

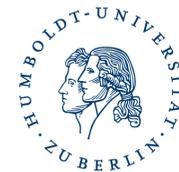


europeana

connect

WP2 – Objectives and Scope of Multilingual Aspects

Vivien Petras, HUB
Anne Schiller, XEROX
Berlin, 26 January 2010



coordinated by the [Austrian National Library](#)  Österreichische
Nationalbibliothek

WP2: Multilingual infrastructure and tools

- **2.1 user preferences** for multilingual access (D2.1.1)
- **2.2 repository of language resources**
- **2.3 multilingual mapping** of controlled vocabularies
- **2.4 translation services** for querying
- **2.5 user behavior & retrieval evaluation** of the services

WP2 – Multilingual Access to Content

- Make Europeana talk European...
 - Multilingual search (query translation)
 - Multilingual browsing (mapping multilingual KOS)

- Core language set

French	English	Spanish
German	Italian	Polish

- Secondary language set

Dutch	Portugese
Hungarian	Swedish

Language Resources Repository

- Focus task on query translation module
- Language resources for:
 - Language detection, stop words, decomposing, lemmatization, named entity detection, phrase detection, dictionaries
- similar resources necessary for vocabulary mapping
 - Wordnet etc. ?
- Licensed resources \leftrightarrow open source resources
- Specification for evaluation / interfacing / maintenance
- Summer 2010: Suggestions for OS language resources

Query Translation

- Dynamic at query-time
- Based on CACAO approach
- Query analysis
- Term translation
- Disambiguation of candidate translations

CACAO User Interface

Demonstrator:

- <http://demo.celi.it:8095/advancedUI/>

Collection:

- CACAO Partner Libraries, TEL
- 5 million metadata records in index

Features:

- Input:
 - Query string in one language
- Output:
 - Search results organized by facets
 - User profile to define settings
 - Refinements of search strings and translations

CACAO User Interface: Query

[Home](#) | [Partners](#) | [Help](#) | [Admin](#) You are connected as cacao [Bookmarks](#) [Settings](#) [Logout](#) | Languages **English** ▾



Cross-Language Access to Catalogues and Online Libraries

[Advanced Search](#) [Virtual Keyboard](#)

Language Search	Language Results	Advanced Options
English <input checked="" type="radio"/>	English <input checked="" type="checkbox"/>	SV expansion <input type="checkbox"/>
German <input type="radio"/>	German <input type="checkbox"/>	WN expansion <input checked="" type="checkbox"/>
French <input type="radio"/>	French <input type="checkbox"/>	W2C disambiguation <input checked="" type="checkbox"/>
Hungarian <input type="radio"/>	Hungarian <input type="checkbox"/>	
Italian <input type="radio"/>	Italian <input type="checkbox"/>	
Polish <input type="radio"/>	Polish <input type="checkbox"/>	

[CACAO searches in any available language across the catalogues of partner libraries](#)

Version 0.1.8

User can select the language in which her search is expressed, which language she would see the results in, as well as some advanced options

CACAO User Interface: Results

Home | Help Bookmark | Settings

CACAO PROJECT
CROSS-LANGUAGE ACCESS TO CATALOGUES AND ONLINE LIBRARIES

New search in English

Searching for **peace** - Total documents retrieved: 3610 [Original Query](#)

Translated terms: **paix(fr) - tranquillité(fr) - peace(en)** - [Translation Options](#)

Add facet

libraryID	language	date	publisher
Bozen University Library	ax	0000-00-00	Amsterdam
Centre de Recherche en Histoire des Sciences et des Techniques	cs	1677	Ann Arbor, Mich. :
Cité des sciences et de l'industrie	de	1897	Humanities Text Initiative
Goettingen State and University Library	en	1923	Baltimore [u.a.] : Johns Hopkins Univ. Press
	es	1969	Basingstoke [u.a.] : Macmillan
	fr	More	

Results 1 - 5 of about 3610 << < / 2 3 4 5 6 7 8 9 10 >> Number of documents per page 5

Aimable paix douce tranquillité / by [André-Marie AMPERE] Publisher: Christine BLONDEL Library: Centre de Recherche en Histoire des Sciences et des Techniques Languages: [fr] Authors: [André-Marie AMPERE] Languages: [fr] Library: Centre de Recherche en Histoire des Sciences et des Techniques Publisher: Christine BLONDEL Table of Contents: Subjects: [Ampère, paix, tranquillité] ISSN: ISBN: URL: http://www.ampere.cnrs.fr/ice/ice_book_detail-fr-text-ampere-ampere_text-163-2.html OAI Sets: TEL URL:	<p>peace paix carte international Paris Europe guerre map conflict woman France dernier ed. Hungary traité world être > Daniel Institute Research SIPRI Stockholm culture nouveau traiter East Glacier International Park United Nations World feminist force gender prince province royaume study violence état Autriche Cobden David Espagne Germany Hongrie Richard agenda build catholique conclure empereur empire late make monde mondial nation perspective plan security temps year & & A. Act Alberta Anville Army Arthur Barth British Campo Civlis Colombie Conference District English Evêque Flandre Formio French Géographi H. Hadtörténelmi Isván Kant Képeskönyve Les> Levéltár Lodgaard Louis Luxembourg M. M.P. May Mike Mister</p> <p>Content All</p>
La grande encyclopédie de la paix / by [Bournier, Isabelle, Pottier, Marc] Publisher: Bruxelles: Casterman Library: Cité des sciences et de l'industrie Languages: [fr]	
Paix / by [Fauchille, Paul] Publisher: Library: Goettingen State and University Library Languages: [fr]	

CACAO User Interface: Translations

The screenshot displays the CACAO Project interface. At the top, there is a search bar with the text "New search" followed by an input field, a language dropdown menu set to "English", and a "Search" button. Below the search bar, the results for the query "peace" are shown, indicating that 3610 documents were retrieved. The interface lists translated terms: "paix(fr) - tranquillité(fr) - peace(en)".

Key interactive elements are highlighted with orange boxes:

- A box containing "Remove term from query" and "Add term to query" buttons.
- A box containing "Suggest a new translation" and "Report a wrong translation" buttons.

Below these buttons, a table shows the current state of the query:

Terms	Action
peace	Remove from query

Two dashed boxes provide user instructions:

- "User can act on the query adding or removing terms" (connected to the first orange box).
- "User can provide feedback to improve overall translation quality" (connected to the second orange box).

CACAO Administrator Interface

The screenshot displays the CACAO Administrator Interface. At the top, there are navigation links: Home, Partners, Help, Admin, and a user connection status 'You are connected as cacao'. On the right, there are links for Bookmarks, Settings, Logout, and a language dropdown menu set to English. Below this is the 'Advanced UI Admin Page' header. A horizontal menu contains tabs for Logs, Database, Web Services, Users, Translations (which is highlighted with an orange box), and Bookmarks. Under the 'Translations' tab, there is a 'Show 10 entries' dropdown and a search field. A table lists translation entries with columns for Source language, Source word, Target language, Target word, Status, and Action. The first entry shows 'fr' source language, 'enfant' source word, 'en' target language, 'baby' target word, and a status of 'REQUEST_FOR_APPROVAL' with 'Accept' and 'Delete' actions. The second entry shows 'fr' source language, 'enfant' source word, 'en' target language, 'kid' target word, and a status of 'APPROVED' with a 'Delete' action. Below the table, it says 'Showing 1 to 2 of 2 entries' and 'First Previous 1 Next Last'. The version number 'Version 0.1.8' is visible in the bottom right corner.

Source language	Source word	Target language	Target word	Status	Action
fr	enfant	en	baby	REQUEST_FOR_APPROVAL	Accept Delete
fr	enfant	en	kid	APPROVED	Delete

An administration screen allows authorized users to perform tasks such as: managing translation request, users, and some configuration parameters

CACAO Metadata: aspects of multilinguality

Multilinguality at different levels:

- Aggregation of libraries
- Library
- Data records
- Data fields

Inherently Multilingual Content

Example: abridged metadata record

Author: Shakespeare, William [?]

Title: Romeo and Juliet [eng]

Subjects: Theaterstück, Selbstmord, Liebe, Drama [ger]

Subject (Geo): Verona, Italien [ger]

- Language information at „class level“ necessary
- commonly, this information is not available

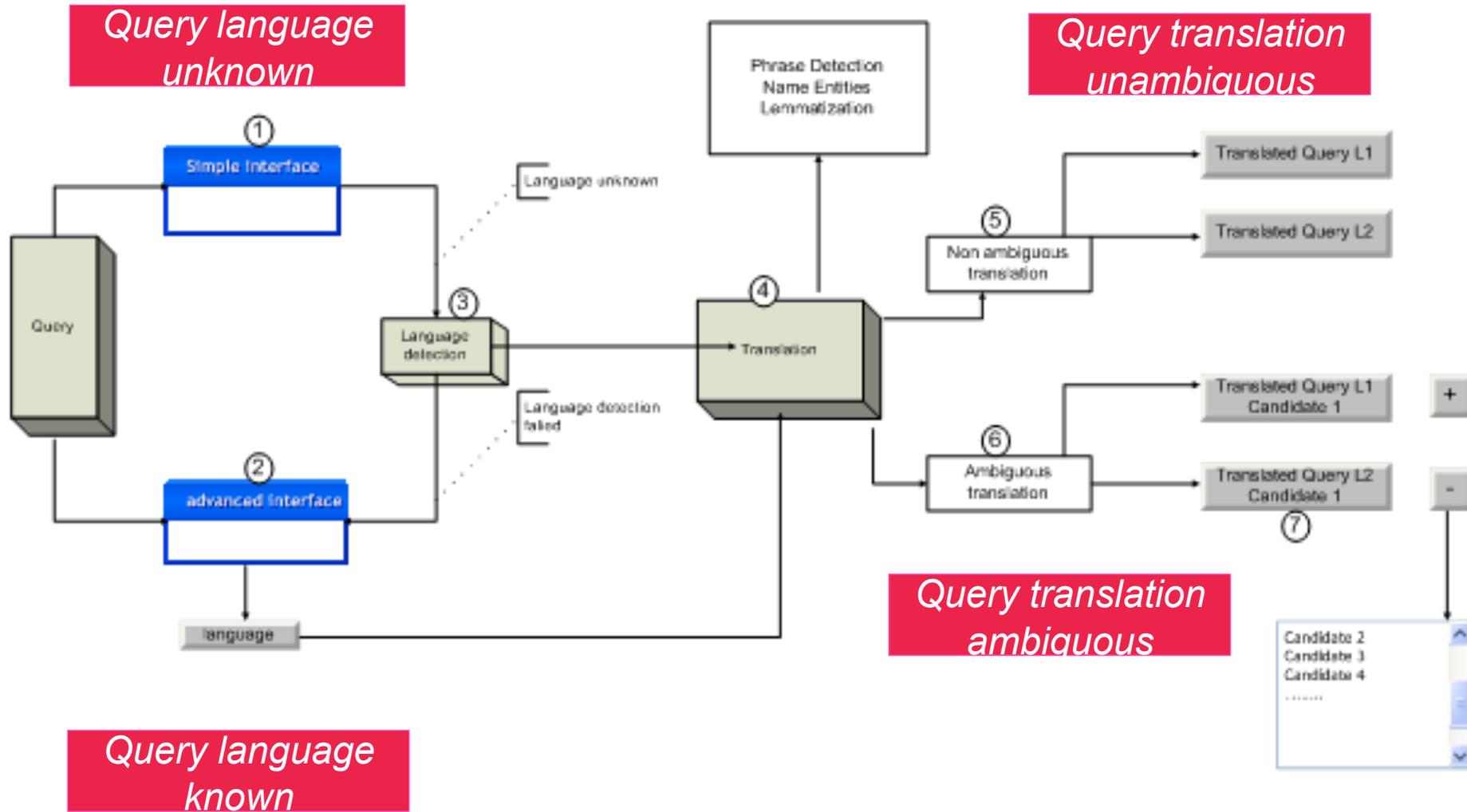
CACAO Application Profile

*As specified in the "CACAO xml:lang status" [...] CACAO recommends that the **language of the metadata** be accurately specified for all elements which contain semantically important text using the attribute `xml:lang`.*

If the language is not specified, the following applies:

- If a default language is set in the general properties, the default language specification will be used.
- Otherwise, a language guesser will be used. The precision of the language guesser may vary according the language and the length of text. However, CACAO cannot guarantee a precision higher than 90%.

General Query Translation Process

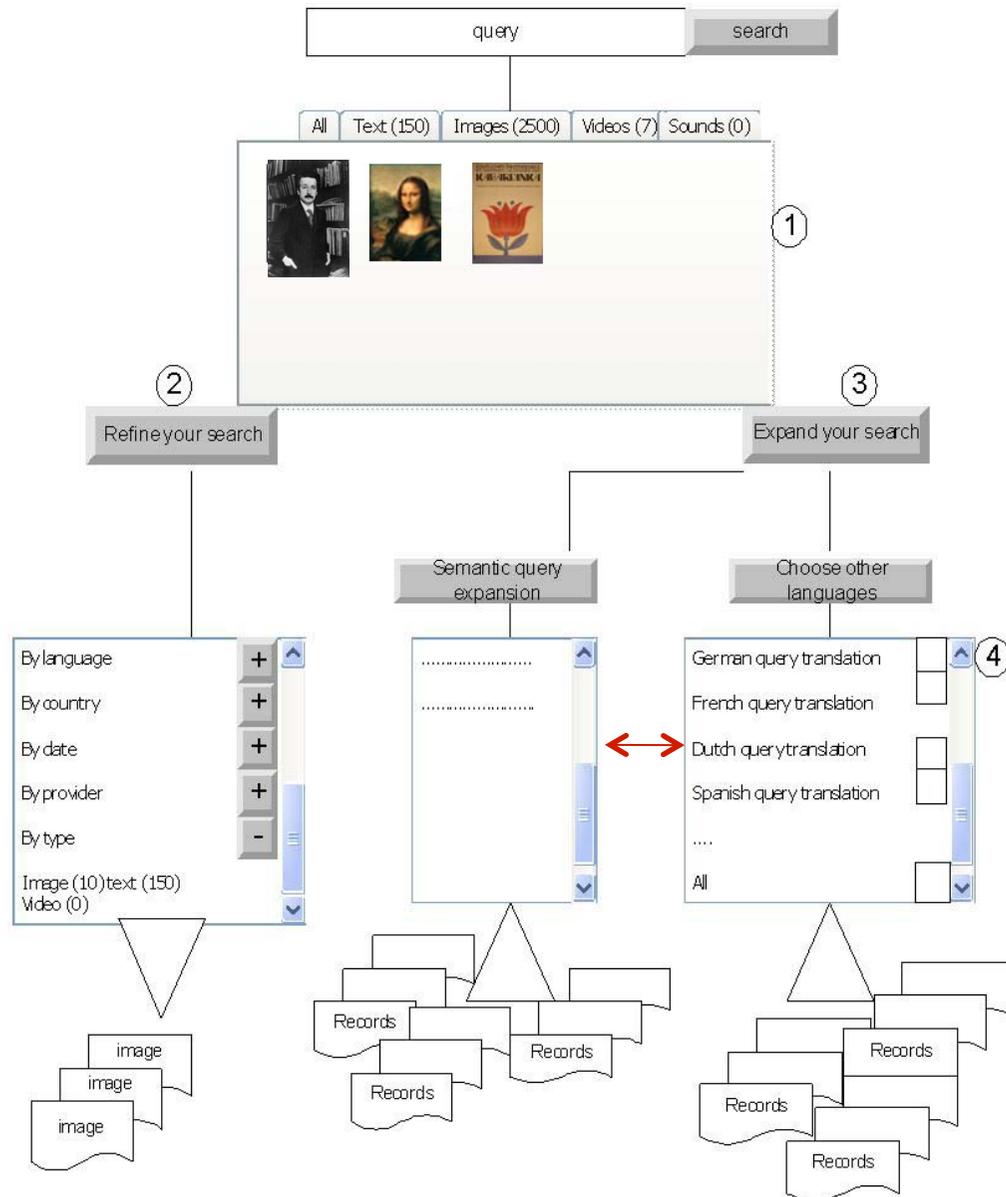


Europeana Query Translation Challenges

- query language unknown (detection OR user-determined?)
- desired target languages unknown
 - all OR user-determined
 - search sequence (query – translation – search OR query – search – translation – search)
- ambiguous translation candidates (present all OR present most likely → user-determined)

- interaction between translation and semantic expansion (Thoughtlab)

Query Translation Interface Scenario



Multilingual mapping of vocabularies

- Multilingual browsing
- Functional Specs for Rhine:
 - Browsing over result sets
 - Index browsing
 - Named entities = person browsing / spatial browsing
→ Special case (somewhat easier with authority files and limited lists)

4.2.2 Clustering on the fly (subject drill-down)



My Europeana Communities Partners Timeline (beta) Thought lab Choose a language

Advanced search

- drill-down works well for normalized, highly controlled-data
- highly controlled data (e.g. type) can be easily translated (static) → NOT YET DONE
- subjects?

Refine your search:

By language

By country

- germany (882)
- france (458)
- poland (134)
- Austria (112)**
- austria (82)
- netherlands (52)
- uk (23)
- hungary (14)

By date

By provider

By type



Mozart, Leopold
SLUB/Deutsche Fotothek
Saxon State Library - Dresden
State and University Library (SLUB)



Mozart, Karl
SLUB/Deutsche Fotothek
Saxon State Library - Dresden
State and University Library (SLUB)



Mozart, Leopold
SLUB/Deutsche Fotothek
Saxon State Library - Dresden
State and University Library (SLUB)



Mozart, Leopold
SLUB/Deutsche Fotothek
Saxon State Library - Dresden
State and University Library (SLUB)

Spec 4.2.2 Clustering on the fly (subject drill-down)

- solution: mapping = semantic equivalence (WP1)
- translation = equivalence (different?)

ENG: movie | film | cinema

GER: Kino | Film

FRA: cinéma | film

POL: ?

- display by language?
- user interaction: representation
- process for partial mappings / translations

Spec 4.2.8 Index browsing

- global index = all terms searched
 - Lemmatized?
 - Separated by language?
 - facet indexes: country, language, provider, type
 - who index
 - where index
 - what index
- separated by language?
 - what if language not available
 - or only partial listing in language available?

Multilingual mapping of vocabularies

- People: RKD / ULAN / VIAF (core languages: probably)
 - Places: TGN / Geonames (core languages: probably)
 - Mapped subject vocabularies (prototype WP1):
 - AAT (English, Dutch)
 - Iconclass (English, Dutch)
 - OSZK tezaurusz (Hungarian)
 - Rameau (French)
 - SCRAN (English)
 - Systematik Österreichische Mediathek (German)
- Missing core languages: Italian, Polish, Spanish
- Mapping criteria by language?

Multilingual mapping of vocabularies

→ Document translation / expansion on subject level



Title: W. MOZART / 1756-1791 / THEATRE IMPERIAL DE VIENNE / DON JUAN 5ème ACTE

Provider: Culture.fr/collections ; france

Subject: carte réclame; Ecole France; estampe ; ethnologie; Paris (lieu d'édition); portrait (Mozart Wolfgang Amadeus, homme de lettres : art de la musique) ; vue d'architecture (théâtre) ; armoiries (Vienne A) ; théâtre lyrique (Don Juan, homme, chandelier, esprit humain : mort : chevalier), publicité (chicorée, A la Ménagère, Duroyon & Ramette), fantôme ; Théâtre impérial de Vienne

Subject (french):

Subject (english):

Subject (german): ...

- support for multilingual users
- expand searchable terms
- partial translations possible
- EDM
- interaction with mappings?

Multilingual mapping of vocabularies challenges

- which vocabularies / which languages
- representation of subject browsing – by language
- solutions for partial translations
- document translation (subject): technical capabilities / interaction with EDM