Metadata Enrichment Services for the Europeana Digital Library

Giacomo Berardi[1], Andrea Esuli[1], Sergiu Gordea[2],
Diego Marcheggiani[1], and Fabrizio Sebastiani[1]

[1] Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, IT
{Giacomo.Berardi,Andrea.Esuli,Diego.Marcheggiani,
Fabrizio.Sebastiani}@isti.cnr.it
[2] Austrian Institute of Technology, 1220 Vienna, AT
Sergiu.Gordea@ait.ac.at

Abstract. We demonstrate a metadata enrichment system for the Europeana digital library. The system allows different institutions which provide to Europeana pointers (in the form of metadata records - MRs) to their content to enrich their MRs by classifying them under a classification scheme of their choice, and to extract/highlight entities of significant interest within the MRs themselves. The use of a supervised learning metaphor allows each content provider (CP) to generate classifiers and extractors tailored to the CP's specific needs, thus allowing the tool to be effectively available to the multitude (2000+) of Europeana CPs.

1 Introduction

Europeana[1] is a digital library that acts as an aggregator of thousands of collections and archives of digitised printed and audiovisual material (including books, paintings, sculptures, movies, and artworks of different nature) [3]. Thousands of private and public institutions spread across the European Union are served by Europeana, ranging from major museums of international fame, to libraries of regional or domain-specific scope. Europeana users can thus search a virtual collection of millions of digital objects which altogether represent a significantly large window on Europe's cultural and scientific heritage.

Europeana does not contain the digitised objects themselves; it contains pointers to them, in the form of searchable metadata records (MRs). MRs are thus the objects around which most of the services that Europeana provides, including searching and browsing, revolve. MRs (compiled according to a format called *Europeana Semantic Elements*[2]) are fed to Europeana by each individual content provider (CP). Each such record describes (and links to) a digital object that resides in the CP's archive.

[1] http://Europeana.eu

[2] http://www.europeana.eu/schemas/ese/

P. Zaphiris et al. (Eds.): TPDL 2012, LNCS 7489, pp. 508–511, 2012.

We here demonstrate a service for the enrichment of Europeana MRs that has been developed in the context of ASSETS, a project funded by the European Commission and aimed at developing new value-added, content-based services for Europeana. The MR is the *de facto* user's gateway to the digital object itself. Enriching the semantics of a MR has thus a beneficial effect on the entire spectrum of the user's experience, including searching and browsing. The metadata enrichment service that we describe here is to be installed on the Europeana portal, and will allow each contributing CP to enrich the semantics of its own MRs prior to contributing them to Europeana. Different CPs are thus encouraged to use the same enrichment tool, thus allowing greater uniformity across MRs of different provenance.

We view metadata enrichment as consisting of essentially two activities: *classification of MRs*, and *information extraction from MRs.* Classification consists in the task of associating to a given MR one or more classes from a pre-specified classification scheme. Information extraction consists instead in the individuation ("extraction") of substrings of text contained in the MR that instantiate one among a set of prespecified concepts of interest. These two services are described in Sections 2.1 and 2.2, respectively.

2 Architecture of the Metadata Enrichment System

2.1 The Metadata Classification Component

Classification refers to the task of associating to a MR one or more classes from a pre-specified classification scheme (i.e., a set of classes, possibly organized as a taxonomy). The chosen classification scheme can be domain-specific or general-purpose. For instance, the Accademia Nazionale di Santa Cecilia (an Italian cultural institution active in the field of classical and contemporary music, and a Europeana CP) will typically be interested in adopting a music-specific classification scheme, while another institution of broader scope might want to adopt a general-purpose scheme such as the Library of Congress Subject Headings[3].

Setting up a classification system for Europeana is challenging, because of the sheer diversity (a) of classification schemes that CPs might choose, and (b) of languages in which the MRs are going to be expressed in. Given this, it would be implausible to provide a classification service based on manually written classification rules, since this would place the burden of rule-writing on the CPs themselves, who would then probably renounce using the service.

As a result, our classification service is based on supervised learning technology: a learning algorithm learns, from a sample of manually preclassified documents that are provided to it, the characteristics that a given MR should have in order to be associated to a given class [5]. This frees the CP from the burden of writing classification rules, and only requires it to provide a sample of manually classified MRs. In many cases these latter may already be available

[3] `http://www.loc.gov/aba/cataloging/subject/weeklylists/`

to the CP as a product of a classification activity that the CP has carried out in its daily operations.

As the supervised learning technology we have used the TreeBoost system, a member of the family of "boosting"-based supervised learning algorithms that has shown state-of the-art accuracy across a variety of datasets [2]. TreeBoost allows the use of classification schemes organized either as a tree or as a directed acyclic graph. In order to cater for the multiplicity of languages in which the MRs can be expressed we use a completely language-independent preprocessing module which only consists of extracting the MRs from the records; stop word removal, stemming, and other types of linguistic analysis that are language-dependent are not used.

Classification accuracy results from experiments on many datasets of metadata records from Europeana CPs are presented in [1].

2.2 The Information Extraction Component

Information extraction refers to the task of identifying ("extracting"), in a given text, substrings that instantiate "concepts" belonging to a prespecified set [4]. Examples of "domain-independent" such concepts may be Person, Location, Organization; examples of domain-dependent concepts (e.g., for the domain of music) may instead be Director, Instrument, or Composer. Identifying instances of such concepts in a MR may be beneficial for browsing, and is ultimately a means of adding semantics to the MR and of enabling semantic search.

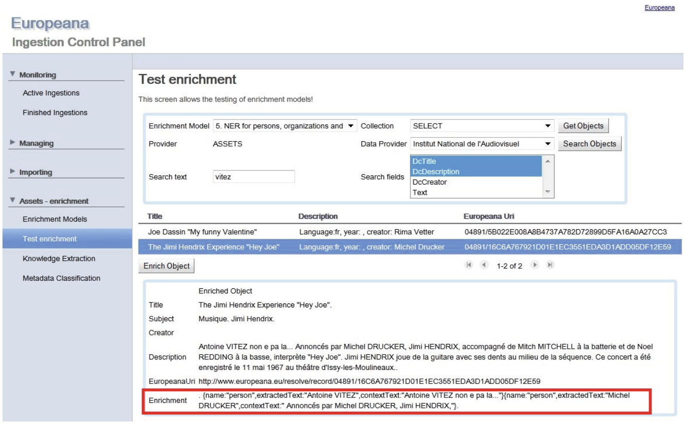
For reasons similar to the ones discussed for classification, it would be inappropriate to deploy an information extraction service based on manually written extraction rules. Again, we have chosen the supervised learning route, according to which the CP provides a general-purpose learning system with a set of texts in which the instances of the concepts of interest have been marked as such; from these annotated texts the system learns to extract the instances of the concepts.

As the supervised learning technology we have used an algorithm belonging to the family of *conditional random fields* (CRFs), which are nowadays considered state-of-the-art for addressing sequence learning tasks. Since syntactic analysis (particularly: POS tagging) is known to be beneficial in information extraction, we here do not avoid language-dependent processing. We first apply an automatic language recognizer to the record in order to determine the language it is written in, and then we submit the record to a POS tagging phase in the cases in which a POS tagger is available for the specified language.

Information extraction accuracy results from experiments on many datasets of metadata records from Europeana CPs are presented in [1], along with additional details about the preprocessing steps we have enacted.

2.3 The Ingestion Control Panel

The backend processing work performed by Europeana follows a complex process that includes operations related to customer relationship management, metadata



Test enrichment

This screen allows the testing of enrichment models!

Enrichment Model: 5. NER for persons, organizations and
Collection: SELECT
Provider: ASSETS
Data Provider: Institut National de l'Audiovisuel
Search text: vitez
Search fields: DcTitle, DcDescription, DcCreator, Text

Title	Description	Europeana Uri
Joe Dassin "My funny Valentine"	Language:fr, year: , creator: Rima Vetter	04891/5B022E008A8B4737A782D72899D5FA16A0A27CC3
The Jimi Hendrix Experience "Hey Joe"	Language:fr, year: , creator: Michel Drucker	04891/16C6A767921D01E1EC3551EDA3D1ADD05DF12E59

	Enriched Object
Title	The Jimi Hendrix Experience "Hey Joe".
Subject	Musique. Jimi Hendrix.
Creator	
Description	Antoine VITEZ non e pa la... Annoncés par Michel DRUCKER, Jimi HENDRIX, accompagné de Mitch MITCHELL à la batterie et de Noel REDDING à la basse, interprète "Hey Joe". Jimi HENDRIX joue de la guitare avec ses dents au milieu de la séquence. Ce concert a été enregistré le 11 mai 1967 au théâtre d'Issy-les-Moulineaux..
EuropeanaUri	http://www.europeana.eu/resolve/record/04891/16C6A767921D01E1EC3551EDA3D1ADD05DF12E59
Enrichment	. {name:"person",extractedText:"Antoine VITEZ",contextText:"Antoine VITEZ non e pa la..."}{name:"person",extractedText:"Michel DRUCKER",contextText:" Annoncés par Michel DRUCKER, Jimi HENDRIX,"}.

Fig. 1. The Ingestion Control Panel

harvesting, metadata processing (i.e. data normalization), thumbnail generation, creation of submission and access information packages, etc. We extend the GUI of the Europeana Ingestion Control Panel by integrating the invocation of enrichment services and supporting the following functionality (see Fig. 1):

- Enrichment model learning: by using a training set that suites their metadata, the content providers or Europeana are allowed to run the learning of enrichment models for metadata classification and/or information extraction;
- Enrichment by metadata classification: the classification of a collection can be performed by selecting an appropriate classification model. The user is also allowed to test the model on a particular collection object;
- Enrichment by information extraction: the extraction of the structured information from the object descriptions can be performed similarly to the classification by selecting an appropriate model and the collection or collection object to be enriched.

References

1. Berardi, G., Esuli, A., Marcheggiani, D., Sebastiani, F., Gordea, S., Täckström, O.: Ingestion services: 2nd release. Deliverable D2.1.3, ASSETS Project ICT PSP 250527, Commission of the European Communities (2012)
2. Esuli, A., Fagni, T., Sebastiani, F.: Boosting multi-label hierarchical text categorization. Information Retrieval 11(4), 287–313 (2008)
3. Purday, J.: Think culture: Europeana.eu from concept to construction. The Electronic Library (6), 919–937 (2009)
4. Sarawagi, S.: Information extraction. Foundations and Trends in Databases 1(3), 261–377 (2008)
5. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)