

The CACAO project: a multilingual interface to Library Catalogues

(Extended Abstract)

Alessio Bosca, Claude Roux
CELI
Xerox Research Centre Europe
{*alessio.bosca@celi.it, claude.roux@xerox.xrce.com*}

Introduction

Managing the development and delivery of multilingual electronic library services is one of the major current challenges for making digital content in Europe more accessible, usable and exploitable. Digital libraries and OPAC-based traditional libraries are the most important source of reliable information, daily used by scholars, researchers, knowledge workers and citizens to conduct their working (and leisure) activities. Facilitating access to multilingual document collections therefore is an important way of supporting the dissemination of knowledge and cultural content.

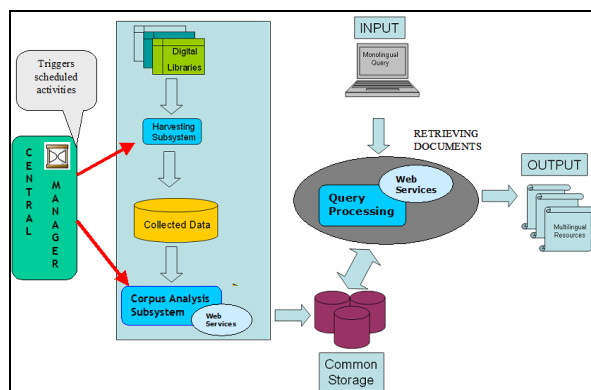
CACAO (Cross-language Access to Catalogues And On-line libraries) project proposes an innovative approach for accessing, understanding and navigating multilingual textual content in digital libraries and OPACs, enabling European users to better exploit the available European electronic content at their disposal. By coupling sound Natural Language Processing techniques with available information retrieval systems the project aims at the delivery of a non-intrusive infrastructure to be integrated with current OPAC and digital libraries. The result of such integration will be the possibility for the user to type in queries in his/her own language and retrieve volumes and documents in any available language.

CACAO aims at offering cross-lingual and cross-border access to the content of classical and digital libraries and enabling users to find digital content irrespective of the language. In fact, in a context of interlaced cross-border libraries, such as the one proposed by META OPAC, the absence of a cross-language perspective is likely to cause a substantial impasse: if a user wanted to access a META OPAC including the National Libraries of France, Germany, Italy, Poland and Hungary, s/he would have to type five queries in five different languages. Much of the advantage of having a unique access point is thus lost. CACAO project proposes a system based on the assumptions that users look more and more at library contents using free keyword queries (as those used with a web search engine) rather than more traditional library-oriented access (e.g. via Subject Heading);

therefore, the only way to face the cross-language issue is by translating the query into all languages covered by the library/collection (rather than, for instance, translating subject headings). The system will then yield results in all desired languages.

Validation is another important aspect in the project: all CACAO core technologies are indeed proven, but they have never been massively deployed in the field of digital libraries. CACAO aims at crossing the chasm between sound innovation and adoption by library institutions for real life purposes.

Architecture overview



CACAO proposes the development of an infrastructure for multilingual access to digital content, including an information retrieval system able to search for books and texts in all the available languages. The core of the search engine takes advantage of information contained in existing catalogues and texts of the digital libraries that is enriched by means of NLP techniques such as word sense disambiguation and named entities recognition. The goal of such integration is to avoid confusing the user by providing irrelevant results due to bad translations and thus enabling a better access to the digital content.

The general architecture of the Cacao system could be summarized as the result of the interactions of few functional subsystems, coordinated by a central

manager and reacting to external stimuli represented by end users queries:

- **Harvesting** subsystem is in charge of collecting data from digital libraries, abstracting from the multiplicity of standards and protocols, and storing them into a repository.
- **Corpus Analysis** subsystem performs specific analysis on the data collected from libraries and infers new information used to support query processing and resource retrieval (e.g. query expansion, terms disambiguation,...).
- **Web Services** subsystem represents third party software providing specific services (e.g. linguistic analysis, translations,...).
- **Query Processing** subsystem: a set of components is devoted to process the original monolingual user query, transforming and enriching it by means of translations and expansions.

Content Enrichment

CACAO approach to multilingual access is based on the integration of a standard IR engine with multilingual thesauri and multilingual lexicons. However a simple, direct integration would provide poor results since records of digital catalogues often contain only small portions of text and the noise brought in by the query translation layer further worsen the situation. Therefore any single fragment of text needs to be linguistically "enriched" in order to guarantee an optimal retrieval.

The strategy adopted by CACAO with respect to content enrichment aims at integrating the search indexes used by the IR system rather than the original records from the libraries; such enrichment is operated by adopting the following technologies:

1. Enrichment of the query via thesauri: the simple query "*plants*" could be enriched by synonyms/hyperonyms/hyponyms such as horticultural, seeds etc. This would allow books such as "*American Horticultural Society Encyclopaedia of Plants and Flowers*" by Christopher Brickell or "*From Seed to Plant*" by Gail Gibbons to receive more emphasis than "*The Parachute Plant*", a thriller by Bill Carrigan.
2. Enrichment of the query via corpus-based expansion lists: from the point of view of the user, this technology is the same as the previous one. The only difference is that such related terms are induced on the basis of the catalogue rather than being stored into a static repository.
3. Tagging of the text in DB records by using a part of speech tagger (i.e. disambiguating the syntactic category of words). As simple as it might seem, this enrichment will allow the system to avoid

retrieving a title such as "*Plant Them Deep*" by Aimee Thurlo, David Thurlo with a query such as *plant*.

Improving Translations

Within CACAO project another aspect where contents enrichment has a strong impact is the improvement of linguistic resources and in particular translation dictionaries. CACAO system is based on translation dictionaries; however, there is probably no single translation dictionary that would be able to cover all digital content either in a library catalogue or in the texts of a digital library.

A first strategy to be adopted in order to compensate for possible lack of translation coverage is query expansion.

The second, probably more innovative approach is based on user input. An analysis of the web logs of a university library, shows that about 40% of the queries are "duplicated" in at least two languages. Indeed, if we could store the translations implicitly provided by the user, we could add items which are I) relevant to users; II) reflecting users' perception of the translation of a given word in a different language.

From a technical point of view, this approach raises some major challenges. Indeed, the fact that two queries issued by the same user are temporally adjacent is not necessarily proof that the second is a translation of the first. Therefore, it is important to set up methodologies to isolate possible translation pairs. In order to detect these cases CACAO system exploits a method based on semantic web vectors. The basic idea is to gather from the web (via queries to search engines) a set of documents strongly related to the original term (**st**).

By using standard NLP technologies, these documents are analyzed, and the terminological items are extracted (let **ST** be these terms). The same operation is performed on the candidate translation (**tt**), thus generating a set **TT** of words in the target language. **ST** is then translated, using the available resources, into a set of translated target words (**STT**). By measuring the intersection between **STT** and **TT**, the system will be able to predict the likeliness of **tt** to be a translation of **st**.