

A System for Uniform and Multilingual Access to Structured Database and Web Information in a Tourism Domain

Feiyu Xu, Klaus Netter, Holger Stenzhorn
DFKI Language Technology Lab
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
{feiyu, netter, holger}@dfki.de

Thematic Session: Application systems using language technology components

Keywords: structured database, web documents, crosslingual information retrieval, machine translation, information extraction and multilingual generation

1 Abstract

We present an information system, which was developed within the project MIETTA (Multilingual Information Extraction for Tourism and Travel Assistance), a project in the Language Engineering Sector of the Telematics Application Program of the European Commission. MIETTA facilitates multilingual information access in a number of languages (English, Finnish, French, German, Italian) to the tourist information (web documents and database information) provided by three different geographical regions: the German federal state of Saarland, the Finnish region around Turku and the Italian City of Rome.

The challenge of the approach is to merge the technologies of crosslingual information retrieval (Jamie Carbonell et al, 1997) and natural language processing to achieve the following goals:

- Provide full access to all information independent of the language the information was originally encoded in and independent of the query language;
- Provide transparent natural language access to structured database information;
- Provide hybrid and flexible query options to enable users to obtain maximally precise information.

For the purpose of cross-lingual retrieval, we apply two different methods. We use offline automatic document translation to be able to construct indices from web documents in others than the original document language. This allows the user to access the content of a document without knowledge of the document language and

provides good retrieval performance within our limited domain. At the same time, multilingual access to the database information is supported by the combination of information extraction (Piskorski and Neumann, 2000) and multilingual generation (Busemann and Horacek, 1998). Information extraction extracts domain-relevant templates from database and normalizes them in a language-independent format, while multilingual generation produces natural language descriptions from templates. As a result, the database content becomes multilingually available for the result presentation, and natural language descriptions can be handled in the same way as web documents, namely, we can apply advanced free text retrieval methods to them.

As for query and navigation options, it can be observed that in many applications, structured database information is accessed by means of forms, unstructured information through free text retrieval. In our approach, we attempt to overcome such correlations by making it completely transparent to the user whether they are searching in a database or a document collection, leaving it open to them what kind of query they formulate. Free text queries, form-based queries and their combination can yield documents and structured database information. The user can formulate their query in their own language, while the retrieved results are presented in a uniform textual representation in their query language too.

The hybrid search options provided in MIETTA are:

- *Free text retrieval:* The user can enter several words or phrases to find both web documents and descriptions generated from templates.
- *Concept based navigation:* The user can navigate through web documents and templates according to the MIETTA concept hierarchy.
- *Form-based search:* The user can select fields in a search form to access templates.

MIETTA uses the existing TNO ISM/VSM search engine for free text retrieval (Hiemstra and Kraaij, 1998). The ISM part makes use of a fuzzy matching algorithm based on trigrams. It allows to match index terms with query words or phrases containing spelling errors or morphological variants. For example, the user can enter “*baroque palaces*” and find documents and template descriptions which contain the phrase “*baroque styled palace*”. In addition to the free text retrieval, the user can also navigate through the concept hierarchy to search for information in a certain category. In contrast to many other search engines, the MIETTA user can also combine the free text retrieval with the concept-based navigation by formulating a query with constraints such as “*find all documents containing the word colosseo belonging to the category Art and Culture*”, see the Figure 1.

Figure 1

Figure 2

More restricted and goal-directed is the form-based query, where the user can select fields in a template form. For example, the user can select the “*Time*” and the “*Location*” fields of a “*Concert*” event template by using a query form. In the following example, the user has formulated a query corresponding to the constraint “*give me all information about concerts in the city center today*” (see Figure 2).

All queries are processed by the query processing component and converted either into a standard SQL query or an ISM/VSM query. The result of the retrieval is presented as a uniform list of links to textual descriptions (generated from templates) and web documents. Both types of information are presented, on the one hand in an absolute ranking order, where only the relevance of the document plays a role, and on the other hand sorted according to the different categories. If the user clicks on a link, they receive either a

web document or a generated text from a template.

To summarize, the MIETTA search engine represents a flexible way of combining crosslingual free text retrieval with standard database access. The hybrid query options and their interaction provide the user with a highly versatile range of options to express their different search requirements, which is also reflected in the presentation of the results and the further navigation options.

Acknowledgements

This system is developed within the MIETTA consortium. We are grateful to Paul Buitelaar, Olga Goldmann and to our colleagues at the MIETTA Partner institutions: Luca Dini, Vittorio Di Tomaso and Giampaolo Mazzini in CELI (Centro per l’Elaborazione del Linguaggio e dell’Informazione), Alessandro Giarante, Marcello Vispi in Unidata S.p.A., Kimmo Koskenniemi, Jyrki Niemi in the University of Helsinki, Elena Baralis and Rosa Meo in Politecnico di Torino, our project coordinators Francesco Bellini and Audrey Boss in Comune di Roma and our user partners.

References

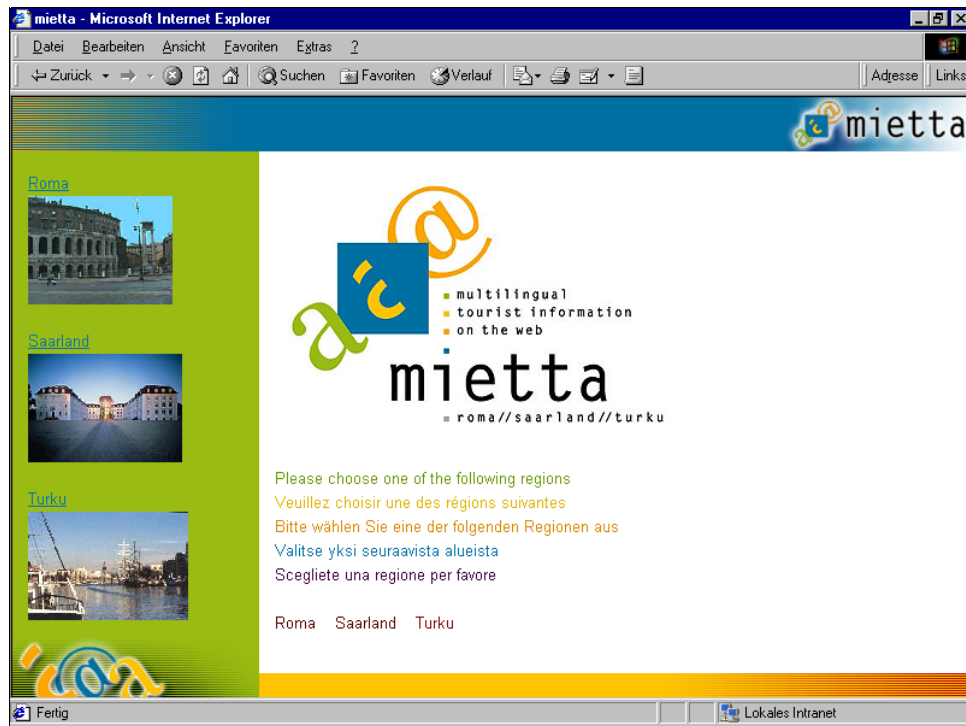
Stephan Busemann and Helmut Horacek (1998). A Flexible Shallow Approach to Text Generation, in: Eduard Hovy (ed.): Proceedings of the Ninth International Natural Language Generation Workshop (INLG 98), Niagara-on-the-Lake, Canada, August 1998, 238-247.

Jaime Carbonell, Yimying Yang, Robert Frederking, Ralf D. Brown, Yibing Geng and Danny Lee (1997). Translingual Information Retrieval: A comparative evaluation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, August 1997.

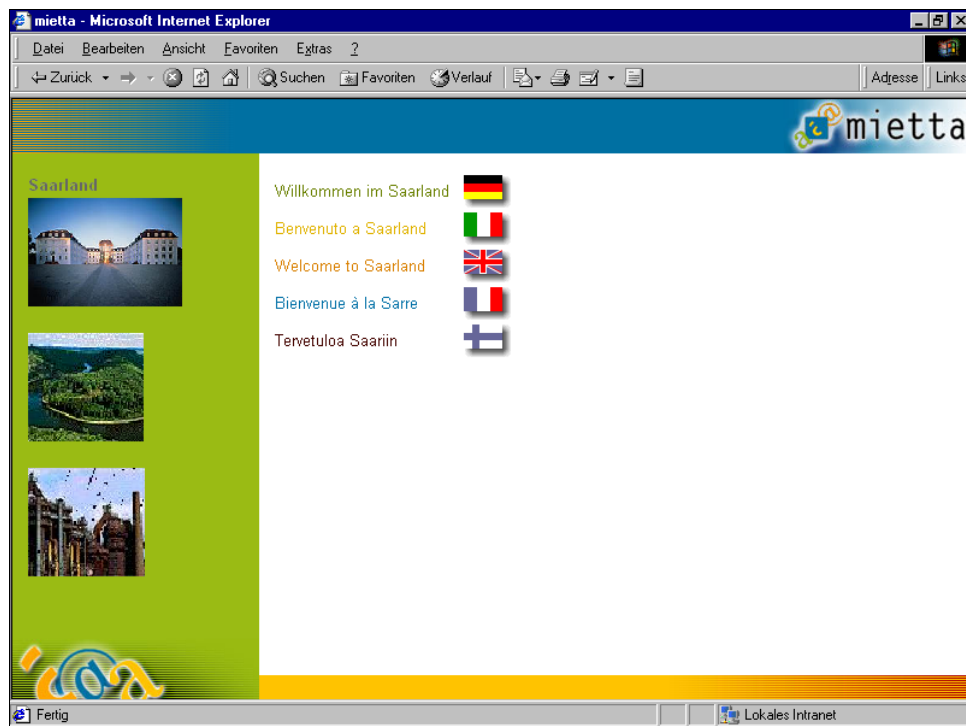
Djoerd Hiemstra and Wessel Kraaij (1998). Twenty-One in ad-hoc and CLIR. In: *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, E.M. Voorhees and D. K. Harman (editors), NIST special publication 500-240.

Jakub Piskorski and Günther Neumann (2000). An Intelligent Text Extraction and Navigation System In proceedings of 6th International Conference on Computer-Assisted Information Retrieval (RIA0-2000), Paris, 2000.

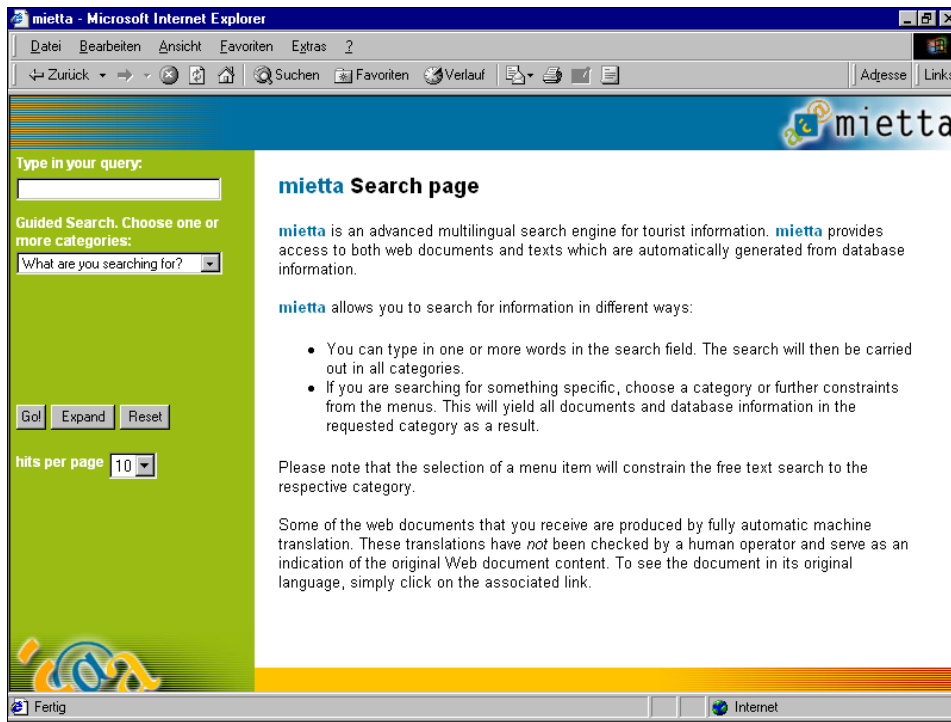
2 Demo Script



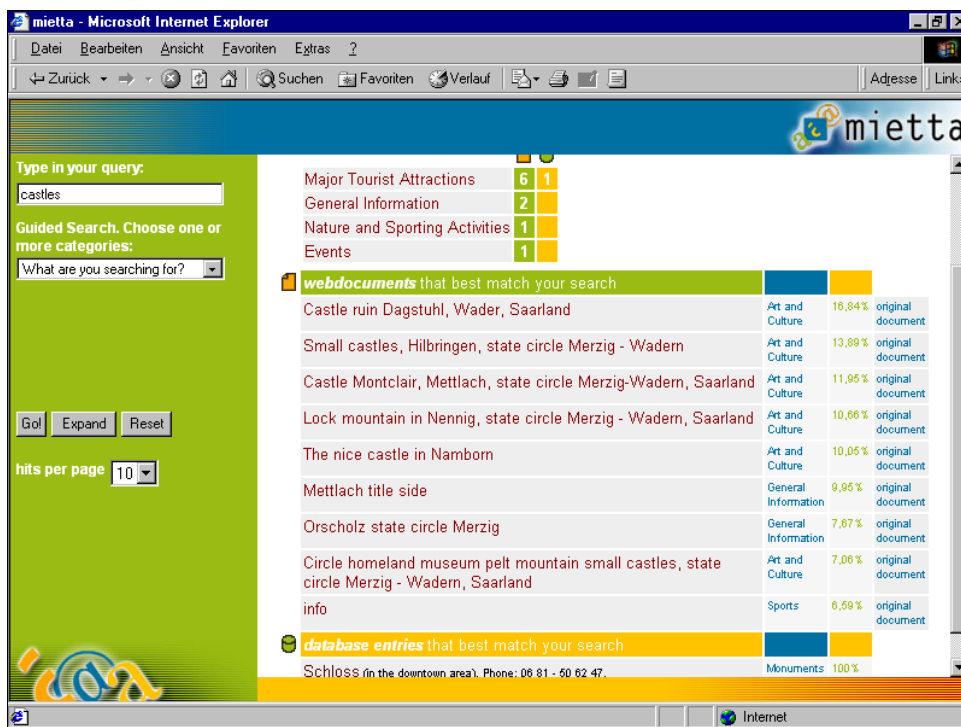
In our demo the user can search information in three regions, the City of Rome in Italy, the German federal state of Saarland and the Finnish City of Turku. The users can choose the region, by clicking on one of the links “Roma”, “Saarland” or “Turku” at the bottom.



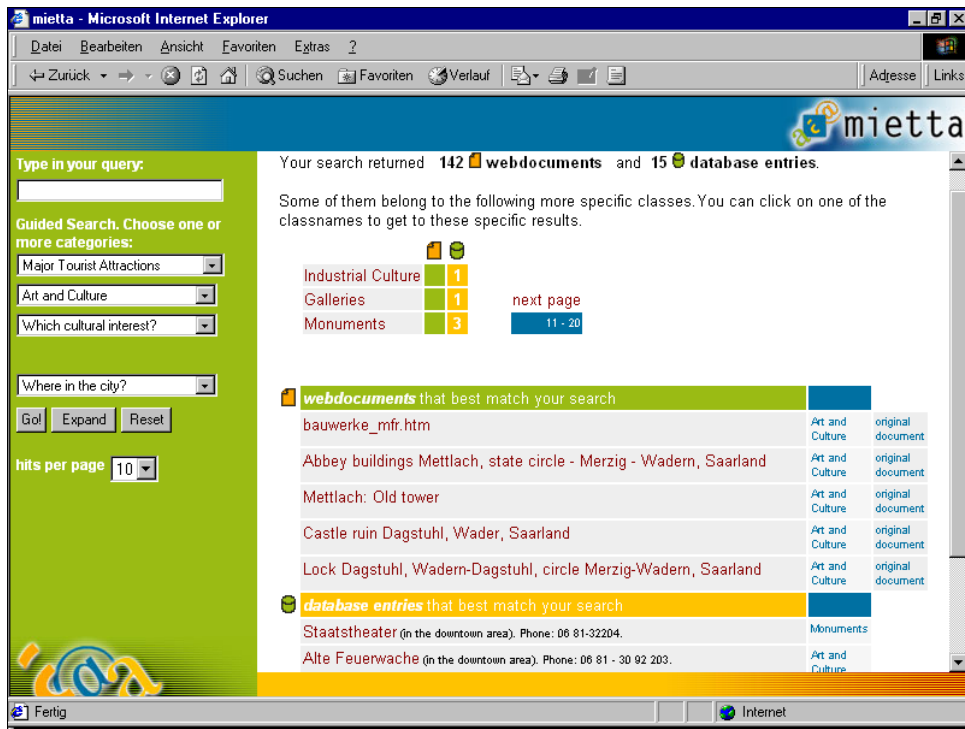
The MIETTA regional server then starts with the language selection. For example, if the user selects the Saarland server, this page will come up for language selection. The interface language, the query language and the language in which the results are (primarily) presented are then set to one of the five languages English, Finnish, French, German and Italian.



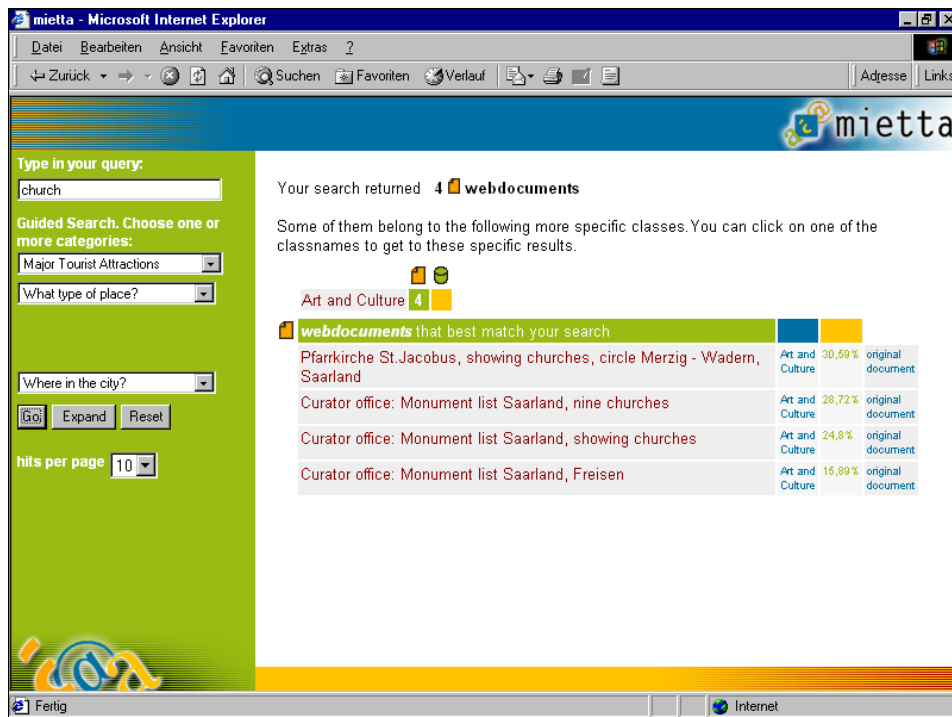
After language selection, the query user interface appears. In the frame on the left hand side, the user can enter some query terms – key words or phrases – and/or select a certain category. The frame on the right hand side describes the functionality of the MIETTA system and its usage.



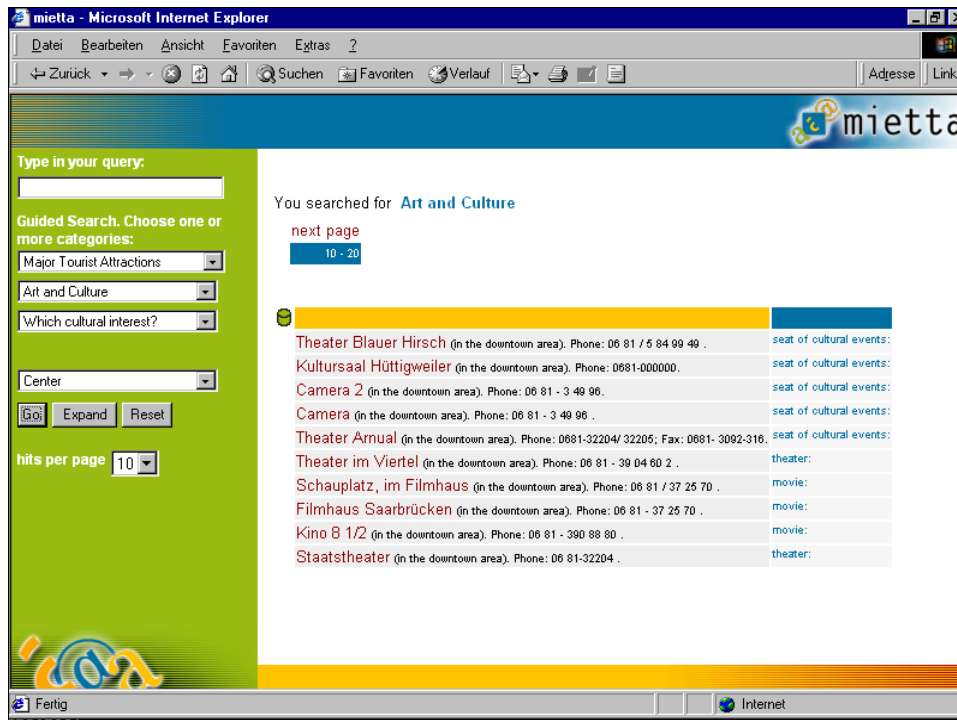
If the user enters some key words, they receive search results containing web documents and texts generated from templates. All the results are ordered according to their relevance and classified into different categories. In the above example, the user query is “castles”. The search engine returns both links to translated web documents (with links to the original documents), as well as database entries matching the query.



The user can also do some guided search by navigating through the category hierarchy. Such a query yields all information within the selected category. For example, the selection of the subcategory “*Art and Culture*” of the main category “*Major Tourist Attractions*” returns all web documents and database entries within this subcategory.



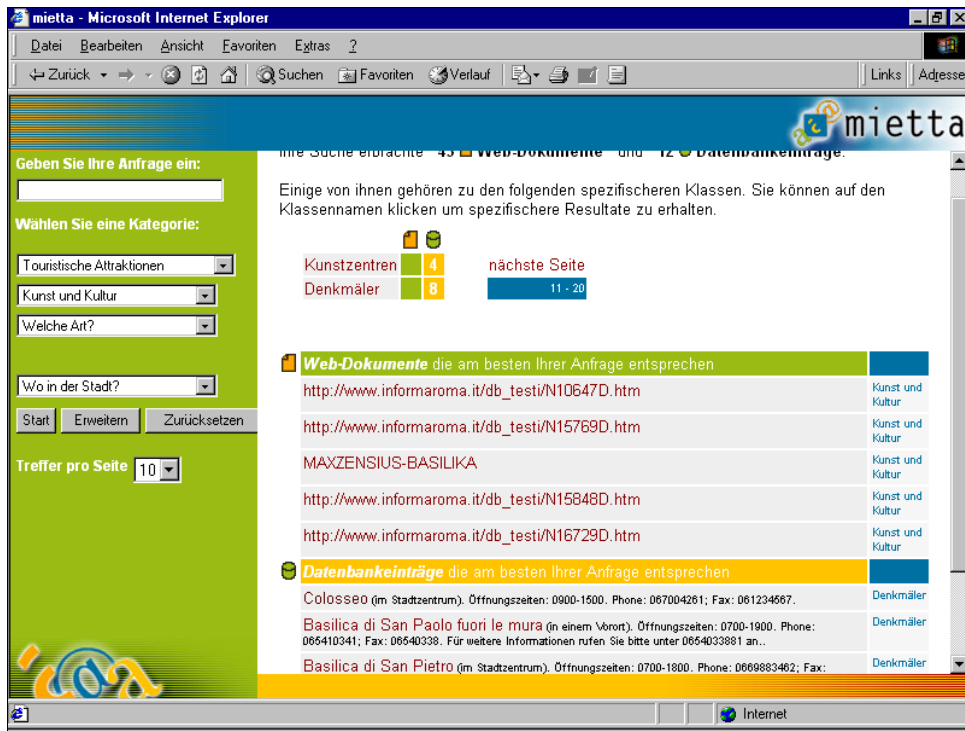
MIETTA also allows a combination of the two query types to restrict each other. The user can enter a free text query and additionally select a certain category. This query combination results in all documents and database entries that match the free text query and also belong to the category.



The user can also enter an even more restricted and goal-directed query – template query – to find more precise facts. Here the user gets all database entries belonging to the subclass “Art and Culture” of “Major Tourist Attractions” in the “Center” of Saarbrücken.

The search results coming from database entries are presented on the basis of multilingual generation. For example, if the users clicks on the link “Staatstheater”, they receive automatically generated, detailed descriptions generated from the corresponding template. The following shows the generation result from one of the above template for the five languages of the MIETTA system.

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.



The above shows the results of a guided search in Rome in German, the table below the generation from the template “Colosseo” in all languages of the system.

English	The Colosseo (roman period) is located in Piazza del Colosseo (in the downtown area). The monument has been built in the first century. The price of the visit is 5 euros. Opening hours: 0900-1500. Phone: 067004261; Fax: 061234567.
Finnish	Colosseo (antiikin Rooman ajalta) sijaitsee osoitteessa Piazza del Colosseo (keskustan alueella). Rakennus on rakennettu viimeisellä vuosisadalla eKr. Hinta käyntiä kohti on 5 euroa. Aukioloajat: 0900-1500. Puhelin: 067004261; faksi: 061234567.
French	Le monument ancienne Colosseo (période romaine) se trouve Piazza del Colosseo (dans la zone du centre). Le monument a été construit au premier siècle. Le prix de la visite est de 5 euros. Horaires d'ouverture: 0900-1500. Téléphone: 067004261; Fax: 061234567.
German	Das Colosseo (römisches Zeitalter) befindet sich in der Piazza del Colosseo (im Stadtzentrum). Das Monument wurde im ersten Jahrhundert gebaut. Der Preis für einen Besuch beträgt 5 Euro. Öffnungszeiten: 0900-1500. Phone: 067004261; Fax: 061234567.
Italian	Il Colosseo (periodo romano) si trova in Piazza del Colosseo (nella zona del centro). Il monumento appartiene al periodo tardo romano (I secolo a.c.) Il costo della visita è di 5 euro. Orario d'apertura: 0900-1500. Telefono: 067004261; Fax: 061234567.

3 Software and Hardware Requirements

Access to power and the Internet is needed.
We intend to bring our own Laptop-PC.

Hardware	PC Pentium II 266 MHz 128 Mbyte memory 300 MByte of free space on hard-disk
Operating System	Microsoft Windows 98/NT/2000
Runtime Environment	Java 2 SDK, Standard Edition, v 1.3 <i>or</i> Java 2 Runtime Environment, Standard Edition, v 1.3
Software	Microsoft Internet Explorer 5 Microsoft Access 97 Microsoft SQL Server 7.0 Apache Jakarta Tomcat 3.1