

# 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

Stefan Farrenkopf

Goettingen State and University Library, Germany

Elisabeth Frasnelli

Library of the Free University of Bozen-Bolzano, Italy

---

CACAO



PROJECT

CROSS-LANGUAGE ACCESS TO CATALOGUES AND ON-LINE LIBRARIES

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



eContentplus Program

 NIEDERSÄCHSISCHE STAATS- UND  
UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN 



FREIE UNIVERSITÄT BOZEN  
LIBERA UNIVERSITÀ DI BOLZANO  
FREE UNIVERSITY OF BOZEN • BOLZANO

# Contents

1. Introduction
2. CACAO Vision and Architecture
3. Translation and False Friends
4. Solution 1: Specification of the Language of the Field
5. Solution 2: Association to a Class
  - a. Term Ambiguity
  - b. False Friends
6. Conclusion

- ➡ 1. Introduction
- 2. CACAO Vision and Architecture
- 3. Translation and False Friends
- 4. Solution 1: Specification of the Language of the Field
- 5. Solution 2: Association to a Class
  - a. Term Ambiguity
  - b. False Friends
- 6. Conclusion

# Introduction

## CACAO (Cross-language Access to Catalogues and Online Libraries)

- 24 months, ending 30 November 2009
- Co-funded by the eContentplus Programme of the European Commission



[http://europa.eu/abc/european\\_countries/index\\_en.htm](http://europa.eu/abc/european_countries/index_en.htm)



# Introduction

## The 23 official languages of the EU

български (Bălgarski) — Bulgarian

Čeština — Czech

Dansk — Danish

Deutsch — German

Eesti — Estonian

Elinika — Greek

English

Español — Spanish

Français — French

Gaeilge — Irish

Italiano — Italian

Latviesu valoda — Latvian



Lietuviu kalba — Lithuanian

Magyar — Hungarian

Malti — Maltese

Nederlands — Dutch

Polski — Polish

Português — Portuguese

Română — Romanian

Slovenčina — Slovak

Slovenščina — Slovene

Suomi — Finnish

Svenska — Swedish

[http://ec.europa.eu/education/policies/lang/languages/index\\_en.html](http://ec.europa.eu/education/policies/lang/languages/index_en.html)

# Introduction

## Multilinguality in the EU

According to a 2006 European Commission/Eurobarometer study of citizens of EU countries

- "56% [...] are able to hold a conversation in a language other than their mother tongue"
- "28% [...] master two languages along with their native language"
- "approximately 1 in 10 [...] has sufficient skills to have a conversation in three languages"

*Europeans and their Languages*, [http://ec.europa.eu/public\\_opinion/archives/ebs/ebs\\_243\\_en.pdf](http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_en.pdf)

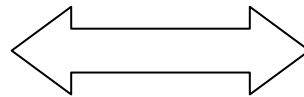
# Introduction

## The users' dilemma

- About 40% of user queries are duplicated in at least two languages.
- In one library, about 20% are duplicated in three languages.



... cat ... Katze ... gatto ...  
chat ... Macska ... kot ...



# Introduction

## The libraries' dilemma and CACAO's challenge

Among the 5 CACAO libraries

- 6 CACAO languages; many more languages represented in catalogs
- 6+ controlled subject vocabularies
- 3+ name authority files
- 5+ classification systems
- 5 bibliographic formats



1. Introduction
- ➡ 2. CACAO Vision and Architecture
3. Translation and False Friends
4. Solution 1: Specification of the Language of the Field
5. Solution 2: Association to a Class
  - a. Term Ambiguity
  - b. False Friends
6. Conclusion

# CACAO Vision and Architecture

## An early implementation

### The Library of the Free University of Bozen-Bolzano

The screenshot shows a search interface with the following elements:

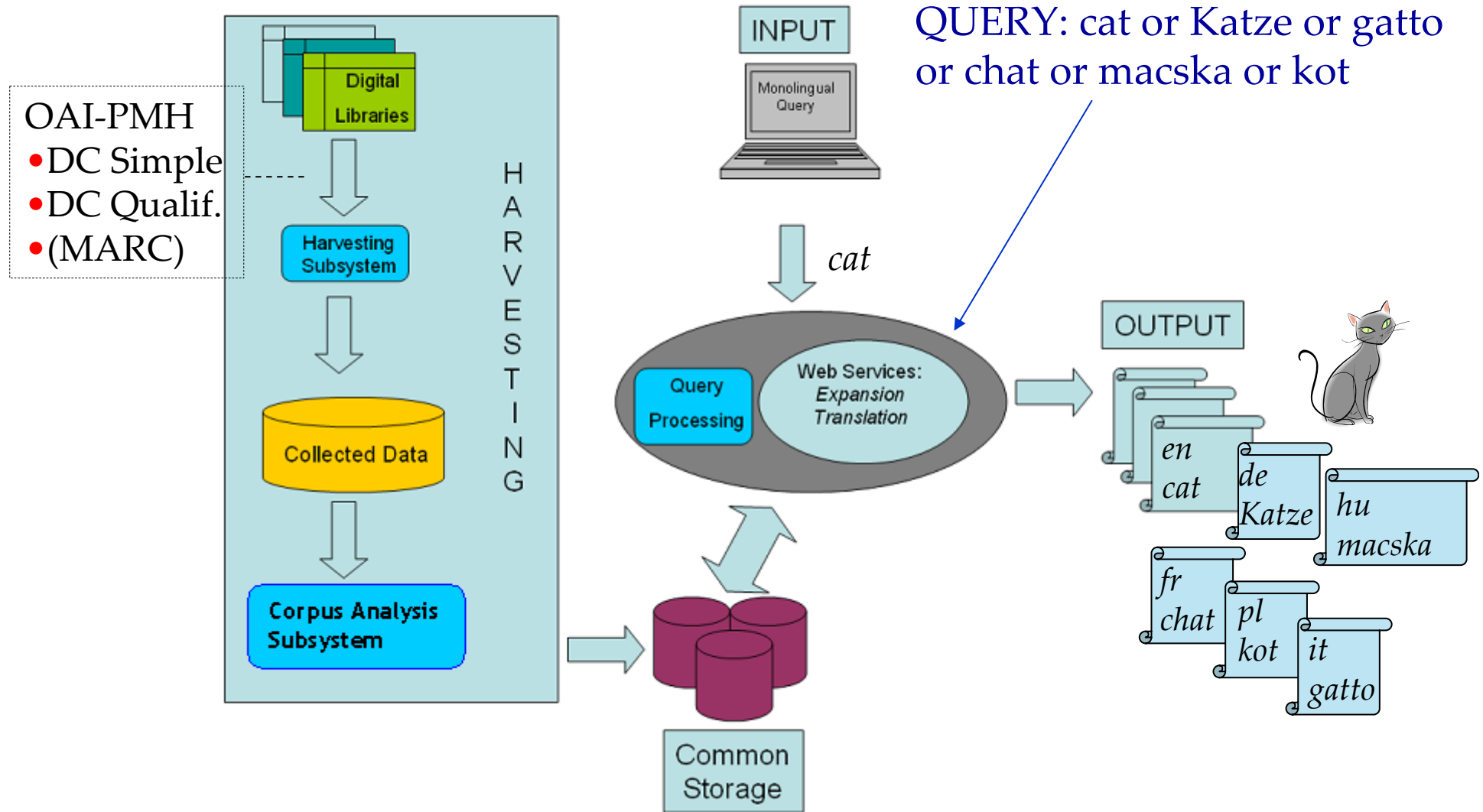
- Search term language:** Radio buttons for 'en' (selected), 'de', and 'it'.
- Search term:** A text input field containing 'cat'.
- Results:**
  - English:** 11 results for 'cat'. A red circle highlights this section.
  - Italian:** 85 results for 'gatta'. A red circle highlights this section.
  - German:** 33 results for 'Katz'. A red circle highlights this section. The results list includes:
    - Author and title: Rinser, Luise: Die rote Katze : Erzählungen / Luise Rinser - 3. Aufl. (Fischer-Bibliothek.)  
Publisher: Fischer Year: 1988  
Relevance: 23
    - Author and title: Seyffert, Sabine: Auf sanften Pfoten schleicht die Katze : Übungen und Geschichten zum Bewegen und Entspannen / Sabine Seyffert  
Publisher: Kösel Year: 2000  
Relevance: 20
    - Author and title: Buttaroni, Susanna... : Ehe, Berge und schwarze Katzen / Susanna Buttaroni ; Andreas Paula.  
Publisher: Alpha & Beta Year: 2006  
Relevance: 20
    - Author and title: Das reflektierende Team : Dialoge und Dialoge über die Dialoge / Tom Andersen (Hrsg.). Mit Beitr. von:... - 4., unver... (Systemische Studien ; 5)  
Publisher: Verl. Modernes Lernen Year: 1996  
Relevance: 20
    - Author and title: Katervaterhasensohn / erzählt von Jana Frey. Mit Bildern von Marlis Scharff-Kniemeyer  
Publisher: Ravensburger Buchverl. Year: 2000  
Relevance: 20

<http://pro.unibz.it/opacdocdigger/index.asp?bSWIN=True&MLSearch=TRUE&Lang=2>

Bernardi, R., D. Calvanese, L. Dini, V. Di Tomaso, E. Frasnelli, U. Kugler, B. Plank. (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.  
<http://www.inf.unibz.it/~bernardi/index.php?page=pub>

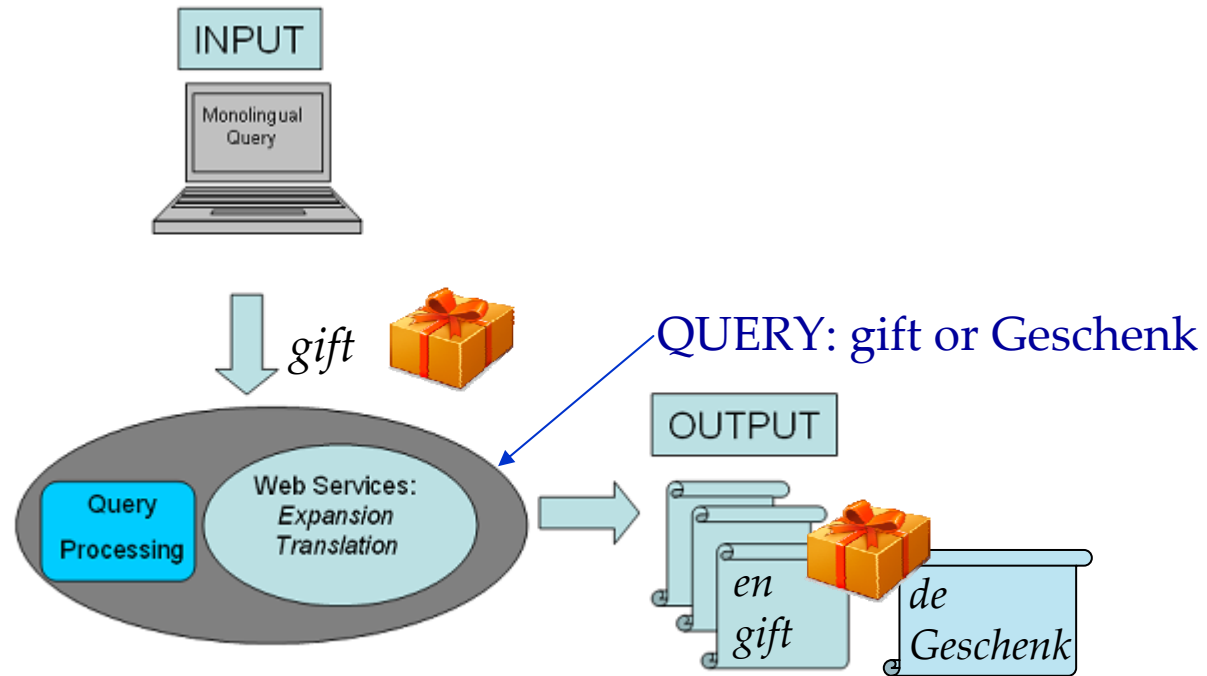
# CACAO Vision and Architecture

## CACAO architecture overview

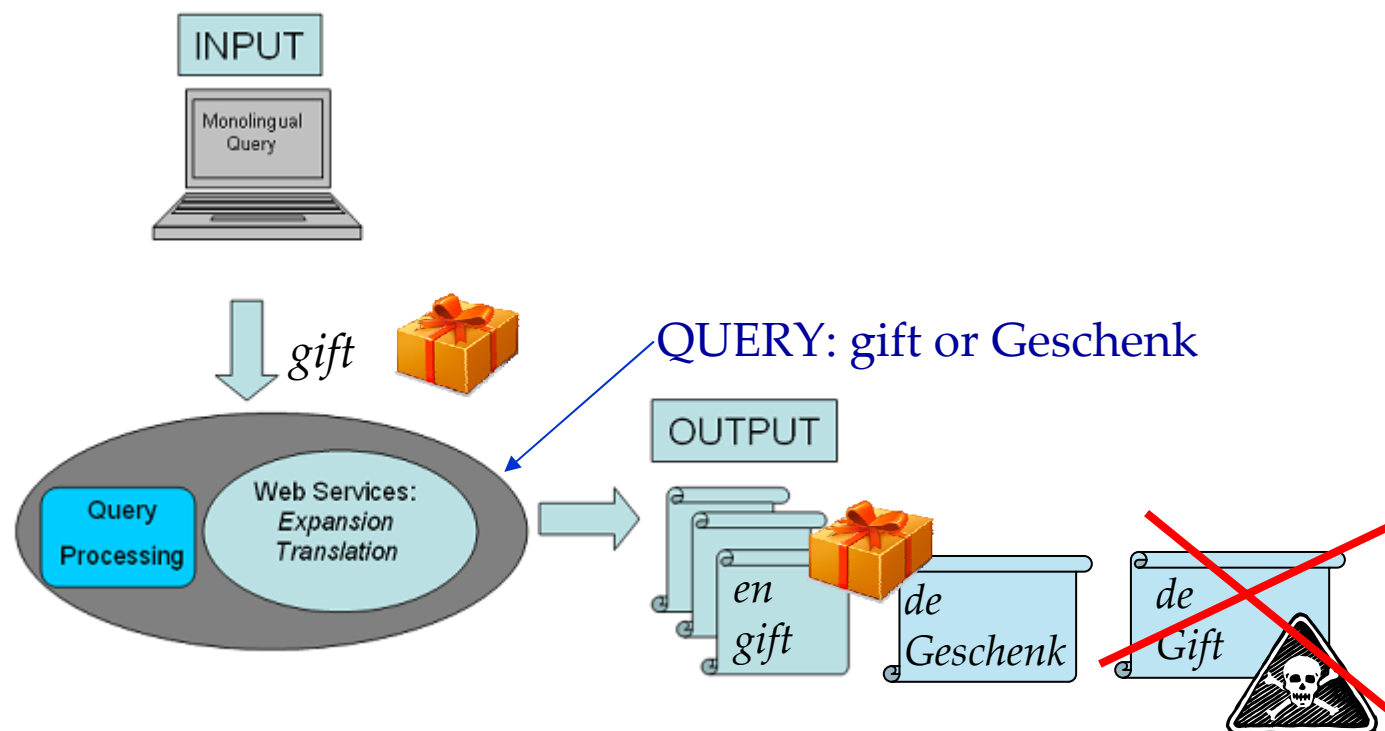


1. Introduction
2. CACAO Vision and Architecture
- ➡ 3. Translation and False Friends
4. Solution 1: Specification of the Language of the Field
5. Solution 2: Association to a Class
  - a. Term Ambiguity
  - b. False Friends
6. Conclusion

# Translation and False Friends



# Translation and False Friends



<dc:title>The Gift book: A sumptuous guide to the world of giving</dc:title>  
 <dc:title>Geschenk-Ideen</dc:title>  
 <dc:title>Gift: Magie u. Realität, Nutzen u. Verderben</dc:title>

## Translation and False Friends

de "falsche Freunde", fr "faux amis"

words in different languages that look similar but that have different meanings

en	<i>gift</i>	de	<i>Gift</i>	"poison"
en	<i>herd</i>	de	<i>Herd</i>	"stove"
en	<i>pain</i>	fr	<i>pain</i>	"bread"
en	<i>cane</i>	it	<i>cane</i>	"dog"
en	<i>farmer</i>	hu	<i>farmer</i>	"jeans"
en	<i>hazard</i>	pl	<i>hazard</i>	"gambling"

1. Introduction
2. CACAO Vision and Architecture
3. Translation and False Friends
- ➡ 4. Solution 1: Specification of the Language of the Field
5. Solution 2: Association to a Class
  - a. Term Ambiguity
  - b. False Friends
6. Conclusion

## Solution 1: Specification of the Language of the Field

### Specification of the language of the field

- a. `<dc:title xml:lang="en">The Gift book: A sumptuous guide to the world of giving</dc:title>`  
`<dc:subject xml:lang="en">Gift</dc:subject>`
- b. `<dc:title xml:lang="de">Geschenk-Ideen</dc:title>`  
`<dc:subject xml:lang="de">Geschenk</dc:subject>`
- c. `<dc:title xml:lang="de">Gift: Magie u. Realität, Nutzen u. Verderben</dc:title>`  
`<dc:subject xml:lang="de">Gift</dc:subject>`



### CACAO has two options

1. Restrict the search terms to the appropriate language  
**QUERY:** `gift[xml:lang="en"]` or `Geschenk[xml:lang="de"]`  
so that (a) & (b) are retrieved but not (c).
2. Perform a simple query  
**QUERY:** `gift` or `Geschenk`  
and sort the hit list so that the hits with the False Friend in (c) are ranked last.

## Solution 1: Specification of the Language of the Field

**Specifying the language of the field is probably the best option to solve the problem of False Friends.**

- However, metadata often do not come with the specification of the language of the field.
- In many cases, this information is simply not available in the catalogs or is not included in the metadata.

Can we infer the language of the field from other information available?

## Solution 1: Specification of the Language of the Field

Can the language of <dc:title> be inferred from <dc:language>?

Problems:

1. Titles that contain foreign/borrowed words. (These records indicate that the (only) language of the document is German.)

<dc:title>Gift - Marcel Mauss' Kulturtheorie der Gabe</dc:title>

<dc:title>Immobiliencontrolling durch Business Intelligence</dc:title>

<dc:title>Business Process Outsourcing: Geschäftsprozesse kontextorientiert auslagern</dc:title>

2. Mismatch between the language of <dc:title> and the language specified in <dc:language>.

<dc:title>Practical business research</dc:title>

<dc:language>ger</dc:language>

## Solution 1: Specification of the Language of the Field

Can the language of <dc:title> be inferred from <dc:language>?

Problems:

3. Monolingual <dc:title> with more than one <dc:language>.

<dc:title>Den Kulturen Raum geben: das Konzept selektiver Kulturräume am Beispiel des deutsch-tschechisch-österreichischen Dreiländerecks</dc:title>

<dc:language>ger</dc:language>

<dc:language>eng</dc:language>

<dc:language>cze</dc:language>

4. A parallel title (2+ languages) that appears in <dc:title>.

<dc:title>Modern problems in pharmacopsychiatry = Moderne Probleme der Pharmacopsychiatrie = Problèmes actuels de pharmacopsychiatrie</dc:title>

## Solution 1: Specification of the Language of the Field

Can the language of <dc:subject> be inferred from the vocabulary encoding scheme?

- Problem: foreign/borrowed terms in subjects

```
<dc:subject xsi:type="dcterms:LCSH*">Bündnis 90/Die  
Grünen</dc:subject>
```

```
<dc:subject xsi:type="cacao:SWD†">Records of Early English  
Drama</dc:subject>
```

```
<dc:subject xsi:type="cacao:SWD†">Consiglio Italiano per le  
Scienze Sociali</dc:subject>
```

\*Library of Congress Subject Headings, English language  
†Schlagwortnormdatei, German-language

## Solution 1: Specification of the Language of the Field Summary

**Specifying the language of the field is probably the best option to deal with the problem of False Friends.**

- However, metadata often do not come with the specification of the language of the field.
- The language of the field is not always inferable.


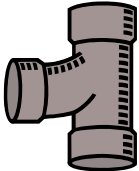
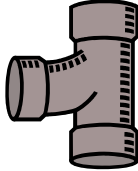

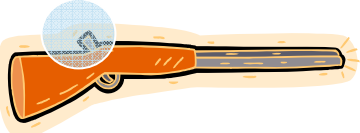

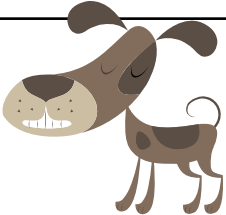
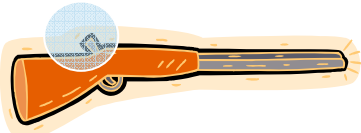
We would like to experiment with the language attribute specification to make best-practice recommendations and improve CACAO performance.

1. Introduction
2. CACAO Vision and Architecture
3. Translation and False Friends
4. Solution 1: Specification of the Language of the Field
5. Solution 2: Association to a Class
- ➡ a. Term Ambiguity
- b. False Friends
6. Conclusion

# Solution 2: Association to a Class

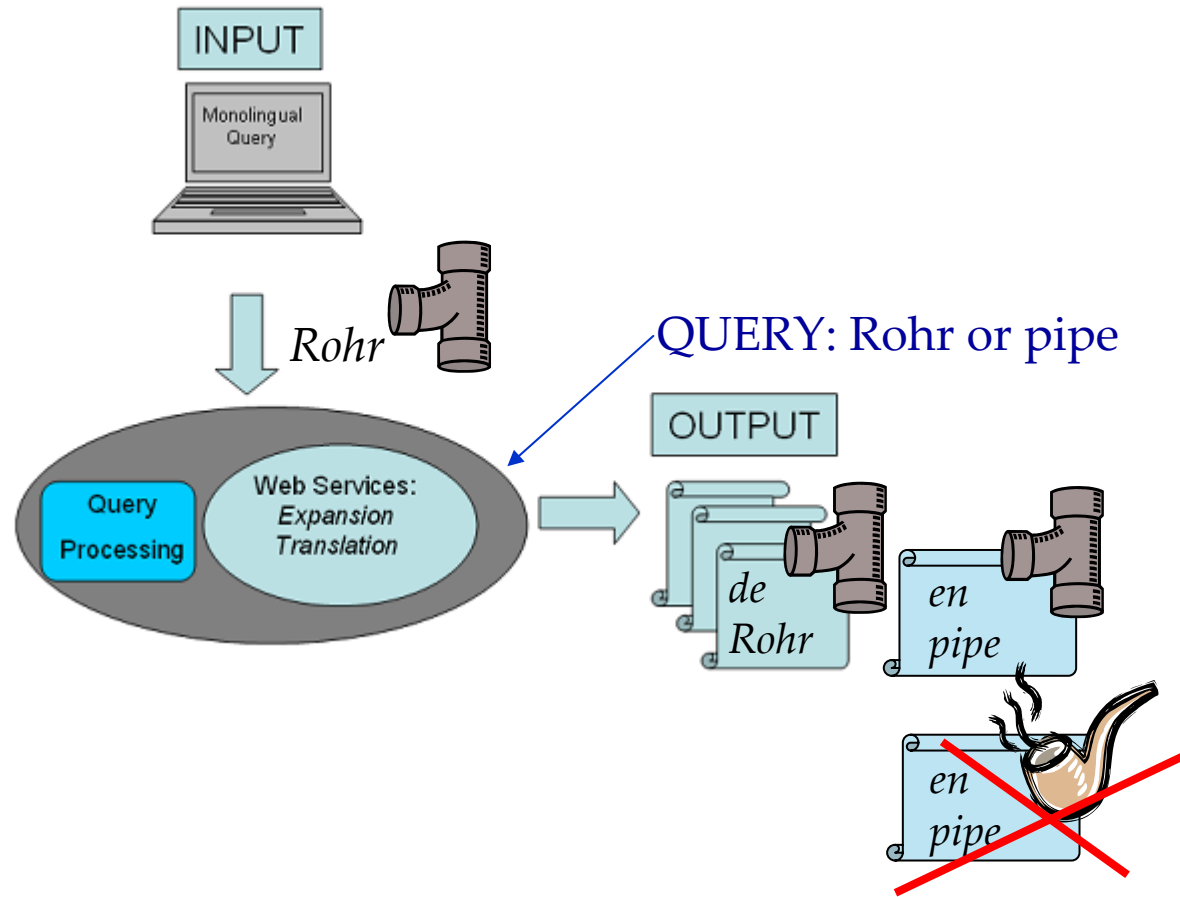
## Term Ambiguity

having 2 or more meanings

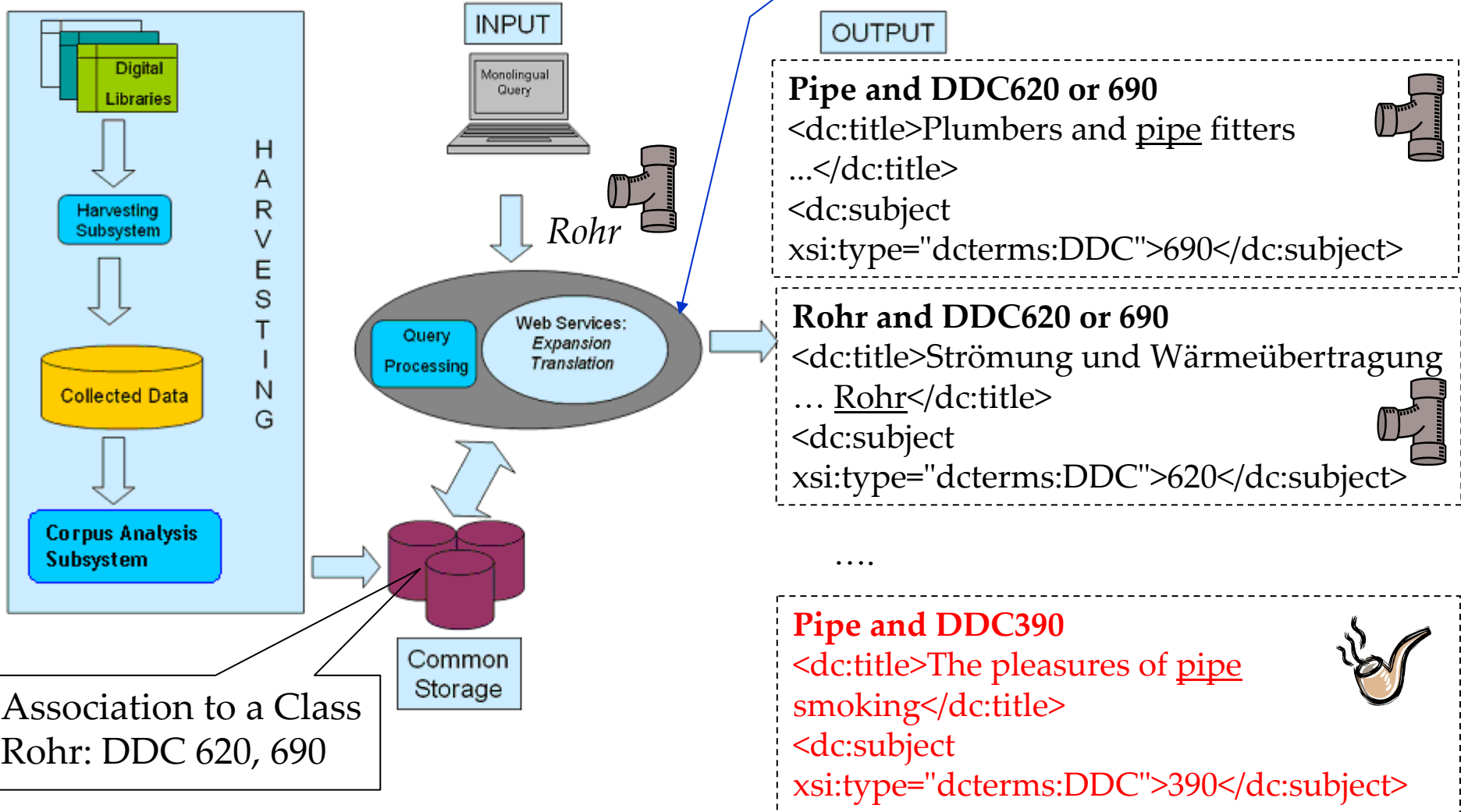
<p>en pipe</p>  	<p>de Rohr</p>  <p>de Pfeife</p> 
<p>it cane</p>  	<p>en dog</p>  <p>en cock (of a weapon)</p> 

# Solution 2: Association to a Class

## Term Ambiguity



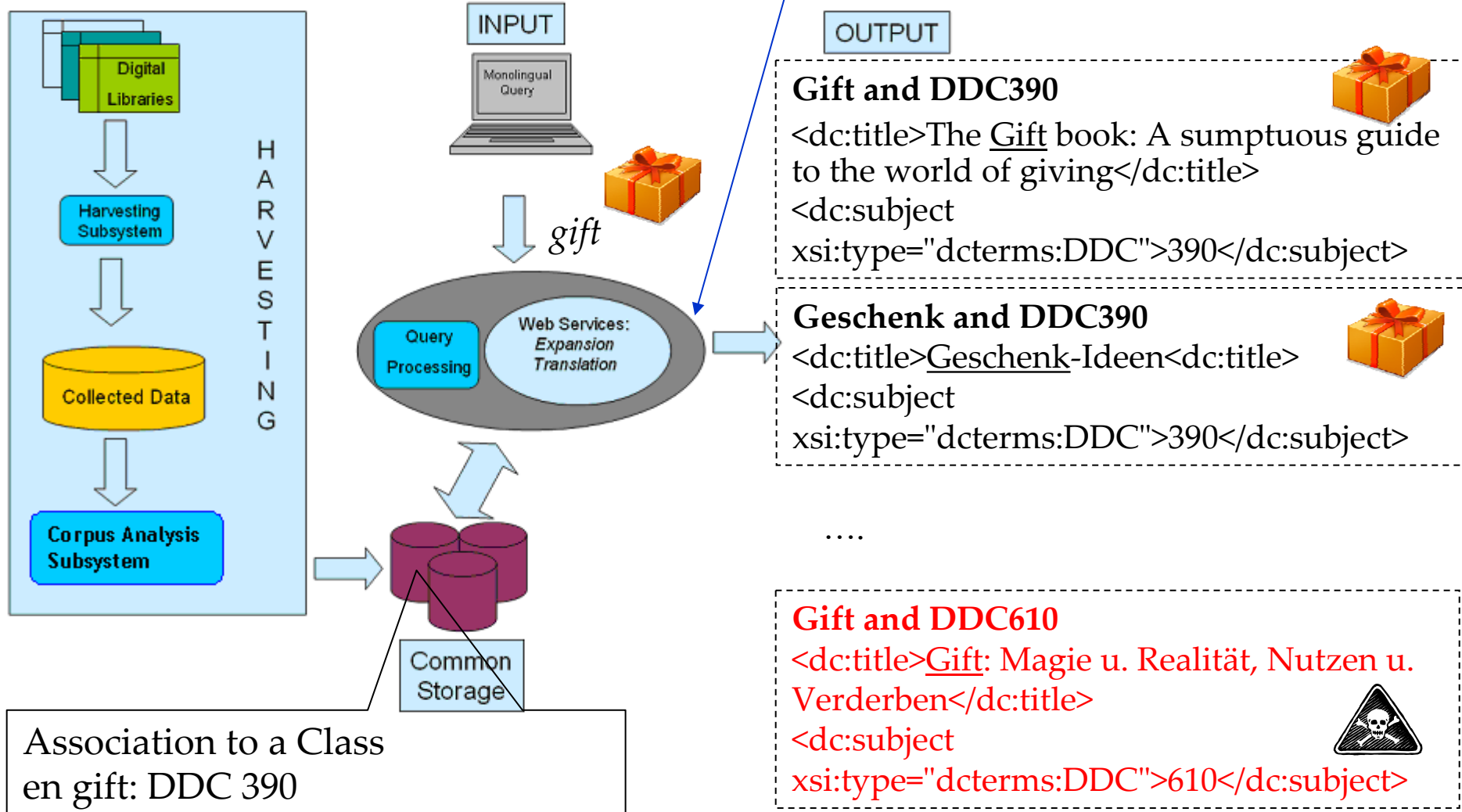
# Solution 2: Association to a Class Term Ambiguity



1. Introduction
2. CACAO Vision and Architecture
3. Translation and False Friends
4. Solution 1: Specification of the Language of the Field
5. Solution 2: Association to a Class
  - a. Term Ambiguity
  - b. False Friends
6. Conclusion



## Solution 2: Association to a Class False Friends



1. Introduction
2. CACAO Vision and Architecture
3. Translation and False Friends
4. Solution 1: Specification of the Language of the Field
5. Solution 2: Association to a Class
  - a. Term Ambiguity
  - b. False Friends
- ➡ 6. Conclusion

## Conclusion

### Summary

The language of the text in the metadata field is important

- so that metadata can be exploited for cross-language purposes or in multilingual settings.
- for dealing with False Friends.

Important fields such as <dc:title> and <dc:subject> often are not provided language attributes.

- The language attribute is not always fully predictable from <dc:language> nor from the vocabulary encoding scheme.

Thus, we would like to experiment with Association to a Class for dealing with False Friends.

- Association to a Class was originally designed for and will be used as a solution for the Term Ambiguity problem.
- It is similar to synsets used in WordNet and EuroWordNet.

## Conclusion

### Interoperability

In order to perform Association to a Class, the metadata must either

- contain the same classification system or
- the classifications (or subject headings) must be mappable to the same system.

CACAO's experience with cross-language access so far supports (perhaps in unexpected ways) the importance of the interoperability of classification systems and of subject vocabularies for information retrieval in cross-domain environments.

- Koch, Neuroth, and Day (2001), NKOS (2001), Chan and Zeng (2002), Harper and Tillett (2007), Mayr and Petras (2008), etc.

CACAO will experiment with already existing mappings and create its own.

## Conclusion

### CACAO in the near future

- The European Library will integrate and evaluate CACAO technologies.
- CACAO libraries
  - will be grouped into a single portal.
  - will create several thematic portals.

# Introduction

## The libraries' dilemma and CACAO's challenge

Among the 5 CACAO libraries

- 6 CACAO languages; many more languages represented in catalogs
- 6+ controlled subject vocabularies
- 3+ name authority files
- 5+ classification systems
- 5 bibliographic formats

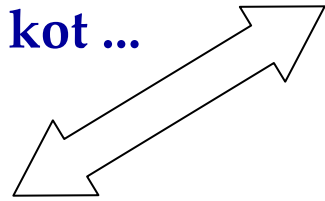


... cat ... Katze ... gatto ...  
chat ... Macska ... kot ...



# Conclusion The CACAO vision

... cat ... Katze ... gatto ...  
chat ... macska ... kot ...



macska



en Thank You.  
de Danke.  
it Grazie.  
fr Merci.  
pl Dziękuję.  
hu Köszönöm.

**Barbara Levergood**

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

**Stefan Farrenkopf**

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

**Elisabeth Frasnelli**

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

*We thank Raffaella Bernardi, Jane Greenberg, Tom Baker,  
the three anonymous reviewers, and the CACAO partners.*

International Conference on Dublin Core and Metadata Applications  
Humboldt-Universität zu Berlin  
Berlin, Germany  
25 September 2008