The Specification of the Language of the Field and Interoperability: Cross-Language Access to Catalogues and Online Libraries (CACAO)

Barbara Levergood
Goettingen State and University Library,
Germany
levergood@mail.sub.uni-goettingen.de

Stefan Farrenkopf
Goettingen State and University Library,
Germany
farrenkopf@mail.sub.uni-goettingen.de

Elisabeth Frasnelli
Library of the Free University of Bozen-Bolzano, Italy
Elisabeth.Frasnelli@unibz.it

Abstract

The CACAO Project (Cross-language Access to Catalogues and Online Libraries) has been designed to implement natural language processing and cross-language information retrieval techniques to provide cross-language access to information in libraries, a critical issue in the linguistically diverse European Union. This project report addresses two metadata-related challenges for the library community in this context: "false friends" (identical words having different meanings in different languages) and term ambiguity. The possible solutions involve enriching the metadata with attributes specifying language or the source authority file, or associating potential search terms to classes in a classification system. The European Library will evaluate an early implementation of this work in late 2008.

Keywords: Multilingual issues; interoperability; Knowledge Organization Systems (KOS) (e.g., ontologies, taxonomies, and thesauri); normalization and crosswalks

1. Introduction

The European Union (EU) has 23 official languages; many more regional and minority languages are spoken in the 27 member states. A 2006 European Commission/Eurobarometer study revealed that "56% of EU citizens are able to hold a conversation in a language other than their mother tongue", "28% state that they master two languages along with their native language", and "approximately 1 in 10 respondents has sufficient skills to have a conversation in three languages".

In this linguistically diverse and multilingual environment in the EU, there is a tremendous need to provide cross-language access to information (i.e., using one language to find information in another). However, European libraries not only do not share a language, they also have no common subject heading system, classification system, authority files, or bibliographic format. Thus, cross-language access to information in library collections is a complex and difficult problem involving not only natural language analysis and translation, but also the mapping of library subject headings, classifications, and bibliographic formats, presenting problems of both syntactic and semantic interoperability.

The CACAO Project (Cross-language Access to Catalogues and Online Libraries), begun in December 2007, is a 24-month targeted project supported by the eContentplus Programme of the European Commission. It is a consortium of nine partners: Cité des sciences et de l'industrie and Xerox Research Centre Europe from France; the Free University of Bozen-Bolzano, CELI, and Gonetwork from Italy; Kórnik Library from Poland; the National Széchényi Library and the Hungarian Academy of Sciences from Hungary; and Goettingen State and University Library of Germany.

OPEN ACCESS

This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952109327

The libraries in the CACAO consortium use a total of at least six different subject heading systems (Library of Congress Subject Headings, Schlagwortnormdatei, Słownik języka haseł przedmiotowych Biblioteki Narodowej [National Library Subject Headings Authority Files], Soggettario per i cataloghi delle biblioteche italiane, 2 local systems) and five different classification systems (Basisklassifikation, Göttinger Online-Klassifikation, Regensburger Verbundklassifikation, 2 local systems). Three of the libraries are multilingual libraries.

CACAO will modify and extend work that has already been implemented at the Library of the Free University of Bozen-Bolzano, a multilingual library having major collections in Italian, German, and English, each with its own subject heading system, as described by Bernardi et al. (2006).

This report reviews two of the important metadata-related challenges that CACAO faces involving the specification of the language of the metadata fields, "false friends" and term ambiguity, and discusses our solutions. We begin with a short description of the CACAO architecture.

2. CACAO Architecture

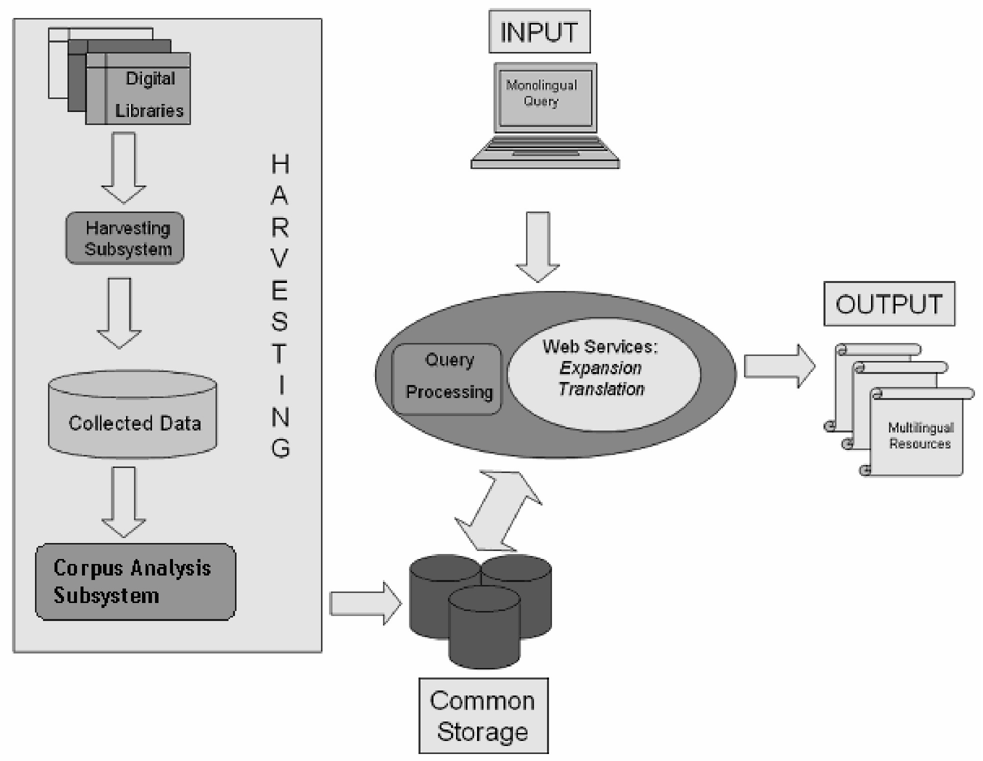

Figure 1 - Architecture Overview (Dini and Bosca (2008), pg. 4)

The CACAO architecture in Figure 1 is designed to support the following vision. A user should be able to enter a monolingual query, say *cat* in English, and retrieve highly relevant records not just in English, but also in any supported language in the database, including records containing, for example, the German word for *cat*, *Katze*, French *chat*, Hungarian *macska*, Italian *gatto*, or Polish *kot*.

As a least-common-denominator solution, CACAO will harvest metadata through library OAI-PMH interfaces, minimally in Dublin Core; MARC 21 may also be accepted if available. The CACAO Corpus Analysis Subsystem performs a variety of analyses on the metadata off-line, the

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952109327

results of which are stored locally and used in support of online Query Processing. When the user enters a query, the Query Processing Subsystem, with the assistance of third-party Web Services providing linguistic analyses, translations, etc., translates and expands the query and matches it against the results of the Corpus Analysis Subsystem. Of course, resources such as lexica, multilingual dictionaries, and thesauri and other controlled vocabularies are accessed by the subsystems.

3. False Friends and Term Ambiguity

We will use a simple example to illustrate some metadata-related problems that arise and some of the possible solutions that we are investigating; these are issues that are challenges not just for CACAO, but also for the library community. Suppose a user enters the query *stove*, wanting to retrieve records containing the English word *stove* or the German translation *Herd*.

USER QUERY: stove

3.1. A Simplistic Solution: Translation

The procedure might seem to be very simple: the Query Processing Subsystem looks up English *stove* in the English-German dictionary, retrieves the German translation *Herd*, and builds a Boolean search query containing those two expressions:

QUERY: stove or Herd

However, this simple query also retrieves false hits containing the English word *herd*:

FALSE HIT: <dc:title>Animal status monitoring and herd management</dc:title>

CORRECT: <dc:title>Herd und Ofen im Mittelalter</dc:title>

CORRECT: <dc:title>The Stove-Top Cook Book</dc:title>

English *herd* and German *Herd* are "false friends", i.e., words in different languages that look similar but that have different meanings. False friends are fairly common, for example English *gift*-German *Gift* ("poison"), English *pain*-French *pain* ("bread"), and English *cane*-Italian *cane* ("dog").

3.2. Solution 2: Enrichment of Metadata

Knowing or being able to determine the language of the terms in a given metadata field increases precision when dealing with false friends. The language would optimally be provided in the metadata itself, as we might find in a German-language catalog which owns the English-language book *The Stove-Top Cook Book* to which German- and English-language subject headings are assigned:

<dc:title xml:lang="en">The Stove-Top Cook Book</dc:title>

<dc:subject xml:lang="de">Herd</dc:subject>

<dc:subject xml:lang="en">Stove</dc:subject>

In the case of a subject term, information about the source of the term in the <dc:subject> field could provide enough information to be able to deduce the language. In this case, we could deduce the language with a fairly high degree of certainty from the fact that the SWD (Schlagwortnormdatei) is a German-language subject heading system:

<dc:subject xsi:type="cacao:SWD">Herd</dc:subject>

This information about the language of the content of the field will be used by CACAO in presenting the ranked results list. Since the German term *Herd* appears in German-language fields in this record:

<dc:title xml:lang="de">Herd und Ofen im Mittelalter</dc:title>

<dc:subject xsi:type="cacao:SWD">Herd</dc:subject>

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952109327

it would be ranked higher than a record in which the false friend of the German translation of the original search term appears in an English-language field. Alternatively, such a record could be excluded entirely from the results list.

```
<dc:title xml:lang="en">Animal status monitoring and herd management</dc:title>
```

3.3. Solution 3: Association to a Class

However, metadata are not always enriched with language or authority attributes as they are in this ideal catalog. CACAO's technical partners are developing a solution for this scenario, the association of terms to a fairly broad class in a library classification system such as the Dewey Decimal Classification (DDC). In our example, the off-line Corpus Analysis Subsystem must have been able to determine that materials about stoves are commonly classed in, e.g., DDC 640 (Home & Family Management), and it has stored this association: stove:DDC 640.

One option would be to organize the results list according to class. For instance, records containing the terms *stove* or *Herd* with a <dc:subject xsi:type="dcterms:DDC"> element having the DDC value provided by the Corpus Analysis Subsystem, 640:

```
<dc:title>Herd und Ofen im Mittelalter</dc:title>
<dc:subject xsi:type="dcterms:DDC">640</dc:subject>
```

would be presented in a group which would be ranked higher than groups of records containing one of those terms with some other DDC value for the <dc:subject> element, including records containing the false friend.

```
<dc:title>Animal status monitoring and herd management</dc:title>
<dc:subject xsi:type="dcterms:DDC">630</dc:subject>
```

4. Association to a Class and Term Ambiguity

The association to a class technique is used in information retrieval and in CACAO for an even more common problem: term ambiguity. The English word *pipe*, for instance, is ambiguous, meaning either "a long tube", German *Rohr*, or "a device for smoking", German *Pfeife*. For purposes of exposition, assume that on entering *pipe* as a search query, the user is asked which meaning is intended and that the user selects the meaning "a long tube". Using the association to a class technique, the Corpus Analysis Subsystem has determined that relevant materials are often classed in DDC 690 (Building & Construction).

Again, one option would be to organize the results list according to class, similar to the *stove/Herd* example. Records containing the terms *pipe* or *Rohr* and including a <dc:subject xsi:type="dcterms:DDC"> element having the DDC value provided by the Corpus Analysis Subsystem, 690:

```
<dc:title>Plumbers and pipe fitters library</dc:title>
<dc:subject xsi:type="dcterms:DDC">690</dc:subject>
```

would be presented in a group which would be ranked higher than groups of records containing one of those terms with some other value for the <dc:subject> element, including records containing the term *pipe* in its unintended meaning:

```
<dc:title>The pleasures of pipe smoking</dc:title>
<dc:subject xsi:type="dcterms:DDC">390</dc:subject>
```

Association to a class can also be used to disambiguate an ambiguous target term. For instance, the English search term *dog* translated into Italian is *cane*. However, Italian *cane* has two senses, "dog" and "cock of a weapon", which would be disambiguated in the same way. Records containing the terms *dog* or *cane* and including a <dc:subject xsi:type="dcterms:DDC"> element having the DDC value 630 (Agriculture) would be presented in a group which would be ranked higher than groups of records containing one of those terms with some other value for the <dc:subject> element, including records containing the term *cane* in its unintended meaning.

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952109327

5. Conclusion

We have argued that the specification of the language of the metadata field, in addition to that of the document itself, is very important so that metadata can be fully exploited for cross-language purposes or in multilingual settings.

If the metadata do not come with or cannot be enriched with the languages of the fields, then CACAO must rely on the association to a class technique, which will be needed in any case. Association to a class was originally designed for and will be used as a solution to the term ambiguity problem; it is similar to synsets used in WordNet and EuroWordNet, which CACAO may also use. The solution involving association to a class may also work as association to a subject heading, although that would require further preparation and testing.

It is important to note that in the association to a class technique, the CACAO Corpus Analysis Subsystem must be able to associate a term such as English *stove* to some class and then the system must be able to match potential hits containing a term such as *Herd* against that same class. In other words, either the systems must contain the same classification system or their classification or subject headings systems must be mappable to the same system. Thus, CACAO's experience with cross-language access so far strongly supports Koch, Neuroth, and Day (2001); NKOS (2001); Chan and Zeng (2002); Harper and Tillett (2007); and many others in the library community who have discussed the importance of the interoperability of subject vocabularies and of classification systems for information retrieval in cross-domain environments. CACAO will rely on already existing mappings such as those provided by the MACS project (Landry (2004, 2006)), which has worked on mappings for RAMEAU (Bibliothèque nationale de France), Library of Congress Subject Headings (British Library), and Schlagwortnormdatei (Deutsche Nationalbibliothek and Bibliothèque nationale suisse).

For optimal performance, even if the metadata of a given collection does not contain the specification for the language of the field as outlined in section 3.2, the Corpus Analysis Subsystem must still have access to such enriched metadata in order to avoid the false friends problem in its off-line analyses. For instance, if the Corpus Analysis Subsystem must determine which class German *Gift* "poison" is most commonly associated with, then it should avoid analyzing fields in which the English *gift* is found. However, we anticipate that the Corpus Analysis Subsystem will have access to a more extensive stored collection of associations between terms and classes than might be available for a given collection.

A prototype of the CACAO information retrieval system was entered in the CLEF 2008 campaign, providing an opportunity to tune and evaluate the system on cross-language library metadata. CACAO's attention will soon turn to related issues involving metadata exchange and interoperability and thereby further explore the characteristics of Dublin Core in its cross-language duties. The European Library, whose Application Profile is Dublin Core-based, will integrate and evaluate CACAO technologies beginning in late 2008. Furthermore, CACAO libraries will be grouped into a single portal and CACAO will additionally create several thematic portals in order to further develop, demonstrate, and promote CACAO technologies.

Acknowledgements

We would especially like to thank Raffaella Bernardi, Jane Greenberg, Tom Baker, and three anonymous reviewers for their careful comments on earlier versions of this report. Of course, all errors are our own. CACAO is funded under the eContentplus programme, ECP 2006 DILI 510035 CACAO Program eContentplus.

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952109327

References

Bernardi, Raffaella, Diego Calvanese, Luca Dini, Vittorio Di Tomaso, Elisabeth Frasnelli, and Ulrike Kugler et.al. (2006). Multilingual search in libraries. The case-study of the Free University of Bozen-Bolzano. *Proc. 5th International Conference on Language Resources and Evaluation - LREC 2006, Genova*. Retrieved, 3 April, 2008 from http://www.inf.unibz.it/~bernardi/index.php?page=pub.

CACAO: Cross-language Access to Catalogues and Online Libraries. (2007). *Annex 1: Description of Work.* 17 November 2007.

CACAO: Cross-language Access to Catalogues and Online Libraries. (2008). *CACAO Project.* Retrieved, April 10, 2008, from http://www.cacaoproject.eu/.

Chan, Lois Mai, and Marcia Lei Zeng. (2002). Ensuring interoperability among subject vocabularies and knowledge organization schemes: A methodological analysis. *68th IFLA Council and General Conference, 18-24 August 2002, Glasgow.* Retrieved, April 10, 2008, from http://www.ifla.org/IV/ifla68/papers/008-122e.pdf.

Dini, Luca, and Alessio Bosca. (2008). *Definition of programmatic interfaces for accessing data storage in digital libraries, e-catalogues and OPAC.* CACAO Deliverable D.3.1.

European Commission. (2008). *The official EU languages.* Retrieved, April 3, 2008, from http://ec.europa.eu/education/policies/lang/languages/index_en.html.

European Commission. Eurobarometer. (2006). *Europeans and their Languages* (p. 8). Retrieved, April 3, 2008, from http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_en.pdf.

Free University of Bozen-Bolzano. (2007) *Multilingual Search.* Retrieved, April 10, 2008, from http://pro.unibz.it/opacdocdigger/index.asp?MLSearch=TRUE.

Harper, Corey A., and Barbara B. Tillett. (2007). Library of Congress controlled vocabularies and their application to the Semantic Web. *Cataloging & Classification Quarterly 43*(3/4), 47-68.

Jurafsky, Daniel, and James H. Martin. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Koch, Traugott, Heike Neuroth, and Michael Day. (2001). Renardus: Cross-browsing European subject gateways via a common classification system (DDC). *IFLA Satellite Meeting on Classification and Indexing, 14-16 August 2001, Dublin, Ohio, USA.*

Koninklijke Bibliotheek. (2005-2008). *The European Library*. Retrieved, May 27, 2008, from http://www.theeuropeanlibrary.org/.

Landry, Patrice. (2004). Multilingual subject access: The linking approach of MACS. *Cataloging & Classification Quarterly 34*(3/4), 177-191.

Landry, Patrice. (2006). Multilinguisme et langages documentaires: le projet MACS en contexte européen. *Documentation et Bibliothèques 52*(2), 121-129.

Networked Knowledge Organization Systems (NKOS). (2001). Classification crosswalks: Bringing communities together. *The 4th NKOS Workshop at ACM-IEEE Joint Conference on Digital Libraries (JCDL), 28 June 2001, Roanoke, Virginia, USA.* Retrieved, April 10, 2008, from http://nkos.slis.kent.edu/DL01workshop.htm.

OCLC. (2008). *Dewey Services: Dewey Decimal Classification.* Retrieved, April 10, 2008, from http://www.oclc.org/dewey/.

Princeton University. Cognitive Science Laboratory. *WordNet.* Retrieved, May 28, 2008, from http://wordnet.princeton.edu/.

Roux, Claude. (2008). *User Requirements.* CACAO Deliverable D.7.1.

TrebleCLEF Coordination Action. *The Cross-Language Evaluation Forum (CLEF).* Retrieved, May 27, 2008, from http://www.clef-campaign.org/.

University of Amsterdam. *EuroWordNet.* Retrieved, May 28, 2008, from http://www.illc.uva.nl/EuroWordNet/.

OPEN ACCESS

This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952109327