

# CACAO: Cross language Access to Catalogues and Online libraries



Die Bibliothek  
La Biblioteca  
The Library



FREE UNIVERSITÄT BOZEN  
LIBERA UNIVERSITÀ DI BOLZANO  
FREE UNIVERSITY OF BOZEN - BOLZANO

Luigi Siciliano<sup>a</sup>, Paolo Buoso<sup>a</sup>,

Daniele Gobbetti<sup>b</sup>, Raffaella Bernardi<sup>b</sup>

<sup>a</sup> Libera Università di Bolzano, Biblioteca universitaria

<sup>b</sup> Libera Università di Bolzano, Centro di ricerca KRDB



## Obiettivi

Il rapporto Eurobarometer "Europeans and Languages" (2005) mostra come in media metà degli europei ritenga di padroneggiare un'altra lingua oltre alla lingua madre ed evidenzia quanto al diminuire dell'età questa capacità sia progressivamente più marcata. In questo contesto sempre più multiculturale e multilingue emerge pertanto la necessità di sviluppare strumenti nuovi, che permettano una ricerca multilingue in maniera trasparente.

Per le biblioteche la Babele europea non si limita alla varietà linguistica ma anche alle pratiche di catalogazione, spesso derivata, ai sistemi di classificazione e di soggettazione, alle liste di autorità e ai formati bibliografici. Il problema della ricerca multilingue risulta quindi di particolare complessità e coinvolge non solo l'analisi e la traduzione automatica della lingua ma anche il mapping tra i sistemi di classificazione e l'interoperabilità dei metadati descrittivi. A questi problemi cerca di dare una risposta CACAO (Cross language Access to Catalogues and Online Libraries), progetto biennale nell'ambito di eContentplus, programma dell'Unione Europea indirizzato al miglioramento delle modalità di accesso alle risorse digitali.

L'obiettivo è la realizzazione di una piattaforma che permetta all'utente di formulare richieste in linguaggio naturale nella propria lingua e ottenere da cataloghi e biblioteche digitali risorse pertinenti in tutte le lingue disponibili, indipendentemente dalla lingua del sistema di indicizzazione semantica adottato. Attualmente si sta lavorando su francese, italiano, inglese, ungherese, polacco e tedesco.

Keyword: fragola

**Title:** Compendium of strawberry diseases  
**Author statement:** ed. by J. L. Maas  
**Persons:** Maas, J. L.  
**Series:** The disease compendium series of the American Phytopathological Society  
**Pages (vol.):** VI, 98 S. : Ill.  
**Edition:** 2. ed.  
**ISBN:** 0-89054-194-9  
**Imprint:** St. Paul, Minn. : APS Pr., 1998  
**Classification:** ZC 55450|ZC 25000

SUBJECTS  
Strawberries / Diseases and pests

**Title:** Erdbeeren ökologisch angebaut  
**Author statement:** Andi Schmid  
**Note:** Umschlagt.  
**1. Author:** Schmid, Andi  
**Series:** Praxis des Ökolandbaus  
**Pages (vol.):** 19 S. : Ill., graph. Darst.  
**Edition:** 1. Aufl.  
**Imprint:** 3-934239-13-7  
Mainz : Bioland Verl.- GmbH, 2003  
**Classification:** ZC 55450

SUBJECTS  
Erdbeeranbau / Biologischer Obstbau

## Partecipanti

### Aziende

- Xerox Research Centre Europe (coordinatore del progetto) (Grenoble, Francia)
- CELI (Torino, Italia)
- Gonetwork (Viareggio, Italia)

### Centri di Ricerca

- Centro di Ricerca KRDB (Knowledge Representation meets Data Bases), Facoltà di Scienze e Tecnologie informatiche, Libera Università di Bolzano (Bolzano, Italia)
- Hungarian Academy of Science - Research Institute for Linguistics (Budapest, Ungheria)

### Biblioteche

- Cité des sciences et de l'industrie (Parigi, Francia)
- Kórnik Library - Polish Academy of Sciences (Kornik, Polonia)
- Göttingen State and University Library (Göttingen, Germania)
- National Széchenyi Library, MEK Hungarian Electronic Library (Budapest, Ungheria)
- Biblioteca Universitaria, Libera Università di Bolzano (Bolzano, Italia)

## Metodo

CACAO applica al mondo degli OPAC tecnologie di traduzione automatica, Natural Language Processing (la richiesta viene interpretata e arricchita) e Information Retrieval (i risultati sono identificati ed elencati in base alla pertinenza) già testate in altri ambiti dai partner tecnologici. L'integrazione degli eterogenei cataloghi delle biblioteche non richiede alcuna modifica degli standard di catalogazione in uso, poiché la condivisione dei record bibliografici avviene tramite set di metadati Dublin Core esposti via OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting).

## Tempi di realizzazione

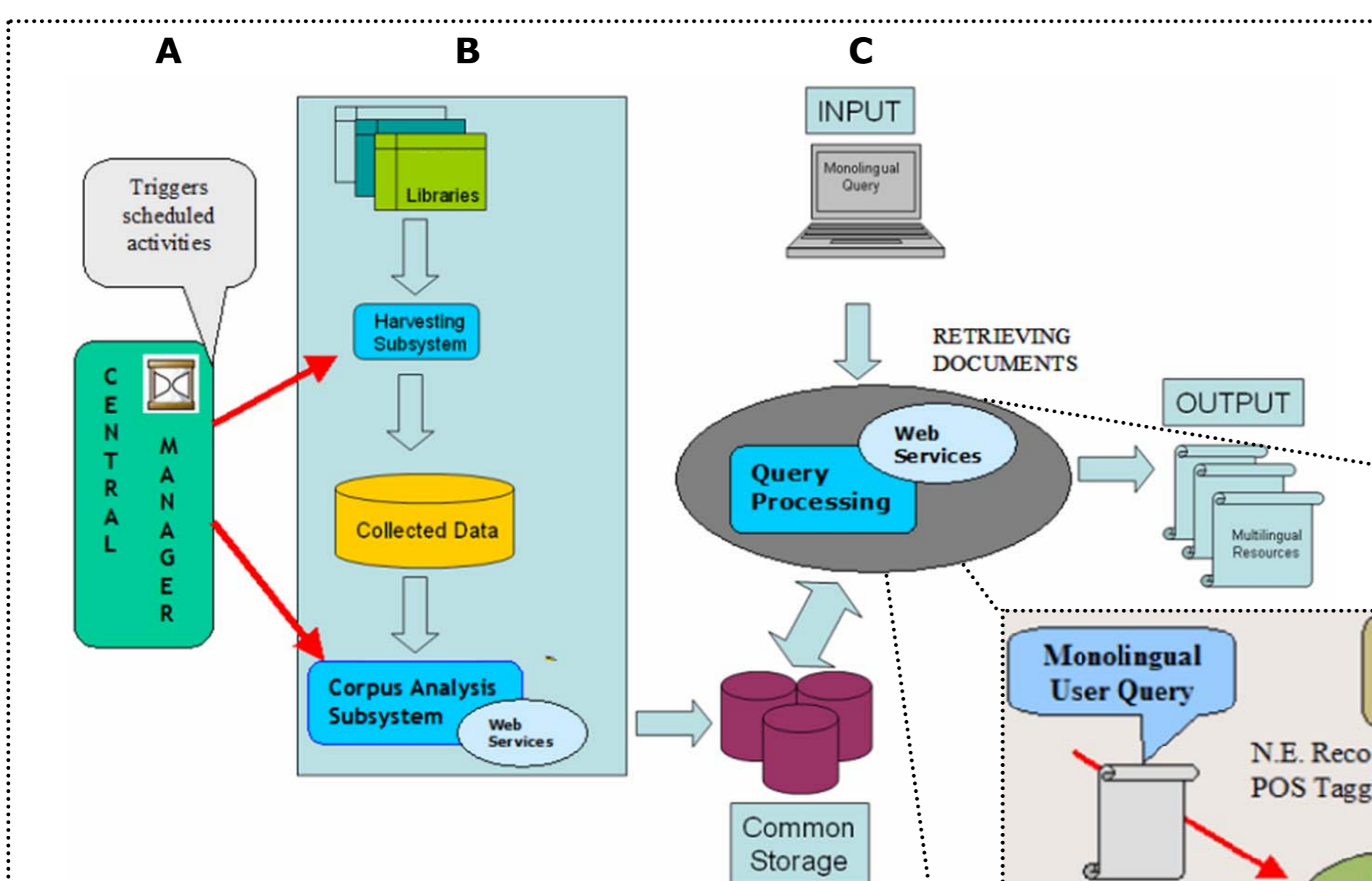
Il progetto è iniziato nel dicembre 2007 e terminerà nel dicembre 2009.

## Stato di avanzamento dei lavori

Alcuni risultati sono già stati raggiunti:

- tutte le biblioteche stanno implementando un server OAI-PMH.
- è stata prodotta la beta release di un DC Application Profile compatibile con il *The European Library Application Profile*.

- Bolzano e Göttingen hanno realizzato un mapping per esporre i propri records bibliografici in formato Dublin Core (DC):  
a) DC simple b) DC qualified.



## Architettura generale

Per garantire la massima modularità del sistema tutti i componenti principali sono *web services*.

- A) Central Manager:** è il pannello di controllo che, tramite una interfaccia web, permette alla biblioteca di gestire la piattaforma.
- B) Le Biblioteche:** forniscono i dati tramite data provider OAI-PMH. I dati raccolti sono salvati in un indice.
- C) L'utente:** formula una query in linguaggio naturale nella propria lingua. La query viene tradotta e arricchita. Il motore di ricerca interroga l'indice e restituisce i risultati nelle diverse lingue.

## Trattamento della query

La query viene innanzi tutto sottoposta al processo di riconoscimento dei nomi propri, alla lemmatizzazione e all'identificazione delle diverse parti del discorso (Named Entity Recognition and Part of Speech Tagging). Dopo tale trattamento la query è pronta per essere tradotta automaticamente. La query stessa viene poi arricchita aggiungendo ai termini della traduzione termini correlati, desunti dai soggetti e dai thesauri disponibili. Tutta l'informazione disponibile (termini originali, termini lemmatizzati, traduzioni, termini correlati) viene infine inoltrata al motore di ricerca, che si incarica di individuare e restituire i dati pertinenti nelle diverse lingue.

## Nuovi ruoli per la classificazione: word to category (W2C)

Nell'interpretazione della query dell'utente, le tradizionali pratiche di soggettazione e classificazione delle risorse trovano un nuovo ruolo nell'interpretazione dei significati e nella traduzione di termini polisemici. La classificazione infatti, grazie alla notazione espressa in caratteri alfanumerici convenzionali, può essere usata in maniera innovativa per il corretto trattamento di termini polisemici nella lingua tradotta. In un contesto per esempio di Classificazione Decimale Dewey (DDC), pensiamo a una ricerca del termine italiano "banca", che intenda ottenere risultati anche in lingua inglese. La traduzione sarà "bank", termine che in inglese significa sia banca che argine o sponda. In tal caso il sistema confronterà la classificazione assegnata al termine banca in italiano (DDC 332.1) con quelli assegnati al termine bank in inglese (sia DDC 332.1, sia DDC 627.133). L'ambito semantico corretto sarà quello desunto dall'intersezione delle notazioni. Nella frequente ipotesi di corrispondenza non completa fra le notazioni, il sistema risale nella struttura della classificazione, cercando un'intersezione fra le notazioni ai livelli gerarchicamente più alti.

Keyword: Banca

Estrazione della notazione  
Banca (DDC 332.1)

Traduzione automatica  
Bank

Disambiguazione

Bank  
DDC 332 Financial economics  
DDC 332.1 Banks

Bank  
DDC 627 Hydraulic engineering  
~~DDC 627.133 Bank protection and reinforcement~~

## CACAO in Europa

Già nel corso del secondo anno di progetto è prevista l'adozione della piattaforma da parte di The European Library ([www.theeuropeanlibrary.org](http://www.theeuropeanlibrary.org)) e la realizzazione di tre portali tematici dedicati rispettivamente alla Storia europea, alla Matematica e alla Geografia.

La **Biblioteca Universitaria di Bolzano** opera da tempo in un contesto multiculturale e plurilingue, si confronta con standard catalografici nazionali diversi e ha precocemente sperimentato, con MuSiL (Multilingual Search in Libraries), l'implementazione di un OPAC multilingue sviluppato in collaborazione con il KRDB e il CELI di Torino.

## Riferimenti principali

- Raffaella Bernardi, Diego Calvanese, Luca Dini, Vittorio Di Tomaso, Elisabeth Frasnelli, Ulrike Kugler, Barbara Planck (2006), *Multilingual search in libraries. The case-study of the Free University of Bozen-Bolzano*, Proceedings 5th International Conference on Language Resources and Evaluation - LREC 2006, Genova.
- Luca Dini, Alessio Bosca (2008), *Definition of programmatic interfaces for accessing data storage in digital libraries, e-catalogues and OPAC*, CACAO Deliverable D.3.1 <[http://www.cacao-project.eu/fileadmin/media/Deliverables/CACAO\\_D3.1.pdf](http://www.cacao-project.eu/fileadmin/media/Deliverables/CACAO_D3.1.pdf)> (14/10/2008).
- European Commission (2006), *Europeans and their Languages*, Special Eurobarometer 243 <[http://ec.europa.eu/public\\_opinion/archives/ebs/ebs\\_243\\_en.pdf](http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_en.pdf)> (14/10/2008).
- Barbara Levergood, Stefan Farrenkopf, Elisabeth Frasnelli (2008), *The Specification of the Language of the Field and Interoperability: Cross-Language Access to Catalogues and Online Libraries (CACAO)*, Proceedings of the International Conference on Dublin Core and Metadata Applications, DC 2008, Berlin.
- Claude Roux (2008), *User Requirements*. CACAO Deliverable D.7.1 <[http://www.cacao-project.eu/fileadmin/media/Deliverables/CACAO\\_D7.1.pdf](http://www.cacao-project.eu/fileadmin/media/Deliverables/CACAO_D7.1.pdf)> (14/10/2008).

## Sito web del progetto

<http://www.cacao-project.eu/>

## Contatti

Luigi Siciliano, [luigi.siciliano@unibz.it](mailto:luigi.siciliano@unibz.it)  
Paolo Buoso, [paolo.buoso@unibz.it](mailto:paolo.buoso@unibz.it)  
Daniele Gobbetti, [gobbetti@inf.unibz.it](mailto:gobbetti@inf.unibz.it)  
Raffaella Bernardi, [bernardi@inf.unibz.it](mailto:bernardi@inf.unibz.it)