HARMONISE: a Solution for Data Interoperability

Mirella Dell'Erba[a], Oliver Fodor[a], Francesco Ricci[a], Hannes Werthner[a b]

[a] *eCommerce and Tourism Research Laboratory, ITC-Irst, Italy*
[b] *University of Trento, Italy*
{dellerba, fodor, ricci, werthner}@itc.it

Abstract: Travel and tourism represents the leading application in b2c e-commerce, due to several reasons: tourism is an information business, it has a rather a long value chain, and since the 60s it represents a classical field of IT and network applications (i.e., computerized reservation systems / global distribution systems – CRS/GDS). Thus, this sector depends heavily on advanced Information Technologies solutions, and the characteristics of this marketplace lead to a crucial challenge: the strong need for interoperability between different information systems, allowing the flow of information in this complex value process. Data heterogeneity is a well known problem, and the existing practical approach to solve this problem is to write ad-hoc data interface programs for each pair of communicating systems. Experience shows that development and maintenance of these programs are expensive in terms of both time and money. This paper presents an alternative solution to the interoperability problem in the travel / tourism marketplace: the creation of a "harmonisation space" in which market players are enabled to exchange data without changing their own systems' schemata.

Key words: Ontology based mediation, interoperability, tourism domain

1. INTRODUCTION

Travel and tourism represents the leading application in b2c e-commerce (see figure 1). There are several reasons: first, tourism is an information business based on contractual agreements about services, the consumption of which necessarily will take place in the future; whereas in advance only information about the product can be gathered (i.e., in economic terms, it is a "confidence good"); second, it has a rather a long value chain; and third, since the 60s it represents a classical field of IT and network applications

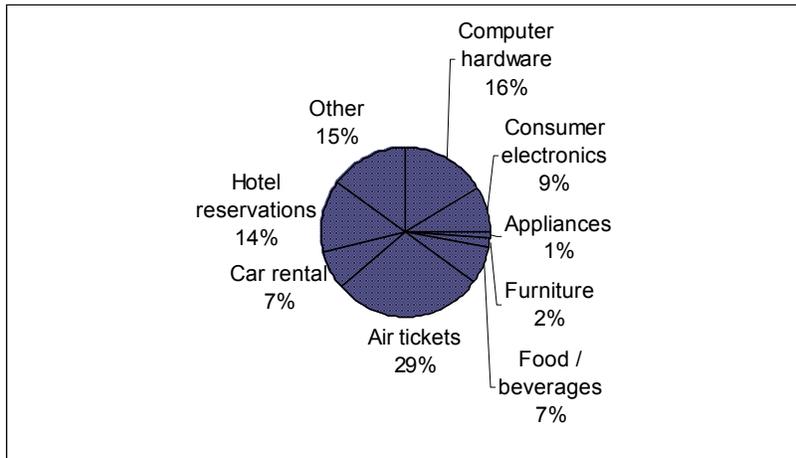(e.g., computerized reservation systems / global distribution systems – CRS/GDS) (Werthner and Klein, 1999).

Figure 1. b2c e-commerce shares (source: Forrester 2001).

In consequence, from an IT point of view, this domain is characterized by a huge set of different Information Systems, with different scopes, basic technologies and architectures, and information schemata. This leads to the interoperability problem, which represents a major barrier for the emerging e-market place, with specific importance for smaller players. In addition, the growing heterogeneity of the Web implies that no single player has enough power to impose one single and accepted standard.

This paper proposes a solution to this interoperability problem, in particular within the travel and tourism domain, by creating a so called "harmonisation space". We foresee that each actor willing to participate can download tools provided by this Harmonisation space, which enable a nearly seamless exchange of information without changing own proprietary data formats.

Since individual organizations are typically reluctant to abandon their own data schemata in favor of a standard schema supplied by someone else, the Harmonise solution allows each partner to keep as much as possible his own "view of the world", that is to see the entire marketplace as it would act in accordance with their individual organization.

In this approach, named "Local As View", (Florescu et al., 1999), the interoperability problem among partners is solved by providing semantic reconciliation between the different tourism information systems with respect to a shared, conceptual reference schema, the domain ontology. This

solution allows market players to view the other participants' information as an extension of its own information system, without concerning for the differences in names and representations of data.

The Harmonise project is an initiative financed by the European Commission, under the 5th Framework RTD Programme. This started in July 2001 and it will end in June 2003.

This paper is organized as follows: the next Section describes the present market scenario and requirements; Section 3 describes the Harmonise solution from the two crucial perspectives: the organizational and the technical one. And we conclude by providing an outlook to future work.

2. PRESENT MARKET SITUATION

As already outlined, the travel/tourism industry is (strongly) dependent on the Information Technology. This dependency is increasing due to the emerging of on-line marketing and information services on Internet. This involves developing and validating new types of work-flow and information exchange between suppliers, intermediaries and customers in the 'tourism value chain'. Whereas in other industries there is a stronger hold on to the traditional way things have been done, in the travel/tourism industry we are witnessing total acceptance to the extent that the structure of the industry and the way business is conducted, is changing. The net is used not only for information gathering; there is an obvious acceptance of ordering services over the Internet. And it's not the case of only trying one or two services; it is all travel and leisure services. There is a new type of user emerging. The Internet users seem to accept to become their own travel agent organizing their trips themselves and build their own travels and leisure trips.

Moreover, technology gives suppliers the flexibility to react to market demands and the capacity to integrate diagonally with other suppliers to provide new combinations of services and improve cost effectiveness. In that sense, electronic tourism is a real global market because it has neither geographic limit nor time restriction. Information generally is up-to-date and potentially much richer, e.g., users can make a virtual visit of a hotel or a resort or they can freely plan vacation activities and choose the means of transport they prefer.

Given this consumer interest many (public) tourism organizations and industry players have arranged their information in an electronic format. However, this process has lead to a plethora of different information systems because each player has adopted its own business model and methodology. The result is a highly fragmented reality in which a high number of diverse actors participate with different cultures and back-grounds.

The problem is mainly the lack of a unique data model to share information among participants to the tourist market. Many attempts have been made in the past for gathering all the diverse systems under a unique standard (see www.rmsig.de for a list of different standardization initiatives) but the intended broad integration of electronic markets has never been reached, due to the great heterogeneity and diversity of tourism market. It represents a strong constraint to the evolution of the electronic tourism market and also to a broader participation of the small and medium sized players.

When taking a closer look at these different initiatives, one can observe that industry (e.g., tour operators, hotel chains, airlines, CRS/GDS) are standardizing their product descriptions and electronic interchange formats, using XML. The major initiative is OTA / TTI (Open Travel Alliance – Travel Technology Initiative). On the other side, destinations with their "obligation" to represents the entire offer within a geographical area, mainly SMEs, are lagging behind. Thus, as similar initiative to the industry field cannot be identified in the destination field, in most of the cases only database schemata are available. Just some exceptions such as TIScover, Austria, are successful enough to provide a XML based exchange formats. However, destinations are not willing to adapt to industry standards since they provide a rather limited view of the richness of a destination's offer. And we want to highlight that diversity is not only a disadvantage: it enables the specialization. Hence, it is also important to preserve it and to allow all the actors providing their contribution even if they are small and medium enterprises.

Data heterogeneity is a well known problem (in the tourism domain) (Stuckenschmidt et al., 2000; Dunzendorfer et al., 1998), and there exist several approaches for integrating heterogeneous databases (see, for example, Bouganim et al., 1999; Sha et al., 1999). However, still the most prominent and practical approach to solve it is to write ad-hoc data interface programs for each pair of communicating systems. Development and maintenance of these programs are expensive in terms of both time and money. The total effort required increases with the number of communicating systems (Omelayenko et al., 2001).

More flexible approaches are required and developments in computer science may provide more promising tools and methodologies to tackle the problem.
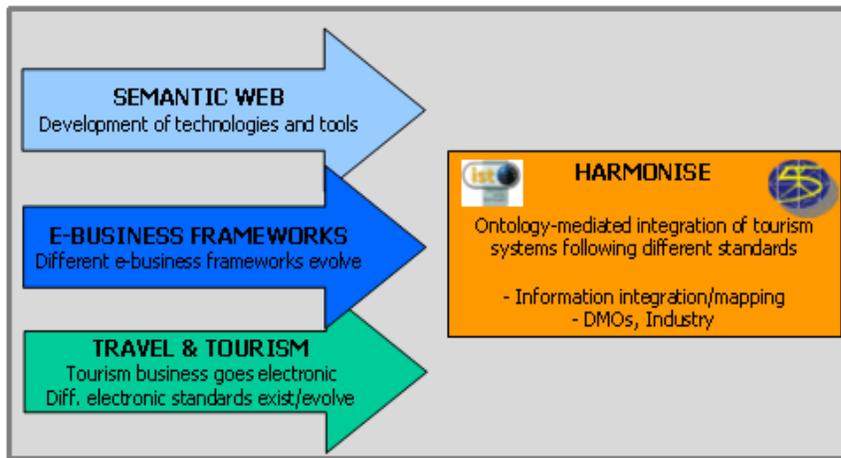
Figure 2. The Harmonise context

Figure 2 shows the context of Harmonise, the fast moving electronic travel and tourism environment together with recent developments in the semantic web domain and e-business frameworks such as ebXML. These developments are based on XML, which is also the base of the Harmonise approach. XML enables unified information representation on the syntactical level which is an important prerequisite for solving higher-level conflicts like structural and semantic heterogeneity. However, neither XML itself nor XML schemas provide sufficient mechanisms for the representation of data semantics. Ontologies can be used for the explication of implicit and hidden knowledge and have been introduced as a key solution for interoperability enabling data exchange (Omelayenko and Fensel, 2001; Visser et al., 2000; Ontoprise 2000; Melnik et al., 2000; Mello et al., 2001). The term ontology refers to an explicit, shared understanding of a given domain. It consists of a set of concepts, properly defined and related to each other. From a technology point of view, it can be seen as a repository of concepts, much like a database represents a repository of data.

ebXML (Electronic business extensible mark-up language), on the other side, provides a framework for modeling business processes and to provide interoperability on the business layer. For this purpose ebXML offers a modular suite of specifications. These specifications provide a standard method to exchange business messages, conduct trading relationships, communicate data in common terms and define and register business processes (see www.ebxml.org). Since it has not been the goal of ebXML – at least at the moment – to develop its own e-business vocabulary, it can be combined with a semantic web approach. This approach is taken in the

Harmonise project, in fact, no modeling of business processes will be designed and the platform will be ebXML compliant.

3.   THE HARMONISE SOLUTION

Setting up an ontology requires a social consensus process where an international consortium of tourism key players agrees on set of relevant concepts in the tourism domain. For this purpose, the Harmonise project is mainly structured w.r.t. to two tasks: an organizational one, which has the goal of gathering and composing a group of tourism experts who will generate the new marketplace, and technical one, which provides the necessary tools.

3.1   Organisational Task: The Tourism Harmonisation Network (THN)

One of the main milestones of Harmonise was to put together an enlarged consortium, called the Tourism Harmonisation Network (THN). This consortium had to involve the major European Tourism organizations and experts; and the major standard and systems owners.

This was a crucial step for the entire project and represents a key requisite for its success. At the moment this "Umbrella Structure" has reached a critical mass (however, it is still open for other participants), and it was officially formalized in January 2002, when many market participants expressed their willingness to become members of the THN, signing the related "Memorandum of Understanding". Among these organizations are National Tourist Board such as TourInFrance (France), SIGRT (Portugal), Finish Tourist Board and Spain Tourist Board; world tourism organizations such as WTO (World Tourism Organisation), besides IFITT (International Federation for IT and Tourism), which was already a project partner form the early start; tourism standard organizations like OTA/TTI (Open Travel Alliance /Travel Technology Initiative); and systems such as WhatsOnWhen or TIScover, etc.

The Tourist Harmonisation Network represents a guiding consortium with a sustainable business model, which allows the entrance of new users/partners, based on minimal rules. Indeed, the objective of Harmonise project is to seek interoperability not only for ontology and semantic issues, but also for business and organizational issues. The THN intends to provide an integration solution without imposing any closed approach to the operators. On the contrary, THN will provide operators with a powerful open

instrument, enabling them to "communicate" without losing out on singularity and avoiding at most any loss in data precision.

The THN will provide a set of products and services such as community services (knowledge centre, discussion forum, newsletters, members' contacts, list of certified enterprises, etc.) and other additional services (including training, annual conference, seminars, workshops, etc.). In addition, it will be responsible for the maintenance of the tourism ontology. THN is also keeping a close relationship with other initiative such as the ebXML group.

3.2 The Harmonise Platform

The tools provided by the Harmonise Platform (HP) allow the different tourism organizations to keep their proprietary data formats while exchanging information in a seamless manner. These proprietary data model formats (e.g. object vs. relational) are transparent to Harmonise, which is only concerned with the XML documents the partners are willing to export.

The data model of a source document (i.e., the XML document) is translated into the terminology specified by the ontology by means of a mediator, the Harmonisation tool. The granularity level of the local data models is also aligned to the representations used by the ontology. In this case the transformation between standards requires a linear number of "mappings" between systems, instead of exponential in the non-mediated scenario. The Harmonisation tool is based on the following three technologies as illustrated in figure 3:

– A tourism ontology, aimed at modeling and storing the basic concepts used in representing the content of information exchanges in tourism transactions (IMHO: Interoperability Minimum Harmonisation Ontology).
– An interchange format suitable to represent the instance data used within interoperable tourism transactions (HIR: Harmonise Interchange Representation).
– A set of mapping rules aimed at the transformation of interoperable transactions data from internal, proprietary formats to the Harmonise Interchange Representation, and vice versa. There will be a set of mapping rules for each different proprietary format; the mapping rules will be defined on the base of the correspondence of the specific proprietary data formats and structures with respect to the IMHO.
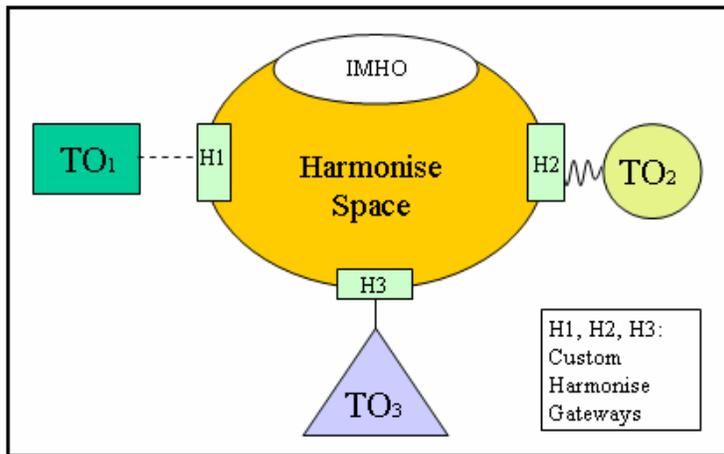
Figure 3. Harmonise space

A tourism ontology contains the relevant concepts, necessary to model the tourism environment (e.g., products, customers, logistics, partners and competitors) where tourism organizations operate in. An ontology, to be really effective, requires a wide acceptance by the (business) community of reference. Hence, in order to set up the Harmonise ontology, an exhaustive analysis of the most relevant standards which have been proposed within the tourism domain has been performed.

Based on the result of such analysis, the partners composing the Tourism Harmonisation Network (THN) agreed on identifying:
- the shared set of interactions which has to be carried out by the class of tourism actors represented (that, in a perspective view, has to be carried out by the respective information systems);
- for each interaction, the minimum common set of information elements to be exchanged and their data components;
- the logical relationships between these information elements;
- a common definition (naming and tagging policy) of the identified information elements and their logical relationships.

According to these common concepts, a first version of the Interoperable Minimum Harmonisation Ontology (IMHO) has been defined which covers some initial concepts. It is planned that the IMHO management process will be carried out within a continuous "validate and revise" approach. This will allow the update and extension of the IMHO content; and to include new tourism definitions or service offers which may be available in the future. The THN partners will agree on a planned release of updated versions of the IMHO to all Harmonise members.

The current adoption of XML, with its domain-oriented tagging to support the exchange of documents, and more generally, information, is a quite new application field for ontologies (Studer et al., 1999). XML mark up tags are simple strings, generally meaningful to humans, which require a precise semantic account to be treated by computers. This can be achieved by using ontologies and adding "semantic annotation" to the XML tags.

"Expressing meaning" is the main task of the Semantic Web approach which was defined as "web of data" in which background knowledge on the meaning of the data is stored directly for machine-understandable process. (Berners-Lee, 1998). The "Resource Description Framework" (RDF) and RDF Schema (the Schema Language for RDF) are W3C Recommendations for describing ontologies: RDF is a simple data model for representing data on the Web and a RDF Schema specification describes how to use RDF to describe RDF vocabularies. The specification also defines a basic vocabulary for this purpose, as well as an extensibility mechanism to anticipate future additions to RDF. The convenience of RDF Schema for conceptual modeling has been discussed in (Melnik, 1999; Decker et al., 2000; Omelayenko et al., 2001) and other sources. The Harmonise Interchange Representation (HIR) is an RDF document format built in accordance with IMHO.

The analysis of relevant standards and database schemata in the domain has shown that two different kinds of conflicts arise when an XML document has to be translated from one format to another one. These conflicts correspond to two different levels of modeling data structures or messages – the conceptual level and the physical level:
– Semantic clashes
– Structural clashes

"Semantic clashes" are clashes between concepts of different standards, or more precisely, between specific conceptual models or ontologies behind different standards. Typical semantic clashes are completely different concepts, different naming of concepts or concepts with a different structure. Some examples:
– Different naming: "postcode" vs. "postalcode";
– Different position: "Postcode" in "Address" instead in "Contact info";
– Different scope: "Telephone number" rather than. "prefix + telephone number".

"Structural clashes" are clashes or differences within the physical representation of concepts. Structural clashes may be caused by the usage of completely different physical representations (like XML-bases messages or fix-structured messages) or different XML-based representation languages (like XML-DTDs or XML schema descriptions). When using the same physical representation, structural clashes may be caused by a different

usage of specific syntactical constructs, e.g., by a different usage of attributes rather than embedded elements within an XML-based representation or by a different usage of domain value, e.g. short vs. int. etc.

A separation between semantic and structural clashes, of course, forces a separation between the corresponding steps in the XML document translation process. In fact, if this process deals with the two different types of clashes in two different phases, the semantic mapping (i.e. mapping between semantic clashes) is independent from the concrete physical representation of concepts. When different physical representations will be used in the future, the semantic mapping will still remain valid.

For this reason, within the Harmonise process the following two steps have been identified to solve the above conflicts:
– C-Normalization (Conceptual Normalization)
– Semantic Transformation (Mapping)

The C-Normalization step (figure 4) is required because of the multiplicity of XML document representation versus a unique conceptual model. Using XML as format for the data representation highlights that the same concept can be expressed in several different ways. Consequently, this step is dedicated to the transformation of the local data organization into a local conceptual schema, represented in a common normalized form of a XML document before performing the semantic mapping step.
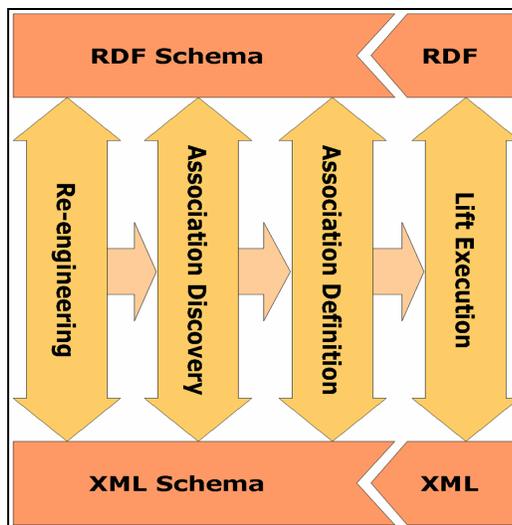

Figure 4. C-Normalization process

C- Normalization supports the semantic "lift" from the lower physical (structural) level to the higher conceptual (semantic) level. The inputs for

the process are heterogeneous logical schemata of local systems, in case of Harmonise participants represented by XML schemata. As output a *Normalized Conceptual Schema* shall be produced together with *C-Normalization Map* (input for the data normalization process at real time).

This second task, "Mapping process" deals with the semantic conflicts of the data being translated. As discussed above, this kind of conflicts is related to the diverse concepts of different standards, independent of the concrete physical representation. Harmonise defines a set of generic rules (reconciliation rules) which are customizable in order to be applied on specific systems.

In fact, tourism organizations who want to participate to this new tourism marketplace by sharing their data in the Harmonization space have to be provided with the Harmonisation Gateways which allow their data to be translated in HIR format. These gateways will be realized by means of two phases (figure 5): the customization phase, in which the user prepares the system to be "mappable", and the cooperation phase, in which the user can communicate with the others users (according to the "Local As View" approach).
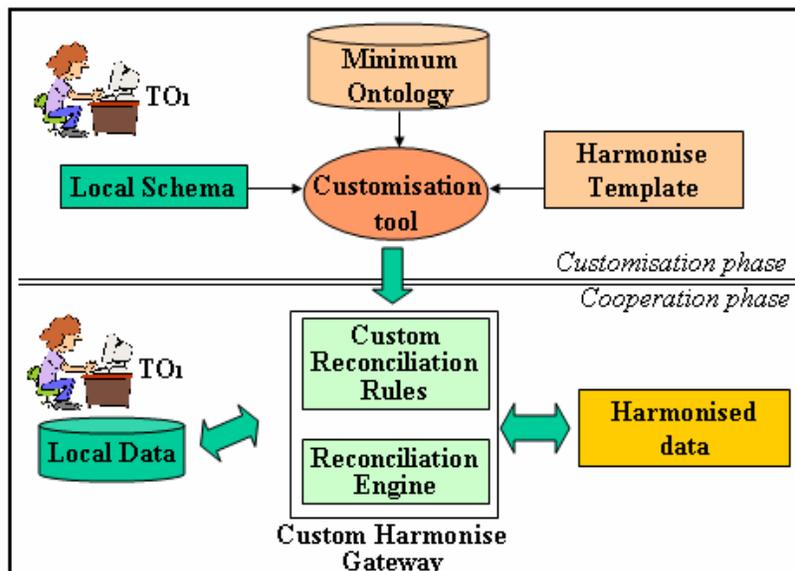

Figure 5. Harmonise phases

During the customization phase, the organization, through a graphical user interface program, accesses the IMHO, as well as the set of reconciliation rules template and identifies the clashes contained in its local

data structure and the rules needed to solve them. Then, the software which generates the personalized gateways, the Harmonise Gateway Generator, extracts the selected reconciliation rules and binds them, together with the reconciliation engine, creating the Custom Harmonise Gateway.

In the cooperation phase, the local data (coded in XML) are transformed by means of the Custom Harmonise Gateway, into the Harmonise Interchange Representation format. Then, this HIR is sent through the net to the other system and will re-transform it by the Custom Harmonise Gateway at the other system into the respective local data format.

The interaction among the users is then direct and not mediated by hard-coded translators and the result is that the Custom Harmonise Gateway acts as a semantic gateway between the systems, permitting the receiver to view the source as an extension of its own database, without concern for the differences in names and representations of data.

4. CONCLUSIONS

In order to solve the interoperability problem we have proposed an Ontology-mediated integration process based on a framework for data integration – instead of a new standard for tourism domain.

The Harmonise Platform allows the different tourism organizations to keep their proprietary data format and to exchange information based on such ontology by the means of a mediator module. This tool is based on *RDF*, for information representation and exchange; on *ontologies*, for knowledge and content management; on the *mapping of heterogeneous information sources*, for the reconciliation method.

A wide social consensus process, the THN will be set up as legal self-sustainable organization in order to be responsible for the maintenance of the knowledge base, the infra-structure and software, for the development of the ontology (IMHO), for supporting the software users and for deepening the research within the interoperability domain. Currently, ongoing technical tasks are:

a) Ontology building
b) Definition of Local Conceptual Schema
c) Formal definition of mapping rules

5.   REFERENCES

Berners-Lee T. (1998): "Semantic Web Road map", Internal note, World Wide Web Consortium. See http://www.w3.org/DesignIssues/Semantic.html.

Bouganim, L., Chan-Sine-Ying T., Dang-Ngoc T., Darroux J., Gardarin G., Sha F. (1999): "Miro Web: Integrating Multiple Data Sources through Semistructured Data Types", VLDB 1999: 750-753

Decker S., Melnik S., van Harmelen F., Fensel D., Klein M., Broekstra J., Erdmann M., Horrocks I., (2000): "The Semantic Web: the Roles of XML and RDF", IEEE Internet Computing, Sept./Oct. 2000

Dunzendorfer, A, Küng, J., Wagner, R. R. (1998): "Data Access to Heterogenuos Tourism Information Systems", In: Buhalis, D., Jafar, J., Tjoa, A M. (eds.): Information and Communication Technologies in Tourism. Springer, Wien, New York, 46-54, 1998.

Florescu, D., Levy, A., Mendelzon, A. (1998): "Database techniques for the World-Wide Web:a survey", SIGMOD Record, 27/3, 1998, 59-74.

Mello R. and Heuser, C. (2001): "A Rule-Based Conversion of DTD to a Conceptual Schema", Lecture Notes in Computer Science, vol. 2224, pag. 133, 2001

Melnik S. and Decker S. (2000): "A Layered Approach to Information Modelling and Interoperability on the Web", Proc. ECDL'00 Workshop on the Semantic Web, Lisbon, Portugal, Sept. 2000.

Melnik S. (1999): "Generic Interoperability Framework", Working Paper, Department of Computer Science, Stanford University. See http://wwwdiglib.stanford.edu /diglib/ginf/WD/ginf-overview.

Omelayenko B., and Fensel D. (2001): "Scalable Document Integration for B2B Electronic Commerce", special issue of ELECTRONIC COMMERCE RESEARCH JOURNAL on B2B RESEARCH, 2001

Omelayenko B., and Fensel D. (2001): "An Analysis of the Integration Problems of XML-Based Catalogues for B2B Electronic Commerce" presented at IFIP 2.6 Working Conference on Data Semantics (DS-9), Hong Kong, China, 2001.

Ontoprise GmbH vision and emerging products document, (2000): "White paper: B3 Semantic B2B Broker", Dec. 2000.

Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation, 22 February 1999, See http://www.w3.org/TR/1999/REC-rdf-syntax-19990222 .

Sha, F., Gardarin, G., Némirovski, L. (2000): "A Semi-Structured Data Cartridge for Relational Databases", ICDE 2000: 192

Stuckenschmidt H. and Wache H. (2000): "Context Modeling and Transformation for Semantic Interoperability", Knowledge Representation meets Databases - Proceedings of the Workshop at ECAI 2000.

Studer R. and Erdmann M. (1999): "Ontologies as Conceptual Models for XML Documents", Proceedings of the 12th Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW'99), Banff, Canada, October 1999.

Visser, U., Stuckenschmidt, H., Wache, H., Vögele, T. (2000): "Enabling Technologies for Interoperability" in Workshop: Information Sharing: Methods and Applications at the 14th International Symposium of Computer Science for Environmental Protection, Visser, U. and Pundt H. (eds.), volume 20, pp. 35–46, Bonn, (2000).

Werthner, H., Klein, S. (1999): "Information Technology and Tourism. A challenging Relationship", Springer Verlag, Vienna – New York, 1999.