The multilingual facet of the CACAO project

Luca Dini, CELI, Italy
Frédérique Segond, Xerox Research Centre Europe, France

eContentplus Program

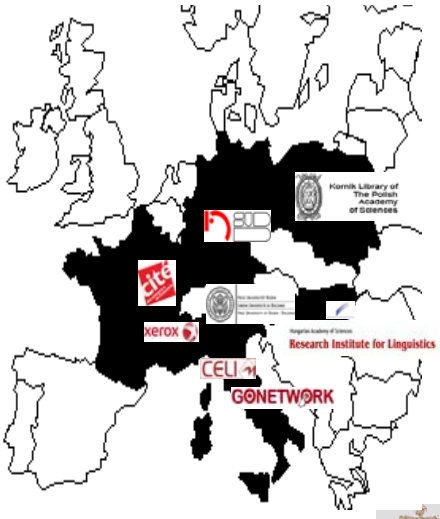
CACAO PROJECT

http://www.cacao-project.eu

CACAO enables international users to better access and exploit the available European multilingual electronic A.O.C. content.

Who are we?

- Xerox Research Centre Europe (France)
- Bibliothèque Cité des Sciences (France)
- CELI (Italy)
- Free University of Bozen-Bolzano (Italy)
- Gonetwork (Italy)
- University of Goettingen (Germany)
- Kórnik Library (Poland)
- Hungarian Academy of Sciences (Hungary)
- National Széchényi Library (Hungary)



CACAO PROJECT

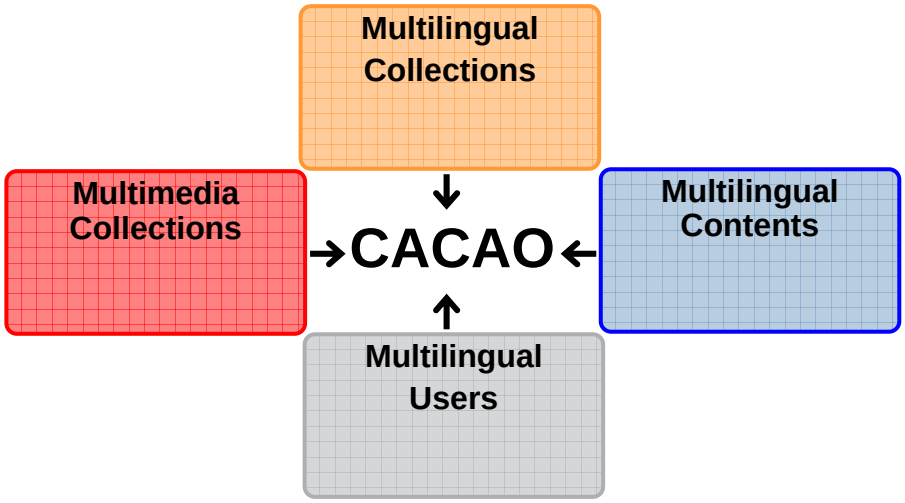
EUROPEANA
connecting cultural heritage

What CACAO does for whom?

CACAO helps individual citizens and librarians ***access, understand and navigate multilingual*** textual digital libraries and Online Public Access Catalogue content.

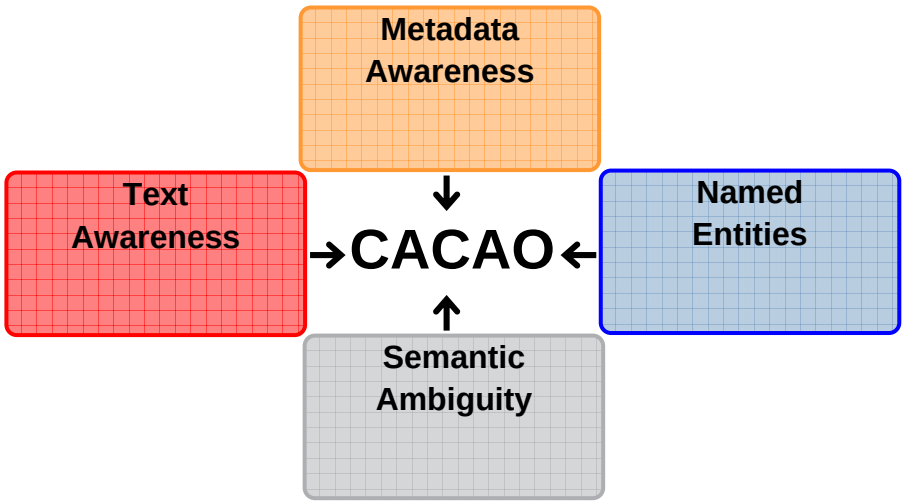
CACAO PROJECT

Facets of Multilinguality



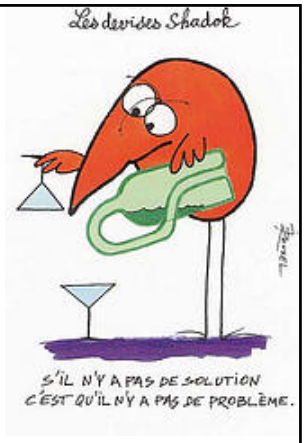
CACAO PROJECT

Facets of Search in DL



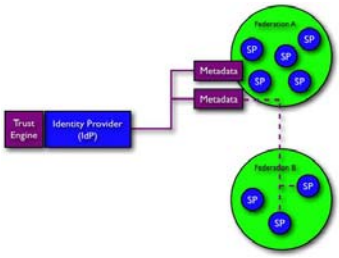
CACAO PROJECT

What are the problems?



- Information search and retrieval
- Metadata search (text)
- Multilingual search

CACAO PROJECT



What are the basic issues of multilingual search?

CACAO PROJECT

Structured documents

While traditionally a document is just a text, structured documents might contained metadata that are completely structured and expressed in different languages (titles, body, footnotes, etc.). These textual metadata go (might go) to a different *index field*.

CACAO PROJECT

Documents and fields

Metadata:

- Title: The Tempest
- Genre: Drama
- Abstract: The sorcerer Prospero, rightful Duke of Milan…

"prospero shakespeare"?

1. Document level : The Tempest, drama, shakespeare. The sorcerer Prospero, rightful Duke of Milan.
2. Metadata level: "+abstract:prospero +author:shakespeare"

As for text indexing it is also possible to give some more weight to certain metadata (e'g to boost generic fields). Metadata often contain free text that can be used for better indexing.

CACAO PROJECT

CACAO Indexing metadata content

Immagini Roma

- I papi, Roma e Dante: l'idea e le immagini di Roma nella Commedia dantescaDi Niccola LongoPubblicato da Bulzoni, 2004



Dante Alighieri

- …. All'angolo di questa via con **corso Dante Alighieri…**

- Books of *Dante Alighieri?*
- Books about *Dante Alighieri?*

CACAO PROJECT

CACAO: Indexing textual content

- CACAO offers extended indexing when free and/or full text is available. This can include an abstract, description or even the whole document.

- CACAO also offers thorough analyses of texts, thanks to automatic detection of personal names, locations, dates etc. which are usually difficult to determine in an approach that is not NLP-based.

CACAO PROJECT

CACAO: Indexing textual content

Title: ***La nascita del teatro moderno***

Abstract: *La "Storia del teatro moderno e contemporaneo" è articolata in tre volumi, ai quali si aggiunge un volume finale che raccoglie le trame dei mille testi teatrali fondamentali dal Cinquecento a oggi. Il primo volume è dedicato alla nascita del teatro moderno, di cui si analizzano i molteplici aspetti della "rottura" rispetto alla tradizione medievale, dove l'esperienza teatrale era ampiamente diffusa ma non veniva riconosciuta come tale a causa del rifiuto ideologico di cui era fatta oggetto da parte della dominante cultura cristiana.*

A query such as: "*teatro storia*" would not return this document, as the only elements which might match are in the abstract and not in the title.

CACAO PROJECT

CACAO: Indexing textual content

pterodattilo

> 0 Hits

> "Manuale di anatomia comparata dei vertebrati" di Emanuele Padoa **('*pterodattilo*' pag.235)**

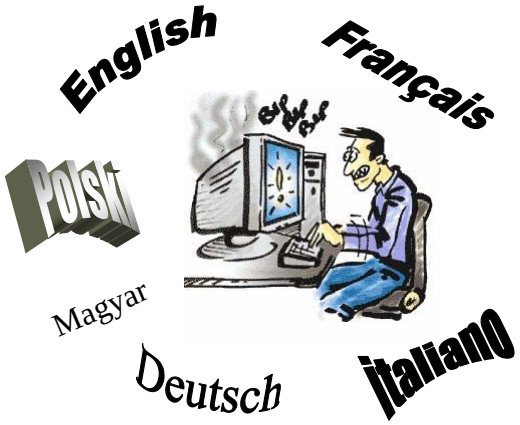
Parma

> "La certosa di Parma" di Stendhal

> "Storia della città di Parma" di Ireneo Affò

CACAO PROJECT

What are the modules for Multilingual and cross-lingual search?



CACAO PROJECT

The basic components to perform multilingual and cross lingual search

- All the basic components of Information retreival and search plus a ***query translator***:
  - It intersects with all other parameters (lemmatization/expansion/stopwording/NER * 2)
  - Additional issue:
    - Compounding
    - NER
    - Lemmatization level
    - Harmony of resources

CACAO PROJECT

A query in one language, answers in different languages

direct answer

Pirandello teatro ?

multilingual resources

query translation

- *Pirandello théatre*
- *Pirandello drámák*
- *Pirandello Theater*
- *Pirandello teatr*
- *Pirandello theatre*

more answers

- Six personnages en quête d'auteur
- Hat szereplo' szerzo't keres
- Sechs Personen suchen einen Autor
- Szes' c' postaci scenicznych w poszukiwaniu autora
- Six Characters in Search of an Author

CACAO PROJECT

Proper names can be ambiguous

A query is not just a sequence of words that can be simply translated

La Fontaine fable ?

The Fountain fable ?

Die Quelle Fabel?

Źródło legenda ?

forrás mese ?

La Fontana favola ?

Proper names and locations should be identified to avoid wrong translations.

CACAO PROJECT

Words are ambiguous

avocat

- ~~"Profession d'avocat" de Armand-Gaston Camus~~
- ~~"Le grand livre des fruits tropicaux" par F. Le Bellec.~~

- "avocat" as a profession?
- "avocat" as a fruit?

lawyer

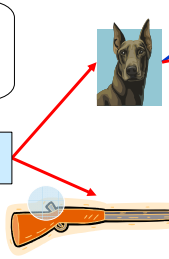
- ~~"Le grand livre des fruits tropicaux" par F. Le Bellec.~~

- "Profession d'avocat" de Armand-Gaston Camus

CACAO PROJECT

Words are ambiguous

How to transform a dog into a nail or a hammer…



Hund (D)
dog (E)
fickó (M)
pies (P)

cane (I)

Italian presents the same ambiguity

Hammer (D)
hammer (E)

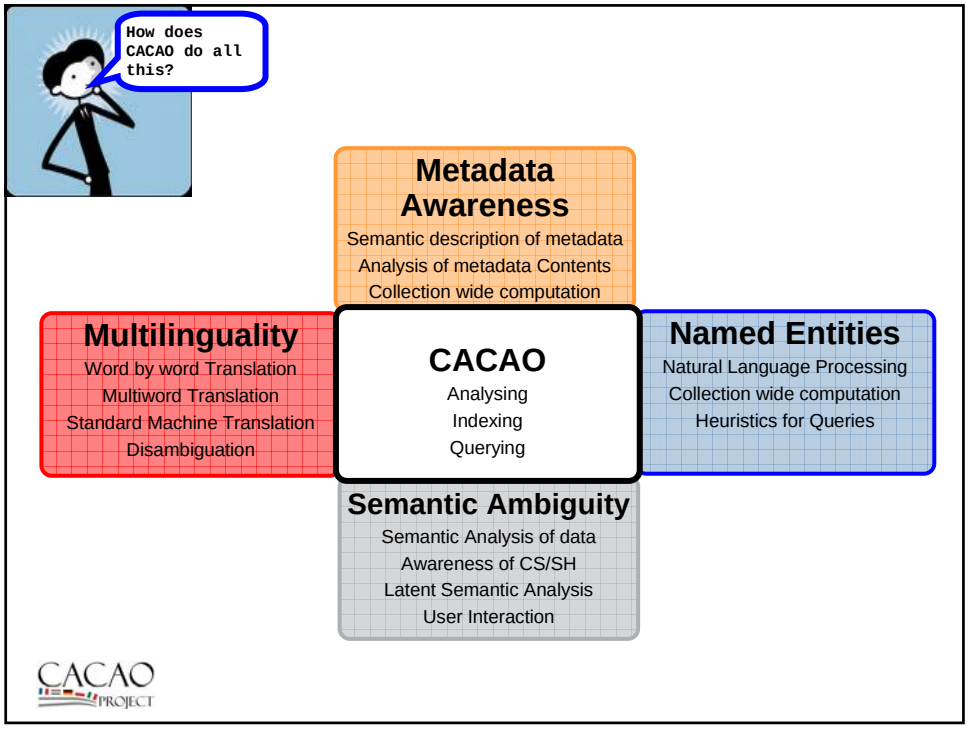
Hahn (D)
kakas (M)
kurek (P)

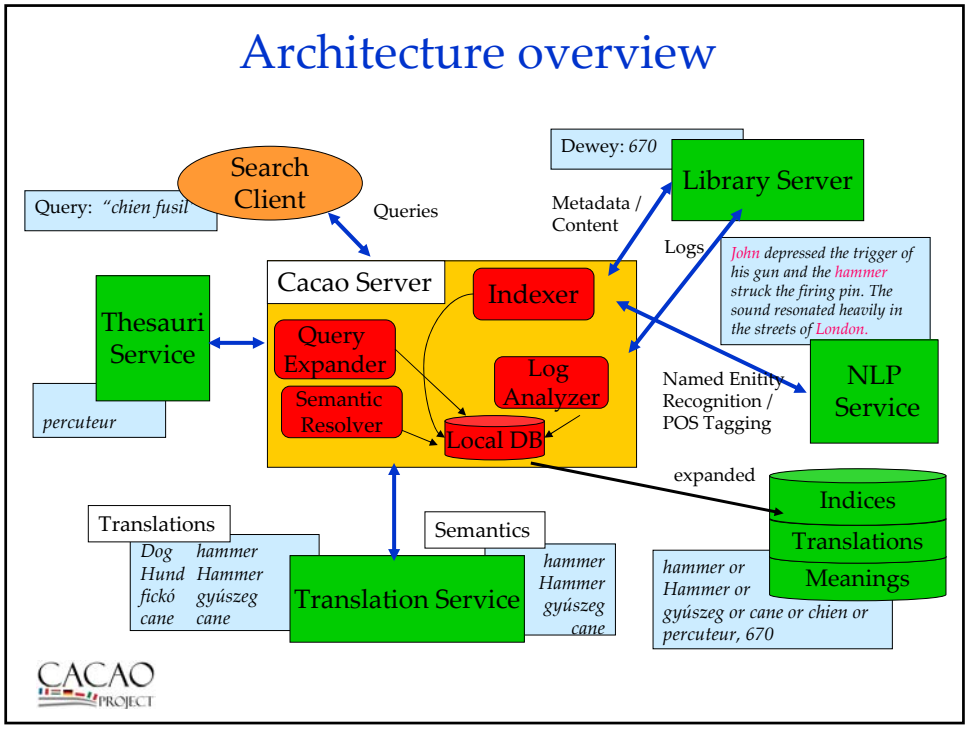
gyúszeg (M)

Translation does not reduce to a simple dictionary search…

D=Deutsch M=Magyar F=Français
E=English P=Polski I=Italiano

CACAO PROJECT


How does CACAO do all this?
Metadata Awareness
Semantic description of metadata
Analysis of metadata Contents
Collection wide computation
Multilinguality
Word by word Translation
Multiword Translation
Standard Machine Translation
Disambiguation
CACAO
Analysing
Indexing
Querying
Named Entities
Natural Language Processing
Collection wide computation
Heuristics for Queries
Semantic Ambiguity
Semantic Analysis of data
Awareness of CS/SH
Latent Semantic Analysis
User Interaction
CACAO PROJECT


Architecture overview
Search Client
Query: "chien fusil
Queries
Dewey: 670
Library Server
Metadata / Content
Logs
John depressed the trigger of his gun and the hammer struck the firing pin. The sound resonated heavily in the streets of London.
Cacao Server
Indexer
Query Expander
Log Analyzer
Semantic Resolver
Local DB
Thesauri Service
percuteur
Named Enitity Recognition / POS Tagging
NLP Service
expanded
Indices
Translations
Meanings
Translations
Dog hammer
Hund Hammer
fickó gyúszeg
cane cane
Translation Service
Semantics
hammer
Hammer
gyúszeg
cane
hammer or Hammer or gyúszeg or cane or chien or percuteur, 670
CACAO PROJECT

http://www.cross-library.com:8080/cacaoUI/

CACAO PROJECT

CACAO PROJECT