

# Application Profiles Supporting Cross-Language and other Functionalities for Library Metadata

Barbara Levergood<sup>1</sup>, Sally Chambers<sup>2</sup>, Luigi Siciliano<sup>3</sup>

<sup>1</sup> Niedersächsische Staats- und Universitätsbibliothek Göttingen, Papendiek 14,  
37073 Göttingen, Germany  
levergood@sub.uni-goettingen.de

<sup>2</sup> The European Library, The National Library of the Netherlands, PO Box 90407, 2509 LK,  
The Hague, The Netherlands  
Sally.Chambers@KB.nl

<sup>3</sup> University Library, Free University of Bozen-Bolzano, Universitätsplatz 1 - piazza  
Università, 1, 39100 Bozen-Bolzano, Italy  
Luigi.Siciliano@unibz.it

**Abstract.** The CACAO project provides an infrastructure that enables cross-language functionality in digital libraries and library catalogues. The European Library is a free service that aggregates the bibliographic and digital collections of Europe's national libraries via a single multilingual interface. The involvement of The European Library in the CACAO project has assisted in the development of the CACAO Application Profile and has furthermore facilitated the development of The European Library Application Profile for Objects to better facilitate cross-language searching.

**Keywords:** CACAO Project, The European Library, application profiles, metadata, Dublin Core, digital libraries, library catalogues, cross-language, multilingual.

## 1 Introduction

The CACAO project<sup>1</sup> provides an infrastructure that enables cross-language functionality in digital libraries and library catalogues using CACAO's information retrieval and natural language processing (NLP) technologies. Through CACAO, the end-user can enter a query in his/her own language and retrieve documents and objects in any supported language. The European Library (TEL) is a free service that offers access to the bibliographic and digital collections of Europe's national libraries via a single multilingual interface.

This paper describes two application profiles (APs) for the metadata to be ingested by CACAO that have helped to facilitate aggregation and improve CACAO performance [1]. We discuss how the involvement of The European Library in

---

<sup>1</sup> CACAO Project (Cross-language Access to Catalogues and Online Libraries) is a 24-month targeted project supported by the eContentplus Programme of the European Commission. <http://www.cacaoproject.eu/>

CACAO has assisted in the development of one of the CACAO APs based on The European Library Application Profile for Objects (TEL-AP for Objects) [2].

## 2 CACAO's Application Profiles

As defined by Heery and Patel, application profiles are “schemas which consist of data elements drawn from one or more namespaces, combined together by implementors, and optimised for a particular local application” [3]. The CACAO AP working group identified several requirements that influence the choice of metadata formats. Since CACAO wanted to harvest metadata from OAI-PMH [4] repositories, the Simple Dublin Core [5] required by the OAI-PMH Guidelines [6] should be among the formats selected by CACAO. The metadata should be in a format that could be reused. The format and encoding of the metadata should be mature, readily available, and easy to implement. The CACAO AP should be based on an AP that CACAO would have to implement anyway. The AP would need to be flexible, offering a sensible requirement for a minimum record, i.e. title and identifier, but also rich metadata supporting CACAO's NLP-based cross-language services, with language specifications, subject headings, classification notations, alternative titles, table of contents, etc. CACAO's solution was to create two APs, one based on Simple Dublin Core and one based on the TEL-AP for Objects, which is in turn based on the Dublin Core Library Application Profile (DC-Lib) [7], itself Qualified Dublin Core-based.

The next task was to identify the requirements relevant at the element- and attribute-levels. First, CACAO needs to analyze the text of certain fields as a part of the indexing process. The language of the metadata is used if available, otherwise the language of the resource or a language guesser may be used. The predominant language of a vocabulary encoding scheme (VES) used in a subject field might also be used as an imperfect substitute. Second, CACAO is experimenting with a word sense disambiguation tool, Word2Category [8], [9], that associates words with classifications in a classification system such as Dewey Decimal Classification [10]. In order to perform this association, the language of the metadata and the classification system must be identified. Third, interfaces for cross-language services need to support those services; searching, facets and the display may all draw on information about the language of the resource.

Based on these requirements, we offer some best practices for how APs might be optimized for cross-language functionalities. All can be implemented in Qualified Dublin Core. Only the identification of the VES is not possible in Simple Dublin Core. All are deviations from the TEL-AP for Objects. (1) Recommend the use of `dc:language` for the language of the resource. (2) Recommend the use of `xml:lang` for the language of the metadata for text elements that contain semantically important content.<sup>2</sup> (3) Recommend the use of `xsi:type` for the

---

<sup>2</sup> Note, however, that one of the Qualified DC XML Schemas, version 2008-02-11, <http://dublincore.org/schemas/xmls/qdc/2008/02/11/dcterms.xsd>, prohibits the use of the `xml:lang` attribute with a number of vocabulary encoding schemes, including LCSH.

identification of the VESs of subject fields. (4) Provide an XML schema which permits local customizations such as the identification of the VESs of subject fields via attributes or the addition of new elements.

### 3 The European Library Application Profile for Objects

In order for The European Library to aggregate the collections from Europe's national libraries, an interoperable metadata format was needed. Within the national library community, it was felt that "TEL will use a collection of namespaces, among which the DC-Lib will be the most important, although this will probably not be sufficient for TEL" [11]. The decision was therefore made that the TEL project<sup>3</sup> would develop its own AP with DC-Lib as the basis.

Three remaining key concerns were identified. (1) The need for new collection-level metadata fields in order to provide collection-level services led to the development of The European Library Application Profile for Collection Descriptions [12]. (2) The need for an identifier for retrieving the metadata record from its originating collection was solved by introducing a new term in the TEL namespace, `tel:recordId`. (3) The need for a link to the digital object to permit direct access was addressed by using the existing Dublin Core term `dc:identifier` in combination with internal coding in the portal itself.

Since the original specification, new requirements have been identified that have necessitated the addition of new terms in the TEL namespace. For instance, as part of the TELplus project [13], 20 million pages of OCR'd material will be made available in The European Library. Partner libraries will provide links in their metadata records to this full-text for indexing purposes, requiring the addition of new metadata elements in v2.0 of the TEL-AP for Objects.

The project Europeana v1.0 [14] will develop the Europeana prototype [15], launched in November 2008, into a full operational service. It is anticipated that The European Library will then become the domain-level aggregator for libraries in Europeana. With this in mind, it is intended that the TEL-AP for Objects will evolve further to ensure that it is interoperable with the Europeana metadata format, European Semantic Elements (ESE) [16]. In addition, any new requirements from the wider library community, such as university and research libraries, expected to be the first group of libraries to join the national libraries in The European Library, will be taken into consideration. In addition, The European Library is actively involved in the Accessible Registries of Rights Information and Orphan Works towards Europeana (ARROW) project [17], whose goal is to develop a digital rights infrastructure for Europe. It is anticipated that the TEL-AP for Objects either will be used as an intermediary within the proposed digital rights infrastructure, or will be extended to

---

<sup>3</sup> The European Library (TEL) Project was a 30 month project beginning in 2001, funded under the European Commission's 5th Framework Programme, [http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive%5Ctelproject\\_archive/telproject\\_archive.html](http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive%5Ctelproject_archive/telproject_archive.html).

provide a link to rights information held elsewhere in a dedicated rights metadata format such as in the ONIX framework.<sup>4</sup>

**Acknowledgments.** The authors wish to thank Raffaella Bernardi, Alessio Bosca, Paolo Buoso, Stefan Farrenkopf, Daniele Gobbetti, Stefanie Rühle, Romain Wenz and each other for the stimulating discussions that led to the development of the CACAO APs.

## References

1. Levergood, B., Siciliano, L., Gobbetti, D., Dini, L., Bosca, A., Buoso, P., Barsanti, I.: Integration with [www.theeuropeanlibrary.org](http://www.theeuropeanlibrary.org) and aggregation of partner libraries. CACAO D5.2 (public) (2009)
2. The European Library Metadata Registry (for objects), <http://www.theeuropeanlibrary.org/handbook/regtable.php>
3. Heery, R., Patel, M.: Application profiles: mixing and matching metadata schemas. *Ariadne* 25 (2000), <http://www.ariadne.ac.uk/issue25/app-profiles/>
4. Open Archives Initiative - Protocol for Metadata Harvesting, <http://www.openarchives.org/pmh/>
5. Dublin Core Metadata Initiative, <http://dublincore.org/>
6. Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: Guidelines for Repository Implementers, <http://www.openarchives.org/OAI/2.0/guidelines-repository.htm>
7. Dublin Core Libraries Application Profile, April 2002, <http://dublincore.org/documents/2002/04/16/library-application-profile/>, was used as the basis.
8. Bosca, A., Gobbetti, D.: Fully integrated CLIR system. CACAO D.1.4 (confidential) (2008)
9. Levergood, B., Farrenkopf, S., Frasnelli, E.: The Specification of the Language of the Field and Interoperability: Cross-language Access to Catalogues and Online Libraries (CACAO). In: Greenberg, J., Klas, W. (eds.) *Metadata for Semantic and Social Applications: Proceedings of the International Conference on Dublin Core and Metadata Applications 22-26 September 2008*, pp. 191-196, Universitätsverlag, Göttingen (2008), [http://webdoc.sub.gwdg.de/univerlag/2008/DC\\_proceedings.pdf](http://webdoc.sub.gwdg.de/univerlag/2008/DC_proceedings.pdf)
10. Dewey Decimal Classification, <http://www.oclc.org/dewey/>
11. Minutes of the TELproject (WP3: Metadata Development) meeting, 1 February 2002. (Unpublished)
12. The European Library Application Profile for Collection Descriptions (v1.5), [http://www.theeuropeanlibrary.org/handbook/Metadata/tel\\_ap\\_cld.html](http://www.theeuropeanlibrary.org/handbook/Metadata/tel_ap_cld.html)
13. TELplus project, <http://www.theeuropeanlibrary.org/telplus>
14. Europeana v1.0, <http://version1.europeana.eu/>
15. Europeana, <http://www.europeana.eu/portal/>
16. Specification for the Europeana Semantic Elements (v3.2), [http://www.version1.europeana.eu/web/guest/provide\\_content](http://www.version1.europeana.eu/web/guest/provide_content)
17. ARROW, Accessible Registries of Rights Information and Orphan Works towards Europeana, <http://www.arrow-net.eu/>

---

<sup>4</sup> The ONIX family includes standards for Books, Serials and Licensing Terms, <http://www.editeur.org/8/ONIX/>