Specifications and interoperability features for open digital content (2nd DRAFT)

Project:	"National Information System of Research and Technology/Social Networks"
Project Id (M.I.S.):	296115
Project:	""A Platform for the Deposit, Management and Delivery of Open Metadata and Digital Content"
Project Id (M.I.S.):	327378

Initial Version:	2009, May
1st Update:	2011, November
2nd Update:	2012, March
3rd Update:	2013, March
Translation to English of the 3rd Update (Draft):	2014, June

Authors:	Dr. Panagiotis Stathopoulos Dr. Nikos Houssos
Translation in English:	Dr. Haris Georgiadis

48 Vassileos Constantinou Av, GR-11635, Athens

Tel.: 210 7273900-02 • Fax: 210 7246824

e-mail: ekt@ekt.gr • http://www.ekt.gr

Permanent electronic address for reference: http://hdl.handle.net/10442/8887

Table of Contents

Introduction

In this study, the main specifications and the features for the interoperability of open digital content are examined, based on the experience of EKT and the international landscape and standards that dominate the respective fields. These specifications cover the implementation of a series of projects being developed by EKT, such as the project "National Information System for Research & Technology/Social Networks-User Generated Content" and the project "A Platform for the Deposit, Management and Delivery of Open Metadata and Digital Content", which leverage resources from the Operational Programme "Digital Convergence" and the Regional Operational Programmes (NSRF 2007-2013).

The initial version of this study as well as its 1st and 2nd updates took place in the context of the "Project "National Information System for Research & Technology/Social Networks-User Generated Content", while the 3rd update was implemented in the context of the Project "A Platform for the Deposit, Management and Delivery of Open Metadata and Digital Content".

The National Documentation Center (EKT) constitutes a scientific e-infrastructure of national use, being part of the National Hellenic Research Foundation. The term "scientific e-infrastructure of national use" reflects the special nature of EKT as a distinct national infrastructure with institutional role in the collection, organization and dissemination of scientific and technological information, within and outside the country. With continuous presence in the national scientific community since 1980, EKT leverages Information and Communications Technology in combination with modern operational methods, in order to develop innovative projects, oriented to the reinforcement of access to digital content. For this purpose, it cooperates with important carriers of authoritative content, such as libraries, repositories, museums and research centers.

Throughout these projects, it has been concluded that it is priority of strategic importance that appropriate tools and services, for the provisioning of digital content and scientific data, are based on both emerging and widespread technologies with wide community support.

Towards this direction, EKT implements the project "National Information System of Research and Technology (NISRT)" which develops a **national research e-infrastructure for the organization and disposal of digital information and content in the domains of science, technology and culture.** The implementation of the first phase of NISRT started at 1996, and was based on the unique expertise and experience of EKT, the existing and emerging international standards and the innovative Information and Communication Technologies.

The e-infrastructure of EKT incorporates state-of-the-art technologies, leverages long-term cooperation with organizations specialized in the areas of research, education and culture, applies international standards across all levels (data organization, content preservation, rendering of services, interoperability across systems) and implements the policy of Open Access to research results. It fulfills well-known user needs, supports the transmission of knowledge and joins the international network of similar infrastructures that are being currently formed.

An additional development tool of EKT is the project "A Platform for the Deposit, Management and Delivery of Open Metadata and Digital Content" that is implemented as part of the Operational Programme "Digital Convergence" (NSRF 2007-2013), co-funded by Greece and the European Union.

The objective of this project is **the development and provisioning of modern digital services to organisations that produce digital content** (libraries, museums, archives and cultural organisations in general). This services aim in the

reinforcement of the digital presence of these organisations, through the utilization of integrated solutions that enable the **documentation, deposition, data protection, organization and disposal of their digital content**.

The services are based on cutting-edge technologies and models, such as the SaaS model (Software as a Service) and Cloud Computing models, and are provided by EKT **without imposing additional costs for the end organisations**, over the Internet, without the need for local setup and **fully tailored to the needs of each organisation**. To ensure their efficient utilization, these services are accompanied by training and supporting services for carriers to whom they are addressed.

This intervention aims in the development of a powerful network of organisations that together will shape the strategy and the specifications for the increase of the authoritative digital content of the country and the encouragement of its reuse.

1. Objectives

The projects of digitization and provisioning of content of the previous programmatic periods have made important contribution to the digitization of the Greek scientific and cultural assets. However, based on previous experience, a series of issues have emerged during the implementation that:

- restrict the dynamics of the generated content, reduce the value of the investment and the potentials for broad reuse and organization of that content
- do not allow the wide adoption of the fast technological developments that occur in the fields of open public data and of open digital content, in accordance with the latest EC requirements for the e-Government.
- do not take into account the new innovative approaches to the provisioning of applications and integrated services, such as Cloud infrastructures, in general, and infrastructures for providing Software as a Service, in particular, for the delivery of applications in the form of services, as defined in the Digital Agenda for Europe 2020 of the European Commission.

At the functional level, the problems that have been identified with regard to the technological aspects are:

- a wide multitude of different systems used for the management and provisioning of digital content
- the increasing maintenance and operation costs in the production phase per individual installation
- the limited potential for reuse of content, expertise and systems
- the inability of custom made software to effectively integrate with international repositories and/or search engines
- the reduced availability and security level of each individual software installation
- the lack of mechanisms that would ensure the long-term preservation, utilization and security of the content.

Part of these problems arise due to the particularities of the Greek environment, which, while characterized by a significant number of cultural and memory organisations that have valuable content that could satisfy a wide range of uses, in most cases, these organisations do not have the necessary technologically critical - size that would allow for he qualitative and sustainable provisioning of complex digital content.

During the previous programmatic period, the project contractors had been given a set of good practices and guidelines in the form of studies for helping them to handle various technological issues. However, the experience captured from the implementation of these projects have shown that technological advices and generic guidelines alone were not sufficient for dealing with the abovementioned problems, especially those regarding interoperability, potential for content reuse and improved user experience.

So, apart from just updating individual components in the existing studies and enhancing them with new components focusing on the issues that have arisen in both technological and operational level, it is necessary to establish a framework and specifications that would ensure compliance of the open digital content systems, combined with the necessary interoperability checkpoints, in order to avoid failures during the implementation of the projects, to reduce their costs and

to make provision for the long-term viability and preservation of the produced content.

For that purpose, the current study specifies the main functional specifications and good practices that the digital repository system that will be used for the organization, deposition, and provisioning of open digital content must follow at a minimum. For critical interoperability specifications, compliance control mechanisms supported by automated tools for compliance checking would give additional value.

The purpose of this document is to define a clear and realistic framework for achieving interoperability across all the systems and content that will result from projects whose main target is the disposal of open digital content. This framework allows for further expansion in order to cover the needs of additional specialized content categories and to incorporate the ongoing international developments in the domain of technology, standardization and infrastructure.

The framework and the corresponding specifications are based on the long national and international experience of the National Documentation Center and the implementation of the national e-infrastructure of open digital content "National Information System of Research and Technology" (http://www.epset.gr) and recognizes as a base the international practices within European Countries (UK, Netherlands, France, Sweden, etc.) and EU (e-content programmes, ICT PSP, e-infrastructure, etc.), successful case studies of large-scale individual international repositories and digital content systems (Open Library of Internet Archive, Google Books and Art Project, etc.) as well as developments, opportunities and capabilities provided by SaaS (Software as a Service) infrastructure for digital content.

The National Documentation Centre has an institutional role in the scientific information and a long-term presence in the disposal of open digital content and services in the academic, research and scientific community.

2. Repositories of open digital content

For the purposes of this study, we use the term "repository" for software systems that are used for the deposition, organization and provisioning of organized content that aim in the collection, disposal and long-term preservation of born digital or digitised content. Repository systems are distinguished from the simple content management systems (CMS) mainly due to the critical importance of a range of features related to interoperability, reuse and preservation of data and rich metadata of digital content.

All content repositories are structured around the digital object, which consists of the body of the main object (full text, image, etc.), the descriptive metadata and their semantic interpretation. This structured information is supported by a series of new but already matured open interfaces, metadata schemata and semantic descriptions for the exchange, searching, aggregation and interconnection of the content in an organized fashion, which have all emerged in recent years in the international scene.

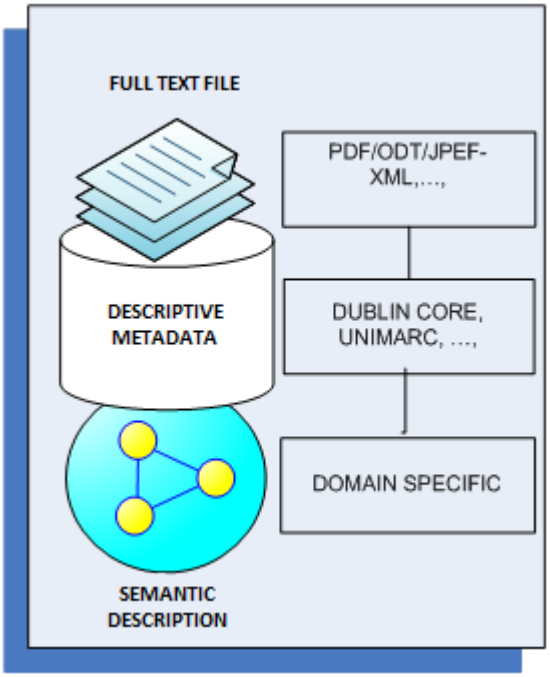

Figure 1 Simplified form of digital object and relevant standards

Based on the type of the content they handle, repositories usually fall into the following categories:

- Repositories / collections of digital cultural and/or historical content (e.g. pictures of works of art, archaeological monuments, manuscripts)
- Repositories of scientific content (e.g. scientific publications such as articles, books, conference proceedings)
- Archives of various types and forms (e.g. personal files of personalities, audiovisual files from broadcast organizations, periodicals)

The forms of content, organization approaches and special standards may vary depending on the nature the content, but all share the same basic principles of interoperability and the requirements of presentation and disposal.

An important role in the spread of usability, search and organization of the content is played by metadata aggregators, which collect metadata from individual repositories using open interfaces. These aggregators can also operate as "registers" of the open digital content, registering its origin, the location of permanent reference and disposal, aggregative statistics etc. From now on, the terms "metadata aggregator" and "register" of open digital content will be used interchangeably.

Next, the basic specification and interoperability requirements at every level of acquisition, process and disposal of digitalized or digital content are presented.

3. Specifications for digitalization, organization and disposal of open digital content

3.1. Digitisation

The procedures, standards and methods of digitisation for the material of interest have been addressed, in significant detail, in the respective part of the study entitled "Guidelines and good practices for digitization and long-term preservation of cultural content" [1], and incorporate material from corresponding widespread international studies and sources [4][5][6][7]. The minimum quality requirements, mentioned in [1], are of great importance and are reproduced below:

Original Object	Minimum resolution	Color depth
Photocopied material (in black and white)	200-300 dpi	8 bit grey
Printed material (in black and white)	400 dpi or 4000 pixels at the largest dimension	8 bit grey
Printed material (in color)	400 dpi or 4000 pixels at the largest dimension	24 bit
Maps and graphics (in black and white)	300 dpi or 4000 pixels at the largest dimension	8 bit grey
Maps and graphics (in color)	300 dpi or 4000 pixels at the largest dimension	24 bit
Photos (in black and white)	600 dpi or 5000 pixels at the largest dimension	8 bit grey
Photos (in color)	600 dpi or 5000 pixels at the largest dimension	24 bit
Works of art (in black and white)	600 dpi or 5000 pixels at the largest dimension	8 bit grey
Works of art, fabrics (in color)	600 dpi or 5000 pixels at the largest dimension	24 bit
35mm slides, negative etc. art (in black and white)	2400 dpi	8 bit grey
35mm slides, negative κλπ (in color)	2400 dpi	24 bit
6cm X 6cm slides (in black and white)	2000 dpi	8 bit grey
6cm X 6cm slides (in color)	2000 dpi	24 bit
Slides or slabs of glass (in black and white)	600 dpi	8 bit grey

Table 1. Minimum requirements for digitalization of various objects, according to [1]

The study also contains a series of additional technical requirements that affect the quality of the final produced material and the user experience (e.g. correction of digital images, alignment, crop, etc.), which, despite being quite obvious and straightforward, are not always employed. In overall, the specifications presented

in [1] are sufficient, although their update should be considered e.g. by adding the possibility for widespread storage of images in lossless JPEG 2000 format.

However, in addition to the aforementioned specifications, it is required **that a system of optical character recognition (OCR) will be used in the digitalization phase for printed material**, in order to make it possible for the produced digital object to be indexed, and therefore, subject to key word searching by search engines in the web, the repositories or the aggregators (full text search).

It is recommended that at least "uncorrected" OCR transcription will be used (the original text produced by the OCR software without any manual corrections) for texts in modern Greek or other widely spoken languages, as this is the most optimal approach that combines sufficient results in low cost. This requirement is critical so as, after the digitalization process, the digital material to be searchable with respect to its full text in the repository, the Internet or the open digital content register, providing that apart from the metadata, data of optical character recognition derived from the OCR process will be also available.

For the processing of the digitalized files by OCR software for all open digital content projects, the following options are available:

A. **The OCR process can be part of the requirements for the contractors that will implement the digitalization projects.** This constitutes the most appropriate solution, since the extra financial burden is proportionally small compared to the total cost (according to draft empirical estimations it ranges from 0.01 up to 0.02€/page). The outputs of the OCR process per digitalized object should include:

   - The plain text in a separated file of type .txt encoded in UTF-8. The text file must include new line and page break special characters.
   - The text in "Image PDF with hidden text" format. This is a separated PDF file that embeds both scanned images and OCR text data allowing for searching and text selection while retaining the look of the original page.
   - Detailed OCR generated text in ABBYY XML or hOCR format that includes the position of each character / world allowing the presentation of the text by browsers or e-book readers with search and hit highlighting capabilities (alternative of opening the file as PDF).

B. If the final beneficiary carries out the digitalization process internally without resorting to external contractors, and for existing material, it is recommended that **the core OCR infrastructure for massive text optical recognition of the National Documentation Center is utilized**.

Γ. An alternative solution is that the individual final beneficiary organisations undertake the entire OCR processing burden. This solution is only viable for organizations for which digitalization constitutes a continuous and ongoing business process.

The aforementioned requirements are summarized in the following table (Table 2) in a form suitable for inclusion in compliance tables.

A/A	Requirement	Description
1.	**Optical Character Recognition (OCR) support for full text indexing and searching**	Uncorrected OCR output for the entire printed content to be digitized.
2.	**OCR output format**	Delivery of text and XML files in UTF-8 encoding. The plain text must be in a separated UTF-8-encoded file and must include new line and page break special characters.
3.	**OCR output format**	Delivery of a detailed file on the standard format ABBYY XML or on the open hOCR format, which should include positions for each character/word, allow the presentation by browser applications and support searching with hit highlighting.
4.	**Delivery of final file**	Delivery of digitized files in the PDF format "Image PDF with hidden text" that incorporates text allowing searching and text selection.

Table 2. Additional specifications for digitalization

3.2. Metadata and Interoperability

The concepts of interoperability and open data are relevant to interoperability at the metadata level and are defined in the recent Public Online Consultation for the general principles and priorities of the Operational Programme Planning "Digital Convergence" [8]:

- **Interoperability** that is ensured by interfaces that make it possible for different organizations and applications to cooperate with each other (open APIs and Web Services, full adoption of the Web2.0 approach). The aim is to create applications that provision their data in a form that is utilizable by all interested parties, whether citizens or organizations, providing the possibility of the exploitation of these data by external applications, without the need of complex heterogeneous systems integration solutions.
- **Open data** freely accessible by all services, companies and citizens. One of the goals of the Programme "Digital Convergence" is the dissemination of information, mainly derived from the cultural/intellectual wealth of the country, that is in or being converted to digital form. Data derived from the joining projects must be **integrated, properly distributed, accessible and in a format suitable for computer processing**.

Particularly in the context of repositories / digital libraries, interoperability at the metadata level can be considered to have the following three dimensions [9]:

- **Interoperability at the level of repository and open content aggregators.** It makes it possible for a repository to dispose its metadata to third party applications and systems. The repository provides its metadata online through an application programming interface that conforms to a specific protocol. The disposal can be done either at a harvesting level, where practical all (or predefined subsets of) the available metadata are downloaded for reuse to other applications, or at the level of meta-search, where only metadata records that match specific search criteria are retrieved. The interoperability at the systems level includes the compliance of the exchanging metadata with a specific encoding (e.g. Unicode).

- **Interoperability at the level of syntax and structure.** It makes it possible for different systems to "read" correctly the data they exchange. Ideally, this requires the following:
  - Use of common language for the encoding of the metadata at the syntax level (e.g. XML).
  - Use of common metadata schemata (e.g. Dublin Core, MODS, CDWA, EAD, etc.)
  - Encoding of the data values according to a common standard. For instance, dates should be serialized using the same format across all systems.
- **Interoperability at the semantic level.** It makes it possible for different systems to "understand" correctly the meaning of the data they exchange. For full interoperability at the semantic level, each metadata field must have a declared and clear meaning. For example, for some work of art, what is the meaning of a date field in its metadata record? (e.g. is it the date of creation, the date of first public exposure, the date of the initial acquisition by the hosting museum or the date when it was registered in the digital library?) The ideal is to declare the meaning of each element using a suitable knowledge representation language such as RDF or OWL and the use of a suitable specialized standard, such as EDM (Europeana Data Model), CIDOC-CRM, LIDO, CERIF or equivalent.

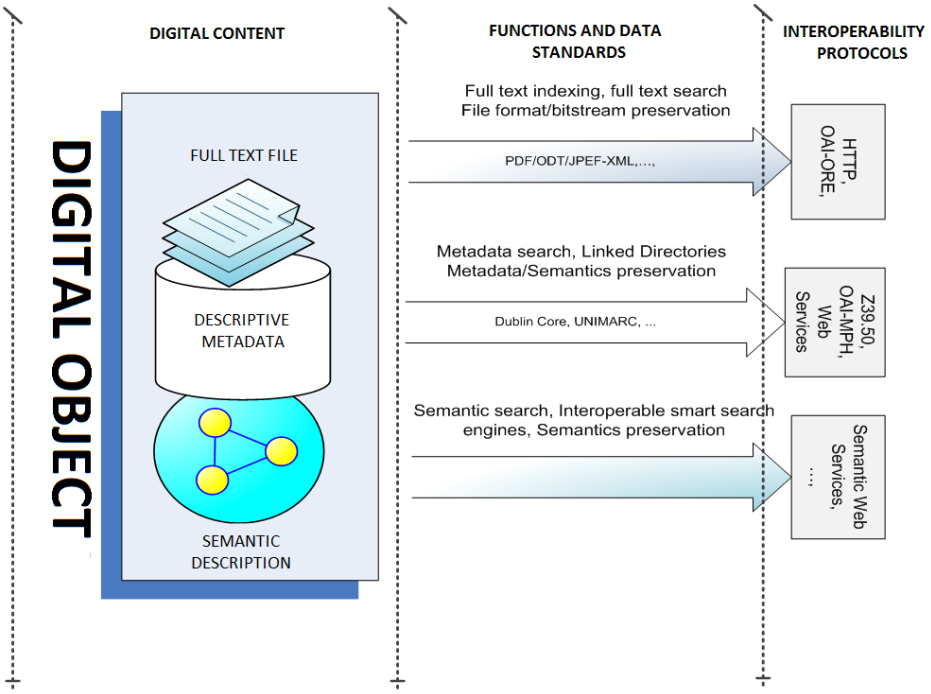

Figure 2 Digital Content, Interoperability and its preservation

Additionally, an important requirement for interoperability at the semantic level is the use of controlled vocabularies, where the value of a metadata field derives from a predefined list of values (e.g. vocabulary, list of standard terms, thesaurus). Common examples are fields such as thematic category, geographic location, chronological period, language and, also, fields referencing real entities, such as individuals and organizations. Ideally, a controlled metadata field should reference only the unique identifier of the actual value that is hosted in an established control vocabulary/thesaurus. In cases where no suitable vocabulary/thesaurus is publically available for a metadata field, the carrier that carries out the documentation is free to create its own vocabulary.

Achieving optimal interoperability at all levels is an extremely difficult task, especially when dealing with a large, unavoidably heterogeneous, open digital content. A holistic and realistic approach is recommended that will define different levels of required features which will support the respective levels of desired services adjusted to the type, the scale and the unique characteristics of the individual system categories. For example, the general rules that apply to all systems will be more relaxed than those applying to a particular bounded class of systems (e.g. digital libraries, museums, archives), for which compatibility with more detailed and specialized metadata formats is required.

The recommended interoperability levels are described thoroughly in Table 3. The highlights are summarized as follows:

- **Fist level:** It ensures the availability of metadata as Open Content and the access to the content with persistent identifiers. It allows the online browsing and retrieval of the content of each repository at any time (e.g. information about the total number of records and digitized objects). It allows retrieval of metadata related to the intellectual property issues and content disposal rules (the corresponding field is under the process of formulation and standardization, by the international community as well, so there should be provision for the gradual compliance with the standards that will emerge and dominate in the years to come). It ensures interoperability at the repository level, enhanced with some basic interoperability features from the levels of syntax and structure interoperability.
- **Second level:** It ensures the online delivery of a broader set of metadata as Public Open Data at a level suitable for the development of third party applications. It supports basic services for common exploration and visual representation of the content by third party applications. It ensures a basic level of compatibility with the semantic web. It corresponds to the interoperability ensuring at the level of syntax and structure.
- **Third level:** It ensures the online delivery of detailed metadata as Public Open Data at a level that supports the development of a variety of third party value-added services. It allows the development of advanced services including central search, exploration and visualization of content. It ensures interoperability with respect to controlled vocabularies. It ensures a sufficient level of semantic compatibility. It corresponds to the semantic interoperability ensuring.

Level	Description	Supported Services
1	Delivery of metadata through web services for harvesting of at least the 15 fields of the basic Duplin Core schema (http://purl.org/dc/elements/1.1), where applicable. Use of unique identifiers per metadata record. Allows indexing by the main general purpose search engines. Dispose of information regarding the state of the content with respect to intellectual property issues.	1. Retrieval of basic metadata for use in external third party applications. 2. Capability to generate reports for the content 3. Capability to perform integrated search on basic fields.

Level	Description	Supported Services
2	Delivery of metadata through web services for harvesting and meta-searching. Disposal of metadata in the Europeana Semantic Elements format. Identification of specific sets for selective harvesting. Encoding of field values according to international standards.	1. Capability to retrieve all metadata for use in external applications. 2. Capability to generate reports for all content 3. Capability to perform integrated search on basic fields. 4. Support for external navigation and visualization services.
3	Support of an international metadata schema suitable for accurate semantic representation. Metadata enhancement with references to established vocabularies of wide use (e.g. thesaurus, thematic categories and spatial data). Disposal of metadata as Linked Data.	1. Retrieval of all metadata for use in external applications. 2. Capability to generate reports for all content over different systems and providers 3. Capability to perform integrated search on basic fields. 4. Advanced services for exterrnal navigation and visualization. 5. Interconnection to other metadata sets (linked data).

Table 3. Interoperability levels for repositories.

This approach is consistent with the recent international experience in aggregating and delivery of digital content in very large scale, including the European portals Europeana[10] and DRIVER[11]. In the these cases, the model that was adopted includes first the identification of a minimum set of requirements applicable to all repositories and then, wherever classification of individual repositories is possible, the identification of more accurate and suitable specifications and standards (e.g. the metadata standard LIDO which is specialized for museums and cultural heritage content[13][14][15]). It also includes the mapping of those accurate specific metadata standards (used per content category) to the Europeana Data Model, provided that the content falls into the suitable category (e.g. cultural content).

Finally, note that the proposed standards, especially those regarding specific metadata fields, regard only interoperability and do not impose any restrictions on the standards that may be chosen for the initial content documentation. This approach is consistent with the principles of shareable metadata, which state clearly that the metadata an organization maintains for its digital content can be very different from the metadata that the organization exposes through public web services [12].

Some indicative interoperability requirements are presented below in the form of a compliance matrix categorized per interoperability level.

Interoperability specifications for repositories

A/A	Requirement	Specification
	Interoperability at systems level	For the encoding of metadata the UTF-8 standard must be used.
2.	Interoperability at syntax and structure level	Metadata of every record are available for harvesting in the format Unqualified Dublin Core (ISO 15836:2003).
3.	Interoperability at systems level	It is required that a permanently available and accessible harvesting service must function conforming to the protocol OAI-PMH, version 2.0 with support to all the verbs of the protocol.
4.	Interoperability at systems level	It is required that at least of the following disposal / meta-searching protocols will be supported: SRU/SRW, Z39.50. This function should be publically available without access restrictions.
5.	Interoperability at systems level	As a minimum, at least the metadata must be indexed by well-known web search engines (Google, Bing, Yahoo).
6.	Interoperability at syntax and structure level	Metadata of every record must be available for harvesting in the Dublin Core format (ISO 15836:2009). All the fifteen basic fields of the Duplin Core (dc) namespace (http://purl.org/dc/elements/1.1) must be considered as mandatory for every record, except for cases where they are not applicable.
7.	Interoperability at syntax and structure level. Interoperability at semantic level.	Metadata of every record must be permanently available for harvesting in the Europeana Semantic Elements (ESE) format (version 3.4.1). In addition, the mapping from the detailed specific schema used to the ESE standard must also be available for retrieval. The ESE representation must include all fields that can be mapped to the standard and not only those intended for feeding Europeana.
8.	Interoperability at systems level. Interoperability at syntax and structure level	The metadata records must be mapped to a collection of logical sets that will correspond to the Sets defined in the OAI-PMH protocol, thus enabling selective harvesting per set. The collection of sets must at least include the following: i. The set of metadata records that include full text ii. One set for each distinct value of the Duplin Core field dc.type.
9.	Interoperability at syntax and structure level.	The Duplin Core metadata of each record must include at least one permanent identifier in the dc.identifier field that uniquely identifies the record. The permanent identifier is not allowed to be modified or assigned to another digital object. The permanent identifier must be generated according to the Handle international standard.

A/A	Requirement	Specification
10.	Interoperability at syntax and structure level.	The metadata of every record that will be available to the central aggregator must include valid and accessible URLs for every associated digital object. These URLs must respond with resources suitable for the audiovisual preview of the corresponding digital objects. For example, a URL may serve a thumbnail if the digital object is an image, a cover image if it is a book, an image of the first page if it is an article or embedded code referring a third party web page, if it is streaming video.
11.	Interoperability at syntax and structure level. Interoperability at semantic level.	The encoding of the dc.type field in metadata records that are available for OAI-PMH harvesting must include a unique identifier and a reference to a standard vocabulary of terms for types. In case where no suitable vocabulary/thesaurus is publically available for a metadata field, the carrier that carries out the documentation is free to define one.
12.	Interoperability at syntax and structure level.	The encoding of the dc.creator field in metadata records that are available for OAI-PMH harvesting must follow a standard for bibliographic references with regards to the names of the creators.
13.	Interoperability at syntax and structure level.	The encoding of the dc.language field in metadata records (if exists) that are available for OAI-PMH harvesting must follow the standard ISO 639-2. For those languages that the standard maintain two different codes, a bibliographical code and a terminological code, the former code must be used (ISO 639-2/B).
14.	Interoperability at syntax and structure level.	The encoding of the date fields (dc.date and qualifiers) in metadata records that are available for OAI-PMH harvesting must follow the standard ISO 8601.
15.	Interoperability at syntax and structure level.	The field dc.date in metadata records that are available for OAI-PMH harvesting will refer to the issue date, if such date exists. Otherwise the date of creation will be filled in the dcterms.created field (http://purl.org/dc/terms/created).
16.	Interoperability at syntax and structure level. Interoperability at semantic level.	New controlled vocabularies, thesauri and terminology related tools developed as part of the documentation of the repository should be available for export to a format compatible with one of the following standards: Simple Knowledge Organization System (SKOS), ISO 2788, ISO 5964, ISO 25964-1
17.	Interoperability at semantic level.	Any reference in metadata records to thematic categories should include a unique identifier pointing to an entry in an established controlled vocabulary/thesaurus.

A/A	Requirement	Specification
18.	Interoperability at semantic level.	Any reference in metadata records to a person name should include a unique identifier pointing to an entry in an established controlled vocabulary/thesaurus. In case where no suitable vocabulary/thesaurus is publically available, the carrier that carries out the documentation is free to define one.
19.	Interoperability at semantic level.	Any reference in metadata records to geographical name should include a unique identifier pointing to an entry in an established controlled vocabulary (e.g. geonames) and/or thesaurus.
20.	Interoperability at semantic level.	Any reference in metadata records to an organization name should include a unique identifier pointing to an entry in an established controlled vocabulary/thesaurus. In case where no suitable vocabulary/thesaurus is publically available, the carrier that carries out the documentation is free to define one.
21.	Interoperability at semantic level.	Any reference in metadata records to a chronological period should include a unique identifier pointing to an entry in an established controlled vocabulary/thesaurus. In case where no suitable vocabulary/thesaurus is publically available, the carrier that carries out the documentation is free to define one.
22.	Interoperability at semantic level.	The repository must use one of the internationally recognized and established schemata and ontologies for the documentation of its Digital Resources that allow the detailed capture of the semantics of the metadata, such as MARC21, UNIMARC, MODS, EAD, Europeana Data Model,CERIF, LIDO, VRA Core, CIDOC-CRM.
23.	Interoperability at systems level. Interoperability at semantic level.	The repository disposes its metadata as Linked Data.

Table 4. Specifications for all repository categories (Level 1)

The specification will be updated for different types of repositories (with respect to the nature of their main content) with the corresponding interoperability levels.

3.3. **Persistent identifiers (addresses) for open digital content resources**

In achieving continuous interoperability, it is essential that, for every digital resource, a constant addressing scheme is established that is independent from the specific software system and internet address used for its provisioning, so as the access to the resource to be independent from the carrier and the software systems it uses for providing the resource. It is required that the repository that hosts the digital objects utilizes a persistent identifier handle service that assigns to the resources permanent locations **that are independent from the repository software and the internet addresses from which the resources are accessible**. The two systems that provide such mechanisms are the Handle System RFC3650 standard and the DOI system, which is based on the Handle system. Unlike the DOI system, the Handle system does not impose any cost for the retrieval of a permanent address per digital resource.

Specifications for persistent identifiers

A/A	Requirement	Specification
1.	**Persistent identifiers for digital objects.**	All digital objects are available from repositories that support persistent identifier mechanism.
2.	**Persistent identifiers for digital objects.**	Each persistent identifier must be independent from the internet domain that hosts the resource, the carrier and the repository software.
3.	**Persistent identifiers for digital objects.**	An appropriate authority must assign and issue a persistent identifier to every metadata record that resides in the repository, according to the Handle standard (RFC3650, RFC3652).
4.	**Persistent identifiers for digital objects.**	The web infrastructure that serves the metadata records and the digital content must permanently provide these resources through accessible URLs derived from their persistent identifiers.
5.	**Persistent identifiers for digital objects.**	The software must support Handle System RFC3650.
6.	**Persistent identifiers for digital objects.**	Each persistent identifier is not permitted at any time in the future to change or be assigned to another digital resource.

Table 5 Specifications for the requirement for persistent identifier assignment to digital resources.

3.4. Specification for the delivery of open digital content

The presentation and delivery of digital content to end users constitutes a critical factor in today's landscape. One minimum condition for the effective presentation of content to end users is the compliance with the specifications defined in the Greek Certification Framework for Public Administration Sites and Portals [16].

However, towards achieving the objectives of the programme, an effort for providing an attractive "user experience" is needed for the widest possible dissemination of content of all types to the interested public and for stimulating the reuse of that content by increasing its potentials.

It is necessary to note that the respective specifications are to a great extend linked to specific types of digital sources (e.g. text, image, audiovisual). Therefore, the specifications can be divided into two categories, (a) those independent from the type of the content and (b) those that apply only to specific types of digital resources.

Detailed requirements tables are presented below for specifications of the category (a) (divided into two levels) and for specifications of the category (b), particularly for two specific content types, text and image. Those specifications derived from accumulated experience, existing content disposal systems as well as successful implementation attempts, both Greek and international.

Mandatory presentation and disposal specifications that apply to all types of content (Level 1)

A/A	Requirement	Specification
	View of the content in the appropriate	For each record a dedicated web page must exist that presents its metadata and the view of

	context	the respective digital object. In some cases, the web page may contain only hyperlinks, instead of actual views, that lead to separated pages dedicated for the viewing of the digital object.
2.	Easy and distinguishable retrieval of material by user.	The record presentation web page allows the download of the digital object (if not forbidden by copyright restrictions) through an easy-to-see selection or hyperlink.
3.	View of the content in the appropriate context. Interconnection from external sources.	The record presentation web pages and the related digital object view pages (where exist) are directly accessible through human-friendly URLs. The URL of the presentation web page of a record must contain its persistent identifier.
4.	Easy and distinguishable retrieval of material by user.	Simple and advance search on the metadata records. The list of search results must contain hyperlinks to the respective record presentation web pages.

Table 6. Presentation and content delivery specifications that apply to all types of content (Level 1)

Presentation and content delivery specifications that apply to all types of content (Level 2)

A/A	Requirement	Specification
	Multiple delivery channels	Apart from the conventional web client devices (desktop/laptop computers), the content is also available - with the appropriate adjustments – for access from alternative devices, such as e-book readers, net-books, tables and smart phones.

Table 7. Presentation and content delivery specifications that apply to all types of content (Level 2)

Mandatory presentation and content delivery specifications for text based material

A/A	Requirement	Specification
1.	User experience improvement.	The digitized text must be presented as a whole and unbreakable – the use of multiple hyperlinks to separated scanned pages, which would require multiple user actions, is not sufficient.
2.	User experience improvement.	The text must be available for online "leafing through" from browsers/e-book readers. The feature must give users the capability to zoom in order to make reading comfortable and must not need any external proprietary software other than the common web browser to function properly.
3.	Easy and distinguishable retrieval of material by user.	Key word search capabilities against full text. Text items derived from scanning, for which optical character recognition (OCR) process was omitted, due to low success rates, are excluded from this specification.

Table 8. Mandatory presentation and content delivery specifications for texts

Mandatory presentation and content delivery specifications for images

A/A	Requirement	Specification
1.	User experience improvement.	Multi-level zooming in/out with "pan" capabilities, wherever not conflicting to copyright restrictions. It is recommended that the magnified portion of the image along with the complete image be hosted together on the same page.
2.	User experience improvement.	At least three-level rotation in steps of 90 degrees, clockwise and counterclockwise, wherever this feature is considered to be useful to the web users.
3.	User experience improvement.	Hyperlink that leads to direct view of the image at the maximum resolution available in the Internet. Defining the maximum resolution in which an image is available is subject to copyright restrictions or to balancing between improving the user experience and keeping the size of the image file at a reasonable level.
4.	User experience improvement.	Support for specifications 1, 2 and 3 should introduce the minimal requirements regarding web browsers. Preferably, there should be a version of the presentation that requires only JavaScript support from the browsers. It is acceptable but not recommended that those features be based exclusively on advanced technologies, such as Flash, Java and Silverlight. The optimal alternative would be providing at least two versions, one of limited requirements (JavaScript) and a more demanding one (e.g. Flash, Java, or Silverlight)
5.	User experience improvement.	It is recommended that the aforementioned features be available within the record presentation web page (which would contain both metadata and image viewing), so as the user does not need to move to another page.

Table 9. Presentation and content delivery specifications for images

3.5. Interoperability with aggregators and the open digital content e-infrastructure of EKT

The usability and impact of the content would benefit from ensuring interoperability with a) a consolidated catalogue (aggregator/open digital content register) and b) a system for the secure preservation and safe keeping of open digital content, with an overall aim of creating a central point for searching digital content and implementing a basic security mechanism and digital backup scheme for the produced open digital content.

This e-infrastructure (register of open documented digital content) may harvest metadata – supporting incremental automatic delta harvesting at regular intervals – from the individual repositories accumulating the content into a central system with a view to creating a consolidated catalogue. It should be noted that the provided specifications ensure interoperability with any aggregator system whose operation is based on the referenced international widespread open standards. Additionally, for purposes regarding basic information systems security and protection of investments in digitalization projects, organisations are encouraged

(but not obliged) to ensure interoperability with a secure content preservation and safe deposit system provided by EKT that supports automatic remote backup of digital content in the EKT e-infrastructure. This system is a prerequisite for future application of basic bit wise preservation rules regarding digital material.

This ensures that at least one remote replica of the data, metadata and original digital content is preserved. It can be viewed as a remote data replication/disaster recovery system in the field of digital content which leverages already required features and interfaces of repositories that support their main functionality (organized metadata, interfaces, persist identifiers etc.) to achieve a basic level of security and availability of content.

A/A	Requirement	Description
1.	**Deposition of repository metadata in an open digital content register**	Providing mechanism for disposing metadata for harvesting according to the protocol OAI-PMH.
2.	**Deposition of repository digital resources in a digital content secure preservation infrastructure.**	METS or OAI-ORE based mechanisms (or any other that is proved to be equivalent) for the delivery of data replicas to a central infrastructure.
3.	**Deposition of repository digital resources in a digital content secure preservation infrastructure.**	Web service with authentication/access restriction capabilities that when given a persist identifier as input it responds with the respective digital object.

Table 10. Interoperability with open digital content aggregator and safe deposit system

4. Indicative software for implementing open digital content provisioning systems

Table 13 enumerates some indicative systems that can be exploited in the implementation of the repository projects based on the aforementioned specifications. The table includes an indicative set of only Free Software / Open Source Software (FS/OSS) systems, thought other relevant open source as well as proprietary systems may exist. The objective of this section is to highlight specialized software for digital library/content repository/content issuing systems that have adopted standards that are under development and leverage the latest technological developments in the respective fields.

These systems are fully interoperable and implement a layered architecture consisting of a "stack" of FS/OSS systems (Figure 3) that provide digital library and repository services, are based on international open standards and are characterized by well scalable and predictable cost and complexity.

A/A	Field	Software and Standards
1.	**Presentation of digital items**	FS/OSS Software ▪ Internet Archive Book Reader ▪ Flexpaper ▪ Multivio Standards ▪ PDF/A ▪ Jpeg ▪ hOCR ▪ ABBYY XML OCR ▪ PDF/A
2.	**Aggregators and national catalogues**	FS/OSS Software and Systems: ▪ DRIVER / D-NET ▪ Europeana Standards ▪ OAI-PMH ▪ OAI-ORE ▪ CERIF
3.	**Handle services**	▪ HANDLE.NET ▪ RFC3652
4.	**Disposal and organization of content:** Repositories e-magazines and issues Digital libraries	FS/OSS Software and Systems: ▪ DSpace, ▪ Fedora ▪ E-prints ▪ Omeka ▪ Greenstone ▪ OJS ▪ openABEKT (under development) Standards ▪ Dublic Core, MODS, UNIMARC, MARC 21, EDM, CDWA, LIDO, CIDOC CRM, CERIF ▪ OAI-PMH, Z39.50, SRU.W, OpenSearch, OpenURL
5.	**System Infrastructure**: Operation Systems, Middleware/DB Virtualization and Cloud Platforms Single Sign On (SSON) Systems Issue management and monitoring systems	▪ Linux Tomcat, JBoss, Postgress,MySQL ▪ XEN, KVM, G-Cloud and even Windows 2008/VMWARE, ESXi ▪ Shibboleth, openid ▪ nagios, cacti, graphite, pupet, awstats

Table 11. Indicative digital content management systems based on FS/OSS and open standards

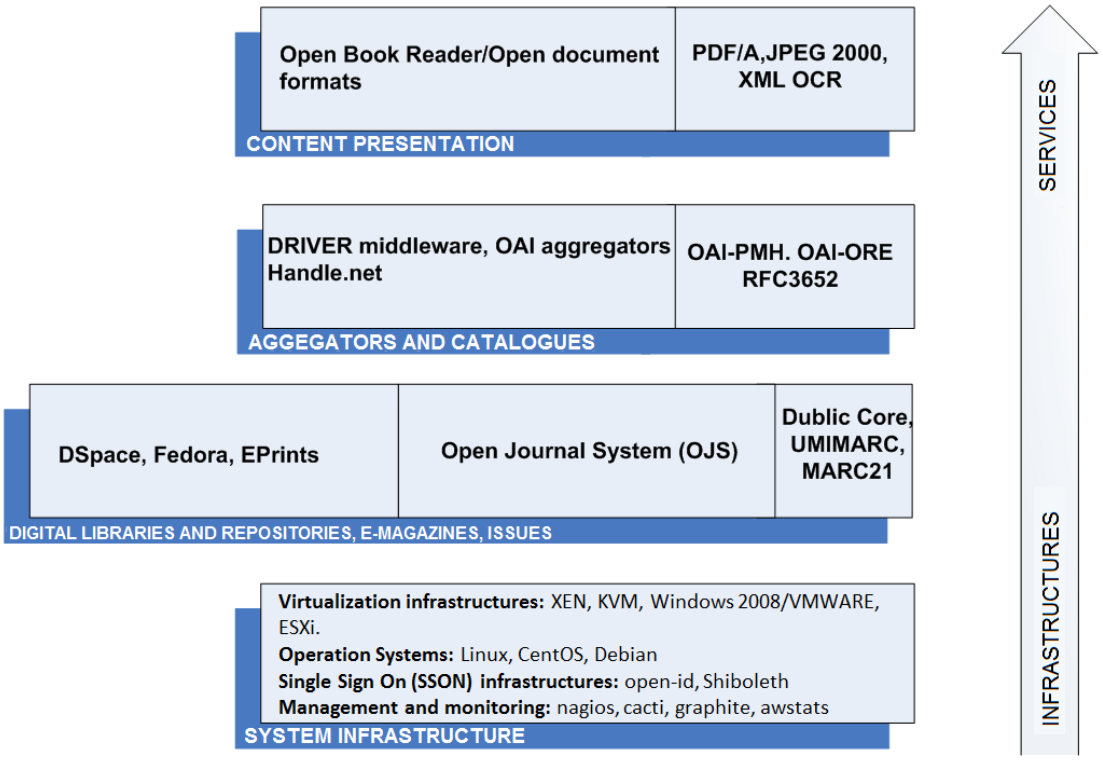

Figure 3 Indicative «stack» of FS/OSS for the management and delivery of digital content in compliance with open standards

5. Summary of functionality and features

The following table summarizes how the aforementioned specifications contribute to the successful implementation of the projects that employ them, over various key areas.

	Increase of the international usage of the content (e.g. inclusion in search engines, full text indexing)	Increase of the domestic usage of the content	Reuse of content	Development of value-added Services	Ensuring of availability, security, preservation and reuse of content (digital preservation)
Optical Character Recognition (OCR) for supporting full text searching and indexing	√	√	√	√	√
Interoperability with open documented content register of EKT	√	√	√	√	√
Interoperability at the level of metadata syntax and structure	√	√	√	√	√
Interoperability at the semantic level.	√	√	√	√	√
Content interconnection from external sources.	√	√		√	√
Multi-channel content delivery	√	√		√	
Web user can easily find and retrieve the material.	√	√	√		
Better user experience. View of material in the appropriate context.	√	√			
Storage of structured content in central location for digital preservation/ remote replication of digital content objects.			√		√
Persistent addresses for digital objects (persistent identifier handle service).			√	√	√

Table 12. Functional requirements for projects in relation to the defined objectives.

References

[1] "Guide on best practices for digitization and long term preservation of cultural content", study in the context of the 3rd CSF, Operational Programme "Information Society", Measure 1.3, available in Greek at: http://digitization.hpclab.ceid.upatras.gr/, http://www.infosoc.gr/infosoc/el-GR/services/elibrary/reports_list/psifiopiisi/synoptikos_odigos.htm

[2] "Study on digitalization of 3-D objects", study in the context of the 3rd CSF, Operational Programme "Information Society", Measure 1.3, available in Greek at the following temporal location: http://www.infosoc.gr/infosoc/el-GR/services/elibrary/reports_list/psifiopiisi/3dimensions.htm

[3] Centre for Cultural Informatics, FORTH (April 2005). Guide for Cultural Documentation and Interoperability, study in the context of the 3rd CSF, Operational Programme "Information Society", Measure 1.3, available in Greek at http://www.ics.forth.gr/CULTUREstandards .

[4] Cornell University Library, *"Moving Theory into Practice: Digital Imaging Tutorial" available at* http://www.library.cornell.edu/preservation/tutorial/contents.html

[5] Harvard University Library, *"Selection For Digitizing: A Decision-Making Matrix,* http://www.clir.org/pubs/reports/hazen/matrix.html

[6] International Federation of Library Associations and Institutions (IFLA), *"Guidelines for digitization projects", available at* http://www.ifla.org/VII/s19/pubs/digit-guide.pdf

[7] National Archives and Records Administration (NARA)*, «NARA Guidelines for Digitizing Archival Materials for Electronic Access»* http://www.archives.gov/research_room/arc/arc_info/guidelines_for_digitizing_archival_materials.html

[8] Public Online Consultation for the general principles and priorities of the Operational Programme Planning "Digital Convergence" http://www.opengov.gr/ypoian/?option=sygglisi.

[9] Ouksel A.M. and Sheth A. (1999) Semantic Interoperability in Global Information Systems, ACM SIGMOD Record, Vol 28(1) March 1999, pp 5-12.

[10] Europeana portal, http://www.europeana.eu

[11] DRIVER portal, http://www.driver-repository.eu/

[12] Shreeves, S. L., J. Riley, and L. Milewicz (2006, August). Moving towards shareable metadata. *First Monday 11*(8).

[13] Angelaki, G., R. Caffo, M. Hagedorn-Saupe, and S. Hazan (2010, April). Athena: A mechanism for harvesting europe's museum holdings into europeana. In *Museum and the Web 2010*, http://www.archimuse.com/mw2010/papers/angelaki/angelaki.html.

[14] Athena project (2009, April). Deliverable 3.1: Report on existing standards applied by european museums, http://www.athenaeurope.org/index.php?en/149/athena-deliverables-and-documents.

[15] Athena project (2009, July). Deliverable 3.2: Recommendations and best practice report, http://www.athenaeurope.org/index.php?en/149/athena-deliverables-and-documents.

[16] (2008, November). Greek Interoperability and eGovernment Services Framework, Certification Framework for Public Administration Sites and Portals, available at http://www.e-gif.gov.gr/portal/pls/portal/docs/216024.PDF

[17] DRIVER Guidelines 2.0: Guidelines for content providers - Exposing textual resources with OAI-PMH, November of 2008, available at: http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_2008-11-13.pdf.

[18] DINI Certificate: Document and Publication Services 2007, Tachnical Report, version 2.0. available at http://nbn-resolving.de/urn:nbn:de:kobv:11-10075687.
[19] The ISO/IEC 27000-series numbering ("ISO27k") has been reserved for a family of information security management standards http://www.iso27001security.com/html/iso27000.html
[20] BS 7799 http://www.bsigroup.com/, British Library Digital Preservation Team – strategy, available at http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/digprestteamstrat/index.html
[21] Office of Government Commerce (2002). ICT Infrastructure Management. The Stationery Office. ISBN 0113308655.
[22] OAIS, Open archival information system, ISO 14721:2003
[23] Trustworthy Repositories Audit and Certification Checklist (TRAC), http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf
[24] Blue ribbon task force Sustainable economics for digital planet ensuing long term access to Digital Information http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf
[25] The Handle System, http://www.handle.net/

List of Tables and Figures

Tables

Figures

ΕΚΤ

ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΤΕΚΜΗΡΙΩΣΗΣ

NATIONAL DOCUMENTATION CENTRE

www.epset.gr

Εθνικό Πληροφοριακό Σύστημα Έρευνας και Τεχνολογίας

Κοινωνικά Δίκτυα - Περιεχόμενο Παραγόμενο από Χρήστες