PATHSenrich: A Web Service Prototype for Automatic Cultural Heritage Item Enrichment

Eneko Agirre, Ander Barrena, Kike Fernandez, Esther Miranda, Arantxa Otegi, and Aitor Soroa

IXA NLP Group, University of the Basque Country UPV/EHU
arantza.otegi@ehu.es

Abstract. Large amounts of cultural heritage material are nowadays available through online digital library portals. Most of these cultural items have short descriptions and lack rich contextual information. The PATHS project has developed experimental enrichment services. As a proof of concept, this paper presents a web service prototype which allows independent content providers to enrich cultural heritage items with a subset of the full functionality: links to related items in the collection and links to related Wikipedia articles. In the future we plan to provide more advanced functionality, as available offline for PATHS.

1 Introduction

Large amounts of cultural heritage (CH) material are now available through online digital library portals, such as Europeana[1]. Europeana hosts millions of books, paintings, films, museum objects and archival records that have been digitised throughout Europe. Europeana collects contextual information or metadata about different types of content, which the users can use for their searches.

The main strength of Europeana lays in the vast number of items it contains. Sometimes, though, this quantity comes at the cost of a restricted amount of metadata, with many items having very short descriptions and a lack of rich contextual information. One of the goals of the PATHS project[2] is precisely to enrich CH items, using a selected subset of Europeana as a testbed[1].

Whithin the project, this enrichment will make possible to create a system that acts as an interactive personalised tour guide through Europeana collections, offering suggestions about items to look at and assist in their interpretation by providing relevant contextual information from related items within Europeana and items from external sources like Wikipedia. Users of such digital libraries may require information for purposes such as learning and seeking answers to questions. This additional information supports users in fulfilling their information need, as the evaluation of the first PATHS prototype shows [2].

In this paper we present a web service prototype which allows independent content providers to enrich CH items. Specifically, the service enriches the items

[1] http://www.europeana.eu/portal/

[2] http://www.paths-project.eu

T. Aalberg et al. (Eds.): TPDL 2013, LNCS 8092, pp. 462–465, 2013.

with two types of information. On the one hand, the item will be linked to similar items within the collection. On the other hand, the item will be linked to Wikipedia articles which are related to it.

There have been many attempts to automatically enrich cultural heritage metadata. Some projects (for instance, MIMO-DB[3] or MERLIN[4]) relate CH objects with terms of an external authority or vocabulary. Some others (like MACE[5] or YUMA [6]) adopt a collaborative annotation paradigm for metadata enrichment. To our knowledge, PATHS is the first project using semantic NLP processing to link CH items to similar items or external Wikipedia articles.

The current service has limited bandwidth, and provides a selected subset of the enrichment functionality available internally in the PATHS project. The quality of the links produce is also slightly lower, although we plan to improve it in the short future. However, we think that the prototype is useful to demonstrate the potential to construct a web service for automatically enriching CH items with high quality information.

2 Demo Description

The web service takes as input one CH item represented following the Europeana Data Model (EDM) in JSON format, as exported by the Europeana API v2.0[7] (a sample record is provided in the interface). The web service returns the following:

- A list of 10 closely related items within the collection.
- A list of Wikipedia pages which are related to the target item.

Figure 1 shows a snapshot of the web service. The service is publicly accessible following the URL `http://ixa2.si.ehu.es/paths_wp2/paths_wp2.pl`.

The enrichment is performed by analyzing the metadata associated with the item, i.e., the title of the item, its description, etc. The next sections briefly describe how this enrichment is performed.

2.1 Related Items within the Collection

The list of related items is obtained by first creating a query with the content of the title, subject and description fields (stopwords are removed). The query is then posted to a SOLR search engine[8]. The SOLR search engine accesses an index created with the subset of Europeana items already enriched offline within the PATHS project. In that way, the most related Europeana items in the subset are obtained, and the identifiers of those related items are listed. Note that the related items used internally in the PATHS project are produced using more sophisticated methods. Please refer to [1] for further details.

[3] http://www.mimo-international.com

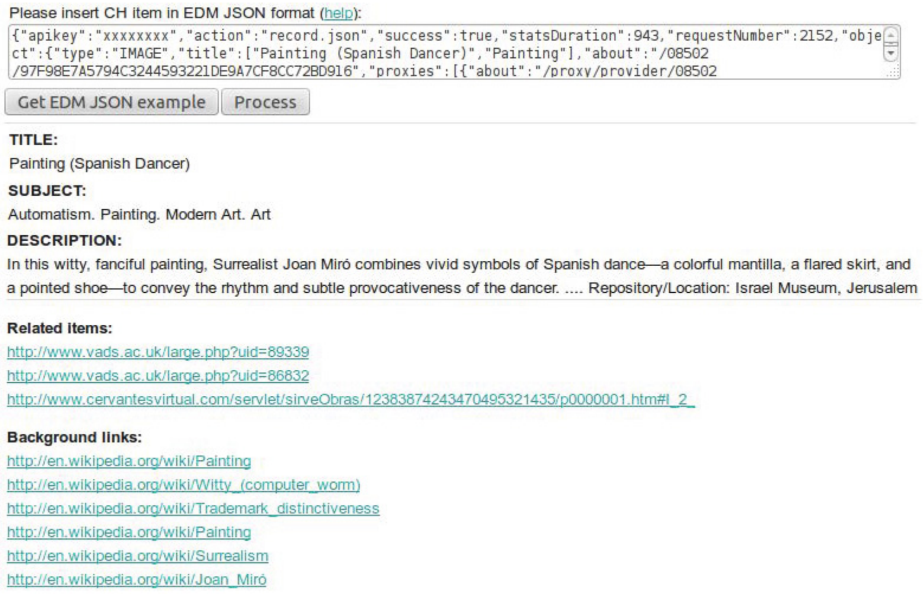
[4] http://www.ucl.ac.uk/ls/merlin

[5] http://www.mace-project.eu

[6] http://dme.ait.ac.at/annotation

[7] http://preview.europeana.eu/portal/api-introduction.html

[8] http://lucene.apache.org/solr/



Please insert CH item in EDM JSON format (help):

```
{"apikey":"xxxxxxxx","action":"record.json","success":true,"statsDuration":943,"requestNumber":2152,"obje
ct":{"type":"IMAGE","title":["Painting (Spanish Dancer)","Painting"],"about":"/08502
/97F98E7A5794C3244593221DE9A7CF8CC72BD916","proxies":[{"about":"/proxy/provider/08502
```

TITLE:

Painting (Spanish Dancer)

SUBJECT:

Automatism. Painting. Modern Art. Art

DESCRIPTION:

In this witty, fanciful painting, Surrealist Joan Miró combines vivid symbols of Spanish dance—a colorful mantilla, a flared skirt, and a pointed shoe—to convey the rhythm and subtle provocativeness of the dancer. .... Repository/Location: Israel Museum, Jerusalem

Related items:

http://www.vads.ac.uk/large.php?uid=89339
http://www.vads.ac.uk/large.php?uid=86832
http://www.cervantesvirtual.com/servlet/sirveObras/12383874243470495321435/p0000001.htm#I_2_

Background links:

http://en.wikipedia.org/wiki/Painting
http://en.wikipedia.org/wiki/Witty_(computer_worm)
http://en.wikipedia.org/wiki/Trademark_distinctiveness
http://en.wikipedia.org/wiki/Painting
http://en.wikipedia.org/wiki/Surrealism
http://en.wikipedia.org/wiki/Joan_Miró

Fig. 1. Web service interface. It consists of a text area to introduce the input item in JSON format (top). The "Get EDM JSON example" button can be used to get an input example. Once a JSON record is typed, click "Process" button to get the output. The output (bottom) consists on a list of related items and background links.

2.2 Related Wikipedia Articles

For linking the items to Wikipedia articles we follow an implementation similar to the method described in [3]. This method creates a *dictionary*, an association between string mentions with all possible articles the mention can refer to. Our dictionary is constructed using the title of the Wikipedia article, the redirect pages, the disambiguation pages and the anchor texts from Wikipedia links. Mentions are lower-cased and all text between parenthesis is removed. If the mention links to a disambiguation page, it is associated with all possible articles the disambiguation page points to. Besides, each association between a mention and article is scored with the prior probability, estimated as the number of times that the mention occurs in the anchor text of an article. Note that such dictionaries can disambiguate any mention, just returning the highest-scoring article for this particular mention.

Once the dictionary is built, the web service analyzes the title, subject and description fields of the CH item and matches the longest substring within those fields with entries in the dictionary. When a match is found, the Wikipedia article with highest score for this entry is returned. Note that the links to Wikipedia in the PATHS project are produced using more sophisticated methods. Please refer to [1] for further details.

3 Conclusions and Future Work

This paper presents a web service prototype which automatically enriches CH items with metadata. The web service is inspired in the enrichment work carried out in the PATHS project, but, contrary to the batch methodology used in the project, this enrichment is performed online. The prototype has been designed for demonstration purposes, to showcase the feasibility of providing full-fledged automatic enrichment.

Our plans for the future include moving the offline enrichment services which are currently being evaluated in the PATHS project to the web service. In the case of related Wikipedia articles, we will take into account the context of the matched entities, which improves the quality of the links [4], and we will include a filtering algorithm to discard entities that are not relevant. Regarding related items, we will classify them according to the type of relation [5]. In addition we plan to automatically organize the items hierarchically, according to a Wikipedia-based vocabulary [6].

Acknowledgements. The research leading to these results was carried out as part of the PATHS project (http://www.paths-project.eu) funded by European Communitys Seventh Framework Programme (FP7/2007- 2013) under grant agreement no. 270082. The work has been also funded by the Basque Government (project IBILBIDE, SAIOTEK S-PE12UN089).

References

1. Otegi, A., Agirre, E., Soroa, A., Aletras, N., Chandrinos, C., Fernando, S., Gonzalez-Agirre, A.: Report accompanying D2.2: Processing and Representation of Content for Second Prototype. PATHS Project Deliverable (2012), http://www.paths-project.eu/eng/content/download/2489/18113/version/2/file/D2.2.Content+Processing-2nd+Prototype-revised.v2.pdf
2. Griffiths, J., Goodale, P., Minelli, S., de Polo, A., Agerri, R., Soroa, A., Hall, M., Bergheim, S.R., Chandrinos, K., Chryssochoidis, G., Fernie, K., Usher, T.: D5.1: Evaluation of the first PATHS prototype. PATHS Project Deliverable (2012), http://www.paths-project.eu/eng/Resources/D5.1-Evaluation-of-the-1st-PATHS-Prototype
3. Chang, A.X., Spitkovsky, V.I., Yeh, E., Agirre, E., Manning, C.D.: Stanford-UBC entity linking at TAC-KBP. In: Proceedings of TAC 2010, Gaithersburg, Maryland, USA (2010)
4. Han, X., Sun, L.: A Generative Entity-Mention Model for Linking Entities with Knowledge Base. In: Proceedings of the ACL, Portland, Oregon, USA (2011)
5. Agirre, E., Aletras, N., Gonzalez-Agirre, A., Rigau, G., Stevenson, M.: UBC_UOS-TYPED: Regression for typed-similarity. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Atlanta, Georgia, USA (2013)
6. Fernando, S., Hall, M., Agirre, E., Soroa, A., Clough, P., Stevenson, M.: Comparing Taxonomies for Organising Collections of Documents. In: Proceedings of COLING 2012, Mumbai, India (2013)