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, decompounding, lemmatization, named entity detection, phrase detection, dictionaries
- similar resources necessary for vocabulary mapping
  - Wordnet etc. ?
- Licensed resources ← → 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:

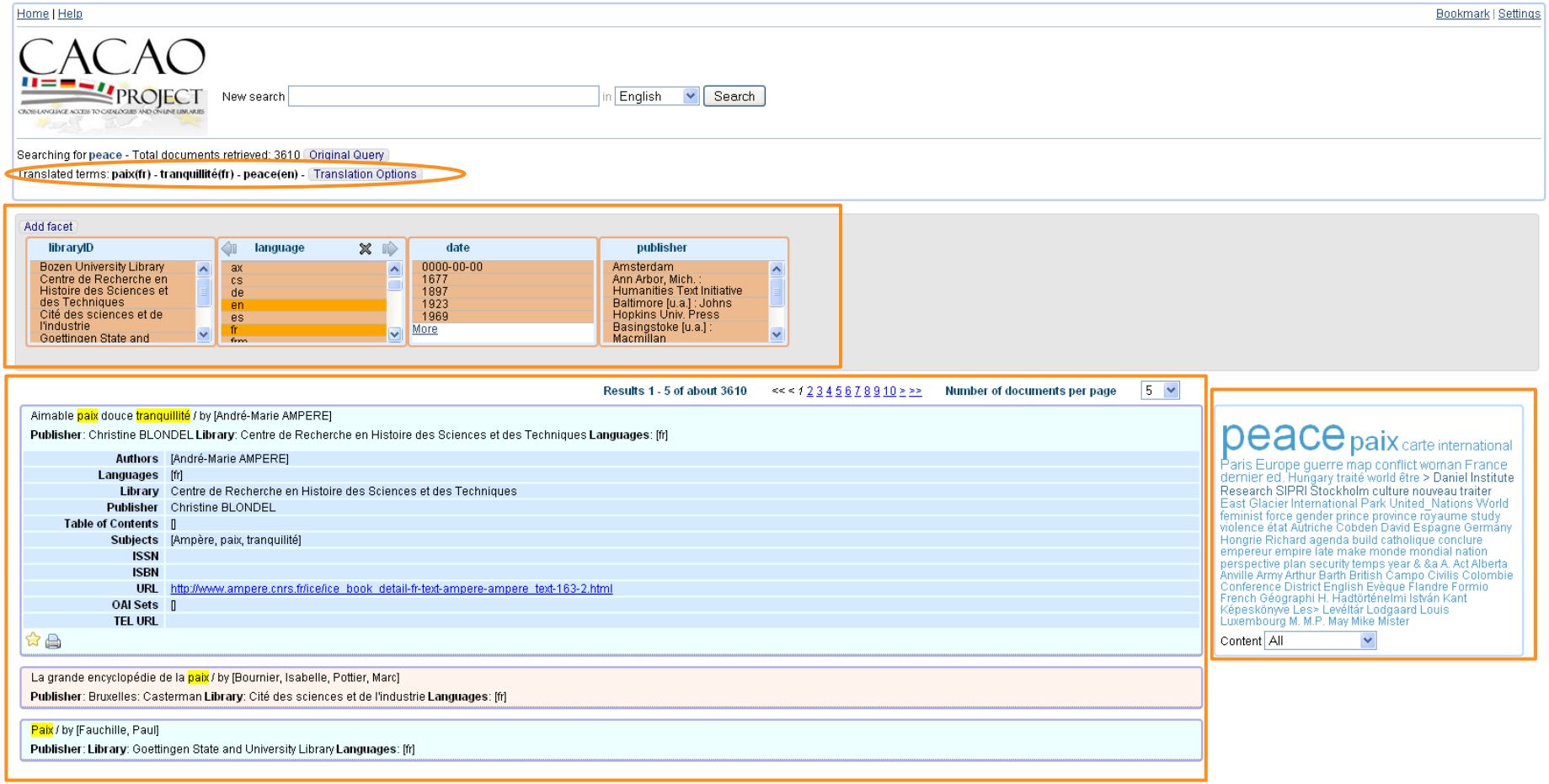
- CACAO Partner Libraries, TEL
- 5 million metadata records in index

Features:

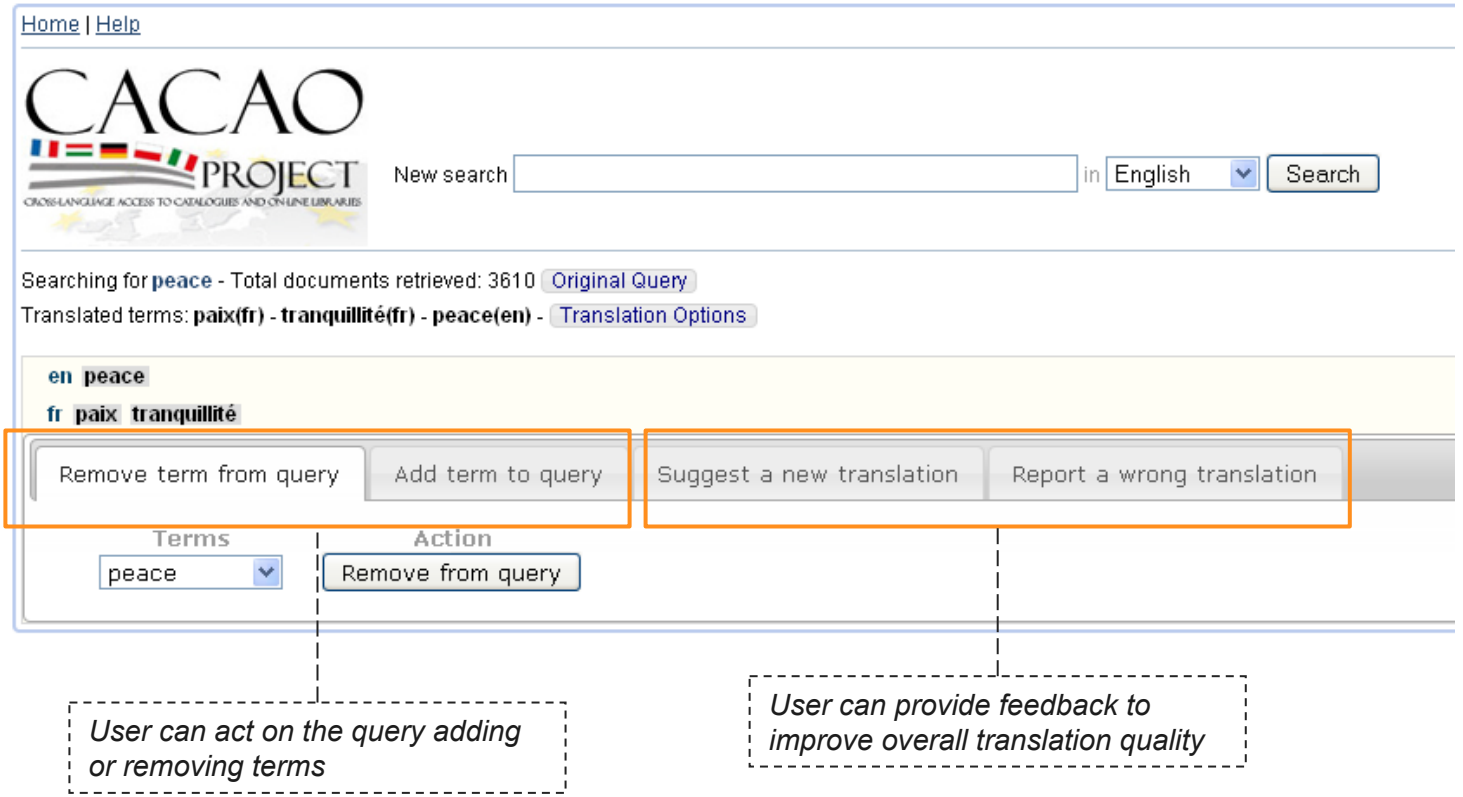
- Input:
  - Query string in one language
- Output:
  - Search results organized by facets
  - User profile to define settings
  - Refinements of search strings and translations

CACAO User Interface: Query

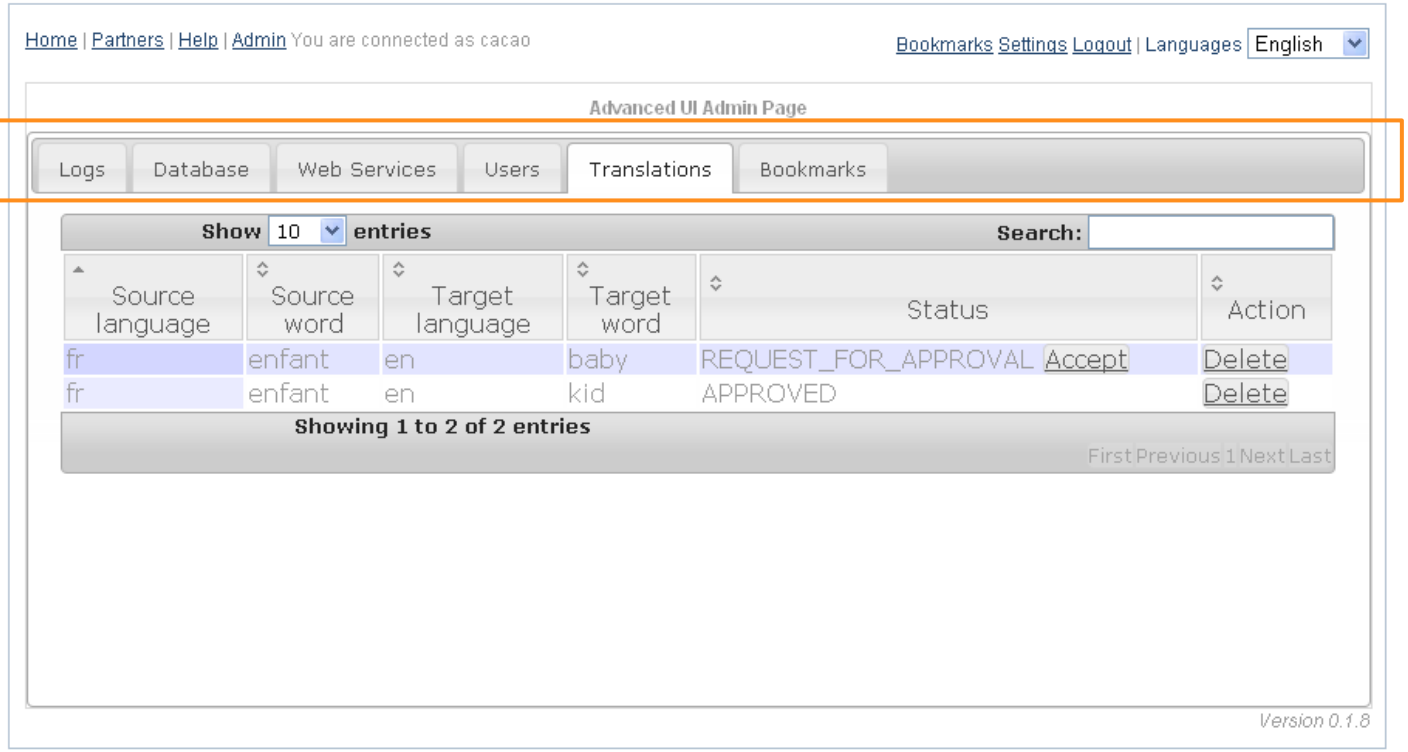
CACAO User Interface: Results



CACAO User Interface: Translations



CACAO Administrator Interface



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

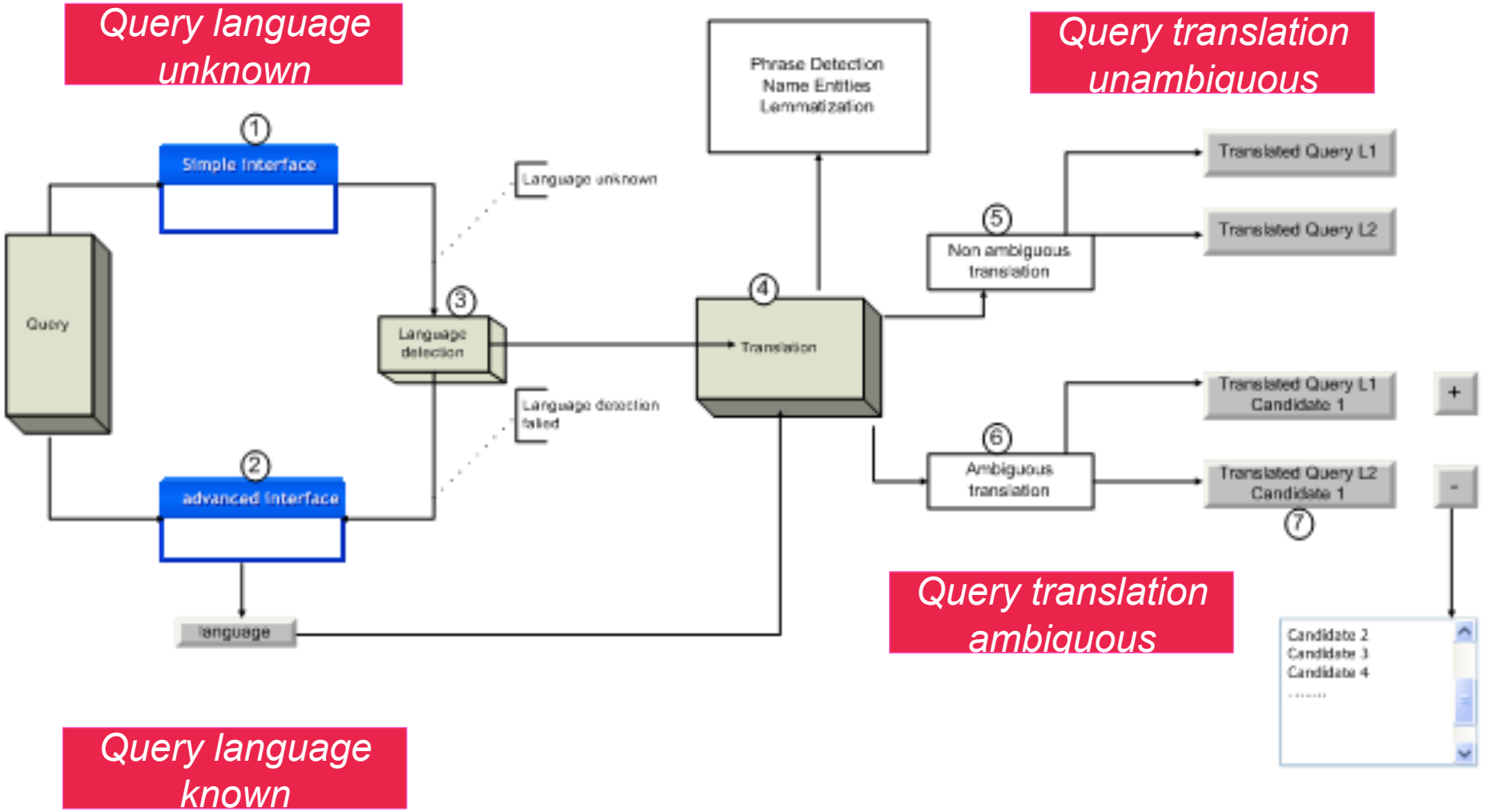
CACAO Application Profile

*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

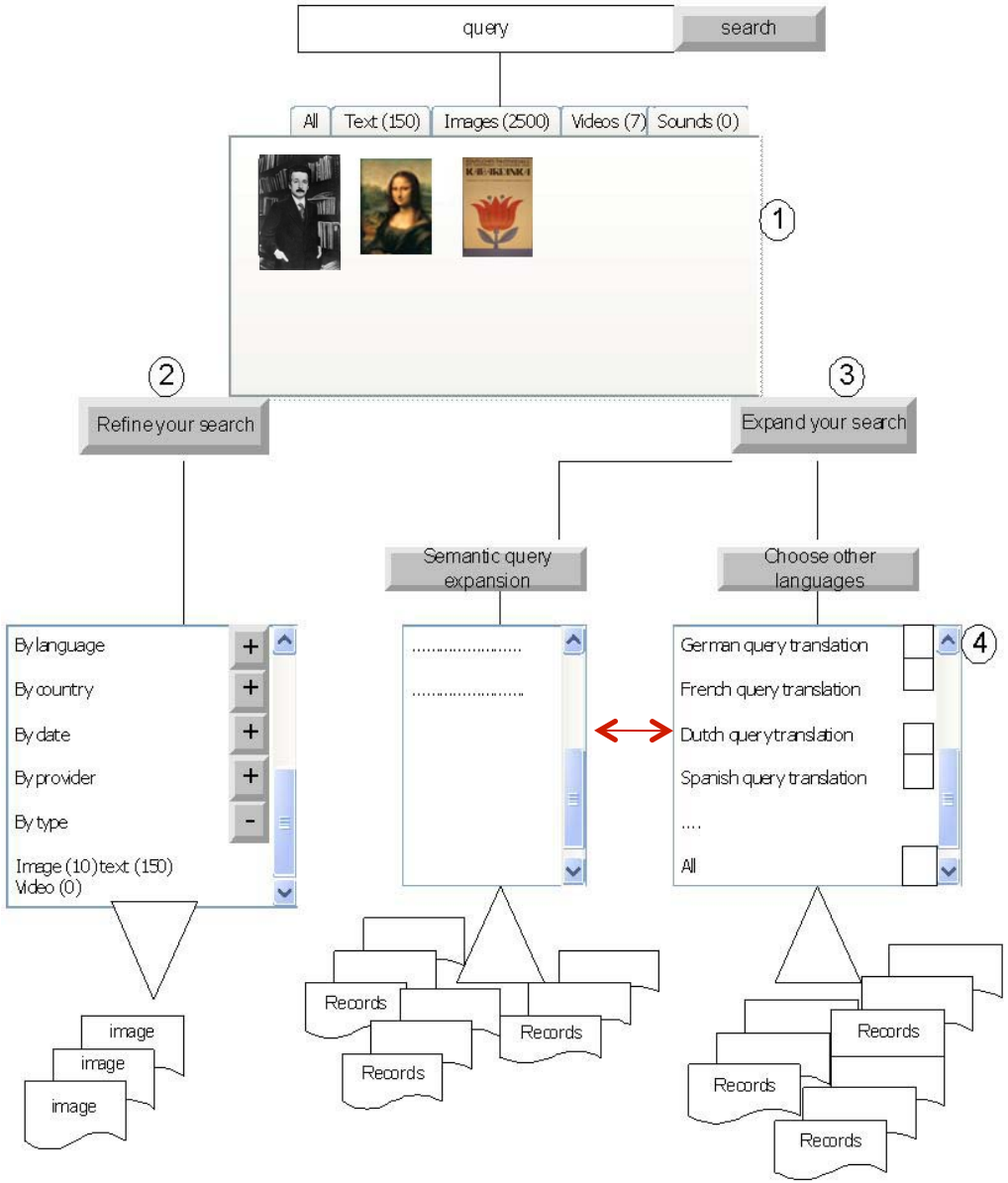


Europeana Query Translation Challenges

- 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

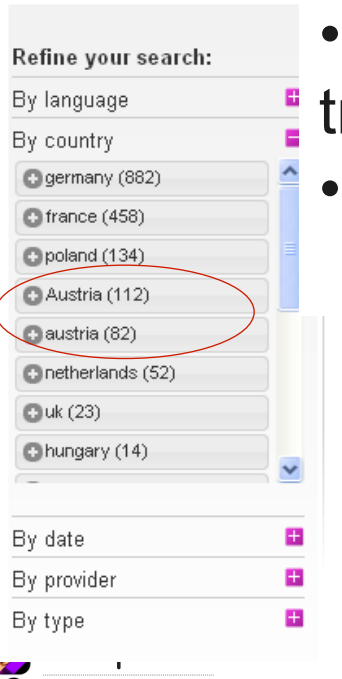
- Multilingual browsing
- Functional Specs for Rhine:
  - Browsing over result sets
  - Index browsing

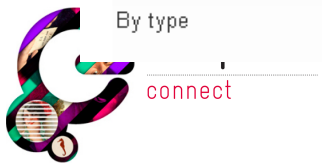
  - Named entities = person browsing / spatial browsing

  → Special case (somewhat easier with authority files and limited lists)

4.2.2 Clustering on the fly (subject drill-down)

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





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 5éme 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?

europeana
connect

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