

Integrating Heterogeneous Tourism Information in TIScover – The MIRO-Web Approach

M. Haller†, B. Pröll†, W. Retschitzegger‡, A. M. Tjoa†, R.R. Wagner†

† Institute for Applied Knowledge Processing (FAW)
University of Linz, AUSTRIA
email: {mhaller, bproell, amt, rwagner}@faw.uni-linz.ac.at

‡ Institute of Applied Computer Science, Department of Information Systems (IFS)
University of Linz, AUSTRIA
email: werner@ifs.uni-linz.ac.at

Abstract. A broad spectrum of tourism information is already distributed over various web sites. However, a major problem for the tourist is to find these web sites and to deal with the differences concerning information presentation and information access. At the same time, it is not feasible to store every kind of information a tourist might be interested in at one web site neither in terms of storage costs nor and even more important in terms of maintenance overhead. Therefore, in the course of the *ESPRIT project MIRO-Web* the official Austrian tourism information and booking system TIScover is extended in order to federate multiple structured and semi-structured tourism information sources on the web. In particular, MIRO-Web supports a homogeneous view on these heterogeneous sources which can be either materialized or defined as virtual. On the basis of this view, appropriate query mechanisms as well as a web-based interface provide the user with a single point of access.

1 Introduction

With the tremendous growth of the web, accessing information on the Internet has become less a question of determining whether the information is out there, but rather, in what form, and how to find it [5]. This situation is especially true for the tourism industry where a broad spectrum of tourism information is already distributed over various web sites and stored using heterogeneous formats. It is obvious that this situation is very undesirable since the tourist is burdened with finding and visiting various web sites in order to gather all the desired tourism information and products. The situation gets even worse since web sites usually differ very much especially concerning information presentation and information access. What is therefore required is that the tourist is enabled to collect all desired tourism information and tourism products at one place, no matter whether, e.g., information about weather and traffic conditions, schedules of trains, planes and buses, or information about actual events like movies at the local cinema including movie reviews as published in different online newspapers are required.

To fulfill this requirement it is of course not feasible to store every kind of information a tourist might be interested in at one web site neither in terms of storage costs nor and even more important in terms of maintenance overhead. A more promising idea would be to take advantage of the huge amount of other relevant tourism information that is already distributed all over the web. In the course of the *ESPRIT project MIRO-Web* [4] which was started at the end of 1997, the official Austrian tourism information and booking system TIScover is extended in order to federate multiple structured and semi-structured tourism information sources on the web. These sources comprise different in-house databases employed in the various tourism offices, Excel files used for managerial purposes, and finally and most important HTML-pages of other web sites [3].

The rest of the paper is organized as follows: After a short overview about TIScover in Section 2, the basic concepts and technologies of MIRO-Web used in order to integrate heterogeneous information sources are presented in Section 3. Section 4 demonstrates the applicability of MIRO-Web to TIScover by presenting two different application scenarios comprising an *Event Agent* and a *Golf Agent*. Section 5 concludes the paper by discussing lessons learned and pointing to future work.

2 An Overview of TIScover

The development of TIScover has been started in 1996 based on the experiences made with the pioneering system TIS@WEB [1]. The aim of TIScover is twofold [11, 12, 13]: first, tourists should be *supplied with comprehensive, accurate and up-to-date tourism information* on countries, regions, villages and all destination facilities they offer like hotels, museums or other places worth seeing. Second, it aims to *attract the tourist to buy certain tourism products* either offline or even more important to allow the tourist to buy them *online*. Originally, TIScover was realized to market the facilities of a certain region of Austria, namely Tyrol, only. Meanwhile, four other Austrian regions have joined TIScover [15]. Besides that, TIScover has been employed in Asia, presenting tourism information about Thailand [16]¹, it is used by the German company START Media Plus, a major player in the area of online reservation systems, to present tourism information about Germany [17], and it is online in Switzerland as KISSwiss, employed by the companies Kümmerly+Frey and Basler Versicherungen [18].

The functionality provided by TIScover can roughly be categorized into three different components, the *public Internet* component, the *Extranet* and the *Intranet* [12]. The *public Internet* component comprises that functionality of the system that is accessible to the public, whereby the most important modules are *Atlas* and *Booking*. The module Atlas allows the customer to browse through all kinds of tourism information by navigating through a geographical hierarchy and to use a *full text search*. The module Booking allows for a *precise structured search* based on a subset of the tourism information presented by Atlas, like villages, hotels, available rooms, events and camping sites along the geographical hierarchy as well as *online booking*

¹ Note, that due to the economical crisis in asia, TIScoverasia is currently not operating.

of these tourism products. Furthermore, TIScover provides an *Extranet* allowing authorized tourism information providers, no matter being a small guesthouse or a large local tourist office to update and extend their tourism information and products directly. Finally, the *Intranet* component of TIScover which is accessible at the system provider's side only allows to configure the whole system in various ways. It is, for example, possible to extend the geographical hierarchy, to specify expiration dates for reports and to define the default language for all system components.

Currently, TIScover stores all tourism information and tourism products within a central relational database. The common database schema of TIScover Austria consists of about 300 database tables and has been constructed on the basis of a domain data model which incorporates all conceptual entities gathered during the process of requirements definition with numerous tourism information providers and from the experiences with the predecessor system TIS@WEB [1]. The database of TIScover Austria comprises about two gigabyte of data. To facilitate performant access, web pages are automatically generated out of the database every time one of the 7.000 tourism information providers ranging from hotels to local tourism offices maintains the content covering among others around 2.000 towns and villages and nearly 40.000 accommodations [14]. As a result, there exist more than 400.000 web pages stored in some million files. Per month, the system has to handle up to 8,6 million pageviews, 2 million visits as well as up to 40.000 requests for information on booking and online bookings.

Although, these figures illustrate that TIScover manages a fairly huge amount of tourism information, it is of course far from being complete. To be able to satisfy also requests for certain information which is not part of the TIScover database, but already available at other web sites, the MIRO-Web project was intended to provide a proper technical basis.

3 Basic Concepts and Technologies of MIRO-Web

The integration of heterogeneous data is a challenging problem when trying to utilize existing information on the web. Many of the problems encountered in building such systems are similar to those addressed in building heterogeneous database systems [5, 8]. However, the integration of information sources on the web poses some fundamentally new challenges:

Heterogeneity of Sources. Web information is stored within distinct heterogeneous sources, whereas the important kind of heterogeneity in our context is *structural heterogeneity* [2, 5]. As an extreme we find *fully structured data* coming, e.g., from databases having a rigid and explicit schema. At another extreme, there is data which is *fully unstructured* having no schema, such as images, sounds, and raw text. But most of the data falls somewhere in between these two extremes, called *semi-structured data* (e.g., HTML-files), where the schema can be implicit and does not have to be rigid.

Evolution of Sources. Information sources at the Internet *evolve at a much higher pace than databases* in a controlled business setting and therefore increase the maintenance overhead of the integration schema.

Autonomy of Sources. Many sources are characterized by a high degree of autonomy, allowing a partial integration only.

Lack of Source Meta Data. Commonly, there is little meta data available about the characteristics of the sources which further complicates their integration.

Most of current web sites do not fully cope with these challenges and do not allow for the integration of multiple data sources beyond simple links between them. The MIRO-Web project [4] focuses on the development of a set of middleware components providing integrated and transparent access from standard web browsers to multiple data sources, ranging from databases to more or less structured files, located on different web sites.

3.1 The Architecture of MIRO-Web

MIRO-Web builds on the technology developed and the knowledge acquired in the course of the ESPRIT project IRO-DB, aiming at the integration of heterogeneous databases [8]. In the spirit of IRO-DB, MIRO-Web is based on a *three-tier architecture*, consisting of a *Data Source Adapter Layer*, a *Mediation Layer* and a *Client Layer* (cf. Fig. 1).

Data Source Adapter Layer. The Data Source Adapter Layer consists of a number of adapters, also known as wrappers, which are needed to mask the heterogeneity of data sources and to transform source data into a structured format [7]. Data source adapters have two main functions: they first translate a query to the underlying query system used by the source and then they translate the results sent by the source after evaluation of the translated query. MIRO-Web adapters can be used in conjunction with a mediator (cf. below) or independently by an application. In the first case they provide an API well suited for the interaction between the mediator and the adapter. In the latter case they provide a JDBC [9] interface through which they support a subset of SQL depending on the query management capabilities of the data source and on those implemented in the adapter itself. MIRO-Web supports several kinds of adapters depending on the nature of the data source, such as relational adapters as well as structured and semi-structured file adapters. To facilitate both, building and maintenance of data source adapters a Java toolkit called *Adapter Development Kit* has been developed.

Mediation Layer. According to [20], mediators “simplify, abstract, reduce, merge and explain data”. The Mediation Layer of MIRO-Web provides the means to combine and integrate these heterogeneous sources into a homogeneous view and supports query possibilities on this view in terms of a single point of access. For a more detailed discussion it is referred to Section 3.2.

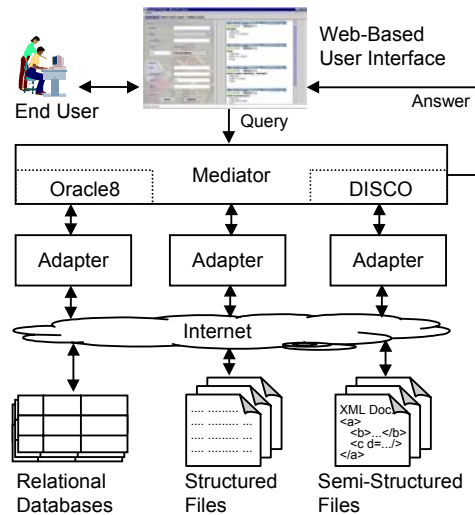


Fig. 1. Overall Architecture of MIRO-Web

Client Layer. The Client Layer provides web-based user interfaces, allowing both, to express queries to the underlying heterogeneous sources and to present the results of these queries in a uniform way. Section 4 presents the user interfaces, which were realized for the application of MIRO-Web to TIScover.

3.2 Materialized Mediation versus Virtual Mediation

An important distinction in building web data integration systems is whether to take a *materialized approach* or a *virtual approach* for mediation. MIRO-Web supports both approaches in order to combine the advantages of each of them. It is the responsibility of the database administrator to decide which approach is suitable for a certain data source.

Materialized Mediation. In the materialized approach, data from multiple web sources is loaded into a warehouse, and all queries are applied to the warehoused data. In particular, a materialized mediator stores data received from the adapters, and provides for integrated views on these imported data. The materialized mediator of MIRO-Web is built on the basis of the object-relational data model of Oracle8 [10], which can represent multi-valued attributes by means of nested tables as well as object references. Appropriate tools are provided for the database administrator to integrate necessary data and to query them through Oracle views.

Where data volume, timeliness, security considerations and frequent changes of the remote data source do not prevent materialization, this approach has the following advantages. First, autonomous sources need not be available all the time and adequate performance can be guaranteed at query time. Second, adapters can concentrate on data mapping and enrichment, and can be limited to rather restricted query capabilities. Finally, queries combining information from multiple sources can be

processed locally, and dedicated indices and representation schemes can be used for improving efficiency.

Virtual Mediation. In the virtual approach, the data remains in the web sources and is not replicated at the mediation layer. To manage query processing in this scenario, the mediator proceeds in several steps at runtime. First the query is decomposed and optimized into sub-queries and a composition query. Each sub-query extracts the necessary information from an adapter. The composition query combines the sub-answers from the sub-queries into the final answer, which is returned to the user through the mediator. Virtual mediation is performed in MIRO-Web by DISCO (Distributed Information Search COmponents) a distributed extensible query engine which provides a uniform query language and data model for declarative access to a heterogeneous collection of data sources [19].

Virtual mediation provides several advantages. First, it is not limited by the storage capacity of the materialized mediator. Thus, arbitrarily large collections of data can be combined. Second, no *a priori* decision must be made by the database administrator *which* data must reside in the materialized mediator. All data is equally available to a mediator using virtual views. Third, data returned through virtual views is topical since data sources are accessed at query time. A materialized mediator must consider the issue of the materialized data becoming out of date with respect to the data sources. Finally, an additional advantage involves security, i.e., each query can be considered independently from a security point of view.

4 Applying MIRO-Web to TIScover

To demonstrate the applicability of MIRO-Web to TIScover in the following two different scenarios are presented, where some kind of *agent* assists the tourist in querying heterogeneous tourism information sources in a transparent way. These agents comprise an Event Agent and a Golf Agent and correspond to the Client Layer of MIRO-Web as described in Section 3.1.

4.1 The Event Agent

A common case is that a tourist wants to find a hotel, which should be near some event locations, such as an exhibition, a sporting event or a cultural event, e.g., a musical or a movie. Furthermore, the weather forecast should be fine and the hotel should cost less than a certain amount of money, issued in an arbitrary currency. To specify such a request, the Event Agent provides a uniform graphical interface in terms of a Java applet (cf. Fig. 2).

The Adapter Layer of MIRO-Web extracts the answer for this request out of different heterogeneous data sources. First, information about hotels is extracted directly out of the TIScover database. Second, information about actual events can be found on different already existing web sites (cf., e.g., www.austria-tourism.at or www.film.at or www.events.at). Third, the weather information is gathered via file transfer protocol in form of a structured file and finally the exchange rates for calculating the requested currency is available in a semi-structured format, namely as an email.

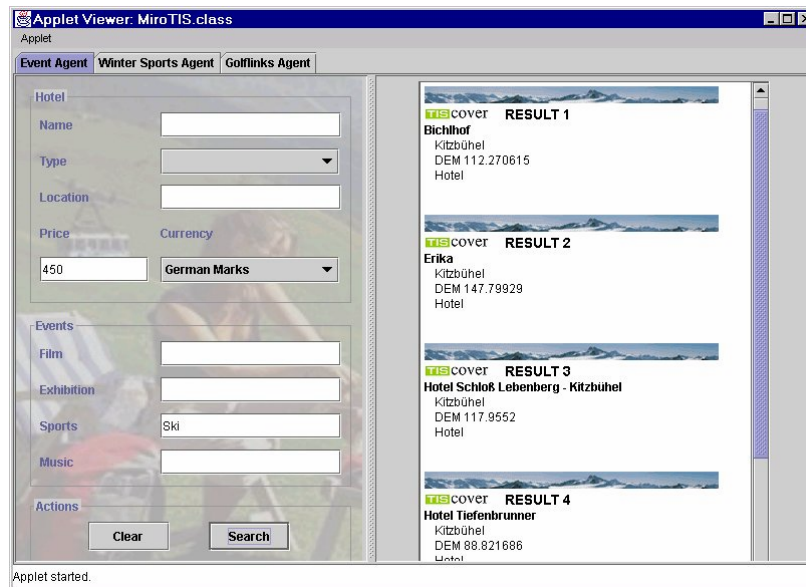


Fig. 2. Event Agent User Interface

The basic join criteria for the query which is used by the Mediation Layer to satisfy an information request is the name of the location. On the basis of this join attribute information about the events can be joined with weather information and hotel information. Finally the price can be calculated and converted by means of the exchange rates and joined with the hotel information. Finally, the Client Layer, i.e., the Java applet presents the result of the information request in a uniform way at the right side of the panel (cf. Fig. 2).

4.1 The Golf Agent

The Golf Agent represents a scenario, where among others, the services of a geographical information system in terms of a route planner are integrated. In particular, the Golf Agent allows a tourist to specify the kind of desired golf courses in a detailed way as well as to define the maximal driving distance to the course (cf. Fig. 3).

As result, the user gets a list of all golf courses taken from a web site providing all golf courses in upper Austria (www.golfweb.at), as well as map, containing the driving directions to each of the proposed golf courses. The latter information is extracted from a German web site, providing online route plans.

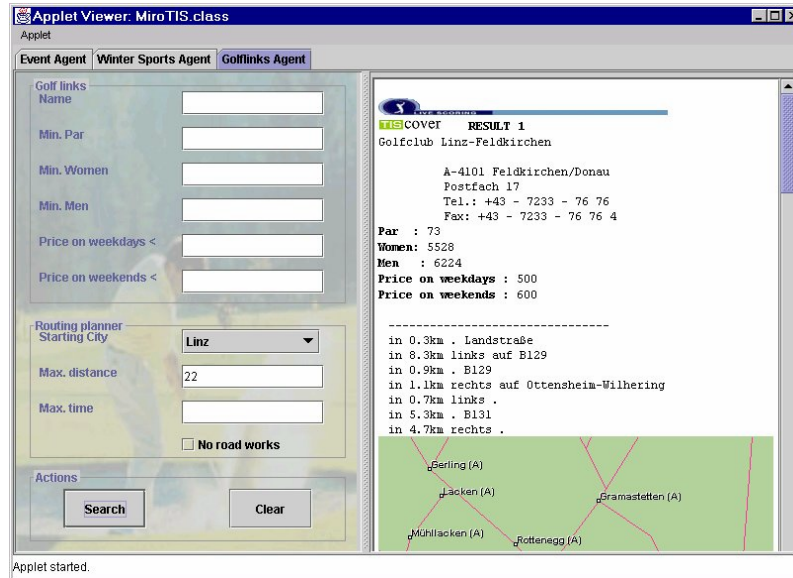


Fig. 3. Golf Agent User Interface

As in the case of the Event Agent, the join criteria for golf courses a driving directions is again the location.

5 Lessons Learned and Future Work

This paper has illustrated the MIRO-Web approach and its application to TIScover. In the course of applying MIRO-Web to TIScover, several interesting issues have been encountered. It has been shown for example, that a lot of web server providers are interested that their data is published not only on their own web sites, but also on other web sites, thus facilitating the federation of web sources. However, a major problem concerning the integration of the various sources is to find a common join criteria for all of the sources. In most of our prototypical realizations, the location or the zip code of a location has been used. One of the most severe problems is the dependency on the web sources to be integrated. If the network is down or slow or if the servers of the provider are even shut down the integration of the data sources is in case of virtual mediation not possible. Furthermore, every time the structure of a data source changes, the corresponding adapter has to be modified too, constituting a maintenance effort which should not be underestimated.

Future work will be done primarily in the course of the follow up ESPRIT project of MIRO-Web called XML-KM (XML-based Mediator for Knowledge Extraction and Brokering). The goal of XML-KM which is strongly based on XML [21] is to enhance the components and tools developed in the course of MIRO-Web in order to be able to collect and disseminate *knowledge* instead of just data. Through a rule-based XML-wrapper, information from corporate databases, HTML files and office applications will be collected in data warehouses [6]. Using classical data mining

tools, knowledge will be extracted out of the warehouses in the form of derived information and rules. Through XML-based query tools, users will be able to subscribe and receive personalized information in an appropriate format on various devices including computers, mobile phones and faxes.

References

1. F. Burger, P. Kroiss, B. Proell, R. Richtsfeld, H. Sighart, H. Starck, *TIS@WEB - Database Supported Tourist Information on the Web*, Proc. of the Int. Conf. on Information and Communication Technologies in Tourism (ENTER'97), A Min Tjoa (ed.), Springer, Edinburgh, 1997.
2. R. Domenig, K.R. Dittrich, *An Overview and Classification of Mediated Query Systems*, SIGMOD Record, Vol. 28, No. 3, Sept. 1999.
3. A. Ebner, M. Haller, K. Plankensteiner, P. Starzacher, E. Stauder, and A M. Tjoa, *Tourism Application Requirements Definition (D7B-1/1)*, MIROWEB (EP 25208), April 1998.
4. P. Fankhauser, G. Gardarin, M. Lopez, J. Munoz, A. Tomasic, *Experiences in Federating Databases: From IRO-DB to MIRO-Web*, Proc. of the 24th Int. Conference on Very Large Data Bases (VLDB'98), A. Gupta et al. (eds.), New York City, USA, Morgan Kaufmann, Aug. 1998.
5. D. Florescu, A. Levy, A. Mendelzon, *Database Techniques for the World Wide Web: A Survey*, ACM SIGMOD Record, Vol. 27, No. 3, Sept. 1998.
6. G. Gardarin, F. Sha, T.D. Ngoc, *XML-based Components for Federating Multiple Heterogeneous Data Sources*, Proc. of the 18th Int. Conf. On Conceptual Modeling, LNCS 1728, Paris, France, Nov. 1999.
7. G. Huck, P. Fankhauser, K. Aberer, E. Neuhold, *Jedi: Extracting and Synthesizing Information from the Web*, Proc. of CoopIS'98, IEEE Computer Society Press, 1998.
8. E. Kapsammer, W. Retschitzegger, R. R. Wagner, *Meta Data-Based Middleware for Integrating Information Systems: A Case Study*, Proc. of the 4th Int. Conf. on Information Systems Analysis and Synthesis (ISAS'98), Orlando, July 12-16, 1998.
9. The JDBC Database Access API, <http://java.sun.com/products/jdbc>, 1999.
10. Oracle Corporation, <http://www.oracle.com/>, 1999.
11. B. Proell, W. Retschitzegger, R.R. Wagner, A. Ebner, *Beyond Traditional Tourism Information Systems - TIScover*, Journal of Information Technology in Tourism (ITT), Vol.1, Inaugural Volume, Cognizant Corp., USA, 1998.
12. B. Proell, W. Retschitzegger, R.R. Wagner, *TIScover - A Tourism Information System Based on Extranet and Intranet Technology*, Proc. of the 4th Americas Conf. on Information Systems (AIS'98), Baltimore, Maryland, 1998.
13. B. Proell, W. Retschitzegger, R.R. Wagner, *Holiday Packages on the Web*, Proc. of the Int. Conf. on Information and Communication Technologies in Tourism (ENTER'99), D. Buhalis et al. (eds.), Springer, Innsbruck, 1999.
14. B. Proell, W. Retschitzegger, H. Sighart, H. Starck, *Ready for Prime Time - Pre-Generation of Web Pages in TIScover*, Proc. of the 8th Int. ACM Conference on Information and Knowledge Management (CIKM), Kansas City, Missouri, Nov. 2-6, 1999.
15. Homepage of TIScover, <http://www.tiscover.com>, TIS Innsbruck, FAW Hagenberg, 1999.
16. Homepage of TIScover ASIA, <http://www.tiscoverasia.com>, GoThailand, 1998.

17. Homepage of TIScover Germany, <http://www.deutschlandreise.de>, START Media Plus, 1999.
18. Homepage of TIScover Switzerland, <http://www.kissswiss.com>, Kümmerly+Frey, Basler Versicherungen, 1999.
19. A. Tomasic, L. Raschid, P. Valduriez, *Scaling heterogeneous databases and the design of DISCO*, Proc. of the 16th Int. Conf. On Distributed Computing Systems, Hong Kong, May 1996.
20. G. Wiederhold, *Mediators in the architecture of future information systems*, Computer, Vol. 25, No. 3, March 1992.
21. The World Wide Web Consortium (W3C), <http://www.w3.org/XML/>, 1999.