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

Report accompanying D2.2: Processing and Representation of Content for Second Prototype

Authors: Arantxa Otegi (UPV/EHU)
Eneko Agirre (UPV/EHU)
Aitor Soroa (UPV/EHU)

Contributors: Nikos Aletras (USFD)
Constantinos Chandrinos (i-Sieve)
Samuel Fernando (USFD)
Aitor Gonzalez-Agirre (UPV/EHU)

Project funded under FP7-ICT-2009-6 Challenge 4 – "Digital Libraries and Content"	
Status	Final
Distribution level	Public
Date of delivery	04/03/2013
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 2.2 “Processing and Representation of Content for Second Prototype”. The deliverable comprises the data produced by WP2 “Content Processing and Enrichment” which is to be used in the second prototype. The data has been released in DVDs and is also available from the subversion server of the project. The data comprises three collections from Europeana: the Culture Grid collection from the UK and the Hispana and Cervantes collections from Spain. The items have been enriched with intra collection links, background links, informativeness scores, normalized dates, event information and sentiment information at item level. This deliverable also includes the outcomes of the intrinsic evaluation performed on the enriched data.

Change Log

Version	Date	Amended by	Changes
v0.1	2012/10/19	Eneko Agirre	First draft with placeholders
v0.2	2012/10/20	Nikos Aletras	Section 6 on similarity
v0.3	2012/11/21	Eneko Agirre	Sections 1, 4, 7, 9, 11
v0.4	2012/11/23	Arantxa Otegi	Sections 2, 4, 10
v0.5	2012/11/23	Aitor Soroa	Sections 3, 5
v0.6	2012/11/26	Aitor Gonzalez-Agirre	Section 6
v0.7	2012/11/27	Eneko Agirre, Aitor Soroa, Arantxa Otegi	First complete draft
v0.8	2012/11/29	Samuel Fernando	Section 5 and minor edits
v0.9	2012/12/03	Konstantinos Chandrinos	Section 3 and 4
v0.10	2012/12/03	Arantxa Otegi	Overall review and minor edits
v0.11	2013/03/04	Eneko Agirre	Section 8

Table of Contents

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.

This second release of the data includes three collections from Europeana: the Culture Grid collection 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 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 at the time being. Additional data will be produced on time to be incorporated in the second prototype.

The contents in deliverable D2.2 are to be used in the second prototype, due month 28. The contents are based partially on the feedback received from WP4 and WP5.

The contents of D2.2 are available from
http://development.paths-project.eu:81/svn/paths/trunk/data/D2.2 or in a DVD on request.

An interface to access the items and part of the enrichment data is available as an aid in the development of ideas in the following urls:
http://ixa2.si.ehu.es/europeanaEN/search
http://ixa2.si.ehu.es/europeanaES/search

1 Introduction

This report accompanies D2.2, which contains the second release of the enriched contents to be used in the second prototype.

The differences with respect to D2.1 are the following, organized according to their respective task and Section.

Regarding task 2.1, Content Collection and Representation Section of this report describes the novelties with respect to D2.1:

- SCRAN was removed from Europeana, and we thus discard it for the second prototype.
- Alinari withdrew from the project, and at the time being the use of their content is being discussed. If included, their data will be enriched and included in the second prototype.
- ESEpaths has been extended to include informativeness scores, normalized dates from <dc:date> field, vocabulary terms for tag clouds, event information, sentiment information at item level, typed related items, title and sentiment of background links.
- For easier ingestion and production, the data will be separated in different ESEpaths. A single item will have several ESE and ESEpaths files, which hold complementary information.
- A proposal for EDMpaths.
- ESEpaths is being considered by the ISO committee.

Regarding task 2.2, Content Analysis Section describes the new enrichment information added in D2.2:

- informativeness scores,
- normalized dates from <dc:date> field,
- vocabulary terms for tag clouds,
- event information,
- sentiment information at item level,
- typed related items,
- title of background links,
- sentiment of background links
- in addition, we also gathered background links from <dc:creator> and <dcterms:spatial>.

Task 2.3, Ontology Extension, is presented in Section 5, describing the analysis of the use of vocabulary terms in the first prototype, and the current efforts to select a vocabulary which fills the requirements of the second prototype. We have evaluated several vocabularies, and will soon select one and apply it to all PATHS collections.

Task 2.4, Intra-collection links, is presented in Section 6, describing the efforts in evaluation of the intra-collection links added in D2.1, and the current efforts to define and deploy typed-similarity, which will be included in the second prototype.

Task 2.5, Background Links, Section 7, reviews current efforts to add background links to educational resources, external CH collections, blogs, news, tweets linked to CH resources. The new background links will cover frequent updates and dynamic links created on the fly based on user profile.

In addition, this report includes three other sections. Section 9 gathers some recommendations from the PATHS perspective for data provides. Section 10 describes the design of a web service which ingests EDM and produces EDMpaths. Section 11 mentions an internal interface which can be used to browse the enriched items.

2 Content of the Data Deliverable

The data in the deliverable comes from the main data source in PATHS project, which is Europeana collection. The main folder is the following:

- **D2.2_europeana**: Comprises the content from Europeana.

The data in this main folder is divided according to the subcollections in the main Europeana collection, as shown here:

- **09405_Ag_UK_ELocal**: Corresponds to one of the three collections of English CULTURE GRID collection.
- **09405a_Ag_UK_ELocal**: Corresponds to one of the three collections of English CULTURE GRID collection.
- **09405b_Ag_UK_ELocal**: Corresponds to one of the three collections of English CULTURE GRID collection.
- **09407_Ag_ES_ELocal**: Corresponds to the Spanish HISPANA collection.
- **90901_L_ES_BibVirtualCervantes_dc**: Corresponds to the Spanish CERVANTES collection.

Each of these folder comprise the same files, which are:

- **ESE.bz2**: The source data represented in the Europeana Semantic Elements (ESE) specification format.
- **ESEpaths.background.other.bz2**: File containing background links to web resources that refer to each of the items in the collection.
- **ESEpaths.background.wiki.bz2**: File containing background links to Wikipedia articles for each of the items in the collection.
- **ESEpaths.events.bz2**: File containing events which are found within each of the items in the collection.
- **ESEpaths.facets.date.bz2**: File containing normalized dates from <dc:date> field for each of the items in the collection.
- **ESEpaths.informativeness.bz2**: File containing informativeness of each of the items in the collection.
- **ESEpaths.intra.bz2**: File containing links to similar items for each of the items in the collection.
- **ESEpaths.sentiments.bz2**: File containing sentiment towards items in the collection.

All the ESEpaths files mentioned above are an extension of ESE. Instead of having all the enrichment information just in one large ESEpaths file (as done for D2.1), the output of each of the content process is stored in a separate ESEpaths file. The contents of each of these files are described in the sections below.

3 Content collection and representation

This second release of the data includes three collections from Europeana: the Culture Grid collection from the UK and the Hispana and Cervantes collections from Spain. Two collections have been removed with respect to D2.1:

- SCRAN was removed from Europeana recently, due to disagreements with the new licensing policy. We thus cannot include it in the second prototype.
- Alinari withdrew from the project, and at the time being the use of their content is being discussed. If an agreement is reached, we will release their enriched metadata asap.

We kept using ESE instead of the newer Europeana Data Model (EDM). The main reason is that most of the providers are not sending their datasets as EDM to Europeana yet, and at the time of preparing the second collection there was no guarantee that they will be available on time (Antoine Isaac, personal communication). In any case that most of the data in Europeana which was in EDM, had been automatically converted from ESE, so there is no functionality or data loss from the perspective of the Second Prototype. In any case, we are aware of the importance of showing that PATHS can work seamlessly with datasets coded in EDM, so we include the following:

- The design of EDMpaths, the extension of EDM which represents all the enrichment done to ESE in ESEpaths
- A web service which ingests EDM and produces EDMpaths has been designed, and will be provided on month 28 (cf. Section 10).

In this section we describe the differences with respect to D2.1:

- ESEpaths has been extended to include informativeness scores, normalized dates from <dc:date> field, vocabulary terms for tag clouds, event information, sentiment information at item level, typed related items, title and sentiment of background links.
- For easier ingestion and production, the data will be separated in different ESEpaths, that is, a single item will have several ESE and ESEpaths files, which hold complementary information.
- The proposal for EDMpaths.
- ESEpaths is being considered by the ISO committee.

In addition, we include a section (cf. Section 9) with recommendations to CH organizations regarding metadata production, with the aim of allowing the PATHS technology to provide more effective data enrichment and thus allow better performance of the prototype.

3.1 Changes to ESEpaths

ESEpaths has been extended to include all the relevant information produced after the content enrichment process. Besides the information described in Deliverable D2.1, ESEpaths now includes fields for:

- Informativeness score (<paths:informativeness>): each item has associated a value indicating the overall "informativeness" of the item, which is related to the amount of text and inversely proportional to the number of items where the text is mentioned. See Section 4 for the more details on the method to compute informativeness.
- Normalized date (<paths:normalized_date>): as the date strings in ESE <dc:date> elements have many formats, we store the normalized data in this element. Only dates before 1990 are considered.
- Vocabulary terms (<paths:vocabulary>): vocabulary terms associated to the item. These terms are used for creating the tag clouds shown to the user. The <paths:vocabulary> element has the following attributes:
    - name: name of the external vocabulary.
    - URI: the address (URI) of the specific category in the vocabulary.
    - confidence: the confidence of the association.
- Event information (<paths:events>): event information associated with the item. The element describes the eventual word form as present in the original field. It has the following attributes:
    - source: the name of the external resource of the event (for instance, WordNet).
    - canonnical_form: the canonical word form of the annotated event.
    - confidence: confidence of the association.
- Sentiment information at item level (<paths:item_sentiment>): sentiment towards the item.
- Typed related items. The <paths:related_item> element has been augmented with these attributes:
    - type: the type of the relation (for instance, same_author, same_location, etc).
- Title and sentiment of background links. The <paths:background_links> element has been augmented with these attributes:
    - title: title of the URL which the background link points to. This title is the one to be shown in the UI.
    - sentiment: polarity of the textual information included in the corresponding link. It has fixed values, namely "pos" for positive results, "neg" for negative and "neu" for neutral.

The XML schema for the ESEpaths is shown in the Annex 1 of this report.

3.2 Proposal for EDMpaths

The Europeana Data Model (EDM) is a new proposal for structuring the information within Europeana. EDM departs from ESE as it offers an open, cross-domain Semantic Web based framework, and tries to accommodate existing standards such as LIDO for museums, EAD for archives, etc.

In PATHS project we are currently exploring the ways to adopt the information represented in ESEPaths to the EDM framework, a new representation we call EDMpaths. Our initial idea was to completely switch to EDM for the second prototype: ingest the input records in EDM and produce EDMpaths as result of the content enrichment process. However, Europeana is not yet receiving new data as EDM for the providers (personal communication). There are ways to automatically convert ESE records to EDM, but the data is not “semantically richer” than the ESE data. For these reasons, we decided to stick to ESE (and, therefore, ESEpaths) within PATHS.

In any case, we still plan to define the EDMPaths format and we will also use it as basis representation for the web service described in Section 10.

At the time being, we have devised the design principles of EDMpaths, and we are currently discussing this design with the Europeana staff.

The main design principles for EDMpaths are:

- Create a new paths "ore:Aggregation", which "ore:aggregates" the original Culture Grid aggregation (much like “edm:EuropeanaAggregation” is built upon an existing aggregation).
- Create a “ore:Proxy” for the paths aggregation and describe all the information produced in paths in this proxy.
- Because many of the ESEPaths elements have attributes, reify these relations and attach all the information to the newly reified concept.

3.3 Possible options for standardisation of ESEPaths

The concept of Paths as a means of navigation through an information space is one that has been explored in different ways. D1.2 includes an extensive survey of such activities. With this in mind, the PATHS partners believe that there is a real opportunity for other memory institutions to benefit from the current work and vice versa - for PATHS to benefit from related work by others - through standardisation.

To begin this process, in the near future the project will prepare the ESEPaths XML schema for publication, including full documentation in the style of a standard specification. These will be hosted in an environment where either the machine readable schema or the human readable documentation will be returned from the namespace URI, depending on the user agent. As an example, the SKOS-XL schema at http://www.w3.org/TR/skos-reference/skos-xl returns either the RDF schema or the HTML documentation for it depending whether the URI is de-referenced with a browser or software wishing to access the schema itself.

Once published in this way, ESEPaths will be publicised through a number of channels through which a community can be brought together. One such channel is available at the World Wide Web Consortium, W3C, where anyone can propose a Community Group in which possible future standardisation work can be discussed. The PATHS partners would actively seek the involvement of related projects in Europe and elsewhere, as well as Europeana staff.

Community Groups can publish their work on the W3C website but an alternative approach might be for Vrije Universiteit, a W3C Member organisation, to publish the ESEPaths specification as a Member Submission. This triggers a W3C Team review and brings more attention to a draft specification than a Community Group is normally able to achieve.

As with any standardisation effort, the key factor is community interest and implementation experience. Should the community so with, then ESEPaths will be in the right place to be the basis of formal standardisation at W3C.

4 Content Analysis

In this section we explain the new enrichment information added with respect to D2.1. The enrichment includes the following:

- informativeness scores,
- normalized dates from <dc:date> field,
- vocabulary terms for tag clouds,
- event information,
- sentiment information at item level,
- typed related items,
- title of background links,
- sentiment of background links

In addition, we also gathered background links from <dc:creator> and <dcterms:spatial>.

4.1 Informativeness scores

In D2.1 we described several methods to select informative items, defining two subsets. For the second prototype it was agreed that it would be more practical to have a measure of informativeness, in order to prioritize more informative items in the interface.

The informativeness of items was calculated as follows:

$$\begin{aligned} \text{informativeness(item)} = & \ \text{length(item title)/avgL title} * \log(N / \text{count(item title)}) \\ & + \text{length(item desc)/avgL desc} * \log(N / \text{count(item desc)}) \\ & + \text{length(item subj)/avgL subj} * \log(N / \text{count(item subj)}) \end{aligned}$$

where title, desc, subj refer to the title, description and subject fields of the metadata, avgL X is the average length of field X over the whole collection, length(item X) is the length of that item field text, and count(item X) gives the frequency of that item field text over the whole collection. The higher the resulting value the more informative the item is. Note that as well as taking into account the length of the fields, this also weights by the inverse document frequency (idf) value, so very frequently occurring terms will be down weighted.

4.2 Normalized dates from <dc:date> field

As the date strings in ESE <dc:date> elements have many formats, we enrich the items with the normalized data. For this purpose we use the Freeling toolkit.

As Freeling recognizes only some date patterns, we first apply some simple regular expressions to the text in <dc:date> field, in order to make this text follow the patterns recognized by Freeling. For example, after applying some regular expressions, the texts “1945” and “01.01.1965” are rewritten, respectively, as “year 1945” and “01/01/1965”, as “year YYYY” and “DD/MM/YYYY” are some of the patterns that Freeling could recognize. After this reformulations, Freeling is applied and the normalized output is used to enrich the item. For the above examples, the normalized outputs will be [??:??/??/1945:??.??:??] and [??:1/1/1965:??.??:??], respectively.

As the dates in this field is referring to the cataloguing date for some of the items, only dates before 1990 are considered, in order to discard the cataloguing dates.

4.3 Vocabulary terms for tag clouds

This information will be included when the ontology extension and mapping is finished (cf. Section 5).

4.4 Event information

In order to discover description text that refers to events, we mined WordNet and the semantic relation there, to produce a list of words that can be used to refer to events. The list includes lemma-PoS pairs, and a canonical way to refer to those events. For instance, walk-V is the canonical event for pass-N, pedestrian-N, walker-N, walkway-N, walking-N, among others.

As an example, an item describing a picture at “Abbey Walk” won’t be enriched with the event walk-V, but an item with the text “Cars are parked along the street and pedestrians can be seen.” will be enriched with the event walk-V.

4.5 Sentiment information at item level

Attributing sentiment at item level has proven a challenging task for the domain of cultural heritage. The two main obstacles are the scarcity of polarity in the sentiment and subtle expression of such polarity, when present.

On one hand, references to cultural objects are mostly of an instructive nature and are thus written in a matter-of-fact language, which will qualify them as neutral. On the other hand when opinion is expressed, it is very subtle.

In commercial applications of sentiment analysis, when tracking sentiment around brands, the identification of polarity is fairly easy. People tend to use strong words, either positive or negative, and these words appear syntactically in the vicinity of the brand name. Thus “shallow parsing” usually suffices.

Our tests in defining sentiment towards items indicate that this is not the case with CH objects and therefore we currently investigate techniques with deeper structural analysis.

4.6 Typed related items

This information will be included when the typed-relations are computed (cf. Section 6).

4.7 Title of background links

The titles for all background links in D2.1 have been added. Newly produced background links will include the title as well.

4.8 Sentiment of background links

The sentiment for all background links in D2.1 is being added. Newly produced background links will include sentiment as well, with an effort to narrow the sentiment detection to the exact CH object.

4.9 Background links from <dc:creator> and <dcterms:spatial>

The <dc:creator> and <dcterms:spatial> fields have been analysed with WikiMiner and linked to corresponding Wikipedia articles. Note that the geospatial links can be used to further enrich the items with geospatial coordinates.

5 Ontology Extension

With increasingly large sets of diverse collections of documents available online a key challenge is organising and presenting these items effectively for information access. To enable the navigation and exploration of collections, content providers typically provide users with free-text search functionalities, along with some form of browsable subject categories or taxonomy, also useful in organising documents. Providing multiple mechanisms for accessing documents enables users to conduct various modes of information seeking activity, from locating specific documents to more exploratory forms of searching and browsing behaviour.

In this task we focused on the proposal and evaluating different taxonomies that could be used to organise and navigate the content in Europeana. Note that Europeana comprises many subcollections taken from different providers, and thus contains a very diverse set of cultural heritage items. This therefore represents a very challenging dataset to organise in a consistent and uniform manner.

The collections in Europeana usually have some hierarchical vocabularies for classifying items. The terms from the vocabulary are used to classify items, including them in “subject” fields. Unfortunately, it is well known that each collection has its own, sometimes proprietary vocabulary, and that mapping vocabularies is an open problem. Having one vocabulary per collection is not suitable for the requirements in WP4.

From WP4 we would need a vocabulary meeting the following requirements:

- It needs to apply to most items
- It needs to be useful for browsing

The first prototype included hierarchical vocabularies[1] from the Cultural Heritage and Digital Libraries domain, focusing on the Library of Congress Subject Headings (LCSH) and the English Heritage-NMR (NMR). The low coverage and practical concerns about how useful the hierarchies were for browsing made us look for other models to organize hierarchically the items in our collections. We examined recent work and proposals, including the following:

- STITCH project[2] on semantic inter-operability, which focuses on ontology mapping and common RDF-style representations.
- The DISMARC/eConnect aggregation platform[3], which provides co-operative elaboration of multilingual vocabularies, and the subsequent Open-up project[4].

[1] Note: we use vocabulary, thesaurus, taxonomy and ontology interchangeably, as our focus is on having a hierarchical structure which organizes the items.
[2] http://www.cs.vu.nl/STITCH/
[3] http://open-up.eu/sites/open-up.eu/files/u22/AT4DL_Koch_Scholz.pdf

Mapping vocabularies automatically would be out of the scope of PATHS, and already pursued in those other projects. Our approach is complementary, as we seek to provide a vocabulary which can be used to organize all items in Europeana. In WP4 WordNet[5] was used as vocabulary, and tested in some experiments with users, with mixed results. Note that the mapping to WordNet was not trivial, and was described in one conference paper[6].

We thus aim at extending and performing more comprehensive experiments with other vocabularies. At the end of the experiments, a vocabulary will be selected and all PATH collection items will be mapped to the vocabulary. The results of the mapping will enrich all items with the corresponding vocabulary terms. Those terms and the vocabulary will be used in the user interface to provide browsing capabilities as hierarchical navigation and tag clouds.

5.1 General approach

We mapped Europeana items to external vocabularies using the approach described in (Fernando et al, 2012), which also evaluates the quality of the considered vocabularies and the mappings. We focus on two main approaches for organising content: the first is to map items from Europeana onto existing manually-created taxonomies; the second is to use data-driven approaches to automatically derive taxonomies from the collection. This requires being able to successfully group items into categories and generate suitable category labels.

5.2 Vocabularies considered

We tested six taxonomies in the experiments. Four of these were based on existing taxonomies which have been mostly manually created: the Library of Congress Subject Headings, WordNet Domains, Wikipedia Taxonomy and the DBpedia ontology. The remaining two taxonomies were automatically derived from the metadata present for the items in the collections: WikiFreq and LDA topics. This section gives a description of each of these taxonomies and how the items in the collection were mapped into them. Statistics for each taxonomy are presented at the end of this section.

Manually created taxonomies

Library of Congress Subject Headings (LCSH)

The LCSH comprises a controlled vocabulary maintained by the United States Library of Congress for use in bibliographic records. They are used in many libraries to organise their collections as well as for organising materials online.

The text from the <dc:subject> field in the Europeana item are used for the mapping. The text is lemmatized using the Freeling toolkit and compared to the category labels for the

[4] http://open-up.eu/

[5] http://wordnet.princeton.edu/

[6] http://www.lrec-conf.org/proceedings/lrec2012/summaries/232.html

LCSH concepts. If the text contains any of the category labels then the item is matched to these categories. If more than one matching label is found, then the longest matching label is used for the mapping.

WordNet domains

WordNet domains comprise a set of labels which has been semi-automatically applied to each of the synsets in WordNet. Each synset is annotated with each one label from a set of about two hundred. The information provided by the domain labels is complementary to the data existing already in WordNet. The domain labels group together words from different syntactic categories (e.g. nouns and verbs), and also may group together different senses of the same word and thus reduce polysemy.

For the mapping process Yago2 is used as an intermediate vocabulary. Yago2 is a knowledge base derived from Wikipedia with more than 10 million entities, and each entity in Yago2 is linked to a WordNet 3.0 synset. We also used a mapping from WordNet 3.0 synsets to WordNet Domain labels as provided by the Multilingual Central Repository (MCR). To perform the mapping, the first step is again to use the <dc:subject> field to link Europeana items to Yago2 entities (using lemmatization and finding the longest possible match). These are then mapped to the WordNet Domain labels via the Yago2 entity-to-synset and the MCR synset-to-WordNetDomain mappings.

Wikipedia Taxonomy

Wikipedia Taxonomy is a taxonomy derived from Wikipedia categories. The authors create the Wikipedia Taxonomy by keeping the is-a relations between Wikipedia categories and discarding the rest. We first get the Wikipedia articles mapped to the item using the background links related to the <dc:subject> field (see Section 7). Then, we link the Europeana item to all Wikipedia Taxonomy categories which are related to these entities.

DBpedia ontology

The DBpedia ontology is a small, shallow ontology manually created based on information derived from Wikipedia. Contrary to the previous vocabularies described above, the DBpedia ontology is a formalised ontology, including inference capabilities. The authors provide the instances of each ontology class, i.e. the set of Wikipedia entities pertaining to this class. For mapping Europeana items to DBpedia ontology classes, we first get the Wikipedia background links for the <dc:subject> field, and then link the item to the classes these entities belong.

Automatically created data-driven taxonomies

LDA topic modelling

Latent Dirichlet Allocation (LDA) is a state-of-the-art topic modelling algorithm that creates a mapping between a set of topics and a set of items, where each item is linked to one or more topics. Each item is input into LDA as a bag-of-words and then represented as a probabilistic mixture of topics. The LDA model consists of a multinomial distribution of items over topics where each topic is itself a multinomial distribution over words. The item-topic and topic-word distributions are learned simultaneously using collapsed Gibbs sampling based on the item - word distributions observed in the source collection. LDA has been used to successfully improve result quality in Information Retrieval tasks and is thus well suited to support exploration in digital libraries.

Wikipedia link frequencies

This is a novel method for taxonomy creation which uses Wikipedia background links as the concept nodes in the taxonomy.

The first step is to run Wikipedia Miner to find links in all Europeana items. Then we find frequency counts for each link. For each item we take the set of links found and create a taxonomy branch (if not already present) with links in order of frequency (most frequent first). The item is then mapped to the least frequent link.

This method is explained in more detail in (Fernando et al., 2012).

5.3 Taxonomy statistics

The following table (Table 1) shows some statistics for each taxonomy:

- The number of items that are mapped into the taxonomy.
- The average number of parents for each item.
- The average depth from the root node to an item.
- The number of top level nodes in the taxonomy.

Taxonomy	Items	Avg. parents	Avg. depth	Top level nodes
LCSH	99,259	1.8	1.97	28,901
DBpedia	178,312	4.2	2	30
Wiki taxonomy	275,379	11.7	1.13	10417
WN domains	308,687	7.1	7.1	6
LDA topics	545,896	1	7.3	9
Wiki freq	66,558	1	3.39	24

Table 1: Statistics for each taxonomy

A problem with some of the manual taxonomies is the very high number of top level nodes, which makes it difficult for users to browse. However there is no obvious way to select suitable top level nodes in these taxonomies. Additionally some of the taxonomies assign items to many parent nodes - this means that the data is repeated across the taxonomy. This is not a problem in itself, but is likely to mean that items may often be assigned to incorrect nodes.

Section 8.1 includes the evaluation of each taxonomy.

Next steps

The work described so far is the first attempt to automatically link Europeana items to a variety of vocabularies, which are both manually and automatically built. We have also conducted an evaluation in order to assess the quality of the vocabulary relations and mappings between those and the Europeana items. In the future, we plan to expand the evaluations to include user studies. A key question is how well these taxonomies assist users when used for browsing large collections, such as Europeana. The aim is to see if there is a correlation between the intrinsic results that were found here with the extrinsic quality judgements when used in real life applications. A promising line of work will be to build on the WikiFreq approach by integrating with the high quality Wikipedia taxonomy knowledge base. The hope is that using this approach will generate highly cohesive units along with a well structured conceptual tree.

After this user evaluation, we will fix a target vocabulary and we will provide the proper mappings to the Europeana items and we will integrate this information in the second prototype.

6 Intra-collection Links

D2.2. contains the same intra-collection links as in D2.1 We have evaluated these links with a manually created dataset, and concluded that a simpler method would also perform well, but given the small difference we decided to keep them (cf. Section 8.2 on evaluation). In addition we are preparing datasets and algorithms for identifying why two items are related, producing an inventory of relevant relations, in a process we dubbed “typed-similarity”. The items enriched with similarity and typed-similarity information will be especially useful for personalization and recommendation in the second prototype.

6.1 Typed-similarity

In addition to the similarity degree between items, we also want to know why two or more items are similar to each other. One of the new features of the second prototype is that the similarity is going to be broken down in types of similarity. PATHS developers have selected seven similarity types: similar author or creator, similar description, similar event or action, similar location, similar people involved, similar subject and similar time period.

The first step is to check with real users whether these types make sense. We manually selected pairs of items from Europeana with the previous similarity types and created a gold standard with them. The gold standard is composed of 175 pairs of items, 25 pairs for each of the similarity types. Of course, a pair can be similar in more than one of these similarity types.

Previously, we had a gold standard with 295 pairs of items. These pairs may not be similar between them. We are going to include these pairs, both related and unrelated, to the gold standard, in order analyze people’s responses.

Now, we plan to use an online survey to ask people about the similarity between two items we are showing. The design of the survey is shown in Figure 1. Note that this survey shown here is a pilot survey, not the final one.

Respondents will be able to select different similarity types for the items shown, one or more, or even none. They also will have a textbox area to write any other similarity type they feel appropriate.

Item A:

Interior of Buddhist chaitya hall, Cave XXVI, Ajanta

Type: Still Image, Photograph
Creator: Gill Robert
Date: 1868-70
Description: Photograph of the interior of the Buddhist chaitya hall, Cave XXVI at Ajanta, taken by Robert Gilll around 1868-7, from the Archaeological Survey of India Collections. The Buddhist cave temples of Ajanta were first excavated into a horse-shoe shaped cliff overlooking the Waghora River in the 2nd and 1st centuries BC. This large chaitya hall belongs to a later group from the 5th century AD. A columned verandah, now almost destroyed, extends across the facade in the courtyard before the cave. The interior of the cave has two rows of fluted columns decorated with bands of floral patterns. At the end of the central nave there is a stupa with a sculpture of a seated Buddha figure.

Item B:

Type: Still Image, Photograph
Creator: Gill Robert
Date: 1868-70
Description: Photograph of the interior of the Buddhist chaitya hall, Cave XIX at Ajanta, taken by Henry Cousens around 1868-70, from the Archaeological Survey of India Collections. Cave 19 is an elaborate rock-cut chaitya hall from the late 5th century. The external facade has a large horseshoe-arched window flanked by figures of yakshas and Buddhas carved in relief. The interior hall has two rows of columns whose capitals are decorated with Buddha figures, flying couples, hermits and musicians. The panels above depict Buddhas surrounded by bands of scrollwork. At the end of the apse there is a Buddha image in a niche carved on the votive stupa.

Why are the items similar? (please do not choose any option if you think they are different)

☐ Similar author/creator ☐ Similar event ☐ Similar people involved ☐ Similar time period
☐ Similar description ☐ Similar location ☐ Similar subject ☐ Other

Figure 1: Design of the survey

Next steps

Given the high interest in the NLP community on similarity and typed-similarity, the dataset annotated in this task will be used in the Semantic Textual Similarity (STS) public evaluation exercise[7] the spring of 2013. This evaluation has been selected as the shared task of *SEM 2013[8], the Joint Conference on Lexical and Computational Semantics. To date, it has attracted the registration of 61 participant teams.

Once we collect the human similarity types, we will start to develop the technology to automatically detect similarity types. The collected ratings will be used for evaluation. The best method will be used to annotate the collections selected for the second prototype.

[7] http://ixa2.si.ehu.es/sts

[8] http://clic2.cimec.unitn.it/starsem2013/

7 Background Links

As mentioned in Section 4 we have added the titles of the background links and added new background links to Wikipedia for <dc:creator> and <dcterms:spatial> fields, using the same software as in D2.1. The links to Wikipedia were used to map items directly to related Wikipedia items, with good results, as presented in a conference paper (Agirre et al. 2012)[9], and also to produce Wikipedia pages enriched with items from Europeana (Hall et al. 2012).

The user studies for the first prototype showed that the interface was showing too many background links, some of which were repeated. Several considerations made us deliver all links and leave the solution at the hands of the interface:

- The number of background links could be limited using a threshold on the weight of the link (available at ESEpaths) or showing a fixed number of links. It's difficult to choose a threshold to eliminate some of the background links in WP2 beforehand (although a threshold of 0.2 was suggested by hand inspection), as the interface might need to show only 5, or 10, or a fixed number.

- The repeated links came from the fact that the target term (e.g. London) occurred several times in different fields, or inside the same field at different offsets. All that information is kept in ESEpaths, so the interface can generate hyperlinks in the respective field, so the user can click directly while reading the item information.

WP2 will thus serve all links with weight information, and it will be up to the back-end or interface to choose how many and which links to show.

In addition I-Sieve is working on the following:

- Background links to educational resources, external CH collections, blogs, news, tweets linked to CH resources.
- Frequent updates, dynamic links created on the fly based on user profile.

Section 8.3 reviews the evaluation of the background links to Wikipedia.

[9] http://www.lrec-conf.org/proceedings/lrec2012/summaries/1021.html

8 Evaluation

In this Section we review the quality of the enriched information in turn.

8.1 Ontology extension

We begin by evaluating the ontology extension techniques (Section 5). Two evaluations were performed on the taxonomies. The first measured the cohesion of the item clustering, and the second gathered human judgements of the relations that were found between child-parent concept pairs in the taxonomy. For both evaluations online surveys were created using an in-house crowdsourcing interface. Links to the surveys were sent out to a mailing list comprising staff and students at a large university.

Cohesion

A cohesive cluster is defined as one in which the items are similar while at the same time clearly distinguishable from items in other clusters. To measure the cohesiveness of the taxonomies we use the intruder detection task. The idea of this is to present 5 items to an evaluator. Four of these are taken from one concept node in the taxonomy and the other (the intruder) is randomly picked from elsewhere in the taxonomy. The more cohesive the concept in the taxonomy the more obvious it should be which is the intruder item.

To generate good quality units for the evaluation the informativeness of items (see Section 4) was calculated and used.

Altogether 134 people attempted the survey. 23 of the users evaluated at least one control unit wrong or evaluated less than 5 units in total, and so their answers were excluded. The remaining 111 participants contributed in total 1255 answers. Each unit received a minimum of 5 answers and an average of 6.97 answers. A unit was judged as cohesive if more than 80% of the annotators agreed on the same intruder.

Taxonomy	Cohesive units	Percentage
LCSH	19	63.3%
DBpedia	17	56.7%
Wiki taxonomy	18	60%
WN domains	15	50%
LDA topics	17	56.7%
Wiki freq	29	96.7%

Table 2: Some results of the cohesion survey

These results in Table 2 show that most of the taxonomies achieved roughly the same level of cohesion for the clusters, roughly between 50 and 63%. However the WikiFreq taxonomy performed far better, with only one unit of the 30 judged as not cohesive. This success

shows that the Wikipedia background links are very effective as a means of grouping together similar items. This might be explained by considering that items grouped together under the same node will share a number of keywords which link to the same Wikipedia articles, thus ensuring that the items are very similar. In contrast the Wikipedia taxonomy and DBpedia ontology use categories rather than articles in Wikipedia as the concept nodes. These are much more loosely defined; each article in Wikipedia can belong to many categories and each category contains many articles. The results also indicate that Wikipedia articles as entities are much more clearly defined than the LDA topic keywords and thus work much better at grouping together the similar items.

Relation classification

For evaluating the taxonomy quality, we analyse what kinds of relations were present in these taxonomies. Given a child-parent pair A,B the evaluators were asked two questions:

- Are the two concepts A and B related? (Yes/No/I don't know) The evaluators were asked to judge the relation within the context of the cultural heritage taxonomy. A positive example was presented: Westminster and London, which were related because Westminster is in London. A negative example was Fish and Bicycle which were unrelated and would not be a useful pair to include in a taxonomy.
- If Yes, then how would you best define the relationship? Is A more specific than B, less specific than B, neither, or don't know? Examples were also given to help with this question. Westminster is more specific than London since Westminster is within London. The term Scientist is less specific than Physicist, since while all Physicists are Scientists, not all Scientists are Physicists (they could be biologists or chemists for example). For the `neither' option consider Physicist and Biologist. The concepts are related (both are scientists) but neither is more specific than the other.

Forty non-control pairs from each taxonomy were presented to the evaluators giving a total of 240 pairs. As for the previous experiment control pairs were manually identified where the answer should be obvious. Five pairs were shown on each page of which one was always a control pair.

Altogether 270 people attempted this survey. 97 people evaluated more than half the control pairs wrong or evaluated less than 5 pairs in total, and so their answers were excluded. Of the 173 remaining participants, a total of 3826 evaluations were made for each pair. A minimum of 8 evaluations and an average of 15.94 evaluations were made for each instance.

Taxonomy	A < B	A > B	Neither	Don't know	Agreement
LCSH	65.4	8.7	23.4	2.5	68.7
DBpedia	76.2	4.9	18.1	0.7	78.9
Wiki taxonomy	**78.3**	4.7	16.0	0.9	**82.8**
WN domains	63.6	6.3	28.0	2.0	67.6
LDA topics	21.4	14.8	**62.1**	1.6	61.0
Wiki freq	30.9	**22.6**	43.6	**2.9**	67.0

Table 3: Some results of the relation classification survey

The results in Table 3 follow a roughly clear pattern: the manually created taxonomies are markedly more likely to contain clearly related pairs of concepts. The A<B case is the most desirable for the taxonomies since we would prefer the most general concepts at the top of the hierarchy narrowing down into more specific concepts. The Wikipedia taxonomy and DBpedia both score relatively highly here, although both contain a surprisingly high number of cases where neither A or B was identified as more specific than the other (16.0 and 18.1\% respectively). For the Wikipedia taxonomy this shows that although almost all child-parent pairs are considered to be related concepts, they are not always easily identified with the child as more specific than the parent. Both WordNet domains and LCSH far worse, again with more relations identified as `neither'. The WikiFreq taxonomy contains a more mixed set of results with quite a high proportion of relations the `wrong way round' with A deemed to be less specific than B, although the highest number falls into the `neither' category. This result is a reflection of the nature of the links within the items. The taxonomy is ordered with the most frequent occurring links at the top going down towards the least. Clearly this is not enough to create the kind of general-to-specific relationships which are desirable. Finally the results for the LDA topics show that the majority are defined as `neither' - the concepts are topically related but mostly without any specificity ordering.

8.2 Intra-collection links

We now turn our attention to the intra-collection links (cf. Section 6). To create a dataset of items in Europeana, we randomly selected 295 pairs of items from Culture Grid and Scran because these collections are in English. They contain different types of items such as objects, archives, videos and audio files.

Each item corresponds to a metadata record consisting of textual information together with a URI. For each item, we extracted its title, description and subjects. The concatenation of these data is the textual information of the item.

After generating the dataset, we need to get reliable judgements of similarity on the pairs of items. This can be done by asking humans to give similarity scores so that to create a set of similarity scores to measure the performance of our similarity methods. To do so, we collect human judgements by making use of Crowdflower[10], a crowdsourcing platform.

We collected 3261 annotations from 99 participants. Each participant was presented with a page of 10 pairs of items and was asked to rate similarity between pairs using the following scale:

- 4: Completely related or identical items.
- 3: Related items.

[10] http://crowdflower.com/

- 2: Partly related items.
- 1: Unrelated items.
- 0: Completely unrelated items.

Participants were free to rate as many pages as they wanted to a maximum of 30 pages. In addition each pair must be rated from 10 different participants at least.

To ensure that annotations were not given randomly, we inserted one question with a predefined answer in each page. Annotations from participants that failed to answer these questions or participants that have given same rating to all of their answers were removed.

The final gold-standard set were generated by averaging the ratings of each pair. Moreover, we computed the inter-annotator agreement to test the level of the agreement between human responses as the average of the Pearson correlation between the ratings of each participant and the average ratings of the other participants. The higher the agreement, both qualitative is the gold-standard. The agreement of our gold-standard is rho=+0.553. In related work, (Grieser, 2011) reported an agreement of rho=+0.507.

We pre-process the data by removing stop words and applying stemming. For both tf.idf and LDA the training corpus was a total of 759896 Europeana items. We have filtered out all items that have no description and have either title length shorter than 4, or have a title which has been repeated more than 100 times. The textual information that we made use of the metadata is the concatenation of title, description and subjects (results included for titles, titles-descriptions, titles-subjects).

	T	**T+D**	**T+S**	**T+D+S**
Overlap	0.466	**0.487**	0.453	**0.487**
n-gram	0.258	0.334	0.362	0.399
tf.idf	0.413	0.441	0.423	0.437
LDA	0.435	0.428	0.426	**0.442**

Table 4: Results on similarity between items

An overview of the results obtained is in Table 4 (T:titles, D: description, S: subjects). Best performance is achieved by Normalised Overlap (0.487). Both LDA and tf.idf perform quite well but lower in comparison to Normalised Overlap (0.442 and 0.437 respectively). The N-gram Overlap fails to perform higher than 0.4 (0.399).

The results are quite surprising because one would expect that more sophisticated techniques such as LDA and tf.idf would have obtained better results than the simple overlap. We believe that the main reason of these results is the nature of the text data in Europeana. Documents are short, often their description is missing and in many cases subjects are identical to the title.

In any case, given the small difference with the LDA method that was used to produce the Intra-Collection links in D2.1, we also use this method in D2.2. This work was published in one workshop and journal (Aletras and Stevenson, 2012a; 2012b).

8.3 Background links

In this Section we present intrinsic evaluation of the background links to Wikipedia articles (cf. Section 7). We first evaluated whether the background links to Wikipedia could be used to link Cultural Heritage items automatically to a corresponding Wikipedia article describing the item. This way we are indirectly evaluating the links, although the evaluation is much more strict that just judging whether the background link is relevant for the item. For the later evaluation, the accuracy is around 74%, please refer to (Milne and Witten, 2008) for further details. In addition, we enriched the queries in the public evaluation CHiC (Cultural Heritage in CLEF).

Evaluating correspondence to Wikipedia articles

We selected a random subset comprising 400 items from the Scran and Cgrid collections in Europeana. The items were then ordered according to the subcollections they came from, so the annotators had a relatively coherent set of such as "The National Museum Record"[11], "The portables Antiquities Scheme"[12] or Scran[13]. Table 5 shows the type of the items in the sample. The majority are photographs, but there are also other types such as paintings or antique coins.

Type	Count
Photographs	276
Coins and Artifacts	57
Books, booklets etc	24
Other	21
Paintings	14
Audio and Video	8
Total	400

Table 5: Types of Europeana items in the sample.

The annotators were given the records with all the metadata. They could also access the item as shown in the Europeana interface and they had to return the URL of a single English Wikipedia article matching the item, or NIL if they could not find any matching entry. The definition of a matching entry provided to the annotators was: "the Wikipedia article and the item must describe the same particular object. In the case of photographs, the article must be about the subject of the photograph, e.g a particular person or location." Note that this definition of matching tries to find equivalent items and articles, and thus does not consider

[11] http://viewfinder.english-heritage.org.uk

[12] http://finds.org.uk/database

[13] http://www.scran.ac.uk

other kinds of relations between item and Wikipedia article, such as for example linking an item to the article about "photography" because it's a photograph, or linking an item to the article of the author.

The random subset of 400 items was independently tagged by two groups of annotators, one in Donostia and another in Sheffield, each one comprising three persons. As a result, the subset was annotated twice and two tags were obtained for each item. We chose one group's answers as gold standard, and used the other for calculating Inter Annotator Agreement (IAA) figures, as explained below. According to the gold standard, 89 items were successfully linked to Wikipedia articles (22% of the sample). Given that the method for matching entries was very strict it came as a surprise that the annotators were able to identify a matching article for so many items. This result suggests that the task of matching Cultural Heritage elements to external resources such as Wikipedia can have a real impact in the richness of the descriptions for that 22% of the sample.

The overall Inter-Annotator Agreement (IAA) between the two tags available for each item are very high: 92.5% in the Cgrid collection and 80.0% in Scran. The agreement takes into account the items which were not associated with an article (i.e. tagged as NIL). Most of the disagreements were due to one tagger not returning any article (NIL) and the other tagger choosing one article. We analysed these disagreements and in general the articles were relevant, and thus well linked. In a few cases, the article is not appropriate, although close. For instance, the item titled "Glyndebourne Opera Company present Le Comte Ory" containing one picture of a performance was linked to an article about an opera festival hold in Glyndebourne. Overall the high IAA numbers show that the annotation is reliable and that the task itself is well-defined.

The analysis of the background links generated in PATHS shows that up to 22% of items in Europeana can be matched with a counterpart in Wikipedia, a remarkable proportion when the vast number of items in Europeana is considered. It was found that up to 75.9% of the items matching a Wikipedia article could be linked automatically, given a perfect algorithm for choosing the correct one among the articles returned by the systems. A simple heuristic based on the weights returned by the systems, length and position in the title attains recall of up to 55.2%. The 75.9% upperbound shows that there is room for improvement. Note that we only used the text in the title, and an analysis of the text in the description could allow to find more and better matching articles. Please refer to (Agirre et al. 2012) for further details on this evaluation.

Evaluating semantic query enrichment in CHiC

The Cultural Heritage in CLEF (CHiC) pilot evaluation lab[14] aims at moving towards a systematic and large-scale evaluation of cultural heritage digital libraries and information

[14] http://www.promise-noe.eu/chic-2012/home

access systems. Data test collections and queries will come from the cultural heritage domain (in 2012 from Europeana) and tasks will contain a mix of conventional system-oriented evaluation scenarios (e.g. ad-hoc retrieval and semantic enrichment) for the CH domain, i.e. a variability task to present a particular good overview ("must sees") over the different objects types and categories in the collection targeted towards a casual user.

The goal of the semantic query enrichment task is to present a ranked list of at most 10 concepts (words or phrases) for each query. These concepts should semantically enrich the query and/or guess the information need or query intent of the user. The concepts can be extracted from the Europeana data that has been provided (internal) or make use of other resources such as the Web or Wikipedia (external). Using the background links provided by PATHS and the use of a random walk algorithm over Wikipedia, we obtained a MAP of 29.05, the third best result on the task. Please refer to (Agirre et al. 2012b) for details.

9 Recommendations for data providers

In this Section we will briefly mention some of the difficulties we found with the metadata in Europeana items, as follows:

- Include the information in the proper field. In many cases the description field includes details about the provider and no textual descriptor about the item.
- Provide informative titles, and some description text. Many items have very short titles (e.g. coin) and no description.
- Provide information about the author of the item, the provider, location and date, which many times are missing.
- Provide good subject descriptions using a vocabulary which is of widespread use (e.g. LCSH).
- Provide links to related Wikipedia articles. These links provide background information. Although PATHS makes an effort to add those automatically, manual links currently have better quality, and can help improve automatic mapping in the future.

Better metadata and more detailed description of the items will mean that the automatic enrichment software in PATHS will be able to produce better quality information, and thus improve the user experience when using the PATHS prototype.

10 Web service for WP2

PATHS has currently focused in the offline processing of selected collections from Europeana. In this Section we briefly mention a web service which will allow independent content providers to enrich their items.

The web service allows enriching items one by one. Given a new item in EDM this web service will return an enriched EDMpaths (see Section 3) with the following information:

- Related items
- Background links to Wikipedia

To get the related items for an item, the API of the SOLR search engine is going to be used. EDMpaths will be enriched with the first ten items retrieved by the search engine using the query formed with the terms of the title, subject and description fields of the item.

In addition, the output EDMpaths will be enriched with background links to Wikipedia for the input item.

The Web service is accessible in the following URL:
http://ixa2.si.ehu.es/paths_wp2/paths_wp2.pl

11 Interface to access items

An internal demo with part of the enrichment data is available as an aid in the development of ideas in the following URL:

http://ixa2.si.ehu.es/europeanaEN/search
http://ixa2.si.ehu.es/europeanaES/search

References

E. Agirre, A. Barrena, O. Lopez de Lacalle, A. Soroa, M. Stevenson and S. Fernando. Matching Cultural Heritage items to Wikipedia. In Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey. 2012.

E. Agirre, Paul D. Clough, Samuel Fernando, Mark Hall, Arantxa Otegi, Mark Stevenson: The Sheffield and Basque Country Universities Entry to CHiC: Using Random Walks and Similarity to Access Cultural Heritage. CLEF (Online Working Notes/Labs/Workshop). 2012b

N. Aletras and M. Stevenson. Computing Similarity between Items in a Digital Library of Cultural Heritage. To appear in ACM Journal on Computing and Cultural Heritage (JOCHH). 2012.

N. Aletras and M. Stevenson. Computing similarity between cultural heritage items using multimodal features. In Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH '12). 2012.

S. Fernando, M. Hall, E. Agirre, A. Soroa, P. Clough and M. Stevenson. Comparing taxonomies for organising collections of documents. In Proceedings of The 24th International Conference on Computational Linguistics (COLING 2012). 2012.

S. Fernando and M. Stevenson. Mapping WordNet synsets to Wikipedia articles. In Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey. 2012.

K. Grieser, T. Baldwin, F. Bohnert, and L. Sonenberg. Using ontological and document similarity to estimate museum exhibit relatedness. ACM Journal on Computing and Cultural Heritage, 3(3):10:1–10:20, 2011

M. Hall, O. Lopez de Lacalle, A. Soroa, P. D Clough, and E. Agirre. Enabling the Discovery of Digital Cultural Heritage Objects through Wikipedia. In Proceedings of the LaTeCH workshop held at EACL 2012. 2012.

D. Milne, and I. H. Witten. Learning to link with wikipedia. In Proceeding of CIKM ’08, pages 509–518, New York, NY, USA. 2008.

Annex 1: XML schema for ESEpaths

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema  xmlns:xs="http://www.w3.org/2001/XMLSchema"
          targetNamespace="http://www.paths-project.eu/"
          version="1.0"
          elementFormDefault="qualified"
          attributeFormDefault="unqualified">

  <xs:annotation>
    <xs:documentation xml:lang="en">
      This Schema extends ESE V3.4 XML Schema for PATHS related items. Specifically it adds the
following elements:
      paths:background_link
      paths:related_item
    </xs:documentation>
  </xs:annotation>

  <xs:element name="informativeness">
    <xs:documentation>
      Item informativeness.
    </xs:documentation>
    <xs:complexType>
      <xs:simpleContent>
            <xs:extension base="xs:float">
            </xs:extension>
      </xs:simpleContent>
    </xs:complexType>
  </xs:element>

  <xs:element name="normalized_date">
    <xs:documentation>
     As the date strings in ESE dc:date elements have many formats,
          we store the normalized data in this element. Only dates
          before 1990 are considered.
    </xs:documentation>
    <xs:complexType>
      <xs:simpleContent>
            <xs:extension base="xs:string">
            </xs:extension>
      </xs:simpleContent>
    </xs:complexType>
  </xs:element>

  <xs:element name="vocabulary">
    <xs:complexType mixed="true">
      <xs:attribute name="source" type="xs:string" use="required">
            <xs:documentation>
              The name of the external vocabulary.
            </xs:documentation>
      </xs:attribute>
      <xs:attribute name="URI" type="xs:anyURI" use="optional">
            <xs:documentation>
```

```
            URI of the specific category in case the vocabulary taxonomy has one (e.g. a dbpedia
category).
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="confidence" type="xs:float" use="optional">
          <xs:documentation>
            Confidence of the association.
          </xs:documentation>
    </xs:attribute>
  </xs:complexType>
 </xs:element>

 <xs:element name="event">
  <xs:documentation>
   This element is used to annotate eventive information within an
   item. The text element within this element is the eventual word form
   as present in the original field.
  </xs:documentation>
  <xs:complexType mixed="true">
   <xs:attribute name="source" type="xs:string" use="required">
          <xs:documentation>
            The name of the external resource of the event (for instance, WordNet).
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="canonical_form" type="xs:string" use="optional">
          <xs:documentation>
            The canonical word form of the annotated event.
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="confidence" type="xs:float" use="optional">
          <xs:documentation>
            Confidence of the association.
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="start_offset" type="xs:integer" use="optional">
          <xs:documentation>
            The offset (in characters) within the field element where the text anchor begins.
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="end_offset" type="xs:integer" use="optional">
          <xs:documentation>
            The offset (in characters) within the field element where the text anchor ends.
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="field" type="xs:string" use="optional">
          <xs:documentation>
            The field of the item where the anchor for this relation is located.
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="field_no" type="xs:integer" use="optional">
          <xs:documentation>
            The index of the field in the item.
          </xs:documentation>
    </xs:attribute>
  </xs:complexType>
 </xs:element>

 <xs:element name="related_item">
  <xs:documentation>
```

```
    This element relates the current item with another, semantically
    similar one. The text element within the element is the id of the
    related item.
   </xs:documentation>
   <xs:complexType mixed="true">
    <xs:attribute name="method" type="xs:string" use="optional">
          <xs:documentation>
           Which method produced the association.
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="type" type="xs:string" use="optional">
          <xs:documentation>
           The type of the relation (for instance, same_author, same_location, etc).
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="confidence" type="xs:float" use="optional">
          <xs:documentation>
           Confidence of the association.
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="field" type="xs:string" use="optional">
          <xs:documentation>
           The field of the item where the anchor for this relation is located.
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="field_no" type="xs:integer" use="optional">
          <xs:documentation>
           The index of the field in the item.
          </xs:documentation>
    </xs:attribute>
   </xs:complexType>
  </xs:element>

  <xs:element name="background_link">
   <xs:documentation>
    This element relates the current item with an external resource. The
    text element within the element is the URI to the external link.
   </xs:documentation>
   <xs:complexType mixed="true">
    <xs:attribute name="source" type="xs:string" use="required">
          <xs:documentation>
           The name of the external resource.
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="title" type="xs:string" use="optional">
          <xs:documentation>
           Title of the url which the background link points to. This title
           is the one to be shown in the UI.
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="confidence" type="xs:float" use="optional">
          <xs:documentation>
           Confidence of the association.
          </xs:documentation>
    </xs:attribute>
    <xs:attribute name="start_offset" type="xs:integer" use="optional">
          <xs:documentation>
           The offset (in characters) within the field element where the text anchor begins.
          </xs:documentation>
```

```
    </xs:attribute>
    <xs:attribute name="end_offset" type="xs:integer" use="optional">
        <xs:documentation>
          The offset (in characters) within the field element where the text anchor ends.
        </xs:documentation>
    </xs:attribute>
    <xs:attribute name="field" type="xs:string" use="optional">
        <xs:documentation>
          The field of the item where the anchor for this relation is located.
        </xs:documentation>
    </xs:attribute>
    <xs:attribute name="field_no" type="xs:integer" use="optional">
        <xs:documentation>
          The index of the field in the item.
        </xs:documentation>
    </xs:attribute>
    <xs:attribute name="method" type="xs:string" use="optional">
        <xs:documentation>
          Which method produced the association.
        </xs:documentation>
    </xs:attribute>
    <xs:attribute name="sentiment" type="xs:string" use="optional">
        <xs:documentation>
          Polarity of the textual information included in the corresponding
          link. It has fixed values, namely "pos" for positive results,
          "neg" for negative and "neu" for neutral.
        </xs:documentation>
    </xs:attribute>
    <xs:attribute name="paths_classification" type="xs:string" use="optional">
        <xs:documentation>
          Category on which the specific result belongs. It has fixed values
          for this case, "academic", "ch" for cultural heritage and "other"
          for results not belonging to the previous categories.
        </xs:documentation>
    </xs:attribute>
  </xs:complexType>
 </xs:element>

</xs:schema>
```