LoCloud

Local content in a Europeana cloud

D1.2: Definition of Metadata Schemas

Authors:
Aitor Soroa, University of the Basque Country (EHU)
Stavros Angelis, Athena Research and Innovation Centre (Athena-RC)
Kostas Pardalis, National Technical University of Athens (NTUA)

Version: Final

LoCloud is funded by the European Commission's
ICT Policy Support Programme

cip competitiveness and innovation framework programme 2007–2013

Revision History

Revision	Date	Author	Organisation	Description
V0.1	2013/07/24	Aitor Soroa, StavrosAngelis, KostasPardavis	EHU, DCU, NTUA	First draft
V0.2	2013/07/31	AitorSoroa	EHU	Include comments from Costis Dallas.
V1.	2013/09/24	Aitor Soroa, StavrosAngelis, KostasPardavis	EHU, DCU, NTUA	Include comments from Kate Fernie and Vangelis Banos
V2.	2014/05/13	AitorSoroa	EHU	Update

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Contents

1. Executive Summary

This deliverable has been produced as part of LoCloud Work Package 1: Planning, preparation and requirements.

A working team including UPV/EHU, ARC and NTUA has been working together to define the metadata schemas to be used in LoCloud as intermediaries to EDM. The team has reviewed the schemas used in prior projects, such as Europeana Local, Athena and CARARE and has defined work needed to support their implementation by LoCloud.

One of the LoCloud's main objectives is to ensure interoperability between the native metadata held by heritage organizations and the metadata used by Europeana. This deliverable describes the identification of the metadata schemas to be used within the project as intermediaries to Europeana Data Model (EDM), the schema introduced by Europeana to improve on ESE, its first data model[1].

EDM is a comprehensive, general-purpose model which accommodates the range and richness of community standards such as LIDO for museums or EAD for archives. The current implementation of EDM by Europeana is part of a roadmap which covers the transition from the initial ESE schema to full implementation of EDM. It is likely that the detail of EDM implementation will change over the next four years.

Following past experiences on domain aggregation services, such as CARARE, Athena and MIMO projects, LoCloud will bridge the gap between the metadata schemas in use by their content partners and the data model being developed by Europeana by establishing or utilizing intermediary schemas and then mapping these to the metadata standard implemented by Europeana. This mapping has enabled harvested metadata to be transformed to EDM for supply to Europeana.

The benefits of this approach are:

- providing an intermediary schema accommodates the need for guidance which is relevant to the needs of a community and the particular characteristics of their data;
- mapping the intermediary schema to EDM builds in flexibility and accommodates any changes by Europeana as it phases in the full EDM model. This approach will be continued and developed in LoCloud.

This deliverable has been revised and updated so that it now takes into account the results of the content survey and planning, and the results of the requirements analysis (deliverables D1.3 and D1.5). The update is described in Section 6.

[1]See http://pro.europeana.eu/documents/900548/f495317b-4557-4a60-9326-723f4618b44c

2. Methodology

In this Section we describe the methodology followed to identify metadata schemas to be used within LoCLoud as intermediaries to EDM. From the beginning of the process, two main choices were considered regarding the role of the intermediate schema in the project. One option is to offer partners a single, big intermediary schema which would include (almost) all the information that content providers want to aggregate. The other option is to provide two or three familiar schemas and let the content providers choose amongst them. Using a single intermediate schema would greatly simplify the aggregation infrastructure, since all native schemas would be integrated to one intermediate schema, which in turn would be finally transformed to EDM. On other hand, the gap between the native schemas used by the partners and this intermediary schema could be too large. Also, local providers usually prefer to support the main domain schemas rather than attempting to define a super-schema. We thus decided to offer a candidate set of three of four choices as intermediate schemas.

In order to achieve this objective, the work has been divided in two main tasks: a state of the art review and a specific survey of LoCloud partners. Thanks to these two main tasks and their respective results, we have made a common recommendation on intermediate schemas. This takes into account the survey and a comparative analysisof all the information sources we have identified through the state of the art.

Additionally, WP1 also conducted a survey (described in Deliverable D1.3) to evaluate and appraise content and metadata among collection holding organisations participating in LoCLoud with regard to fitness-for-purpose, completeness and quality. As a result, we have a clear picture of the kind of cultural information that will be ingested into Europeana, which will in turn help assessing the suitability of the intermediate schemas in the process.

2.1. Study on state of the art on metadata schemas

The state of the art parses a set of existing projects and initiatives regarding metadata standards. The goal is to identify standard metadata export formats commonly used in the information systems deployed by the cultural institutions. This study will give an exposition of a wide range of standards being used by cultural institutions, thus providing a "landscape" of the current situation. The analysis of each of them has allowed us to pick up among their results and experiences what can enrich our own study about the use of metadata within the project.

2.2. Specific survey

We conducted a little survey in order to gather information from the content partners to help us on making an informed decision about the intermediate schemas to choose. In the survey the partners were asked to choose an intermediate schema from a selected set of three possible choices. The aim of the survey was to show a clear picture of the preferences among partners regarding which intermediate schema is more convenient for them.

3. Existing Schemas

Many institutions have been working towards the development of standards to make their documentation uniform, in order to systematically document, keep and consult objects and archive records, not only for the physical preservation of the Cultural Heritage (CH)asset but also for the preservation of its related information for future reference.

These institutions have defined guidelines and indications helping in the gathering of information about an asset, such as adopting thesauri and controlled vocabularies for the standardization of the terms. Some of these data standards have been defined within a national framework, such as the ICCD schema (Italy), the MIDAS standard (England); others aim at guaranteeing data interoperability, such as LIDO and the EuScreen schema, which are CIDOC-CRM and EBUCore compliant respectively.

The rapid growth of Internet resources and digital collections has been accompanied by a proliferation of metadata schemas, each of which has been designed according to the requirements of particular user communities, intended users, types of materials, subject domains, project needs, etc. Several National Bodies and Cultural Institution that deal with cataloguing and documentation of Cultural Heritage developed different forms and standards. In the following section we summarize the most important.

3.1. ICCD

The ICCD [2]defines a number of standards and tools for the Cataloguing and Documentation of Italian national archaeological, architectural, art, history and ethno-anthropological heritage. These standards have been defined in agreement with the Italian Regions. These standards and best practices are related but not limited to, photographic documentation for catalogue recording[3], application of norms on digitization of photographic images[4], norms on digitization of graphic images[5], norms for multimedia documentation[6] and norms on data transfer[7].

3.2. MIDAS Heritage standard

The Midas Heritage standard[8] is a data standard for information about the historic environment, which is developed and maintained by English Heritage, for and on behalf of the UK Forum on Information Standards in Heritage (FISH).

[2]http://www.iccd.beniculturali.it/index.php?en/95/institute
[3]http://www.iccd.beniculturali.it/getFile.php?id=292
[4]http://www.iccd.beniculturali.it/getFile.php?id=290
[5]http://www.iccd.beniculturali.it/getFile.php?id=291
[6]http://www.iccd.beniculturali.it/getFile.php?id=293
[7]http://www.iccd.beniculturali.it/getFile.php?id=160
[8]http://www.english-heritage.org.uk/publications/midas-heritage/

3.3. VRA Core

VRA core[9] is a data standard that consists of a metadata element set (units of information such as title, location, date, etc.), as well as an initial blueprint for how those elements can be hierarchically structured. The element set provides a categorical organization for the description of works of visual culture as well as the images that document them. VRA core is developed and maintained by the VRA Core Oversight Committee[10].

3.4. CIDOC CRM

The CIDOC CRM[11] is a formal ontology and an ISO[12], that defines CH documentation concepts and the relationship between them, used to clarify the documentation process, and to ensure no loss of semantic content when integrating heterogeneous Cultural Heritage data sources. More specifically CIDOC-CRM is intended to promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework to which any cultural heritage information can be mapped. It is intended to be a common language for domain experts and implementers to formulate requirements for information systems and to serve as a guide for good practice in conceptual modeling. In this way it can provide the "semantic glue" needed to mediate between different sources of cultural heritage information, such as that published by museums, libraries and archives.

3.5. EBUCore

EBUCore[13] has been purposefully designed as a minimum and flexible list of attributes to describe audio and video resources for a wide range of broadcasting applications including archives, exchange and production in the context of a Service Oriented Architecture. It is also a metadata schema with well-defined syntax and semantics for easier implementation. EBUCore is based on the Dublin Core to maximise interoperability with the community of Dublin Core users such as the European Digital Library 'Europeana'. EBUCore expands the list of elements originally defined in EBU Tech 3293-2001 for radio archives, also based on Dublin Core. EBUCore 1.3 takes into account latest developments in the Semantic Web and Linked Data community. EBUCore 1.3 is available as a RDF ontology entirely compatible with the W3C Media Annotation Working Group ontology, which model is common and based on the EBU Class Conceptual Data Model. EBUCore has been developed and maintained by the European Broadcasting Union (EBU) and it is mainly used by broadcasters and audiovisual archives.

3.6. MPEG-7

The MPEG-7 standard, formally named "f", provides a rich set of standardized tools to describe multimedia content. MPEG-7 standardizes so-called "description tools" for multimedia content: Descriptors (Ds), Description Schemes (DSs) and the relationships between them. Descriptors are

[9] http://www.loc.gov/standards/vracore/
[10] vracore@vraweb.org
[11] http://www.cidoc-crm.org/
[12] http://www.iso.org/iso/catalogue_detail?csnumber=34424
[13] http://tech.ebu.ch/lang/en/MetadataEbuCore

used to represent specific features of the content, generally low-level features such as visual (e.g. texture, camera motion) or audio (e.g. melody), while description schemes refer to more abstract description entities (usually a set of descriptors). These description tools as well as their relationships are represented using the Description Definition Language (DDL), a core part of the language. Both human users and automatic systems that process audiovisual information are within the scope of MPEG-7[14].

3.7. MPEG-21

The MPEG-21[15] standard aims at defining a framework for multimedia delivery and consumption which supports a variety of businesses engaged in the trading of digital objects. The MPEG-21 standard is focusing on filling the gaps in the multimedia delivery chain. MPEG-21 was developed with the vision in mind that it should offer users transparent and interoperable consumption and delivery of rich multimedia content. The MPEG-21 standard consists of a set of tools and builds on its previous coding and metadata standards like MPEG-1, -2, -4 and -7, i.e. it links them together to produce a protectable universal package for collecting, relating, referencing and structuring multimedia content for the consumption by users (the digital item). The vision of MPEG-21 is to enable transparent and augmented use of multimedia resources (e.g. Music tracks, videos, text documents or physical objects) contained in digital items across a wide range of networks and devices.

3.8. DUBLIN CORE

The Dublin Core Metadata Element Set[16] was proposed as a minimum number of metadata elements required to facilitate the creation of simple descriptive records for electronic documents. The set consists of a flat list of fifteen elements describing common properties of resources. To promote global interoperability, a number of the element descriptions may be associated with a controlled vocabulary for the respective element values. It is assumed that other controlled vocabularies will be developed for interoperability within certain local domains. Dublin Core is the result of the Dublin Core Metadata Initiative (DCMI)[17].

3.9. SPECTRUM

SPECTRUM[18] documentation standard is more than a metadata schema used mainly in museums. It is a guide to documenting all procedures a museum might need to undertake in managing its collections (e.g. acquisition, cataloguing, auditing, and loans). SPECTRUM recommends several “units of information” that can be recorded to support each of these procedures, some of which are required, others recommended. In terms of cataloguing museum objects, SPECTRUM suggests that sometimes it will be appropriate to catalogue at a collection level, at other times, at the item level. It suggests that any catalogue record should include at least: an identity number, name of the

[14]http://mpeg.chiariglione.org/standards/mpeg-7
[15]http://mpeg.chiariglione.org/standards/mpeg-21
[16]http://dublincore.org/
[17]http://dublincore.org/about-us/
[18]http://www.collectionslink.org.uk/spectrum-standard

object, number of items or parts, physical description, and details about its acquisition, location and any associated images. SPECTRUM does not prescribe particular elements for digital reproductions, so those developing museum collection management systems often use SPECTRUM as the basis for the object information and Dublin Core to record information about any associated digital images. SPECTRUM has been developed and is maintained by Collections Trust[19] and its members.

3.10. CATEGORIES FOR THE DESCRIPTION OF WORKS OF ART (CDWA)

The CDWA[20] describes the content of art databases by articulating a conceptual framework for describing and accessing information about works of art, architecture, other material culture, groups and collections of works, and related images. The CDWA includes 381 categories and subcategories. A small subset of categories are considered core in that they represent the minimum information necessary to identify and describe a work. CDWA has been developed and maintained by the GETTY institute.

3.11. MACHINE-READABLE CATALOGUING (MARC21)

MARC21[21] is a format standard for the storage and exchange of bibliographic records and related information in machine-readable form. Its format supported by the majority of library systems and offers participation in an international bibliographic community following common standards, and the advantage of copy cataloguing at much reduced cost and with no need to maintain conversion programs.

3.12. INTERNATIONAL PRESS TELECOMMUNICATIONS COUNCIL (IPTC)

IPTC Core[22] is a set of metadata primarily for digital images to be used by Adobe's Extensible Metadata Platform XMP. IPTC metadata were employed by Adobe Systems Inc. to describe photos already in the early nineties. A subset of the IPTC "Information Interchange Model - IIM" was adopted as the well-known "IPTC Headers" for Photoshop, JPEG and TIFF image files which currently describe millions of professional digital photos.

3.13. EUScreen schema

The EUScreen is harvesting schema developed by the EUscreen (www.euscreen.eu) project for harvesting multimedia metadata into the service environment of Europeana. It was implemented based on EBUCore, which is an established standard in the area of audiovisual metadata. An extensive evaluation of alternative standards in this area (MPEG7, DCMI, TV Anytime) was conducted before choosing the EBUCore.

[19] http://www.collectionslink.org.uk/index.php
[20] http://www.getty.edu/research/publications/electronic_publications/cdwa/
[21] http://www.loc.gov/marc/bibliographic/
[22] http://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata-201007_1.pdf

3.14. METS

Metadata Encoding and Transmission Standard schema[23] is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation.

3.15. EAD

Encoded Archival Description[24] is an XML standard for encoding archival finding aids, maintained by the Library of Congress in partnership with the Society of American Archivists. The EAD standard's document type definition (DTD) specifies the elements to be used to describe a manuscript collection as well as the arrangement of those elements (for example, which elements are required, or which are permitted inside which other elements). EAD 1.0 was an SGML DTD; EAD 2002, the second and current incarnation of EAD, was finalized in December 2002 and is an XML DTD. The EAD tag set has 146 elements and is used both to describe a collection as a whole, and also to encode a detailed multi-level inventory of the collection. Many EAD elements have been, or can be, mapped to content standards (such as DACS and ISAD(G)) and other structural standards (such as MARC or Dublin Core), increasing the flexibility and interoperability of the data.

3.16. Europeana Semantic Elements (ESE)

ESE is a Dublin Core-based schema that provides a generic set of elements (mostly derived from Dublin Core DC), specifically added to support the description in the cultural heritage domain as needed by Europeana. ESE records are flat objects and its expressiveness is considered to be low. For instance, it is not possible in ESE to distinguish information about the original cultural heritage objects from its digital representation. Therefore, Europeana is moving from ESE to a more expressive metadata schema (EDM, see Section 3.17).

3.17. The Europeana Data Model

The Europeana Data Model (EDM)[25]is designed to provide an integration medium for collecting, connecting and enriching the descriptions provided by Europeana content providers. It is a major improvement on the Europeana Semantic Elements (ESE), the basic data model that Europeana began life with. Each of the different heritage sectors represented in Europeana uses different data standards, and ESE reduced these to the lowest common denominator. EDM reverses this reductive approach and is an attempt to transcend the respective information perspectives of the sectors that are represented in Europeana –the museums, archives, audio-visual collections and libraries. EDM is not built on any particular community standard but rather adopts an open, cross-domain Semantic Web-based framework that can accommodate the range and richness of

[23]http://www.loc.gov/standards/mets/
[24]http://www.loc.gov/ead/
[25]http://pro.europeana.eu/edm-documentation

particular community standards such as LIDO for museums, EAD for archives or METS for digital libraries.

3.18. The CARARE Schema

CARARE schema is a metadata schema based on existing standards from the archaeology and architecture domain. The CARARE metadata schema was released in autumn 2010 and updated in spring 2011 following testing by content partners.

The CARARE metadata schema is a harvesting schema intended for delivering metadata to the CARARE service environment about an organization's online collections, monument inventory database and digital objects. The strength of the schema lies with its ability to support the full range of descriptive information about monuments, building, landscape areas and their representations. It does not support activities such as monument management and protection.

The focus of the CARARE schema is on the detailed description of heritage assets (monuments, buildings, landscapes or artifacts) and related digital resources and events in which the heritage asset is represented. The Schema is based on MIDAS Heritage with additional elements from LIDO and ESE (Europeana Semantic Elements) to cover the information needed for the digital resources being made accessible to the CARARE and Europeana service environments.

The root element of a CARARE record is the CARARE wrap, which wraps one or many CARARE records. The CARARE schema's core is:

- 4 themes (Heritage Asset, Digital Resource, Activity and Collection information) and
- a series of global types (record information, spatial, temporal, rights, appellation, etc) which may be used across the schema to define the information elements.
- The schema specifies whether themes and elements are mandatory, strongly recommended or optional.

The four major concepts which are wrapped into a main entity – the CARARE record - are:

- Heritage asset – this includes descriptive and administrative metadata for archaeological monuments, historic buildings, industrial monuments, archaeological landscape areas, shipwreck, artifacts and ecofacts. This entity is unique in the CARARE record.
- Digital resource – this includes metadata for images, drawings, plans, maps, archives, publications and 3D models representing a heritage asset.
- Collection information – this describes the collection which holds the content being provided.
- Activity – this includes metadata for both historical events which took place at the heritage assets (such as building, alternation, demolition, battles, etc) and archaeological events (such as excavations, surveys, etc).

In addition the CARARE schema allows, by means of specific element values, identification of additional place and agent entities which can be used to document contextual information

related to a specific site or monument. CARARE entities can be explicitly related through appropriate element values to denote a relationship between heritage assets (for instance, in the case one asset is part of another asset, which is typical in architectural and archaeological heritage), or a relationship between a heritage asset and a digital resource representing it. The CARARE Schema has been extended in the context of the 3d-Icons projects to version 2.0.[26][27]

3.19. LIDO Schema

LIDO is the result of a collaborative effort of international stakeholders in the museum sector, starting in 2008, to create a common solution for contributing cultural heritage content to portals and other repositories of aggregated resources. Being an application of the CIDOC Conceptual Reference Model (CRM) it provides an explicit format to deliver museums' object information in a standardized way. Led by the CDWA Lite Advisory Committee and the Documentation Committee of the German Museums Association, it was agreed to create a single schema that met the requirements articulated by CDWA Lite, museumdat, and feedback received from the greater community of information and technology professionals. As part of this effort, compliance with CIDOC-CRM was a major requirement. A working group was established for the development of LIDO. Resulting from the report on existing standards applied in European museums, it was concluded, within the ATHENA project, that a metadata format for ATHENA would have to meet the needs of both museumdat and SPECTRUM. Consequently ATHENA decided to join the LIDO initiative and support further development that would subsequently integrate SPECTRUM requirements into the LIDO schema. Since then it has been used very effectively in many Europeana feeder projects like Athena[28], Linked Heritage[29] and JUDAICA[30] while extensive documentation can be found for it at http://network.icom.museum/cidoc/working-groups/data-harvesting-and-interchange/what-is-lido/ .

LIDO can be used for delivering metadata, for use in a variety of online services, from an organization's online collections database to portals of aggregated resources – as well as exposing, sharing and connecting data on the web. In addition it is intended to represent the full range of descriptive information about museum objects, e.g. art, cultural, technology and natural science while it supports multilingual environments. Similar characteristics are offered in a simple and effective representation by LIDO in contrast to other metadata schemas (like MPEG7, EBUCore, EADS) that are quite complex to use.

[26] http://3dicons-project.eu/eng/content/download/2382/17896/file/The%20CARARE%20metadata%20schema2.pdf
[27] http://3dicons-project.eu/eng/content/download/2383/17901/file/carare-v2.0.1.xsd
[28] http://www.athenaeurope.org/ ATHENA (ECP-2007-DILI-517005), Access to cultural heritage networks across Europe.
[29] http://www.linkedheritage.eu/ Linked Heritage (270905), Coordination of Standards and Technologies for the enrichment of Europeana.
[30] http://www.judaica-europeana.eu/

LIDO only requires 4 mandatory elements (ObjectWorkType, RecordID, RecordSource, Title) and in that way the data providers can decide on how light – or how rich – they want their contributed metadata records to be, while also allowing for delivering data and resources relating to their objects. LIDO can also include links from contributed metadata back to records in the providers' 'home' context while it allows for identification of each referenced entity, e.g. provide references to controlled vocabulary and authority files. In general LIDO structure is divided in the Descriptive and administrative information groups, the first includes metadata about the cultural object while the second one administrative metadata. In addition another very important characteristic of LIDO is that it has been extensively used for delivering content to Europeana. Hence mapping between LIDO and ESE is available at Athena project website (http://www.athenaeurope.org/index.php?en/1/home). A mapping between LIDO and EDM can be automatically achieved since the mapping between ESE and EDM has been made by Europeana.

4. Criteria for selecting intermediate schemas

The main criterion for selecting intermediate schemas is for the intermediate schemas to be *general enough* (C1) and suitable to contain all original information from the content providers' collections with *minimum (preferably none) information loss* (C2). The intermediate schemas should be *well accepted* by the community and the content providers. The intermediate schemas should be *domain-specific* (C3), meaning that their coverage should extend to the appropriate domains of the content providers' information and the nature of their collections. They should also be able to *capture specific metadata* (C4) like geographical information, semantic relations, preservation specific information etc.

The first and most important step in the process of selecting intermediate schemas is to know the types of the content providers' resources, as most schemas have been developed to better handle specific types of resources. *Backwards compatibility* (C5) between the current version of an intermediate schema and possible past versions should exist; as this would greatly assist in easily getting information from providers that have their collections described using some past version. The *ease to create mappings* (C6) is also an important criterion. These mappings should be able to work in both ends, from the source content providers' schema towards the intermediate schemas, as well as from the intermediate schemas to EDM, with a requirement for minimum information loss. An optional requirement would be for the intermediate schemas to share a common structure and characteristics, or are expressed in a similar way (e.g. XML), or in the absence of a common structure, good known crosswalks.

A good starting point is to review schemas that have been successfully used and validated in earlier projects, like CARARE, ATHENA, MIMO, etc. Finally we should determine the content providers' staff expertise regarding the work with the metadata schemas and provide them with intermediate schemas appropriate for their level of familiarization.

5. Proposal for Intermediate Schemas

5.1. Comparative analysis

In this section we do a comparative analysis of the metadata schemas described in Section 4. The purpose of this analysis is to assess the quality and suitability of these schemas to be used as intermediate schemas within LoCloud.

The analysis is depicted in Table 1. Rows correspond to schemas and columns correspond to the selection criteria as defined in Section 5. In summary, these criteria are as follow[31]:

- C1: Schema generality.
- C2: suitable to contain all original information from the content providers' collections with minimum (preferably none) information loss.
- C3: Domain-specific.
- C4: Ability to capture specific metadata.
- C6: Ease to create mappings to EDM.

Schema	C1	C2	C3	C4	C6
ICCD	-	X	X	-	-
MIDAS	-	X	X	X	-
VRA Core	-	X	X	X	-
CIDOC CRM	X	X	X	X	-
EBUCore	-	-	X	X	-
MPEG-7	-	-	-	-	-
MPEG-21	-	-	-	-	-
Dublin Core	X	-	-	X	X
SPECTRUM	-	X	X	X	-
CDWA	-	X	X	-	-
MARC21	-	X	X	-	-
IPTC	-	-	X	-	-

[31]We excluded criteria C5 “backwards compatibility” from the table because we did not find the information of whether records that are well-formed according to anearlier version of a schema "break" under new versions of the schema.

Schema	C1	C2	C3	C4	C6
EUScreen	-	-	X	-	-
METS	-	-	-	-	-
EAD	-	X	X	-	X
ESE	-	-	-	X	X
EDM	X	-	-	X	X
CARARE	X	X	X	X	X
LIDO	X	-	X	X	X

Table 1 Comparative analysis of Schemas

The results of the analysis show CARARE and LIDO as good candidates for acting as intermediate schemas into LoCloud. The study suggests that the ESE schema is a poor intermediate schema, because it is not rich enough and lacks the ability of representing specific metadata information. Nevertheless, the partners involved in the task decided to include ESE as a choice offered to partners in the survey (5.2 below). The reasons for this choice were twofold. First, there was a general feeling that ESE schema has been widely used among the content providers within the project. Secondly, if ESE were used as intermediate schema it would cause little impact into LoCloud's aggregation workflow, because the mappings between ESE and EDM are available.

There is no other schema, among those considered, which completely fulfils the project's requirements.

MARC21 and EAD schemas are also good candidates, especially for library and archival materials. However, we need more detailed information regarding collections to be contributed to LoCloud in order to make a decision on whether to recommend them as intermediaries.

5.2. Specific Survey

Based on the analysis of the previous Section, we conducted a survey among the content providers and ask them to describe their preferences regarding intermediate schemas. In the survey we asked the partners to choose from a selected set of three possible schemas:

- The CARARE schema (see Section 3.18).
- The LIDO schema (see Section 3.19).
- ESE (see Section3.16).

We also allowed partners to choose any other schema of choice, so that we do not discard some important schema used by providers and missing from the state of the art survey.

Almost all providers (20 out from 23 providers) responded to the survey. In summary, CARARE got the majority of votes (9), followed by LIDO (4) and ESE (4). One partner chose EDM.

Some partners raised some important concerns on adopting ESE as intermediate schema. The concerns could be summarized as follow:

- ESE is good for helping people find information, but not for making quality metadata or for exchanging data. Most systems contain more detailed information than ESE can represent in particular geospatial and temporal data. First scaling down to ESE and then expanding again to EDM in Europeana does not make much sense. Intermediary schema should be a rich schema that can render the source data as well as possible (without loss of data).
- ESE was designed to feed the Europeana website and is not very useful outside Europeana.
- ESE will soon entirely be replaced by EDM at Europeana, and will have no further use; conversion from ESE will only be offered for a while for backward compatibility reasons.

Partners also raised some concerns about using EDM as intermediate schema, as EDM is still under development and it is likely that the detail of EDM implementation will change over the next four years, which would render the mappings out of date.

5.3. Recommendationfor Intermediate Schemas

Based on the previous sectionsand the outcomes of content survey carried out in Work Package 1, whose aim is to appraise content and metadata among organizations participating in LoCLoud, we recommend the following intermediate schemas:

1. **For material which is moveable** (like museum items, etc.),**the use of the LIDO schema as intermediate schema.**
2. **For material which is territory based** (like monuments, archeological items, etc.)**, the use of the CARARE schema (versions 1.x and 2.x) as intermediate schema.**
3. **For archive materials** (collections of manuscripts, etc)**, the use of EAD as intermediate schema.**
4. **Use EDM as intermediate schema for those providers who are currently exporting metadata following EDM to make mappings to the *current* implementation of EDM by Europeana.**
5. **Expand the aggregation workflow in LoCloud and provide mappings to automatically convert the information from ESE to EDM. This option will be useful for those providers who are currently exporting metadata following ESE**.

6. Amendment

Deliverable “D1.5: Requirement Analysis” presents the technical aspects of the user requirements that have been collected through a series of workshops and surveys with the aim of facilitating the design of the technical infrastructure of the LoCloud project.

The D1.5 deliverable assesses the relevance of the intermediate schemas described in this deliverable, further expanded by the results of the content provider workshops. According to the survey,content providers indicated that the most suitable metadata schemas they could use to provide their content into are, primarily, CARARE (12 partners), LIDO (11 partners) and EAD (10 partners). Besides, 10 partners showed a preference towards ESE as intermediate schema.

In general, we think that the outcomes of the deliverable D1.5 are compatible with the proposal made in this document. In this regard, we think that the survey and user workshops conducted under D1.5 assess that the claim that the proposed intermediate schema effectively suit the content providers. However, the survey also revealed that the initial recommendation did not suit all content providers, particularly those who are exporting the metadata following either ESE or EDM . Therefore, we decided to expand our recommendation by adding two more options (points 4 and 5 of Section 5.3) to completely fulfill the requirements of the content providers.