

Managing Diversity in Knowledge Representation and Information Retrieval

University of Geneva

Gilles Falquet, Jacques Guyot, Claire-Lise Mottaz, Saïd Radhouani

<http://cui.unige.ch/isi>

Supported by Geneva International Academic Network

WSIS 2005

Diversity in knowledge

Communication needs:

Common referential
Common representation
→ UNL Knowledge base

But a common knowledge base cannot and must not be defined centrally by a single institution or group of persons

Diversity needs:

Different contributors representing
→ different nationalities
→ different cultures
→ different social origins
→ etc..

Diversity

World views:

dog is a kind of **pet** (in Europe)

dog is a kind of **food** (in Asia)

Time:

slide (1970) (sheet of plastic)

slide (2005) (Powerpoint)

Use:

car (for the driver, for the mechanics)

Religion/philosophy/values:

pork is a kind of forbidden food (Islam)

pork is a kind of food (...)

Languages:

dog (en), chien (fr), kalb (ar), ...

table (furniture or information display ?)

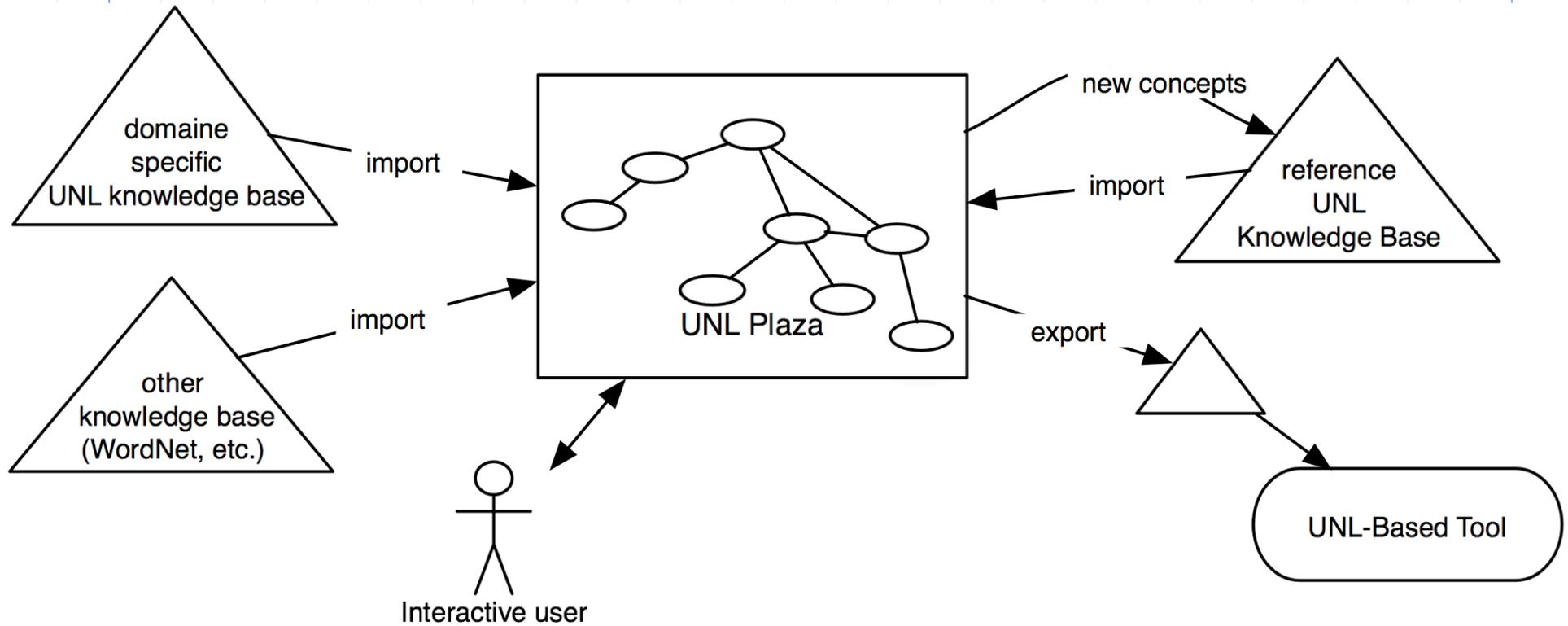
UNL Plaza for knowledge management

Free software platform

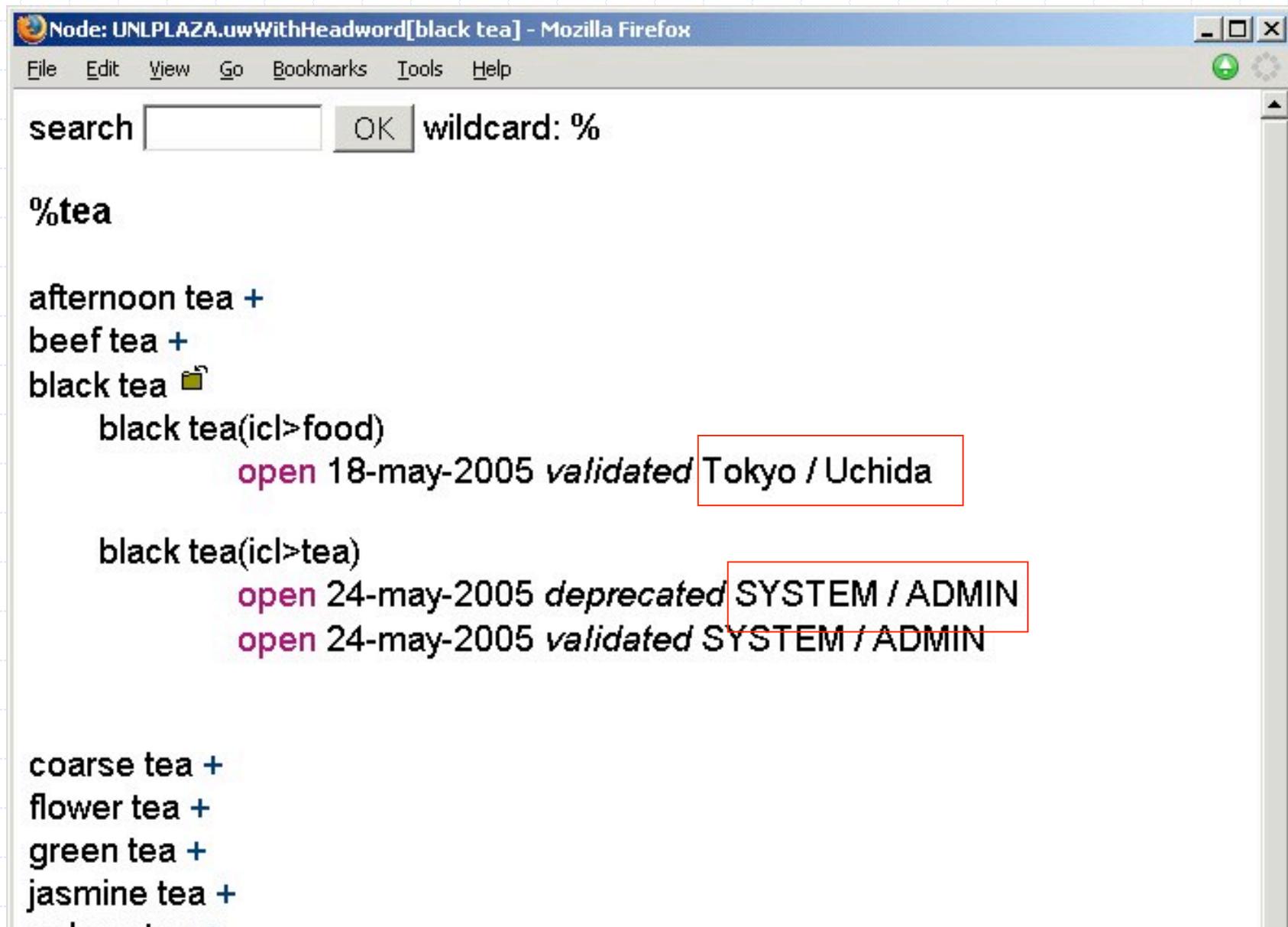
Collaborative development of UNL knowledge bases
on the Web

Support for knowledge diversity

Operation principles



Importing definitions from 2 sources



The screenshot shows a Mozilla Firefox browser window with the title "Node: UNLPLAZA.uwWithHeadword[black tea] - Mozilla Firefox". The address bar is empty, and the menu bar includes "File", "Edit", "View", "Go", "Bookmarks", "Tools", and "Help". The search bar contains the text "search" followed by an empty input field, an "OK" button, and "wildcard: %".

The search results are displayed as follows:

- %tea**
- afternoon tea +**
- beef tea +**
- black tea** 

 - black tea(icl>food)**
 - open** 18-may-2005 *validated* **Tokyo / Uchida**
 - black tea(icl>tea)**
 - open** 24-may-2005 *deprecated* **SYSTEM / ADMIN**
 - open** 24-may-2005 *validated* **SYSTEM / ADMIN**

- coarse tea +**
- flower tea +**
- green tea +**
- jasmine tea +**

On the left side of the browser window, the letters "I", "S", and "I" are stacked vertically.

Linguistic diversity

The screenshot shows a web browser window with the following content:

- Address bar: <http://jg2:1880/lazy/ns?a=UNLKB.focusONseqid&u=Japanes...>
- Menu: [メニュー](#), [探す](#), [ツリ](#), [search-U](#), [new-U](#), [Japanese](#)
- Header: **UNL: table [12276]icl>furniture**
- Section: テーブル icl>家具
- Section: information uw
- Text: uw:(icl>furniture), md:(icl>furniture(icl>functional thing))
table(icl>furniture)
- Section: translation
- List:
 - Arabic: [طاول](#)
 - English: [table](#)
 - Japanese: [テーブル](#)
 - Japanese: [卓](#)
- Section: child

Interface

Terms

Application: Multilingual Information Retrieval

No common international language

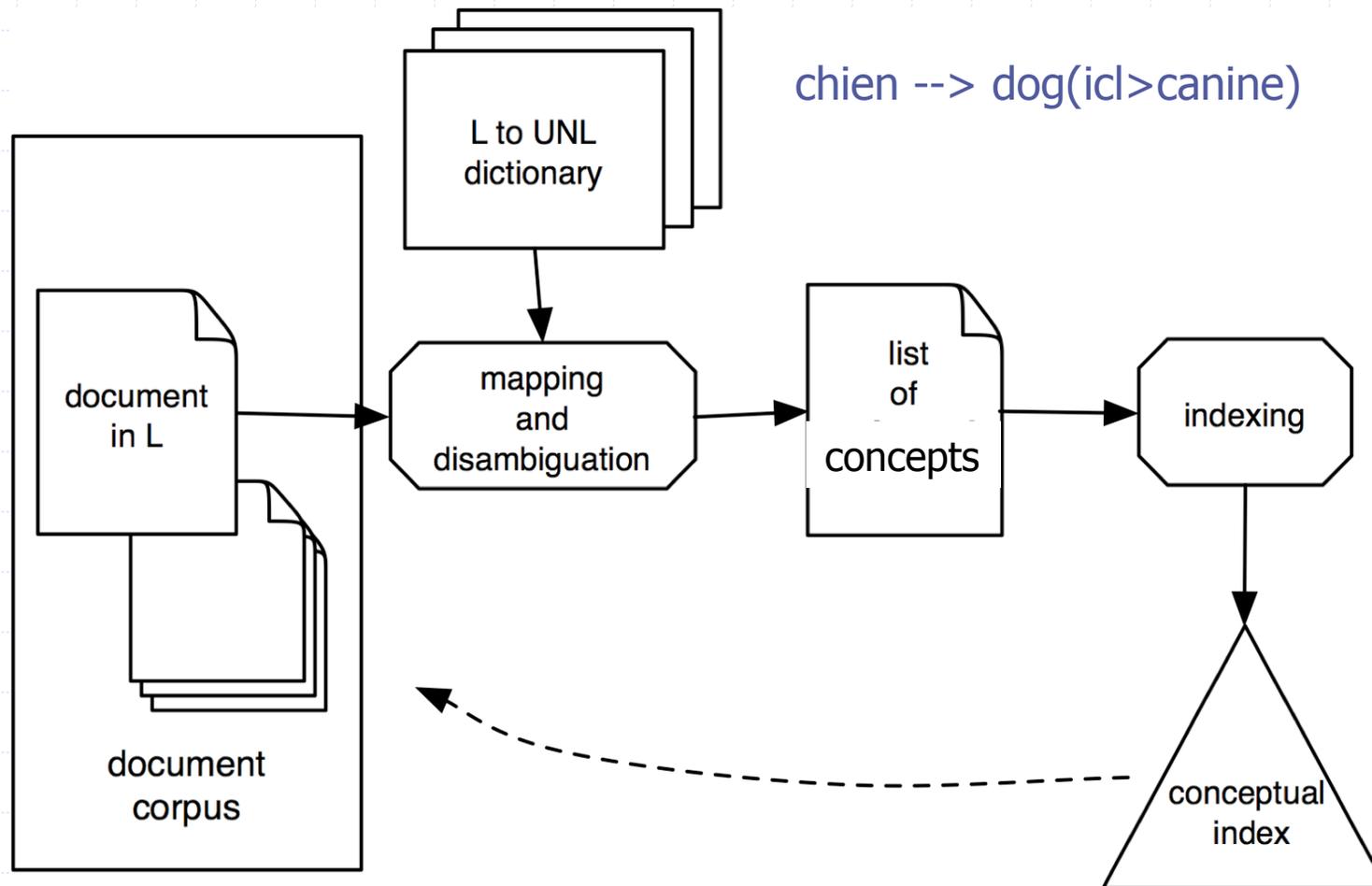
Users can often read and understand documents in several different languages

Documents in non-dominant languages are currently hidden to most users

→ Querying in one language

→ Finding relevant documents in many languages

Technical solution: Conceptual indexing

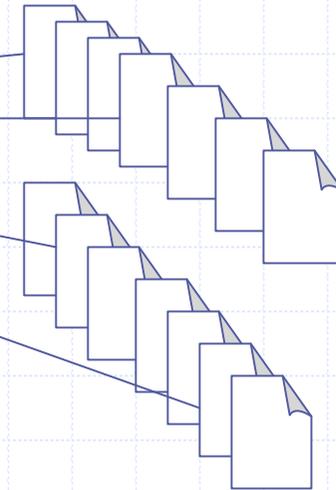


Conceptual indexing

concept id --> documents

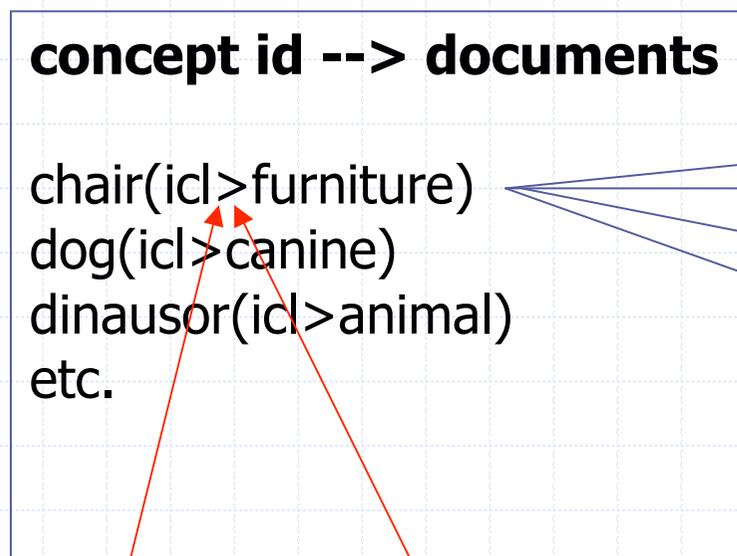
chair(icl>furniture)
dog(icl>canine)
dinausor(icl>animal)
etc.

English documents

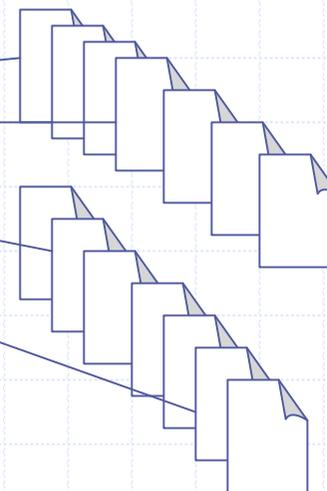


French documents

Conceptual indexing and querying



English documents



French documents

IT --> concept
dictionary

DE --> concept
dictionary

sedia

Stuhl

I
S
I

Example : Query in arabic (dinosaur egg)

The screenshot shows a search engine interface with the following components:

- Search Bar:** Contains two terms: "بيضة" (Egg) and "دينصور" (Dinosaur).
- Results List:** A list of search results with language codes and document IDs. The entry "AR_\$AGZ.950316.0128_\$IT" is highlighted with a red circle.
- Document View:** Displays the content of the selected document. The title is "Cina: estratto dna da uova dinosauro 70.000.000 anni." (China: extracted dna from dinosaur egg 70,000,000 years). The text describes the extraction of DNA from a dinosaur egg in China, mentioning the film 'Jurassic Park' and the professor Zhang Yun.
- Footer:** Shows the search criteria: "[le resultat : 105 docs [19ms = "دينصور" AND "بيضة"]".

Evaluation

CLEF campaign 2005

1.500.000 documents (Newspaper articles)

8 languages: Fr, En, Du, De, It, Fi, Es, Sw

Queries in 10 languages: + Ru and Ar

→ Good results compared to languages dependent approaches

Last minute (DVD - UNITAR-UNESCO 2004)

Printing

term1 <بحث>
term2 <شهادة>

Search

FR Results
EN D:/FR/droit/856/conventions.coe.int/Treaty/FR/Reports/Html/185.htm
ES D:/FR/droit/fgassocies/www.droitweb.com/dw2/articles0814.html
DE D:/FR/droit/fgassocies/www.droitweb.com/dw2/articles441a.html
DU D:/FR/droit/fgassocies/www.droitweb.com/dw2/versionimprimableda7f.html
PT D:/FR/droit/fgassocies/www.fgassocies.com/dw2/articles0814.html
FI D:/FR/droit/fgassocies/www.fgassocies.com/dw2/articles441a.html
SW D:/FR/droit/fgassocies/www.fgassocies.com/dw2/versionimprimableda7f.html
FI D:/FR/droit/gouvernance/loi2000/jorf_nor.htm
IT D:/FR/enjeux/122/www.fao.org/docrep/v9886f/v9886f04.htm
ZH D:/FR/enjeux/93/93.htm
RU D:/FR/enjeux/delgi-bf/www.delgi.gov.bf/Tic/Réglementation/Médias-Texte2.htm
AR D:/FR/enjeux/delgi-bf/www.delgi.gov.bf/Tic/Réglementation/Médias-Texte4.htm
D:/FR/enjeux/delgi-bf/www.delgi.gov.bf/Tic/Réglementation/Protection-Texte7.htm
D:/FR/technique/conception/eve-serveurs/4_4_3.html

Nearest
D:/FR/enjeux/869/forumautremali.htm
D:/FR/droit/donneperso/delis/www.delis.sgdg.org/menu/nir/uja2000.htm
D:/FR/droit/fgassocies/www.droitweb.com/dw2/articles1f55.html
D:/FR/droit/fgassocies/www.fgassocies.com/dw2/articles1f55.html
D:/FR/droit/fgassocies/www.droitweb.com/dw2/articlesfcc3.html
D:/FR/droit/fgassocies/www.fgassocies.com/dw2/articlesfcc3.html
D:/FR/droit/donneperso/commerce/www.juriscom.net/uni/mem/03/introduc.htm
D:/FR/droit/fgassocies/www.droitweb.com/dw2/versionimprimablebd65.html
D:/FR/droit/fgassocies/www.fgassocies.com/dw2/versionimprimablebd65.html
D:/FR/droit/fgassocies/www.droitweb.com/dw2/articles5c06.html

Document



Convention sur la cybercriminalité (STE n° 185)

[English](#)

Rapport explicatif
(adopté le 8 novembre 2001)

I. La Convention et son Rapport explicatif ont été adoptés par le Comité des Ministres du Conseil de l'Europe à l'occasion de sa 109^e Session, le 8 novembre 2001. La Convention a été ouverte à la signature à Budapest, le 23 novembre 2001, à l'occasion de la Conférence Internationale sur la Cybercriminalité.

II. Le texte de ce rapport explicatif ne constitue pas un instrument d'interprétation authentique du texte de la Convention, bien qu'il puisse faciliter la compréhension des dispositions qui y sont contenues.

I. Introduction

1. La révolution des technologies de l'information a changé radicalement la société et continuera vraisemblablement de le faire dans un avenir prévisible. Cette révolution a simplifié bien des tâches. Alors qu'initialement, seuls certains secteurs de la société avaient rationalisé leurs méthodes de travail en s'appuyant sur les technologies de l'information, il ne reste pour ainsi dire plus aucun secteur qu'elles n'aient marqué de leur empreinte. Les technologies de l'information se sont insinuées, d'une manière ou d'une autre, dans tous les aspects des activités humaines.

2. Les technologies de l'information se singularisent notamment par l'impact qu'elles ont eu et continueront d'avoir sur l'évolution des

le résultat : 14 docs [1ms = "<شهادة>" AND "<بحث>"]

Other use of conceptual indexing: bitext alignment (Don Quixote En -- Es)

The screenshot shows a bitext alignment window with the following components:

- Search Bar:** 'search term1' contains 'windmills' (circled in red). 'term2' is empty.
- Language Selection:** 'from' is set to '_EN' and 'to' is set to '_ES'.
- Results Panel:** Shows a list of files under 'C:/AAA/BILANG/DOC'.
- Main Text Area:** Split into 'From' (English) and 'To' (Spanish) columns.
 - From (English):** Contains text from Don Quixote. The word 'windmills' is circled in red.
 - To (Spanish):** Contains the corresponding Spanish text. The phrase 'molinos de viento' is circled in red.
- Nearest Panel:** Lists several files from 'C:/AAA/BILANG/D'.
- Status Bar:** Displays '("windmills")IN["_EN"] = le resultat : 1 docs [21ms]'.

Work in progress and next steps

Field test of the UNL Plaza

Term disambiguation

- based on language diversity

Collaboration to add new dictionaries

Thanks for your attention
and
Welcome for more demonstration