ICT-PSP Project no. 270905

LINKED HERITAGE

Coordination of standard and technologies

for the enrichment of Europeana

Starting date: 1st April 2011

Ending date: 31st October 2013

Deliverable Number:	D2.1
Title of the Deliverable:	*Best practice report on cultural heritage linked data and metadata standards*
Dissemination Level:	Public

Contractual Date of Delivery to EC:	Month 06
Actual Date of Delivery to EC:	November 2011

Project Co-ordinator

Company name:	Istituto Centrale per il Catalogo Unico (ICCU)
Name of representative:	Rosa Caffo
Address:	Viale Castro Pretorio 105, I-00185 Roma
Phone number:	+39.06.49210427
Fax number:	+39.06. 06 4959302
E-mail:	rcaffo@beniculturali.it
Project Web site address:	http://www.linkedheritage.org

Context

WP	2
WP Leader	Collections Trust
Task	2.1
Task Leader	Collections Trust
Dependencies	None

Author(s)	Gordon McKenna (Collections Trust, UK) Regine Stein (Philipps Universitaet Marburg, DE)
Reviewers	Graham Bell (Editeur, UK) Michael Hopwood (Editeur, UK) Dov Winer (Digital Heritage, UK)
Approved by:	

History

Version	**Date**	**Author**	**Comments**
0.1	November 2011	As above.	
Final	December 2011	As above.	

TABLE OF CONTENTS

1 INTRODUCTION

1.1 THE PURPOSE OF WORK PACKAGE 2

Work package 2 of the *Linked Heritage* project (WP 2) is tasked with:

1. Exploring the state of the art in linked data and its applications and potential;
2. Identifying the most appropriate models, processes and technologies for the deployment of cultural heritage information repositories as linked data;
3. Considering how linked data practices can be applied to cultural heritage information repositories, to enrich them and to allow them to align with other linked data stores and applications;
4. Exploring the state of the art in persistent identifiers (both standards and management tools);
5. Identifying the most appropriate approach to persistent identification, e.g. a unique standard or a set of different standards;
6. Designing a feasibility model and realising a demonstrator of a flexible, scalable, secure and reliable infrastructure for a network of 'linked data enabled' cultural heritage information repositories;
7. Exploring the state of the art in cultural metadata models, and in particular their interoperability across libraries, museums, archives, publishers, content industries, and the Europeana models (ESE and EDM);
8. Outlining the potential benefits that richer cultural heritage metadata could bring to Europeana, and to the other services which will use it.

1.2 ROLE OF THIS DELIVERABLE IN THE PROJECT

This deliverable has three roles in the project:

- Educate the partners, and the wider cultural heritage community, about linked data. This includes linked data's associated technical standards;
- Give advice based on the use of linked data in the cultural heritage community;
- Inform the subsequent work of WP 2 in the rest of the project In particular it will inform:
    - **Task 2.3** – *Technical specifications*: Deliverable D2.3 - *Specification of the technologies for large-scale implementation of cultural heritage linked data*;
    - **Task 2.4** – *Enabling linked cultural heritage data*.

1.3 APPROACH

This deliverable was created based on a process for creating similar deliverables that was developed, and successfully used, during the *ATHENA* project. Its steps are:

1. ***Carry out research*** – Look at what already exists in the environment under discussion. Perhaps survey the project partners on what they are using and or their opinions;
2. ***Make an analysis of the research*** – Look for patterns and trends which can be explained;
3. ***Give simple advice*** – This should be practical and implementable by the partners in the project, and beyond;
4. ***Reuse or create tools*** – Tools should be: easy to use; relevant to the cultural sector audience; and be adaptable, with an open licence, which allows for derivatives to be created (e.g. multilingual versions);

5. ***Identify further needs*** – Leading to further work in the project, and later.

1.4 STRUCTURE OF THE DELIVERABLE

Section 2 of the deliverable gives an overview of linked data: how it came about; what it is; and why it is important.

An analysis of a survey[1] of Linked Heritage partners (and providers) is carried out in *Section 3.* This covered: their knowledge of linked data; their experience in using linked data; their views on the licensing of linked data; and the use of metadata for describing their collections.

In the *Section 4* we look at the *Linking Open Data Cloud* which is a major source of information about this area. This work allowed us to answer the questions:

- Is *The Cloud* 'open'?
- Which IPR licences are used for linked data?
- How big is *The Cloud*?
- What are the subjects in the data?
- Which formats are used to encode data?
- How is *The Cloud* linked?
- What cultural heritage data is in *The Cloud*?

Section 5 explores the *Standards Landscape for Linked Data*. This describes all the major standards, including those for creating and licensing the use of linked data.

Cultural metadata standards are looked at in *Section 6*. The results of the partners' metadata survey are given, and this leads to the selection of standards for use during the *Linked Heritage* project.

Section 7 contains work package 2's best practice advice for linked data and metadata, and *Section 8* gives conclusions including suggestions for further work.

Finally there are two appendices:

1. *The linked heritage survey questions*;
2. *Short bibliography of introductory material on linked data*.

[1] The questions asked can be found in *Appendix 1* at the end of the deliverable.

2 OVERVIEW OF LINKED DATA

2.1 PUBLISHING STRUCTURED DATA ON THE WEB

There have been a number of attempts to publish structured data on the Web. This section looks at these attempts and highlights why they are limited.

Classic Web scenario

The classic Web scenario has a number of features:

- A single global information space (made up individual web servers holding documents and other resources linked together);
- URLs as:
    - Globally unique IDs;
    - Retrieval mechanism.
- HTML – shared content format for documents;
- Hyperlinks – links between documents and other web accessible resources.

This can be represented as:

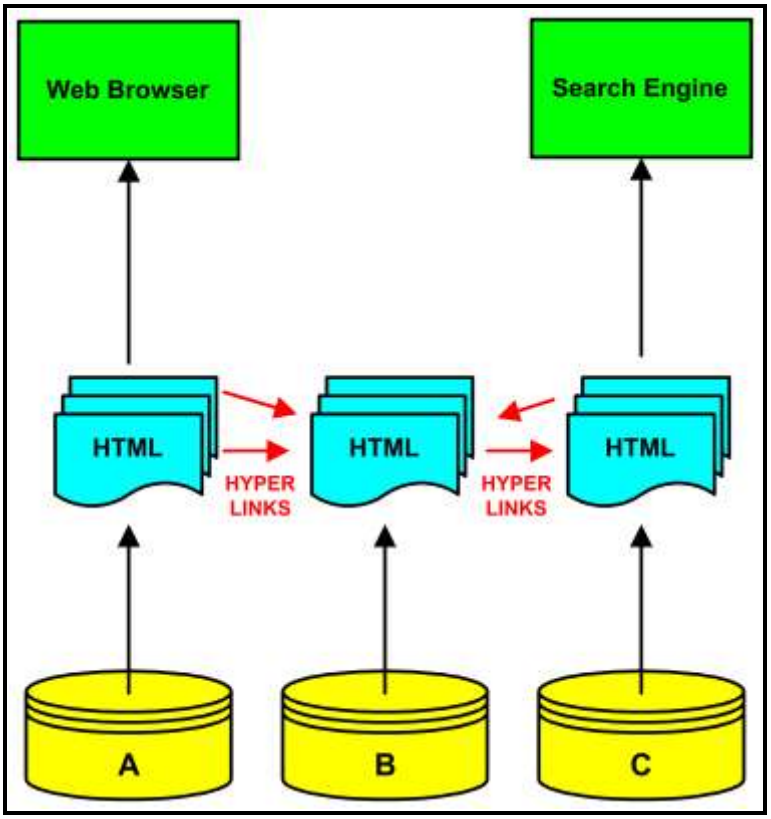

Figure 1: Classic Web scenario

Here the web servers, A, B and C, give access to their own sets of web pages (i.e. documents and other resources). These pages and resources are linked together by HTML hypertext links.

However this scenario has issues:

- Content may generally: not be well structured; with inexplicit semantics; not interoperable; or may be behind a protective 'wall';
- Applications cannot process the content;
- Expressive questions cannot be asked by the user.

What we actually want to do is use the Web like a single global database. This can be done by publishing **structured data** directly on the Web. This extension of the classic Web is usually called the **Semantic Web**. This is a group of methods and technologies to allow machines (applications) to understand the meaning (or "semantics") of information (data) on the Web.

API/mashup scenario

One popular way of publishing structured data on the Web is to use APIs and mashups.

In general terms an Application Programming Interface (API) is a set of rules and specifications, embedded in 'code', which software programs can use to communicate with each other. Therefore an API acts as an interface facilitating interaction in a way similar to how the user interface aids interaction between humans and computers.

Usually in the context of the Web[2] an API is defined as:

- A set of HTTP (HyperText Transfer Protocol) request messages;

  and

- A definition of the structure of response messages,

The responses are usually delivered as XML[3] (EXtensible Markup Language) or JSON[4] (JavaScript Object Notation) format. Web APIs have until recently been web services, such as SOAP[5] (Simple Object Access Protocol). However there has been a move towards services with a more direct style of communications, such as REST[6] (REpresentational State Transfer).

In terms of applications, Web APIs enable the use of content created in one place on the Web to dynamically appear in another place. Typical applications include:

- Photographs from sites like *Flickr* appearing in social networks like *Facebook*;
- Presentations from *SlideShare* embedded in a profile on *LinkedIn*;
- Comments made on *Twitter* posted in *Facebook*;
- Videos, from *YouTube* embedded in a blog.

[2] For further information on API see: **http://www.programmableweb.com**

[3] See *Section 7* below for a description of the standard.

[4] Pronounced 'jason'. Defined by: **Crockford, Douglas**. (2006). *The application/json Media Type for JavaScript Object Notation (JSON)* [RFC 4627]. The Internet Society. Download from: **http://tools.ietf.org/html/rfc4627**

[5] Defined in: **Gudgin, Martin** (*et al*) [Eds.] (2007). *SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)*. World Wide Web Consortium (W3C). Download from: **http://www.w3.org/TR/soap12-part1**

[6] For a technical introduction see: **Rodriguez, Alex**. (2008). *RESTful Web services: The basics*. IBM. Download from: **https://www.ibm.com/developerworks/webservices/library/ws-restful/**

In the cultural heritage sector examples of APIs include:

Organisation or Service	URL for API information
Amsterdam Museum	http://www.appsforamsterdam.nl/wp-content/uploads/2011/02/AmsterdamMuseum.txt
Brooklyn Museum [New York]	http://www.brooklynmuseum.org/opencollection/api
Culture Grid [UK]	http://www.culturegrid.org.uk/wp-content/uploads/2010/06/Culture-Grid-search-service-v2.pdf
Organisation or Service	**URL for API information**
DigitalNZ [New Zealand]	http://www.digitalnz.org/developers
Europeana OpenSearch [Europe]	http://europeanalabs.eu/wiki/EuropeanaOpenSearchAPI
LAARC (London Archaeological Archive and Research Centre) Catalogue Search	http://www.museumoflondon.org.uk/laarcWS/v1/doc
Muselius [international]	http://www.muselius.com/api
Museum Victoria [Melbourne]	http://museumvictoria.com.au/collections/help/api
National Maritime Museum [London]	http://collections.nmm.ac.uk/api.html
Open Context [international]	http://opencontext.org/about/services
Oxford Celtic Coin Index	http://www.finds.org.uk/CCI/blog/accessing-the-api/
Powerhouse Museum [Sydney]	http://www.powerhousemuseum.com/collection/database/download.php
Reciprocal Research Network [Canada]	http://www.rrnpilot.org/api
Science Museum [London]	http://api.sciencemuseum.org.uk/documentation/
Steve In Action [international]	http://tagger.steve.museum/api-docs/api.php
Victoria and Albert Museum	http://www.vam.ac.uk/api

Typically access to an API in controlled by an API key. This is a code, generated to be unique, by the API provider. API keys can be used to track how the API is being used. This is done to prevent malicious use or abuse of the APIs 'terms of service'.

Data from individual APIs can be used to create new applications, however data from multiple Web APIs can be combined together to form new web applications called 'mashups'. Mashups often combine a 'general' API, e.g. Google Maps, with a domain specific API, e.g. *Europeana OpenSearch*. This scenario is shown by the figure on the next page:

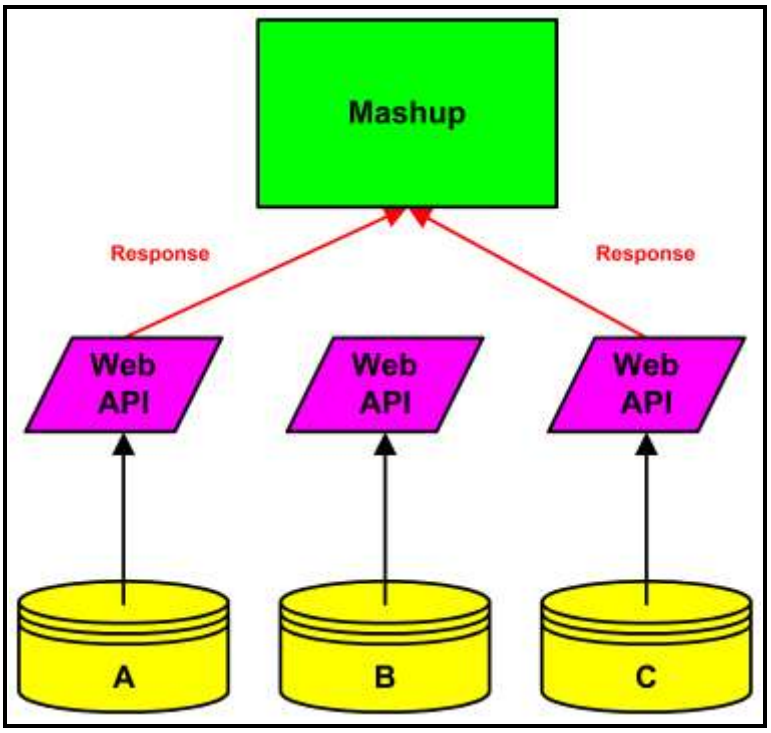

Figure 2: API/mashup scenario

Recent work[7] with the *Europeana OpenSearch* API, including 'hack days', has produced:

- Europeana search;
- Related items widget[8];
- Thesaurus search;
- *WordPress* search widget;
- Separation of national and non-national results from Europeana search;
- Extract keywords from Europeana to aid search;
- Search by a box on a map;
- Batch upload data to *Wikimedia*;
- Semantic enrich Europeana using a game;
- Random Europeana image extractor.

Despite all this activity the API/mashup scenario has *pros and cons*:

For:

- APIs expose structured data;
- APIs enable new applications.

[7] See: **http://www.version1.europeana.eu/web/api/application-gallery**

[8] A widget is a small executable application that can be installed in a web page to provide additional functionality, such as a search.

Against:

- Proprietary interfaces – each API is unique;
- Mashups are based only on fixed set of sources;
- Hyperlinks cannot be set between data objects;
- APIs puts data on the Web into separate ‘silos’.

Publishing ‘linked data’ is suggested as a solution to these issues. It is explored in the next section.

2.2 LINKED DATA

The solution, proposed by Sir Tim Berners-Lee[9], to the issues of the classic Web and API/mashups scenarios, is to publish structured data as 'linked data' and thereby enable the Semantic Web. In his suggestion he says:

> "*The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.*"

The paper then goes on to define four 'principles' or 'rules' for linked data[10]:

1. Use URIs as names for things;
2. Use HTTP URIs so that people can look up those names;
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL);
4. Include links to other URIs, so that they can discover more things.

In outline this scenario can be represented as:

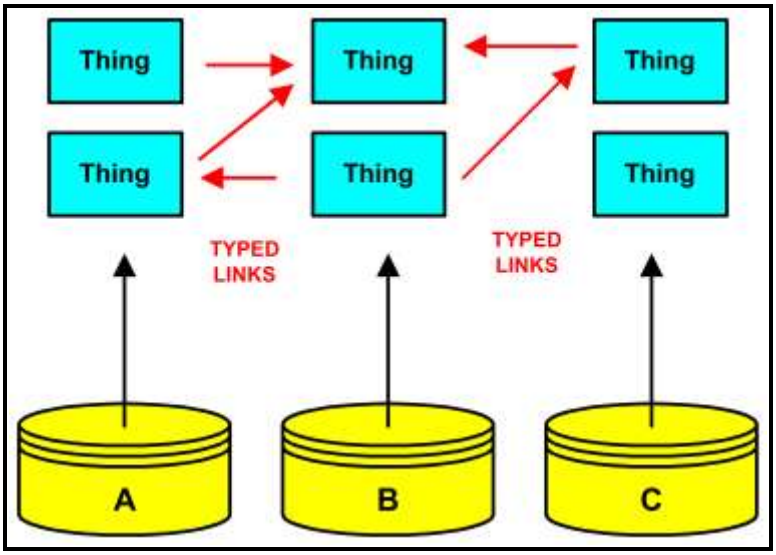

Figure 3: Linked data scenario

'Things' can be anything, including: objects; archival material; books and other document types; media files; places; persons; and organisations; events and concepts (such as material, colour, and style).

So why publish linked data? A number of reasons come to mind:

- It builds on the classic architecture of the Web;
- The data becomes part of the Semantic Web;
- People can use various data browsers to explore the data;
- The data is can be crawled by Semantic Web search engines and is used by various applications;
- People will start setting links to the data, which might make more people find and use the data.

[9] **Berners-Lee, Tim**. (2006–2009). *Design Issues: Liked Linked Data*. World Wide Web Consortium (W3C). Download from: **http://www.w3.org/DesignIssues/LinkedData.html**

[10] For the URI, RDF and SPARQL see *Section 5*.

Linked data is more generic than APIs because it:

- Builds on standards in contrast to proprietary Web APIs;
- Enables applications that use an unbound set of data sources and incorporate new data sources.

The 3rd and 4th principles talk about giving useful information using the standards RDF* and SPARQL in the next sections we look at these. Note that these will not be detailed technical primers in RDF or SPARQL. It aims to give a non-technical person a 'taster' of the subject.

2.3 PUBLISHING LINKED DATA WITH RDF

RDF (Resource Description Framework) is the standard in linked data that is used to describe the 'things' (known as 'resources'). It is an abstract data model that is based on classic conceptual modelling. The approaches it uses include entity-relationship or class diagrams, making statements about things in the form of **subject-predicate-object** expressions. These expressions are known as **triples** in RDF. Looking at the different parts of a triple:

- **Subject** – the thing being described;
- **Predicate** – a trait, aspect, or property of the thing, which expresses a relationship between the subject and object;
- **Object** – the thing that is the value of the predicate (trait, aspect or property) of the object thing.

So in the statement "The Kiss was created by Gustav Klimt":

- **Subject** – The Kiss;
- **Predicate** – Created by;
- **Object** – Gustav Klimt.

In terms of representation:

- **Subject** – must be a URI;
- **Predicate** – must be a URI;
- **Object** – may be a URI or a constant value or 'literal' (e.g. "oil on canvas")

If there are maintained URIs for the all parts of the expression then publishing linked data will be relatively simple. However it is expected that this will not the case with cultural heritage sector. Therefore the first task for a cultural heritage organisation may be to create and maintain suitable URIs for the things that they will be describing in their linked data.

For a museum this will include URIs for the objects in their collections (usually the subject in an expression). These identifiers should also be maintained ensure that they are persistent[11].

The predicate URIs will have to exist before publication. This is because they point to the elements of a format which embody the descriptive traits, aspects or properties needed to describe the subject of the expression. To give an example:

If an organisation chooses[12] to use the Dublin Core elements in their linked data and they wish to publish information about the creation of something then URI **`<http://purl.org/dc/elements/1.1/creator>`** will embody the predicate "Created by". The choice of 'format' is an important consideration for the publication of linked data. This will be looked at later in the deliverable.

[11] See D2.2 – *State of the art report on persistent identifier standards and management tools for a further discussion of this issue*.
[12] See *Section 5* for a discussion about which formats are available.

For the object URIs, these might already exist in the body of linked data already published. However if they do not the publisher must create and maintain them[13].

So for the example the graph[14] for the *The Kiss* might look this:

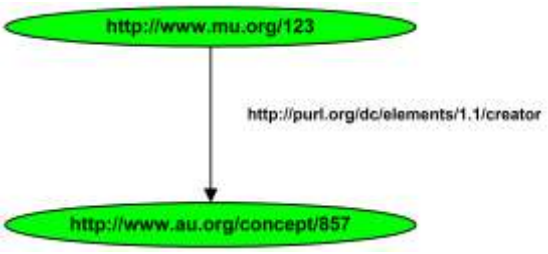

Figure 4: A simple RDF graph

While written out it can look like this:

```
<http://www.mu.org/123> <http://purl.org/dc/elements/1.1/creator> <http://www.au.org/concept/857>
```

The transformation from a RDF graph to a textual representation is called serialisation. There are four common, and interchangeable, serialisations:

- RDF/XML;
- Notation3 (or N3);
- Turtle;
- N-Triples.

For the serialisations used in cultural heritage linked data see *Section 4.8*. Full details of the standards can be found in *Section 5*.

2.4 CONSUMING LINKED DATA WITH SPARQL

The other standard mentioned in the In Berners-Lee's *Principles* is SPARQL[15] (SPARQL Protocol and RDF Query Language). This is a query language for RDF.

Data is made accessible through a standards-compliant service called a 'SPARQL endpoint' which sits on top of the knowledge base of linked data. This allows four query types:

- **`SELECT`** – Returns data values in a table defined in the query;
- **`CONSTRUCT`** – Returns data values which are transformed into valid RDF defined in the query. This can then be serialised into, for example RDF/XML;
- **`ASK`** – Gives a Boolean 'yes' or 'no' result for a query;
- **`DESCRIBE`** – Returns an RDF graph with information that meets the query. The form of the graph is decided by the provider of the knowledgebase and not by the query itself.

Each type takes a '**`WHERE`**', which restricts the query. **`WHERE`** is optional for the **`DESCRIBE`** type.

Here is a simple example using RDF data:

```
<http://example.org/book/book1> <http://purl.org/dc/elements/1.1/title> "The Bible" .
```

[13] See *Section 5* for a discussion about what linked data, with URIs is available.
[14] Note the object and subject URIs are not real.
[15] Pronounced like 'sparkle'.

With the SPARQL query:

```
SELECT ?title
WHERE
{
  <http://example.org/book/book1> <http://purl.org/dc/elements/1.1/title> ?title .
}
```

Gives the result:

title
"The Bible"

In summary the usual linked data setup consists of:

- Data base in the form of a RDF 'triplestore';
- SPARQL endpoint service giving access to the knowledge base in a number of RDF serialisations.

Other ways to consume linked data

An overview of the consumption of linked data was given recently by Juan F Sequeda[16]. In it he identified some linked data consuming applications

Generic applications:

- ***Linked data browsing*** –
  - View data (using a URI link) in a tabular form from within a web browser;
  - Services include: *The Tabulator*; *Zitgist*; *Marbles*; *Explorator*; and *Disco*;
  - Navigate between things using RDF links;
  - Not very usable?
- ***Linked data search engine*** –
  - Crawls and searches the Semantic Web of RDF documents;
  - Follows RDF links;
  - Human focused examples: *Falcons*, *SWSE*, and *VisiNav*;
  - Machine focused examples: *Sindice*; *Swoogle*, *Watson*, and *Uberlic.*
- **Semantic Search Engine Optimisation**
  - Markup HTML using RDFa;
  - Use standard formats: *Google Vocabulary*, *Dublin Core*;
  - Result will be that Google and Yahoo will crawl it and render it better.
- ***On-the-fly mashups*** – e.g. SIG MA (http://sig.ma)

[16] Sequeda, Juan F. (2011). Consuming Linked Data. Semantic Technology Conference, June 2011. View at: **http://www.slideshare.net/juansequeda/consuming-linked-data**

Domain specific applications:

- Government: Data.gov, Data.gov.uk, http://data-gov.tw.rpi.edu/wiki/Demos;
- Music: Seevl.net;
- Dbpedia Mobile;
- Life Science: LinkedLifeData;
- Sport: BBC World Cup;
- Faceted browsers: http://dbpedia.neofonie.de/browse, http://dev.semsol.com/2010/semtech.

3 LINKED HERITAGE SURVEY ON METADATA & LINKED DATA

3.1 HOW THE SURVEY WAS CARRIED OUT

WP 2 decided that it would be informative to survey the Linked Heritage partners (and associated providers). This was undertaken to find out:

- Information about the respondents to the survey (organisation type, and country);
- Their knowledge and use of linked data, and linked data initiatives;
- Their attitude to Europeana's proposed open data agreement;
- Their use of metadata.

The survey had 22 questions and was in three parts:

1. *Respondent Information*;
2. *Metadata*;
3. *Linked Data*.

The survey was created in two versions:

- RTF document – which allowed the use by project partners and/or distribution of the questions to content providers before using the:
- On-line system – where project partners entered the data they had collected and allowed the easier collection of data and analysis.

The rest of this section details the results of parts 1 and 2 of the survey. Part 3 is dealt with in the section on metadata.

3.2 RESPONDENT INFORMATION

Here are the figures for the types of organisations that responded to the survey:

3.2.1 Types

Here are the figures for the types of organisations that responded to the survey:

Respondent type	Number of respondents	%
Museum	4	10.3
Library	5	12.8
Archive	4	10.3
Sound archive	1	2.6
Publisher	0	0
Aggregator	10	25.6
Other	23	59.0

These figures show that the content being supplied to Europeana through the *Linked Heritage* project covers all of the cultural domains including aggregators. However there is also significant number of responses from organisations which are not contributing content and therefore they will not appear in the metadata section of the survey.

Nearly 60% of the respondents are not one of the 'standard' types. Therefore it is useful to list what was the response was to the question: "*If you ticked 'Other' please give organisation type*":

- Mediator between providers and Linked Heritage project;
- Group of museums;
- Governmental organisation for the protection of immovable cultural heritage and of the movable and living cultural heritage associated with it.
- National *Books in Print*;
- Technical partner;
- University;
- DOI [Digital Object Identifier] registration agency;
- Centre for research and innovation;
- Ministry of Culture;
- Company in cultural heritage field;
- Scientific research institute with museum collections;
- Management and quality services company;
- National contact point;
- SME – consultancy;
- Public broadcaster and media archive (video, sound, and photographs);
- Publishing standards body
- Theatre documentation (photographs);
- Public organisation;
- Regional public administration responsible for the cultural heritage information system;
- Technology provider;
- Association and information centre;
- Cultural agency.

In contrast the ATHENA project was much more museum-centric.

3.2.2 Countries

Here are the figures for the countries where respondents are based:

Country	Number of respondents
Austria	1
Belgium	4
Bulgaria	1
Cyprus	1
Czech Republic	1
Estonia	1
France	3
Germany	4
Greece	3
Hungary	1
Ireland	2
Israel	1
Italy	6

Country	Number of respondents
Poland	1
Russian Federation	1
Slovakia	1
Slovenia	1
Spain	2
Sweden	1
United Kingdom	2

Obviously these figures reflect the partners of the project, but there is a spread throughout Europe, with a couple of respondents outside the EU.

Taken as a whole, the information about respondents leads the authors of the deliverable to conclude that the sample is fairly representative of the sector.

3.3 LINKED DATA

3.3.1 Awareness

To "*Are you or your organisation familiar with the concept of linked data?*"

Response	Number of respondents	%
Yes	30	75.0
No	10	25.0

The 'No' surprised the authors, but shows that there is a 'market' for information and tools about linked data!

3.3.2 Use

To "*Have you or your organisation had experience of using linked data in connection with your collections?*"

Response	Number of respondents	%
Yes	7	17.5
No	33	82.5

Those who answered '*Yes*' were asked "*... please give details of which source(s) of linked data you use ... and why you use it ...*" The sources used were:

Linked data source	Number of respondents
DBpedia	4
GeoNames	3
Freebase	1
IPTC	1
Thesauri in SKOS	1

Only two respondents gave information as to why they used a source:

- *DBpedia* - interesting information source;
- *GeoNames* - for place name disambiguation.

3.3.3 Publication

To "*Have you or your organisation had experience of publishing linked data in connection with your collections?*"

Response	Number of respondents	%
Yes	4	10.0
No	36	90.0

Those who answered '*Yes*' were asked "... *please give details ...*" Three gave details:

- **http://data.kunstkamera.ru/sparql** and **http://data.kunstkamera.ru/**
- Full bibliographic records of OPAC and Digital Library (OSZKDK) in DC. Name authority in FOAF. Thesaurus in SKOS. Details and URLs are: **http://nektar.oszk.hu/wiki/Semantic_web**. Support RDFa in Digital Library (OSZKDK).
- The Department for the French Archives had published its thesaurus in SKOS in a linked data reuse perspective. An ongoing national project will bring together all the vocabularies in use in the ministry in order to get a network of concepts that would be connected to other initiatives such as *RAMEAU* in SKOS.

3.3.4 Linked data projects and initiatives

To "*Do you or your organisation know of any linked data projects or initiatives in your country in the field of cultural heritage?*"

Response	Number of respondents	%
Yes	15	37.5
No	25	62.5

Those who answered '*Yes*' were asked "... *please give details ...*" Ordered by country the responses were:

Country	Project or initiative
France	*RAMEAU*: **http://www.cs.vu.nl/STITCH/rameau/index-fr.html**
	ISIDORE: **http://rechercheisidore.fr**
	Pactols: **http://www.frantiq.fr/thesaurus-pactols**
	BABEL: **http://babel.alienor.org**
	COLLECTIONS: **http://www.culture.fr/fr/sections/collections/moteur_collections**
	PALISSY: **http://www.culture.gouv.fr/public/mistral/dapapal_fr?ACTION=NOUVEAU&USRNAME=nobody&USRPWD=4%24%2534P**
	EROS: **http://www.c2rmf.fr/pages/page_id18479_u1l2.htm**
	PATRIMOINE LOT: **http://www.patrimoine-lot.com**
	WIKIMEDIA COMMONS FRANCE: **http://commons.wikimedia.org/wiki/Accueil**
	Centre Pompidou Virtuel: **http://www.centrepompidou.fr**

Country	Project or initiative
Germany	*Linked data service of the German National Library*: **http://www.d-nb.de/eng/hilfe/service/linked_data_service.htm**
	"*Several initiatives throughout the country*"
Israel	*Vocabularies of the Israel Museum Jerusalem that have been migrated to SKOS:* **http://www.imj.org.il/imagine/thesaurus/allobject.htm** **http://www.imj.org.il/imagine/thesaurus/objects/objectTOC.htm**
Italy	*Linked Open Data Italia*: http://www.linkedopendata.it/en-home
	SPAR ontologies: **http://opencitations.wordpress.com/2010/10/14/introducing-the-semantic-publishing-and-referencing-spar-ontologies**
	Datagov.it. Associazione italiana per l'Open Government: **http://www.datagov.it**
	http://www.linkedopencamera.it
	http://www.spaghettiopendata.org
Russia	*Open Kunstkammer*: **http://www.kunstkamera.ru** (**http://ercim-news.ercim.eu/en86/special/the-open-kunstkammer-data-project**)
Sweden	*LIBRIS (joint catalogue of the Swedish academic and research libraries)*: **http://www.kb.libris.se**
Spain	*Open Data Gencat*: **http://dadesobertes.gencat.cat/en/index.html**
	Euskadi: **http://opendata.euskadi.net/w79-home/es/**
	Patmapa: **http://patmapa.gencat.cat/**
	Cantabria's Cultural Heritage Ontology: **http://hdl.handle.net/10760/13938**
United Kingdom	*Various government data sets*: **http://data.gov.uk**

All these figures are not untypical, as will be shown in section on the *Linking Open Data Cloud* (see below).

3.4 EUROPEANA OPEN DATA AGREEMENT

We asked the respondents:

> "*Europeana's new licence requires that providers will have to agree to have the metadata that they provide to Europeana published as Linked Open Data. This means that any 3rd party use, including commercial, is permitted. Does your organisation agree to this?*"

They answered:

Response	Number of respondents	%
Yes	11	29.7
Not sure	20	54.1
No	6	16.2

Respondents were also asked to explain their answer.

Those who answered "*Yes*" said (with numbers):

- 1 – *Publishing on Web means Open Data;*
- 1 – *Participated in the ATHENA project;*
- 1 – *Metadata provided to Europeana specifically selected for open linked data.*

Those who answered "*Not sure*" said:

- 4 – *Metadata not ours (our providers' decision);*
- 4 – *Under discussion;*
- 2 – *Under discussion (possible legal obstacles);*
- 2 – *Decision not ours (made at a higher level);*
- 1 – *Will provide minimal data;*
- 1 – *Against commercial reuse.*

Those who answered "*No*" said:

- 3 – *Against 3rd party commercial use;*
- 1 – *National policy does not allow commercial use;*
- 1 – *Do not contribute to Europeana.*

4 ANALYSIS OF THE *LINKING OPEN DATA CLOUD*

4.1 INTRODUCTION

The *Linking Open Data Cloud*[17] (*The Cloud*) is the best known representation of linked data. It shows 'packages' of linked data and the links between packages. In May 2007 it looked like this (with 12 packages):

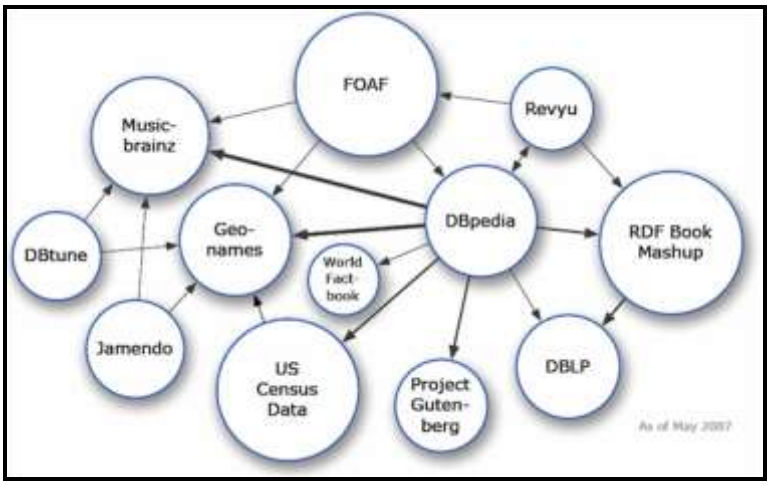

Figure 5: *The Cloud* in May 2007

In September 2011 the version that is colourised to represent the domain of the package looked like this (with 311 packages):

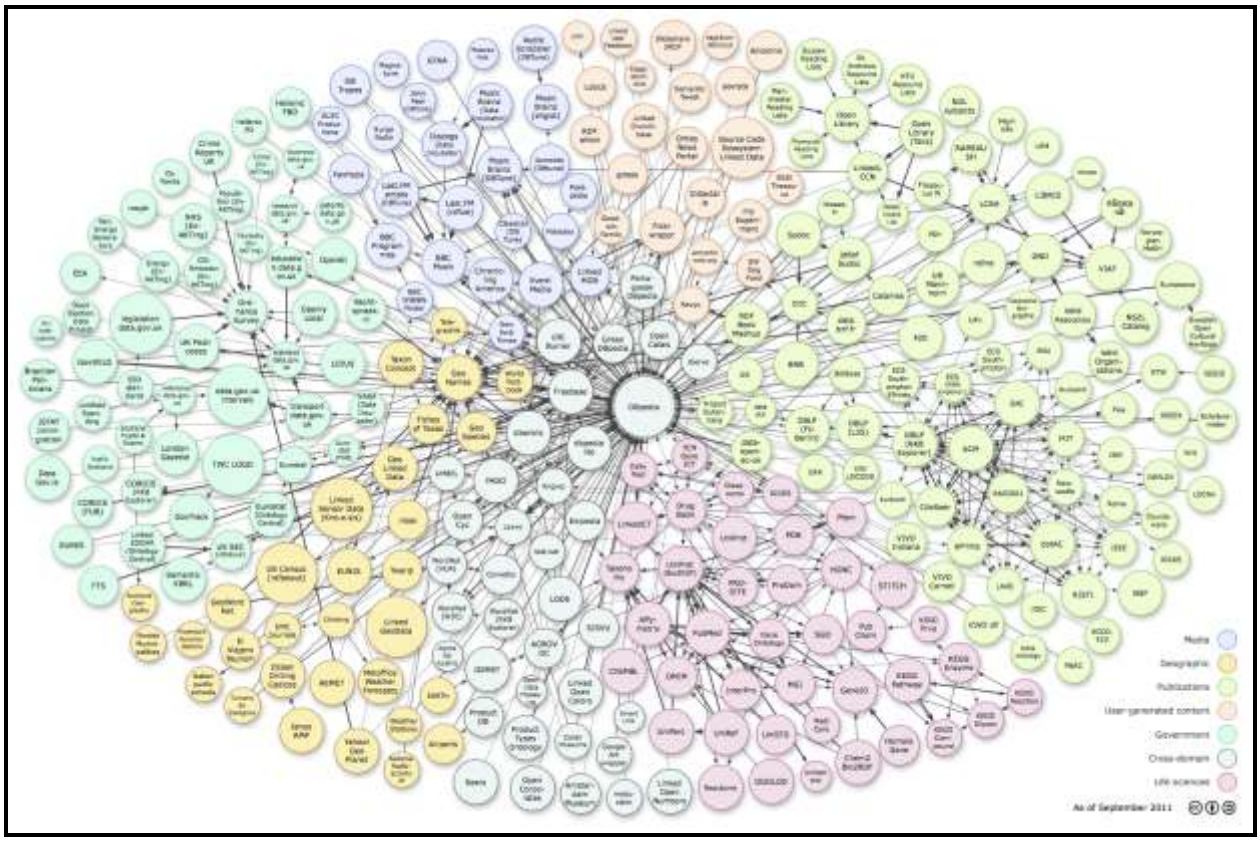

Figure 6: *The Cloud* in September 2011

[17] See: **http://linkeddata.org**

It can be seen that *The Cloud* is growing very quickly and, in its latest form, it is becoming very difficult to get a proper overview of what it made up of. Luckily *The Cloud* is maintained using a wiki which is maintained on *The Data Hub* website[18]. This effort is part *Linking Open Data* community project[19] which is part of the W3C's Semantic Web Education and Outreach Interest Group (SweoIG).[20] Therefore it may be considered as representing a significant proportion of the linked data available.

The Data Hub is a registry of open (and not open) knowledge with information on packages and projects (including the *LOD Cloud* 'group'). Once the *LOD Cloud* group is chosen a user is presented with the first of a set (currently seven) of result screens:

Figure 7: *The Data Hub* search results screen

For each package this screen gives information about:

- Name of the package (as a link to the full record);
- Description of the package;
- Links to the resources (including examples) available for the package;
- IPR status of the package.

For each package there is a full record:

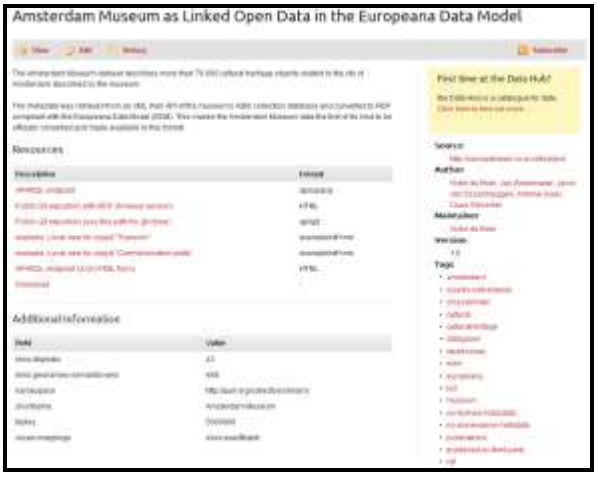

Figure 8: *The Data Hub* package record

[18] See: **http://thedatahub.org** and **http://thedatahub.org/group/lodcloud**

[19] See: **http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData**

[20] See: **http://www.w3.org/wiki/SweoIG**

For each package this screen includes additional information about:

- Which other packages are linked to (including number of links);
- The number of 'triples' in the package (a measure of size)
- Further details (not visible in the screenshot) about the IPR situation of the package;
- In Tags:
    - Subject information;
    - Which 'formats' are used.

Such information allows us to examine some aspects of *The Cloud*, and from this we can discover the 'emergent' best practice for linked data based on common practise.

4.2 IS *THE CLOUD* 'OPEN'?

This may seem to be a strange question to ask. However when first examining the information on *The Data Hub* website it became apparent that there is a significant component of *The Cloud* that is not open. In *The Cloud*:

Open = able to be re-used commercially.

Examining the data showed:

In terms of packages (311)

IPR Status	**%**
Open	42.6
Not open	57.4

In terms of triples (c38 billion)

IPR Status	**%**
Open	30.9
Not open	69.1

This result is rather surprising as it shows that the majority of *The Cloud* is not open. One reason for this anomaly may be that *The Cloud* is rather like a historic landscape with the evidence of many different time periods apparent at the surface. In this case the assumption is that we are seeing many packages which are early components of *The Cloud*, at time when IPR and having a licence was not considered important. That being said the latest update still has 'Not open' packages. Other insights can be gained by looking at the licences being used in more detail.

4.3 WHICH IPR LICENCES ARE USED?

4.3.1 Open licences

Of the 132 packages (c11.9 billion triples) with open licences:

Licence type	% by Package	% by Triples
Creative Commons Attribution (CC BY)	28.8	45.8
Creative Commons Attribution Share Alike (CC BY-SA)	18.2	10.2
Open Data Commons Public Domain Dedication and Licence (ODC PDDL)	10.6	0.2
Creative Commons CC Zero (CC0)	9.1	2.9
UK Crown Copyright with data.gov.uk rights	7.6	27.4
Other (Public Domain)	6.8	7.0
Other (Open)	5.3	5.0
Other (Attribution)	3.0	0.4
UK Open Government Licence (OGL)	3.0	0.1
GNU Free Documentation Licence (GNU FDL)	3.0	0.0
Open Database Licence (ODbL)	2.3	0.9
GNU General Public Licence (GNU GPL)	0.8	<0.1
New BSD license and Simplified BSD licence	0.8	<0.1

The dominant use of CC BY for an open licence is to be expected. It is an obvious choice, together with CC BY-SA and ODC PDDL and CC0. The latter is a relatively new option, and is the choice made by Europeana, and at second hand by its providers, for its publication of linked open data. It is the most permissive of the open licences with attribution being a 'recommendation' rather than mandatory.

One national initiative is worth mentioning, is that in the United Kingdom. Much data is being published by the UK government using its own open data licences. At the moment these make up over 10% of *The Cloud*. The UK Open Government Licence is interoperable with CC BY.

4.3.2 Not open licences

Of the 178 packages (c26.7 billion triples) with licences that are not open, or with no licence information:

Licence type	% by Package	% by Triples
[not given]	69.1	89.4
None	14.6	0.3
Creative Commons Attribution Non-commercial (CC BY-NC)	7.3	5.8
Other (Not Open)	6.7	<0.1
Creative Commons Attribution (CC BY)	1.1	0.6
Other (Non-Commercial)	0.6	3.9
Creative Commons Attribution Share alike (CC BY-SA)	0.6	<0.1

From the above[21] it can be seen that for over 80% of packages and nearly 90% triples of the 'not open' part of *The Cloud* or there is no information about the IPRs.

It is interesting to note that this situation does not seem to impact on the use of *The Cloud*, and that some of the newest packages do not have licences.

For those who publish their data in *The Cloud* with a licence, but do not want their data to be open, then one of two options is taken:

- CC BY-NC;
- Their own 'non-standard' licence with, presumably, special requirements.

4.4 HOW BIG IS *THE CLOUD*?

As mentioned above there are c38 billion triples in *The Cloud*. There is a large distribution in size. 9 packages (2.89%) have over a billion triples. Nearly a quarter of the packages are relatively small with less than 100,000 triples. The smallest has only 368 triples. This suggests that there is an element of 'test' linked data in *The Cloud*, which is confirmed by some packages being described as 'test'. The average number triples in a package is c124 million.

The ten largest packages with open licences are:

Package	Number of triples
LinkedGeoData	3.00 billion
UK Legislation	1.90 billion
Linked Sensor Data (Kno.e.sis)	1.73 billion
data.gov.uk Time Intervals	1.00 billion
DBpedia	1.00 billion
Open Library data mirror in the Talis Platform	0.54 billion
The Open Library	0.40 billion
Freebase	0.34 billion
transport.data.gov.uk	0.33 billion
Data Incubator: MusicBrainz	0.18 billion

LinkedGeoData (CC BY licence) is a knowledge base of spatial obtained from the *OpenStreetMap*[22] project. Its aim is to give a semantic element to the Semantic Web.

Three packages – *UK Legislation*, *data.gov.uk Time Intervals*, and *transport.data.gov.uk* – are part of an UK Government initiative to publish their public data in an open manner. All of them are published under the "UK Crown Copyright with data.gov.uk rights", a UK specific open licence.

Linked Sensor Data (Kno.e.sis) (CC BY licence) has data on information on weather stations and observations from a US university-based centre.

DBpedia, *Open Library data mirror in the Talis Platform*, *The Open Library*, and *Freebase* are well-known sources of encyclopaedic information on a wide range of topics. They also have a range of different open licences: CC BY-SA, Other (Open), Other (Public Domain), and CC BY.

Data Incubator: MusicBrainz (Other (Public Domain) licence) contains information about music, specifically: albums, artists, tracks, labels and their relationships.

[21] Please note that CC BY and CC BY-SA are open but in the data are described as not open. We have preserved this in the table
[22] See: **http://www.openstreetmap.org**

The ten largest packages without open licences are:

Package	Number of triples
TWC: Linking Open Government Data	9.80 billion
Data.gov	6.40 billion
Source Code Ecosystem Linked Data	1.50 billion
2000 U.S. Census in RDF (rdfabout.com)	1.00 billion
PubMed	0.80 billion
DBTune.org MySpace RDF Service	0.66 billion
UniParc	0.63 billion
DBTune.org AudioScrobbler RDF Service	0.60 billion
Linking Italian University Statistics Project	0.59 billion
UniProt UniRef	0.49 billion

TWC: Linking Open Government Data is the largest package in *The Cloud* and is an aggregation of US government data. It includes data published in the *Data.gov* package. *The Data Hub* does not have any information about the licence for this data. *2000 U.S. Census in RDF (rdfabout.com)* is also US government data about population statistics, and has a CC BY-NC licence.

The following packages have no licence information on *The Data Hub*:

- *Source Code Ecosystem Linked Data* contains structured source code facts from open source projects. It is authored by a Canadian university.
- *PubMed* is a US-based source of medical publications.
- *DBTune.org MySpace RDF Service* and *DBTune.org AudioScrobbler RDF Service* are part of a mini-cloud of nine music-related packages.
- *UniParc* and *UniProt UniRef* are parts of life science knowledge bases from US academic institutions.
- *Linking Italian University Statistics Project* is the publication of Italian Government data abou university students.

4.5 WHAT ARE THE SUBJECTS IN THE DATA?

Within the descriptions for each package within *The Data Hub* wiki are a number of different 'tags'. Some of these tags are obviously subject-based and give an indication of the content of the packages. There does not seem to be a controlled terminology that is being used. So the same subject may be represented by a different tag in different packages. In our analysis we have combined a number of tags which appear to be the same subject. Note also packages can have more that on subject.

After this process the following are the ten most common subjects in *The Cloud*:

Subject tag	Number of packages with tag	% of packages with tag
publications	94	30.23
government	54	17.36
life sciences	46	14.79
geographic	40	12.86
media	32	10.29
library	22	7.07
United Kingdom	22	7.07
education	20	6.43
user generated content	19	6.11
bibliographic	15	4.82

This result generally follows the categories illustrated by the colourised version of *The Cloud* diagram. It is also a 'snapshot' of the current state of the content. *The Cloud* is dominated by data in these areas. By comparison there is very little cultural heritage data. This is probably because, until the advent of Europeana, there has been no interest in linked data in this community.

The appearance of 'United Kingdom' as a tag shows largely the effect of the UK Government's policy of publishing linked data. The role of the USA is not apparent, but this because packages are not tagged 'United States' even when potentially they could be.

4.6 WHICH FORMATS ARE USED TO ENCODE DATA?

In order to encode data for *The Cloud* various formats are used. In most of the literature on linked data the term used for them is 'vocabulary'. We continue to use 'format' here to avoid confusion with the cultural heritage use of vocabulary as being the descriptive terms being used rather than the metadata elements. Also of note is that some of the formats are called 'ontologies'.

The most commonly used are:

Format[23]	**Number of packages using the format**	**% of packages using the format**
Resource Description Framework (rdf)	261	83.92
Dublin Core (dc)	97	31.19
Friend of a Friend (foaf)	84	27.01
Simple Knowledge Organization System (skos)	57	18.33
RDF Schema (rdfs)	42	13.50
Web Ontology Language (owl)	34	10.93
Basic Geo (geo)	25	8.04
Advanced Knowledge Technologies Reference Ontology (akt)	22	7.07
eXtensible HyperText Markup Language (xhtml)	19	6.11
Bibliographic Ontology (bibo)	14	4.50
[none given]	13	4.18
Music Ontology (mo)	13	4.18
DBpedia Ontology (dbpedia)	12	3.86
vCard (vcard)	11	3.54
Semantically-Interlinked Online Communities (sioc)	10	3.22
Creative Commons (cc)	8	2.57
Functional Requirements for Bibliographic Records (frbr)	6	1.93
GeoNames Ontology (geonames)	6	1.93
XML Schema (xsd)	6	1.93
Event Ontology (event)	5	1.61

There seems to be three types of format:

- ***Basic*** – Those that generally organise the entities in *The Cloud*, including links between the entities. They are found in use in nearly all the packages in it, as might be expected. Therefore it is likely that any cultural heritage package will also use them.

  They are: *Resource Description Framework*; *RDF Schema*; *Web Ontology Language*; and *XML Schema*.

- ***Descriptive*** – Those whose elements hold descriptive data about the entities for use in many packages. They are generally developed by a set of interested parties who want to publish their information as linked data. Quite often they have their origins in a specific project or initiative.

[23] The abbreviation in brackets after a format's name is the 'namespace' for that format.

They are: *Dublin Core* (for web resources); *Friend of a Friend* (persons); *Simple Knowledge Organization System* (terminologies); *Basic Geo* (geographical); *Bibliographic Ontology*; *Music Ontology*; *vCard* (business cards); *Semantically-Interlinked Online Communities* (social networks); *Creative Commons* (IPR); *Functional Requirements for Bibliographic Records and Event Ontology*.

- ***Package specific*** – Those whose elements represent the specific data held in a particular package. They were developed in the context of the publication of a single package as linked data. However they can be used in the publication of other packages which may lead to them becoming *de facto* standards.

  They are: *Advanced Knowledge Technologies Reference Ontology*, *DBpedia Ontology*, and *GeoNames Ontology*.

  That there are some formats of this type that are used by more than one package is significant. It suggests that these 'parent package' is playing a significant role in *The Cloud*. Obvious examples of this are *DBpedia* and *GeoNames*, and we shall see a similar pattern when we look at linking in *The Cloud* in the next section.

  It is surprising, when Berners-Lee suggests using a 'standard' format, to find that 75 formats are used by two or less packages. What we are seeing is perhaps, taking a biological analogy, is an evolutionary explosion in 'species' in a new environment. For the sake of interoperability it may be hoped that 'survival of the fittest' will begin to act. It seems that linked data is still in an experimental phase.

4.7 HOW IS *THE CLOUD* LINKED?

The most important part of *The Cloud* is how the packages are linked together. *The Data Hub* site allows us to see the detail of the links.

The ten most commonly linked to packages, in terms of the number of packages linking, are:

Package being linked to	Number of packages linking	Number of links
DBpedia	158	31,531,365
GeoNames Semantic Web	42	9,353,935
[none]	34	0
DBLP Computer Science Bibliography (RKBExplorer)	27	1,338,927
Association for Computing Machinery (ACM) (RKBExplorer)	26	1,487,410
ePrints3 Institutional Archive Collection (RKBExplorer)	26	281,385
Freebase	25	10,452,728
CiteSeer (Research Index) (RKBExplorer)	24	805,921
School of Electronics and Computer Science, University of Southampton (RKBExplorer)	24	37,996
ReSIST Project Wiki (RKBExplorer)	24	408

The clear 'winners' are *DBpedia*, *GeoNames Semantic Web*, and *Freebase*. These are linked to by 50.8%, 13.5% and 8.0% of the other packages in *The Cloud*. It is supposed that this success is due their being well-known.

The six packages in the list with '*(RKBExplorer)*' at the end of names are part of a mini-cloud of about 50 packages. RKBExplorer[24] is a system for publishing linked data, developed during the EC-funded *ReSIST*[25] project. It has a browser that allows users to explore the interlinked data sets.

It is interesting, and perhaps at first glance surprising, to note that over 10% of the packages in *The Cloud* do not link to other packages. They are generally linked to, or have been published in order to be linked to. Included in this group are some of the largest packages, e.g. *Data.gov*, *2000 U.S. Census in RDF (rdfabout.com)*, *data.gov.uk Time Intervals*, *UniParc*, *The Open Library*, and *GeneID*.

The ten most commonly linked to packages, in terms of number of links, are:

Package being linked to	Number of packages linking	Number of links
UniProtKB Taxonomy	6	46,630,898
MARC Codes List	3	42,409,958
QDOS	1	40,000,000
UniProtKB	10	33,447,122
DBpedia	158	31,531,365
Ordnance Survey Linked Data	16	29,717,902
UniParc	1	27,534,215

[24] See: **http://www.rkbexplorer.com**
[25] See: **http://www.resist-noe.org**

Package being linked to	Number of packages linking	Number of links
IdRef: Sudoc authority data	3	20,040,000
Sudoc bibliographic data	1	20,000,000
flickr™ wrappr	4	16,358,998

DBpedia is the only package to appear in this and the previous list, which reinforces its 'popularity'. *flickr™ wrappr* is extensively linked from DBpedia to provide images for its concepts.

Packages with 'UniProt' at the beginning of their name, and the *UniParc* package, are part of a mini-cloud of the subject of proteins.

Sudoc is the French academic union catalogue, and the links here are between packages related to it.

Ordnance Survey Linked Data is geographical data for the UK, and linked to by packages from that country, especially UK government data packages.

QDOS is connected to a package dealing with popular music.

This analysis shows that the linking of packages is not something that is, at least at the moment, growing in an 'organic' way. There are initiatives which are responsible for creating large parts of *The Cloud*. The implication is that for the cultural heritage sector that such an initiative needs to happen too. Europeana is taking a leading role in such an initiative[26].

[26] See: **http://version1.europeana.eu/web/lod**

4.8 CULTURAL HERITAGE DATA IN *THE CLOUD*

There are 18 packages in *The Cloud* that could be identified as having 'cultural heritage' as their subject or related to it:

Package	IPR	Number of triples
VIAF: The Virtual International Authority File	[not given]	200,000,000
Europeana Linked Open Data	[not given][27]	185,000,000
British National Bibliography (BNB)	CC0	80,249,538
Hungarian National Library (NSZL) catalog	[not given]	19,300,000
Amsterdam Museum as Linked Open Data in the Europeana Data Model	CC BY-SA	5,000,000
Library of Congress Subject Headings	[not given]	4,151,586
Swedish Open Cultural Heritage	Other (Open)	3,400,000
Calames	[not given]	2,000,000
RAMEAU subject headings (STITCH)	[not given]	1,619,918
data.bnf.fr - Bibliothèque nationale de France	[not given]	1,400,000
National Diet Library of Japan subject headings	[not given]	1,294,669
Gemeenschappelijke Thesaurus Audiovisuele Archieven – Common Thesaurus Audiovisual Archives	ODbL	992,797
Gemeinsame Normdatei (GND)	Other (non-commercial)	629,582
Archives Hub Linked Data	CC0	431,088
Thesaurus for Graphic Materials (t4gm.info)	CC BY-SA	103,000
Italian Museums (LinkedOpenData.it)	CC BY-SA	49,897
Thesaurus W for Local Archives	[not given]	11,000
MARC Codes List Open Data	Other (Public Domain)	8,816

Two of the packages are directly related to Europeana: Amsterdam Museum and Europeana itself.

There is evidence of a French effort with linked data, especially terminologies: *Calames*, *RAMEAU subject headings (STITCH)*, *data.bnf.fr - Bibliothèque nationale de France*, *Thesaurus W for Local Archives*. This was also seen in the Linked Heritage partners' survey. Sweden is also doing something similar with *Swedish Open Cultural Heritage.* Italy is also starting to follow the same path.

There is an additional terminology and authority file component with: *VIAF: The Virtual International Authority File*, *British National Bibliography (BNB)*, *Library of Congress Subject Headings*, *National Diet Library of Japan subject headings, Gemeinsame Normdatei (GND)*, *Thesaurus for Graphic Materials (t4gm.info)* and the *MARC Codes List Open Data*.

Finally there is a contribution from the domains of libraries (*Hungarian National Library (NSZL) catalog*), archives (*Archives Hub Linked Data*), and audio-visual archives (*Gemeenschappelijke Thesaurus Audiovisuele Archieven – Common Thesaurus Audiovisual Archives*).

The part of *The Cloud* from cultural heritage is still rather small (c500m triples or <1.5%). However developments from Europeana are planned to significantly increase its size. Linked Heritage will be a significant component of it.

Let us further explore further details about the cultural heritage mini-cloud.

[27] This will eventually be published as *CC0*.

Cultural heritage packages use these formats:

Format	Number of packages using the format
Resource Description Framework	13
Simple Knowledge Organization System	11
Dublin Core	7
eXtensible HyperText Markup Language	4
Friend of a Friend	3
Basic Geo	1
Bibliographic Ontology	1
DBpedia	1
Music Ontology	1
Object Reuse and Exchange	1
RDF Schema	1
vCard	1
Web Ontology Language	1
XML Schema	1

The general picture is similar to *The Cloud* as a whole, except that the use of *SKOS* is much more significant, indicating the importance of terminological resources and authority files in the sector;

Of note is the absence of a format for museum information specifically. Also the Europeana Data Model is not mentioned in *The Data Hub*, but from other sources was used by Amsterdam Museum, and probably by the Europeana packages.

Cultural heritage packages in *The Cloud* link to:

Package being linked to	Number of packages linking	Number of links
DBpedia	5	82,308
Library of Congress Subject Headings	4	108,135
VIAF: The Virtual International Authority File	2	1,820,684
GeoNames Semantic Web	2	510,658
Dewey Decimal Classification (DDC)	2	200,543
RAMEAU subject headings (STITCH)	2	83,530
Swedish Open Cultural Heritage	1	100,489
Gemeinsame Normdatei (GND)	1	20,000
IdRef: Sudoc authority data	1	10,000
[DCMI Type Vocabulary – not in *The Cloud*]	1	10,000
UK Postcodes	1	5,000
AGROVOC	1	700
Hungarian National Library (NSZL) catalog	1	136
[none]	1	0

As one might expect *DBpedia* is the most popular package to link to. Another 'general' package linked to is *GeoNames Semantic Web.* Both of these were also identified in the Linked Heritage survey, and represent well known sources of cross-domain and geographical information to link to this.

Apart from this the rest of the linked packages are mainly other cultural heritage packages, and especially standard terminologies and authority files.

Looking at the use of serialisations:

Serialisation	Number of packages using (%)
RDF/XML	16 (88.9%)
N-Triples	5 (27.8%)
Turtle	1 (5.5%)
[none given]	1 (5.5%)

RDF/XML is used by all but two of the packages: *Europeana Linked Open Data* uses mentions only *N-Triples*, and the *Calames* Package does not mention any serialisation. *N-Triples* are usually published together with *RDF/XML*. The one occurrence of *Turtle* is in combination with *RDF/XML*.

This suggests that cultural heritage linked data should be, at least, published as *RDF/XML* and possibly as *N-Triples* in order to be compatible to existing data. However there is no reason why all the serialisations cannot be used.

5 STANDARDS LANDSCAPE FOR LINKED DATA

5.1 INTRODUCTION – DESCRIBING STANDARDS

Following a model developed during the *ATHENA* project, we describe each standard in a Dublin Core (DC) derived format. 9 out of the 15 DC elements are used in the descriptions.

These elements are:

Title	The name (or names) under which the standard is known. In most cases both the abbreviated and the full name is listed.
Creator	The name of the organisation or individual who originally created the standard.
Publisher	The name of the organisation that makes the standard publicly available.
Date	The date on which the standard was originally published.
Identifier	A number or other identifier under which standard is published or a URL which points to the definition of the standard.
Rights	Whether rights restrictions apply to the standard.
Description	A textual description explaining the standard and its usage.
Subject	Keywords that identify the nature of the standard.
Relation	Other standards that this standard relates to, and associated websites.

The descriptions are aimed at a general reader in the cultural heritage sector. More technical details for the majority of standards discussed can be found in various places on the Web. The purpose of this section is to allow the reader to have an easy reference to the range of relevant standards in one place.

The standards described here are **open** and mostly *de facto*.[28] We have classified the standards into:

- Key standards;
- Format standards;
- IPR licence standards.

Note that in a description not all DC elements will be present and some elements will be repeated.

[28] For a discussion of different types of standard see: **McKenna, Gordon** and **De Loof, Chris**. 2009. *ATHENA D3.1 – Report on existing standards applied by European museums*, pp4-5. ATHENA Project.
Download from: **http://www.athenaeurope.org/getFile.php?id=396**

5.2 BASIC STANDARDS

The creation of linked data uses a number of standards, including some that are called 'formats' in *The Data Hub* website package records:

Title	***URI (Uniform Resource Identifier)***
Creator	Berners-Lee, T (W3C/MIT); Fielding, R (Day Software); Masinter, L (Adobe Systems)
Publisher	The Internet Society
Date	2005 (current standard) [original concepts in 1990]
Identifier	**http://www.rfc-editor.org/rfc/rfc3986.txt** (generic syntax)
Rights	[Open Standard]
Description	String of characters used to identify a name or a resource on the Internet. ***Form:*** The syntax of a URI is: `[scheme name]:[scheme-specific part]` • `scheme name` – includes examples as "`http`", "`ftp`", "`mailto`", file, or "`urn`" followed by a colon character, and then by a scheme-specific part • `scheme-specific part` – these are specified in the rules of the scheme. However they must conform to the general requirements for URIs. These include the rules on the use of particular characters. URLs and URNs are URIs. Berners-Lee's *Principles* say that with linked data the URI should be an HTTP URI
Subject	identifier (Internet)
Relation	**http://tools.ietf.org/html/rfc1738 (**URL)
	http://tools.ietf.org/html/rfc2141 (URN)

Title	***URL (Uniform Resource Locator)***
Creator	T Berners-Lee (CERN), L Masinter (Xerox Corporation) & M McCahill (University of Minnesota) [Eds.]
Publisher	Internet Engineering Task Force (IETF)
Date	1994 [original]
Identifier	**http://tools.ietf.org/html/rfc1738**
Rights	[Open Standard]

Description	A URI (i.e. a string) that specifies: • Where a resource is available; • The mechanism for retrieving it. ***Form***: **`scheme://domain:port/path?query_string#fragment_id`** • **`scheme`** – defines the namespace, purpose, and the syntax of the remaining part, examples: http, https, gopher, wais, ftp. • **`domain:port`** – gives the destination location for the resource (domain name or IP address). Port is optional, if absent the default is used (for http default port = 80). • **`path`** – used to specify and find the resource • **`?query_string`** – used to pass data to a piece of software to enable retrieval • **`fragment_id`** – used to specify a part or a position within the overall resource E.g. **`http://www.athenaeurope.org/index.php?en/91/information-on-the-project`** (the 'About us' page on ATHENA project website
Subject	identifier (Internet);
Relation	**http://www.rfc-editor.org/rfc/rfc3986.txt (**URI)
	http://tools.ietf.org/html/rfc2141 (URN)

Title	***URN (Uniform Resource Name)***
Creator	Network Working Group (ed. R Moats, AT&T)
Publisher	Internet Engineering Task Force (IETF) (syntax); IANA, the Internet Assigned Numbers Authority (namespace assignment).
Date	1997
Identifier	**http://tools.ietf.org/html/rfc2141** (Syntax)
Rights	[Open Standard]
Description	String acting as persistent, location-independent, resource identifiers, designed to make it easy to map other namespaces. Note that they do not point to a location and therefore might not be resolvable. ***Form***: **`urn:<NID>:<NSS>`** **`<NID>`** is the Namespace Identifier, and **`<NSS>`** is the Namespace Specific String. The Namespace ID determines the syntactic interpretation of the Namespace Specific String.

Description [Continued]	E.g. **`urn:isbn:0451450523`** is URN for *The Last Unicorn*, identified by its book number. Example namespaces: ISBN; ISSN; ISAN; NBN[29]
Subject	identifier (Internet)
Relation	**http://www.rfc-editor.org/rfc/rfc3986.txt (**URI)
	http://tools.ietf.org/html/rfc1738 (URL)

RDF (including RDFS) is the key standard for linked data. It is specified by a set of W3C recommendations that were published in 2004. There was an earlier recommendation in 1999, but there are few applications now based on this. Here we describe each specification separately:

Title	***Resource Description Framework (RDF): Concepts and Abstract Syntax***
Creator	Graham Klyne (Nine by Nine) and Jeremy J. Carroll (Hewlett Packard Labs) [Eds.]
Publisher	World Wide Web Consortium (W3C)
Date	2004
Identifier	**http://www.w3.org/TR/2004/REC-rdf-concepts-20040210**
Rights	World Wide Web Consortium (W3C) [Open Standard]
Description	Defines an abstract syntax on which RDF is based. The standard includes: • Motivation; • Design Goals; • RDF Concepts; • RDF Vocabulary URI and Namespace (Normative); • Datatypes (Normative); • Abstract Syntax (Normative); • Fragment Identifiers.
Subject	semantic web
	linked data
Relation	**http://www.w3.org/2003/03/Translations/byTechnology?technology=rdf-concepts** [links to: Chinese, Simplified; Chinese, Traditional; French; Hungarian; Japanese; and Russian translations]

[29] National Bibliography Number. These are identifiers used by national libraries for those documents (e.g. web pages) where there is no identifier given by the publisher (e.g. an ISBN). The URN namespace for NBNs is described in RFC 3188 (**http://tools.ietf.org/html/rfc3188**). Some national libraries have resolution services for these URNs.

Title	***RDF Semantics***
Creator	Patrick Hayes (IHMC) [Ed.]
Publisher	World Wide Web Consortium (W3C)
Date	2004
Identifier	**http://www.w3.org/TR/2004/REC-rdf-mt-20040210**
Rights	World Wide Web Consortium (W3C) [Open Standard]
Description	The precise semantics and corresponding complete systems of inference rules. The standard includes: • Interpretations; • Simple Entailment between RDF graphs; • Interpreting the RDF vocabulary; • Interpreting the RDFS Vocabulary • Interpreting Datatypes; • Monotonicity of Semantic Extensions; • Entailment Rules (Informative).
Subject	semantic web
	linked data
Relation	**http://www.w3.org/2003/03/Translations/byTechnology?technology=rdf-mt** [links to: French; and Hungarian]

Title	***RDF Primer***
Creator	Frank Manola and Eric Miller (W3C) [Eds.]
Publisher	World Wide Web Consortium (W3C)
Date	2004
Identifier	**http://www.w3.org/TR/2004/REC-rdf-primer-20040210**
Rights	World Wide Web Consortium (W3C) [Open Standard]

Description	A technical introduction to RDF. The standard includes: • Making Statements About Resources; • An XML Syntax for RDF: RDF/XML; • Other RDF Capabilities; • Defining RDF Vocabularies: RDF Schema; • Some RDF Applications: RDF in the Field; • Other Parts of the RDF Specification.
Subject	semantic web
	linked data
Relation	http://www.w3.org/2003/03/Translations/byTechnology?technology=rdf-primer [links to: Chinese, Simplified; French; Hungarian; and Japanese translations]

Title	***RDF Vocabulary Description Language 1.0: RDF Schema***
Creator	Dan Brickley (W3C) and R.V. Guha (IBM) [Eds.]
Publisher	World Wide Web Consortium (W3C)
Date	2004
Identifier	http://www.w3.org/TR/2004/REC-rdf-schema-20040210
Rights	World Wide Web Consortium (W3C) [Open Standard]

Description	Describes how to use RDF to describe RDF vocabularies. It defines a vocabulary for this and defines other built-in RDF vocabulary. The standard includes:
	• Classes:
	  ○ rdfs:Resource;
	  ○ rdfs:Class;
	  ○ rdfs:Literal;
	  ○ rdfs:Datatype;
	  ○ rdf:XMLLiteral;
	  ○ rdf:Property.
	• Properties:
	  ○ rdfs:range;
	  ○ rdfs:domain;
	  ○ rdf:type;
	  ○ rdfs:subClassOf;
	  ○ rdfs:subPropertyOf;
	  ○ rdfs:label;
	  ○ rdfs:comment.

Description [continued]	• Using the Domain and Range vocabulary (Informative)
	• Other vocabulary:
	  ○ Container Classes and Properties;
	  ○ RDF Collections;
	  ○ Reification Vocabulary;
	  ○ Utility Properties.
	• RDF Schema summary (Informative)

Subject	semantic web
	linked data
Relation	**http://www.w3.org/2003/03/Translations/byTechnology?technology=rdf-schema** [links to: French; Hungarian; and Japanese translations]

Title	***RDF/XML Syntax Specification*** [see below in serialisations section]

Title	***RDF Test Cases***
Creator	Jan Grant (ILRT, University of Bristol) and Dave Beckett (ILRT, University of Bristol) [Eds.]
Publisher	World Wide Web Consortium (W3C)
Date	2004
Identifier[30]	**http://www.w3.org/TR/2004/REC-rdf-testcases-20040210**
Rights	World Wide Web Consortium (W3C) [Open Standard]
Description	A set of test cases corresponding to technical issues addressed by the RDF Working Group. The test case themselves are machine processable.
Subject	semantic web
	linked data
Relation	**http://www.w3.org/2003/03/Translations/byTechnology?technology=rdf-testcases** [links to: French; and Hungarian translations]

[30] Contains the definition of the RDF serialisation *N-Triples* (see below).

RDF graphs have to be transformed into a form that can be processed by machines. The process is called serialisation. RDF is serialised in four main ways:

Title	***RDF/XML Syntax Specification***
Creator	Dave Beckett (University of Bristol) [Eds.]
Publisher	World Wide Web Consortium (W3C)
Date	2004
Identifier	**http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210 [Revised]**
Rights	World Wide Web Consortium (W3C) [Open Standard]

Description	Defines a XML syntax for RDF.
	The standard includes:
	• An XML Syntax for RDF:
	  ○ Introduction;
	  ○ Node Elements and Property Elements
	  ○ Multiple Property Elements
	  ○ Empty Property Elements
	  ○ Property Attributes
	  ○ Completing the Document: Document Element and XML Declaration
	  ○ Languages: xml:lang
	  ○ XML Literals: rdf:parseType="Literal"
	  ○ Typed Literals: rdf:datatype
	  ○ Identifying Blank Nodes: rdf:nodeID
	  ○ Omitting Blank Nodes: rdf:parseType="Resource"
	  ○ Omitting Nodes: Property Attributes on an empty Property Element
	  ○ Typed Node Elements
	  ○ Abbreviating URI References: rdf:ID and xml:base
	  ○ Container Membership Property Elements: rdf:li and rdf:_n
	  ○ Collections: rdf:parseType="Collection"
	• Reifying Statements: rdf:ID3 Terminology;
	• RDF MIME Type, File Extension and Macintosh File Type;
	• Global Issues;
	• The RDF Namespace and Vocabulary:
	  ○ Identifiers;
	  ○ Resolving URIs;
	  ○ Constraints;
	  ○ Conformance.
	• Syntax Data Model:
	  ○ Events;
	  ○ Information Set Mapping;
	  ○ Grammar Notation.
	• RDF/XML Grammar:
	  ○ Grammar Summary;
	  ○ Grammar Productions;
	  ○ Reification Rules;
	  ○ List Expansion Rules.

Description [continued]	• Serializing an RDF Graph to RDF/XML; • Using RDF/XML with HTML and XHTML; • Using RDF/XML with SVG.
Subject	semantic web
	linked data
	RDF serialisation
Relation	**http://www.w3.org/2003/03/Translations/byTechnology?technology=rdf-syntax-grammar** [links to French, Hungarian and Japanese translations]

Title	***Notation3 (or N3)***
Creator	Tim Berners-Lee (W3C) and Dan Connolly (W3C) [Eds.]
Publisher	World Wide Web Consortium (W3C)
Date	2011
Identifier	**http://www.w3.org/TeamSubmission/n3**
Rights	World Wide Web Consortium (W3C) [Open Standard]
Description	Defines a non-XML syntax for RDF. It is designed so that it easier to write by hand, and sometimes easier to follow than RDF/XML. It uses a tabular notation, and therefore makes RDF triples easily recognisable. The standard includes: • Grammar; • Syntax details; • Semantics; • Notes on Numbers; • Appendix: N3 Subsets. It is closely related to the other serialisations: *Turtle* and *N-Triples*
Subject	semantic web
	linked data
	RDF serialisation
Relation	**http://www.w3.org/TeamSubmission/turtle** [Turtle]
	http://www.w3.org/2001/sw/RDFCore/ntriples [N-Triples]

Title	***Turtle***
Creator	David Beckett and Tim Berners-Lee (W3C) [Eds.]
Publisher	World Wide Web Consortium (W3C)
Date	2011
Identifier	**http://www.w3.org/TeamSubmission/turtle** [working draft]
Rights	World Wide Web Consortium (W3C) [Open Standard]
Description	Defines a non-XML syntax for RDF. A sub-set of Notation3 (restricted to RDF) and a super-set of N-Triples. The standard includes: • Turtle Syntax; • Turtle Grammar; • Examples; • Identifiers for the Turtle Language; • Conformance; • Media Type and Content Encoding; • Turtle compared to N-Triples; • Turtle compared to Notation3; • Turtle compared to SPARQL. It is closely related to the other serialisations: *Notation3* and *N-Triples*
Subject	semantic web
	linked data
	RDF serialisation
Relation	**http://www.w3.org/TeamSubmission/turtle** [Notation 3 or N3]
	http://www.w3.org/2001/sw/RDFCore/ntriples [N-Triples]

Title	***N-Triples***
Creator	Jan Grant (ILRT, University of Bristol) and Dave Beckett (ILRT, University of Bristol) [Eds.]
Publisher	World Wide Web Consortium (W3C)
Date	2004
Identifier[31]	**http://www.w3.org/TR/rdf-testcases/#ntriples**

[31] The definition forms part of the *Test Cases* document of RDF (see above).

Rights	World Wide Web Consortium (W3C) [Open Standard]
Description	Defines a non-XML syntax for RDF. Designed to be simpler than *Notation3* and *Turtle*, and therefore easier for software to generate and parse. However this means that it can be difficult to write and read by humans. The standard includes: • Extended Backus-Naur Form (EBNF) Grammar; • Strings; • URI References; • Example; • Tests.
Subject	semantic web
	linked data
	RDF serialisation
Relation	**http://www.w3.org/TeamSubmission/turtle** [Notation 3 or N3]
	http://www.w3.org/TeamSubmission/turtle [Turtle]

Title	SPARQL Protocol and RDF Query Language (SPARQL)
Creator	RDF Data Access Working Group[32] (part of W3C Semantic Web Activity) (Eds. Eric Prud'hommeaux (W3C) and Andy Seaborne (Hewlett-Packard Laboratories, Bristol)
Publisher	World Wide Web Consortium (W3C)
Date	2008
Identifier	**http://www.w3.org/TR/rdf-sparql-query**
Rights	World Wide Web Consortium (W3C) [Open Standard]
Description	A query language for RDF. It was designed to meet the use cases and requirements identified by the RDF Data Access Working Group. These test cases included: • Finding: An email address; information about motorcycle parts; new things about people; film soundtracks; what people say about news stories; learning resources. • Activities like: Browsing patient records; monitoring news events; avoiding traffic jams; exploring the neighbourhood; sharing photographs with a friend;

[32] Group now called – *SPARQL Working Group*. See**:** **http://www.w3.org/2001/sw/DataAccess/homepage-20080115**

Description [continued]	The standard includes: • Making Simple Queries (Informative); • RDF Term Constraints (Informative); • SPARQL Syntax; • Graph Patterns; • Including Optional Values; • Matching Alternatives; • RDF Dataset; • Solution Sequences and Modifiers; • Query forms; • Testing Values; • Definition of SPARQL. Data is made accessible through a standards-compliant service called a 'SPARQL endpoint' which sits on top of the knowledge base of linked data. This allows four query types: • **`SELECT`** – Returns data values in a table defined in the query; • **`CONSTRUCT`** – Returns data values which are transformed into valid RDF defined in the query. This can then be serialised into, for example RDF/XML; • **`ASK`** – Gives a Boolean yes/no result for a query; • **`DESCRIBE`** – Returns an RDF graph with information that meets the query. The form of the graph is decided by the provider of the knowledgebase and not by the query itself. Each type takes a '**`WHERE`**', which restricts the query. **`WHERE`** is optional for the **`DESCRIBE`** type.
Subject	query language (RDF)
	semantic web
	linked data
Relation	**http://www.w3.org/2003/03/Translations/byTechnology?technology=sparql-query** [Links to translations]
	http://www.w3.org/TR/1999/REC-rdf-syntax-19990222 [RDF]
	http://www.w3.org/TR/rdf-dawg-uc [use case document]

Title	eXtensible Markup Language (xml)
Creator	World Wide Web Consortium (W3C)
Publisher	World Wide Web Consortium (W3C)
Date	2008
Identifier	**http://www.w3.org/TR/REC-xml/** (Fifth Edition)
Rights	[Open Standard]
Description	XML started as 'SGML light', to overcome some of its complexities and terseness. Soon if overhauled its ancestor in popularity and is now the most widely used structuring language for electronic documents. XML structures a document by 'tagging' texts. The tags can be freely defined, but can be controlled by a Document Type Definition (DTD) or an XML-schema. XML uses the Unicode character set, so that it is very usable in multi-lingual and international applications. Several XML derivates have been standardized, such as XSLT (eXtensible Style Language and Transformation) and the xPath query syntax.
Subject	document structure
	document encoding
Relation	SGML
	HTML
	Unicode

Title	***XML Schema Definition Language (xsd)***
Creator	World Wide Web Consortium (W3C) [various ediitors]
Publisher	World Wide Web Consortium (W3C)
Date	2004
Identifier	**http://www.w3.org/TR/xmlschema-0** [Primer]
	http://www.w3.org/TR/xmlschema-1 [Structures]
	http://www.w3.org/TR/xmlschema-2 [Datatypes]
Rights	World Wide Web Consortium (W3C) [Open Standard]
Description	Defines a class of XML documents in terms a set of rules (structure and data types) to which a document must conform in order to be considered 'valid'. The standard is defined in a set of separate documents (for two versions):

	W3C XML Schema 1.0 Specification: • Primer; • Structures; • Datatypes.
Subject	document structure
	document data types
	document encoding

Title	***Object Reuse and Exchange (ore)***
Creator	Open Archives Initiative
Publisher	Open Archives Initiative
Date	2008
Identifier	**http://www.openarchives.org/ore/1.0/toc.html** [Specifications and User Guides]
Rights	[Open Standard]
Description	Used for the description and exchange of 'aggregations' of web resources. It defines: • ***Aggregations*** – Conceptual resources, i.e. not 'concrete', identified by an URI. As such have relationships with other resources, including aggregations of aggregation. • ***Aggregated resources*** – Resources that are part of an aggregation; • ***Resource maps*** – Resource describing an aggregation using a set of assertions. Which aggregation is being described is a mandatory assertion. Other assertions indicate the aggregated resource(s). Certain metadata are mandatory as well, such as the map creator. Dublin Core is used for mandatory metadata, like the creator of the map. • ***Proxies*** – Optional virtual resources that are proxies for aggregated resource in an aggregation. The standard consists of a set of documents: • A primer; • User guides on: o Resource map implementation in Atom; o Resource map implementation in RDF/XML; o Resource map implementation in RDFa; o HTTP implementation; o Resource map discovery • Specifications: o Abstract Data Model; o Vocabulary

	• Tools and resources. It is important here because EDM uses ORE.
Subject	linked data (aggregation)
Relation	**http://dublincore.org/documents/dc-rdf** [Expressing DC as RDF]

Title	***Web Ontology Language 2***[33] ***(owl)***
Creator	W3C OWL Working Group
Publisher	World Wide Web Consortium (W3C)
Date	2009
Identifier	**http://www.w3.org/TR/owl2-overview** [Overview of OWL 2]
Rights	World Wide Web Consortium (W3C) [Open Standard]
Description	A Semantic Web computational logic-based language designed to represent rich and complex knowledge about: • Things; • Groups of things; • Relations between things. Knowledge expressed in OWL can be reasoned with by computer programs to: • Verify the consistency of that knowledge; • Make implicit knowledge explicit. OWL documents are known as 'ontologies', and: • Can be published on the Web; • May refer to or be referred from other OWL ontologies. The **`owl:sameas`** property is especially useful when one wants to assert that one thing in a published package is the same as another thing in a different package.
Subject	RDF (ontololgy)
	linked data (ontology)
Relation	**http://www.w3.org/TR/2009/REC-owl2-syntax-20091027** [Structural Specification and Functional-Style Syntax]
	http://www.w3.org/TR/2009/REC-owl2-mapping-to-rdf-20091027 [Mapping to RDF Graphs]
	http://www.w3.org/TR/2009/REC-owl2-direct-semantics-20091027 [Direct Semantics]

[33] A description only OWL2 is given.

	http://www.w3.org/TR/2009/REC-owl2-rdf-based-semantics-20091027 [RDF-Based Semantics]
	http://www.w3.org/TR/2009/REC-owl2-conformance-20091027 [Conformance]
Relation [continued]	**http://www.w3.org/TR/2009/REC-owl2-profiles-20091027** [Profiles]
	http://www.w3.org/TR/2009/REC-owl2-primer-20091027 [Primer]
	http://www.w3.org/TR/2009/REC-owl2-new-features-20091027 [New Features and Rationale]
	http://www.w3.org/TR/2009/REC-owl2-xml-serialization-20091027 [XML Serialization]
	http://www.w3.org/TR/2009/REC-rdf-plain-literal-20091027 [rdf:PlainLiteral: A Datatype for RDF Plain Literals]

5.3 FORMAT STANDARDS

Each package in *The Cloud* publishes RDF in a number of different formats, often in more than one format in each package. The most often used ones by cultural packages are:

5.3.1 Descriptive formats

Title	***Simple Knowledge Organization System (skos)***
Creator	Alistair Miles (STFC Rutherford Appleton Laboratory / University of Oxford) and Sean Bechhofer (University of Manchester) [Eds.]
Publisher	World Wide Web Consortium (W3C)
Date	2009
Identifier	**http://www.w3.org/TR/2009/REC-skos-reference-20090818** [Reference]
Rights	World Wide Web Consortium (W3C) [Open Standard]
Description	Designed for the publication of controlled structured vocabularies for the Semantic Web. These include: thesauri, classification schemes, taxonomies, and subject-headings. SKOS supports the publication of multilingual resources, and is built on RDF and RDFS.
Subject	linked data (controlled vocabulary)
	linked data (thesaurus)
	linked data (classification scheme)
	linked data (taxonomy)
	linked data (subject heading)
Relation	**http://www.w3.org/2003/03/Translations/byTechnology?technology=skos-reference** [Links to translations]

Title	***Dublin Core (dc)***
Creator	Mikael Nilsson (KMR Group, NADA, KTH (Royal Institute of Technology), Sweden); Andy Powell (Eduserv Foundation, UK); Pete Johnston (Eduserv Foundation, UK); and Ambjörn Naeve (KMR Group, NADA, KTH (Royal Institute of Technology), Sweden)
Publisher	Dublin Core Metadata Initiative
Date	2008
Identifier	**http://dublincore.org/documents/dc-rdf** [Expressing DC as RDF]

Rights	Dublin Core Metadata Initiative
Description	A simple metadata element set intended to facilitate discovery of electronic resources. Elements can be grouped into those having data on: Content – Coverage, Description, Type, Relation, Source, Subject, Title; Intellectual Property – Contributor, Creator, Publisher, Rights; Instantiation – Date, Format, Identifier, Language. Its use has been mandated by several governments in Europe (e.g. UK) and throughout the world (e.g. Australia).
Subject	linked data (resource)
Relation	**http://dublincore.org/documents/1999/07/02/dces/** [Dublin Core Element Set]

Title	***Friend of a Friend (foaf)***
Creator	Dan Brickley and Libby Miller
Publisher	FOAF Project
Date	2000 onwards
Identifier	**http://xmlns.com/foaf/spec** [Vocabulary Specification 0.98]
Rights	Creative Commons Attribution
Description	A format, using RDF and OWL, for describing persons, their relations to other persons and things, and their activities. It can describe: • Basic information (e.g. surname, given name, and image of); • Personal information (e.g. where based, interests, and who they know); • Online accounts and instant messaging (a person's IDs in various online services); • Projects and groups (defining membership); • Documents and images (describing related resouces).
Subject	linked data (person)
Relation	**http://www.foaf-project.org** [FOAF Project website]

Title	***Basic Geo (geo)***
Creator	W3C Semantic Web Interest Group
	Dan Brickley [Ed.]
Publisher	World Wide Web Consortium (W3C)
Date	2006

Identifier	**http://www.w3.org/2003/01/geo**
Rights	[?]
Description	A RDF vocabulary for basic geographical information: latitude, longitude, and altitude.
Subject	linked data (geographic)

Title	***Bibliographic Ontology (bibo)***
Creator	Frédérick Giasson (Structured Dynamics) [ed.]
	Bruce D'Arcus and Frédérick Giasson (Structured Dynamics)
Publisher	Structured Dynamics LLC
Date	2009
Identifier	**http://bibliontology.com/specification** [Revision: 1.3]
	http://bibotools.googlecode.com/svn/bibo-ontology/trunk/doc/index.html [Documentation pages]
Rights	Creative Commons Attribution
Description	A format, using RDF, for describing bibliographic items like books, magazines, and newspaper pages.
Subject	linked data (bibliographic)
	linked data (book)
	linked data (magazine)
	linked data (newspaper)
Relation	**http://bibotools.googlecode.com/svn/bibo-ontology/trunk/doc/index.html** [Ontology]

Title	***Music Ontology (mo)***
Creator	Yves Raimond (BBC) and Frédérick Giasson (Structured Dynamics) [Eds.]
	Yves Raimond (BBC); Frédérick Giasson (Structured Dynamics); Kurt Jacobson (Centre for Digital Music, Queen Mary, University of London); George Fazekas (Centre for Digital Music, Queen Mary, University of London); Thomas Gängler (Faculty of Computer Science, University of Technology Dresden, Germany); and Simon Reinhardt [Authors]
Publisher	SourceForge

Date	2006 onwards
Identifier	**http://musicontology.com**
Rights	Creative Commons Attribution
Description	Contains the concepts and properties for describing music, for example: • Artists; • Albums; • Tracks; • Performances; • Arrangements.
Subject	linked data (music)

Title	***vCard (vcard)***
Creator	S. Perreault (Viagenie)
Publisher	Internet Engineering Task Force (IETF)
Date	1998 onwards
Identifier	**http://tools.ietf.org/html/rfc6350** [Version 4.0 specification]
Rights	IETF Trust and the persons identified as the document authors
Description	Standard for describing electronic business cards. It can contain: • Names; • Addresses; • Phone numbers; • E-mail addresses; • Web URLs; • Logos; • Photographs; • Media clips.
Subject	linked data (business card)

We include here information for the Europeana Data Model even though it is not specifically mentioned in *The Data Hub*.

Title	***Europeana Data Model***
Creator	Europeana v1.0 Project
Publisher	Europeana Foundation
Date	2011
Identifier	http://group.europeana.eu/c/document_library/get_file?uuid=aff89c92-b6ff-4373-a279-fc47b9af3af2&groupId=10605 [Version 5.2]
	http://group.europeana.eu/c/document_library/get_file?uuid=718a3828-6468-4e94-a9e7-7945c55eec65&groupId=10605 [Primer]
Rights	Europeana Foundation [Open Standard]
Description	Created for structuring data for Europeana ingestion, management and publication, and improves on Europeana's basic data model the Europeana Semantic Elements (ESE). EDM attempts to go beyond information perspectives of the cultural heritage domains in Europeana: museums, archives, audiovisual collections and libraries. It uses a cross-domain, Semantic Web based approach, which takes account of community standards like: LIDO for museums, EAD for archives or METS for digital libraries. It has classes: • From other namespaces: ORE **Aggregation**; ORE **Proxy**; RDFS **Resource**; SKOS **Concept**. • Of its own: **Agent; Europeana Aggregation; Europeana Object; Event; Information Resource; Non-Information Resource; Physical Thing; Place; Time Span; Web Resource.** It has properties: • From other namespaces: ORE **Aggregates**; ORE **Proxy For**; ORE **Proxy In**. • Of its own: **Aggregated Cultural Heritage Object; Current Location; Happened At; Has Met; Has Type; Has View; Incorporates; Is Annotation Of; Is Derivative Of; Is Next in Sequence; Is Related To; Is Representation Of; Is Similar To; Is Successor Of; Landing Page; Occurred At; Realizes; Was Present At.** It also integrates the elements of ESE.
Subject	linked data (cross-domain)
	linked data (Europeana)

5.3.2 Package specific formats

Title	***DBpedia Ontology (dbpedia)***
Creator	Wikipedia Community
Publisher	Wikipedia Community
Date	2007 onwards
Identifier	**http://mappings.dbpedia.org/index.php?title=Special%3AAllPages&from=&to=&namespace=200** [Classes]
	http://mappings.dbpedia.org/index.php?title=Special%3AAllPages&from=&to=&namespace=202 [Properties]
Rights	Creative Commons Attribution Share Alike
Description	Cross-domain ontology, based on the 'infoboxes' in Wikipedia. It has over 320 classes described by 1650 properties.
Subject	linked data (cross-domain)
	linked data (DBpedia)
Relation	**http://wiki.dbpedia.org/Ontology** [webpages]

5.4 LICENCE STANDARDS FOR LINKED DATA

An organisation may choose to publish linked data under a licence of its own devising. However the analysis of *The Cloud* has identified some standard licences.

5.4.1 Open data licences

Title	***Creative Commons Attribution (CC BY)***
Creator	Creative Commons
Publisher	Creative Commons
Date	2007 [Version 3.0]
Identifier	**http://creativecommons.org/licenses/by/3.0/** [licence deed]
	http://creativecommons.org/licenses/by/3.0/legalcode [legal code]
Rights	Creative Commons Attribution 3.0
Description	Allows the work to be: • Shared – copied, distributed and transmitted; • Remixed – adapted; • Used commercial purposes. However the licensee must: • Attribute the author of the work – In the way specified by the author or licensor. This attribution must not suggest that the licensee's use is endorsed.
Subject	IPR licence (open data)
Relation	**http://creativecommons.org** [Creative Commons website]

Title	***Creative Commons Attribution Share Alike (CC BY-SA)***
Creator	Creative Commons
Publisher	Creative Commons
Date	2007 [Version 3.0]
Identifier	**http://creativecommons.org/licenses/by-sa/3.0/** [licence deed]
	http://creativecommons.org/licenses/by-sa/3.0/legalcode [legal code]
Rights	Creative Commons Attribution

Description	Allows the work to be: • Shared – copied, distributed and transmitted; • Remixed – adapted; • Used for commercial purposes (but note the share alike, commercial, derivatives must also be available for re-use.). However the licensee must: • **Attribute** the author of the work – In the way specified by the licensor. This attribution must not suggest that the licensee's use is endorsed. • **Share Alike** any derivative works – These must be under the same (or similar licence) to that obtained from the original author.
Subject	IPR licence (open data)
Relation	**http://creativecommons.org** [Creative Commons website]

Title	***Creative Commons CC Zero (CC0)***
Creator	Creative Commons
Publisher	Creative Commons
Date	2009
Identifier	**http://creativecommons.org/publicdomain/zero/1.0/** [waiver deed]
	http://creativecommons.org/publicdomain/zero/1.0/legalcode [legal code]
Rights	Creative Commons Attribution 3.0
Description	A permanent waiver (as opposed to a licence) of all rights to the work being published. It allows any use, without attribution.
Subject	IPR waiver (open data)
Relation	**http://creativecommons.org** [Creative Commons website]

Title	***GNU Free Documentation Licence (GNU FDL)***
Creator	Free Software Foundation, Inc.
Publisher	Free Software Foundation, Inc.
Date	2008 (version 1.3)
Identifier	**http://www.gnu.org/copyleft/fdl.html**

Rights	Free Software Foundation, Inc (Can freely copy and distribute copies, but no changes are allowed)
Description	A 'copyleft' licence designed for the free documentation of software, but which can be used for other text works. The work licensed can be used for any purpose, including commercial. However there are conditions: • Attribution – All earlier authors of the work must be attributed; • Changes – All changes to the work must be recorded; • Derivatives – must be licensed under the same licence; • Licence maintenance – The licence text must be kept up to date, and include previous versions; • Digital Rights Management – is not allowed. It is similar to CC BY-SA.
Subject	IPR licence (open data)
Relation	**http://www.gnu.org** [GNU website]

Title	***GNU General Public Licence (GNU GPL)***
Creator	Free Software Foundation, Inc. (Original author: Richard Stallman)
Publisher	Free Software Foundation, Inc.
Date	2007 (Version 3)
Identifier	**http://www.gnu.org/copyleft/gpl.html**
Rights	Free Software Foundation, Inc.
Description	A free software licence granting the licensee the right to change and redistribute the software free of the prohibitions of copyright law. Note that 'free' equates with 'freedom' rather than with 'free price'. It is similar to CC BY-SA.
Subject	IPR licence (open data)
Relation	**http://www.gnu.org** [GNU website]

Title	***New BSD licence***
Creator	Regents of the University of California
Publisher	[Public Domain]
Date	1999
Identifier	**http://en.wikipedia.org/wiki/BSD_licenses#3-clause_license_.28.22New_BSD_License.22_or_.22Modified_BSD_License.22.29** [on Wikipedia]
Rights	Public Domain
Description	Originally used for the Berkeley Software Distribution (BSD), a Unix-like operating system. The relevant part of the licence says: Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met: • Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer. • Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution. • Neither the name of the <organization> nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.
Subject	IPR licence (open data)

Title	***Simplified BSD licence***
Creator	FreeBSD Project
Publisher	[Public Domain]
Date	[?]
Identifier	**http://en.wikipedia.org/wiki/BSD_licenses#2-clause_license_.28.22Simplified_BSD_License.22_or_.22FreeBSD_License.22.29** [on Wikipedia]
Rights	[Public Domain]
Description	Originally used for the Berkeley Software Distribution (BSD), a Unix-like operating system. The relevant part of the licence says: Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met: 1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer. 2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

Subject	IPR licence (open data)

Title	***Open Data Commons Public Domain Dedication and Licence (ODC PDDL)***
Creator	Open Knowledge Foundation
Publisher	Open Knowledge Foundation
Date	2008
Identifier	**http://opendatacommons.org/licenses/pddl/1-0/**
Rights	Creative Commons Attribution
Description	Aimed at data in databases, the licensee can: • Share – copy, distribute and use the database; • Create – works from the database; • Adapt – change and build on the database. There is no attribution requirement.
Subject	IPR waiver (open data)
Relation	**http://opendatacommons.org** [Open Knowledge Foundation website]

Title	***Open Database Licence (ODbL)***
Creator	Open Knowledge Foundation
Publisher	Open Knowledge Foundation
Date	2009
Identifier	**http://opendatacommons.org/licenses/odbl/1.0/**
Rights	Creative Commons Attribution
Description	Aimed at data in databases, the licensee can: • Share – copy, distribute and use the database; • Create – works from the database; • Adapt – change and build on the database. However licensees must: • Attribute – Any public use of the database, or works produced from the database, in the manner specified in the original licence. Notices on the original database must be preserved; • Share-Alike – Any derivatives must use this licence;

	• Keep open – Digital Rights Management (DRM), or other measures may not be used with derivatives.
Subject	IPR licence (open data)
Relation	**http://opendatacommons.org** [Open Knowledge Foundation website]

Title	***UK Open Government Licence (OGL)***
Creator	Controller of Her Majesty's Stationery Office (HMSO) [developer]
	National Archives [deliverer]
Publisher	Her Majesty's Stationery Office (HMSO)
Date	2010
Identifier	**http://www.nationalarchives.gov.uk/doc/open-government-licence**
Rights	UK Open Government Licence (OGL)
Description	The licensee can: • Copy, publish, distribute and transmit the information; • Adapt the information; • Exploit the information commercially. However the licensee must: • Acknowledge the source of the Information (by an attribution statement specified by the provider(s) and, where possible, provide a link to the licence). • Not use the information in a way that suggests any official status or that the provider endorses licensee or the use; • Not mislead others or misrepresent the information or its source; • Not breach the UK *Data Protection Act 1998* or *the Privacy and Electronic Communications (EC Directive) Regulations 2003*. It is interoperable with Creative Commons Attribution (CC BY). It supersedes the UK Crown Copyright with data.gov.uk rights licensing arrangements.
Subject	IPR licence (open data)
Relation	**http://www.nationalarchives.gov.uk/documents/information-management/uk-government-licensing-framework.pdf** [UK Government Licensing Framework]

5.4.2 Un-open data licences

There is only one standard licence used in *The Cloud* that is not open:

Title	***Creative Commons Attribution Non-commercial (CC BY-NC)***
Creator	Creative Commons
Publisher	Creative Commons
Date	2007 [Version 3.0]
Identifier	**http://creativecommons.org/licenses/by-nc/3.0/** [licence deed]
	http://creativecommons.org/licenses/by-nc/3.0/legalcode [legal code]
Rights	Creative Commons Attribution 3.0
Description	Allows the work to be: • Shared – copied, distributed and transmitted; • Remixed – adapted. However the licensee must: • **Attribute** the author of the work – In the way specified by the author or licensor. This attribution must not suggest that the licensee's use is endorsed. • Must use the work for **non-commercial** purposes only.
Subject	IPR licence (not open)

6 CULTURAL METADATA STANDARDS

6.1 INTRODUCTION

Metadata standards were explored in the *ATHENA* project[34] which ended in April 2011. *Linked Heritage* has many partners in common with *ATHENA*, including the authors of this deliverable. A similar partner survey was carried out in earlier project with broadly similar results as will be given below. *ATHENA* also created deliverables[35] and a tool[36] for a cultural heritage audience which are relevant to this document.

6.2 STANDARDS LANDSCAPE

We will not repeat the work undertaken in ATHENA but make links to definition of the standard and relevant supporting documentation.

6.2.1 Museum (descriptive)

Standard	Link
CDWA	**http://www.getty.edu/research/institute/standards/cdwa/index.html**
LIDO	**http://www.lido-schema.org/schema/v1.0/lido-v1.0.xsd** (XML schema)
	http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf (specification document)
museumdat	**http://museum.zib.de/museumdat/museumdat-v1.0.xsd** (XML schema)
	http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf (specification document)
Object ID	**http://www.object-id.com/guide/guide_index.html**
SPECTRUM	**http://www.collectionstrust.org.uk/spectrum**
	http://www.collectionstrust.org.uk/schema (XML schema)
VRA	**http://www.vraweb.org/projects/vracore4/vra-4.0.xsd** (XML schema)
	http://www.vraweb.org/projects/vracore4/VRA_Core4_Element_Description.pdf

[34] See: **http://www.athenaeurope.org**

[35] **McKenna, Gordon** and **De Loof, Chris** (2009). *Report on existing standards applied by European Museums*. ATHENA Project. See: **http://www.athenaeurope.org/getFile.php?id=396**

McKenna, Gordon and **De Loof, Chris** (2009). *Recommendations and best practice report regarding the application of standards, including recommendations for a harvesting format and fact sheets for dissemination*. ATHENA Project. See: **http://www.athenaeurope.org/getFile.php?id=538**

[36] **McKenna, Gordon** and **De Loof, Chris** (2009). *Digitisation: Standards Landscape for European Museums, Archives, Libraries*. ATHENA Project. See: **http://www.athenaeurope.org/getFile.php?id=435**

6.2.2 Archive (descriptive)

Standard	Link
EAD	**http://www.loc.gov/ead/ead.xsd** (W3C schema)
ISAD(G)	**http://www.ica.org/sites/default/files/isad_g_2e.pdf**

6.2.3 Publisher (descriptive)

Standard	Link
ONIX	**http://www.editeur.org/files/ONIX%203/ONIX_BookProduct_XSD_schema+codes_Issue_15.zip** (W3C schema)
	http://www.editeur.org/files/ONIX%203/ONIX_for_Books_Release3-0_docs+codes_Issue_15.zip (documentation)

6.2.4 Technical (non-descriptive)

Standard	Link
METS	**http://www.loc.gov/standards/mets/mets.xsd**

6.2.5 Resource discovery

Standard	Link
Dublin Core	**http://dublincore.org/documents/1999/07/02/dces/**

6.2.6 Conceptual

Standard	Link
CIDOC-CRM	**http://cidoc.ics.forth.gr** (CRM website)
FRBR	**http://www.ifla.org/VII/s13/frbr/frbr.pdf** (Final report)
Indecs	**http://www.doi.org/topics/indecs/indecs_framework_2000.pdf** (Indecs framework)

6.3 PARTNERS SURVEY

As part of the survey partners' use of metadata standards was explored in questions 10-13[37]. The table below shows how many provider organisations use which metadata schemas. The schemas have been arranged by 'domain audience'. Where a schema does not appear in answers to the survey it is not included in the table. Figures in the columns ***Standard adapted*** and ***Standard not adapted*** do not straight forwardly relate to the respective standard. It means that providers using this standard (often amongst others) have adapted or not some (or all) of the (various) standards they use. However the information allows for drawing some important conclusions.

6.3.1 Results

Standard [domain audience]	All	Standards adapted	Standards not adapted	Provider Type: Museum	Library	Archive	Sound Archive	Publisher	Aggregator	Other
Number of provider types represented	47			4	5	4	1	0	10	23
Number of standards used by providers	98									
Museum (descriptive)	***17***									
CDWA	1	1							1	
LIDO	9	1	4	2		1			3	3
museumdat	1	1							1	
Object ID	1		1	1						
SPECTRUM	4	1	1	1					1	2
VRA	1	1								1
Archive, descriptive	***5***									
EAD	3	3	1			1			1	1
ISAD(G)	2	2				1				1
Publisher, descriptive	***4***									
ONIX	4	2	1						1	3
Technical, non-descriptive	***5***									
METS	5	3	2		1	2				2

[37] See *Appendix 1* below.

Resource Discovery	***20***										
Dublin Core	20	12	5	1	5	2	1		3	8	

Conceptual	***4***										
CIDOC CRM	3	1	2						1	2	
FRBR	1	1			1						

Other	***28***										
Other formats	22	1	9	1	3	4	1		5	8	
None	6									6	

6.3.2 Interpretation

Use of metadata

Unsurprisingly there is a variety of different types of metadata standard in use. However these types serve different needs. They can be classified into:

- Descriptive domain standards for museums, libraries, archives, and publishers;
- Standards of a technical, non-descriptive nature, some which are conceptual;
- A standard, Dublin Core, which aims at the discovery of resources, originally only web accessible resources.

From the table it can be seen that Dublin Core or non-standard, 'Other formats', are used by 50% of respondents. However a closer look shows that:

- Those who use Dublin Core serve also other standards. Dublin Core is used as a secondary standard.
- Those who use 'Other formats' also metadata standards. This means that they can export different types of metadata.

In the museum domain the LIDO harvesting schema is the leading one, while in the library and archive domains the picture is more diverse.

Adaptation of metadata

50% of those who answered the question: "*Did you adapt (change) the standard when you used it?*" said 'Yes'. Looking more closely at the results it can be concluded that adaption takes place:

- Widely with Dublin Core, library and archive standards;
- Not widely with museum standards, e.g. LIDO. This may underline the ability of LIDO to act as an interchange standard.

6.3.3 Evaluation

Looking at the different cultural domains it can be observed:

Museums

- A number of descriptive metadata standards are used: CDWA, LIDO, museumdat, Object ID, SPECTRUM, and VRA.
- LIDO was built on CDWA, museumdat and SPECTRUM. Therefore it can be asserted that all of them are compatible with and can be represented in LIDO.
- LIDO is an application of the CIDOC Conceptual Reference Model and so conforms with the ISO standard conceptual model for the domain.
- Many of those organisations who said that they used an 'Other standard' are from the museum domain. Therefore it is highly likely that their metadata can be expressed in LIDO.

Archives

- EAD and ISAD(G) are both in use, but are not suitable for data from the other domains.

Libraries

- Various specifications of the MARC world, as well as MODS (often in conjunction with METS), and TEI are in use within the library and within the archive domain.
- However best practice is difficult to identify because of the heterogeneity of MARC. Moreover the library community in general is moving on to the relatively new RDA (Resource Description and Access) model to cope with semantic web and linked data needs.

Looking at the other standards:

Technical (non-descriptive)

- METS can be seen as a technical, non-descriptive model which reflects the structure of a digital object. It is often used in conjunction with other models.

Conceptual:

- CIDOC CRM and FRBR are the conceptual models for the museum and the library domains respectively. They are not meant to serve as explicit metadata models themselves.

  Much work has already been carried out to harmonize both models, and the relevant concepts from the FRBR model were integrated into the CIDOC CRM model. A similar effort is currently underway with the archive domain so that the CIDOC CRM can in fact be seen as the leading conceptual reference model for cultural heritage.

6.4 SELECTION OF STANDARDS FOR USE IN LINKED HERITAGE

In the *Description of Work* for *Linked Heritage* gives the following criteria for the selection of a suitable metadata model for the project:

- The established user base;
- Adherence to standards and/or standards status in its own right;
- Demonstrated interoperability with other metadata models, including ESE;
- Demonstrated and/or potential ease of integration with the technologies selected for linked data, persistent identification, and public private partnerships);

- Maturity and quality of available technical implementation, documentation and support.

Based on the survey results, LIDO will be the primary metadata standard for aggregation within the Linked Heritage project.

LIDO (Lightweight Information Describing Objects) is the result of a collaborative effort of international stakeholders in the museum sector to create a common solution for contributing cultural heritage content to portals and other repositories of aggregated resources, as well as exposing, sharing and connecting data on the web. Being an application of the CIDOC Conceptual Reference Model (CIDOC CRM / ISO 21127) it provides an explicit format to deliver a museum's object information in its full descriptive richness.

LIDO Version 1.0 was developed with the support of the *ATHENA* project and delivered to the community during the ICOM/CIDOC conference in November 2010 in Shanghai/China.

The information and activities that are related to LIDO are centralized within the *Data Harvesting and Interchange Working Group* of CIDOC[38], the International Committee for Documentation of ICOM, the International Council of Museums.

Looking at further criteria for the selection LIDO:

- Being built upon previous work, and the large experience of international stakeholders in the museum documentation area, LIDO gained a widespread adoption in a very short amount of time. It has established a large user base and support within the CIDOC community (recently seen in CIDOC's annual conference where many contributions referred to the LIDO standard). Having its home within CIDOC ensures in particular the close connection with the group sustaining the CIDOC CRM standard, the *CRM-Special Interest Group* (CRM-SIG)[39].
- LIDO was or is used for aggregation model in the EC-funded projects: *ATHENA*, *MIMO*[40] and *Judaica Europeana*[41]. Through these projects LIDO's interoperability has been proved with metadata used by the different content providers, as well as interoperability with both Europeana's ESE and EDM standards.
- The technical implementation of LIDO in the metadata interoperability (MINT)[42] services that will be used in the *Linked Heritage* project, were developed during the ATHENA project. The solution has proved successful already for the ingestion of large amounts of data into Europeana.
- The schema design process for LIDO v1.0 took into account from the beginning the requirements for implementing the linked data concept, and in particular persistent identification so it is a suitable choice for integration with linked data technologies.

Therefore LIDO fulfills all requirements expressed for the metadata model and is easy to apply within the Linked Heritage project.

However for metadata from the library and archive domains the respective domain standards should be accepted as a delivery format, and suitable solutions for the ingestion process should be examined. It should be confirmed that EAD data, based on the *APEnet EAD* specification, can be ingested through *APEnet* [43]to Europeana.

[38] See: **http://cidoc.icom.museum**
[39] See: **http://www.cidoc-crm.org**
[40] See: **http://www.mimo-project.eu**
[41] See: **http://www.judaica-europeana.eu**
[42] See: **http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki**
[43] See: **http://www.apenet.eu**

For the library domain there seems to be no established ingestion workflow beyond Dublin Core / ESE data. Therefore since an important goal of the Linked Heritage project is the enrichment of Europeana, e.g. through the provision of as rich metadata as available, it will be examined what the library community is planning for future ingestion into Europeana, and a mapping template will be provided for transforming data from MARC variants used by providers in the Linked Heritage project, into LIDO.

7 BEST PRACTICE RECCOMENDATIONS

The publication of linked data is still at the experimental stage. Best practice can only be said to be emerging. Therefore the recommendations given in this section are based on:

- Common practice in the general linked data community, as represented by *The Cloud*;
- The practice of cultural heritage organisations that have published linked data;
- The general practice of the cultural heritage sector.

Some of the recommendations offer a range of options, with no 'right' choice. The choice an organisation makes is dependent on individual circumstances, and may be affected by legal and ethical considerations.

The recommendations can be separated into three 'choice areas':

7.1 WHAT INFORMATION TO PUBLISH AS LINKED DATA

Looking at what kind of information is being published as linked data in *The Cloud*, and especially the relatively small part which is about cultural heritage, two main types of information should be considered:

Collections information

This will be the bulk of the information that will be published by cultural heritage organisations. However they should also consider publishing information about:

- ***Surrogates*** – the results of digitisation;
- ***Supporting material*** – including exhibition catalogues, books, history files, and learning units;
- ***User generated content*** – reactions to the collections (permissions having been gained to publish).

Terminological information

Looking at *The Cloud* a large component is from terminological resources being used by cultural heritage organisations. These can be the result of international, national, thematic, organisational initiatives. The effort to do this is strong in the library and archive domains. It includes the publication of name authorities.

Also this work gives the opportunity for cooperative, possibly international and multilingual, publication, perhaps in the context of EC-funded projects. Topics for terminological publication include: object types; event methods (e.g. creation method); places; organisations; events; materials; iconography; and many others.

The primary advice in choosing what kind of data to publish as linked data is:

- **Consider publishing information about all aspects of collections and their related materials;**
- **Consider publishing terminological information, and seek partners to cooperate with in order to avoid duplication.**

7.2 WHAT LICENCE SHOULD THERE BE FOR THE LINKED DATA

This section deals with the licensing arrangements that are associated with the publication of linked data. Choices made in this are affected by general considerations of how much control the publisher of linked data wants to have over its data, but are also affected by what kind of data is being published.

As was seen by the analysis of *The Cloud* a large part of published linked data does not seem have a licence for its use. The result is that it is unclear what can be done with this data. In these litigious times users are particularly careful not to do anything that will leave them exposed to a possible loss of organisational reputation or even a lawsuit.

The primary advice about licensing is:

- **Any publication of linked data must be accompanied by a licence which makes it clear what uses can be made of the data.**
- **The licence may be standard, e.g. provided by *Creative Commons*, or one created specifically by the publisher.**

In general terms the two classes for the licence are:

- ***Open licence*** – This allows any use of the data, especially including commercial use, sometimes with restrictions about attribution and misuse.
- ***Not-open licence*** – This restricts uses to non-commercial only, with similar requirements for attribution and misuse.

With both classes there are a range of standard licences, e.g. those provided by *Creative Commons* and *GNU*, and the option of a specific organisational licence.

For an organisation to decide which particular licence it should give with its publication of linked data it is suggested that they follow these steps:

Step 1 – Decide what uses of the metadata you want to allow

An organisation may use the chart below to come to a decision about the licence it should use:

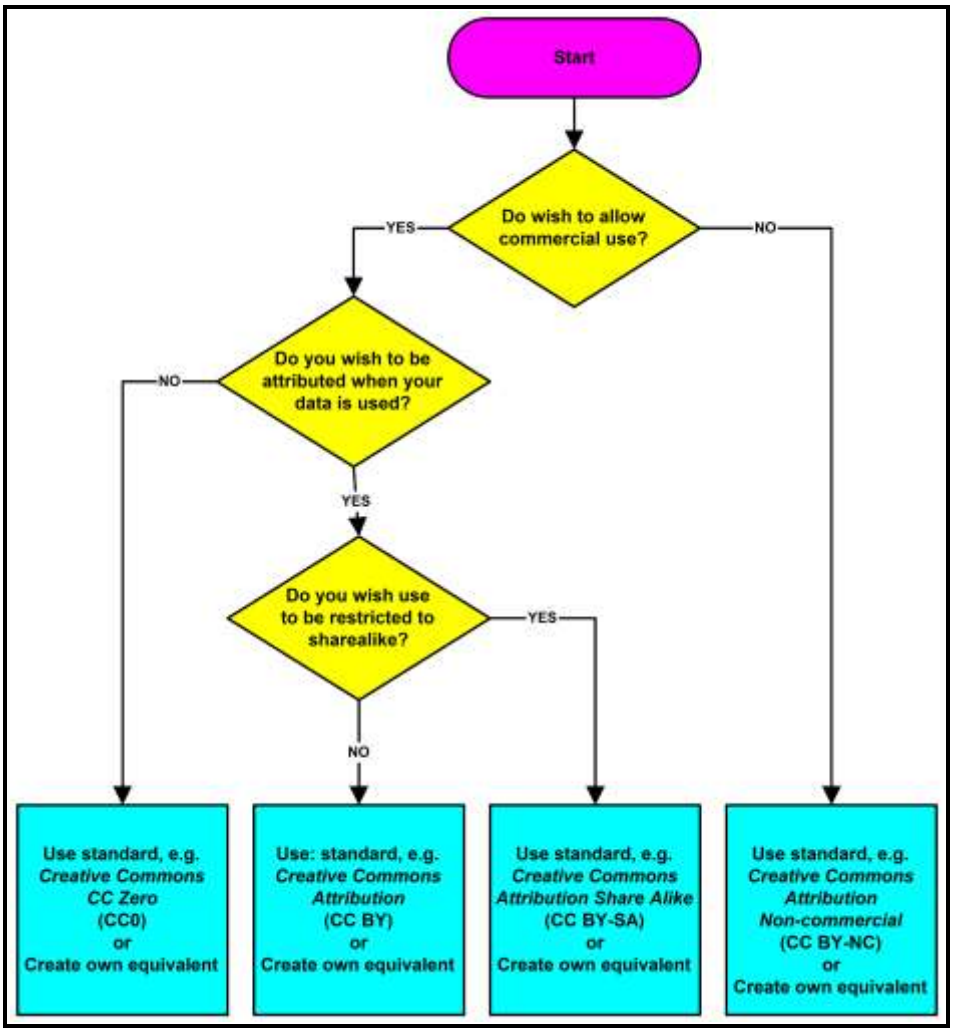

Figure 9: Linked data licence decision tree

Note that it is helpful for a user of the linked data to know if non-standard licence is compatible with a standard one.

Step 2 – Examine the rights environment of the data to be published

Step 1 assumes that an organisation is able to make a choice of licence without restriction. To test this assumption the organisation should seek to find answers to these questions:

- ***Is the data the organisation's intellectual property?***

  Usually the data that an organisation uses for describing its collections is its intellectual property. However some partners in the *Linked Heritage* project have said that this is not the case. This is the situation for aggregators where the metadata comes from their providers. The aggregator should already have a licence for its use of its providers' data. This can restrict the licence the aggregator can give for its linked data publication.

  It is also sometimes the same situation where volunteers (i.e. not members of staff) have been involved in creating the data. Best practice here would be for volunteers to have assigned their rights, or given an open licence to the organisation.

  Another situation that arises is where data from two sources is mixed. An organization may begin with externally-supplied data and enrich it. Rights over the enriched data are complex

If either situation is present then the advice is:

The organisation must either accept the restrictions imposed by the original creation of the data or seek to renegotiate the licence it has.

- ***Are there any legal or other restrictions on the type of licence that can be offered?***

  It may be that an organisation is operating in a rights environment which forbids the use a type of licence. This seems to be particularly the situation where a standard licence, e.g. *Creative Commons* is being considered.

  If this situation is present then the advice is:

  Consider using a non-standard licence that meets local needs.

  Also commercial may be specifically excluded for some types of data by law.

  If this situation is present then the advice is:

  The organisation cannot use a licence which allows commercial use.

- ***Is the organisation able to make a decision about licensing even when it has the rights in the data?***

  The survey of partners also brought to light the situation where a cultural heritage organisation does not have the authority to decide on licensing independently of its superior body. This is particularly the case where the cultural heritage body is owned by a national or regional government.

  The superior body may mandate a more **or** less restrictive licence than wished for by the cultural heritage organisation. It is possible that the cultural heritage organisation's data is viewed as an exception to general rules, and it might be possible to negotiate an exception.

  If this situation is present then the advice is:

  The organisation must use the licence that its superior body supports.

7.3 HOW TO PUBLISH THE LINKED DATA

In this area a potential publisher of linked data has three choices to make:

Which format standards to use

It is inconceivable that they will not use the basic standards like: RDF, RDFS, and OWL. However for the 'descriptive' formats it is advised to:

- **Not to create a proprietary format which is only intended to be used for your package;**
- **Use standard format(s) appropriate for the type of data being published. Looking at what is being used a few formats seem to be good suggestions:**
  - **Web resources: *Dublin Core*;**
  - **Persons: *Friend of a Friend*;**
  - **Terminological resources: *Simple Knowledge Organization System*;**
  - **Bibliographic resources: *Bibliographic Ontology*;**
  - **Music: Music Ontology.**

These recommendations are based on the current, in-use, formats. However there is a 'gap in the market' for a format for cultural heritage linked data.

Consider[44] using a cultural heritage specific format for linked data. Possible candidate formats, ones based on: *EDM*, *CIDOC CRM*, and *LIDO*.

RDF serialisations to publish

On the basis of the common practice it is advised that to:

Publish the linked data in the RDF/XML and N-Triples serialisations.

How to link the package into The Cloud

One issue that was brought out by discussions of the WP 2 Working Group was: Which are the 'trusted' packages in *The Cloud?* A measure of trust is if one knows the publisher of a package. This type of linking seems to be very common in all parts of *The Cloud* and leads to the formation of mini-clouds of interlinked packages. There seems to be a cultural heritage mini-cloud forming. A possible reason for this formation is the Europeana initiative.

Other very important issues are:

- The identification of resources. Are the identifiers you use compatible with the identifiers used in a potential package to link to;
- How compatible are the semantics of the packages. For example, if one wishes to identify 'personas' (public identities), is that the same as FOAF, which says it identifies people.
- A package has to be accessible to queries of it.

[44] The *Linked Heritage* project gives the community an opportunity to look at these possibilities. In particular if offers the possibility of using *LIDO*. See next section.

Therefore we advise:

- **Link to packages, of a general nature, which are often linked to: *DBpedia; GeoNames Semantic Web;* national sources of terminology (e.g. *UK Postcodes*)**;
- **Link to known packages in the cultural heritage, e.g.: *Library of Congress Subject Headings*; *VIAF: The Virtual International Authority File*; and *Dewey Decimal Classification*);**
- **Provide a SPARQL endpoint to the package.**

Obviously the final task is to make an entry for the package into *The Data Hub* registry!

7.4 CONTRIBUTING TO EUROPEANA

In the context of the *Linked Heritage* project (or any *Europeana Group* project) the requirements of Europeana are important as they will be publishing the metadata that it has aggregated as linked data.

From December 2011 contribution of metadata will be governed by the *Europeana Data Exchange Agreement*[45]. Metadata aggregated before this date will have to conform to this by the end of June 2012 at the latest.

It is worth stating those requirements as they impact on providers:

Licensing requirements

Europeana wishes to publish providers' metadata as linked data using the *CC0* licence. As mentioned above this means that any use of the metadata, including commercial use, is possible. Also the use does not require attribution.

If a provider intends to publish their metadata, e.g. as linked data, with the *CC0* licence then there are no difficulties and they can easily sign the agreement without difficulty.

However this situation seems, from the *Linked Heritage* partners' survey, to be not common. The result is that providers to Europeana have two options:

1. Remove all their metadata from Europeana.
2. Only give Europeana the metadata that agree to publish under the *CC0* licence.

Option 1 is a difficult option to take, both reputational and sometimes contractually. Partners may have contractual obligations with the Commission which means they must give metadata to Europeana.

Option 2 is more attractive, but providers should be aware that the *Agreement* requires that metadata supplied to Europeana must conform to their published metadata specifications.

Metadata specifications

The current metadata specifications refer to ESE. The relevant *ESE* elements are:

Namespace:Element	***Definition and notes***	***Data requirements***
`europeana:isShownBy`	An unambiguous URL reference to the digital object on the content provider's web site in the best available resolution/quality. (i.e. a link to the content as a text, image, sound, or video file **not** to the webpage with it on) Data here will allow the full functionality of Europeana and the automatic generation of a thumbnail by them. If this cannot be given then you **must** provide data for **isShownAt**.	Must be a valid URI (e.g. URL).

[45] Downloaded from: **http://www.version1.europeana.eu/c/document_library/get_file?uuid=deb216a5-24a9-4259-9d7c-b76262e4ce55&groupId=10602**

Namespace:Element	***Definition and notes***	***Data requirements***
europeana:isShownAt	An unambiguous URL reference to the digital object on the content provider's website in its full information context. If this cannot be given then you **must** provide data for **isShownBy**.	Must be a valid URI (e.g. URL).
europeana:object	For image thumbnails, if you can give a URL to a thumbnail on your website then give that URL here. However if these thumbnails are smaller than 110 pixels high then they will be scaled up to that size by Europeana. If you do not have a thumbnail then you may give the same data as in **isShownby** element.	Must be a valid URL.
europeana:type	The Europeana material type of the resource.	Must be: TEXT or IMAGE or SOUND or VIDEO
europeana:provider	Name of the organisation that is delivering content directly to Europeana.	Europeana maintains a standard list of organisations.
europeana:dataProvider	Name of the organisation that is delivering content to the aggregator who is the providing directly metadata to Europeana.	Europeana maintains a standard list of organisations.
dc:title or **dc:description**	A provider must supply data in one (or both) of these elements	
dc:language	If the content being described is of **europeana:type** TEXT then the provider must supply data in this element.	Use *ISO 639-2* (the three character code).
dc:subject or **dc:type or** **dc:coverage or** **dcterms:spatial**	A provider must supply data in one of these elements	It suggested that providers use a set of standard terms.

Each provider to Europeana will have to decide how to conform to these specifications. Note that the specifications refer to the presence of an XML element not to the data in those elements.

8 CONCLUSIONS

8.1 WORK CARRIED OUT

In this deliverable we have:

- Given an overview of linked data:
  - Publishing structured data on the Web;
  - Linked data;
  - RDF and SPARQL.
- Surveyed *Linked Heritage* partners for their:
  - Awareness of linked data – general knowledge of, use with their own collections, and publication both by them and by others;
  - Views on the Europeana Data Agreement;
  - Use of metadata for describing their collections.
- Analysed the *Linking Open Data Cloud* ('*The Cloud*'), focussing on the:
  - Licensing arrangements for the use of the published data;
  - Size of *The Cloud* and the packages in it;
  - Formats used to publish data;
  - Subjects of the data;
  - Formats used to encode data;
  - How *The Cloud* is linked;
  - Cultural heritage data in *The Cloud*.
- Gave best practice advice on:
  - What kind of linked data to publish;
  - Which licence to publish linked data under;
  - How, in terms of technical standards, to publish cultural heritage linked data;
  - Contributing to Europeana.

This represents an introduction to the topic, as well as acting as guidance for the rest of the project.

8.2 THE NEEDS OF *LINKED HERITAGE* PARTNERS

The partners' survey showed that the majority have some knowledge of linked data, but most have no experience of it. Therefore the project should:

- Provide tools for partners which will give them the basic information and some limited skills in this area. This deliverable, and later ones, can act as tools. However there is a need for more user-friendly ones. Therefore we suggest that pocket guides should be created in the same format as those for the ATHENA project. Also there is the opportunity to create training material in this area within the framework of work package 7.
- Further their experience by involving them in the publication of linked data (by partners CT and PUM). The aim here would be to encourage other partners to take part, at least in a limited way.

This limited way would include making the decisions needed to publish linked data (e.g. licensing) even if they choose not actually publish linked data.

8.3 PUBLISHING LINKED DATA

The *Linked Heritage* project is committed to publishing linked data from CT and PUM metadata. This should be carried out following the advice given in this deliverable in the areas of:

- Data to publish;
- Licence to give;
- Technical standards use;

Most of this advice is framed in terms of making decisions. It will be useful if the all partners take part in the decision-making process, however in the final decisions will have to be made by those who actually are publishing linked data.

8.4 FURTHER WORK IN THE *LINKED HERITAGE* PROJECT

In the next 12 months work package 2 must create the following deliverables:

- D2.2 - *State of the art report on persistent identifier standards and management tools* (Month 09).

  This will be a similar document to this deliverable and will have similar aims.

- D2.3 - *Specification of the technologies for large-scale implementation of cultural heritage linked data* (Month 18).

  This deliverable will be based on a series of use cases for the publication and consumption for the linked data. The creation of these case studies will be the responsibility of CT, but will involve the thematic working group, who will input their experience.

APPENDIX 1: THE LINKED HERITAGE SURVEY QUESTIONS

Provider Information

Please give the following information about the organisation providing the content:

1. Country

2. Organisation name

3. Address (include postcode)

4. Is the provider a (tick ☑ all that apply)?

☐ Museum
☐ Library
☐ Archive
☐ Sound archive
☐ Publisher
☐ Aggregator
☐ Other

5. If you ticked 'Other' please give organisation type

Contact (who filled in the form)

6. Name

7. Job title

8. Telephone number

9. E-mail address

Metadata

10. Please check the all boxes for the metadata formats that are used to describe the objects in this collection.

- ☐ CDWA
- ☐ CIDOC-CRM
- ☐ Dublin Core
- ☐ EAD
- ☐ FRBR
- ☐ ISAD(G)
- ☐ MAB
- ☐ MARC
- ☐ METS
- ☐ MIDAS
- ☐ MODS
- ☐ museumdat
- ☐ Object ID
- ☐ ONIX
- ☐ SPECTRUM
- ☐ TEI
- ☐ VRA
- ☐ Other

11. If you answered 'Other' please give details.

12. Did you adapt the standard?

☐ Yes
☐ No

13. If 'Yes' then please say how you adapted the standard.

14. Please give the language(s) that your metadata is in:

Linked Data

15. Are you or your organisation familiar with the concept of linked data?

☐ Yes
☐ No

16. Have you or your organisation had experience of using linked data in connection with your collections?

☐ Yes
☐ No

17. If you answered 'Yes' please give details of which source(s) of linked data you use (e.g. "GeoNames") and why you use it (e.g. "it is a trusted source").

18. Have you or your organisation had experience of publishing linked data in connection with your collections?

☐ Yes
☐ No

19. If you answered 'Yes' please give details. Include: the URL which gives access to the linked data you have published; What type of data you publish (e.g. "full records", "only basic information"); and what kind of licence for reuse do you give (e.g. "any use"; "non-commercial", "Creative Commons [type]"

20. Do you or your organisation know of any linked data projects or initiatives in your country in the field of cultural heritage?

☐ Yes
☐ No

21. If you answered 'Yes' please give details. Include the URL which gives access to project or initiative you know about.

22. Europeana's new licence requires that providers will have to agree to have the metadata that they provide to Europeana published as Linked Open Data. This means that any 3rd party use, including commercial, is permitted. Does your organisation agree to this?

☐ Yes
☐ No
☐ Not sure

APPENDIX 2: SHORT BIBLIOGRAPHY OF INTRODUCTORY MATERIAL ON LINKED DATA

This list is not a complete list of what is available, but contains sources, all of which are accessible on the Web, that the authors found useful:

Subject area	***Title*** **(notes)**	**URL**
Linked data	*D2R Server: Publishing Relational Databases on the Semantic Web*	**http://www4.wiwiss.fu-berlin.de/bizer/d2r-server** [Website]
		http://www4.wiwiss.fu-berlin.de/bizer/pub/Bizer-Cyganiak-D2R-Server-ISWC2006.pdf [Concepts & 'cookbook']
		http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/#publicservers [Public servers & projects]
Linked data	*Linked data: A practical introduction*	**http://www.slideshare.net/mediasemanticweb/linked-data-michael-hausenblas-2009-03-05**
Linked data	*The next Web of open linked data* (April 2009)	**http://www.youtube.com/watch?v=OM6XIICm_qo**
Linked data	*The year open data went worldwide* (March 2010)	**http://www.youtube.com/watch?v=3YcZ3Zgk0a8&feature=relmfu**
Linked data	*Consuming linked data*	**http://www.slideshare.net/juansequeda/consuming-linked-data**
OWL	*OWL 2 Web Ontology Language: Primer* (Technical)	**http://www.w3.org/TR/2009/REC-owl2-primer-20091027**
RDF	*RDF Primer* (Technical; available in French and Hungarian)	**http://www.w3.org/TR/2004/REC-rdf-primer-20040210**
RDF format	*Europeana Data Model Primer*	**http://group.europeana.eu/c/document_library/get_file?uuid=718a3828-6468-4e94-a9e7-7945c55eec65&groupId=10605**
SPAQL	*SPARQL Tutorial*	**http://openjena.org/ARQ/Tutorial/index.html**