Grant Agreement No. ICT-2009-270082
Project Acronym PATHS
Project full title **Personalised Access To cultural Heritage Spaces**

Report accompanying D2.1: Processing and Representation of Content for First Prototype

Authors: **Eneko Agirre (UPV/EHU)**
Oier Lopez de Lacalle (UPV/EHU)

Contributors: **Aitor Soroa (UPV/EHU)**
Mark Stevenson (USFD)
Samuel Fernando (USFD)
Nikos Aletras (USFD)
Antonis Kukurikos (i-Sieve)
Kate Fernie (MDR)

Project funded under FP7-ICT-2009-6 Challenge 4 – "Digital Libraries and Content"	
Status	Final
Distribution level	Public
Date of delivery	02/12/2011
Type	Other
Project website	http://www.paths-project.eu
Project Coordinator	Dr. Mark Stevenson University of Sheffield

Keywords	PATHS, adaptive systems, cultural heritage, content curation, information access, information systems, learning, natural language processing, thesaurus, representations, content analysis, ontology extension, intra-collection links, background links
Abstract	This report accompanies and describes the contents of Deliverable 2.1 "Processing and Representation of Content for First Prototype". The deliverable comprises the data produced by WP2 "Content Processing and Enrichment" which is to be used in the first prototype. The data has been released in DVDs and is also available from the subversion server of the project. The data comprises a private collection as delivered by Alinari, and four collections from Europeana: the Culture Grid and SCRAN collections from the UK and the Hispana and Cervantes collections from Spain. The items have been enriched with the results of the content analysis, as well as terms from vocabularies, intra-collection links and background links.

Change Log

Version	Date	Amended by	Changes
0.1	31-10-2011	Eneko Agirre Oier Lopez de Lacalle	Outline version
0.2	11-11-2011	Eneko Agirre Oier Lopez de Lacalle Aitor Soroa Samuel Fernando Nikos Aletras Mark Stevenson	Sections: 1, 3.2, 3.5; 2, 4, 5, 8; 3; 7 and 8; 6 and 8;
0.3	13-11-2011	Eneko Agirre	Overall review and edits
0.4	14-11-2011	Antonis Kukurikos Oier Lopez de Lacalle	7 (other background links) Overall review and edits
0.5	22-11-2011	Eneko Agirre Oier Lopez de Lacalle	Overall review and minor edits
0.6	23-11-2011	Oier Lopez de Lacalle	Formatting issues
0.7	28-11-2011	Eneko Agirre Oier Lopez de Lacalle	Overall review and minor edits
0.8	29-11-2011	Oier Lopez de Lacalle	Formatting issues

Table of Contents

List of Tables

List of Figures

Executive Summary

The objective of WP2 is to collate content from Cultural Heritage sources, format it to allow convenient processing and augment it with additional information that will enrich the user's experience. The additional information includes links between items in the collection and from items to external sources like Wikipedia. The resulting data forms the basis for the paths used to navigate the collection, providing a collection of content for the first PATHS prototype and defining the standards for content format and descriptive metadata.

In this first release of the data, we include a private collection as delivered by Alinari, and four collections from Europeana: the Culture Grid and SCRAN collections from the UK and the Hispana and Cervantes collections from Spain.

The working package includes the following tasks:

- Task 2.1: Content Collection and Representation.
- Task 2.2: Content Analysis.
- Task 2.3: Ontology Extension.
- Task 2.4: Intra-Collection Links.
- Task 2.5: Background Links.

UPV/EHU, USFD and i-Sieve participate in this WP.

Tasks 2.1 and 2.2 are active up to month 28. Tasks 2.3 through 2.5 are active in two rounds. Tasks 2.3 is active months 7-16 and 25-30, and tasks 2.4 and 2.5 are active months 11-16 and 25-30. The deliverable includes the data produced on those tasks, but note that tasks 2.4 and 2.5 have run only for one month.

The contents in deliverable D2.1 are to be used in the first prototype, due month 16. Therefore, WP2 will produce updates of the data produced in tasks 2.2 - 2.5 according to the feedback received from WP4.

The contents of D2.1 are available from http://paths.avinet.no:81/svn/paths/tags/D2.1 or in a DVD on request.

Data Content

This is the structure of the contents.

- D2.1_alinari: Alinari 24 Ore Spa in several formats. In total, the size of the folder is about 300M when the data is compressed.
    - PATHS-14072011.zip Original data
    - ESE.tar.bz2 Data mapped to ESE
    - SAF.tar.bz2 Output of the linguistic analysis
    - ESEpaths Folder with intra-collection links and background links
        - BackgroundLinksWiki: background links to Wikipedia pages.
        - IntraLinks: similar items for each item
- D2.1_europeana: It comprises the content from Europeana in several formats. The total size (compressed) is about 650M.
    - ESE.tar.bz2 Original data. Once decompressed has the following structure:
        - SCRAN: 00401_Ag_UK_Scran_oai_scran.xml.gz
        - CULTURE GRID (divided in three collections)
            - 09405_Ag_UK_ELocal/09405_Ag_UK_ELocal.xml.gz
            - 09405a_Ag_UK_ELocal/09405a_Ag_UK_ELocal.xml.gz
            - 09405b_Ag_UK_ELocal/09405b_Ag_UK_ELocal.xml.gz
        - CERVANTES: 90901_L_ES_BibVirtualCervantes_dc.xml.gz
        - HISPANA: 09407_Ag_ES_ELocal.xml.gz
    - SAF.tar.bz2 Output of the linguistic analysis
    - ESEpaths Folder with background links
        - BackgroundLinksWiki: background links to Wikipedia pages.
    - Filtered2.tar Forlde with the list of the defined subset of items for the PATHS project.

The Background links to resources other than Wikipedia will be incorporated in the following months. The intra-collection links for the Europeana data will be provided in the following months. This is due to the fact that the respective tasks were planned to start in November.

1 Introduction

This report accompanies and describes the contents of Deliverable 2.1 "Processing and Representation of Content for First Prototype". The deliverable comprises the data produced by WP2 "Content Processing and Enrichment" which is to be used in the first prototype. The data has been released in DVDs and is also available from the subversion server of the project:

http://paths.avinet.no:81/svn/paths/tags/D2.1

Note that the release makes reference to the contents in these subversion directories:
http://paths.avinet.no:81/svn/paths/trunk/data/D2.1_alinari/
http://paths.avinet.no:81/svn/paths/trunk/data/D2.1_europeana/

The objective of WP2 is to collate content from Cultural Heritage sources, format it to allow convenient processing and augment it with additional information that will enrich the user's experience. The additional information includes links between items in the collection and to external sources like Wikipedia. The resulting data forms the basis for the paths used to navigate the collection, providing a collection of content for the first PATHS prototype and defining the standards for content format and descriptive meta-data.

In this first release of the data, we include a private collection as delivered by Alinari, and four collections from Europeana: the Culture Grid and SCRAN collections from the UK and the Hispana and Cervantes collections from Spain.

The content of the target collection, both meta-data and textual information, was analyzed with Natural Language Processing tools to identify relevant pieces of information that can be used for generating links. The information identified includes specific attributes of the items as typically found in the meta-data and also additional information that gives context to the item as found in the textual information that accompanies the items (people, organizations, dates and locations).

The subject field was also processed and automatically compared with some relevant vocabularies (two for English and one for Spanish). The items in the target collections were linked to each other based on the information provided by the data analysis task (links available for the Alinari data alone). In addition, the items have been linked to background information as made available in Wikipedia. The background links to other sources and the intra-collection links for Europeana items will be incorporated later in the project. The respective tasks were planned to start in November and the first results will be ready soon.

This report is structured as follows. Section 2 describes the organization of the data contained in the deliverable. Section 3 describes the collections included in the deliverable, including the Europeana and Alinari datasets, detailing the subset of the Europeana collections, the representation of the meta-data based on ESE, the mapping of Alinari records to ESE and some issues with Europeana versions. Section 4 reports the linguistic processing of all collections. In Section 5 we review the automatic enrichment of items with terms from some of the vocabularies used in Europeana collections. Sections 6 and 7 review, respectively, the enrichment of items with lists of similar items and with background information. Finally, Section 8 presents the quantitative and qualitative evaluation of the automatically produced data.

2 Contents of the data deliverable

The data in the deliverable is divided according to the main two data sources in PATHS project: Alinari and Europeana collections. Data from both sources follow the same organizational structure. The main folders are:

- **D2.1_alinari**: Comprises the content provided by Alinari 24 Ore Spa in several formats. In total, the size of the folder is about 300M when the data is compressed.
- **D2.1_europeana**: Comprises the content from Europeana in several formats. The total size (compressed) is about 650M.

The data in the deliverable has the following structure:

- **PATHS-14072011.zip** (only in D2.1_alinari): Original data from Alinari 24 Ore Spa. It is first set of Alinari images available for download (11.268 photos and data) http://fototeca.alinari.it/download/PATHS-14072011.zip (244 MB). Specifically it contains 11.268 images in 128 px without watermark and 11.268 images in 256 px with small watermark. The meta-data are in Italian and English.

- **ESE** (ESE.tar.bz2): Contains the source data for the linguistic processing. The data, divided in different collections, is represented in the Europeana Semantic Elements (ESE) specification format. The collections in Europeana (the data below D2.1_europeana folder) were downloaded on 2011-01-28 (there might be some changes in the collections since that date). The PATHS project will focus on the next four collections regarding Europeana (each stored in one folder), the first two are in English, while the last two are in Spanish:
  - SCRAN: 00401_Ag_UK_Scran_oai_scran.xml.gz
  - CULTURE GRID (divided in three collections)
    - 09405_Ag_UK_ELocal/09405_Ag_UK_ELocal.xml.gz
    - 09405a_Ag_UK_ELocal/09405a_Ag_UK_ELocal.xml.gz
    - 09405b_Ag_UK_ELocal/09405b_Ag_UK_ELocal.xml.gz
  - CERVANTES: 90901_L_ES_BibVirtualCervantes_dc.xml.gz
  - HISPANA: 09407_Ag_ES_ELocal.xml.gz

  The Alinari folder contains the source data mapped into ESE format (Europeana Semantic Elements) (see Section 3.6 for further details about the mapping). The collection consists of 11,268 items (PATHS-14072011.zip).

- **SAF** (SAF.tar.bz2 in the future): This folder comprises the output of the different linguistic processors (part of speech tagging, multiword recognition, name entity

classification...). All the information will be stored in SAF format (Simple Annotation Format), an internal representation to share in a unify way the information extracted from the different linguistic analysis.

The output files are organised hierarchical structures distributed in directories. Each file contains the information about a single record in the collection. In order to avoid problems with the IDs of the records from Europeana (some of them are URLs) we have created an internal ID for each record. The folder of each collection is organised as follows:

```
00/
        0000/
                0000000.saf
                0000001.saf
                ...
                0000099.saf
        0001/
                0000100.saf
                ...
                000199.saf
...
99/
…
```

The mapping between internal IDs and Europeana/Alinari ID are stored in "map_id" file, which is located on the top of the collection's folder.

- **ESEpaths** (ESEpaths.tar.bz2): This folders stores the output of the WP2 presented in D2.1. In it base, the format is an extension of ESE, and includes the intra-collection links (T2.4, described in Section 6) and the background links (T2.5, described in Section 7). This folder is divided in the following sub-folders
    - **BackgroundLinksWiki**: It stores the background links to Wikipedia pages.
    - **IntraLinks**: It stores the similar items for each item

The contents themselves are described in the sections below.

3 Content collection and representation

3.1 List of Europeana sources

The PATHS project has focused on the following four collections in Europeana, where in total the University of the Basque Country has processed 2,112,993 items.

- **SCRAN**: The SCRAN collection (http://www.scran.ac.uk) is an online resource with over 360,000 images and media from different museums, galleries and archives in Scotland.
- **CULTURE GRID**: The Culture Grid (http://www.culturegrid.org.uk) is a digital content provider service of Collection Trust (http://www.collectionstrust.org.uk). The main objective of The Culture Grid is to bring the wide range of collection available to as many people as possible. It contains over one million item records from 40 different UK collections (national and regional museums, libraries...)
- **CERVANTES:** Biblioteca Virtual Miguel De Cervantes (http://www.cervantesvirtual.com) comprises digitalized Spanish text in various formats. In total, the online library contains about 75,000 works from different periods.
- **HISPANA**: The electronic headquarters of the Biblioteca Nacional de España (http://www.bne.es) allows the access to the digitalised diverse content that collates from audio and video to texts and drawing. The material is collected from the different national and regional museums, libraries and other type of providers.

The dataset used within the PATHS project comprises Europeana items from two collections for English (Culture Grid and SCRAN) and two collections for Spanish (Hispana and Cervantes). The data was collected by downloading the items using the Europeana subversion server, located at on the 28th of January of 2011:

http://sandbox08.isti.cnr.it/svn/trunk/sourcedata/xml/COLLECTION_NAME/

Specifically, we used the following sub-collections:

- Culture Grid (English)
  - 09405_Ag_UK_ELocal
  - 09405a_Ag_UK_ELocal
  - 09405b_Ag_UK_ELocal
- SCRAN (English)
  - 00401_Ag_UK_Scran_oai_scran

- Hispana
  - 9407_Ag_ES_ELocal
- Cervantes Virtual Library:
  - 90901_L_ES_BibVirtualCervantes_dc

The following table (Table 1) compares the contents of each collection.

Collection	No. + type of items
Culture Grid	
09405_Ag_UK_ELocal	381450 images
09405a_Ag_UK_ELocal	93105 images
09405b_Ag_UK_ELocal	73226 images
Scran	
00401_Ag_UK_Scran_oai_scran	2144 texts 267814 images 38618 video 2226 sound
Hispana	
09407_Ag_ES_ELocal	1129640 texts 105493 images
Cervantes Virtual	
90901_L_ES_BibVirtualCervantes_dc	19278 texts

Table 1: Collection summary

The first column in Table 1 specifies the sub-collection of the main collection. The second column shows the number and type of the items in the sub-collection. All in all, we have nearly 2.1 million items (2,112,993 items).

3.2 Alinari Collection

Regarding the Alinari collection (http://www.alinari.it), the University of the Basque Country has processed about 11,268 records from Alinari photographic archives, which in total comprise 4,000,000 photographs. Due to different data formats, previously Alinari data was converted to ESE format in order to harmonise the whole process. For that we have established mapping between Alinari meta-data representation and ESE (see Section 3.6).

3.3 Defining a subset for Europeana

Some of the items in Europeana have short and uninformative titles and lack a description. We considered filtering those out for the first prototype. We tested two possibilities:

- **Filter 1**: discarding all items that have no description, **or** have the length of the title shorter than 4 words, or have a title which has been repeated more than 100 times.
- **Filter 2:** discarding all items that have no description **and** have either the length of the title shorter than 4 words, or have a title which has been repeated more than 100 times

As an example Table 2 shows one of the item with short and uninformative meta-data, related to the item shown in Figure 1. This lack of accurate meta-data form a challenge to language processing techniques since it is difficult to extract enough information to model the item accurately.

<record>
<dc:format>text/html**</dc:format>**
<dc:identifier>http://www.fitzmuseum.cam.ac.uk/opacdirect/76397.html**</dc:identifier>**
<dc:language>en-GB**</dc:language>**
<dc:publisher>The Fitzwilliam Museum, Cambridge, UK**</dc:publisher>**
<dc:source>Fitzwilliam Museum**</dc:source>**
<dc:subject>lead-glazed earthenware, figure**</dc:subject>**
<dc:subject>lead-glazed earthenware**</dc:subject>**
<dc:subject>figure**</dc:subject>**
<dc:title>Winter**</dc:title>**
<dcterms:isPartOf>Fitzwilliam Museum**</dcterms:isPartOf>**
<dcterms:provenance>bequeathed by Glaisher, J.W.L., Dr, 1928 [C.930-1928]**</dcterms:provenance>**
<europeana:country>uk**</europeana:country>**
<europeana:isShownAt>http://www.fitzmuseum.cam.ac.uk/opacdirect/76397.html**</europeana:isShownAt>**
<europeana:language>en</europeana:language>
<europeana:object>http://www.peoplesnetwork.gov.uk/dpp/resource/2476506/stream/thumbnail_image_jpeg
</europeana:object>
<europeana:provider>CultureGrid**</europeana:provider>**
<europeana:type>IMAGE**</europeana:type>**
<europeana:uri>http://www.europeana.eu/resolve/record/09405b/C94EFC46CD0C5258DA4F02FAA814187
1F89F9AD5**</europeana:uri>**
</record>

Table 2: Example of an item with a short and uninformative title and lack of a description.

Figure 1:Object related to the information given in Table 2.

The results for each filter are shown in Table 1. The first filter was too strict, and would have discarded approximately 50% in all collections, except in 09405b_Ag_UK_ELocal, where all

the items would be discarded, as they don't have descriptions. Thus we applied the second filter, which is more conservative and only discards 19% of the items.

	Total	Filter1	Filter2
09405_Ag_UK_ELocal	381450	108195	344791
09405a_Ag_UK_ELocal	93105	36811	91186
09405b_Ag_UK_ELocal	733226	0	30981
SCRAN	320107	156282	292938
Hispana	1235133	605428	1219731
Cervantes	19278	2756	14983

Table 3: Results for each filter on Europeana datasets. The figures denote the number of accepted items.

3.4 Europeana Semantic Elements Specifications (ESE)

Items in the collection follow the Europeana Semantic Elements Specifications (ESE) described in http://www.version1.europeana.eu/c/document_library/get_file?uuid=104614b7-1ef3-4313-9578-59da844e732f&groupId=10602. ESE is an XML format for describing Europeana items, formally specified in the XML Schema found at http://www.europeana.eu/schemas/ese/ESE-V3.4.xsd.

Here is an example of an Europeana record described following the ESE guidelines:

```
<record>
<dc:identifier>http://www.beamishcollections.com/collections/display.asp?ItemID=679</dc:identifier>
<europeana:uri>http://www.europeana.eu/resolve/record/09405/51160DC81B1E1EF56DCE20D42D886792
    E0388025</europeana:uri>
<dc:title>RMS Mauretania</dc:title>
<dc:source>Beamish Treasures</dc:source>
<dc:description>Line of boilers to be fitted to the RMS Mauretania. From album from The Wallsend Slipway
    and Engineering Co. 1906.</dc:description>
<dcterms:isPartOf>Beamish Treasures</dcterms:isPartOf>
<dc:subject>Industry</dc:subject>
<dc:subject>Shipbuilding</dc:subject>
<dc:type>Image</dc:type>
<europeana:object>http://www.peoplesnetwork.gov.uk/dpp/resource/2061112/stream/thumbnail_image_jpeg
    </europeana:object>
<europeana:provider>CultureGrid</europeana:provider>
<europeana:isShownAt>http://www.beamishcollections.com/collections/display.asp?ItemID=679</europeana:isShownAt>
<europeana:hasObject>true</europeana:hasObject>
<europeana:country>uk</europeana:country>
<europeana:type>IMAGE</europeana:type>
<europeana:language>en</europeana:language>
</record>
```

Table 4: A sample record following ESE.

Recently the Europeana Foundation decided to switch to a new format called "Europeana Data Model (EDM)"[1] for structuring the data that Europeana will be ingesting, managing and publishing. The Europeana Data Model is a major improvement on the Europeana Semantic Elements (ESE), as it supports the full richness of the content providers' metadata, and also enables data enrichment from a range of third party sources.

We anticipate switching to the EDM data model for the second prototype of the PATHS system.

3.5 ESEPaths

The result of the content processing stage is twofold:

- We create new intra-links between Europeana records, with the aim of connecting "similar" records together.
- We create a set of background links which relate Europeana records with entities of some external resource (e.g. Wikipedia)

The new relations are describes using a new format called "ESEPaths", which enhances ESE records with new elements. Specifically, ESEPaths is a superset of ESE, including also the following elements and attributes:

- <paths:related_item> relates one <record> with another. It has the following attributes:
  - confidence (float, optional): confidence of the association
  - method (string, optional): which method produced the association
  - field (string, optional): the field of the ESE record where the anchor for this relation is located.
  - field_no (integer, optional): the index of the field in the ESE record.
  - start_offset (integer, optional): the offset (in characters) within the field element where the text anchor begins.
  - end_offset (integer, optional): the offset (in characters) within the field element where the text anchor ends.
- <paths:background_link> relates one <record> with an entity of some external resource. It has the following attributes:
  - source (string, required): the name of the external resource.
  - confidence (float, optional): confidence of the association
  - method (string, optional): which method produced the association
  - field (string, optional): the field of the ESE record where the anchor for this relation is located.

[1] http://version1.europeana.eu/c/document_library/get_file?uuid=48b552e1-c71d-4f3f-a1ae-c1929cb7de76&groupId=10605

- field_no (integer, optional): the index of the field in the ESE record.
- start_offset (integer, optional): the offset (in characters) within the field element where the text anchor begins.
- end_offset (integer, optional): the offset (in characters) within the field element where the text anchor ends.

Here is an example of an ESEPaths record. Only the new elements are shown:

```
<record>
<dc:identifier>http://www.beamishcollections.com/collections/display.asp?ItemID=679</dc:identifier>
<!-- … -->
<!-- related items -->
<paths:related_item confidence="0.8" start_offset="0" end_offset="11"
                    field="dc:subject" field_no="0" method="LDA">
http://www.europeana.eu/portal/record/09405t/A6F9A3871B29568A2CE46E74C32C30C31E752BA8
</paths:related_item>
<paths:related_item confidence="0.2" start_offset="0" end_offset="8"
                    field="dc:description" field_no="0" method="LDA">
http://www.europeana.eu/portal/record/00401/FBD53036D38EF4F34D317C13AE6426CDE5A818BE
</paths:related_item>

<!-- background links items -->
<paths:background_link source="wikipedia" start_offset="0" end_offset="11"
                       field="dc:subject" field_no="0"
                       confidence="0.015" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Archaeology
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="0" end_offset="8"
                       field="dc:description" field_no="0"
                       confidence="0.274" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Scotland
</paths:background_link>
</record>
```

Table 5: A sample record following ESEPaths

3.6 Mapping Records to ESE

The table below shows the mapping used to convert Alinari's source data into ESE standard. In the table we indicate which meta-data type in Alinari corresponds to Europeana ESE specification meta-data.

Alinari	**Europeana ESE**
Image ID	dc:identifier
Caption	dc:title
Date of photography	dc:temporal
Place of photography	dcterms:spatial
Detailed place of photography	dc:coverage
Photographer	dc:creator
Object	dc:format
Technique	dc:subject
Date of artwork	dc:date
Artist	dc:creator
Period and style	dc:subject
Artwork support	dc:medium
Location	dcterms:spatial
Events	dc:description
People	dc:description
Credit	dc:source
Keywords	dc:subject
Permission and Restrictions	dc:rights
Labels	dc:rights
FORMAT (b/w – col)	dc:format
ORIENTATION (Portrait – Landscape)	dc:description

Table 6: Mapping table of Alinari metadata to ESE specification

3.7 Europeana versions and accessing thumbnails

The version of the Europeana dataset downloaded in the beginning of the project has three relevant pointers: europeana id (<europeana:uri>), europeana thumbnails (<europeana:object>) and local collection id (<dc:identifier>).

The problem is that the data contained in each of these pointers is not constant:

- The <dc:identifier> field (according to Europeana this is the most stable) has changed for about 70,000 items between the first Paths version and the most recent version.
- The <europeana:object> field (which provides the link to thumbnail images) has changed in about 220,000 instances (but in a systematic way).

Fortunately, it is relatively easy to create a mapping between the old and new version of the <europeana:uri> fields using the resolve web service, as in the following example:
http://www.europeana.eu/resolve/record/09405a/DCD497B6DB6BAD9BE9F3C82BCAE7E723245FC398

It is also possible to access old URIs for thumbnails using the following API:
http://europeanastatic.eu/api/image?uri=<europeana:object>

For instance:
http://europeanastatic.eu/api/image?uri=http://www.peoplesnetwork.gov.uk/dpp/resource/2060284/stream/thumbnail_image_jpeg

Given the fact that the linguistic processing of all the items is a costly process, and that, at any given time, the Europeana datasets can be assigned new ids, the board decided to stick to the version of the dataset downloaded in the beginning of the project for the sake of this deliverable. If the prototype building on top of this data needs to access the updated URI, it can use the mapping above. If the prototype would need to access the thumbnails, it can also use the API mentioned above.

4 Content analysis

The University of the Basque Country carried out linguistic processing of the item records in Europeana and Alinari collections. The aim of the content analysis is described in the technical annex Task2.2, in which the analysis will drive to identify the relevant pieces of information that can be used to generate links. The processing has focus on linguistically interesting fields, which are those with text enough to be processed, in the Europeana Semantic Elements specification. Nevertheless, the PATHS project does not discard processing other information fields for the next versions. The processed fields are the following:

- ***Title***: A name given to the resource. Typically, a Title will be a name by which the resource is formally known.
- ***Description***: An account of the resource.
- ***Subject***: The topic of the resource.

In order to extract useful linguistic information (lemmas, part of speech, multiwords terms, named entities, etc.), after comparison of different linguistic processors (see Section 4.2) two well-know text analyzers were applied:

- We made use of **FreeLing** (http://garraf.epsevg.upc.es/freeling) to process data in Spanish and English. We performed the lemmatization, part of speech tagging and multiword recognition. In addition, named entity classification was carried out for the content in Spanish.
- **Stanford Named Entity Recognizer and Classifier** (http://nlp.stanford.edu) was deployed to extract the occurring named entities. The extracted entity types (we did the same with Freeling for Spanish) were: PERSON, LOCATION, ORGANIZATION, DATE.
- In addition, we also run specific vocabulary matching over the items in the collections. See Section 5 for further details.

The output of the whole process comprises the information given by the different linguistic processors. All the information is stored in SAF format (Simple Annotation Format), the internal representation to share the information extracted from the different linguistic analysis in a unify way. The folders (one for each collection) are organised in a hierarchical fashion, and the linguistic information of an item is stored in one single SAF file, as the following example shows.

```
<?xml version="1.0" encoding="utf-8"?>
<!-- sAF: simple Annotation Format :-) -->
<sAF>
  <!-- Store all information for every record -->
  <record rid="000-000-005-549-C" repository="Europeana" collection="SCRAN">
    <!-- general terms -->
    <terms>
      <term lemma="instrumentation" pos="N.NN" field="description"/>
      <term lemma="print" pos="V.VBN"/>
      <!-- ... -->
    </terms>
    <!-- multiwords -->
    <mws>
      <mw lemma ="additional_parts" pos="N"/>
      <mw lemma ="st_martin's_church_yard" pos="N"/>
      <!-- ... -->
    </mws>
    <!-- entities -->
    <entities>
     <!--
      possible netypes:

      PER: person
      LOC: location
      ORG: organization
      DATE: date
      ...
      Also, we could add more information gathered from other sources:
      Wikipedia category
      Specialized vocabulary
      ...
      -->
     <entity netype="PER" lemma="Picasso"/>
   </entities>
  </record>
 </sAF>
```

Table 7: Example of SAF record.

4.1 Content Processing

This section describes the steps given from the input ESE format to the output SAF format, where all the essential information is agglutinated in a simple and legible way. First, the given and the obtained contents will be described: The selection of interesting fields, the application of different linguistic processors, and the obtained output. Next, some of the main issues are described with the proposed solutions. Finally, organizational description of the output file structure is given.

Processing Steps

The main steps are described as follows:

- **Input ESE files.** In this step ESE item records are extracted in order to process each item independently. Each item is stored in one single file, in which the file name is obtained from its identifier. In addition, this would allow simple parallelization techniques to speed-up the whole process.

- **Identifier mapping.** Due to some Identifiers are complex strings like URLs (e.g.: "http://www.vads.ac.uk/large.php?uid=1") could difficult the whole process. Specially, the naming of the item-files. Thus, in order to avoid such problem we build a new identifier index, and store apart the mappings to recover the original identifiers. Moreover, mapping to original identifiers to integers allows hierarchical organization and avoid problems of storing too many files in a unique folder (e.g.: Ext3 file system limits up to 1,000,000 files per directory).

- **Text extraction.** As mentioned in previous sections, for the first version only fields which a priori are interesting have been processed. That is, fields with enough text content are taken into account. All in all, in this first run we have focused on *description, subject* and *title* meta-data fields.

- **Linguistic processing.** As an output we obtain two types of file for each record ID and field. The first output consists of an output of XML specified by the Kyoto Annotation Format (KAF) (Bosma et al, 2009) file with different linguistic information obtained from FreeLing. The second file type is the output obtained by Stanford NERC tool. Finally, for each item in the collection a file with all the interesting features are comprised in SAF format: word-form, lemmas, PoS, NERC (organisation, location, person), other entities (dates, misc).
- As a summary, the obtained files are named as follows:
    - Output 1:
        - Get one single KAF file for each item and record
        - File naming: *internalID.field.kaf*
    - Output 2:

- Get one NERC file for each item and record. This only is for English content, since NERC information is already extracted from Freeling.
- File naming: *internalID.field.ner*

o Output 3:

- Get one SAF file for each item.
- File naming: *internalID.saf*

Figure 2 summarizes all the steps done from the input ESE format to the ouput SAF format.

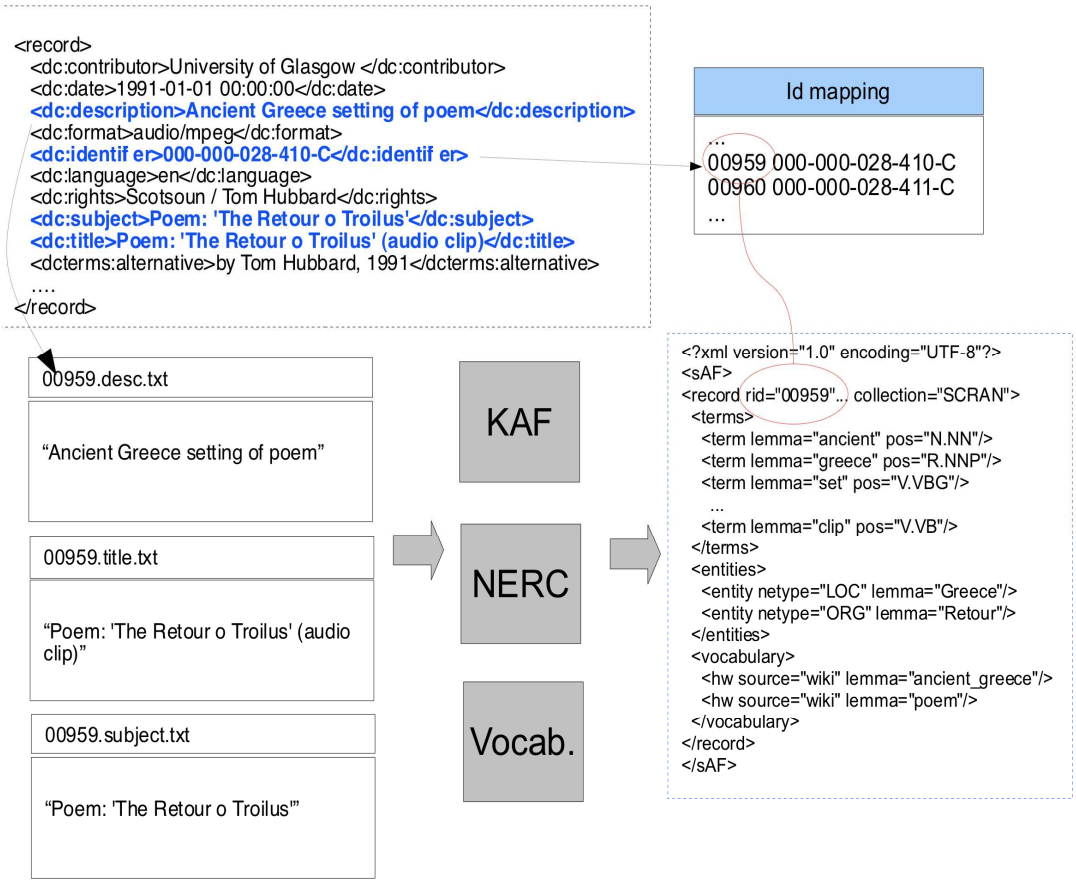


Figure 2: Content Processing Example. This figure summarizes the steps done from input ESE format to output SAF files.

Issues

There are some issues that have been taken into account and solutions are being considered. Some of the solutions have been taken for the first data release, but some others will be taken into consideration for the next version:

- Regarding the item identifiers, we have taken the following decisions:
  - If there are more than one identifiers take the first. For instance, the data from SCRAN collection is given with two different item identifier.

- If there is any collision, like the same field value in different records, what have to be done? We do not take any decision, since we do not find any collision between items.
- Some identifiers are complex strings like URLs (e .g: http://www.vads.ac.uk/large.php?uid=1): we decided to build a new identifier index based on integers.

- For the repeated fields (except identifiers) we decided to concatenate the content data of the field type.

- NERC with FreeLing (only for Spanish). Currently FreeLing takes the longest entity. E.g: "biblioteca_américa_de_la_universidad_de_santiago_de_compostela", althoug at this point of the project is enough, the extraction of "sub-named entities" ("biblioteca_américa" and "universidad_de_santiago_de_compostela") could be interesting to improve the quality of the data processing.

Directory Structure

The output files will be organized hierarchical structures distributed in directories. The solution given to the data storage allows the 1) identification of each item with an integer, simplyfing the indexing of the data, and 2) avoids the problem of storing too many file in a unique folder. For instance, Ext3 file system does not allow more than 100,000 file one directory. Thus, as example, a possible output structure could be like this for each output type (KAF, NER, and SAF):

```
00/
        0000/
                0000000.desc.kaf
                0000000.desc.ner
                0000000. title.kaf
                0000000. title.ner
                0000000.saf
                ...
                0000001.saf
                ...
                0000099.saf
        0001/
                …
                0000100.saf
                ...
                000199.saf
...
99/
```

...

Note that while KAF and NERC outputs are field-wise, SAF output join all the information of each field in one file per item.

4.2 Comparison of the linguistic processors

We experimented with two different language processing systems (and their associated formats) to determine which might be better for the task of processing the data in Europeana and other collections.

The first system is FreeLing (Padro et al, 2010). This is an open source library of language processing tools which contain the following functions in the library (as specified on the FreeLing website). The output of this system is in XML as specified by the Kyoto Annotation Format (KAF) (Bosma et al, 2009).

The second system is GATE (Cunningham et al, 2002). This is also an open source framework comprising similar language processing tools to FreeLing. The output gives GATE annotations in XML.

Comparison methodology

The key tasks we want to perform with the linguistic processing are following:

- Identification of nouns via part of speech (PoS) tagging.
- Lemmatization
- Multiword recognition
- Named entity classification: Recognition of dates, places, people, organisations etc.

The evaluation was carried out comparing the accuracy of each system performing the task listed above. For testing we used a selection of texts from the description metadata field from the English and Spanish collections in Europeana. In the case of English, we based on 3 texts extracted from the SCRAN collection and 4 texts from the Culture Grid collection. For Spanish, we used 3 texts taken from the Cervantes collection. The texts are listed below in the "Details of comparison" section.

Results for English

- **Noun identification and lemmatizer** are broadly similar for the two systems.
- **Multi-word recognition.** FreeLing seems to do this better. For example it correctly identifies the phrase "St Martin's Church Yard" which GATE misses. It also does so with more precision: GATE seems to throw up many false positives. *Note: FreeLing*

automatically groups together words into noun phrases, while GATE does this in separate steps using the multi-word and noun phrase chunking tools.

- **Named entity recognition:**
  - **Dates**: GATE seems to do better at recognising dates, finding 1810-11 as a date in one case, and "between 1888 and 1955" in another, which FreeLing misses.
  - **People**: Both systems seem to be generally similar for recognising people.
  - **Locations**: GATE seems to do better at this, correctly finding more places, such as Aberdeen and London in two separate texts which FreeLing misses. GATE also seems more precise, with KAF throwing up more false positives (spring and summer as places).

Results for Spanish

- **Identification of nouns and lemmatization.** Both systems perform similarly well, and there are no significant differences.
- **Multi-word recognition**. The performance of both processors is similar this task, although Freeling is more actively developed for Spanish than GATE and might be considered to be slightly superior in performance. Since Freeling multi-word recognizer is based on syntactic chunks it is able to return "*Vicerrectorado de Extensión Universitaria de la Universidad de Alicante*" as unique multiword, while GATE returns *"Vicerrectorado de Extensión Universitaria*" and "*Universidad de Alicante*" separately.
- **Named entity recognition.** GATE is not trained to extract named entities in Spanish and extracts them deploying gazetteers and it does not carry any kind of disambiguation. By contrast, FreeLing recognizes and classifies named entities based on a machine learning algorithm. Overall, the extraction of FreeLing seems to be more accurate and robust for Spanish.

Conclusions of the comparison

Overall, both systems perform similarly when extracting PoS tags and lemmas in both languages. Nevertheless, depending on the language one system outperforms the other in the rest of the tasks. Specifically, the main difference comes from the named entity recognition and classification task, where FreeLing do better for Spanish and GATE for English (since FreeLing does not have any NERC module).

Regarding English, GATE seems to better at finding dates and locations, while FreeLing has a better multi-word recogniser. Therefore to maximize accuracy the best approach seems to be to use components from both systems and combine the results.

Work is now proceeding using FreeLing to annotate the texts from Europeana. The idea is that if we find that FreeLing is not producing good enough results we can augment with some GATE components at a later stage.

According to the results for Spanish, it is difficult to assess both processors with only three short texts, but we believe that in Spanish Freeling looks more robust than GATE. For Spanish NERC module is active in Freeling, which it performs well, while GATE seems to be based on gazetteers and it does not apply any kind of disambiguation technique. Gazetteers can produce many false positive.

5 Ontology extension

The University of the Basque Country has made the first attempt to extend the information in Europeana collection with information from an ontology. In this first attempt, we run a string matching algorithm on the different collections to check the existing overlap between the items and different vocabularies/thesauri.

The analysis has focused on two collections for English and one for Spanish content:

- The Library of Congress Subject Headings (LCSH) and English Heritage-NMR for English
- The Spanish version of the LCSH vocabulary for Spanish

As an overview the coverage is higher than expected:

- when using NMR we cover 42% of the items (27% do not have dc:subject, 31% have dc:subject, but no matching term)
- when using LCSH we cover 46% of the items (27% do not have dc:subject, 27% have dc:subject, but no matching term)
- when using Spanish LCSH we cover 48% of the Spanish the items (71% do not have subject, 11% have subject, but not term matching)

For the first prototype this means that we can browse a significant subset of the collection using LCSH and NMR already.

5.2 Vocabulary Coverage on Europeana Collections

Item Coverage

Taking into account only the dc:subject field in ESE, these are the figures in percentage of the item with at least one term linked to the vocabulary:

	LCSH	NMR	Overlap	Join
CULTURE GRID	31%	36%	26%	43%
SCRAN	69%	51.%	43%	78%
CERVANTES	57%	.--	--	--
HISPANA	14%	--	--	--

Table 8: Results of the coverage of the analyzed vocabularies over the 4 collection from Europeana.

Following some absolute figures:

- Total English records: 858,582
  - Records with dc:subject: 626,948 (73.0% of total)
- Total Spanish records: 1,254,409
  - Records with dc:subject: 325,598 (25% of the total)

- **NMR**:
  - Records mapped: 360,660
    - 42.0% of total items (dc:subject, dc:title, dc:description)
    - 57.5% of dc:subject
  - **Culture Grid**: 199,425 mapped out of 548,047
  - **SCRAN**: 161,235 mapped out of 310,802

- **LCSH**:
  - Records mapped: 391,290
    - 45.6% of total items (dc:subject, dc:title, dc:description)
    - 62.4% of dc:subject
  - **Culture Grid**: 174,779 mapped out of 548,047
  - **SCRAN**: 216,511 mapped out of 310,802

- **Spanish LCSH:**
  - Records mapped: 604072 (out of 1,254,409)
    - 48.1% of total items (dc:subject, dc:title, dc:description)
    - 57.9% of item with dc:subject
  - **Cervantes**: 13571 mapped out of 19,276
  - **Hispana**: 590501 mapped out of 1,235,133

Details on EnglishHeritage-NMR (NMR)

We extracted a dictionary of 12,474 headers.

SCRAN collection

- Number of item covered taking into account only "dc:subject" field: 161,235 (out of 310,802), which in total means that the 51% of the items are covered.
- number of different terms matched over “dc:title”, “dc:description” and “dc:subject”: 4,795
- total number of strings matched: 1,015,480

Terms	Frequency	Summed frequency	Coverage (%)
street	35981	35981	3.54
architecture	25378	61359	6.04
house	23947	85306	8.40
castle	20413	105719	10.41
road	17881	123600	12.17
fife	16718	140318	13.81
church	16646	156964	15.45
transport	16378	173342	17.06
headquarters	16308	189650	18.67
book	11943	201593	19.85
bridge	10564	212157	20.89
...	...	...	...
cavetto	1	1015470	99.99
Pan pipe	1	1015471	99.99
Cold bath	1	1015472	99.99
Commemorative tablet	1	1015473	99.99
Crop mark	1	1015474	99.99
Hay house	1	1015475	99.99
Surveyors office	1	1015476	99.99
Sussex stone	1	1015477	99.99
Horse trapping	1	1015478	99.99
Rolling mill	1	1015479	99.99
Sailcloth mill	1	1015480	100

Table 9: Sample of matched words in SCRAN collection as regards of NMR vocabulary.

Culture Grid collection

- Number of items covered regarding "dc:subject": 199425 (out of 548047), which means that in total the 36% of the items are covered
- number of different terms matched over “dc:title”, “dc:description” and “dc:subject”: 6,120
- Total number of strings matched: 2634719

Terms	Frequency	Summed frequency	Coverage (%)
coin	136038	136038	5.16
street	77516	213554	8.10
road	61127	274681	10.42
house	50848	325529	12.35
copper	40990	366519	13.91
church	32609	399128	15.14
can	28995	428123	16.24
sculpture	25556	453679	17.21
cast	22797	476476	18.08
window	22420	498896	18.93
token	18697	517593	19.64
...	...	...	...
Ridge piece	1	2634709	99.99
Symbol stone	1	2634710	99.99
Storage pit	1	2634711	99.99
Vehicle repair workshop	1	2634712	99.99
Resettlement camp	1	2634713	99.99
District library	1	2634714	99.99
Roadside settlement	1	2634715	99.99
slasher	1	2634716	99.99
Portal dolmen	1	2634717	99.99
Heel stiffener	1	2634718	99.99
Coal fired power station	1	2634719	100

Table 10: Sample of matched words in Culture Grid collection as regards of NMR vocabulary.

Details on Library of congress (LCSH)

We extracted a dictionary of 750,750 headers in which:

- 410,521 are tagged as "prefLab" in skos
- 340,229 are extracted from "altLab" (similar to synonyms)

SCRAN collection

- Number of items covered regarding "dc:subject": 216511 (out of 310802), which in total means that the 69% of the item are covered.
- Number of different terms matched over “dc:title”, “dc:description” and “dc:subject”: 11,150
- Total number of strings matched: 1,476,145

Terms	Frequency	Summed frequency	Coverage (%)
photograph	85699	85699	5.80
street	35993	121692	8.24
location	30356	152048	10.30
architecture	24378	176426	11.95
scottish	24140	200566	13.58
east	21345	221911	15.03
st	20752	242663	16.43
church	19994	262657	17.79
road	19970	282627	19.14
black	18974	301601	20.43
fife	16718	318319	21.56
...	...	...	...
schism	1	1476136	99.99
histori	1	1476137	99.99
astronautics	1	1476138	99.99
Higher education	1	1476139	99.99
vases	1	1476140	99.99
Roman baths	1	1476141	99.99
British dominions	1	1476142	99.99
Harmonica players	1	1476143	99.99
County services	1	1476144	99.99
Zambesi river	1	1476145	100

Table 11: Sample of matched words in SCRAN collection as regards of LCSH vocabulary.

Culture Grid collection

- Number of items covered regarding "dc:subject": 174779 (out of 548047), which in total means that the 31% of the item are covered.
- number of different terms matched over “dc:title”, “dc:description” and “dc:subject”: 14,228
- Total number of strings matched: 2,902,084

Terms	Frequency	Summed frequency	Coverage (%)
coin	136823	136823	4.71
street	77851	214674	7.39
road	62173	276847	9.53
date	49641	326488	11.25
copper	40910	367398	12.65
church	39062	406460	14.00
design	37723	444183	15.30
st	35914	480097	16.54
can	28997	509094	17.54
britain	26023	535117	18.43
number	25366	560483	19.31
...	...	...	...
Face painting	1	2902075	99.99
weightlifting	1	2902076	99.99
financiers	1	2902077	99.99
ducklings	1	2902078	99.99
Vernon family	1	2902079	99.99
Super heroes	1	2902080	99.99
dharmachakra	1	2902081	99.99
Verdon family	1	2902082	99.99
refrigerants	1	2902083	99.99
Donnelly family	1	2902084	100

Table 12: Sample of matched words in Culture Grid collection as regards of LCSH vocabulary.

Details on the Spanish version of Library of congress

Cervantes Collection

- Number of items covered regarding “dc:subject”: 11040 (out of 18120), wich int total means 54% of the items are covered.
- Number of different terms matched over “dc:title”, “dc:description” and “dc:subject”: 1,276
- Total number of string matched: 24224

Terms	Frequency	Summed frequency	Coverage (%)
Español	4744	4744	19.58
Poesía	2472	7216	29.78
La	2105	9321	38.47
teatro_español	1908	11229	46.35
Novella	1088	12317	50.84
Cine	432	12749	52.62
Lengua	387	13136	54.22
comedia	366	13502	55.73
Catalán	303	13805	56.98
literatura_infantil	288	14093	58.17
Arte	288	14381	59.36
…	...	...	...
teatro_popular	1	24214	99.95
arzobispos	1	24215	99.96
Domino	1	224216	99.96
negociación	1	24217	99.97
lógica formal	1	24218	99.97
orden sagrado	1	24219	99.97
Nitrates	1	24220	99.98
servicio militar obligatori	1	24221	99.98
proceso penal	1	42222	99.99
gestión de memoria	1	4223	99.99
Bili	1	24224	100

Table 13: Sample of matched words in Cervantes collection as regards of Spanish LCSH vocabulary.

Hispana Collection

- Number of items covered regarding “dc:subject”: 177,589 (out of 307,478), wich int total means 57% of the items are covered.
- Number of different terms matched over “dc:title”, “dc:description” and “dc:subject”: 1,960
- Total number of strings matched: 828,875

Terms	Frequency	Summed frequency	Coverage (%)
prensa	177569	177569	21.42
interés	176977	354546	42.77
la	65866	420412	50.72
correspondencia	63215	483627	58.34
moral	37868	521495	62.91
comercio	31518	553013	66.71
industria	24695	577708	69.69
color	17225	594933	71.77
correo	16871	611804	73.81
cabeza	13586	625390	75.45
enseñanza	13100	638490	77.03
...	...	...	...
lengua española-modismos	1	828865	99.99
aranceles de aduanas	1	828866	99.99
pallium	1	828867	99.99
difuntos	1	828868	99.99
nepotismo	1	828869	99.99
vicios	1	828870	99.99
régimen parlamentario	1	828871	99.99
salmo gairdneri	1	828872	99.99
máquinas	1	828873	99.99
conciencia_moral	1	828874	99.99
godos	1	828875	100

Table 14: Sample of matched words in Hispana collection as regards of Spanish LCSH vocabulary.

6 Intra-collection links

Sheffield University has explored analysing the Europeana and Alinari collections to identify similar items, and thereby add intra-collection links.

6.1 Background and Motivation

Europeana and Alinari collections contain a vast amount of items stored in an unstructured way. This makes navigation difficult for users who may want to browse collections and find relevant items to one that they have already viewed.

Europeana items are stored in a meta-data format including information about their title, description and subject. In addition the Paths Alinari data contains information about caption and subject keywords. This textual information can be analysed and used to compute semantic similarity between items. Computing similarity can help as to relate items and create links between them. For example an item can be linked to the top N most relevant ones. Therefore, these links can be used to create paths of similar items so that a user could easily access relevant items. This is important because improves navigation and may lead to better user experience.

In recent years, techniques have been developed for estimating semantic similarity between texts. These are classified into two broad categories. The first category includes corpus-based methods (Blei et al. 2003; Jurafsky and Martin 2008; Mihalcea et al. 2007) that use the distribution of words in a corpus. The methods in second category rely on knowledge resources such as thesauri, dictionaries and semantic networks (Agirre et al. 2009; Grieser et al. 2011; Harrington 2010).

The aim of this task was to create links between items in the collection. To achieve this we apply both corpus and knowledge-based methods to compute similarity. We show that LDA (a probabilistic topic for text) has the best performance over other methods. In addition, we find for each item its top 25 relevant items together with confidence values and converting those results to the ESE format.

6.2 Computing Similarity between Items

We apply both corpus-based and knowledge-based semantic similarity measures on Europeana and Alinari data to identify intra-collection links. In this section, we define four corpus-based methods which are the Jaccard coefficient, tf.idf, N-gram overlap and LDA. In addition we show how we adapted a knowledge-based measure (WLVM) to measure text similarity. We assume that each item contains textual information and we consider them as discrete documents.

Corpus-based Methods

Jaccard coefficient is defined as the number of common words between sets of words of two texts divided by the total number of words in both sets.

$$sim_{JC}(a,b) = \frac{|A \cap B|}{|A \cup B|}$$

where a,b are two documents and A, B their sets of words respectively.

The tf.idf weighted cosine measure (Jurafsky and Martin 2008) is the cosine similarity between two documents represented as vectors. Each vector contains a sequence of words in a document together with their frequencies. Moreover, each word is weighted by its inverse document frequency (idf) which is the logarithmic value of the total number of documents in a collection divided by the number of documents that contain the specific word. The tf.idf measure is calculated as follows:

$$sim_{tf.idf}(a,b) = \frac{\sum_{w \in a,b} tf_{w,a} tf_{w,b} (idf_w^2)}{\sqrt{\sum_{a_i \in a} (tf_{a_i,a} idf_{a_i})^2} \times \sqrt{\sum_{b_i \in b} (tf_{b_i,b} idf_b)^2}}$$

where $tf_{w,x}$ is the frequency of the term w in $x \in \{a,b\}$ and idf_w is the inverted document frequency of the word w.

The N-gram overlap measure (Patwardhan et al. 2003) matches common word sequences between documents. The similarity score computed is the sum of the squares of the length of the common word sequences:

$$sim_{ngram}(a,b) = \frac{\sum_{n \in N} n^2}{|A| + |B|}$$

where a,b are two document texts, n is the length of an overlapping subphrase and N is the set of the overlaps found in both texts.

The semantic similarity between documents can be compared by analysing the underlying topics described within them. Methods relying on statistical modelling for discovering topics in a collection of documents are called topic models. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is such a method which clusters words in topics. The basic assumption is that subsets of words contained into documents can share a common topic. It is an unsupervised method and assigns words to topics creating a uniform Dirichlet prior probability distribution of each topic over words. Each document is represented as a set of various topics. Each topic contains a set of words and it is modelled as an infinite mixture over a set of topic probabilities. The output of LDA is a distribution of topic probabilities of each word in a document. Similarity can be computed as the cosine of the angle between document vectors.

An LDA model is trained using a corpus of documents and each document as a mixture of topics obtained from the model. Gensim is a Python framework that offers several text

probabilistic modelling algorithms. We used gensim to create LDA probability distributions over topics. Firstly, we used all the items of Europeana or Alinari to create the corpus. Secondly, we created a dictionary mapping words to their frequency in the collections. By constructing the dictionary, we transformed all documents to bags-of-words. The next step was to create the LDA model. Two LDA models consisting of 100 and 25 topics were trained over the Europeana and Alinari corpus respectively. All the bag-of-words were transformed to LDA vectors. LDA vectors consist of topics corresponding to an item and a probability of that topic. Finally, the similarity between two items computed as the cosine of the angle between their vectors.

Knowledge-based methods

Finally, we use a knowledge-based measure called Wikipedia Link Vector Model (WLVM) (Milne 2007). WLVM uses both the link structure and the article titles of Wikipedia to measure the similarity between two concepts (words). We first identify Wikipedia concepts contained in the item data and then we perform pairwise comparisons between all concepts in the items.

We adapt WLVM to a measure for estimating text similarity. First, we run Wikifier which is a service for identifying Wikipedia concepts within text segments. Then, we carry out pairwise comparison between concepts of two documents.

$$sim_{WLVM}(A,B)=\frac{1}{2}\left(\frac{\sum_{a\in A} maxSim_{WLVM}(a,B)}{|A|}+\frac{\sum_{b\in B} maxSim_{WLVM}(b,A)}{|B|}\right)$$

where A,B are the set of Wikipedia concepts detected for two items and $maxSim_{WLVM}(w,X)$ is the maximum WLVM similarity score between term w and the set of Wikipedia concepts X where $X=\{A,B\}$.

6.3 Contents in D2.1

We applied the LDA-vector approach to the Paths Alinari data. We used caption and keywords of each item to compute similarity. For each item we stored the top 25 most similar items. Figure 1 shows a sample image from the Alinari data set and the top 5 most similar images identified by the approach.

Figure 3: Alinari pictures.

After computing similarity, all the data were converted to the ESE format to be integrated to the PATHS prototype. The output size of Alinari data is 27.8MB and it took approximately 1.5 hours to compute.

7 Background links

7.1 Wikipedia links

This section describes how items in Europeana and Alinari have been automatically enriched with background links into relevant Wikipedia articles by Sheffield University, a process referred to here for brevity as ‘wikification’.

Motivation

Items in Alinari and Europeana often have very limited meta-data descriptions. Adding links to relevant articles in Wikipedia provides background which is of interest to users, thus presenting a richer experience for those browsing these collections. Additionally the links may also help to categorise and organise the collections using the Wikipedia category hierarchy.

Background

The first work to address this specific task was Mihalcea and Csomai (2007). Their procedure for wikification used two stages. The first stage was detection, which involved identifying the terms and phrases from which links should be made. The most accurate method for this was found to be using link probabilities. Formally the link probability of a phrase is defined as the number of Wikipedia articles that use it as an anchor, divided by the number of Wikipedia articles that mention it at all.

The next stage, disambiguation, ensures that the detected phrases link to the appropriate article. For example the term plane usually links to an article about fixed wing aircraft. However it sometimes points to a page describing the mathematical concept of a theoretical surface, or of the tool for flattening wooden surfaces. To find the correct destination, Mihalcea and Csomai (2007) train a classifier using features from the context. Although the quality of results obtained is very good, a large amount of pre-processing is required, since the entire Wikipedia must be parsed.

The problem of topic indexing is similar to that of wikification. The aim of this is to find the most significant topics; those which the document was written about. Medelyan et al. (2008) describes an approach to topic indexing with Wikipedia. The same approach is used for link detection as Mihalcea and Csomai (2007). However a different approach is used for disambiguation, which achieves similar results, but with much less computational expense. This is done by balancing the commonness (prior probability) of each sense with the relatedness of the sense (how well it relates to its surrounding context).

Milne and Witten (2008) builds upon this previous work. A set of 500 articles was used as training data - thus the software learns to disambiguate and detect links in the same way as Wikipedia editors. Disambiguation of terms within the text is performed first. A machine learning classifier is used with several features. The main feature used is the commonness of a target sense, which is defined by the number of times it is used a destination from some anchor text. For example the token ‘tree’ links to the woody plant 93% of the time, 3% of the time to the type of graph, and 3% to the computer science concept. The disambiguation algorithm is predisposed to select the first of these senses rather than the more obscure ones.

The next feature used is relatedness. This computes the similarity of two articles by comparing their incoming and outgoing links. Only unambiguous terms (terms which only link to one article) are used as context terms. The relatedness of a candidate sense is computed as the weighted average of its relatedness to each of the unambiguous context articles. It is recognized that some context terms are more useful than others. For example the term ‘the’ is unambiguous, but is not useful for disambiguation. Thus each context term is weighted by link probability, which measures how often the term is used as a link within Wikipedia articles. Thus millions of articles contain the term ‘the’ but do not use it as a link, and thus would have a very low link probability weighting. Other context terms may be outliers which do not relate to the central thread of the document, and therefore a further weighting measure used is the relatedness of each context term to the others. The C4.5 algorithm (Quinlan 1993) is then used for the classifier.

Once terms have been disambiguated against articles in this way, the next step is to determine which terms should be linked and which should not. Again a machine learning classifier is used. The relatedness and link probability values are again used as features. Additionally the disambiguation confidence from the previous step is used as a feature. Since it is more useful to provide links for very specific terms, the generality of each topic is also included as a feature. Finally a set of features that define the anchors location and spread within the document are used.

Application to PATHS data

To find appropriate Wikipedia links within the PATHS meta-data, the Wikipedia Miner API (Milne and Witten 2008) is used. This loads an instance of Wikipedia into a database, and then provides functions to detect and disambiguate relevant Wikipedia links from plain text. The PATHS data comprises the Europeana and Alinari collections with associated meta-data. The Alinari data and the English sub-collections from Europeana are used as input for the wikification process.

The English Europeana data set contains 858,582 meta-data records in the ESE (Europeana Semantic Elements) XML format. The Alinari data set is far smaller with 11,628 records. The Alinari data has been mapped by the PATHS team into the same ESE format for convenience in processing.

The most descriptive and meaningful fields in ESE are the title, subject and description fields. All records have a title. However the subject and description fields can be quite brief. About 200,000 items have no description at all. Half of the items have less than 13 tokens in the description field. For the subject field, half of the items have less than 2 tokens. It may therefore be necessary in future to filter out items with very little meta-data.

However at this stage all items are processed with the title, subject and description fields aggregated together and input into the Wikipedia Miner, version 1.2.0. The default training models for the disambiguation and link detection are used, along with default parameters. The background links are appended into the ESE XML records, following the ESEpaths scheme suggested by the PATHS team. This preserves all the original content of each record, and adds the background links as XML fields which use offset annotation to identify the relevant parts of the record as anchor text for the links. The attributes also point to the field where the link is found (for example ”dc:title” or ”dc:subject”). Since sometimes there exists more than one field with the same name, the attribute also gives the field number to identify which of these is referenced.

An example record is shown here:

<record>
<dc:contributor>Historic Scotland</dc:contributor>
<dc:date>1301-01-01 00:00:00</dc:date>
<dc:date>1700-12-31 00:00:00</dc:date>
<dc:date>Medieval/Early modern 14th century - 17th century</dc:date>
<dc:description>Craigmillar Castle (Armorial Panel)</dc:description>
<dc:description>Scotland, Edinburgh& Lothians Two and a half miles south east of central Scotland, Edinburgh</dc:description>
<dc:format>image/jpeg</dc:format>
<dc:identifier>000-000-004-177-C</dc:identifier>
<dc:identifier>000-000-004-177-R</dc:identifier>
<dc:language>en</dc:language>
<dc:rights>Crown Copyright reproduced courtesy of Historic Scotland</dc:rights>
<dc:subject>Architecture and Buildings</dc:subject>
<dc:title>Craigmillar Castle (Armorial Panel)</dc:title>
<europeana:country>uk</europeana:country>
<europeana:isShownAt>http://www.scran.ac.uk/000-000-004-177-C</europeana:isShownAt>
<europeana:language>en</europeana:language>
<europeana:object>http://images.scran.ac.uk/RB/images/thumb/0033/00330127.jpg</europeana:object>
<europeana:provider>Scran</europeana:provider>
<europeana:rights>http://www.europeana.eu/rights/rr-p/</europeana:rights>
<europeana:type>IMAGE</europeana:type>
<europeana:uri>http://www.europeana.eu/resolve/record/00401/3366F300459094DED0AB3AE72555B3772DA206F0</europeana:uri>
<paths:background_link source="wikipedia" start_offset="0" end_offset="11" field="dc:title" field_no="0" confidence="0.906" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Craigmillar_Castle
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="0" end_offset="11" field="dc:title" field_no="0" confidence="0.753" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Craigmillar

</paths:background_link>
<paths:background_link source="wikipedia" start_offset="0" end_offset="18" field="dc:title" field_no="0" confidence="0.906" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Craigmillar_Castle
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="12" end_offset="18" field="dc:title" field_no="0" confidence="0.017" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Castle
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="20" end_offset="28" field="dc:title" field_no="0" confidence="0.017" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Roll of arms
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="0" end_offset="12" field="dc:subject" field_no="0" confidence="0.017" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Architecture
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="0" end_offset="11" field="dc:description" field_no="0" confidence="0.906" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Craigmillar Castle
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="0" end_offset="11" field="dc:description" field_no="0" confidence="0.753" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Craigmillar
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="0" end_offset="18" field="dc:description" field_no="0" confidence="0.906" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Craigmillar Castle
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="12" end_offset="18" field="dc:description" field_no="0" confidence="0.017" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Castle
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="20" end_offset="28" field="dc:description" field_no="0" confidence="0.017" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Roll of arms
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="0" end_offset="8" field="dc:description" field_no="1" confidence="0.029" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Scotland
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="21" end_offset="29" field="dc:description" field_no="1" confidence="0.580" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Lothian
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="65" end_offset="81" field="dc:description" field_no="1" confidence="0.017" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Central Belt
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="73" end_offset="81" field="dc:description" field_no="1" confidence="0.029" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Scotland
</paths:background_link>

```
<paths:background_link source="wikipedia" start_offset="83" end_offset="92" field="dc:description"
field_no="1" confidence="0.064" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/Edinburgh
</paths:background_link>
<paths:background_link source="wikipedia" start_offset="83" end_offset="92" field="dc:description"
field_no="1" confidence="0.030" method="wikipedia-miner-1.2.0">
http://en.wikipedia.org/wiki/University of Edinburgh
</paths:background_link>
</record>
```

Table 15: Example of a record enriched with background links

For completeness, all candidate topic links identified by Wikipedia Miner are included. Typically when these links are used in applications low confidence scores would be filtered out. Also it is worth noting that sometimes the same article is assigned to different parts of the text, and also multiple articles are assigned to the same piece of text. So for example the article Cragmillar Castle is assigned to the text 'Craigmillar' and also to the text 'Craigmillar Castle'. The article Craigmillar is also assigned to the text 'Craigmillar'. The task of finding the best anchor text for each article is a separate problem, and thus would require further post-processing.

7.2 Other Background Links

. The purpose of the specific task is to provide a notion of the impact of the items included in the Europeana and Allinari collections examined by the PATHS project.
In further detail, the results of the task will provide links to web resources that refer to an item in the collection, indicating the perceived relevance of the resource, the impact of the resource via the "buzz factor" metric created by i-sieve and the expressed sentiment towards the item, via comments or articles contained in the resource.

Methodology

The analysis process used for the task is summarised in the following figure.

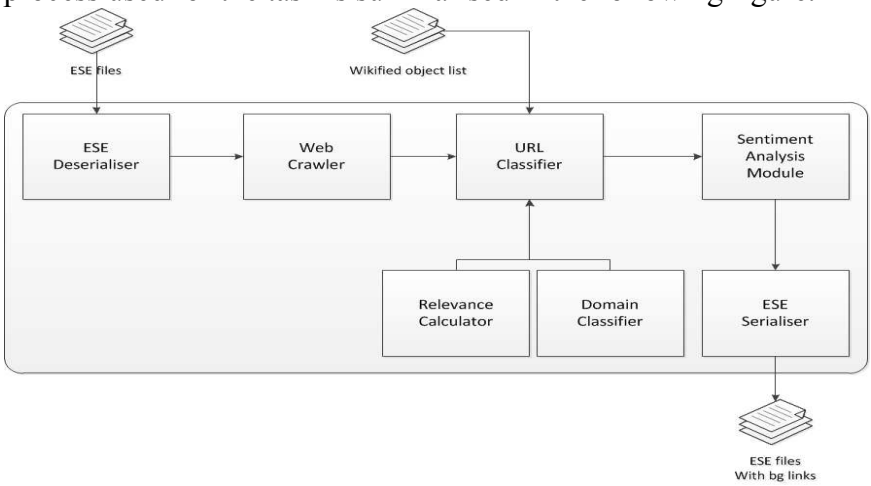


Figure 4 Summary of the analysis of extraction of the external resources background link.

The input of the module is an ESE-formatted file, containing the description of an arbitrary amount of collection items.

During the first stage of the process, the records contained in the ESE file are de-serialised to custom Record Java objects, encapsulating the available data for each item. The **De-serialiser** is built as an XML parser and uses native Java resources.

The second step realises the retrieval of possibly relevant web resources, using a **Web Crawler** that implements the Google Search and Bing APIs. The input for the crawler is the title property of each Record object, as it provides the more restrictive description of the corresponding item. The topmost 200 results are taken into account for further processing, since it is unlikely that results below the threshold will be relevant or significant.

The URL Classification process provides a metric for the relevance of a web resource with a collection item. Moreover, it classifies the resources as cultural or academic.

The **Relevance Calculator** associates a score with each resource. The calculation relies on the presence of various features in the text content of the resource. In more detail, we take into account the following:

- The presence of the title and description item properties.
- The existence of data from the wikibox of the wikipedia resources associated with the item (see section 7.1).
- The presence of terms included in a custom ontology, designed by i-sieve. The ontology defines phrases that indicate the academic or cultural natural of the resource. The terms are not of equivalent significance, thus weights are introduced for each term. The weights are subject to modifications, depending on the feedback from human annotators.

At the end of the process, each resource is associated with a relevance score. A sample of the resources are to be analysed by human annotators, in order to validate the weights of the terms in the ontology, as well as the threshold over which a resource should be characterised as relevant with adequate confidence. The system is retrained based on the remarks of the annotators and the Relevance Calculation is performed again.

The resources approved after the previous steps are subjected to sentiment and impact analysis. The sentiment analysis is performed by modules implementing the Rule Based Analysis algorithms developed by i-sieve, which have been tested in numerous projects of the company. The impact analysis provides the “buzz factor” metric, which takes into account the relevance of the resource and the visibility of the specific article as well as the visibility of the web site that contains it.

The process is finalised with the creation of a ESEPaths document via the **ESE Serialiser** module. The ESE Serialiser extracts the properties of the Record objects that represent the items for which relevant background links were found and creates an XML file containing the corresponding information.

Results

The input for the module were:

- An ESE file containing the representation of 11,628 items from the Allinari collection.
- An ESE file containing the representation of 858,852 items from the English Europeana collections.

As mentioned, the topmost 200 search results were taken into account for processing, resulting to 2,325,600 links for the Allinari dataset and 17,1770,400 links for the Europeana dataset.

The links for each item which had a score greater than a certain threshold were considered for approval as a background link. If there were more than 10 background links , the 10 with the biggers "buzz factor" were selected. Following the ESE Paths format, an example for a file containing background links can be seen in the following table:

<record>
<dc:identifier>ACA-F-002452-0000</dc:identifier>
<dc:title>The public gardens on Viale Strozzi, in Florence</dc:title>
<dc:temporal>1890 ca.</dc:temporal>
<dcterms:spatial>Florence</dcterms:spatial>
<dc:creator>Alinari, Fratelli</dc:creator>
<dc:format>Glass plate</dc:format>
<dc:subject>N Silver salt gelatin</dc:subject>
<dc:source>Alinari Archives-Alinari Archive, Florence</dc:source>
<dc:subject>Flowerbed;;Element;;For Landscapes;;Structure;;Architecture;;Avenue;;Fountain;;Garden and Park;;Landscape</dc:subject>
<dc:rights>WARNING: Permission must be required for non editorial use. Please contact Alinari Archives;</dc:rights>
<dc:format>B/W</dc:format>
<paths:background_link source="google" start_offset="0" end_offset="0" field="dc:title" field_no="0" confidence="0.884" method="i-sieve crawler">
http://www.casasantapia.com/images/gardens/mediciriccardigarden.htm
</paths:background_link>
<paths:background_link source="google" start_offset="0" end_offset="0" field="dc:title" field_no="0" confidence="0.778" method="i-sieve crawler">
http://www.answers.com/topic/strozzi-family</paths:background_link>
<paths:background_link source="google" start_offset="0" end_offset="0" field="dc:title" field_no="0" confidence="0.672" method="i-sieve crawler">
http://www.italyguide.com/index2.php?m=menu&h=exhibitions&sp=spot_tick </paths:background_link>
<paths:background_link source="bing" start_offset="0" end_offset="0" field="dc:title" field_no="0" confidence="0.514" method="i-sieve crawler">
http://www.expedia.co.uk/florence-hotels-strozzi-palace-hotel.h899169.hotel-information</paths:background_link>
</record>

Table 16 An exemplary ESEPaths file containing web background links

It is expected that the analysis of the discovered web resources will be completed within M12 of the project. A summarisation and statistical analysis of the results will be performed as soon as the complete data is available.

8 Quality Assurance and Evaluation

In this section we describe the quality assurance procedures set up for the data in D2.1. These procedures have been set up jointly with the participants in Work Package 5. The procedures are the following.

- Content analysis: we will study the scalability, reporting throughput (items per minute, KB of content per minute) and size of the information produced in SAF (absolute, per item and relative to the size of the input ESE). We will also produce an estimate of the speed and sizes for collections 10 times larger.

- Ontology extension: we will provide a quantitative study of the amount of matches found with some of the existing vocabularies. Regarding scalability we will report throughput (items per minute, KB of content per minute) and size of the information produced in SAF (absolute, per item and relative to the size of the input ESE).

- Intracollection links: we will provide a quantitative study of the amount of similarity links between items. Regarding quality, we will evaluate how the automatic links compare to manual annotations.

- Background links: we will provide a quantitative study of the amount of links to wikipedia articles. Regarding quality, we will check manually a sample of the links between equivalent items and articles.

Below we describe the results for each of the tasks.

8.1 Content analysis

The content analysis carried out consists of several expensive subtasks. Each of the steps requires sophisticated methods to analyse and extract the linguistic information as mentioned in Section 4.

In total the processing of the whole dataset took 5 weeks divided in 7 CPUs. The production of KAF, in which the main linguistic information is extracted, was performed in 4 weeks, and the classification of named entities (NERC production) was done also in 4 weeks. Regarding NERC production only for English collections done, since for Spanish NERC is performed when producing KAF. These two subtasks where parallelized in order to speed the whole process up. Finally, in the last week the extracted information was filtered out to SAF files. This last step was run in a single CPU.

Table 16 shows the estimation of the throughput for different linguistic processing and SAF converting. Note that the ratio of items per minute is estimated on a single CPU

	#collections	#CPU	#weeks	Item/minutes
KAF	4	7	4	7.5
English NERC	2	7	4	3.1
SAF	4	1	1	295
TOTAL	-	-	-	7.4

Table 17: Estimation of the throughput for different linguistic processing and SAF converting. Note that the ratio of items per minute is estimated on a single CPU.

Regarding SAF size, the 4 collections are 9GB big, being on average about 4.9KB the size of each item. The relative to the size of ESE each SAF item is almost 4 times bigger that an ESE item.

Taking into account the actual estimations, the pre-processing of datasets 10 times larger would be tedious, while the size would still manageable. More specifically, in that case, we would need about 50 weeks to process the whole dataset using 7 CPUs.

The figures show that the process is as expected for the size and nature of the collection. We have identified some possibilities to make the process faster, which we plan to deploy for the next version.

8.2 Ontology extension

The quantitative study of the amount of matches is shown in Section 5, where the number of matches per vocabulary and collection are given. Moreover, a detailed description of coverage and the different examples of matched words is also depicted.

Regarding the size, vocabularies SAF files are about 0.2KB each, which in total would take around 0.5GB. On average, the relative size to ESE input item is about 6 times smaller than an ESE item.

The performance in time of matching algorithm is not as expensive as the linguistic processing. Nevertheless, the algorithm could be expensive enough for bigger datasets, as for example the whole set of Europeana, in the case we are using a large vocabulary for matching. Roughly, the throughput of the system is about 3,500 items per minute.

8.3 Intracollection links

The various methods used are evaluated against a gold-standard data set of 30 pairs of items selected randomly from Europeana collections. This data set includes human judgements on the similarity between the items.

We obtained human judgements by conducting an online survey. Annotators were asked to rate similarity of a pair of items from 0 to 4 where 0 denotes completely unrelated items and 4 indicates completely related ones. A total of 74 responses received from which 38 were partially completed and apparently discarded. We considered the rest of 36 responses for creating the gold-standard. A number of obvious similar pairs (same title, image) were used to ensure that the answers have not been given randomly. Therefore, for each pair of item the gold-standard score is the average of all the human ratings.

We applied similarity measures using item titles and descriptions. We measure performance by computing Pearson's correlation coefficient between results of each method and the gold-standard. Table 7 shows the results of each method.

Measure	**Pearson's Correlation Score**
sim_{JC}	0.722
$sim_{tf.idf}$	0.705
sim_{ngram}	0.67
$sim_{LDA-vector}$	**0.807**
sim_{WLVM}	*0.651*

Table 18: Evaluation of similarity.

Results show that LDA-vector outperforms the rest of the measures and is chosen as the similarity measure.

The output size of Alinari data is 27.8MB and it took approximately 1.5 hours to complete. Therefore, the throughput ratio is about 125 items/min and 309 KB/min.

8.4 Background links

It takes approximately 24 hours to annotate all the Europeana items with Wikipedia links. This means a single item takes on average 0.1 seconds to process. Using a probability threshold of 0.5, the number of links found per item in Europeana follows this distribution:

Num. of links	Freq. (%)
0	33.56
1	27.27
2	19.56
3	10.1
4	5.03
5	2.16
6	1.07
7	0.54
8	0.26
9	0.15
>=10	0.29

Table 19: Number of links per item in Europeana.

More than 90% of the items have 3 or fewer links. Over a third has no links at all.

Evaluation

For evaluation purposes, a gold standard set of Europeana items was created. In total 100 items were randomly selected from Europeana. 25 of the items were filtered out since there was insufficient text to be processed, leaving 75 remaining. For each the title, subject and description text was aggregated together. The text was then manually annotated with a set of links to relevant Wikipedia articles (if any). It was found that 65 of these 75 had one or more relevant Wikipedia links (in the judgement of the annotator). The output of the Wikipedia Miner system was then judged against this gold standard. The evaluation methodology used here follows that of (Milne and Witten 2008), where only the set of links is judged, and not the anchor positions of text. As mentioned earlier, finding the correct anchors for the links in the text is a separate problem in itself. The system was evaluated using different probability cut-off points. Precision and recall and F-measure were calculated.

Probability cut-off	Precision	Recall	F
p > 0 (all links)	0.29	0.87	0.44
p > 0.2	0.46	0.31	0.37
p > 0.5	0.44	0.25	0.32

Table 20: Evaluation of links according to cut-off.

Using all links gives a high recall value of 87%. Eliminating low probability links results in a large drop in recall performance, but precision is improved.

9 Summary

This report accompanies and describes the contents of Deliverable 2.1 "Processing and Representation of Content for First Prototype". The deliverable comprises the data produced by WP2 "Content Processing and Enrichment" which is to be used in the first prototype. The data has been released in DVDs and is also available from the subversion server of the project:

http://paths.avinet.no:81/svn/paths/tags/D2.1-2011-11-30

The objective of WP2 is to collate content from Cultural Heritage sources, format it to allow convenient processing and augment it with additional information that will enrich the user's experience. The additional information includes links between items in the collection and to external sources like Wikipedia. The resulting data forms the basis for the paths used to navigate the collection, providing a collection of content for the first PATHS prototype and defining the standards for content format and descriptive metadata.

In this first release of the data, we include a private collection from Alinari, and four collections from Europeana: the Culture Grid and SCRAN collections from the UK and the Hispana and Cervantes collections from Spain.

The content of the target collection, both metadata and textual information, was analysed with Natural Language Processing tools to identify relevant pieces of information that can be used for generating links. The information identified includes specific attributes of the items as typically found in the metadata and also additional information that gives context to the item as found in the textual information that accompanies the items (people, organizations, dates and locations).

The subject field was also processed and automatically compared with some relevant vocabularies (two for English and one for Spanish). The items in the target collections were linked to each other based on the information provided by the data analysis task (links available for the Alinari data alone). In addition, the items have been linked to background information as made available in Wikipedia. The background links to other sources and the intra-collection links for Europeana items will be incorporated later in the project. The respective tasks were planned to start in November and the first results will be ready soon.

The contents of the deliverable will be updated with the outcome of WP2, which will continue to work concurrently to the development of the first prototype. The final outcome of WP2 will be released on the beginning of April, so it can be used in the first prototype.

Glossary

- **ESE**: Europeana Semantic Elements (ESE) specification format.
- **KAF**: Kyoto Annotation Format.
- **Latent Dirichlet Allocation**: A generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.
- **LDA**: Latent Dirichlet Allocation
- **Lemmatization**: The process of grouping together the different inflected forms of a word so they can be analysed as a single item.
- **Linguistic Processor**: A program which make use of natural language processing techniques in order to extract information from text.
- **Named Entity**: Usually predefined categories of atomic elements such as the names of persons, organization, locations and temporal expresions.
- **Named Entity Recognition and Classification**: Also known as entity identification and entity extraction, is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.
- **NERC**: Named Entity Recognition and Classification.
- **Multiword expression**: Made up of a sequence of two or more lexemes that has properties that are not predictable from the properties of the individual lexemes or their normal mode of combination.
- **Part of Speech tagging**: Part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context.
- **SAF**: Simple Annotation Format.

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M. and Soroa, A. (2009), "A study on similarity and relatedness using distributional and WordNet-based approaches", Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09 , Association for Computational Linguistics, Morristown, NJ, USA, p. 19.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), "Latent Dirichlet allocation", Journal of Machine Learning Research , Vol. 3, JMLR.org, pp. 993–1022.

Bosma, W. E. , P. Vossen, A. Soroa, G. Rigau, M. Tesconi, and A. Marchetti, and M. Monachini, and C. Aliprandi. KAF: A Generic Semantic Annotation Format. *In Proceedings of the GL2009 Workshop on Semantic Annotation, 2009.*

Budanitsky, A. and Hirst, G. (2006), "Evaluating wordnet-based measures of lexical semantic relatedness", Computational Linguistics , Vol. 32, pp. 13–47.

Cunningham, H., D. Maynard, K. Bontcheva and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002*.

Grieser, K., Baldwin, T., Bohnert, F. and Sonenberg, L. (2011), "Using ontological and document similarity to estimate museum exhibit relatedness", Journal on Computing and Cultural Heritage (JOCCH) , Vol. 3, ACM, p. 10.

Harrington, B. (2010), A semantic network approach to measuring relatedness, in 'International Conference on Computational Linguistics', pp. 356–364.

Mihalcea R. and A. Csomai. Wikify!: Linking Documents to Encyclopedic Knowledge. In CIKM, volume 7, pages 233–242, 2007.

Mihalcea, R., Corley, C. and Strapparava, C. (2006), Corpus-based and knowledgebased measures of text semantic similarity, in 'National Conference on Artificial Intelligence'.

Medelyan, O. , Witten, I.H. , and D. Milne. Topic Indexing with Wikipedia. In Proceedings of the AAAI WikiAI workshop, 2008.

Milne, D. (2007), Computing semantic relatedness using wikipedia link structure, in 'Proceedings of the New Zealand Computer Science Research Student Conference', Citeseer.

Milne, D. and I.H. Witten. Learning to Link with Wikipedia. In Proceeding of the 17th ACM conference on Information and knowledge management, pages 509–518. ACM, 2008.

Padró , L., M. Collado, S.Reese, M. Lloberes, and I. Castellón. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. *In Proceedings of 7th Language Resources and Evaluation Conference (LREC), 2010.*

Patwardhan, S., Banerjee, S. and Pedersen, T. (2003), Using measures of semantic relatedness for word sense disambiguation, in 'Proc. of the 4th Int'l. Conf. on Intelligent Text Processing and Computational Linguistics', pp. 241–257.

J.R. Quinlan. C4. 5: Programs for Machine Learning. Morgan Kaufmann, 1993.

Annex 1: Summary of Europeana data

Summary of English collections

Total records = 985346

dc:contributor

Values for 421786 records (42.8%)
Top values:

Value	Freq.
North East Midland Photographic Record	80765
The Scotsman Publications Ltd	66101
Royal Commission on the Ancient and Historical Monuments of Scotland	28112
Kirklees	26132
The Scotsman newspaper publisher	21060
The Scotsman' newspaper publisher	17774
National Museums of Scotland	16983
Scottish Motor Museum	10722
National Library of Scotland	9958
Newsquest (Herald& Times)	9909

Tokens: Mean=3.6, Mode = 0, Median = 0

dc:creator

Values for 463540 records (47.0%)
Top values:

Value	**Freq.**
Root	62703
Taunt, Henry	12914
Unknown	7594
Donna Nicoll	7394
unknown	7011
Lambeth Libraries	5700
Minter, Faye - Portable Antiquities Scheme	5695
Abigail Evans	5601
Andrews-Wilson, Liz - Portable Antiquities Scheme	5347
Daubney, Adam - Portable Antiquities Scheme	4714

Tokens: Mean = 2.0 Mode = 0, Median = 0

dc:date

Values for 470214 records (47.7%)
Top values:

Value	Freq.
[2008]	26833
[2007]	25199
[2006]	16121
[2005]	15503
[2003]	12020
[2004]	9560
[2009]	8974
01/01/90 00:00	6257
31/12/10 00:00	4444
31/12/36 00:00	4241

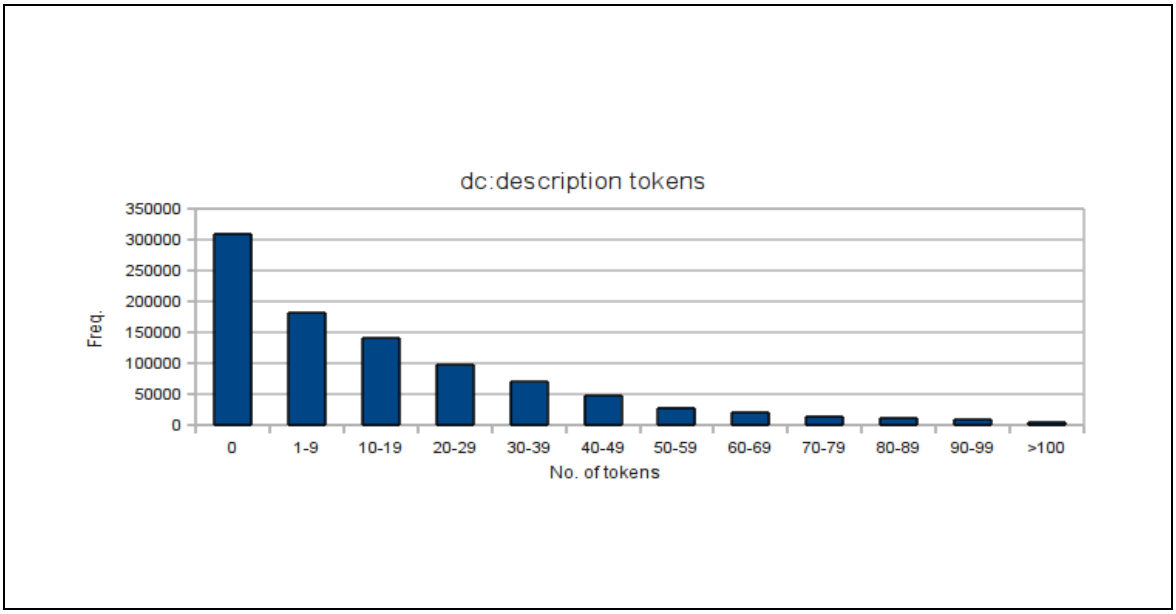
Tokens: mean = 2.4, Mode = 0, Median = 1

dc:description

Values for 674933 records (68.5%)
Top values:

Value	Freq.
Scotland, Edinburgh headquarters of The Scotsman newspaper	14907
Scotland, Glasgow, River Clyde	3609
Single-decker	3511
Scotland, Edinburgh location	2700
World War I taken during	1903
Scotland, City of Edinburgh, Edinburgh locality	1704
Double-decker	1473
Scotland, Dundee location of publication	1469
Scotland locality	1375
Scotland, Angus, Dundee depicted	1328

Tokens: Mean = 26.7, Mode = 0, Median = 10



dc:format

Values for 602710 records (61.2%)
Top values:

Value	Freq.
image/jpeg	271865
text/html	162356
JPEG/IMAGE	112082
application/imgzoom	31396
Image	15788
video/mpeg	5832
audio/mpeg	2226
application/pdf	2212
video/quicktime	400
Christmas card	128

Tokens: Mean = 0.61, Mode = 1, Median = 1

dc:publisher

Values for 298565 records (30.3%)
Top values:

Value	Freq.
The Fitzwilliam Museum, Cambridge, UK	162356
North East Midland Photographic Record	80765
Kirklees	26132
Wolverhampton Archives	7301
Gateshead Council	4144
Wolverhampton Arts and Museums Service	4048
Tyne & Wear Archives & Museums	3916
Dudley Archives	3686
Dudley Museums Service	2371
Walsall Archives / Local History Centre	1627

Tokens: Mean = 1.36, Mode = 0, Median = 0

dc:rights

Values for 545452 records (55.4%)
Top values:

Value	Freq.
Copyright 2005 Portable Antiquities Scheme	125487
The Scotsman Publications Ltd	66028
Reproduced by permission of English Heritage.NMR	23644
National Museums Scotland	23064
Royal Commission on the Ancient and Historical Monuments of Scotland	16427
Corpus Vitrearum Medii Aevi / Courtauld Institute of Art	15043
English Heritage.NMR	14468
Scottish Motor Museum Trust	10722
National Library of Scotland	10315
Newsquest (Herald& Times)	9911

Tokens: Mean = 3.03, Mode = 0, Median = 3

dc:source

Values for 385167 records (39.1%)
Top values:

Value	Freq.
Fitzwilliam Museum	162356
Portable Antiquities	125562
Vads	93105

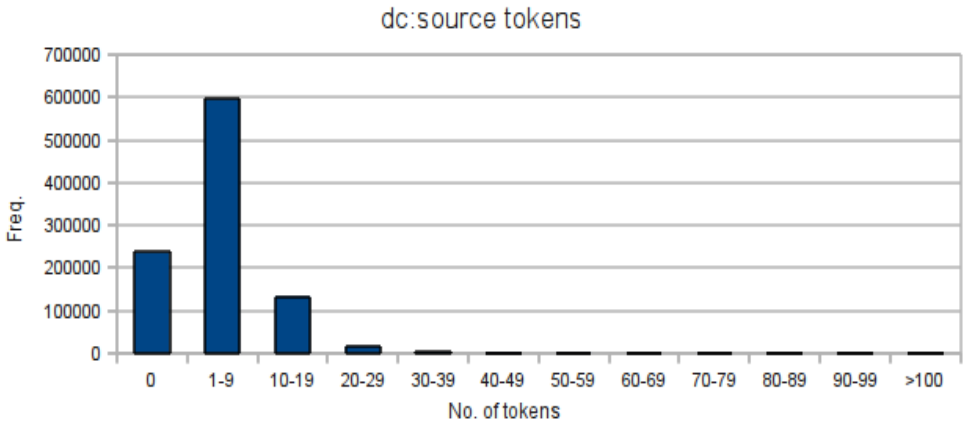
Tokens: Mean = 0.69, Mode = 0, Median = 0

dc:subject

Values for 716078 records (75.6%)
Top values:

Value	Freq.
coin	35530
print	33081
Kirklees	26127
drawing	17666
letter	14971
Unknown	14832
Photograph	14708
letter, handwriting	13367
Transport	10702
Photograph from The Herald newspaper archive	9918

Tokens: mean = 4.7, Mode = 0, Median = 3



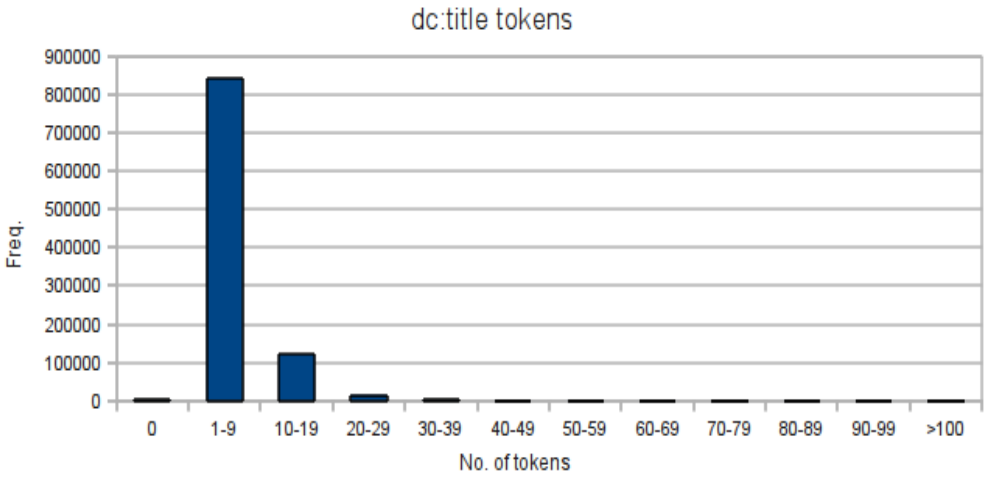
dc:title

Values for 982521 records (99.7%)

Top values:

Value	Freq.
Coin	26818
The Corpus of Romanesque Sculpture in Britain and Ireland	18964
coin	15417
letter	14252
COIN	7874
Corpus Vitrearum Medii Aevi (CVMA)	6992
Design Council Slide Collection	4479
Buckle	4345
Brooch	4298
Vessel	4110

Tokens: Mean=5.6, Mode = 1, Median = 5



dc:type

Values for 474547 records (50.1%)

Top values:

Value	**Freq.**
Image	413538
Physical Object	93105
Unknown	2369
PhysicalObject	1158
Sound	1038
MovingImage	515
Text	465

Tokens: Mean = 0.61, Mode = 0, Median = 1

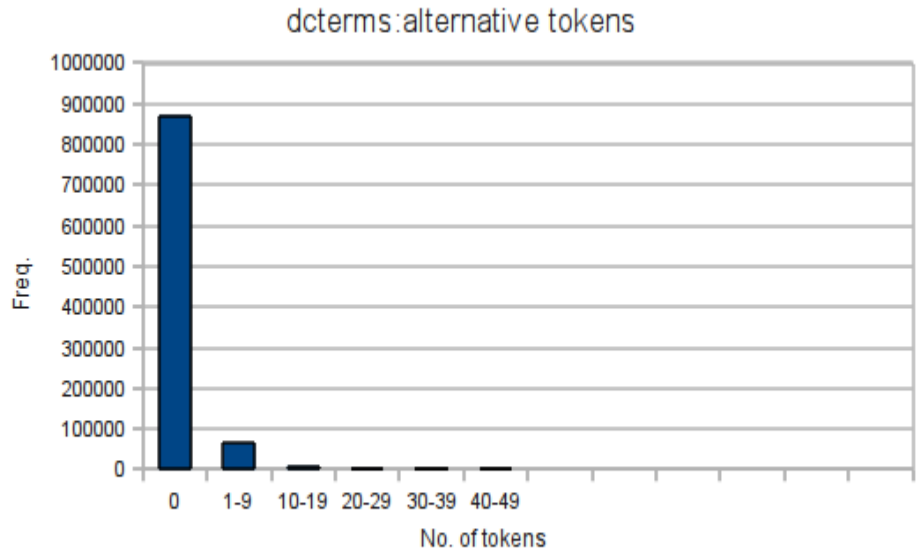
dcterms:alternative

Values for 76146 records (7.7%)

Top values:

Value	Freq.
from James Wotherspoon's 'In The Track Of The Comet'	3608
Jean Jenkins collection	1499
As described in Ordnance Gazetteer of Scotland, 1882-1885, by Francis H. Groome	1339
From the Roy Map, 1747-1755, 1:36,000	1034
a photograph by Thomas Annan of Glasgow (1829-1887)	671
Scanned from the 1853 edition of the 'Scots Musical Museum', James Johnson and Robert Burns (Edinburgh and London: W. Blackwood& Sons, 1853)	600
Printed plate	550
Single-decker	538
from the Hutton Collection	534
minted in Rome	457

Tokens: Mean=0.46, Mode = 0, Median = 0



dcterms:isPartOf

Values for 674544 records (68.5%)

Top values:

Value	Freq.
Fitzwilliam Museum	162356
Portable Antiquities	125562
Vads	93105
Picture the Past	80774
Leodis	55976
English Heritage	50208
Kirklees Image Archive	31326
Black Country History	21152
Durham County Council	15788
Bowes Museum	14780

Tokens: Mean = 1.4, Mode = 2, Median = 2

dcterms:medium

Values for 132935 records (13.5%)
Top values:

Value	Freq.
Paper	20784
Photograph	18682
Print, monochrome	8261
Video	4082
Engraving	3039
Broadside	1874
Glass negative	1741
Paper, ink	1634
Ink on paper	1620
Book	1577

Tokens: Mean = 0.38, Mode = 0, Median = 0

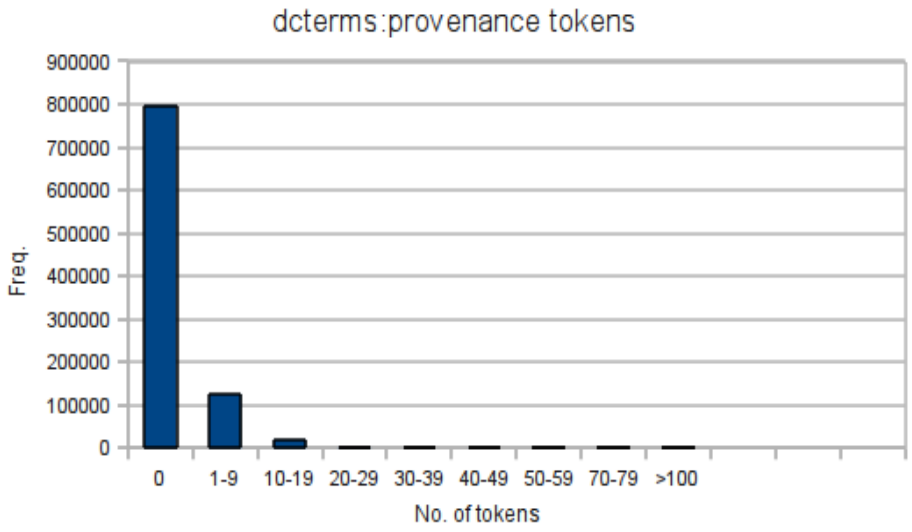
dcterms:provenance

Values for 150308 records (15.3%)

Top values:

Value	Freq.
bequeathed by Reitlinger, Henry Scipio, 1950, Bequeathed by H.S. Reitlinger, 1950	252
bequeathed by Reitlinger, Henry Scipio, 1950, H. S. Reitlinger Bequest 1950	113
bequeathed by Reitlinger, Henry Scipio, 1950, Bequeathed by Henry Scipio Reitlinger, 1950	102
bequeathed by Reitlinger, Henry Scipio, 1950, H.S. Reitlinger Bequest, 1950	10
given by Ricketts and Shannon	6
bequeathed by Reitlinger, Henry Scipio, 1950, H.S. Reitlinger Bequest 1950	6
bequeathed by Reitlinger, Henry Scipio, 1905-05-03, H. S. Reitlinger Bequest 1950	4
bequeathed by Reitlinger, Henry Scipio, 1991-04-29, H.S. Reitlinger Bequest, 1950	2
bequeathed by Harris, Peter, 1976, with a life interest to his widow	2
bequeathed by Leverton Harris Fund, 1926 [3777]	1

Tokens: Mean = 1.2 Mode = 0, Median = 0



dcterms:spatial

Values for 95034 records (9.6%)

Top values:

Value	Freq.
Europe	62410
United Kingdom	62409
World	62408
England	6817
London	5713
Greater London	5700
Lambeth	5700
Darlington	3128
Durham City	2190
World, Europe, United Kingdom, England, Tyne and Wear, Gateshead, Gateshead	2174

Tokens: Mean = 0.39, Mode = 0, Median = 0

dcterms:temporal

Values for 76294 records (7.7%)
Top values:

Value	Freq.
2001 to date	6114
1901 - 1925	2503
1951 - 1975	2002
1976 - 2000	1607
name=Edwardian; start=1901; end=1910;	1391
1926 - 1950	1384
1951-01-01T00:00:00/1970-12-31T23:59:00	1359
1926-01-01T00:00:00/1950-12-31T23:59:00	1141
1901-01-01T00:00:00/1925-12-31T23:59:00	1077
1876 - 1900	730

Tokens: Mean = 0.24, Mode = 0, Median = 0

europeana:provider

Values for 947712 records (100%)
Top values:

Value	**Freq.**
CultureGrid	674544
Scran	310802

europeana:type

Values for 947712 records (100%)
Top values:

Value	**Freq.**
IMAGE	851210
TEXT	91739

VIDEO	39133
SOUND	3264

europeana:year

Values for 222999 records (23.5%)
Top values:

Value	Freq.
2008	26843
2007	25199
2006	16607
2005	16283
2003	12074
2004	9567
2009	8974
1900	6560
1920	3958
1951	3851

Summary of Spanish records

Total records = 1254411

dc:contributor

Values for 231494 records (18.5%).
Top values:

Value	Freq.
Santa Ana, Manuel María de (1820-1894)	50692
Cámara Agrícola de Carrión de los Condes (Palencia)	18605
Partido Liberal (Gerona)	11573
Federación de Sindicatos Católicos-Agrarios (Palencia)	10053
Confederación Regional del Trabajo de Baleares	9599
Sindicato Único de Trabajadores de Mahón	9597
Comité Comarcal de Sindicatos de Menorca	9591
Partido Liberal Dinástico (Tarragona)	8815
Tipografía de Antonio Arqueros (Badajoz)	8199
Tipografía La Minerva Extremeña (Badajoz)	8197

dc:coverage

Values for 35475 records (2.8%)

Top values:

Value	**Freq.**
-19º-20º siglos	10330
-19º-20º séculos	10330
-1887-1938	9970
-1900-1919	4362
-1897-1947	1927
-Galicia	1927
-S.XIX	1371
-1906-1938	1076
-S.XIX-XX	981
-S. XIX	898

dc:creator

Values for 81852 records (6.5%)

Top values:

Value	**Freq.**
Anonymous	3378
Anónimo	2059
Lorente Páramo, Francisco Javier	1772
Sorolla Bastida, Joaquín (Lugar de Nacimiento: Valencia, 27/08/1863 - TL04 Cercedilla, 10/08/1923)	1286
Gadir	757
Pardo Bazán, Emilia Condesa de 1851-1921	655
Atribuido a	430
PLOSSU, Bernard (Lugar de Nacimiento: Dalat, 26/02/1945)	392
SERRANO AGUILAR, Pablo (Lugar de Nacimiento: Crivillén, 1908 - TL04 Madrid (m), 1985)	385
Real Fábrica de Cristales de la Granja	362

dc:date

Values for 1150382 records (91.7%)
Top values:

Value	Freq.
1913	24136
1914	23036
1917	22766
1912	22701
1916	22128
1909	21939
1915	21787
1910	21778
1908	21472
1911	21426

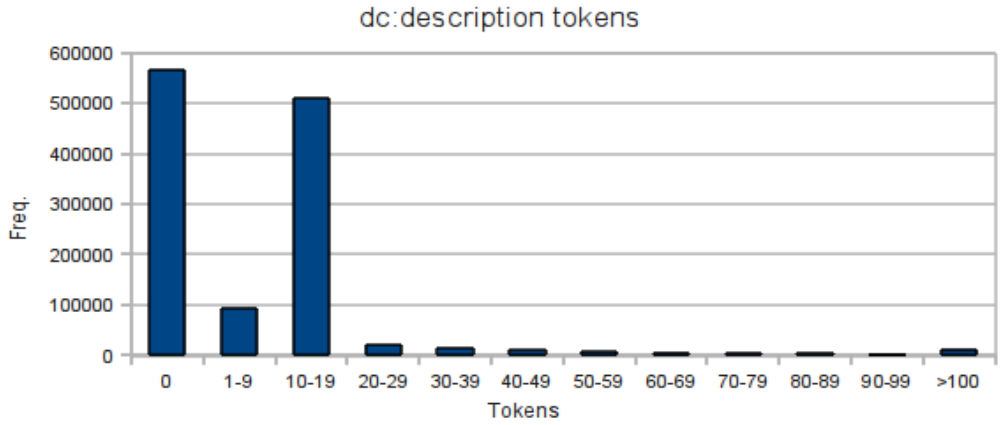
dc:description

Values for 687940 records (54.8%)

Top values:

Value	Freq.
Copia digital. Madrid: Ministerio de Cultura. Subdirección General de Coordinación Bibliotecaria, 2004	331786
Copia digital. Madrid: Ministerio de Cultura. Subdirección General de Coordinación Bibliotecaria, 2003	155921
Copia digital: realizada por la Biblioteca de Andalucía	85408
Copia digital. Valladolid: Junta de Castilla y León. Consejería de Cultura y Turismo. Dirección General de Promociones e Instituciones Culturales, 2009-2010	6883
Copia digital: realizada por la Fundación Provincial de Cultura. Diputación Provincial de Cádiz	2229
Digitalización. Vitoria-Gasteiz: Fundación Sancho el Sabio, 2008	925
Texto a dos col.	917
Port. con orla tip.	714
Copia digital. Madrid: Ministerio de Cultura. Subdirección General de Coordinación Bibliotecaria, 2009	567
Rústica	485

Tokens: Mean=9.7, Mode=0, Median=8



dc:format

Values for 1131916 records (90.2%)

Top values:

Value	**Freq.**
image	939507
application/pdf	807069
image/tiff	129659
image/jpeg	60620
video/flv	121
audio/mp3	96
text/plain	53
application/msword	4
imagen	3
application/epub+zip	1

dc:language

Values for 1252646 records (99.9%)

Top values:

Value	**Freq.**
spa	1226712
cat	19927
lat	2453
fre	1192
eng	764
baq	755
val	349
glg	346
por	220
ita	196

dc:publisher

Values for 282242 records (22.5%)

Top values:

Value	**Freq.**
Ministerio de Cultura	101884
Zuloaga, Hilarión de	50721
Hilarión de Zuloaga	50295
Diputación Provincial de Guadalajara]	16123
Alicante Biblioteca Virtual Miguel de Cervantes	14450
: [s. n.	13333
[s. n.	6172
Asociación de la Prensa	5457
[FET.JONS	5342
Prensa y Radio del Movimiento	5340

dc:rights

Values for 101892 records (8.1%)

Top values:

Value	**Freq.**
Ministerio de Cultura	101884
VEGAP	1426
VEGAP, Visual Entidad de Gestión de Artistas Plásticos	51
Cabré Aguiló, Juan	16
Todos los públicos.	8
Cifuentes, Ramón	2
Pérez Rioja, Aurelio	1
Alviach, Manuel	1
Huerta, M.	1
Cánovas del Castillo Vallejo, Antonio	1

dc:source

Values for 1235133 records (98.5%)

Top values:

Value	**Freq.**
Biblioteca Virtual de Prensa Histórica	848932
Galiciana: Biblioteca Digital de Galicia	129674
Hispana; Spain	101884
Repositorio Biblioteca virtual de Andalucía	98020
Biblioteca Digital de Madrid	22856
CER.ES: Red Digital de Colecciones de museos de España (Museo de América)	17341
CER.ES: Red Digital de Colecciones de museos de España (Museo Nacional de Artes Decorativas)	16823
Biblioteca Virtual del Patrimonio Bibliográfico	9201
CER.ES: Red Digital de Colecciones de museos de España (Museo del Traje. Centro de Investigación del Patrimonio Etnológico)	8387
Biblioteca Digital de Castilla y León	6927

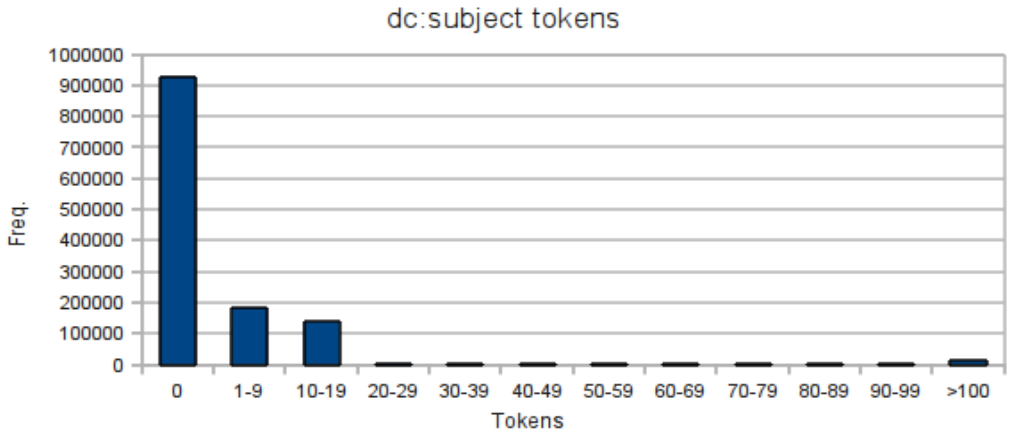
dc:subject

Values for 325598 records (26.0%)

Top values:

Value	Freq.
Secciones - Hemeroteca - Prensa	91537
Catálogos - Publicaciones periódicas - Prensa	91504
Prensa	33658
Prensa gallega	29837
Prensa galega	29287
Cerámica	17808
Edad Contemporánea	14349
-Administración-Publicaciones periódicas	14204
Bronce	9357
Numismática	8722

Tokens: Mean = 1.96, Mode =0, Median = 0



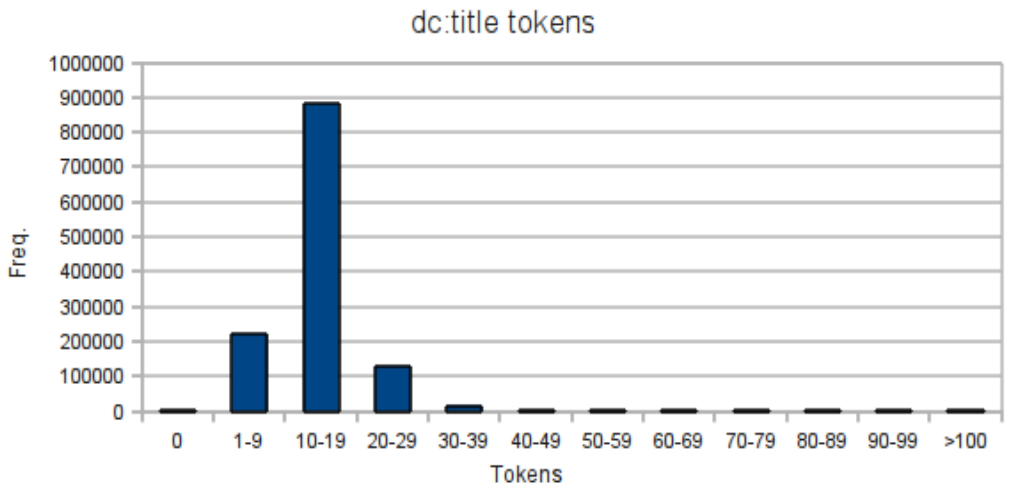
dc:title

Values for 1254410 records (100.0%)

Top values:

Value	**Freq.**
Moneda	7270
Vasija	3206
Plato	2777
Aguja	2164
Fotografía	2074
Cuenco	1722
Vaso	1325
Fusayola	1226
Figura antropomorfa	888
Moeda	888

Tokens: Mean=14.2, Mode = 13, Median =14



dc:type

Values for 124653 records (9.9%)

Top values:

Value	**Freq.**
text	17107
Moneda	7276
Cuadro	6096
Fotografía	3998
Book	3590
Vasija	3206
Estampa	2961
Plato	2814
Escenografía	2474
Aguja	2164

dcterms:created

Values for 85484 records (6.8%)

Top values:

Value	Freq.
1801=1900	2076
1900-1965	1622
1974-1978	1552
1701=1800	1473
1601=1700	1433
100 AC-700 DC	1303
1901=2000	1247
200[ac]=1[ac]	665
1100-1450 DC	659
1970[ca]	594

dcterms:isPartOf

Values for 1206964 records (96.2%)
Top values:

Value	Freq.
En: La Correspondencia de España : diario universal de noticias. - [Madrid] : Hilarión de Zuloaga , 1859-[1925] = ISSN 1137-1188	50072
En: El Defensor de Granada : diario político independiente. - [S.l. : s.n. , 1880-1984]	23541
En: Boletín Oficial de la Provincia de Oviedo . - Oviedo : [s.n.] , 1836-1982 = ISSN 1579-7279	23216
En: Diario de Córdoba de comercio, industria, administración, noticias y avisos . - [S.l. : s.n. , 1849-1938]	22708
En: Crónica Meridional : diario liberal independiente y de intereses generales. - [S.l. : s.n. , 1860-]	19646
Museo de América	17382
Museo Nacional de Artes Decorativas	16873
En: Boletín Oficial de la Provincia de Guadalajara . - [S.l. : Diputación Provincial de Guadalajara] ,	16075
En: El regional : diario de Lugo . - Lugo : [s.n.] , 1884-	15085
En: Boletín de Segovia . - [S.l. : s.n.] , 1833-	14076

Tokens, Mean = 15, Mode = 15, Median = 15

dcterms:issued

Values for 1129989 records (90.1%)

Top values:

Value	Freq.
[s.a.]	733
s.a.]	720
[19--?]	397
[18--?]	297
[1906]	178
[1903]	171
1884	153
1885	148
[S. XVIII]	146
[1917]	142

dcterms:medium

Values for 100858 records (8.0%)

Top values:

Value	**Freq.**
Bronce	9355
Cerámica	8562
Arcilla	8224
Papel	8056
Madera	6420
Lienzo	5037
Plata	4698
Hueso	3994
Pasta cerámica	3529
Loza	3371

dcterms:provenance

Values for 28070 records (2.2%)

Top values:

Value	**Freq.**
Donación	15430
Depósito	5654
Dación	3317
2004	3104
Fábrica Pickman S.A.	3046
Legado fundacional	1416
Legado fundacional de la Vda. de Sorolla, 1931	1070
Asignación: Donación	866
Egner, Günther Adolf	832
03/05/2007	825

dcterms:spatial

Values for 1095231 records (87.3%)

Top values:

Value	**Freq.**
España-Comunidad de Madrid-Madrid-Madrid	70855
España-Andalucía-Granada-Granada	63628
-Galicia	62834
España-Comunidad Valenciana-Alicante-Alicante	49968
España-Andalucía-Córdoba-Córdoba	48398
España-Illes Balears-Illes Balears-Palma de Mallorca	47519
España-Andalucía-Almería-Almería	44929
España-Canarias-Santa Cruz de Tenerife-Santa Cruz de Tenerife	34383
España-Cataluña-Tarragona-Tarragona	32956
España-Cataluña-Girona-Girona	31691

dcterms:temporal

Values for 96249 records (7.7%)

Top values:

Value	**Freq.**
Edad Contemporánea	14349
Edad Moderna	5480
Cultura romana	4620
Andes centrales	4384
1801=1900	2076
Intermedio temprano	2000
Alto Imperio Romano	1738
Andes septentrionales	1678
1900-1965	1622
Alto Imperio Romano. Romano	1591

europeana:hasObject

Values for 1254411 records (100.0%)

Top values:

Value	**Freq.**
true	946515
false	307896

europeana:provider

Values for 1254411 records (100.0%)

Value	**Freq.**
Hispana	1235133
Biblioteca Virtual Miguel de Cervantes	19278

europeana:type

Values for 1254411 records (100.0%)

Value	Freq.
TEXT	1148918
IMAGE	105493

europeana:year

Values for 1150388 records (91.7%)

Top values:

Value	Freq.
1913	24151
1914	23048
1917	22774
1912	22710
1916	22139
1909	21957
1915	21810
1910	21805
1908	21497
1911	21450

Annex 2: Alinari records

Currently we have a set of 11628 records provided by Alinari. This contains the following fields:

Image ID
Caption
Date of photography
Place of photography
Detailed place of photography
Photographer
Object
Technique
Date of artwork
Artist
Period and style
Artwork support
Location
Events
People
Credit
Keywords
Permission and Restrictions
Labels
Format (b/w - col)
Orientation (portrait or landscape)

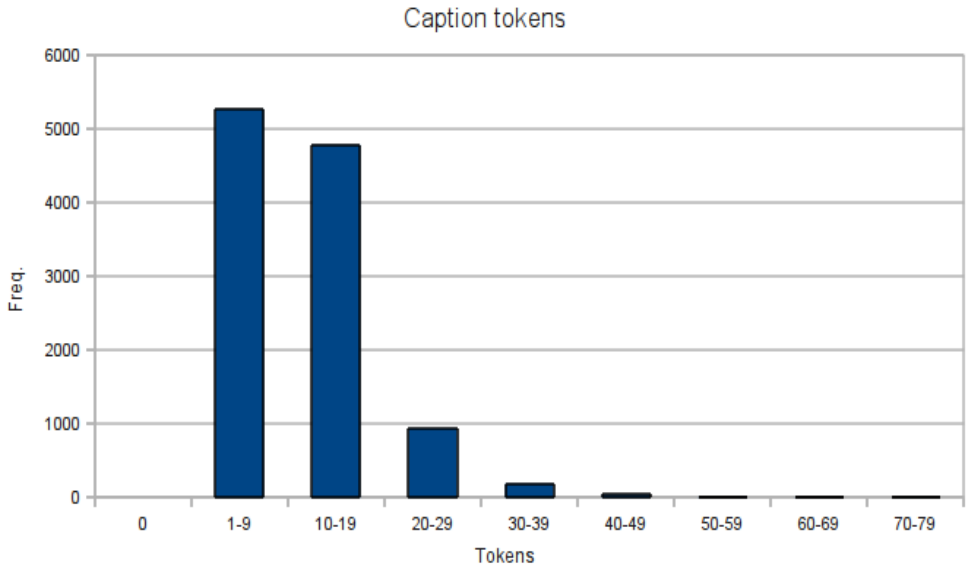
Below is a list of some of the fields and summary statistics of occurring values for each field.

Caption

Values for 11235 records (99.7%)
Top values:

Value	Freq.
The Pine trees of Rome	23
Construction of the Transiberian railway around Lake Baikal	14
View of Genoa	9
Electric poles of the S.A.E., Società Anonima Elettrificazione of Milan, on the Milano-Domodossola line	8
A construction phase of the Cignana Dam, work of the Umberto Girola building firm	8
View of the city of Perugia	8
A part of the historic parade on the occasion of the ""Regata delle antiche Repubbliche Marinare"" held in Amalfi in 1957	8
Water consortium of Agro Monfalconese: work on the water pump on the Isonzo river in Sagrado	7
Genoa's port	6
Hunting on horseback with dogs	5

Tokens: Mean=11.45, Median = 10.0, Mode=7



Date of photography

Values for 11267 records (100.0%)
Top values:

Value	Freq.
1890 ca.	993
1900 ca.	832
1920-1930 ca.	511
1915-1920 ca.	341
1920 - 1930 ca.	339
1880 ca.	331
1890 - 1910 ca.	242
1910 ca.	207
1930 ca.	197
1870 ca.	160

Place of photography

Values for 11058 records (98.1%)
Top values:

Value	Freq.
France	567
Rome	466
Naples	347
Great Britain	303
Florence	279
Greece	264
Switzerland	259
United States	216
Italy	206
Africa	191

Detailed place of photography

Values for 6612 tokens (58.7%)
Top values:

Value	Freq.
Environs	184
environs	142
Paris	98
Somalia	53
New York	52
Lake Como	47
Athens	43
Via Appia Antica	35
Vallombrosa	34
Corfù	34

Photographer

Values for 11256 tokens (99.9%)
Top values:

Value	Freq.
Unidentified Author	3803
Alinari, Fratelli	2510
Brogi	1478
P.z	358
Ojetti, Ugo	217
Graham, James	87
Mauri, Achille	73
Fürst Lwoff, Eugen	71
Sommer, Giorgio	64
Baldinetti, Modestino	63

Object

Values for 11267 records (100%)
Top values:

Value	Freq.
Print on double-weight paper	6256
Glass plate	4052
Photomechanic print	713
Carte de visite	151
Stereoscopic Photography	43
Panoramic	24
Cabinet mount	7

Photomechanic matrix	6
Engraving from a daguerreotype	5
Postcard	4

Technique

Values for 9938 tokens (88.2%)
Top values:

Value	Freq.
N Silver salt gelatin	3541
P Albumen	2883
P Silver salt gelatin	1762
P Collodio-chloride paper	431
P Chromolithography	380
P Collotype	211
P Phototype	161
P Photomechanics	156
P Aristotype	143
P Albumenised salt paper	83

Date of artwork

Values for 2087 rcords (18.5%)
Top values:

Value	Freq.
XIX sec.	42
XVIII sec.	28
XIV sec.	28
XVI sec.	27
XV sec.	26
XII sec.	23
XIII sec.	22
1837-1846	18
I sec.	17
II sec.	16

Artist

Values for 690 records (6.1%)
Top values:

Value	Freq.
Ammannati Bartolomeo (1511-1592)	19
Canzio, Michele	18
Junker, Carl	15
Gizdulich, Riccardo	11
Fanzago, Cosimo	11
Kallikrates	11
Brizzi, Emilio	11
Fidia	10
Vignola	10
Roebling, John Augustus	10

Period and style

Values for 2051 records (18.2%)
Top values:

Value	Freq.
Europe	1910
First and Second Millennium A.D.	1458
Renaissance-Baroque styles and periods	583
Ancient Civilization	544
Middle Ages	405
Modern European Styles and Movements	362
Roman Art	358
Renaissance	299
Gothic	154
Imperial Period	153

Artwork support

Values for 2078 records (18.4%)
Top values:

Value	Freq.
Stationary Modern Work	1297
Stationary Archaeological Work	517
Castle	255
Movable Modern Work	229
Building	179
Church	176
Temple	130
Bridge	125
Villa	109
Painting	66

Location

Top values:

Value	Freq.
Rome,	68
Florence,	60
Venice,	36
Pompeii,	36
Greece, Athens	31
Naples,	31
Turin,	30
Great Britain,	27
France, Paris	23
Bellagio,	23

Events

Values for 533 records (4.7%)
Top values:

Value	Freq.
World War I, 1914-1918	302
The Paris Commune, March-May 1871	32
Eruption of Vesuvius, April 1906	24
World War II, 1939 - 1945	21
Universal Exhibition of Paris, 1900	18
Spanish Civil War, 1936-1939	18
Earthquake of Diano Marina e Diano Castello, 23 February 1887	16
Official visit of Adolf Hitler to Italy, 3-9 May 1938	14
Earthquake of Messina (Sicilian and Calabrian Earthquake), 1908	13
Flood in Verona, 22 september 1882	12

People

Values for 52 records (0.5%)
Top values:

Value	Freq.
Lorena Karl Stephan of	10
Ojetti Ugo	5
Antinori Piero	3
Victor Emanuel III of Savoy	3
Barrès Maurice	3
Hitler Adolf	3
Vittorio Emanuele Prince of Naples	2
D'Annunzio Gabriele	2
Gallone Soava	2
Maria of Savoy	2

Credit

Values for 11268 records (100%)

Value	Freq.
Fratelli Alinari Museum Collections, Florence	7162
Alinari Archives-Alinari Archive, Florence	2537
Alinari Archives-Brogi Archive, Florence	1517
Gabba Raccolta Acquisto Fratelli Alinari Museum Collections, Florence	20
Aragozzini Vincenzo Archive Fratelli Alinari Museum Collections, Florence	12
Verchi Marialieta Collection Fratelli Alinari Museum Collections, Florence	6
Abetti Fondo Acquisto Fratelli Alinari Museum Collections, Florence	5
Reteuna Collection Fratelli Alinari Museum Collections, Florence	5
Ojetti Archive Fratelli Alinari Museum Collections, Florence	1
Pasta Fernando Archivio Fratelli Alinari Museum Collections, Florence	1

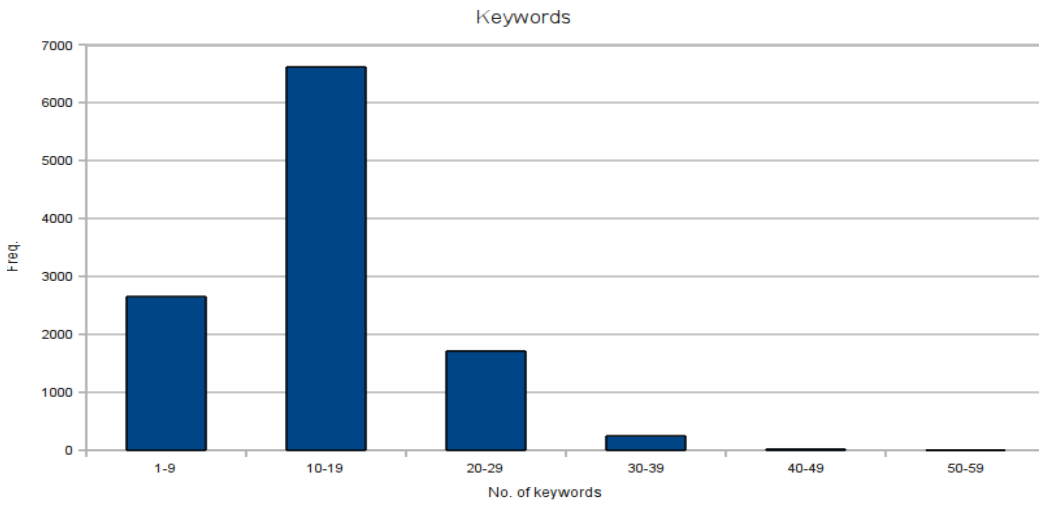
Keywords

Values for 11268 records (100.0%)
Top values:

Value	Freq.
Landscape	11249
Architecture	7660
Structure	7323
Vegetation	3687
View	3677
Housing	3517
Tree	3093
Building Elements,Decoration and Sections	2878
Transportation	2283
Landscape (Seascape)	2191

Note: The meta data contains a list of distinct keywords for each item. Therefore it makes more sense to consider number of keywords rather than number of tokens:

Keywords: Mode = 12, Median = 13, Mean = 14.1



Labels

Values for 11268 records (100.0%)

Value	Freq.
WARNING: Permission must be required for non editorial use. Please contact Alinari Archives;	11233
Copyright: ARTIST'S COPYRIGHT MUST ALSO BE CLEARED; WARNING: Permission must be required for non editorial use. Please contact Alinari Archives;	30

Format (b/w - col)

Values for 11268 records (100.0%)
Top values:

Value	Freq.
B/W	10781
Col	487

Orientation (portrait or landscape)

Values for 11268 records (100%)
Top values:

Value	Freq.
L	9796
P	1472