JLIS.it

Linked heritage experience in linking heritage information

Gordon McKenna

Introduction

This paper will look at the experience of the EC-funded Linked Heritage project in the area of linked data. It will cover:

- the project in context;
- work package 2 - Linking Cultural Heritage Information;
- the results of research into the use of linked data in the cultural heritage sector;
- a look forward to the further work of the project.

Overview of the Linked Heritage Project

The Linked Heritage project is part-funded by the ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme. The project began in April 2011 and lasts for 30 months. The project has three objectives:

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-6304

- to contribute large quantities of new content to Europeana, from both the public and private sectors;
- to demonstrate enhancement of quality of content, in terms of metadata richness, re-use potential and uniqueness;
- to demonstrate enable improved search, retrieval and use of Europeana content.

Linked Heritage aim to facilitate and deliver large-scale, long-term enhancement of Europeana and its services. It addresses the problems associated with:

- non-standard descriptive terminologies;
- the lack of private sector and 20th Century content:
- the preservation of complex metadata models within the Europeana metadata schema.

Project partners include all the key stakeholder groups from 20 EU member states, with Israel and Russia. They include ministries and responsible government agencies, content providers, aggregators, leading research centres, publishers and SMEs.[1]
The objectives of the project are:

[1]Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche (IT); Università Degli Studi di Padova (IT), Consiglio Nazionale delle Ricerche (IT), Ministère de la Culture et de la Communication (FR), Eesti Vabariigi Kultuuriministeerium (ER), Hellenic Ministry of Culture (GR), National Technical University of Athens (GR), University of Patras (GR), Collections Trust LBG (UK), An Chomhairle Leabharlanna Ireland (IE), Pintail Ltd (IE), Fundacio Privada I2CAT, Internet i Innovacio Digital A Catalunya (SP), Philipps Universitaet Marburg (GW), Stiftung Preussischer Kulturbesitz (GW), Central Library of the Bulgarian Academy of Sciences (BU), Javni Zavod Republike Slovenije za Varstvo Kulturne Dediscine (Slovenia), The Cyprus Research and Educational Foundation (CY), Stowarzyszenie Miedzynarodowe Centrum Zarzadzania Informacja (PL), Riksarkivet (SW), MEDRA S.R.L. (IT), Gottfried Wilhelm Leibniz Universitaet Hannover (GW),

- to contribute large quantities of new content to Europeana, from both the public and private sectors;
- to prepare for the enhancement of the quality of both new and existing Europeana content, in terms of its metadata richness, its re-use potential and its uniqueness;
- to demonstrate improved search, retrieval and use of Europeana content, both within the Europeana portal and by third parties via the Europeana API.

It is doing this by:

- assembling representative stakeholder groups (content providers, aggregators, ministries and policy making bodies, technologists, private sector companies, and associations);
- consultation, consensus building, networking, and the sharing of perspectives and priorities;
- the identification and promotion of best practice. This is the most appropriate and useful standards, specifications and recommendations for the contribution, ingestion and enhancement of Europeana content;
- large scale implementation (including the necessary technology integration in compliance with the Europeana standards)

Editeur Limited (UK), MVB (Marketing und Verlagsservice des Buchhandels) (GW), Orszagos Szechenyi Konyvtar (HU), Koninklijke Musea voor Kunst en Geschiedenis (BE), Institutu Umeni - Divadelniho Ustavu (Czech Republic), Instituto Superior Tecnico (PO), Valsts Agentura Kulturas Informacijas Sistemas (Latvia), PACKED (Platform voor de Archivering en Conservering van Audiovisuele Kunsten) (BE), CORDIA (Slovakia), Universita Degli Studi di Roma La Sapienza (IT), C.T.F.R. SRL (IT), Departament de Cultura i Mitjans de Comunicació (SP),Promoter di Masi Pietro & C S.N.C. (IT), Université de Savoie (FR), Association Dedale (FR), UMA Information Technology (AU), Digital Heritage LBG (UK).

and validations of the identified best practice standards and specifications. These will serve to provide to Europeana 3 millions new objects;

- the preparation of a demonstrator how the improved specifications are to be applied and how to implement the enrichment of Europeana content;

- training and dissemination to build capacity and awareness in the cultural heritage sector, particularly in the use of Linked Heritage technical outputs, but also in Europeana technologies.

The project is split into seven work packages:

WP 1 – Project management and coordination Deals with the basic project management of linked heritage, e.g. monitoring progress and managing the relationship with the Commission. Also manages the setting up and maintenance of working groups, both national and thematic.

WP 2 – Linking Cultural Heritage Information Looks at the potential use of linked data in the cultural sector (see next section).

WP 3 – Terminology Works on the enabling of the use of terminologies with the project and in a wider cultural heritage context.

WP 4 – Public Private Partnership Explores the standards in use in the non-heritage commercial cultural sectors, and the possibilities with integrating this with the cultural heritage sectors, especially with Europeana.

WP 5 – Technical Integration Enables the technical tools and requirements of the project.

WP 6 – Coordination of Conten Manages the process of giving access to Linked Heritage's partners to Europeana.

WP 7 – Dissemination & Training Making the wider community aware of the project's work, and producing learning tools to enable that community to use the results.

Work Package 2 – Linking cultural heritage information

This paper is part of the results of this work package. Its objectives are:

- to explore the state of the art in linked data and its applications and potential;
- to identify the most appropriate models, processes and technologies for the deployment of cultural heritage information repositories as linked data;
- to consider 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;
- to explore the state of the art in persistent identifiers (both standards and management tools);
- to identify the most appropriate approach to persistent identification, e.g. a unique standard or a set of different standards;
- to design a feasibility model and to realised a demonstrator of a flexible, scalable, secure and reliable infrastructure for a network of 'linked data enabled' cultural heritage information repositories;

- to explore 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: Europeana Semantic Element (ESE); and Europeana Data Model (EDM);
- to outline the potential benefits that richer cultural heritage metadata could bring to Europeana, and to the other services which will use it.

Linked data in the cultural heritage sector

Partner Survey

As part of the tasks the work package carried a survey of Linked Heritage partners, and providers. This covered, amongst other things, their knowledge of linked data and their experience in using linked data.

Respondent information

Table 1 on the facing page shows 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;

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

Table 1: Here are the figures for the types of organisations that responded to the survey.

- 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.

Countries

Country	Number of respondents	Country	Number of respondents
Austria	1	Ireland	2
Belgium	4	Israel	1
Bulgaria	1	Italy	6
Cyprus	1	Poland	1
Czech Republic	1	Russian Federation	1
Estonia	1	Slovakia	1
France	3	Spain	2
Germany	4	Sweden	1
Greece	3	United Kingdom	2
Hungary	1		

Table 2: Here are the figures for the countries where respondents are based.

Obviously, figures in table 2 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.

Linked data

Awareness

Response	Number of respondents	%
Yes	30	75.0
No	10	25.0

Table 3: To ”Are you or your organisation familiar with the concept of linked data?”

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

Use

Response	Number of respondents	%
Yes	7	17.5
No	33	82.5

Table 4: To ”Have you or your organisation had experience of using linked data in connection with your collections?”

Those who answered ”Yes” were asked to give details of which source(s) of linked data they use and why they use it’. The sources used were: 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).

Publication

Those who answered "Yes" were asked to give details. Three respondents 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, 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.

Response	Number of respondents	%
Yes	4	10.0
No	36	90.0

Table 5: To "Have you or your organisation had experience of publishing linked data in connection with your collections?"

Linked data projects and initiatives

Response	Number of respondents	%
Yes	15	37.5
No	25	62.5

Table 6: To "Do you or your organisation know of any linked data projects or initiatives in your country in the field of cultural heritage?"

Those who answered "Yes" were asked to give details. The responses, ordered by country, are listed in table 7.

Country	Project or initiative [2]
France	RAMEAU
	ISIDORE
	Pactols
	BABEL
	COLLECTIONS
	PALISSY
	EROS
	PATRIMOINE LOT
	WIKIMEDIA COMMONS FRANCE:
Germany	Linked data service of the German National Library
	"Several initiatives throughout the country"
Israel	Vocabularies of the Israel Museum Jerusalem (SKOS)
Italy	Linked Open Data Italia
	SPAR ontologies
	Datagov.it
	LinkedOpenCamera
	Spaghetti Open Data
Russia	Open Kunstkammer
Sweden	LIBRIS
Spain	Open Data Gencat
	Euskadi
	Patmapa
	Cantabria's Cultural Heritage Ontology
United Kingdom	Various government data sets

Table 7: Linked data projects and initiatives - Responses details

Europeana Open Data Agreement

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

Table 8: To "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?"

[2]Details of responses listed in table 7. 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, 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", Vocabularies of the Israel Museum Jerusalem that have been migrated to SKOS:http://www.imj.org.il/imagine/thesaurus/allobject.htm and http://www.imj.org.il/imagine/thesaurus/objects/objectTOC.htm, ItalyLinked 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, RussiaOpen Kunstkammer: http://www.kunstkamera.ru, Sweden LIBRIS (joint catalogue of the Swedish academic and research libraries): http://www.kb.libris.se, SpainOpen 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.

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.

The Linking Open Data Cloud

The Linking Open Data Cloud[3] (The Cloud) is the best known representation of linked data. It shows ’packages’ of linked data and the

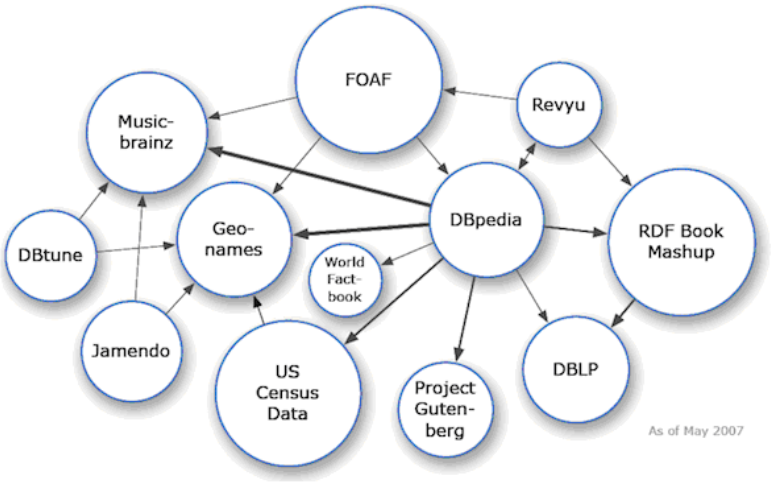

Figure 1: The Cloud in May 2007

links between packages. In May 2007 it looked like in figure 1 (with 12 packages).

By September 2011 the version that is colourised to represent the domain of the package looked like in figure 2 on the facing page (with 311 packages). 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.[4] This effort is part Linking Open Data community project[5] which is part of the W3C's Semantic Web Education and Outreach Interest

[3] http://linkeddata.org.

[4] http://thedatahub.org, http://thedatahub.org/group/lodcloud.

[5] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

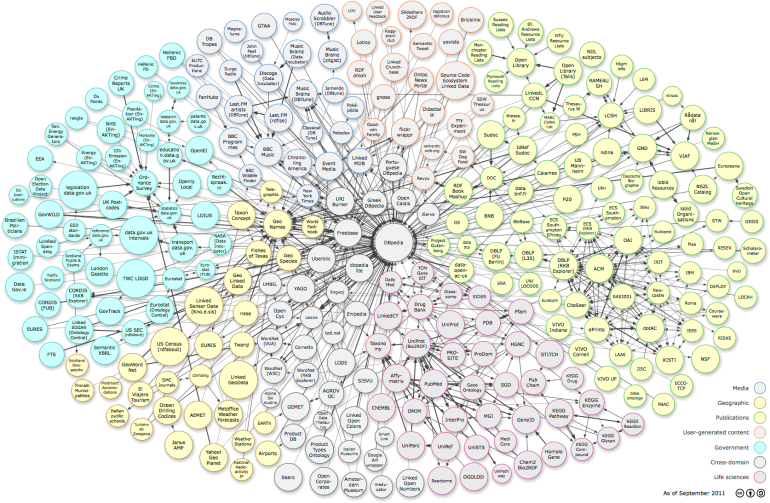

Figure 2: The Cloud in September 2011

Group (SweoIG).[6] 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, as shown in figure 3. For each package the results 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.

Figure 3: The Data Hub search results screen

For each package there is a full record, as shown in figure 4 on the facing page. For each package the full record screen includes additional information about:

[6] http://www.w3.org/wiki/SweoIG.

- 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.

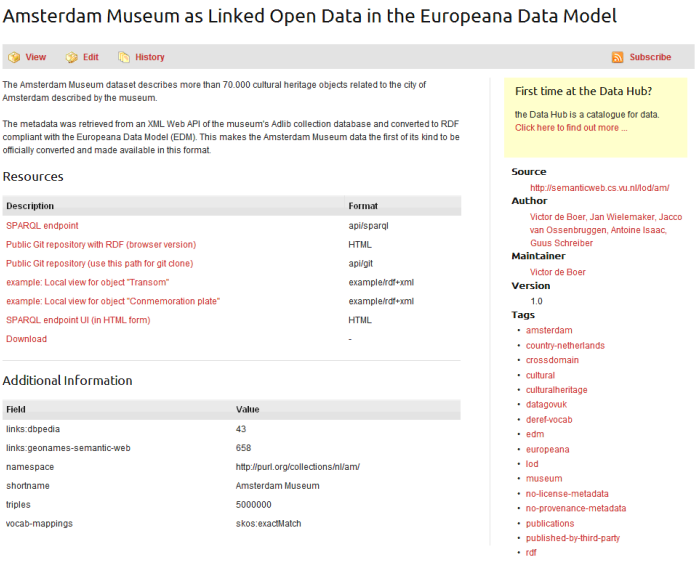

Figure 4: The Data Hub package record

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" means "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

Table 9

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.

Which IPR licences are used?

Open licenses

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

Table 10

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.

Not open licenses

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

Table 11

From the above[7] 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.

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

[7] 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.

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

Table 12

LinkedGeoData (CC BY licence) is a knowledge base of spatial obtained from the OpenStreetMap[8] 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 ency-

[8] http://www.openstreetmap.org.

clopaedic 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.

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

Table 13

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 Audio-Scrobbler 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 about university students.

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 theten most common subjects in The Cloud are shown in table 14 on the next page. 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.

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

Table 14

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 listed in table 15 on the facing page.
There seem 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

Format	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

Table 15: The abbreviation in brackets after a format's name is the 'namespace' for that format.

in a specific project or initiative.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.

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

Table 16

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[9] is a system for publishing linked data, developed during the EC-funded ReSIST[10] 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:

[9] http://www.rkbexplorer.com.
[10] http://www.resist-noe.org.

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
IdRef: Sudoc authority data	3	20,040,000
Sudoc bibliographic data	1	20,000,000
flickr™wrappr	4	16,358,998

Table 17

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.[11]

[11] http://version1.europeana.eu/web/lod.

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)[12]	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

Table 18

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

[12] This will eventually be published as CC0.

(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. Cultural heritage packages use formats listed in table 19.

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

Table 19: Formats used

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 targets listed in table 20.

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

Table 20: Targets of links in The Cloud

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 listed in table 21 on the next page. RDF/XML is used by all but two of the packages: Europeana Linked Open Data uses mentions only N-Triples, and the Calames Package do not mention

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

Table 21: Serialisations

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.

Best practice recommendations

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':

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.

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.

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;
- RDF serialisations to publish;
- How to link the package into The Cloud.

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[13] 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;

[13] 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.

- 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.

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!

Future Work on linked data

In the next stage of the project work package 2 will be working on two tasks which will show the potential of linked data:

Task 2.3 – Technical specifications

This will specify how cultural heritage information can be enriched by, and can enrich, the 'Cloud'. We will identify: models, processes and technologies which offer the best potential. Selection criteria will include:

- existing use of linked data in cultural heritage and the humanities;

- the use of standards;
- being able to interoperate with other linked data stores. These will include 'major actors' already identified, such as DBpedia and GeoNames;
- show integration with the technologies selected in other Linked Heritage thematic work packages (i.e. public private partnerships and terminologies);
- maturity and quality of a technical implementation, documentation and support.

Task 2.4 – Enabling linked cultural heritage data

This will demonstrate how to extend existing ingestion procedures to enable content providers to publish their content as linked data, in addition to publishing it in Europeana. The demonstrator will:

- enable content providers to contribute content to the linked data repository and maintain their existing linked data information;
- enhance the ingestion processes with tools for:
  - browsing the linked data repository and its connections to external sources;
  - creating and editing links between entities;
  - extending retrieval to include preferred sources for links and textual information.

GORDON MCKENNA, Collections Trust.
gordon@collectionstrust.org.uk

McKenna, G. "Linked heritage experience in linking heritage information". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #6304. DOI: 10.4403/jlis.it-6304. Web.

ABSTRACT: The Linked Heritage Project started in April 2011. It is funded by the European Commission under the IST Policy Support Programme (ICT PSP), and runs for 30 months. Its main objective is to contribute a large quantity of new content to Europeana, from both the public and private sectors (c3 million items). In addition, the project will show how: 1. To enhance the quality of Europeana content, in terms of its metadata richness, its re-use potential and its uniqueness; 2. To enable improved search, retrieval and use of Europeana content. The author is working in this project, specifically as lead partner in the work package Linking Cultural Heritage Information. This will, amongst other things, be exploring best practice report on cultural heritage linked data and metadata standards. This paper will give some results of the research that has been undertaken. Questions that will be answered include: "What is linked data? Is all of it Open? Which standards are being used? What use is being made of linked a data in cultural heritage at the moment?".

KEYWORDS: Library linked data; Linked Heritage Project; Europeana; Cultural Heritage Project

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15