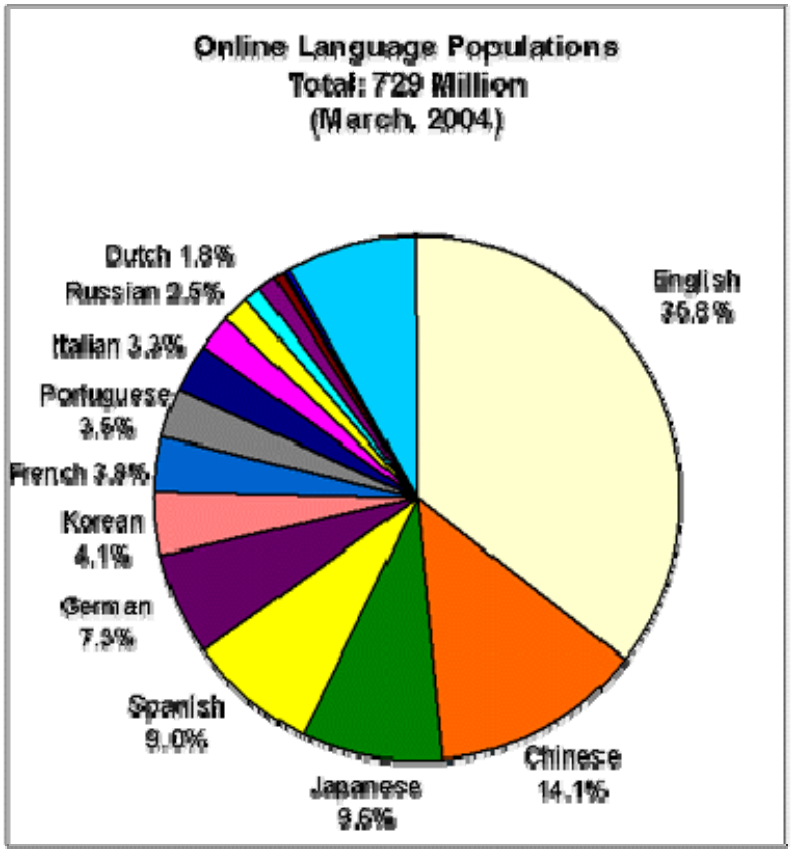
DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

Multilingual Information Access for Digital Libraries

Carol Peters
ISTI-CNR, Pisa

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

There's a lot of "other" languages out there



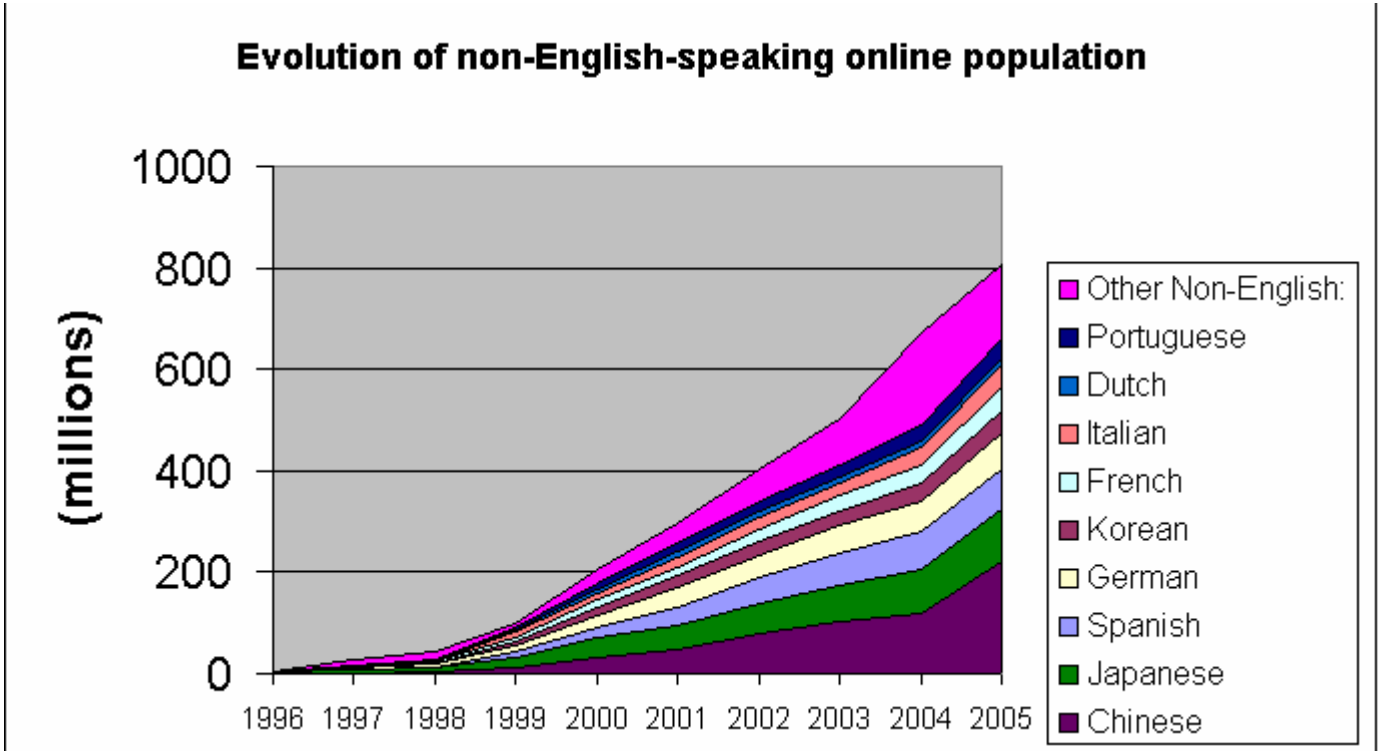
Multilingual users

Multilingual content (e.g. web)

English	68.4%
Japanese	5.9%
German	5.8%
Chinese	3.9%
French	3.0%
Spanish	2.4%
Russian	1.9%
Italian	1.6%
Portuguese	1.4%
Korean	1.3%
Other	4.6%
Total Web pages:	313 B

Source: Vilaweb.com, as quoted by (2003)

Evolution of non-English speaking population



DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

Europe's Linguistic Diversity

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

Multilingual Information Society

- Web as platform for knowledge dissemination and acquisition
  - Distance Learning…..
  - **Digital Libraries…..**
- Content available in many languages

- information providers and seekers should have equal opportunities
- preservation of national languages

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

Europe's Cultural Heritage

- Europe's collective memory is **multilingual**
- Making our historic and cultural heritage available to all citizens for studies, work, leisure via the Internet implies efficient functionality to represent, store, access, interpret and reuse this material whatever the **form, media,** and **language**
- Impact is social, cultural and economic

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

Multilingual Information Access

- Increasing pressure for access to information without language or cultural barriers:
  - Find information in foreign languages
  - Read and interpret that information
  - Merge with information in other languages
- Need for Multilingual Information Access

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

What is MLIA?

- MLIA related research regards the storage, access, retrieval and presentation of information in any of the world's languages.
- Two main areas of interest:
  - multiple language access, browsing, display
  - cross-language information discovery and retrieval

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

Multi-Language Access, Browsing, Display

The enabling technology:

- character encoding
- specific requirements of particular languages and scripts
- localization and presentation

Cross-Language Information Retrieval

Crossing the language barrier…

- querying of multilingual collection in one language against documents in many other languages…
- filtering, selecting, ranking retrieved documents
- presenting retrieved information in an interpretable and exploitable fashion

CLIR methods

- How is it done?
  - Translate: search requests, documents (or both)
- Translation resources
  - Machine Translation (MT)
  - Parallel/comparable corpora
  - Bilingual Dictionaries
- Example problems
  - Handling non-ascii character sets
  - Morphology: inflection, derivation, compounding, …
  - OOV terms, e.g. proper names
  - Multi-word concepts, e.g. phrases and idioms
  - Ambiguity, e.g. polysemy

Users of CLIR systems

- Obvious question…
  - Why do users want to retrieve documents they presumably can't read?
- A few users are truly multilingual
  - Can formulate searches and judge relevance in many languages
  - Want convenience of a single query
- Many users know more than one language
  - Want to query in their native language
  - Can judge relevance even if results not translated
  - Have access to document translation
  - Objects retrieved are language-independent (e.g. images)
- Some users are only monolingual but need access to information in other languages
- Results must be presented in a useful form according to the user needs/profile

Involving the user

- Interactive CLIR systems can help users locate and identify relevant foreign-language documents
  - Formulate and translate the query (e.g. entering diacritics, selecting translation alternatives)
  - Query re-formulation (e.g. selecting query expansion terms)
  - Browsing/navigating results (e.g. translating metadata)
  - Identifying relevant documents (e.g. summarising and translating results)

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

MLIA and Digital Libraries

The neglected problem!

- It's complex!
- It's resource demanding!
- There are other issues to solve!

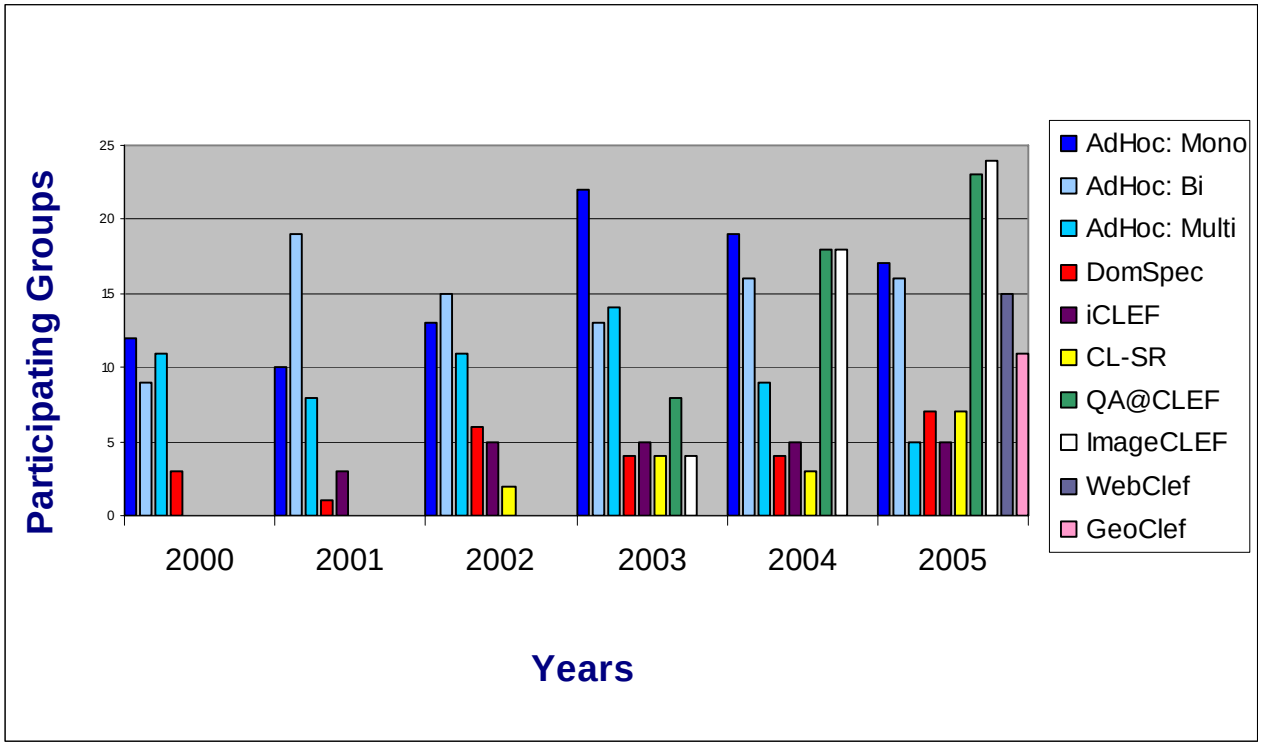
- BUT – everyone agrees - it's important!

Implementing MLIA is Complex

- Multilingual Portals
  - How many languages / how many levels should be multilingual / how to handle updates /linguistic and cultural dependent issues
- Monolingual Search for Multiple Languages
  - encoding and representation issues / language identification / indexing issues (stop words, stemmers, morphological analysers, named entity recognition, ..)
- Cross-Language Search
  - translation resources (dictionaries, corpora, MT systems)
- Presentation of Results
  - in form interpretable and exploitable by user

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

DELOS supports MLIA research via CLEF



From cross-language textual document retrieval towards all aspects of multilingual multimedia information access

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

Existing DL Software Systems

Some kind of multilingual support

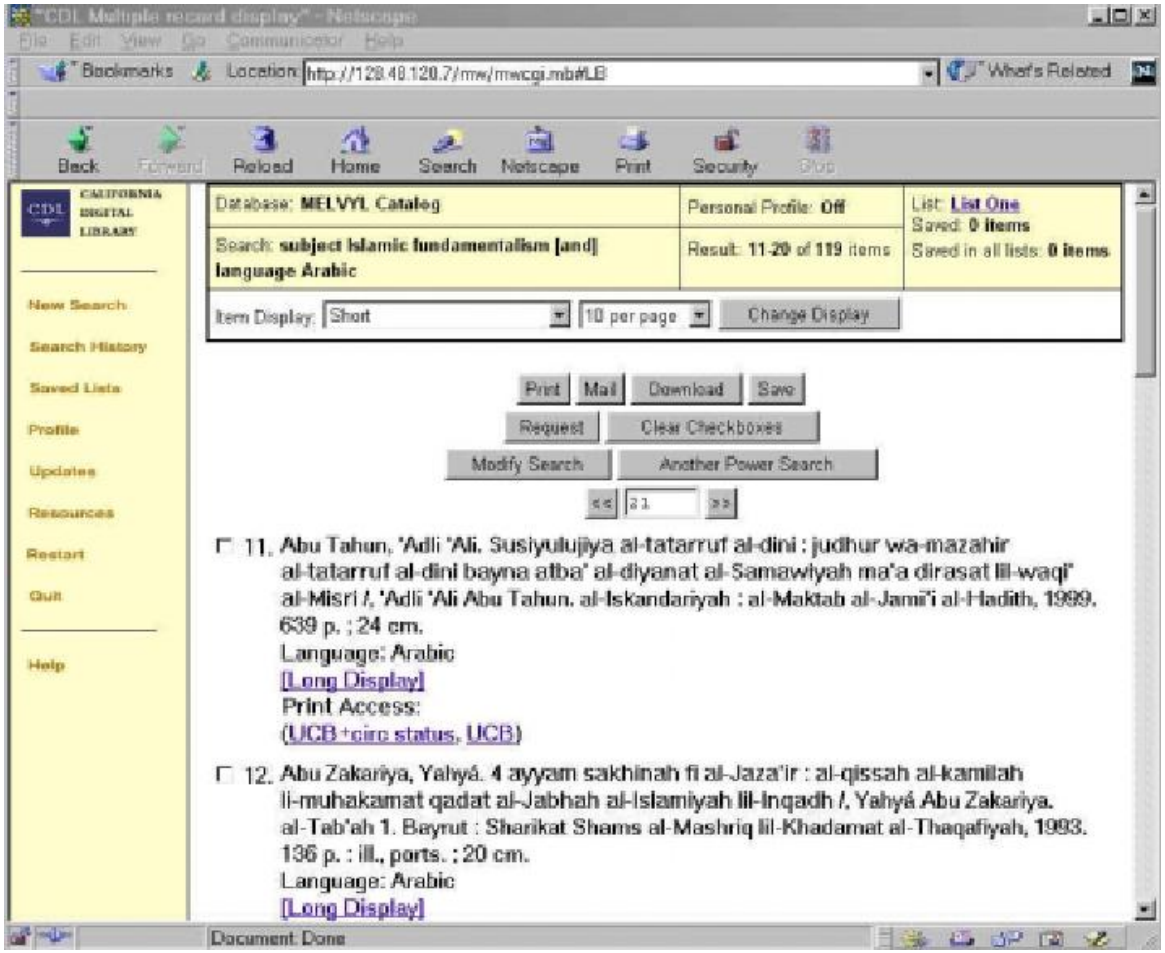
- D-Space
- Greenstone
- Open-DLib
- NSDL

Cross-language functionality

- Cheshire

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

Cheshire Interface



DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

The Challenge

- Bridge the gap between research and application
- Transfer research results to real world
- Make existing resources and methodology generally available
- Raise awareness
- Tackle the un-solved technological issues
- Move from technology-centred to user-centred research

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

Particular Requirements of a European Library

- Languages covered
- Classes of users
- Types of search
- Results presentation

- Solutions must be realistic
- Solutions must be scalable / extendible

What could we have now?

- Multilingual centralised portals
- Support monolingual search in multiple languages
  - Character encoding issues / stopword lists / stemmers / morphological analysers
- Support simple cross-language search
  - querying on metadata (central metadata registry) and keywords
  - dictionary-based search / interlingua or pivot language
  - thesauri for domain-specific search
  - interactive search / browsing functionality
- Present results in a simple fashion

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

A Few Recommendations

- Content description and representation
  - Automatic text classification (in multilingual context)
  - Automatic extraction of metadata and mapping onto a common vocabulary
  - Multilingual ontologies
- User-centred research
  - interfaces (studied to facilitate interaction with user according to linguistic and cultural diversities)
  - results presentation (extraction and merging from multilingual collections/summarization/translation ….)
- Centralised repository for language resources and language processing tools

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

Importance of Standards

- Unicode (http://www.unicode.org/)
- Multilingual Dublin Core http://dublincore.org/groups/languages/
- RDF Encoding of Multilingual Thesauri http://www.w3.org/2001/sw/Europe/reports/thes/8.3
- OWL (Web Ontology Language) http://www.w3.org/TR/2004/REC-owl-features-20040210/

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

i2010: Digital Libraries

MLIA is a key issue and impacts not only on

- Online accessibility of Europe's cultural heritage

but also

- Digitisation
- Preservation and storage