Daguerreobase

Collective cataloging tool for daguerreotypes and daguerreotype literature

cip competitiveness and innovation framework programme 2007–2013

DELIVERABLE

Project Acronym:
Daguerreobase

Grant Agreement number:
ICT_PSP 297250

Project Title:
DAGUERREOBASE

D5.4 Linked Open Data

Version 2.0

Author(s): **Mark Lindeman (PIM)**
Herman Maes (NFM)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	PU
RE	Restricted to a group defined by the consortium and the Commission Services	
C	Confidential, only for members of the consortium and the Commission Services	

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision	Date	Author	Organisation	Description
First Draft	18/05/2013	Herman Maes	NFM	
Second Draft	24/07/2013	Herman Maes	NFM	
First Revision	12/09/2013	Mark Lindeman	PIM	
Version 1	18/09/2013	Herman Maes	NFM	
Draft version 2	20/02/2014	Mark Lindeman	PIM	
Version 2	31/3/2014	Herman Maes	NFM	Final version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise.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

"This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community". http//ec.europe.eu/ict_psp

Table of Content

1. INTRODUCTION

As Daguerreobase will act as an (project) aggregator of content for Europeana, the realization of the objectives of the Daguerreobase project are strongly dependent on good metadata communication and connectivity between the Daguerreobase aggregator and the Europeana portal.

To make the Daguerreobase into a valuable source of information that will contribute to and benefit from Europeana, several methods will be used.
The Linked Open Data protocol was an integral part used in the development of the new Daguerreobase. An API is developed to prepare the necessary mappings for the metadata conversions to the renewed Daguerreobase structure and to handle the alignment of the digital content with Europeana but also with other aggregators and providers.
A DAM (Digital Asset Management system) and data entry tool will be used to provide the metadata and derivatives to the Daguerreobase aggregator.
More detailed descriptions of these methods and tools are described in separate documents and deliverables as part of the WP5 tasks.

- Collection management set up (D5.2);
- API (D5.3);
- Linked Open Data (D5.4);
- Linking Europeana Collection (D5.5);
- Data Import (D5.6).

This document will describe the use of Linked (Open) Data.

2. DEFINITIONS[1]

API: Application Programming Interface.

Content: Digital (digitised or digital born) information about objects. In this project, content is restricted to digital images and metadata descriptions of daguerreotype objects.

Content Aggregator: An aggregator in the context of Europeana is an organization that collects metadata from a group of data providers and transmits it to Europeana. Aggregators also support the data providers with administration, operations and training.

Project aggregator: Project aggregators are organizations that have joined a project consortium with a specific aim and purpose. Project aggregators aim to aggregate within a specific theme or by domain (single or cross).

Daguerreobase: project/thematic content aggregator on daguerreotypes. EU-CIP, ICT-PSP funded project #297250.

Daguerreobase API: technical interface for the purpose of searching and retrieving Daguerreotype Metadata.

Europeana portal: Stichting Europeana, the Foundation providing the Europeana API.

Linked Data: In computing, **linked data** (often capitalized as Linked Data) describes a method of publishing structured data so that it can be interlinked and become more useful. It builds upon standard Web technologies such as http, rdf and URI's, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried. [2]

Linked *Open* Data: Linked *Open* Data (LOD) is Linked Data which is released under an open licence, which does not impede its reuse for free. Creative Commons CC-BY is an example open licence, as is the UK's Open Government Licence. Linked Data does not of course in general have to be open -- there is a lot of important use of linked data internally, and for personal and group-wide data. You can have 5-star Linked Data without it being open. However, if it claims to be Linked Open Data then it does have to be open, to get any star at all.[3]

[1] According to Europeana definitions. Europeana Aggregators'Handbook Edition 1 May 2010.
[2] Bizer, Christian; Heath, Tom; Berners-Lee, Tim (2009). "Linked Data, The Story so far". *International Journal on Semantic Web and Information Systems* **5** (3): 1–22. doi:10.4018/jswis.2009081901. ISSN 15526283. Retrieved 2010-12-18. Solving Semantic Interoperability Conflicts in Cross–Border E–Government Services.
[3] http://www.w3.org/DesignIssues/LinkedData.html

Metadata: metadata is information about Content, describing its characteristics to aid in its identification, discovery, interpretation and management.

Preview: general term including smaller and/or lower resolution version of still image Content or shorter and/or lower resolution extract of audio or moving image Content.

3. LINKED (OPEN) DATA

WP5 is responsible for the actual ingestion of data from the renewed Daguerreobase that will serve as an aggregator for containing a critical mass of historical relevant, multilingual and high quality content of mainly historical and European style daguerreotype objects and literature. The current consortium partners, new institutional partners and private collectors or owners will deliver the main part of the actual content.

Addressing Europeana, Daguerreobase developed multilingual terminology and thesauri/entries lists to improve semantic web-based access and retrieval of cultural and historical information within Europeana. The technical partners in the Daguerreobase project explored the use of Linked Data (LD) and Linked Open Data (LOD). Applying LD and LOD will support the semantic processing and interoperability within and with the Europeana data model (EDM).

3.1. Principles and use of Linked (Open) Data

Tim Berners-Lee outlined four principles of linked data in his *Design Issues: Linked Data* note[4] :

1. Use URI's to denote a 'thing',
2. Use http URI's so that these things can be referred to and looked up by people and user agents,
3. Provide useful information about the 'thing' when its URI is de-referenced, leveraging standards such as RDF or SPARQL.
4. Include links to other related things (using their URIs) when publishing data on the Web.

In 2010 he introduced a (5) star rating scheme in order to encourage people to deliver good linked (open) data:

Under the star scheme, you get one (big!) star if the information has been made public at all, even if it is a photo of a scan of a fax of a table -- if it has an open licence. Then you get more stars as you make it progressively more powerful, easier for people to use.

★	Available on the web (whatever format) *but with an open licence, to be Open Data*
★★	Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★★★	as (2) plus non-proprietary format (e.g. CSV instead of excel)
★★★★	All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★	All the above, plus: Link your data to other people's data to provide context

Use of Linked (Open) Data

[4] http://www.w3.org/DesignIssues/LinkedData.html

In November 2011, the ***Linked Heritage*** project[5] finalized Deliverable 2.1, '*LH_D2-1_Bestpracticereportonculturalheritagelinkeddataandmetadatastandards_final*' in which the use and potential of linked data was explored. The project identified the most appropriate models, processes and technologies for the deployment of cultural heritage information repositories as linked data. The Daguerreobase project is reusing some of this information about the practise and use of LD and LOD in making decisions for the own project.

Other purposes of deliverable D2.1 of the Linked Heritage project were:

- Considering how linked data practices could be applied to cultural heritage information repositories;
- Exploring the state of the art in persistent identifiers (both standards and management tools);
- Identifying the most appropriate approach to persistent identification;
- 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;
- 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);
- Outlining the potential benefits that richer cultural heritage metadata could bring to Europeana, and to the other services which will use it.

In Section 4 of deliverable D2.1 of the Linked Heritage project, the *Linking Open Data Cloud* was highlighted and commented. The project researched the next 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*?

The 'Cloud'

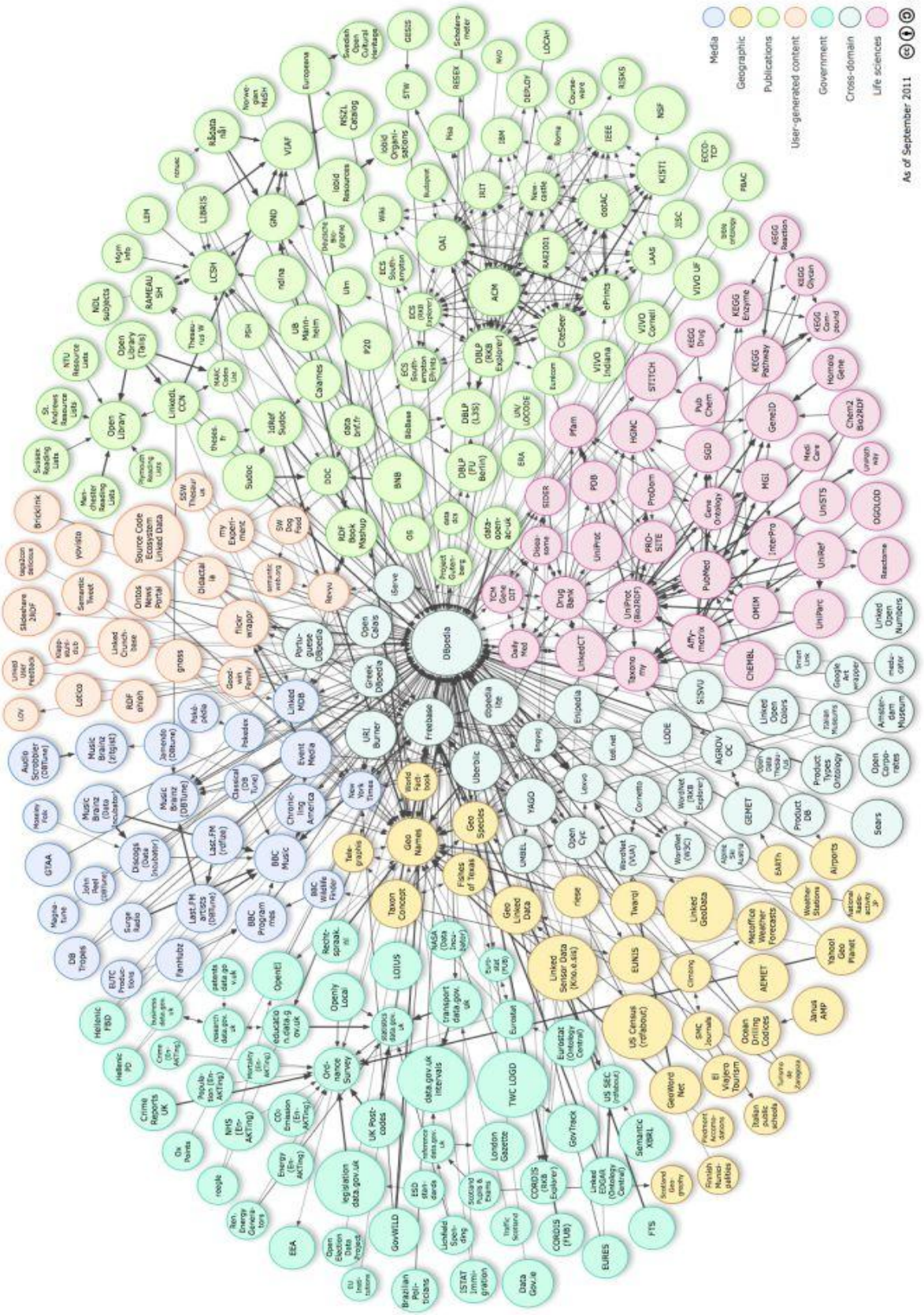
The next diagram shows datasets in the different areas or fields that have been published in the *Linked Data* format, by contributors to the *Linking Open Data* community project and other individuals and organisations. It is based on metadata collected and curated by contributors to the *Data Hub*[6]. The diagram is published on http://lod-cloud.net/ and is maintained by Richard Cyganiak (DERI[7], NUI Galway) and Anja Jentzsch (HPI[8]). The theme´s in the cloud are Media, Geographic, Publications, User-generated content, Government, Cross-domain, Life sciences. This diagram was last updated: 2011-09-19.

[5] LINKED HERITAGE, ICT-PSP Project no. 270905, Coordination of standard and technologies for the enrichment of Europeana. LH_D2-1_Bestpracticereportonculturalheritagelinkeddataandmetadatastandards_final. This deliverable was created based on a process for creating similar deliverables that was developed and successfully used, during the *ATHENA* project.

[6] http://datahub.io/group/lodcloud

[7] http://www.deri.ie/ , the Digital Enterprise Research Institute.

[8] http://www.hpi.uni-potsdam.de/willkommen.html , the Hasso Plattner Institute.



[9]

Next conditions and steps are required to get a metadata set into the diagram.

[9] "Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

Make sure that data are published according to the Linked Data principles. We interpret this as:

- There must be *resolvable http:// (or https://) URIs*.
- They must resolve, with or without content negotiation, to *RDF data* in one of the popular RDF formats (RDFa, RDF/XML, Turtle, N-Triples).
- The dataset must contain *at least 1000 triples*.
- The dataset must be connected via *RDF links* to a dataset that is already in the diagram. This means, either your dataset must use URIs from the other dataset, or vice versa. We arbitrarily require at least 50 links.
- Access of the *entire* dataset must be possible via *RDF crawling*, via an *RDF dump*, or via a *SPARQL endpoint*.

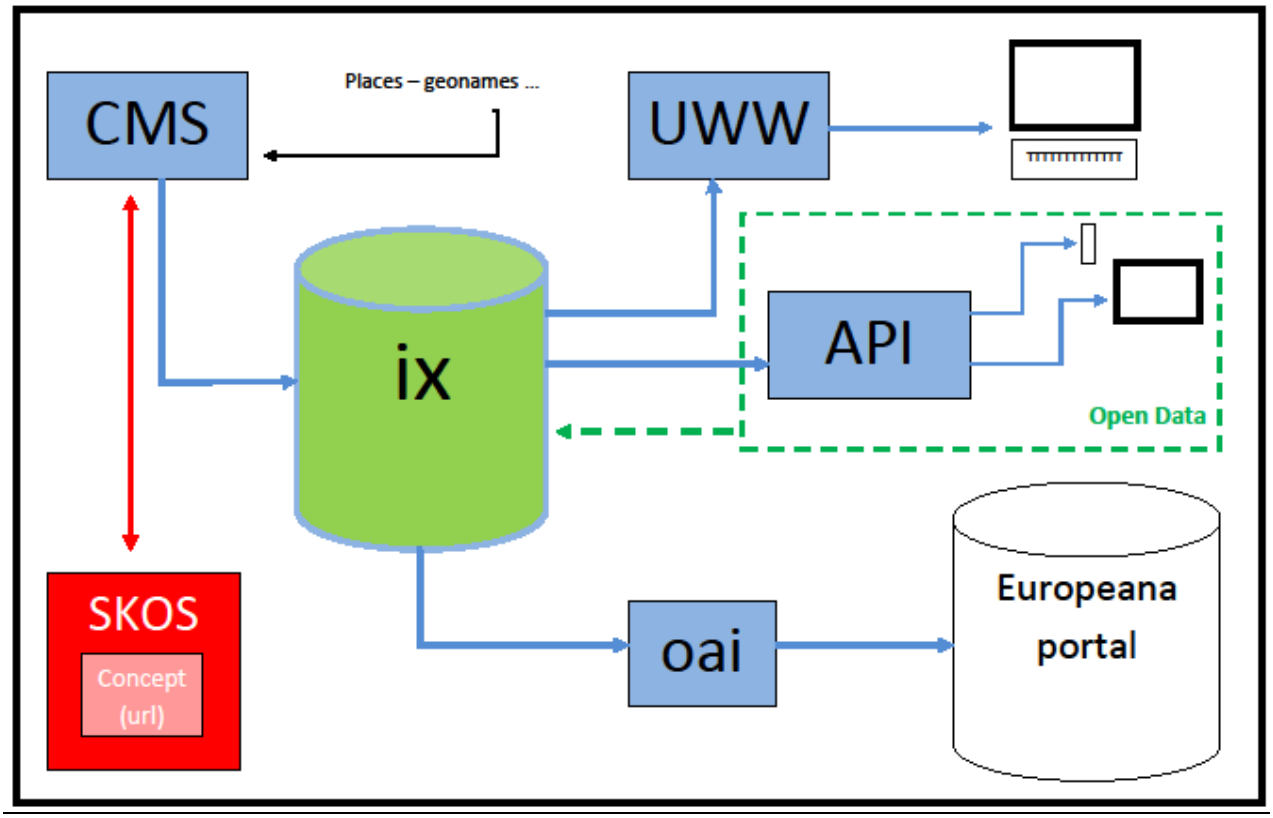
If your dataset meets these criteria:

1. Please add it to the Data Hub, the open registry of data and content packages. See the Guidelines for Collecting Metadata on Linked Datasets in the Data Hub for more details. (Before creating a new Data Hub record, please double-check whether a record already exists for your dataset.)
2. We provide a handy record validator; use it to check that at least the minimum required information is present.
3. Email richard@cyganiak.de and mail@anjajentzsch.de.
4. We will review the Data Hub record, and add it to the lodcloud group.
5. The dataset will be included in the next update of the diagram.

In Annex 1, Section 5 to 8 of deliverable 2.1 of the Linked Heritage project are given.

> *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.

3.2 Structure of the Daguerreobase aggregator repository



3.2.1. Metadata

A daguerreotype is much more than just a photographic image. An important outcome of the Daguerreobase project is the realisation of a community standard for the description of the complex structure of a daguerreotype object. The standard for the description of a daguerreotype object is described in a separate deliverable, D2.3.a-d[10], and includes the entire metadata model and applied general descriptive standards.

3.2.2. Presentation of the collection

The Daguerreobase image repository is integrated in *Joomla! 2.5*. The image repository has a 'simple search' and. The OpenSearch tools can be used to search on specific fields to provide all needs of the different user groups. The search engine will be based on *SolR*, an open source Apache project to which Picturae (PIM) has made several contributions. This will result in the possibility of faceted search capabilities.

A single search will lead to a result page, which can be viewed in four different layouts:

- Gallery – a view with large images and per image a title;

[10] D3.2.a-d Standards ...

- List - in the list view there are only a few images shown (max 10), all the metadata is shown at the side of the images;
- Presentation – a full screen view of the individual records of the search results. This will be shown in an overlaid design.

All views will be developed in '*HTML 5*'. HTML 5 will be used as a solid base for the website that will have a so called "responsive design", which means the site will be accessible from mobile devices and tablet PC's (such as iPad, etc.).
A detailed view of every record is available and has a unique URL and is fully search engine optimized (SEO). By using this method all the records will be indexed by search engines (Google, Bing, etc.) and thus it will be easier to find a result in a regular browser. The possibility to view the images up close is provided by converting the images to a tiled view format using JPEG2000.

The full image repository will be exported to a site index, which will be periodically used by search bots to index the full collection.
The Linked Open Data initiative provides semantic web capabilities. Data from GeoNames (see geonames.org) will be available thus providing contextual information about spatial information for an image or a subject.

3.2.3. Enrichment of the collection

To facilitate and encourage the aggregation of content by not only the experts, but also by the general public, the Daguerreobase project will provide the following capabilities in the database:

- Commenting:

A comment can be placed with every record and provide a base for discussion and community building.
The capability of placing comments will provide possibilities for enriching metadata.
All this User Generated Meta Data will be searchable by the Search Engine and will be saved separately but still connected to the object in the collection. The User Generated Meta Data can be used by Linked Open Data protocols providing a valuable and rich source for information for other search queries.

- Linking to External Information (Triples, RDF, Linked Open Data):

DBpedia:
DBpedia is considered the Semantic Web mirror of Wikipedia. It will be used to tag items from Daguerreobase to topical items from Wikipedia.
GeoNames:
The GeoNames geographical database covers all countries and contains over eight million. It will be used to tag geospatial information on Daguerreobase.
By using different kinds of methods to search other databases the contextual information of the Daguerreobase will be significantly enhanced. Within the Resource Descriptive Framework (RDF) so called triples are used to create links between Linked Open Data from GeoNames and DBpedia (both part of the LOD cloud).

3.2.4. Collection management

With its extensive experience in facilitating the cultural heritage sector, PIM implemented a collection management system built for connecting metadata and collections.

- Data-Entry:

To create advanced metadata the collection management system uses a data entry tool.

- Triples:

The collection management system provides the ability to use and manage triple relations. Triple relations are a core part of connecting the information in the Daguerreobase to the semantic web.

- Thesauri/entries lists:

To make use of the possibilities of Linked Open Data, thesauri/entries lists are built following the SKOS (Simple Knowledge Organization System) methodology. The proposed collection management system will make use of the Open Source SKOS management system OpenSKOS for the user to be able to manage all used thesauri.

3.3 Specifications implemented Linked Open Data

PIM implemented the Open Search API as described in D5.3. Results are published in RDF/XML format, which is recommended for the Linked Open Data.
The result page contains the items with all fields required by the open search standard as well as the rdf description of the document. This rdf description is used consistently throughout the website and all the services related to it. This rdf description might contain links to sources from other linked open data projects like the Geonames thesaurus. It also contains links to all images associated with the daguerreotype.

An example of such an item is shown below.

```
− <rdf>
  − <edm:Timespan default:about="#5375673902754">
      <edm:begin>1855</edm:begin>
      <edm:end>1855</edm:end>
    </edm:Timespan>
  − <edm:ProvidedCHO>
      <dc:date rdf:resource="#5375673902754"/>
    − <dc:description>
        Pendant of CRE2; J.W. Blom and S.W.V. Kiebert married in 1956; [referencenumber acquisition:] AWI-2007-4;
      </dc:description>
    − <dc:description>
        Framed daguerreotype. Portrait of J.W. Blom in 1855.
      </dc:description>
      <dc:title>Justina Wilhelmina Blom (1832-1916)</dc:title>
      <dc:subject>portrait;woman</dc:subject>
      <dc:subject>Blom, Justina Wilhelmina</dc:subject>
      <dc:subject>person</dc:subject>
      <dc:subject>Annotation</dc:subject>
      <dc:identifier>57f8e2da-c9fc-11e3-9221-432519f2dbd2</dc:identifier>
      <dc:identifier>NFM-CRE1</dc:identifier>
      <edm:type>IMAGE</edm:type>
      <dc:type>daguerreotype</dc:type>
      <dc:language>English</dc:language>
      <dc:creator rdf:resource="#agent_63a4ac22-c9fc-11e3-abc4-b3beaef03542"/>
      <dc:creator rdf:resource="#agent_6455e758-c9fc-11e3-8bf9-f7b98bd58009"/>
    </edm:ProvidedCHO>
  − <edm:Agent rdf:about="#agent_63a4ac22-c9fc-11e3-abc4-b3beaef03542">
      <skos:prefLabel>Dr. Schneider</skos:prefLabel>
      <foaf:name>Dr. Schneider</foaf:name>
      <skos:note>Dr. Schneider Linkstr. 9 Berlin</skos:note>
    </edm:Agent>
  + <edm:Agent rdf:about="#agent_6455e758-c9fc-11e3-8bf9-f7b98bd58009"></edm:Agent>
  − <ore:Aggregation rdf:about="http://daguerreobase.org/type/57f8e2da-c9fc-11e3-9221-432519f2dbd2">
      <edm:isShownAt>http://images.memorix.nl/dag/thumb/250x250/</edm:isShownAt>
    − <edm:isShownBy>
        http://daguerreobase.org/type/57f8e2da-c9fc-11e3-9221-432519f2dbd2
      </edm:isShownBy>
      <edm:provider>Daguerreobase</edm:provider>
      <edm:dataProvider>Daguerreobase</edm:dataProvider>
      <edm:rights rdf:resource="http://creativecommons.org/licenses/by/4.0/">Creative Commons – Attribution – (BY)</edm:rights>
      <edm:hasView rdf:resource="http://images.memorix.nl/dag/thumb/250x250/bf9a0e6f-8e8c-5979-e583-c98c0ebb6cf4.jpg"/>
      <edm:hasView rdf:resource="http://images.memorix.nl/dag/thumb/250x250/d7dc9b4e-d8b0-3920-2b22-f0277b59db75.jpg"/>
    </ore:Aggregation>
  + <edm:WebResource rdf:about="http://images.memorix.nl/dag/thumb/250x250/bf9a0e6f-8e8c-5979-e583-c98c0ebb6cf4.jpg"></edm:WebResource>
  + <edm:WebResource rdf:about="http://images.memorix.nl/dag/thumb/250x250/d7dc9b4e-d8b0-3920-2b22-f0277b59db75.jpg"></edm:WebResource>
</rdf>
```

Annex 1, Copy of Chapters 4.6 – 4.8 of D2.1 of the Linked Heritage project on LD and LOD

Chapter 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[1]	**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

1 The abbreviation in brackets after a format's name is the „namespace" for that format.

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.
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 (See: **http://www.rkbexplorer.com**) is a system for publishing linked data, developed during the EC-funded *ReSIST* project (See: **http://www.resist-noe.org**). 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
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 (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][2]	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	CC BY-SA	103,000

(t4gm.info)		
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

[2] This will eventually be published as *CC0*

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.

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 *NTriples*, 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.