

Multilingual System for Web-Information: The State of The Art

Feiyu Xu
DFKI, LT-Lab
Germany



Multilingual Information System

- Motivation
- Strategies
- MIETTA System



Motivation

- ❑ More and more web information are encoded in other languages than English, for example, Chinese 7%
 - ★ English is loosing its dominance
- ❑ Improve the inter-culture and business communication
 - ★ Quick access to information is important for every day life
- ❑ Text REtrieval Conference (TREC) (<http://trec.nist.gov/>)
 - ★ English, Spanish, Chinese, etc.
- ❑ Information Retrieval for Asian Language Conference (IRAL)
- ❑ European ESPRIT consortium (French, Belgian, German)



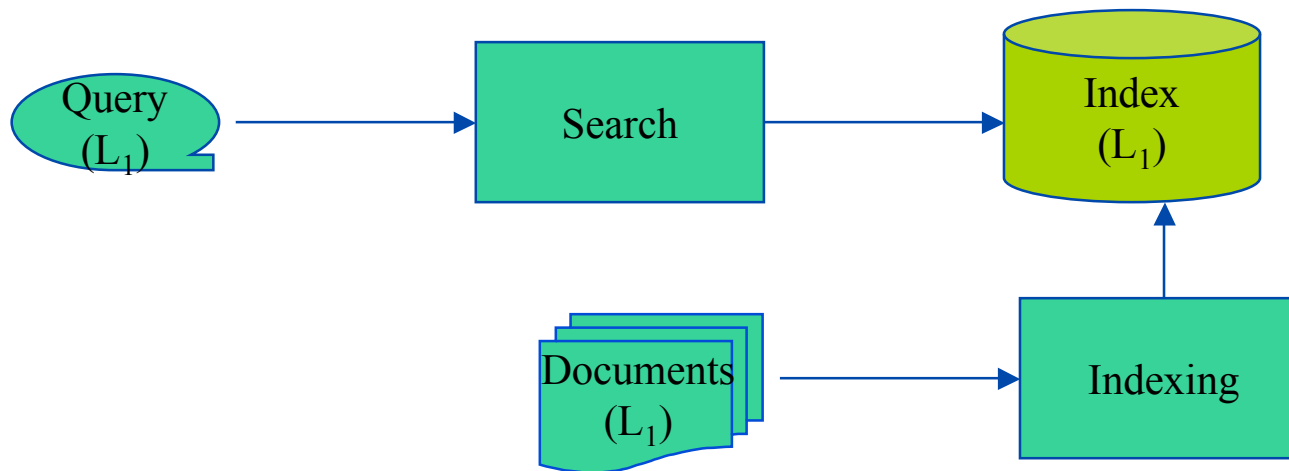
What is Information Retrieval (<http://www.it-world.org>)

- _ **Synonyms:** document retrieval
- _ **Definition:** Information Retrieval is the process of locating information that fits a user's requirements, where the requirements are usually expressed as a search query. The fit of the retrieved information with the information need is referred to as "relevance" ...
- _ http://www.it-world.org/HLT_Survey/ltw-chapter7-2.pdf



What is Monolingual Information Retrieval?

- Query and information to be looked for are encoded in a same language



What is Multilingual Information Retrieval?

- ❑ An extension of the general information retrieval problem
- ❑ Finding information, e.g., web documents which are not encoded in the same language as the query is encoded in
- ❑ Similar terms: “crosslingual information retrieval” and “translingual information retrieval”

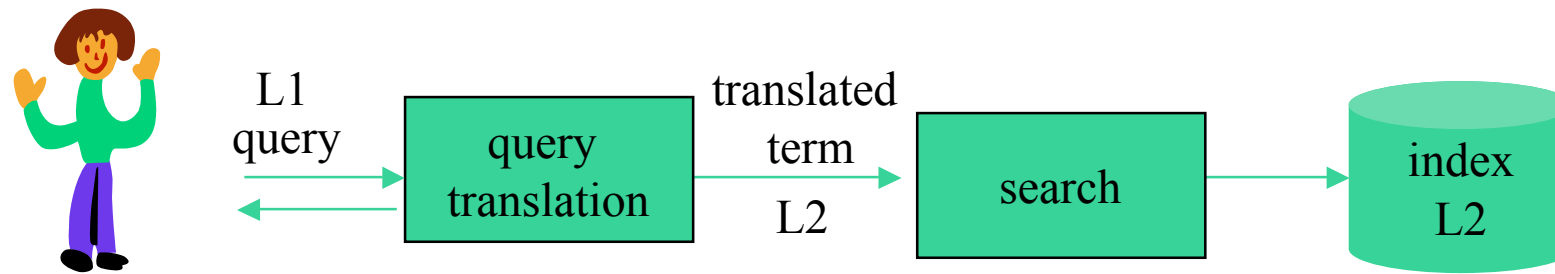


Different Multilingual Information Retrieval Strategies Supported by Language Technologies

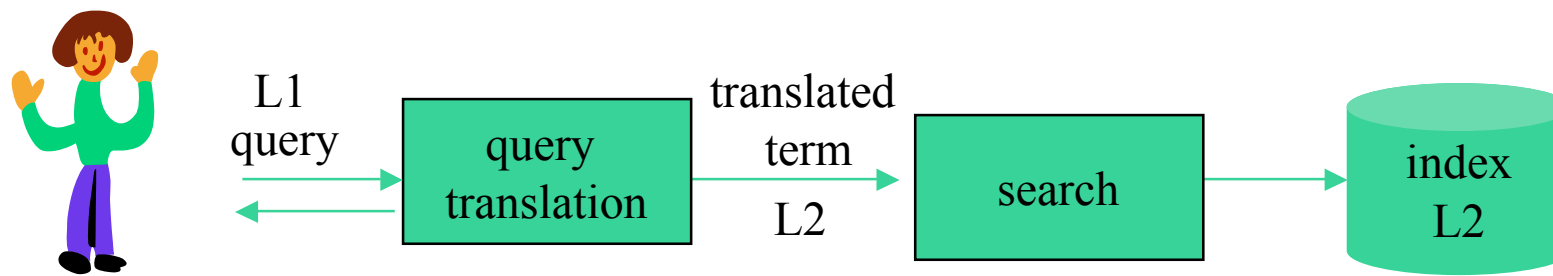
- ❑ Online query translation
 - ★ Help user to formulate his query in a foreign language
- ❑ Online document translation
 - ★ Translate the found document into the query language
- ❑ Offline document translation
 - ★ Make web documents multilingual available
- ❑ Combination of information extraction and multilingual generation
 - ★ Make database information multilingual available and allow the free text retrieval of database information

Query Translation

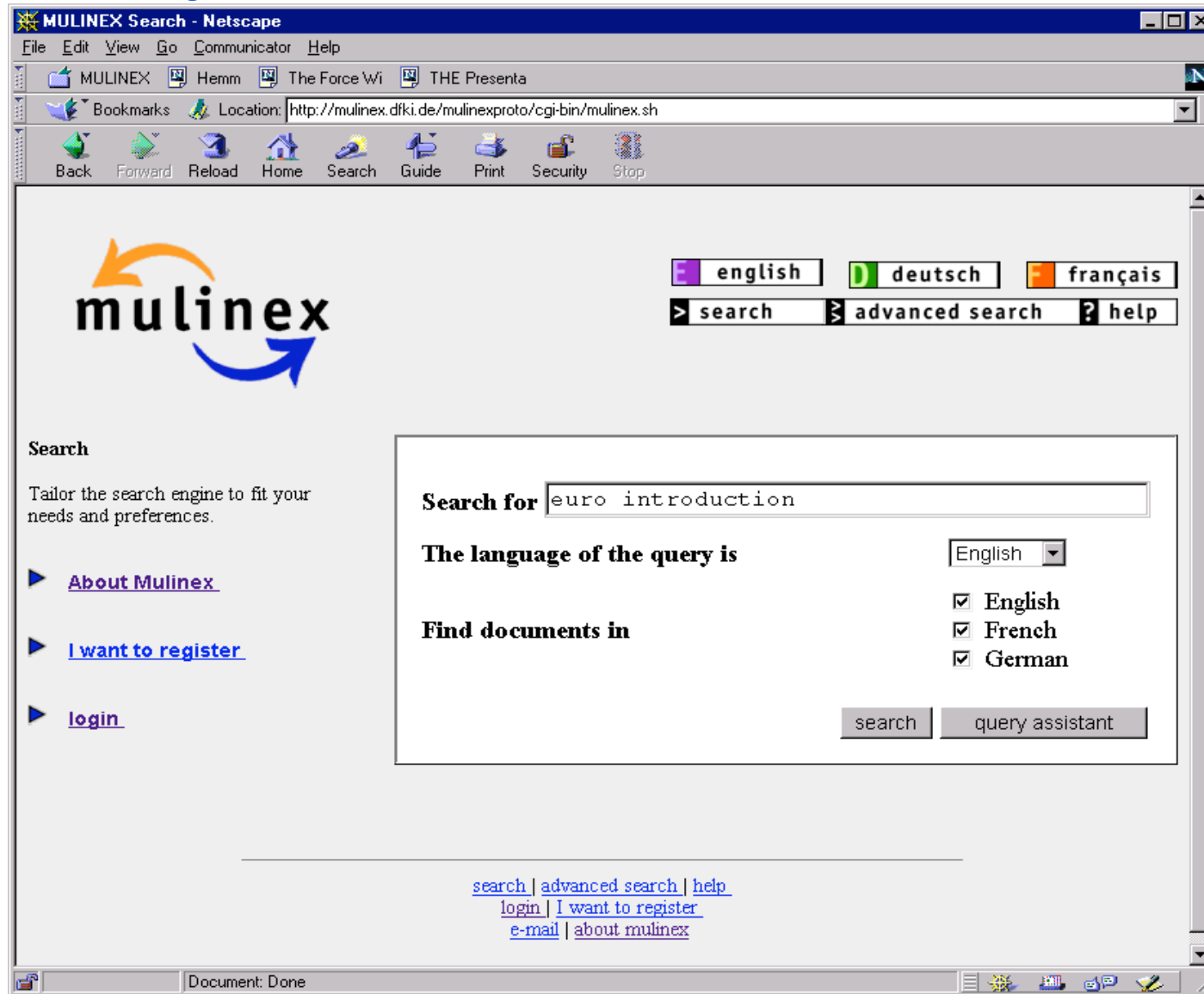
- Help user to formulate their query in another language



- The primary problem is that short queries provide less context for word sense disambiguation, and inaccurate translations lead to bad recall and precision
- How can the user access the content of the found document?



MULINEX System



Source: Feiyu Xu, 2002



MULINEX System

MULINEX Query Assistant - Netscape

Location: <http://mulinex.dki.de/mulinexproto/cgi-bin/mulinex.sh>

Back Forward Reload Home Search Guide Print Security Stop

mulinex

english deutsch français
search advanced search help

Query translation

Your search will be carried out with the following translations of your query. You can modify the translation by:

1. turning off unwanted translations
2. adding your own translations in the text fields

English query terms	French translations	German translations
<input checked="" type="checkbox"/> euro	<input checked="" type="checkbox"/> euro	<input checked="" type="checkbox"/> Euro
<input checked="" type="checkbox"/> introduction	<input type="checkbox"/> instauration <input checked="" type="checkbox"/> introduction <input type="checkbox"/> présentation	<input type="checkbox"/> Empfehlungsschreiben <input checked="" type="checkbox"/> Einleitung <input checked="" type="checkbox"/> Einführung
Germany		

search query assistant

[search](#) | [advanced search](#) | [help](#)
[login](#) | [I want to register](#)
[e-mail](#) | [about mulinex](#)

MULINEX System

The screenshot shows a Netscape browser window titled "MULINEX Search Results - Netscape". The address bar shows the URL "http://mulinex.dfk.de/mulinexproto/cgi-bin/mulinex.sh". The browser interface includes a menu bar (File, Edit, View, Go, Communicator, Help), a toolbar with navigation buttons (Back, Forward, Reload, Home, Search, Guide, Print, Security, Stop), and a bookmarks bar. The main content area features the MULINEX logo, language selection buttons for "english", "deutsch", and "français", and search options: "search", "advanced search", and "help".

Under "Personal search", there are links for "Feedback", "I want to register", and "login". A search form contains the query "euro introduction Germany", a language dropdown set to "English", and checkboxes for "English", "French", and "German". "search" and "query assistant" buttons are at the bottom of the form.

Below the search form, a message states: "You searched for the English query **euro introduction Germany** in German, French, English." A navigation bar shows document counts: "all documents" (100), "deutsch" (22), "français" (57), and "english" (21). A "hide summaries" button is on the right.

A sidebar on the left lists categories: Cooking, Finance, Legal, Macintosh, Medicine, Politics, Taxes, Travel, and Other. The "Legal" category is selected, showing a list of results:

- 1 [Charte PME OEC](#)
Category: Politics, , Legal, Macintosh, Finance
Summary: CHARTE DE LA PREPARATION DES PME □ L'EURO. Questionnaire PME-EURO. Ventes, politique commerciale, marketing. Achats, politique diapprovisionnement, logistique. Gestion financière. Ressources humaines. Système d'information et informatique. Comptabilité, comptes annuels et information de gestion. '
Summary in: English German
<http://www.finances.gouv.fr/euro/charte/98-111-2.htm> Size 18 K
- 2 [Europarl: der Euro - VIERTE WAHLPERIODE \(1994-1999\)](#)
Category: Legal, Politics, Finance, Travel
Summary: VIERTE WAHLPERIODE (1994-1999). 1. Legislativberichte. 2. Nichtlegislative Berichte. 3.LAUFENDE ARBEITEN. Europarl: der Euro - VIERTE WAHLPERIODE. VIERTE WAHLPERIODE Legislativberichte Nichtlegislative Berichte Laufende Arbeiten 1. Legislativberichte A4-0379/96 - PV 28. 11. 96 - AB1

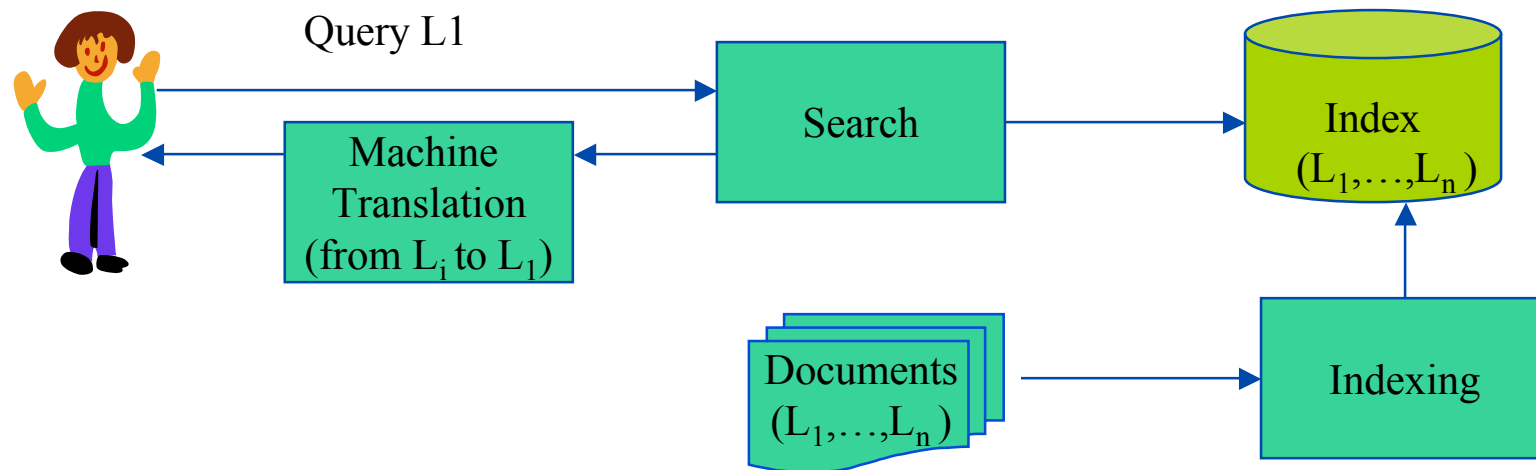
The browser status bar at the bottom shows "Document: Done".

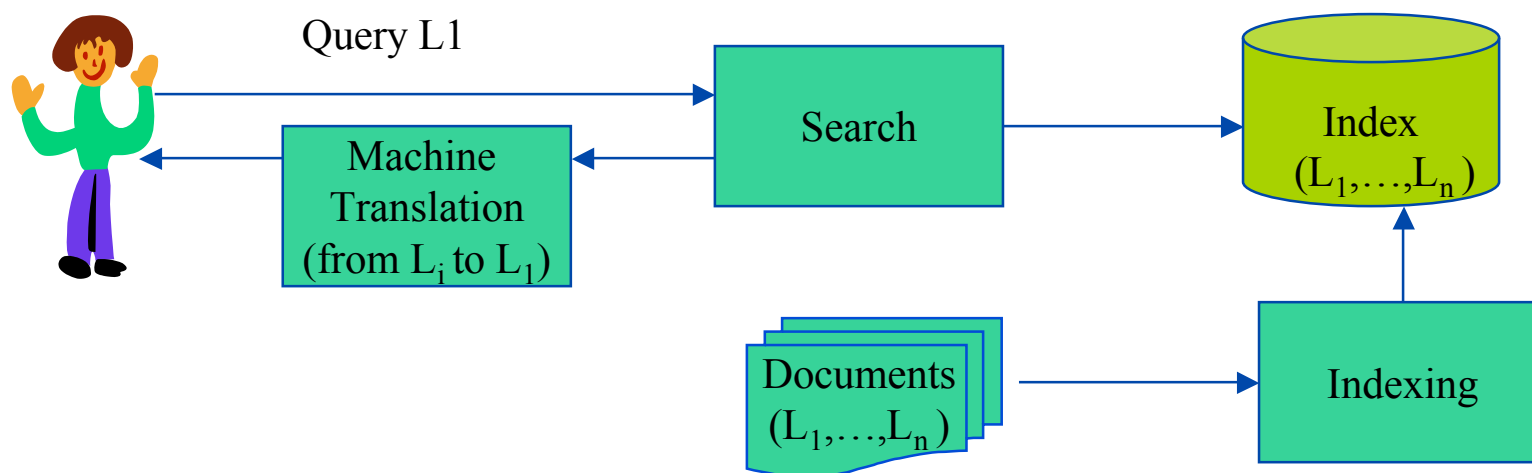
Source: Feiyu Xu, 2002



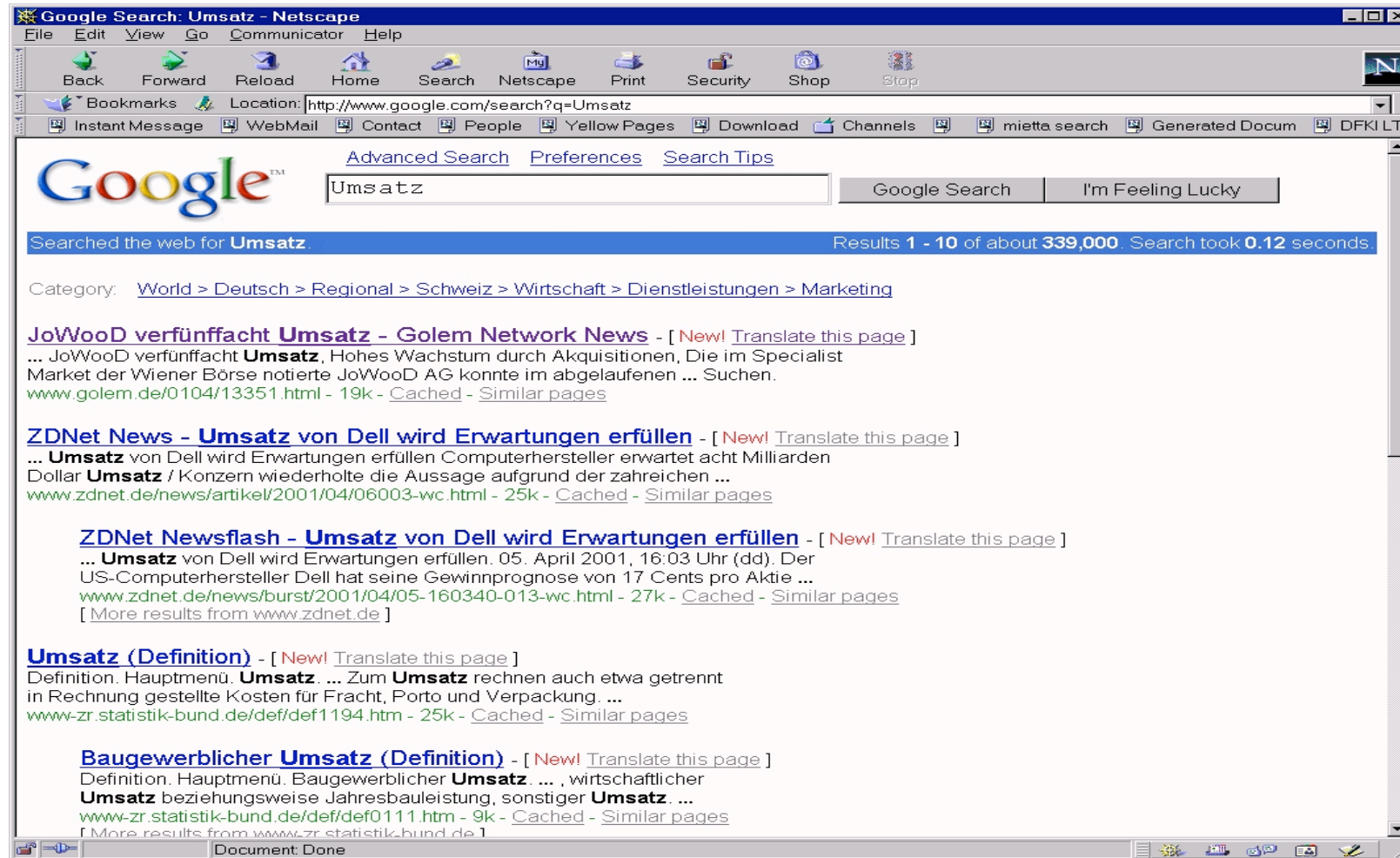
Online document translation

- Translating the found the documents into query language, for example, google





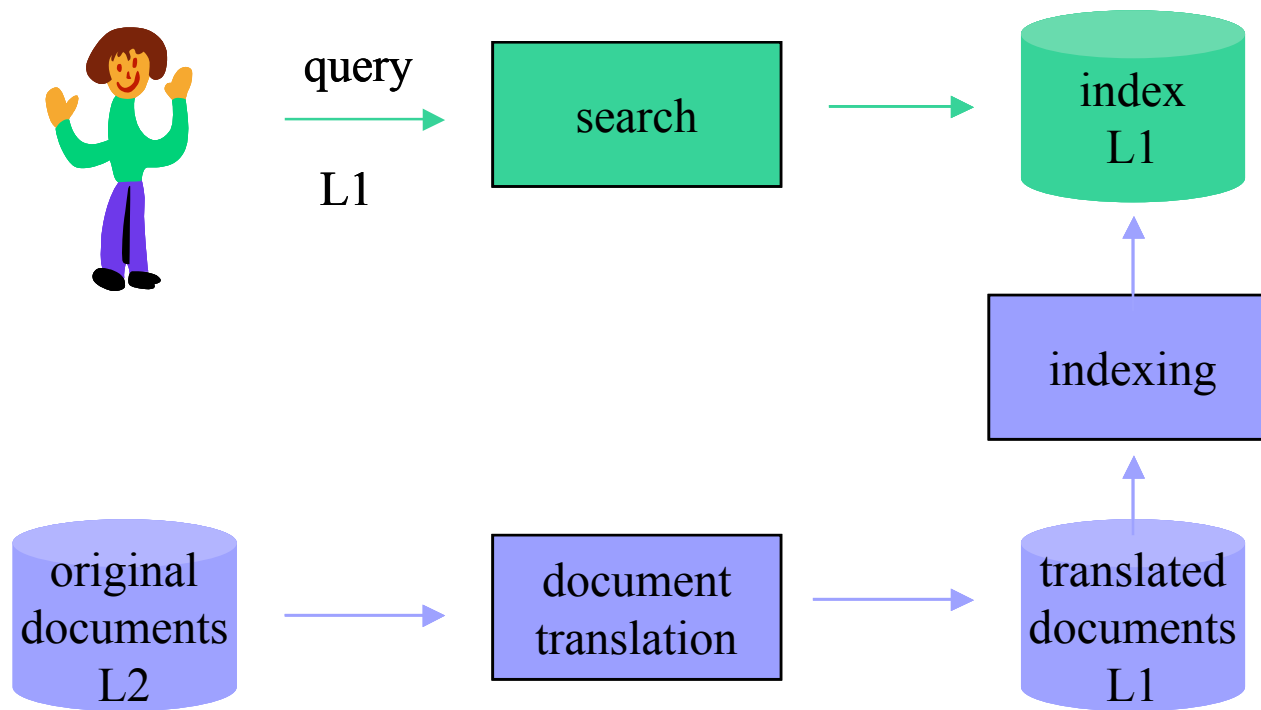
Online document translation (Google)

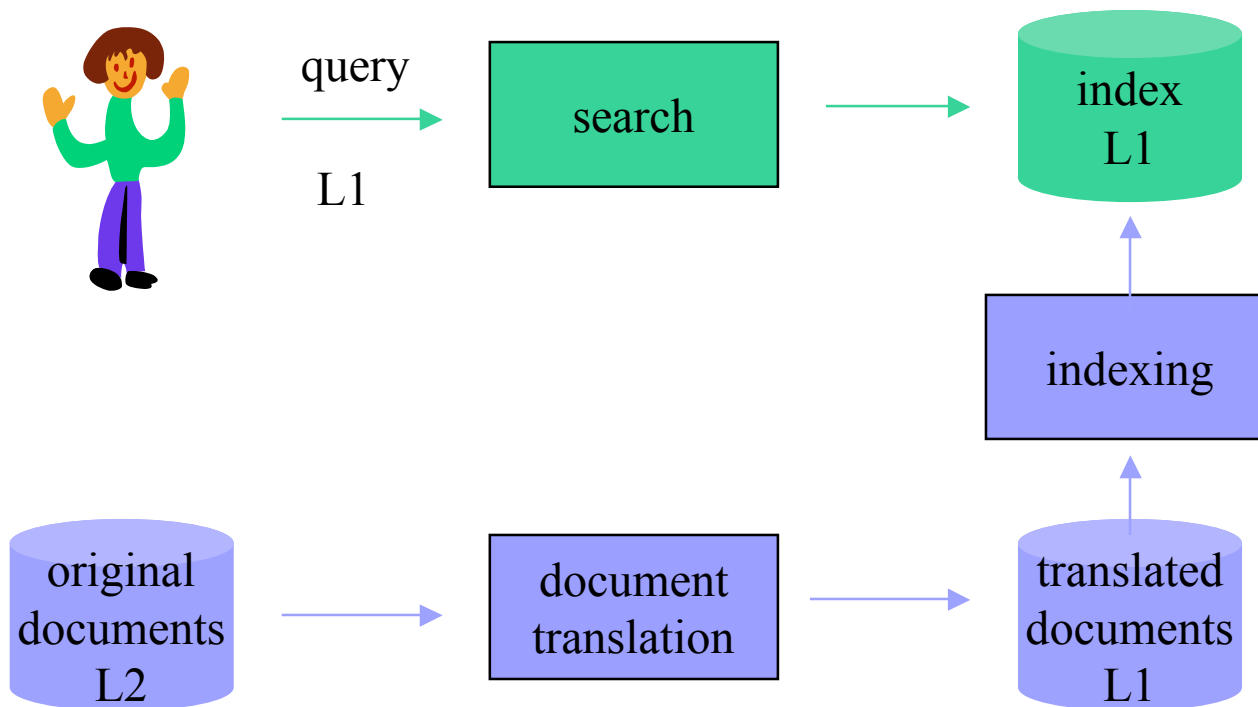


Offline Document Translation

□ Automatic offline translation

- ★ Source text is translated into target languages
- ★ Index is constructed from translation
- ★ Search term in one language yields original and translated documents





Offline Document Translation

- ❑ A higher translation and retrieval performance, since the full original document provides more context for disambiguation. The word sense disambiguation problem is less severe than query translation
- ❑ The main limitation is the duplication of the indices, and the translated documents also need to be stored
- ❑ The offline translation is practically not viable due to big cost of computation and storage for the general search engines like Alta-Vista, Yahoo, etc.



Facts Sheet - MIETTA

- ❑ Title: MIETTA -Multilingual Information Extraction for Tourism and Travel Assistance
- ❑ Funding: EU Language Engineering Sector of TAP (HLT-IST)
- ❑ Technical Partners: DFKI, Celi, University of Helsinki, Polito, Unidata
- ❑ User Partners: Commune DI Rome, City of Turku, Staatskanzlei of the Saarland



Objectives

- ❑ Multilingual internet portal and specialised information system for tourist information
 - Five languages: English, Finnish, French, German, Italian
 - Three regions: Rome, Saarland and Turku

- ❑ Integrated access to heterogeneous data sources and make it fully transparent to end users whether they are searching in
 - ★ WWW documents or
 - ★ Databases

Offline Document Translation in MIETTA

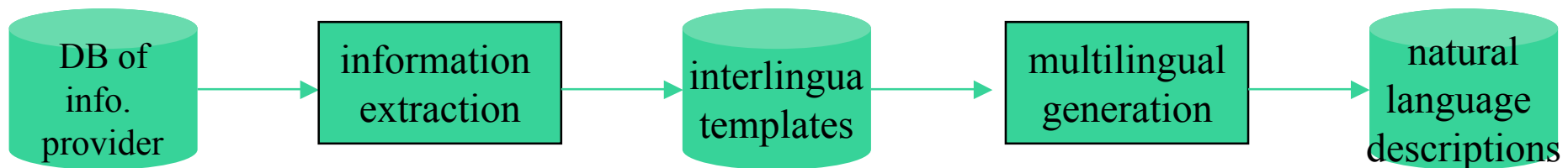
- ❑ Use document translation as the main strategy. The reason is that it allows direct access to the content, it provides better performance within a restricted domain
- ❑ Use LOGOS for document translation, which covers the following directions:
 - ★ German⇒ English, French, Italian
 - ← English⇒ French, German, Italian, Spanish
- ❑ The final document collection in MIETTA after the document translation yielded an almost fully covered multilingual setup.

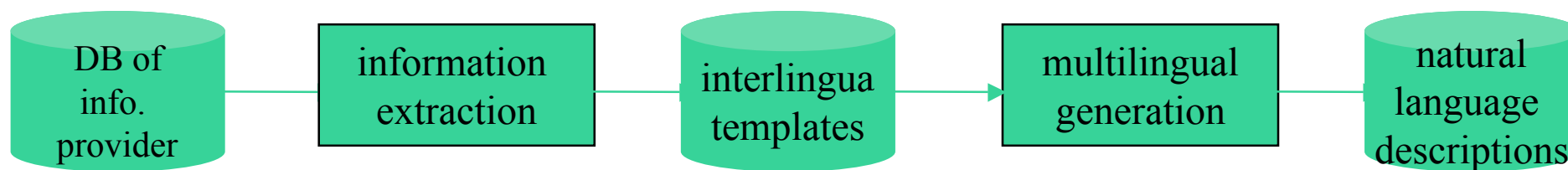


Information Extraction and Multilingual Generation

□ Motivation

- ★ Make the database content more structured and multilingual accessible.
- ★ Apply the same free text retrieval method to the generated descriptions as to the web documents





Source: Feiyu Xu, 2002



- The objective of information extraction is twofold:
 - ★ To extract the domain relevant information (templates) from the unstructured data so that the user can access more facts and more accurately
 - ★ To normalise the extracted data in a language independent format to facilitate the multilingual generation
- Three steps for template extraction in MIETTA
 - ★ Natural language shallow processing: named entities, np, vp
 - ★ Normalisation: converting information into a language independent format
 - ★ Template filling: mapping the extracted information into template slots by employing specific template filler rules

Example of IE

German text from an event calendar in Saarland

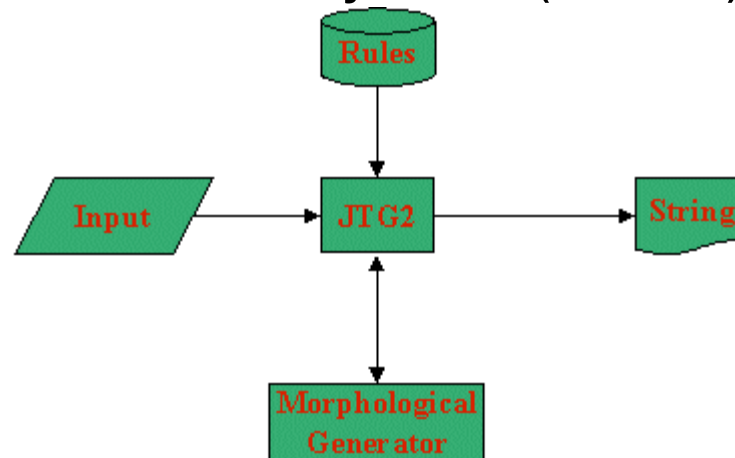
St. Ingbert: -Sanfte Gymnastik für Seniorinnen und Senioren, **montags** von 10 bis 11 Uhr im **Clubraum, Kirchengasse 11.**

English: St. Ingbert: -Gentle Gymnastic for seniors, every Monday from 10:00 to 11:00 am, in Club room, Kirchengasse 11

Event:	Name:	gymnastic
	Addressee:	seniors
	time:	start time: 10 end time: 11 weekly: yes weekday: 1
	location:	city name: St. Ingbert address: Club room Kirchengasse 11

Multilingual Generation

□ Template Generation system (JTG/2)



- Language independent input allows for easy extension of the generation component to other languages

Example

Level1: Event
Level2: Theater
Level3:
Event-Name: Faust
StartDate: 21.10.99
PlaceName: Staatstheater
Address: Schillerplatz, 66111 Saarbrücken
Phone: 0681-32204

English:

The theater show Faust will take place at the Staatstheater in Schillerplatz 1, 66111 Saarbrücken (in the downtown area).

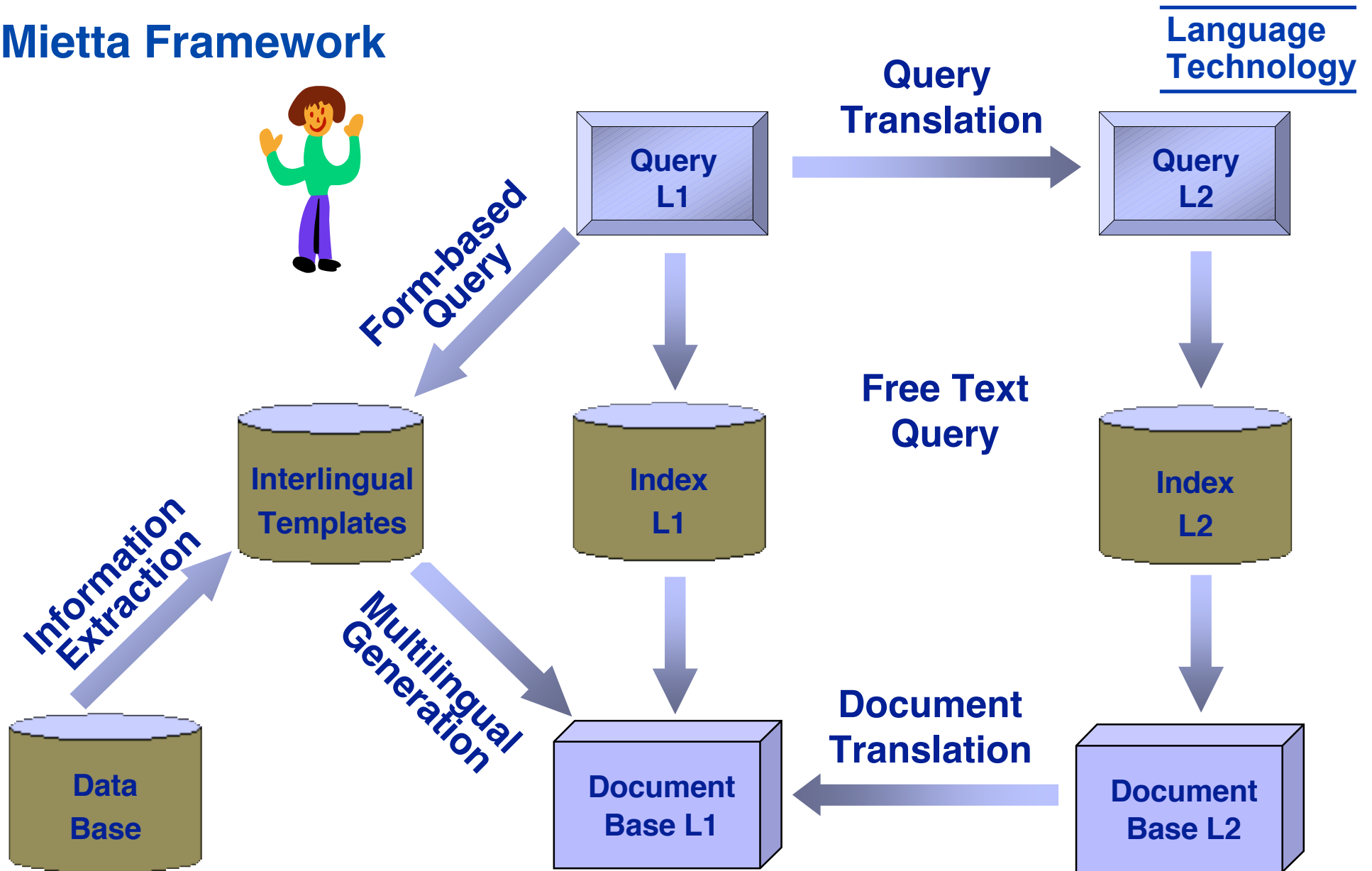
The scheduled date is Thursday, October 21, 1999. Phone: 06 81-32204

Finnish:

Teatteriesitys Faust järjestetään Staatstheaterissa, osoitteessa Schillerplatz 1, 66111 Saarbrücken (keskustan alueella). Tapahtuman päivämäärä on 21. lokakuuta 1999. Puhelin: 06 81-32204.



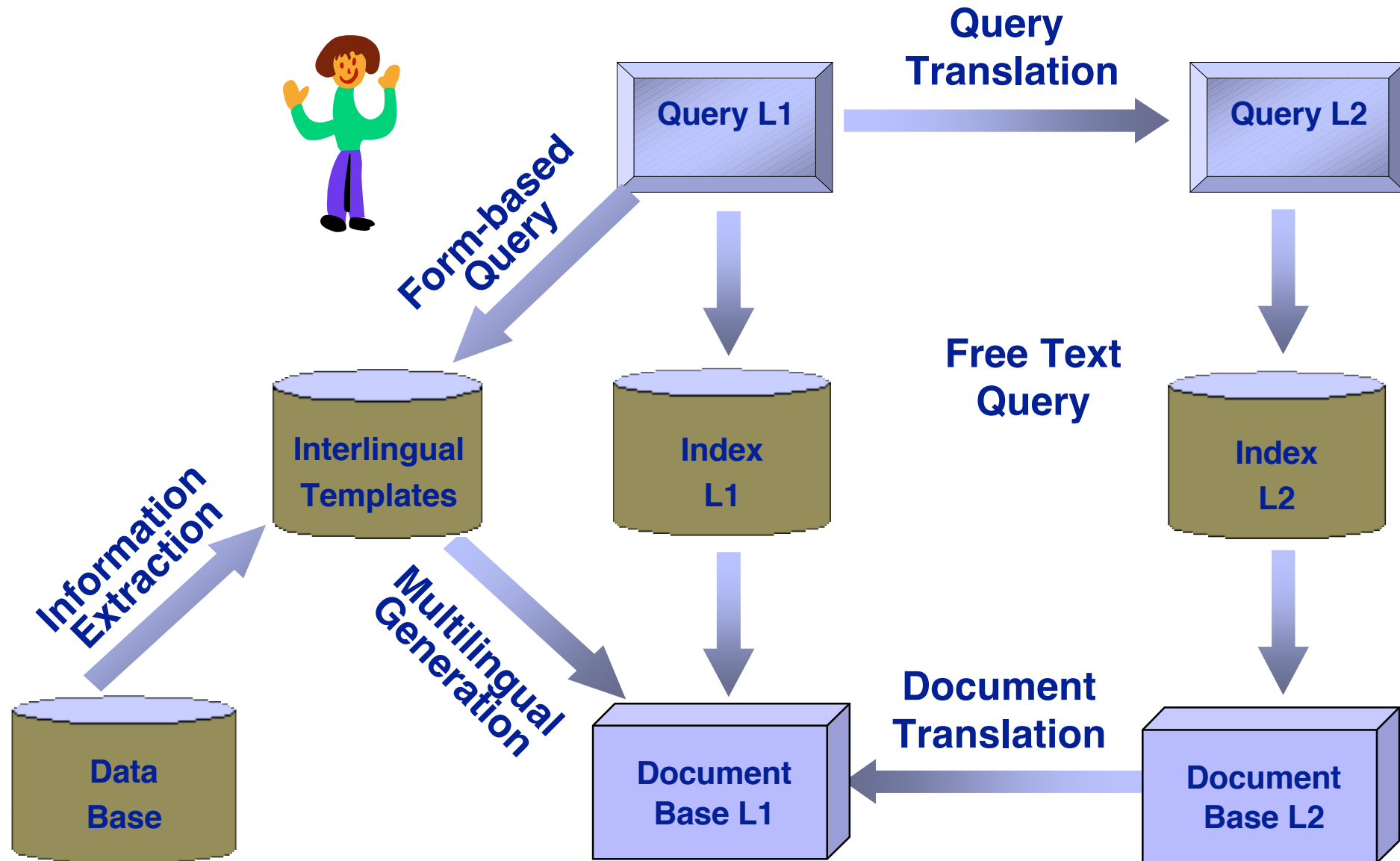
Mietta Framework



Source: Feiyu Xu, 2002

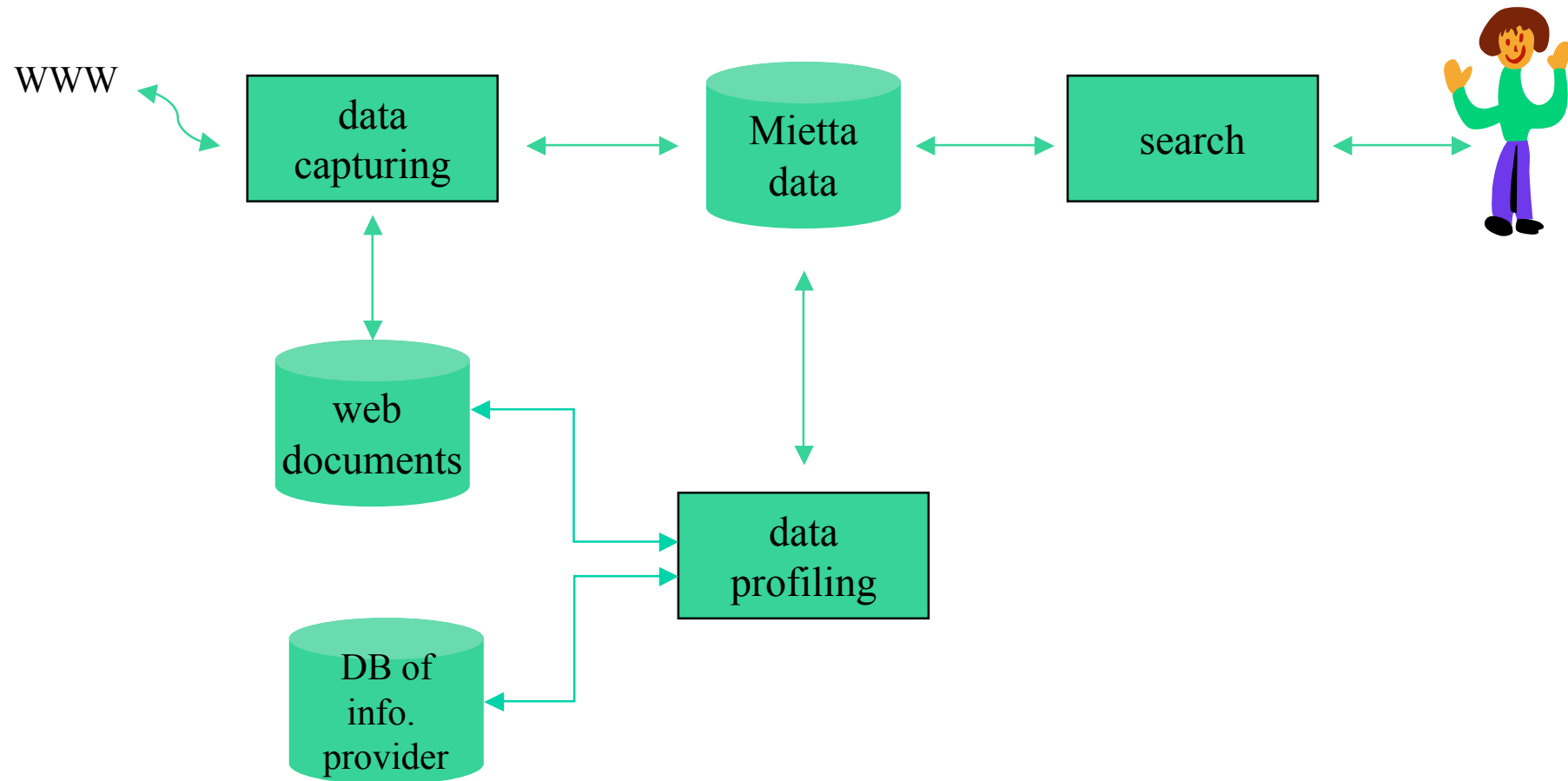


Language Technology



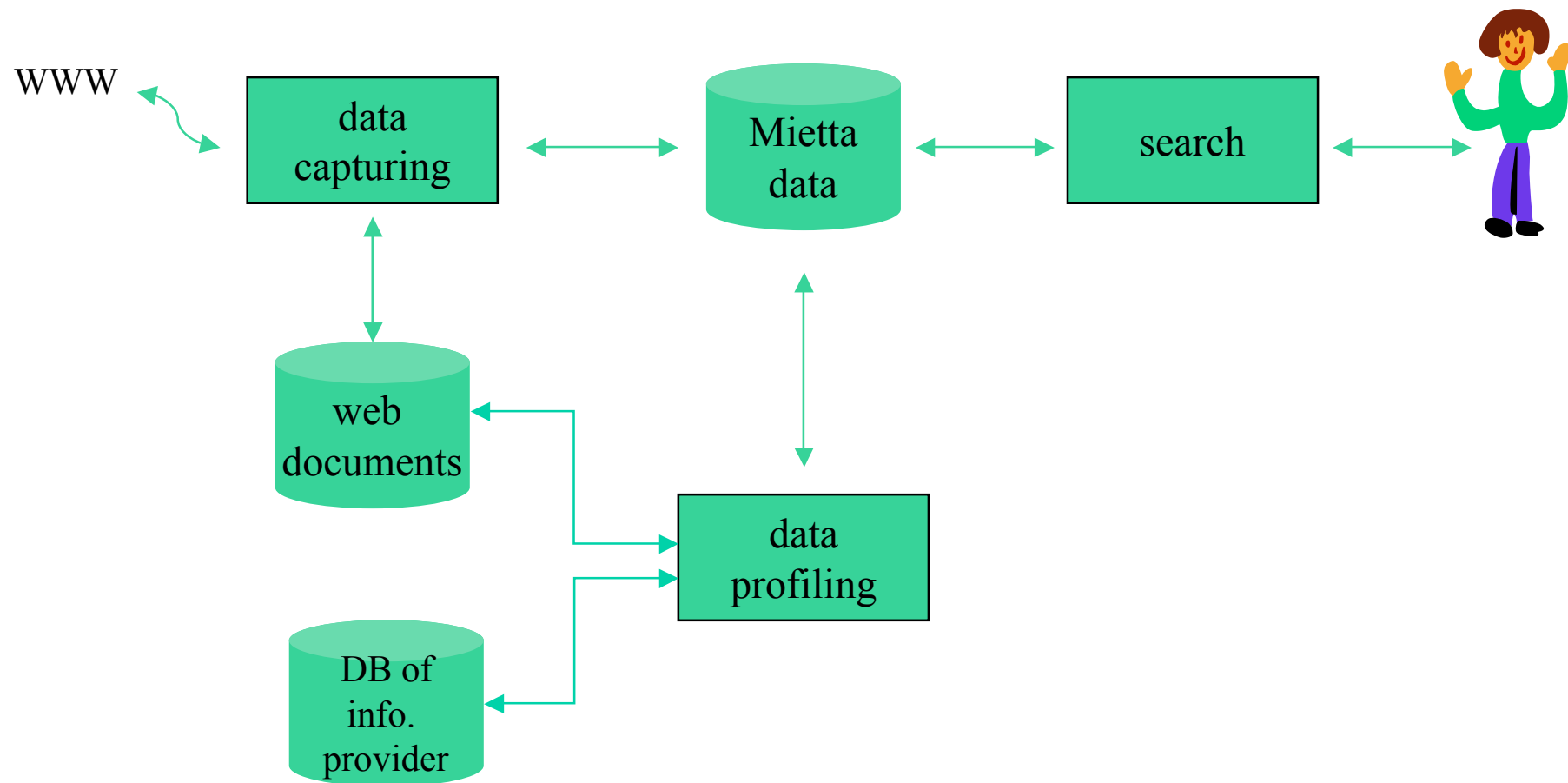
The Overall MIETTA System

Language
Technology



Source: Feiyu Xu, 2002

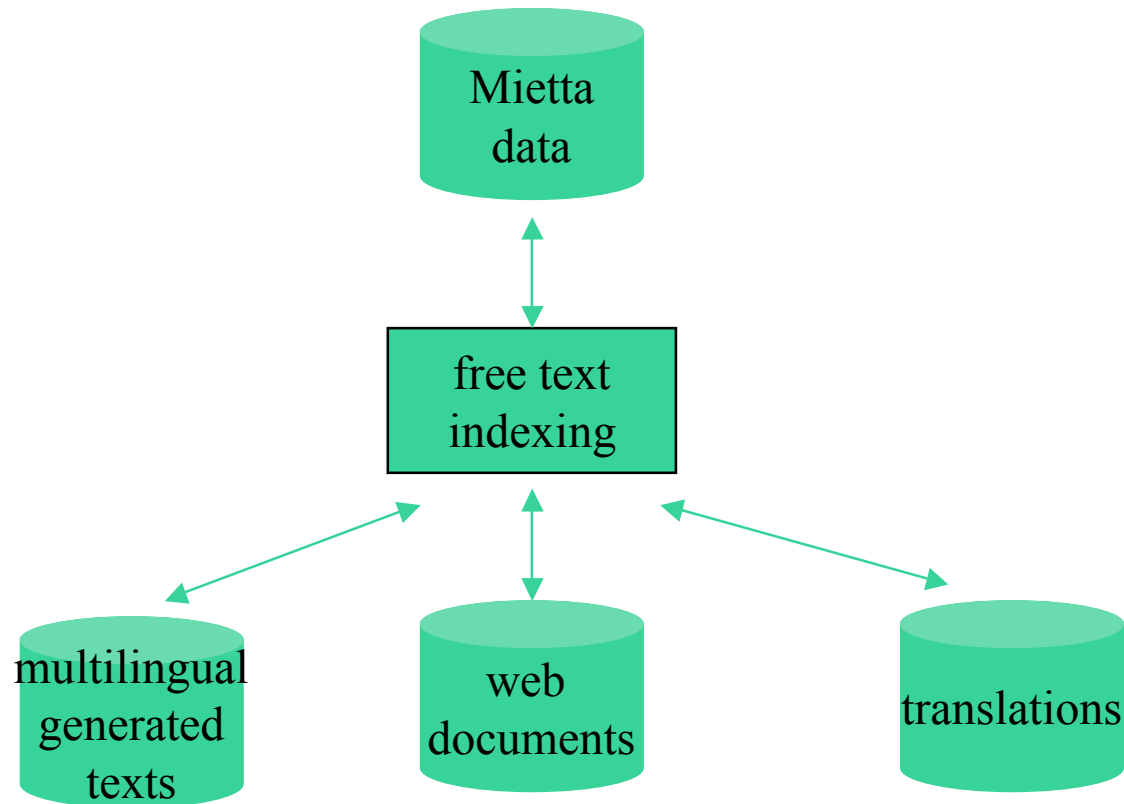


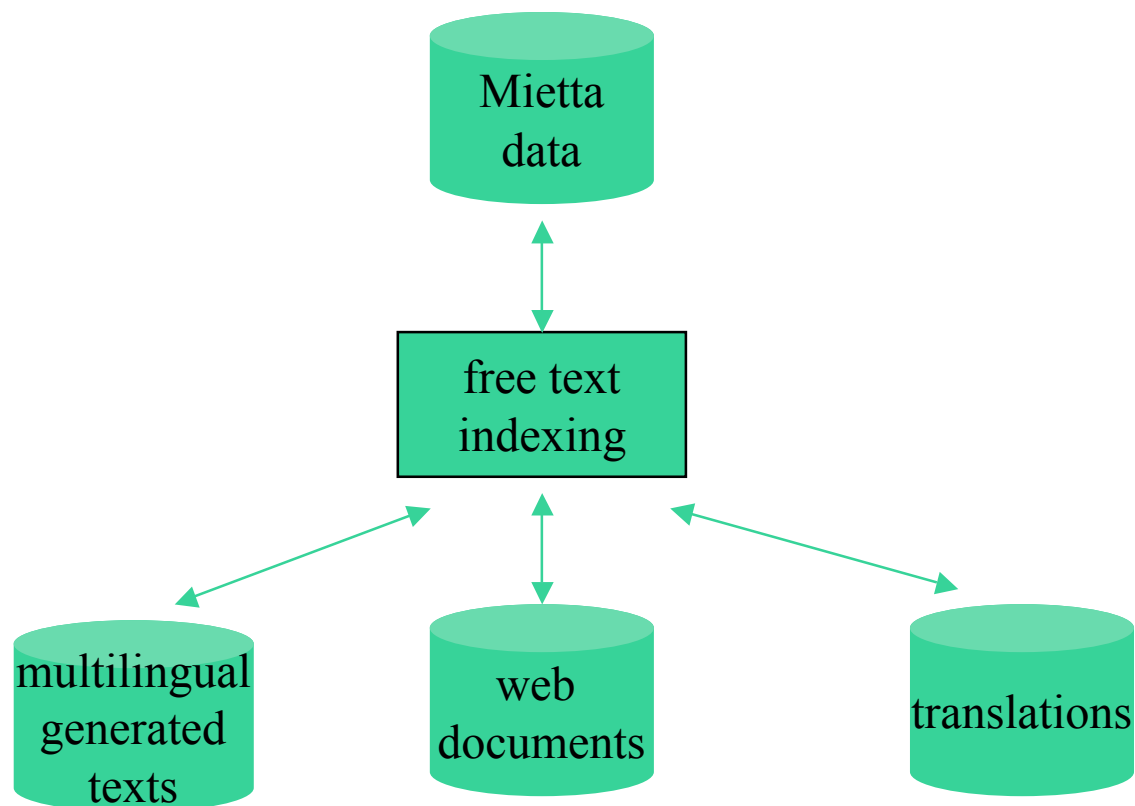


- ❑ Document translation, based on LOGOS machine translation system
- ❑ Information Extraction from database entries for template construction
- ❑ Multilingual generation from templates to obtain natural language descriptions
- ❑ Free text indexing

TNO (ISM, VSM) Indexing Toolkit

- ❑ ISM: A lemma-based fuzzy index based on trigrams
- ❑ VSM: A vector space model index based on lemmatas





- Adaptation to other domains
 - ★ Domain specific templates
 - ★ Domain Concept hierarchy
 - ★ Domain specific template filler rules
 - ★ Domain specific generation grammars
- Extension to other languages
 - ★ Natural language generation tool requires less effort for the development of a grammar rule set in a language
 - ★ Information extraction requires available language specific resources
 - ★ Document translation is dependent on the machine translation system

Evaluation of the MIETTA System

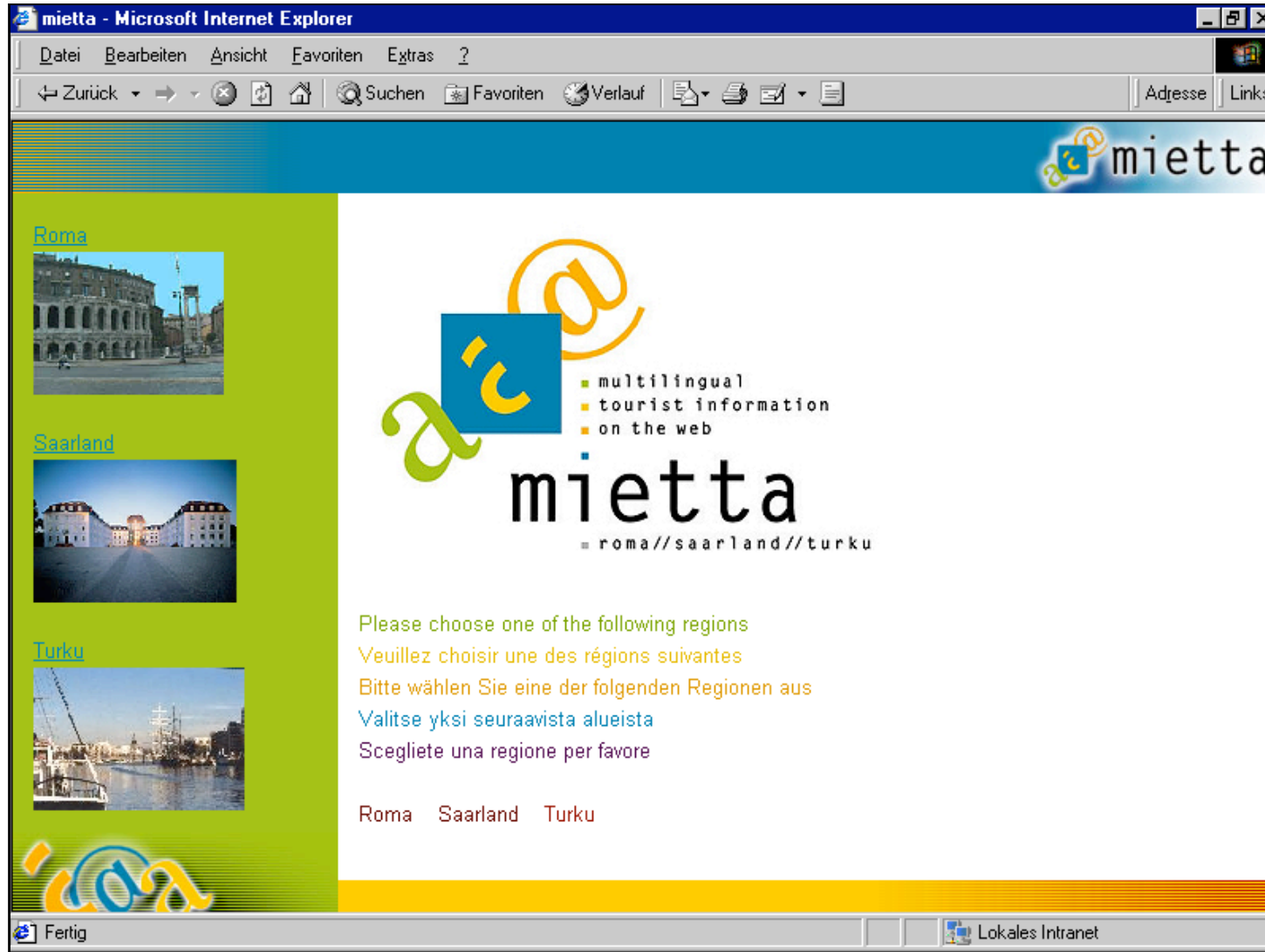
- The standard relevance assessment model used in ad hoc and routing forums of TREC is difficult to apply to the complete MIETTA system because of
 - ★ Broad variety of search strategies
 - ★ Heterogenous data sources
- MIETTA is evaluated as technically “excellent” by EU
- Two projects are derived from MIETTA
 - ★ Natural science foundation of China Project in SJTU
 - ★ EU project of MIETTA to transfer the idea into product in XtraMind in Saarbrücken

Conclusion: Innovative Technical Features

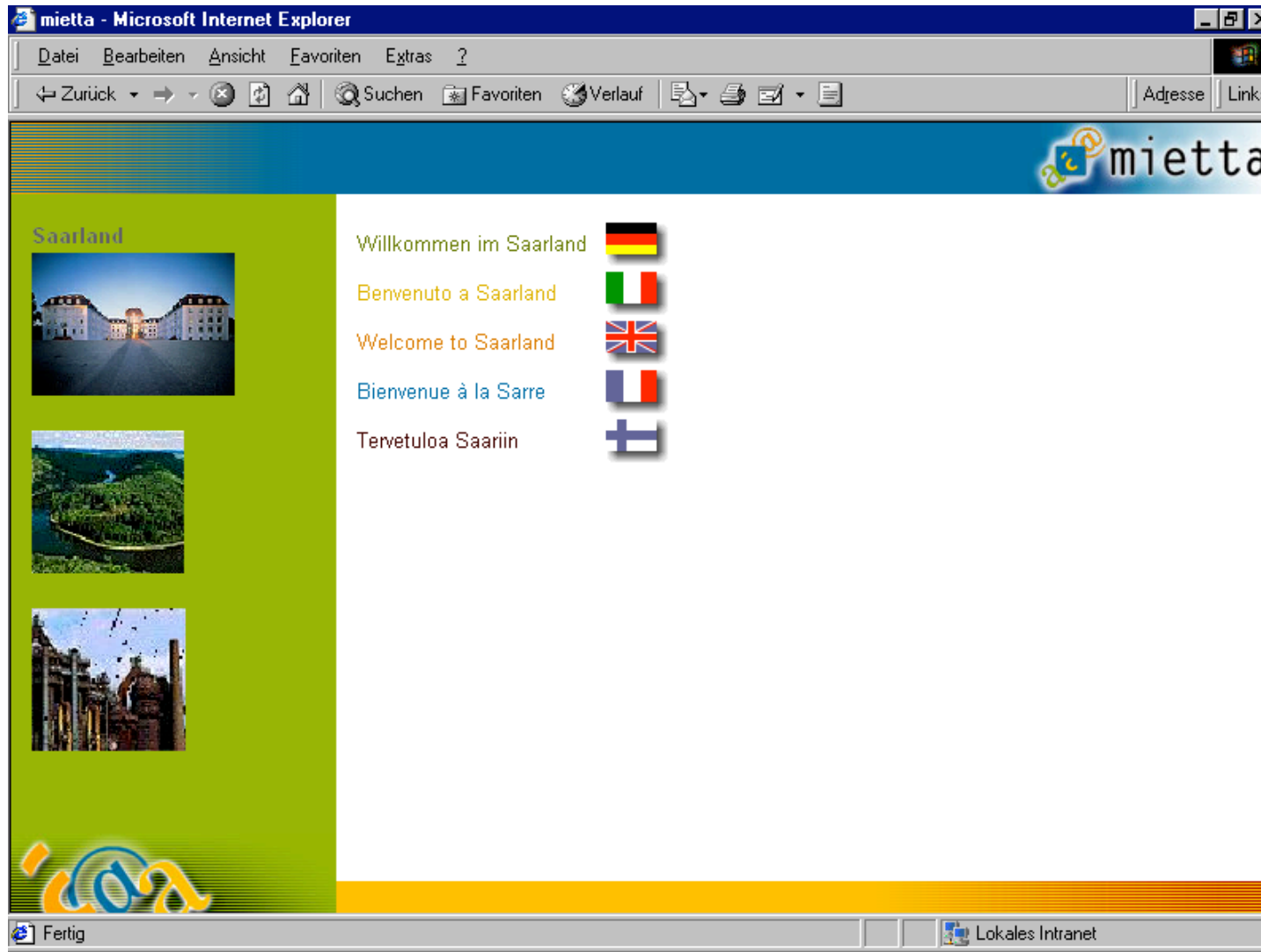
- ❑ Integration of different multilingual and crosslingual search technologies
- ❑ Combination of IE and multilingual generation
- ❑ Integration of DB and text document access
- ❑ Intelligent User Interface
- ❑ XML for advanced information management
- ❑ Localisation technologies for user interface and multilingual generation
- ❑ Highly suitable as a domain-specific information system and internet portal



MIETTA Start Page: Choose Region



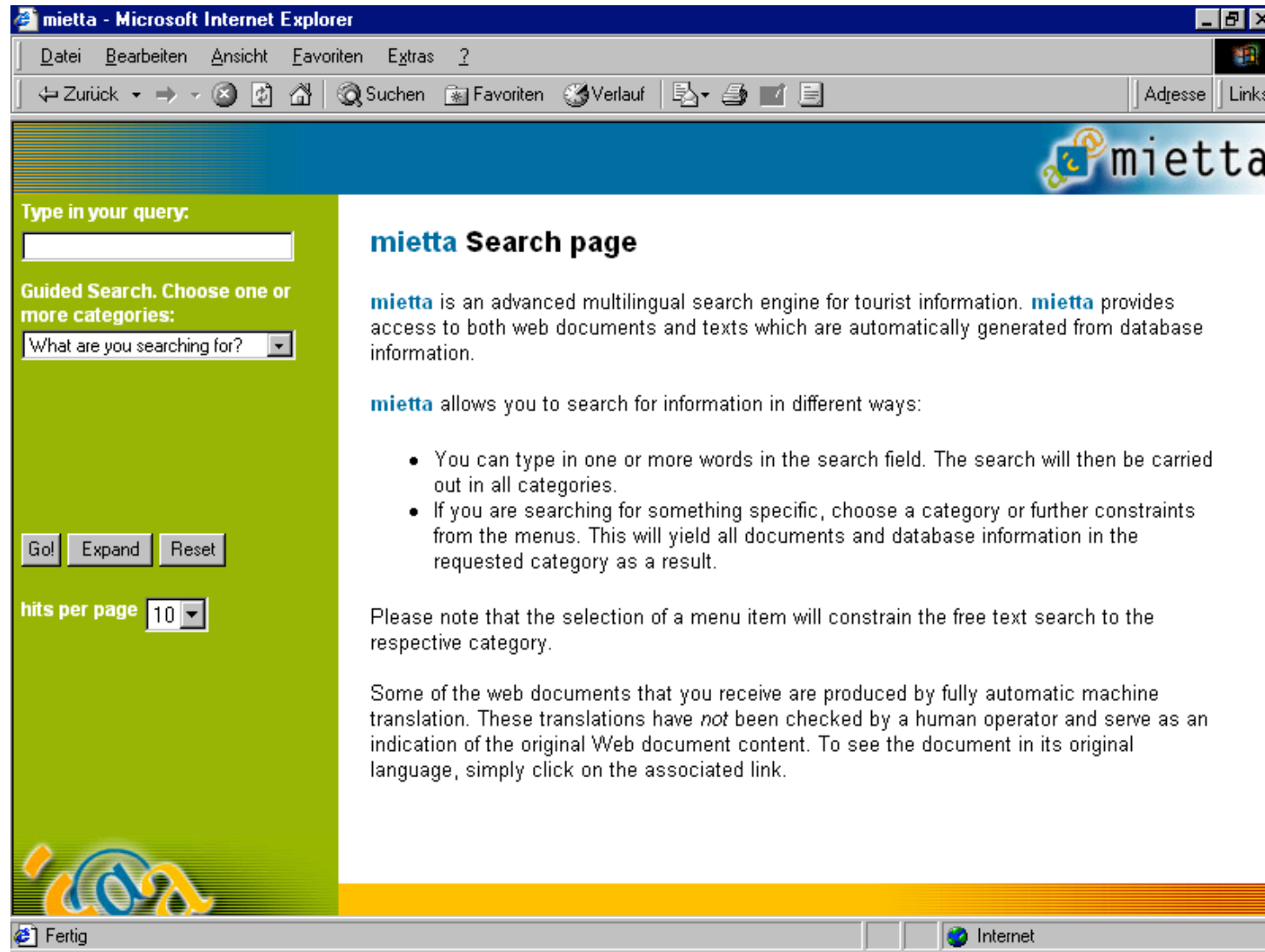
Choose Language



Source: Feiyu Xu, 2002



MIETTA Search Menu



MIETTA Free Text Retrieval

mietta - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Suchen Favoriten Verlauf Adresse Links

mietta

Type in your query:
castles

Guided Search. Choose one or more categories:
What are you searching for?

Go! Expand Reset

hits per page 10

Major Tourist Attractions	6	1
General Information	2	
Nature and Sporting Activities	1	
Events	1	

webdocuments that best match your search

Castle ruin Dagstuhl, Wader, Saarland	Art and Culture	16,84%	original document
Small castles, Hilbringen, state circle Merzig - Wadern	Art and Culture	13,89%	original document
Castle Montclair, Mettlach, state circle Merzig-Wadern, Saarland	Art and Culture	11,95%	original document
Lock mountain in Nennig, state circle Merzig - Wadern, Saarland	Art and Culture	10,66%	original document
The nice castle in Namborn	Art and Culture	10,05%	original document
Mettlach title side	General Information	9,95%	original document
Orscholz state circle Merzig	General Information	7,67%	original document
Circle homeland museum pelt mountain small castles, state circle Merzig - Wadern, Saarland	Art and Culture	7,06%	original document
info	Sports	6,59%	original document

database entries that best match your search

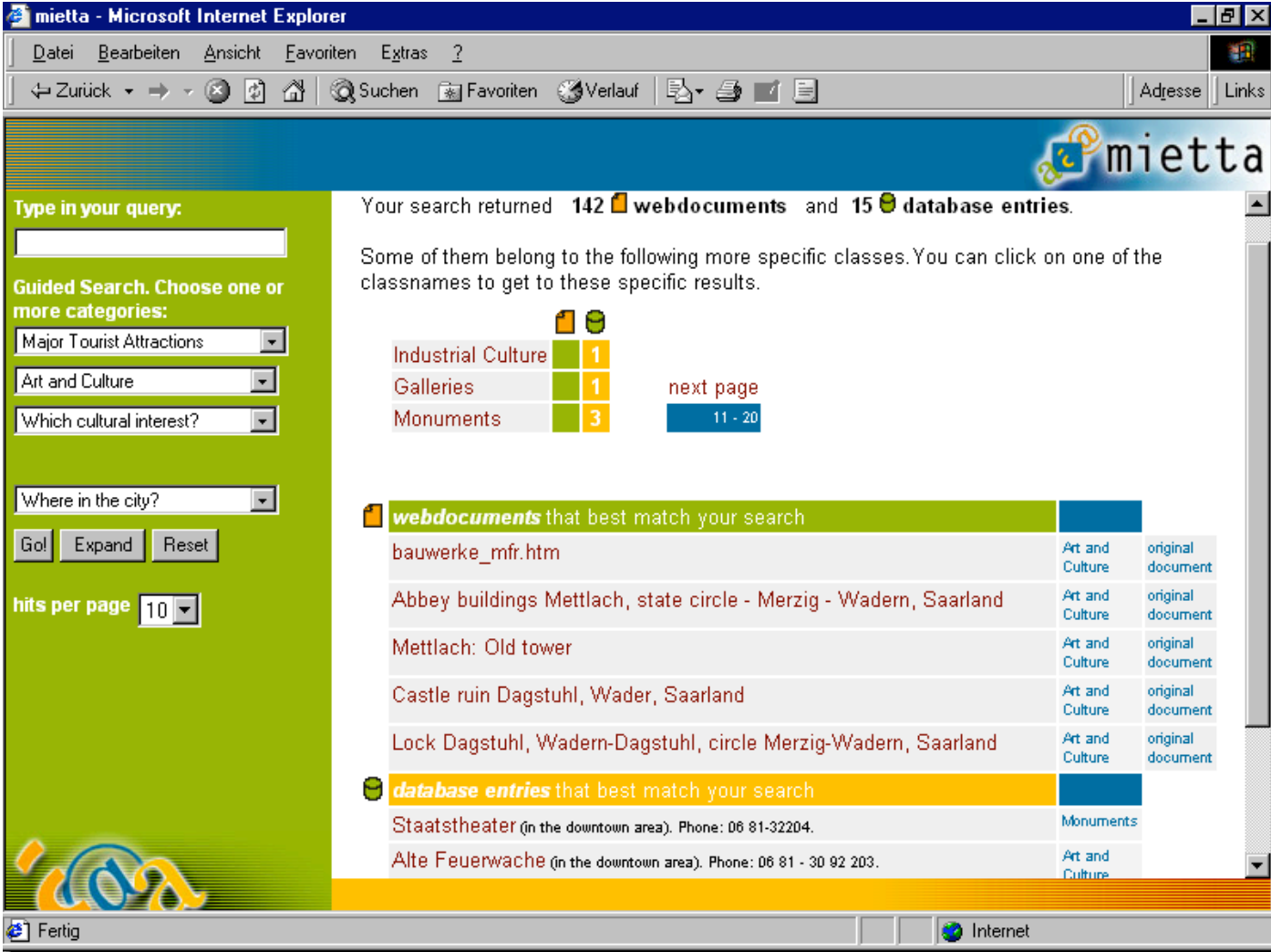
Schloss (in the downtown area). Phone: 06 81 - 50 62 47.	Monuments	100%
--	-----------	------

Internet

Source: Feiyu Xu, 2002



MIETTA Class-based Navigation



Source: Feiyu Xu, 2002



MIETTA Class-based Navigation with Free Text

Type in your query:
church

Guided Search. Choose one or more categories:
Major Tourist Attractions
What type of place?
Where in the city?

Go Expand Reset

hits per page 10

Your search returned **4** web documents

Some of them belong to the following more specific classes. You can click on one of the classnames to get to these specific results.

Art and Culture **4**

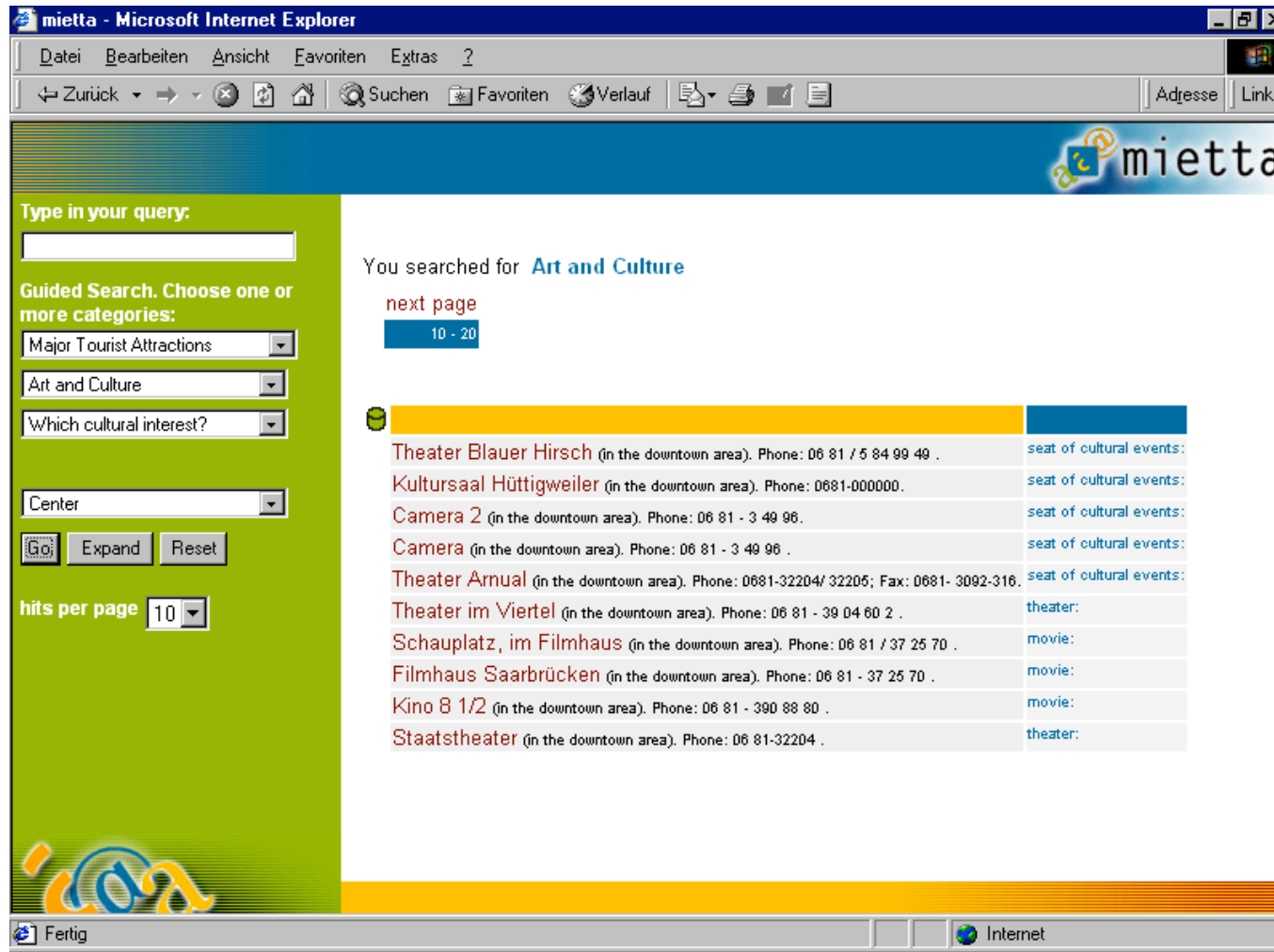
webdocuments that best match your search

Document Title	Class	Percentage	Type
Pfarkirche St.Jacobus, showing churches, circle Merzig - Wadern, Saarland	Art and Culture	30,59%	original document
Curator office: Monument list Saarland, nine churches	Art and Culture	28,72%	original document
Curator office: Monument list Saarland, showing churches	Art and Culture	24,8%	original document
Curator office: Monument list Saarland, Freisen	Art and Culture	15,89%	original document

Source: Feiyu Xu, 2002



MIETTA Form based Query





Online Text Generation

English	The theater Staatstheater is located in Schillerplatz 1, 66111 Saarbrücken (in the downtown area). Phone: 06 81-32204 .
Finnish	Teatteri Staatstheater sijaitsee osoitteessa Schillerplatz 1, 66111 Saarbrücken (keskustan alueella). Puhelin: 06 81-32204.
French	Le théâtre Staatstheater se trouve Schillerplatz 1, 66111 Saarbrücken (dans la zone du centre). Téléphone: 06 81-32204 .
German	Das Theater Staatstheater befindet sich in der Schillerplatz 1, 66111 Saarbrücken (im Stadtzentrum). Phone: 06 81-32204 .
Italian	Il teatro Staatstheater si trova in Schillerplatz 1, 66111 Saarbrücken (nella zona del centro). Telefono: 06 81-32204.











Result Presentation

- Result contains both database entries and documents
- All information is presented in uniform format
 - ★ Classified
 - ★ Ordered according to the relevance

Your search returned 3  **webdocuments** and 1  **database entries**.

Some of them belong to the following more specific classes. You can click on one of the classnames to get to these specific results.

Major Tourist Attractions	 	1
Transport	 	2
Events	 	1

 webdocuments that best match your search		
LINE B	Transport	63,64%
FASCIA BLU (BLUE SECTOR)	Transport	30,71%
EASTER	Public and Town Festivals	5,65%
 database entries that best match your search		
Colosseo (in the downtown area). Opening hours: 0900-1500. Phone: 067004261; Fax: 061234567.	Monuments	100%