensions chiefly involve appropriate specializations of the standard concepts and introducing relevant terminology. The latter is an important, evolving constituent of every documentation system. In order to support consistency in object documentation and information retrieval, it is recommended that terminology is organized as a thesaurus according to the ISO2788 [3] and ISO5964 [4] standards. Ontologies and term thesauri usually form large semantic networks, the representation of which for logical processing in the Web can be done in RDF / RDFS

[2], preferably in the form proposed by SKOS [5]. Table 4 summarizes the normative framework for interoperability.

Conclusion

We have presented our approach to developing and employing information structures for cultural documentation and for the integration and preservation of a digital cultural inventory. A basic condition for the creation of a viable and useful digital cultural inventory is employing appropriate and compatible information structures. Of course, this is complementary to establishing good practices and procedures and adopting policies for sustained funding and organizational support.

We have followed a dual strategy: on one side we propose specific standard (meta)data structures for specific application areas, on the other all those structures are related to the common core ontology of the CIDOC CRM, which provides semantic interoperability in the long term. For the description of moveable objects and site monuments in particular, we have integrated concepts from multiple, relevant standards into a new, comprehensive, common XML DTD compatible with the CIDOC CRM, resulting in an information structure of unprecedented genericity. In other cases we simply adopt existing international standards.

We thus hope to have set the foundations for long-term validity and exploitation of the information to be captured in documentation projects following these guidelines. We have deliberately ignored finding aids such as Dublin Core [26], because they do not provide structuring of the information itself, while they can be added at schema level at any time, provided more expressive information structures exist. As information from cultural organisations becomes increasingly available, next generations of integrated information access systems will find a rich ground of well defined (meta)data structure semantics and uniform information syntax.

References

[1] Extensible Markup Language (XML) 1.0 (Third Edition), W3C Recommendation 04, February 2004, http://www.w3.org/TR/REC-xml/, http://www.w3schools.com/xml/default.asp

[2] Resource Description Framework (RDF), W3C , http://www.w3.org/RDF/

[3] ISO2788: Monolingual thesauri standard, http://www.collectionscanada.ca/iso/tc46sc9/standard/2788e.htm

[4] ISO5964: Multilingual thesauri standard, http://www.collectionscanada.ca/iso/tc46sc9/standard/2788e.htm

[5] SKOS-Core 1.0 Guide, An RDF Schema for thesauri and related knowledge organisation systems, http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/

[6] ISO/DIS 21127, The ICOM/CIDOC Conceptual Reference Model. Current official edition: Nick Crofts, Ifigenia Dionissiadou, Martin Doerr, Matthew Stiff

(editors), Definition of the CIDOC object-oriented Conceptual Reference Model, April 2004 (version 4.0). http://cidoc.ics.forth.gr/official_release_cidoc.html

[7] International guidelines for museum object information : the CIDOC information categories, ICOM/CIDOC, Paris, 1995. ISBN 92- 9012-124-6.

[8] International Core Data Standard for Ethnology/Ethnography, ICOM/CIDOC, Paris, 1996, ISBN 960-214-012-7, http://cidoc.icom.museum/ethst0.htm

[9] Draft International Core Data Standard for Archaeological Sites and Monuments, ICOM/CIDOC, Paris, 1995, ISBN 92-9012-125-4, http://cidoc.icom.museum/arch0.htm

[10] Object-ID, http://www.object-id.com/about.html

[11] MPEG7 multimedia content description, ISO/IEC TR 15938-8:2002, http://archive.dstc.edu.au/mpeg7-ddl/

[12] DIG35 Metadata Specification Version 1.1, Digital Imaging Group, http://www.i3a.org/i_dig35.html

[13] Text Encoding Initiative (TEI), http://www.tei-c.org/

[14] User Manual for the Electronic Documentation of Monuments, Ministry of Culture, National Monuments Record Directorate (in Greek), http://www.ics.forth.gr/CULTUREstandards/phges.htm

[15] POLEMON National Monuments Record System, Ministry of Culture, National Monuments Record Directorate (in Greek), http://www.ics.forth.gr/CULTUREstandards/phges.htm

[16] Metadata Standards for Museum Cataloguing http://www.chin.gc.ca/English/Standards/metadata_description.html

[17] MIDAS: A manual and Data Standard for Monument Inventories., http://www.english-heritage.org.uk/Filestore/nmr/standards/Midas3rdReprint.pdf

[18] OpenGIS Reference Model (Kurt Buehler, ed.), Open GIS Consortium Inc., Ref. No. OGC 03-040 Ver. 0.1.2, http://www.opengeospatial.org/specs/?page=orm

[19] SPECTRUM: The UK Museum Documentation Standard, 2nd Edition, Museum Documentation Association, Cambridge, United Kingdom, 1998, http://www.mda.org.uk/spectrum.htm

[20] Encoded Archival Description (EAD) Version 2002, Society of American Archivists and the Library of Congress., http://www.loc.gov/ead/

[21] Digital Preservation Coalition Handbook , http://www.dpconline.org/graphics/handbook/

[22] A Metadata Framework to Support the Preservation of Digital Objects, OCLC/RLG Working Group on Preservation metadata., http://www.oclc.org/research/projects/pmwg/pm_framework.pdf

[23] IEEE Final 1484.12.1-2002 LOM Draft Standard for Learning Object Metadata, http://ltsc.ieee.org/wg12/20020612-Final-LOM-Draft.html

[24] Sharable Content Object Reference Model (SCORM), http://www.adlnet.org/index.cfm?fuseaction=SCORMDown

[25] AHDS, http://ahds.ac.uk/collections/index.htm

[26] Dublin Core http://dublincore.org/documents/dcmi-terms/

[27] C. Bekiari, D. Calomirakis, P. Constantopoulos, P. Pantos, "POLEMON: A project to computerize the monuments records of the Greek Ministry of Culture", in Proc. Computer Applications and Quantitative Methods in Archaeology Conference, CAA 1996, 1996.

[28] C. Bekiari, D. Gritzapi, D. Kalomoirakis, POLEMON: A Federated Database Management System for the Documentation, Management and Promotion of Cultural Heritage, Proc. CAA 1998, Barcelona, 1998.

[29] P. Constantopoulos, D. Chronaki, D. Agelakis, E. Fritzalas, C. Bekiari, MAISTOR: A Structural Documentation System for Buildings, 2000, Proc. CAA 2000, Ljubljana, Slovenia, 2000.

[30] P. Constantopoulos, M. Doerr, M. Theodoridou and M. Tzobanakis, "Historical documents as monuments and as sources", in Proc. Computer Applications and Quantitative Methods in Archaeology Conference CAA 2002, Heraklion, 2002.

[31] M. Doerr , Semantic interoperability : Theoretical Considerations, TR 345, Institute of Computer Science, FORTH, October 2004.

[32] Standards for cultural documentation and technologies for supporting and integrating the digital cultural inventory and for system interoperability, Institute of Computer Science – FORTH, Information Society Operational Programme, Action 1.3, Project No. 92402.

[33] METS Official Website, http://www.loc.gov/standards/mets/.

[34] The European IST Project SCHEMAS http://www.schemas-forum.org/project-info/objectives.htm

[35] M. Patel, M. White, N. Mourkoussis, J. Darcy, P. Petridis, F. Liarokapis, P. Lister, K. Walczak, R. Wojciechowski, W. Cellary, J. Chmielewski, M. Stawniak, W. Wiza, J. Stevenson, J. Manley, F. Giorgini, P. Sayd, F. Gaspard, "ARCO -An Architecture for Digitization, Management and Presentation of Virtual Exhibitions" . In: Proceedings Computer Graphics International, Hersonissos, Crete, Greece, June 16-19 2004.

[36] TMX http://www.lisa.org/tmx/tmx.htm