MINERVA

Ministerial Network for Valorising activities in digitisation

MINERVA - IST-2001-35461

Deliverable Number	D3.2
Deliverable Title	Feasibility survey of the common platform
Deliverable Type	Report
Responsible Partner	Relais Culture Europe & Ministère de la Culture et de la Communication (France)
Authors	Editor : Muriel Foulonneau Contributors: Alicia Moreno, Erminia Sciacchitano, Cristina Magliano, Dimitrios Koutsomitropoulos, David Dawson, Pete Johnston
Status	Public
Submission Date	September 2003 final revision
Notes	None

Contents

CONTENTS **2**

1. INTRODUCTION **3**

2. OBJECTIVES AND AUDIENCE **4**

3. DESCRIPTION OF DIGITAL CONTENT CREATION **5**

3.1 A DATA MODEL **5**

3.2 REFERENCE TO EXISTING STANDARDS AND TERMINOLOGIES **7**

4. PUBLISHING **14**

4.1 DISTRIBUTED DATA COLLECTION 14
4.2 ACCESS POINTS 14
4.3 ACCESS TO ON-LINE RESOURCES 14
4.4 CONTENT QUALITY 15
4.5 TECHNICAL SPECIFICATIONS 15

5. SET A PROPER OBSERVATORY TO COLLECT INFORMATION **16**

5.1 AN OBSERVATORY TO STRUCTURE THE DIGITISATION FIELD 16
5.2 SELECTING INFORMATION 16
5.3 GATHERING INFORMATION 16
5.4 CROSS DOMAIN OBSERVATORIES 17

6. MULTILINGUAL ACCESS TO DIGITAL CULTURAL CONTENT - CURRENT ACTIVITIES AND MAJOR CHALLENGES **18**

6.1 THESAURUS AND TERMINOLOGIES 18
6.2 FULL TEXT AND EXPLOITATION OF NATURAL LANGUAGE 19
6.3 PRODUCING MULTILINGUAL RESOURCES 20
6.4 THE ROLE OF EUROPEAN EXPERTS 20

7. THE EUROPEAN COLLABORATION **21**

7.1 BUILDING OAI REPOSITORIES 21
7.2 VARIOUS USER PROFILES 21
7.3 ACCESS TO MULTILINGUAL RESOURCES 23

ANNEX 1 LIST OF REFERENCES **25**

ANNEX 2 MODEL/EXAMPLE OF QUESTIONNAIRE TO COLLECT INFORMATION **26**

1. Introduction

European countries have already invested significantly in programmes for digitising cultural and scientific content. These activities have included several areas, such as museum objects, archaeological and environmentally important sites, music and audio-visual archives, bibliographic materials, documents and manuscripts. The main challenges are now promoting the uptake of new technologies for the digitisation of cultural and scientific content, ensuring lasting accessibility and preservation, the development of new services and job opportunities. Other important objectives include strengthening the European content industry and stronger support for its co-operation with educational communities, with consequent mobilisation of material and immaterial resources.

The National Representatives Group on digitisation of cultural and scientific heritage has supported all European countries in setting and explaining policies, defining the scope and structuring of this field and the Minerva project aims at providing common tools to support those policies.

Policies need indicators and scoreboards as decision-making tools, which can help set the priorities for digitisation, evaluate the progress, impact and results of the frameworks implemented in each country.

In order to facilitate the structuring and the management of the digitisation activity and to facilitate digital resource discovery, national inventories of digitisation projects and digitised collections have been set up in various countries. They describe institutions which lead such projects, projects, digital collections created and services and products made to access digitised content. Such strategic tools help identifying what is going on with digitisation and which are the results of policies implemented, they allow to measure the outcomes of the framework and possibly set priorities for future programmes.

2. Objectives and audience

The **objectives** and the audience of inventories of digitisation activities must be well-defined. Most systems usually come from the library sector and are then extended to other cultural heritage collections. Their main focus is either management of digitisation activities or citizen access to existing resources, or both :

Management issues

- Providing examples of practices to decision-makers intending to launch a new project;
  Decision-makers need tools to identify competence centres and support the dissemination of standards. Project operators can benefit from other institutions' experience on specific types of projects and material.
- Creating complementary collections and avoiding duplications;
  The overview of digital collections and digitisation projects are used as a basis for setting up a policy on digitisation through priorities set on physical collections to digitised. Directories of digitised resources and digitisation projects may refer to equivalent inventories of physical collections. Those tools guarantee to save efforts by not digitising twice the same material. This issue is particularly important for libraries which deal with non unique objects.

Citizen access to cultural heritage

- Providing access to web sites containing digitised collections;
  Inventories of digitisation initiatives may be a means of setting up intermediary tools for providing item-level access to cultural heritage resources, all the same as a Web directory which is used as a guide to individual resources / Websites.
- Facilitating new services for item-level access.
  They can also be intermediary products, allowing to identify collections to build metasearch tools (eg. systems such as TASI precise Z39.50 compliance) in heterogeneous catalogues.

The audience of directories of digitised collections and digitisation projects may be either professional (project leaders) or organisational (strategic decision-makers in the digitisation sector) or the general public to access digital resources.

3. Description of digital content creation

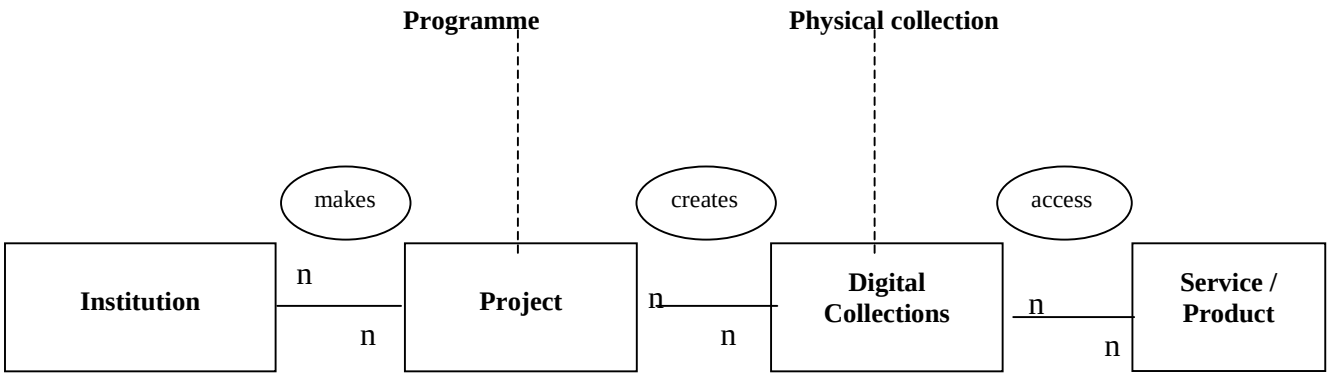
The inventorying activity usually cover all types of heritage resources and the directories are an opportunity to build cross-domain service for the whole heritage field. National inventories on digitisation must provide an opportunity to take into account cross- domain issues, notably in the creation of digital cultural content based on heritage collections. A major challenge is to manage all types of cultural heritage collections, whether in libraries, archives, museums, monuments …
The purpose is to gather information in order to gather the following information :

- who launches projects (institutions);
- projects (objectives, duration, funding…);
- material (collections subject, size, format);
- on-line or off-line locations : services for exploiting the content.

3.1 A data model

Description of digitisation activities shall be supported by a proper data model and include the description of key entities, such as a digital collection, a digitisation project, a Website and a cultural heritage institution .

A digitisation project creates one or more digital collection(s)



4 core entities

- The ***institution*** which is important to record key information on digitisation activities is the one that launches or coordinates the digitisation project. Although many other institutions may be associated either to the project or to the collection or the service/product as content holder, service publisher or technical subcontractor, this

institution is the creator of digital resources, receives the money for carrying out a project and develops competencies in the field.

- The ***projects*** launched have a start date and an end date, a clear defined budget, defined goals and operational objectives. Projects can be recorded even if not started if the funding is certain. They should include the creation of digital surrogates of digital resources which have a heritage value.
- The ***digital collections*** created are descriptions of sets of items aiming to provide an overview of digital content created, unlike item-level descriptions.
  A collection is "an aggregation of items"[1] but it may have a different meaning according to its curatorial environment : "Almost always, the collections of 'archives' delineate themselves: they relate normally to a specific person or institution. The collections of 'libraries', on the other hand, should be delineated by the purpose for which the library exists: by the information needs of their user populations. In contrast, the collections of 'museums', are - again - delineated somewhere between those two extremes."[2].
  There is no comprehensive cross-domain definition of a collection but certainly a need to define sets of digital resources created. Existing experiments do not impose any definition to the people involved in the projects. Those have rather applied pragmatic approaches to a collection according to its definition in the digitisation process (eg. the collection 3D objects from the museum X) and/or according to the service or product therefore created (eg. the collection of images digitised for the YYY virtual exhibition). Referring to the content creation process may appear as an adequate starting reference to describe digital cultural datasets in this context.
  Access to the collection may be provided offline, this does not exclude future on-line service.
- The descriptions must include the ***services and products*** which allow to access that content, typically an online database, a virtual gallery or a CDRom - "A service is a system that provides one or more functions of value to the end-user. Examples include: […] interlibrary loans, a Z39.50 or Web server."[3]

The digital content creation environment

However, the reference to the **physical collection(s)** which have been used to create the digital collection allows to draw an equivalence with the physical collections management strategies. The full **programme framework** description includes information on rules and standards applied for technical, legal as well as project management issues.
Multiple **agents/institutions with multiple roles** in the digitisation process may act in the content creation process as a project coordinator, as a project partner, content holder, content owner, technical sub-contractor, physical collection holder, physical collection owner, rights holder, service publisher…
Inventory systems may or may not include those entities to report on digitisation activities. Those descriptions may usefully complete the description of projects and content by providing contextual information.

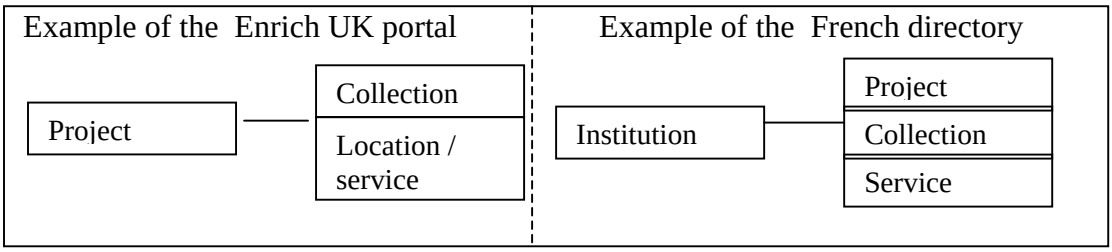
1 Dublin Core terminology

2 Andy Powell, "Collection Level Description - a review of existing practice", 1998, <http://www.ukoln.ac.uk/metadata/cld/study/>, p. 8

3 Still at draft stage <http://dublincore.org/usage/terms/dcmitype/>

A modular approach

Referring to the data model, it appears very important to focus the data collection according to its own needs. The entities (project/collection/service) may not be separated in the data collection process. For example, a system may only consider a relationship 1-1 for a collection and a service, describing a Website rather than digital collections, thus modelling a single entity in the information system.



The model presented is modular, as long as an entity is adequately defined, it is meaningful, even if not all entities are described, for example, the descriptions of services and collections may allow to build a classical Web directory, with no consideration for the digitisation process.

3.2 Reference to existing standards and terminologies

The reference to existing metadata and terminology standards is a key issue to ensure service sustainability and interoperability with complementary applications such as benchmarking systems.

However, to include cross-domain information and facilitate new data recording (when collection level descriptions do not already exist for example), it appears easier to keep simple organisation of data, prefer simple comprehensive terminologies rather than elaborated thesauri and avoid sector-specific categories, notably for physical material formats and subjects.

Metadata sets

The following metadata sets have been designed from the Dublin Core metadata set, the RSLP model for collection level description and several other examples and practices which have been implemented to describe related material.

In order to ensure data re-usability, the reference to standard metadata sets is important. Indeed, the equivalence between RSLP schema for collection description and the Learning Objects Metadata schema are being mapped[4]. The elements on the content creation process, notably quality and formats used and the audience of either the service or the collection must ensure data re-usability and relevance of the inventory to all activity types.

4 work in progress in UKOLN

Institution		
Minerva Attribute	**RDF property**	**Definition**
General attributes		
Institution Name	**vcard:fn**	The name of the institution in charge.
Identifier	**dc:identifier**	A formal identifier for the institution.
Type		Institution activity sector, whether archive, library, museum, local community…
Legal Status		Administrative status of the institution, whether public, commercial or non profit
Jurisdiction		The organisation the institution is affiliated with eg. an institution under the umbrella of a ministry.
Role	**vcard:role**	The role (typically an organisational role) fulfilled by the institution in the digitisation process.
Telephone number	**vcard:voice** (sub-property of vcard:tel)	The telephone number of the institution contact.
Fax number	**vcard:fax** (sub-property of vcard:tel)	The fax number of the institution contact.
Email address	**vcard:email**	The electronic mail address of the agent.
Address	**vcard:adr:pobox** **vcard:adr:street** **vcard:locality** **vcard:region** **vcard:pcode** **vcard:country**	Physical location of the institution
URL	**vcard:URL**	Institutional Website

Project		
Minerva Attribute	**RDF property**	**Definition**
General attributes		
Title	**dc:title**	The name of the project.
Identifier	**dc:identifier**	A formal identifier for the collection.
Status		Project status, whether planned, on-going or completed
Start date	**dc:date**	Start date of the project
Completion date	**dc:date**	Completion date of the project
Description	**dc:description**	A description of the project, including objectives, textual or keywords.
Digitisation process	**dc:description**	The technical feature of the digitisation process, whether direct or indirect, which technical choices…
Funding		Funding sources, amount, duration, including Associated programme
Institution in charge	**dc:creator**	Institution responsible for the project
Contributing institution	**dc:contributor**	Other institutions involved in the project
Associated project	**dc:relation**	Related projects
Contact person	**vcard:fn**	Name of contact person, project manager
Telephone number	**vcard:voice (sub-property of vcard:tel)**	The telephone number of the contact person.
Fax number	**vcard:fax (sub-property of vcard:tel)**	The fax number of the contact person
Email address	**vcard:email**	The electronic mail address of the contact person.
Mail address	**vcard:adr:pobox vcard:adr:street vcard:locality vcard:region vcard:pcode vcard:country**	Physical address of the contact person

Digital Collection		
Minerva Attribute	**RDF property**	**Definition**
General attributes		
Title	**dc:title**	The name of the collection.
Identifier	**dc:identifier**	A formal identifier for the collection.
Description	**dc:description**	A description of the collection.
Digital document format	**dc:format**	The physical or digital characteristics of the collection.
Language	**dc:language**	The language of the items in the collection.
Digital document type	**dc:type**	The type of the collection content : Text / Image / Interactive resource / sound
Access Control	**cld:accessControl** (sub-property of dc:rights)	A statement of any access restrictions placed on the collection, including allowed users, charges, etc.
Size	**dc:format**	An evaluation of the size of the collection. Is dc:format as physical description of the collection in DC Collection
Accrual Status	**cld:accrualStatus** (sub-property of dc:description)	A statement of accrual policy (closed, passive, active, partial/selective), accrual method (purchase, deposit)) and accrual periodicity (closed, irregular, periodic). Especially important for harvesting purpose to foreseen evolution of aggregated resources.
Standards	**dc:description**	Descriptive and terminology standards used.
Legal Status and IPR	**cld:legalStatus** (sub-property of dc:description)	A statement of the legal status of the collection.
Associated service / product	**cld:hasLocation** (sub-property of dc:relation)	The identifier for the online location of the collection.
Subject		
Subject	**dc:subject**	A concept (keyword) of the items in the collection.
Spatial coverage	**dcq:spatial** (sub-property of dc:coverage)	The spatial coverage of the items in the collection.
Temporal coverage	**dcq:temporal** (sub-property of dc:coverage)	The temporal coverage of the items in the collection.
Associated agents		
Owner	**cld:owner**	The identifier for an agent or institution who has legal possession of the collection.
Associated Project	**dc:creator**	The identifier for an agent who gathers (or gathered) the items in a collection together.

External relationships		
Illustration	**dcq:hasPart** (sub-property of dc:relation)	Resources contained in the collection.
Physical collection	**cld:hasAssociation** (sub-property of dc:relation)	Source physical collection(s)
Associated collection	**cld:hasAssociation** (sub-property of dc:relation)	The identifier or name of a second collection that is associated [by provenance] with the current collection.
Audience		Audience of the collection if the aggregation of items is done for a specific audience

Physical Collection		
Minerva Attribute	**RDF property**	**Definition**
General attributes		
Title	**dc:title**	The name of the collection.
Identifier	**dc:identifier**	A formal identifier for the collection.
Description	**dc:description**	A description of the collection.
Physical document formats	**dc:format**	The physical characteristics of the source objects for digitisation.
Institution	**cld:hasLocation** (sub-property of dc:relation)	The identifier for the physical location of the collection, institution which holds (s) the physical collection
Associated digital collection	**cld:hasAssociation** (sub-property of dc:relation)	The identifier or name of a second collection that is the digital representation with part of or the whole current collection.

Service / Product [RSLP Location]		
Minerva Attribute	**RDF property**	**Definition**
General attributes		
Name	**dc:title**	The name of the location.
Identifier	**dc:identifier**	A formal identifier for the location.
Language	**cld:accessConditions**	Language(s) in which the service / product is available, eg. multilingual Website
Type	**dc:type**	DC:type service but may be better defined (eg. database, virtual exhibition….)
Description	**dc:description**	
Technical environment	**dc:description**	Technical features of the service / product
Communication protocol	**Dc:description**	Mention any communication protocol such as OAI-PMH, ZING, Z39.50
Access Conditions	**cld:accessConditions**	Hours of access, classes of permitted user, fees…, technical requirements, if a plug-in or any type of specific technical device or software is needed to access the service / product. WAI compliance ….
Collection contained	**cld:isLocationOf** (sub-property of dc:relation)	The identifier for a collection held online (digital) location.
Audience		Audience of the service, which is designed for one or multiple category (ies) of users
Maintenance		Maintenance conditions (data update…)
Associated agents		
Publisher	**dc:publisher**	Institution or person responsible for making the resource available
Administrator	**cld:administrator** (sub-property of dc:publisher)	The identifier for an agent who has responsibility for the physical or electronic environment in which the collection is held.
Online location		
Location	**cld:locator** (sub-property of dc:identifier)	The online location (URL) of an online (digital) collection.

Element	Terminology
Spatial coverage	1 level down country level for the country considered (regions / provinces / states) + country level for main European countries + continents Best practice would include the reference to a thesaurus
Temporal coverage	Centuries or periods which can be translated into proper dates to ensure interoperability with additional systems Paleolithic III millennium bc VI c. bc I c VII c XIII c XIX c Mesolithic II millennium bc V c. bc II c VIII c XIV c XX c Neolithic I millennium bc IV c. bc III c IX c XV c XXI c Calcolithic IX c. bc. III c. bc IV c X c. XVI c Bronze age VIII c. bc II c. bc V c XI c XVII c Iron Age VII c. bc I c. bc VI c XII c XVIII c
Subject	Archaeology Buildings, Architecture & Monuments Cinema & audiovisual Contemporary art Costume & Textiles Decorative and applied arts Economics Ethnology & anthropology Fine arts Geography Health & medicine Industry & commerce landscapes Law Literature Local history Music Performing arts Philosophy Politics Religion & Beliefs Science & Technology Transports Urbanism & planning (regional, town, country)
Physical document format	Books Articles Newspapers Serials Yearbooks Manuscripts Illuminated manuscripts Bibliographic records Numerical / statistical data Archival records Notes, coins and medals Maps Moving image Printed Images Drawings Engravings Photography Postcard Sound track Archaeological site Monument Incunabulum Installation Musical instrument Music score painting Jewellery Sculpture Textile Cadaster
Digital document types	Still Image Moving image Interactive resource Sound Text 3D
Digital document formats	AVI video Bitmap Image GIF image HTML JPEG file Macromedia Director Macromedia Flash Midi audio Mpeg audio Mpeg video PDF Plain Text PNG Quicktime video Real Audio Sun audio (au) TIFF Image WAV (audio) XML SVG
Audience	formal education life-long learning professional leisure / tourism academic research general public children
Project status	completed / on-going / planned
Access services type	CDRom Interactive terminal Virtual exhibition / gallery DVDRom Online database Offline database Paper Website Teaching aid Portal
Services Communication protocol	Z39.50 OAI-PMH
Institution types	library museum archive academic / university information centre other
Country	ISO 3166-1
Language	ISO 639-2

4. Publishing

To publish a directory on digitisation activities, the main issues are to ensure high quality and up-to-date content, relevant access functionalities and comprehensible reference to digital objects.

4.1 Distributed data collection

To collect data from various sources, it is possible to allow data recording and data update to various persons and possibly to project managers, with a control process by qualified agents from an observatory or by heritage sectors.

4.2 Access points

When publishing the content of the directory, several access points should allow powerful retrieval functionalities, on each entity. Here are the meaningful access points to the institutions, projects, collections and services/products.

Proposed access points

INSTITUTION	***PROJECT***	***COLLECTION***	***SERVICE***
Institution name	Project name	Name	Name
Institution type	Project status	Description	Type
Institution country	Project description	Subject	Language
		Physical formats	Description
		Digital document types	
		Spatial coverage	
		Temporal coverage	

4.3 Access to on-line resources

The directory may be a stage in a larger process to provide on-line access to digital objects. It may be a gateway and optionally help building metasearch functionality on available resources, through metasearch protocols or metadata harvesting for example. The information on availability of data for searching or harvesting is therefore important to allow further functionalities to be implemented either in the directory itself or in specialised services.

The recording of services to access collections and possibly the access to sub-sets of a digital library may be very useful and gateway-style information add information which is complementary to resources display. A search engine may therefore provide retrieve collections and projects' descriptions and possibly individual objects as an extended functionality of the same directory.

Service descriptions shall therefore mention the conditions and limits to access, including WAI standard compliance for disabled users and the necessity of plug-in and material requirements.

4.4 Content quality

The value of an inventory is the quality of the information which is offered and the accessibility of that information to the target audience.
In order to facilitate record browsing, several rules may be applied :

- if the relationship is 1-1 between 2 entities, it may be easier to display both descriptions in a single page.
- if various entities are recorded all together in a single database table for example, they shall be identified when displayed, possibly with sub-titles in the page.
- the presentation of the directory in various languages at least for the interface is very important in a European context. The mention of the language of collections in collection description may also help foreign users to access cultural textual resources.

The quality of the records can be assessed if some information is added :

- last update of a description should be recorded and displayed in order to ensure data reliability. However, it is important to not consider a spell-check for example as an update. The last update element may therefore not be automatic.
- the author of a description should be identified and recorded so that information authenticity can be assessed.

4.5 Technical specifications

The platform should be open-source and should store (or export) XML formatted files[5]. It must implement a repository of the Open Archives Initiatives Protocol for Metadata Harvesting.

It should preferably support WAI compliance.

If containing images, sound tracks, video …. As illustrations of collections, reference to standard file formats (JPEG, PNG) is advised.

Further information on usability and access of cultural Websites in the European quality grid[6].

5 see "Technical Guidelines for Digital Cultural Content Creation Programmes", Minerva interoperability focus.

6 http://www.minervaeurope.org/structure/workinggroups/userneeds/docindex.htm

5. Set a proper observatory to collect information

The main challenge of an inventory on digitisation is to ensure data reliability, exhaustiveness, update and value. The information is therefore gathered by an institution acting as an observatory. It leads to provide an overview of what is going on in the field of digitisation of heritage and which digital knowledge is created.
The efficiency of the data collection directly depends upon the stage of the inventorying activity, the clear definition of data collected and the processes implemented to ensure data collection.

5.1 An observatory to structure the digitisation field

Efforts to inventory initiatives are often led at a higher organisational level, state, province, national agencies, programme managers or by domain, national libraries, or sectoral consortia : those who have a role in structuring a sector or funding digitisation activities can decide to foster the knowledge of those activities and help resource discovery.

5.2 Selecting information

The objective of an inventory may not be exhaustiveness and the data collected must be clearly defined.

5.3 Gathering information

Various strategies can be adopted to gather information on digitisation activities:

- **a survey** : a questionnaire is sent to institutions which may have launched digitisation projects
- In several cases, that survey is then published on the Web so as to set up an **online directory** of projects and/or collections and/or services/Websites
- Based on that survey, a proper **service** can then be set up, which allows to collect information on regular basis on projects, collections and services

A survey
It is important to start by an ***overview of information sources*** :

- competence centres
- professional mailing lists (eg. librarians)
- experts mailing lists (eg. digital images)
- funding programmes
- professional consortia
- national / local authorities on cultural heritage institutions

On that basis, it is possible to launch a larger mailing to custodial institutions with a questionnaire, whether online form or paper format.

An operational service
Several conditions enable such a directory to be kept up-to-date :

- to know about new initiatives;
- to know about the evolution of directory records (eg. changes in project status, amount of surrogates created…).

A proper observatory or an institution acting as an observatory on digitisation can therefore ensure the maintenance of such a service by getting support from existing information sources.
In order to ensure data update it is easier to rely on existing information flows :

- **competence centres** : specialists of technical issues which have a regular knowledge of the persons who need support or face specific difficulties. Those can also be virtual communities of experts (through mailing lists).
- **funding authorities**, notably programme managers.

Regular data update through all information sources can of course complement the data collected through competence centres and funding authorities.

5.4 Cross domain observatories

The analysis of collections, service types and institution types is an important management issue for setting future policies and programmes. This shall rely on key information sources to ensure high quality content on a regular basis to a department/agency/institution acting as an observatory on digitisation. The observatory should have legitimacy in the related field(s), access to the relevant information, dissemination capacity and a cross-domain scope of action.

The digitisation process has many features in common from one heritage sector to another. The management of digitisation activities can be an opportunity to build valuable and meaningful cross-domain services, as a preparation of cross-domain services to access cultural objects in the future. Therefore, the collections descriptions should not re-create categories of material for each sector when not necessary, but rather encapsulate all heritage information in common categories.

The committees gathering main sectors representatives and main institutions involved in digitisation should certainly lead to build proper cross-domain observatories to inventory initiatives, help managing digital content creation and support the setting up of intelligent collaborative and interactive service to access cultural content. That strategy should also take into account data re-usability for research or education purposes and be common to all sectors involved in digitisation activities.

6. Multilingual access to digital cultural content - Current activities and major challenges[7]

The coordination of digitisation policies in Europe will lead to common online services to access heritage resources, raising the European added value for citizens. However, a major challenge is to respect cultural and linguistic diversity while ensuring a coherent access to resources produced in various countries. Technologies can certainly help building efficient services on the Internet and allow citizens to access European treasures.
In order to validate solutions to deal with multilingual descriptions of collections and services, and more generally to build common access services in Europe, it is important to consider existing solutions, the state of the art of research on automatic cross-lingual information retrieval, finally to facilitate the production of multilingual resources. The meeting taking place in Paris on October 3rd, 2003 has allowed to consider several issues for multilingualism and cultural asset.

> "Lack of simple, common forms of access for the citizen
> Access by the citizen to the different resources, at national and at EU level, is compromised by the lack of common approaches and technical standards as well as by lack of support and systems for multilingual access. "
>
> Lund Principle, April 2001

6.1 Thesaurus and terminologies

The use of thesaurus and terminologies is widespread to build multilingual systems but it appears difficult to consider resources on all subjects and many languages.
The ISO standard **ISO 5964:1985** Documentation - Guidelines for the establishment and development of multilingual thesauri is the basis of such a work. A multilingual thesaurus shall implement interfaces to **manage relationships** and to **retrieve equivalences**. Relationship management can be handled through collaborative work interface (MACS project).
It is possible, like for the HEREIN project (on heritage policies) or the NARCISSE project (on museum objects restoration) to build a common thesaurus from scratch and translate the common concepts or to translate an existing terminology.
The major challenge comes when creating interoperability between existing systems. It is possible that there is no equivalence between concepts : that is one of the reason why the MACS project does **not use pivot terminology** for subject headings. Their choice was :

- Equal consideration for all languages : no pivot language
- No new thesaurus : equivalences between each existing thesaurus to all other ones
- Relations are conceptual clusters = mappings and numeric identifiers
- Results shall be coherent for the end-user

This multilingual thesaurus can be considered as **a service** in a larger framework of information retrieval, such as implemented in the TEL project. The TEL project gathers 8 European National libraries. In 2004, it will allow to access catalogues (through Z39.50 gateways or OAI-PMH repositories) and uses MACS to allow cross-language retrieval.

[7] http://www.cordis.lu/ist/ka3/digicult/lund_p_browse.htm

Another approach is to use a series of concepts (**ontology**) instead of terms as a pivot, thus mapping the terms to concepts, with definitions and relationships (including mathematical representation of similarities). Those concepts are based on a metamodel (representation) of a given domain. As an example, the Harmonise project has built on ontology on tourism.
Ontologies can also be used for full text or image retrieval if linking strings or images features to concepts. Phraseology expresses lexical alternatives of a concept.

6.2 Full text and exploitation of natural language

Other techniques, less common in the cultural field can however help discovering digital cultural content in a multilingual environment. Those techniques may not always benefit from the highly structured metadata models in use in the heritage sector, but they may be coupled with the above mentioned terminologies. This is the case of language technologies and information retrieval.
Language processing technologies allow :

- to develop intelligent interfaces for natural language information retrieval and cross-language information retrieval
- text analysis for translation, text cartography, automatic summaries …
- multilingual generation of text

Many applications of language processing technologies, notably in the area of patents allow a less specialised public to access information. For patents, the system avoids low quality automatic translations, but rather extracts concepts and translates them according to an international taxonomy of 70.000 terms. Phrases are extracted and linked to a concept. As an example, it has been demonstrated that when, in a patent, it is written “it is desirable that …”, then it means that the following text is a solution. The text is tagged according to those information.
To apply a similar methodology to heritage resources, it does not appear useful to build complex thesaurus systems, but rather to generate multilingual versions of descriptive files. A project is being built with the museum sector on that basis.
Another strategy is to develop systems above research on **cross language information retrieval**, based on unstructured textual information. The Cross-language evaluation forum (CLEF), supported by DELOS (European network on digital libraries) gathers each year many labs from all over the world to compare results of queries in one language for corpora in one or more other languages. Those researches are mainly oriented towards unstructured content (for the Web). The algorithms work well and can reach 90% of the results gathered by a monolingual query. A major task is the manual evaluation of results.
A problem with that type of research is that if the user cannot search content in another language, he/she cannot understand the result neither. However, that difficulty does not exist with images, then the functionality seems more relevant to the users. Then, the EUROVISION project to retrieve images, thanks to captions has been extended as an exercise of cross-language evaluation. The EUROVISION project has started with a collection of images held in a library and has applied information retrieval algorithms to not only captions, but to the whole metadata set, although merged as a single element. Results reach 70-78% of monolingual queries.
Still, there is a limit to those results :

- several words are not understood by the automatic translator,
- in several cases, the image does not mean anything if not understanding captions and captions may be easier to translate automatically since they have a style simpler than

large texts. Moreover, the structured metadata have been merged and considered as a single caption. Future work could help taking benefit out of the structuring of those data and related terminologies.

Images retrieval can be supported by refine and expand functionalities when browsing results. The CIQUEST project (Concept-based Interactive QUery Expansion Support Tool)[8] has demonstrated this efficiency with images clusters.

6.3 Producing multilingual resources

Such as demonstrated above, a major problem for the end-user is to understand the resource displayed. It is important, when creating a cultural product or service, to take into account the possibility to localise a Website or part of a Website.

The Minerva quality principles refer to a minimum level of multilingual access to cultural Websites. Indeed, the European experts encourage the creation of at least a presentation page, metadata and a "guide" to the Website in at least one other language, different from the main language of the Website. However, translation is not enough to ensure usability of the same service in another country so that it appears very important to well identify specificities of European citizen browsing cultural Websites.

The Mudicu project (eContent) has started such a study with highly interactive content. It has notably demonstrated the necessity to localise symbols and all interfaces, not only translations of textual content (also true for multilingual thesaurus interfaces described by ICCD). The specificity of European citizens is important according to the degree of inclusion of the societies in the Internet (complex or simple presentations of information), the rate of large band-width equipment, also according cultural features such as perception of colours.

6.4 The role of European experts

Common European service shall take into account the progresses of existing experiments, notably with a focus on multilingual access to European common heritage.

For inventories of digitised content :

- ensure the possibility to record information in various languages
- provide the possibility to browse multilingual content through controlled terminologies translated in various languages
- test ontology-based access to descriptions
- provide online automatic translation to ensure a minimum level of access to collection descriptions.

More generally, the work of European experts should :

- ensure a good understanding of on-going research and their potential impact on the cultural heritage sector, notably with DELOS network on digital libraries;
- include multilingual access to programmes creating new heritage services, including at least metadata and a presentation page;
- disseminate knowledge on localisation of cultural Websites;
- disseminate good practices of multilingual cultural services;
- improve thesaurus translations and mappings;
- encourage the use of translated international terminologies.

[8] http://dis.shef.ac.uk/ciquest/

7. The European collaboration

The compliance with the guidelines and specifications described in this document should ensure interoperability of the system both with international related standards and other European services, semantic (metadata sets and terminologies), technical (open source systems) and organisational since the methodologies of such observatories shall be identical across Europe.

European services may be either professional : describing all entities, or largely related to resource discovery by end-user. The major issue would be therefore to build ***a portal to European digitised cultural resources*** for the largest possible audience.

7.1 Building OAI repositories

The possibility to implement export formats or OAI-PMH repositories[9] according to the schemas mentioned for collections, projects, services and institutions descriptions ensure the possibility to build metasearch tools, European gateways and possibly European catalogues in the future for European digital heritage discovery.
Schemas to be exposed are unqualified Dublin Core for each item, the schemas mentioned above for institutions, projects, physical collections, digital collections and services and DC Culture.
The "about" section of the records shall refer to standard **provenance schema**[10]. Another schema must be defined to allow **intellectual property and usage** statement on the descriptions[11].
From different systems exposed in Open Archives Repositories, a European platform can be set up to gather information on digitisation activities and digital resources from all over Europe.
This material should allow to build browsing interfaces according to different user profiles and according to various search scenarios.

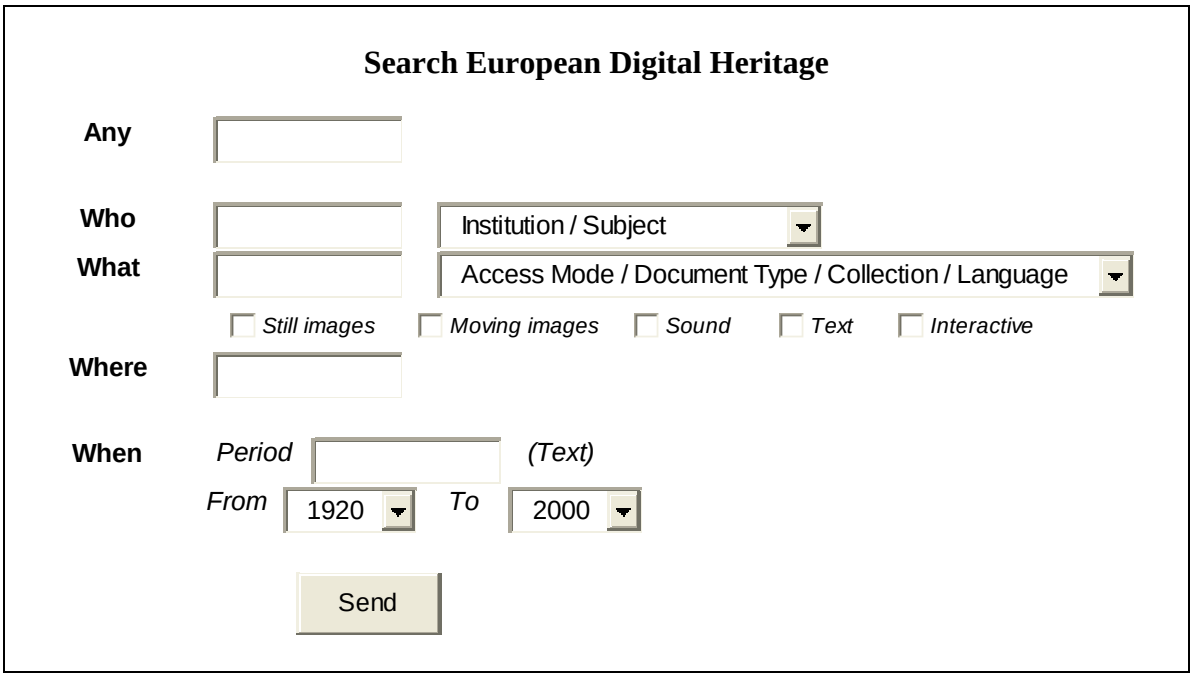
7.2 Various user profiles

It is important to build relevant interfaces to content which is partially heterogeneous, collected from diverse sources and which must be displayed in a comprehensible way for all partners.
For professional use, a model based on all entities and the access points described above can provide full information to European heritage professionals. The terminologies in use in all systems can optionally be mapped to the one proposed in this document so as to improve the quality of access points.
Another challenge is the access to the end-user looking for digital resources. The application type would therefore be related to portals of Internet resources and the unit would be the "access service", then described thanks to the other related entities. The DC Culture provides a logical path to set up those interfaces for end-users who are looking for digital resources on the following model :

9 http://www.openarchives.org

10 http://www.openarchives.org/OAI/2.0/guidelines-provenance.htm

11 ongoing work in the interoperability focus



The interface should allow to only search **resources which are accessible online,** otherwise, the professional interface would certainly be more adapted. The selection can therefore be done according to the service type element in original descriptions.

The search interface should allow to refer to **controlled terminologies** according to the common terminologies and possible crosswalks set by the different systems but it shall not be mandatory for content providers to use the terminologies. However, the system should translate the queries rather than proposing to select in a controlled list[12]. This allows to access resources through the controlled terminology even if a content provider has not set crosswalks. Thus, the reference to those terminologies are invisible to the user. For geographical names, interpretation by the system according to a thesaurus appears a very important feature.

The audience-driven approach has led the interoperability focus to define the **DC.Culture schema**[13], as an attempt to use a cross-domain schema based on DC elements with a clear focus on building valuable access points. The extraction would therefore mix the collection and service schemas, taking the "service" entity as a reference (if a service provides access to various collections, elements on collections are mixed, and a single DC.Culture record is created) and defined as following :

12 on this issue, see the work led by the Minerva interoperability focus and the deliverable 4.1.

13 http://www.minervaeurope.org/DC.Culture/XMLSchema/1.0

Equivalence to DC.Culture schema

DC Culture elements	Equivalence
Who	
Creator	*Project – institution in charge (institution name)*
Publisher	*Service – publisher*
Contributor	*Project – contributing institution (institution name)*
Rights	*Digital collection – Legal status and IPR*
What	
Title	*Service – title*
Description	*Physical collection - description* *Digital collection - subject* *Digital collection - description* *Service – description*
Type	*Digital collection – digital document types* *Service – type*
Format	*Physical collection – format* *Digital collection – format*
Identifier	*Service – location*
Source	*Physical collection – name*
Language	*Digital collection – language*
Relation	*Project – name* *Digital collection – has association* *Project – has association*
Where	*Digital collection – spatial coverage*
When	*Digital collection – temporal coverage*

7.3 Access to multilingual resources

As national system are built in various languages, the access to multilingual resources is a major issue to build a common platform.
The descriptions shall mention the language of resources if any and the **language of the description**. The system should accept that elements are partially translated and that the schema[14] include the value of an element in various languages according to the following model :

<dc:subject xml:lang=”FR”>archéologie</dc:subject>
<dc:subject xml:lang= “EN”>archaeology</dc:subject>

It is then possible to propose to the persons recoding information to provide descriptions in various languages for non controlled fields, such as in the Inventory of Canadian Digital Initiatives[15].

The standard terminologies used by the European platform shall be translated in each language of descriptions to allow resource discovery. The quality of information retrieval is certainly improved if equivalences are designed from each source language to each target language but the use of English as a **pivot language** for controlled terminologies and terminology equivalences appear as more feasible approach in the context of European enlargement. If referring to an internal **thesaurus**, then the element shall contain the whole path of the category such as in the following :

<dc:subject xml:lang>monument/catholic architecture</dc:subject>

14 all schemas can be modified to add the language attribute for the use of the European platform.

15 http://www.nlc-bnc.ca/initiatives-bin/cindi?Bilingual=Oui&project_nbr=0

The system for information retrieval, adopted from the above, which includes equivalence lists between terminologies is also valuable for multilingual information retrieval.

The European system shall translate part of the data, for example to set the equivalence between a period and a proper date interval (specific element only used by the European platform to ensure the connection of different systems).

All European **character sets** shall be managed by the platform, including Cyrillic and Latin characters, for example thanks to XML encoding (Unicode).

Cross language information retrieval techniques, based on internal dictionaries to translate either the descriptions or the queries[16], could successfully be applied to structured heritage resource descriptions. Even if using automatic translation, it is important to take into account the fact that peer to peer automatic translation does not exist for all language, so that it is difficult to avoid the use of one or various pivot language(s).

Finally, interfaces and user understanding are also a difficulty. When connecting resources from various countries, the results are produced in various languages, so that, even with multilingual retrieval functionalities, the user may not be able to understand content. Online **automatic translation** can help understanding descriptions[17]. Although free online softwares managing all European languages are rather rare and of unequal quality[18], a link to get the translation of the result automatically generated into English or another available language can allow to assist the end-user in European heritage retrieval.

Finally, the schemas used in the platform, when gathering descriptions, should perform :

- de-dupping
- add specific elements for date intervals and other elements with a specific use in the platform
- add elements on the content provider
- add information on last update of the description

The platform therefore collects RSLP, DC Collection and DC Culture type XML files enriched with the language attribute on each element, it then re-processes them and adds specific elements for internal use. Presentation of European Heritage shall take into account at least two interfaces types, possibly in all languages used in the directory, according to two user profiles : professional and all public portal, which should lead to the establishment of various applications.

In the future, this platform shall allow to search digital objects by collecting metadata from the online services described in the platform.

16 see Cross-Lingual Evaluation Forum workshop 2002 http://clef.iei.pi.cnr.it:2002/, notably activities on image corpora

17 such as in the TEL project

18 see a quite rich example http://www.tranexp.com:2000/InterTran?

Annex 1 List of references

Initiator	Directory	Address	Comment
NINCH	International Database of Digital Humanities Projects	http://www.ninch.org/programs/data/	Project-oriented, survey, prototype
University of Arizona	Clearinghouse of Image databases	http://www.library.arizona.edu/images/image_projects.html	Project-oriented, relies on a mailing list
Unesco/IFLA	Unesco / IFLA directory of digitized collections	http:// www.unesco.org/webworld/digicol/	Collection-oriented, aim at creating a virtual library - Mostly with IFLA : only provides a reference to the collection.
Australian National Library	Australian digitisation projects	http:// www.nla.gov.au/libraries/digitisation/	
Canadian National Library	Inventory of digitised collections	http://www.nlc-bnc.ca/initiatives/	
Information Highway Applications Branch Industry Canada	Canada's digital collections	http://collections.ic.gc.ca/e/index.php	Programme providing a gateway to its collections – To build Canadian treasures on the Web
Digital Libraries Federation	Public Access Collection	http://www.hti.umich.edu/cgi/b/bib/bib-idx?c=dlfcoll	Includes a reference to source collections
Association of Research Libraries (US)	Digital Initiatives Database of the	http://www.arl.org/did/	Link to Websites
French Minsitry of Culture & Coimmunication	French directory of digitised collections	http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f_02.htm	Points to national databases and other information sources For general public and professionals
Spanish Ministry of Science and Technology	IFIgenia catalogue of cultural digitised collections	To open in 2003	Survey For general public and professionals
UK New Opportunities Fund	New Opportunities Fund portal	http://www.enrichuk.net/	Link to Websites For the general public
UK Technical Advisory service for Images	TASI Image sites	http:// www.tasi.ac.uk/imagesites/images.html	Few information are published. Relies upon a competence centre Professional audience
Kulturnet Sweden	digitaliseringprojekt inom arkiv, bibliotek och museer / ABM-området	http://www.kultur.nu/rapporter/digitaliseringsprojekt.html	Survey
State and University Library of Lower Saxony	Digitised European Periodicals	http://gdz.sub.uni-goettingen.de/dieper/	For periodicals, European project based on 10 major libraries in Europe. Database gathered with EROMM http://www.eromm.org, access not free
Göttinger Digitalizierungs Zentrum	DGF Projects	http://gdz.sub.uni-goettingen.de/en/vdf-e/	Funded Projects in the Deutsche Forschunsgemeinschaft Programme

Annex 2 Model/example of questionnaire to collect information

Corresponding schema element is highlighted close to the field name.

Institution

possibility to multiply number of institutions

Name : …… name ……..…….....……

Address : vcard/adr.street.....
.....

Zipcode : .... **vcard/adr.pcode**.... **City :** .. **vcard/adr.locality** .....
Region…… **vcard/adr.region** ……

Telefone: +39 (0) **vcard/tel.voice**	**Fax :** +39 (0) **vcard/tel.fax**	**E-mail : vcard/email** @
Website: vcard/URL		

Legal status : legalstatus

☐ Public ☐ Commercial ☐ Non profit

Institution type : Type

☐ library ☐ museum ☐ local community / administration ☐ Academic
☐ archive ☐ religious ☐ research centre ☐ media (TV, radio ....)
☐ professional association
☐ other specify : ……

Jurisdiction : Jurisdiction

☐ Ministry of culture ☐ Local community
other : ……

" if there are additional institutions associated to the project, please fill the file again"

Project

Title name

Description description

Status status

☐ completed ☐ on-going ☐ planned

dateStart **dateEnd**

Starting date (year) :…… Date of completion (planned – year) :……..…

Funding programme funding

☐ Nation digitisation programme of the French ministry of culture ☐ Regional programme ☐ European commission
☐ Academic …
Other : ……

Digitisation mode digitisation process

☐ Direct	☐ Indirect - Microfiche	☐ Indirect - Positive transparent color Phototype
☐ Indirect - Window card (carte à fenêtre)	☐ Indirect - Microfilm	☐ Indirect - Positive transparent B&W Phototype
☐ Indirect - Audio disc	☐ Indirect - Negative transparent color Phototype	☐ Indirect - Print run (tirage papier)
☐ Indirect - Film	☐ Indirect - Negative transparent B&W Phototype	☐ Indirect - Video
☐ Indirect - Tape recording	☐ Indirect - Microfiche	☐ Indirect - Tape recording

Other : ……

Performance of digisation digitisation process

☐ Internal

☐ External

Contact information

Name: .....**contact vcard fn**.....

Telefone : +39 (0) **contact vcard tel voice** **E-mail : contact vcard email** @

Description of digital collections

Name of digital collection : .. **name** .....

Description: description

Language language.....

Subject subject

Archaeology	Buildings, Architecture & Monuments	Cinema & audiovisual
Contemporary art	Costume & Textiles	Decorative and applied arts
Economics	Ethnology & anthropology	Fine arts
Geography	Health & medicine	Industry & commerce
Landscapes	Law	Literature
Local history	Music	Performing arts
Philosophy	Politics	Religion & Beliefs
Science & Technology	Transports	Urbanism & planning (regional, town, country)

Date temporal coverage

☐ Paleolithic	☐ III millennium bc	☐ VI c. bc	☐ I c	☐ VII c	☐ XIII c	☐ XIX c	
☐ Mesolithic	☐ II millennium bc	☐ V c. bc	☐ II c	☐ VIII c	☐ XIV c	☐ XX c	
☐ Neolithic	☐ I millennium bc	☐ IV c. bc	☐ III c	☐ IX c	☐ XV c	☐ XXI c	
☐ Calcolithic	☐ IX c. bc.	☐ III c. bc	☐ IV c	☐ X c.	☐ XVI c	☐	
☐ Bronze age	☐ VIII c. bc	☐ II c. bc	☐ V c	☐ XI c	☐ XVII c	☐	
☐ Iron Age	☐ VII c. bc	☐ I c. bc	☐ VI c	☐ XII c	☐ XVIII c	☐	

other : ……

Zone geografiche spatial coverage : all regions

☐ Abruzzo	☐ Lazio	☐ Puglia	☐ Valle d'Aosta
☐ Basilicata	☐ Liguria	☐ Sardegna	☐ Veneto
☐ Calabria	☐ Lombardia	☐ Sicilia	
☐ Campania	☐ Marche	☐ Toscana	
☐ Emilia Romagna	☐ Molise	☐ Trentino Alto Adige	
☐ Friuli Venezia Giulia	☐ Piemonte	☐ Umbria	☐ Italia as a whole

Other European country

Austria	Denmark	Netherlands	Malta	Spain
Belgium	Estonia	Hungary	Norway	Sweden
Great Britain	Finland	Iceland	Poland	Switzerland
Croatia	France	Ireland	Portugal	
Cyprus	Germany	Italy	Romania	
Czech Republic	Greece	Luxemburg	Slovenia	

☐ Central Asia	☐ South America	☐ Middle East
☐ Northern America	☐ Northern Africa	☐ Far East
☐ Central America	☐ Sub-saharian Africa	☐ Oceania

Standard **standard**

☐ DTD EAD	☐ Dublin Core Culture
☐ AMICO-DTD	☐ METS
☐ MARC-XML	☐ MODS
☐ DTD Biblio-ML	☐ LCSH
☐ SPECTRUM-XML	☐ AAT
☐ TEI	☐ TGN
☐ Dublin Core	☐ Garnier thesaurus
☐ MARC	☐
☐ CDWA	☐
☐ ISAD (G)	☐

other : ……

Types Type	**and formats Format**	
☐ Text (OCR)	ASCII PDF XML	Other format : ……
☐ Text (image)	PDF TIFF JPEG/JFIF PNG	Other format : ……
☐ Still image	TIFF BMP JPEG/JFIF GIF SVG PNG PDF	Other format : ……
☐ Vectorial map / drawing	DXF SVG	Other format : ……
☐ Moving image	MPEG3 AVI video QTVR	Other format : ……
☐ Sound	WAV MP3 Sun audio Real audio MPG	Other format : ……
☐ 3D		
☐ Interactive resources	SVG Macromedia Macromedia Flash	

	Current size	planned accrualstatus
Number of images, pages, hours of sound or video	……	……

IPR and legal status

IPR holder ……

Usage rights granted……

Associated collection Associated collection

Collection title ……...

Illustration(s) Illustration

Limited to 3 illustrations. Images must be in 192 x 128 pixels, at 72 dpi definition
Please mention caption, photograph and copyright :
Illustration 1 : .....
Illustration 2 : .....
Illustration 3 : .....

Associated physical collection

possibility to multiply number of physical collections

Name

Description :

Physical document format Physical document format

	Books		Sound track
	Articles		Cadaster
	Newspapers		Archaeological site
	Serials		Monument
	Yearbooks		incunabulum
	Manuscripts		Installation
	Bibliographic records		Musical instrument
	Numerical / statistical data		Music score
	Archival records		painting
	Notes, coins and medals		jewelry
	Maps		Sculpture
	Moving image		Textile
	Printed Images		Illuminated manuscripts
	Drawings		
	Engravings		
	Photography		
	postcard		

Institution location
Institution

Access service and product

Name

Description： **.....service description**

.....

.....

Service type
service type

☐ CDRom	☐ Interactive terminal	☐ Virtual exhibition / gallery
☐ DVDRom	☐ Online database	☐ Offline database
☐ Portal	☐ Website	☐ Learning service

Location

URL： ……

Technical requirements

Technical requirements for accessibility

☐ WAI compliance
☐ Plug-ins requirements……

Communication protocol Communication protocol

☐ Z39.50
☐ OAI-PMH
☐ Other

Audience

Target Audience

☐ formal education
☐ life-long learning
☐ professional
☐ leisure / tourism
☐ academic research
☐ general public
☐ chidren

Publisher

Publisher .....