

Ontology mapping using multiple dimension approach

Nuno Silva and João Rocha

Departamento de Engenharia Informática

Instituto Superior de Engenharia – Instituto Politécnico do Porto

Rua Dr. António Bernardino de Almeida, 431

4200-072 Porto – Portugal

{Nuno.Silva, Joao.Rocha}@dei.isep.ipp.pt

Abstract: Semantic Web is a new emergent paradigm for the Web. It considers that current Web potentialities are not being exploited completely, specially because its use is currently user-centered. It suggests that the computer become part of web processing and not only the user interface. One of the most relevant problem is the partial or complete incompatibility between information communities. This incompatibility is currently manually solved. Semantic Web paradigm suggest to complement data with machine processable meta-data which allowed semantic-web aware systems to reason and infer new knowledge, and eventually define semantic bridges between two information communities permitting to transform on the fly data between those communities. In this paper we analyze several dimension of the problem, revising related approaches and proposing a consistent mapping process, which is the base of the prototype service on ontology mapping being developed.

Keywords: Semantic Web, Ontology, Interoperability, Mapping, Web Services.

1. INTRODUCTION

Information¹ integration has been the focus of substantial research for several years in different communities like Databases (Bergamaschi et al., 2001, Madhavan, Bernstein, and Rahm, 2001), Data-warehouse (Critchlow, Ganesh and Musick, 1998), Ontology (Neumann et al., 2001, Visser and Tamma, 1999). Different goals, terminology and contexts lead to some confusion and miss-understandings respecting approaches, techniques and overall solutions proposed. In (Rahm and Bernstein, 2001) a good analysis and classification of different dimensions of the problem is presented, and in (Pinto, Gomez-Perez and Martins, 1999) it is proposed a clear and pragmatic terminology of adopted strategies.

Semantic Web² vision propose the representation of data semantics in a machine processable way, and its exploitation in order to cope with the discrepancy between the huge amount of existent data and minor knowledge perceived by humans. Such vision boosted research in information integration in automatically aligning data according to user semantics requirements. The World Wide Web Consortium³ proposed several standards for serializing (XML), structuring (RDF/S), transforming (XSL/ XSLT/XPATH) information, arising as an important framework for access and exchange information across heterogeneous information system. Additionally, complementary standards were proposed (DAML+OIL: DARPA Agent Markup Language+Ontology Inference Layer⁴) and others being specified (OWL: Ontology Web Language⁵) with ontology as common central concept. Ontology has not a straight forward acceptable definition (Ushold and Gruninger, 1996; Guarino, 1994; Gruber, 1993), but in this context it should be understood as the explicit, formal and partial specification of a conceptualization. The specification of data accompanying ontologies allows both machines and humans to reason and understand data, and it permits to relate data from heterogeneous source at semantic level, minimizing meaning errors in aligning results.

Besides existence of many standards to exchange data between systems (EDIFACT, PIF, cXML, ebXML, BizTalk, etc.), and many ontologies for almost all type of organization and its processes, systems integrators are feeling the same problems formerly known as automation islands. UDDI initiative⁶ for example tries to cope with some interoperability problems like service discovery, service interface specification and communication protocols. Providers specify interfaces and interoperability protocols for their services, which in turn are published and made available in a unique (although possible distributed) repository. Consumers consult the repository and access the specification of interfaces and protocols, which are then implemented in their system to access services. However, no further facilities are provided in developing interfaces and protocols, which severely limits dynamicity. Additionally, both development and operation phases are data oriented, which means the interoperability semantics is not explicitly defined and thus depends on the developer interpretations. Some related projects are WSDL and SOAP. Indeed, not enough research has been carried out in integrating information systems in virtual enterprises, e-business, B2B, B2C, etc. scenarios, characterized by a extremely dynamic behavior, composition, structure, organization, behavior, goals, strategies, etc. It is necessary to boost the application of Semantic Web potentialities and developed the so called Intelligent Web Services⁷.

We suggest the development of a special kind of alignment service that dynamically aligns data from two information sources described by ontologies. The proposed strategy, not mentioned in (Pinto, Gomez-Perez and Martins, 1999) is known

¹ This work is partially supported by the FCT project POCIT/2001/GES/41830.

² <http://www.semanticweb.org>

³ <http://www.w3.org>

⁴ <http://www.ontoknowledge.org/oil/oilhome.shtml>

⁵ <http://www.w3.org/2001/sw/WebOnt/>

⁶ <http://www.uddi.org>

⁷ <http://www.cs.vu.nl/~dieter/wese>

as mapping between ontologies (Neumann et al., 2001; Park, Gennari and Musen, 1997; Doan et al., 2002). It considers that completely accuracy of alignment is not always required, which allows to process it faster and focus in parts of data relevant for a specific interoperability. Additionally, this strategy permits to align heterogeneous data sources or heterogeneous ontologies in an inexpensive and effective way. In contrast, merging or integrating data sources in such conditions does not make much sense, and respective results would be extremely poor.

In next section will describe generic scenarios and related requirements challenging web services and specially ontology mapping service. In third section we propose several alternative conceptual models for ontology mapping. In section 4 we propose a conceptual mapping process and describe strategic and technical approaches to each of them. In section 5 some summary is presented and future work suggested.

2. SEMANTIC WEB SERVICES

Some recent analysis (e.g. (Rahm and Bernstein, 2001)) shows relevant similarities between some research initiative in the context of Semantic Web and other projects and initiatives from the 80's and early 90's on integrating information systems. In the 80's the Web was only a mirage but nowadays it is extremely relevant for people and business in exchanging private or business contents. However, data and services populate the internet in such quantity that searching for a specific item is no longer a straight forward task and user goals are easily swamped. More formal Internet users like vendors and consumers or B2B entities are facing problems in aligning communication contents in a meaningful and effective way. Search engines evolved a lot from the early times and are no longer a simple catalogue of web pages. More and more content analysis, categorization, indexing, etc. is performed which allows its use with minimal changes since its roots. Even considering the extremely powerful techniques being developed and applied to search engines it seems clear that currently paradigm (data oriented) will not fulfill the upcoming requirements in application domains like e-business and knowledge management, where semantic and dynamicity play a crucial role in interoperability processes. Semantic Web was proposed to cope with this gap (Berners-Lee, 1999) and suggests to characterize the data being supplied, i.e. define data over the data. (Melnik and Decker, 2000) suggests a distinction and decomposition of information in the Semantic Web into three layers: syntax, data and semantic; in the same sense as protocol stack in internetworking.

According to the previous considerations, two scenarios are now introduced allowing to present describe a real world requirements and constraints:

1. Information retrieval, using a search engine. In a Semantic Web context, ontologies complement data, which allows search engine to better analyze and classify it internally, and supply it to the user according to its specific requirements. In this scenario, accuracy may not be the main requirement, but the process must be fast and computational inexpensive;
2. Interoperability in marketplace. Two or more entities must align their heterogeneous information systems prior to any business activity. In contrast to previous scenario, the alignment accuracy is crucial to business interoperability, which in turn may need to use superior computational power.

We believe that in near future interoperability should be extremely fast and emergent, in contrast to some similar processes already running, that rely on very complex setup integration processes prior to any business interoperability. In that sense we suggest to apply a light alignment process, defining associations and translation methods between elements from two ontologies. Notice that no single database or integration view is created but on contrast, the entities information system are kept separate and heterogeneous.

It is necessary that main actors in this powerful infrastructure (search engines or marketplaces for example) allow and exploit this new (data-) dimension, but it is also necessary to develop mechanisms that permit the user to specify semantic requirements, drive the process and automate it as much as possible through the entire process. In that respect we identified six dimensions of the alignment problem that should be considered:

- Extension of the mapping, which depends both on the heterogeneity of the two aligning ontologies (the overlapping extension) and the requirements of the application (how much should be aligned to permit interoperability);
- Accuracy of the mapping, i.e. how much correct should be the mapping. A 100% correct mapping is conceptually impossible except if both source and target ontologies are the same. It depends on three factors:
 - The application scenario which can request high level accuracy (e.g. e-business application) or less accuracy (e.g. searching information);
 - Ontologies heterogeneity. The much ontologies are semantically different, the less accurate the mapping can be;
 - The mapping extension required. The longer the mapping extension is required the less accurate the mapping can be.
- Time necessary to process the mapping, which depends on the previous referred constraints: heterogeneity, application scenario, accuracy and extension;
- Dynamicity refers to the manner instances are delivered to the final target, and it mainly depends on the application scenario. Two alternatives are possible:
 - All the instances are delivered at once. Both mapping and transformation process can be perform at once and could be closely integrated in some implementations;
 - Instances are delivered on demand, which evidence the distinction between the mapping and the transformation processes. The mapping process is than considered a setup phase while the transformation process is an operational process during the business interoperability.

- Evolution. Consider a long time interoperation scenario or a library of previously processed mappings. In these cases, the underlying ontologies may evolve, implying mappings became outdated. The mapping process must be aware of this dimension;
- Cooperation between ontologies owners in the mapping process. This dimension clearly depends on the application scenario, but specially on the ability ontologies owners have derive mappings and to cooperate with interoperability partners or the mapping process. A higher level participation is typical when the interoperation depends on a human agent or when intelligent software agents represents the physical entities. Lower or none cooperation typically occur in scenarios when the accuracy level is not so relevant or when one the partners (eventually the mapping service) decisions are mandatory.

Most of times these dimensions are mutually contradictory, meaning that in order to achieve one the others are prejudiced. Priorities or limits must be established for each of them to preserve each dimension according to its relevance in each specific mapping.

A service that would negotiate these issues without (or minimal) human intervention would be a huge improvement during interoperability, and it represents one of the most important steps in automatizing the semantic-based interoperability process.

3. CONCEPTUAL MODELS

This section intends to describe conceptual strategies of such service, respecting entities roles in the process and two fundamental models two derive mappings between multiple ontologies.

One of the most important issue in dealing with multiple ontologies scenarios is the possibility to establish a common unifying ontology, which would reduce the number of mappings necessary to perform interoperability. In a community of n entities, this corresponds to $\sum_{i=1}^n (n-i)$ mappings, considering that mappings are bi-directional. In case exists a common ontology, the number of mappings reduces to n , because all ontologies map to the common one, and from this to the others. However, a two step translation is necessary which requiring greater computation effort. Additionally this two step procedure will cause higher loss of semantics than a single step procedure which in some case would be unacceptable. CYC and IEEE SUO are two examples of higher level context independent ontologies that could eventually be used as a base in a *lingua franca* approach, but in pragmatic scenarios like marketplaces, specific ontologies would be supplied which reduces semantic discrepancies. Figure 1 depicts these two approaches.

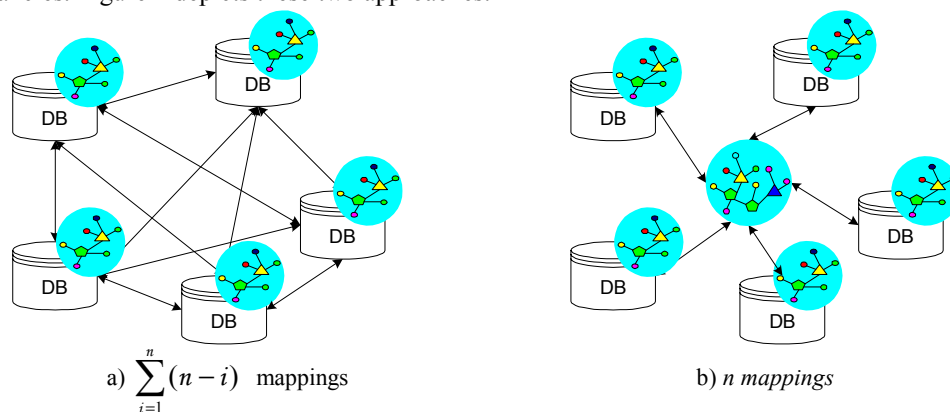


Figure 1 – Conceptual models according to participation of entities

Using one model or the other is a matter of application scenario, but nevertheless we believe that the common ontology approach would be very much used in the future, because it would be the natural evolution of current systems. In fact, users expect that the interface to the service remains unchanged, not worrying with the internal process. For example, in querying a search engine, nowadays the user has the chance to specify a searching starting point in the information system taxonomy, which corresponds to constraining the requested information. In the future, user would have the possibility to specify complex ontology, which will be used to specify information requirements. The search engine will translate user requirements to its own ontology and than to the information source one, and query it accordingly. This is a typical common ontology application scenario, and similar situation would occur in e-business and marketplaces.

An ontology mapping service may assume three different approaches in what concerns its relation to the entities (Figure 2):

1. External independent service, In this approach entities agreed to delegate in a third party entity the mapping process. This model is described in Figure 2 between Entity B and Entity C. Each of the entities deliver its ontologies to the external service and all the mapping process is carried out by the external service;
2. Internal service in both entities. Each entity regulate the mapping process (eventually) in different proportions, where an extreme situation would be if one of the entities would rule the entire process. This model is represented in the relation between Entity A and Entity B. In this situation both entities must implement the mapping service internally;
3. Delegation in an external entity. One (or both) of the entities delegates in a external service its representation in the mapping process thus performing an interactive mapping process like in previous model, but owners are not directly

involved. This situation is represented in Figure 2 in the relation between Entity A and Entity D. Entity D delivers its ontology to the external service, with whom the Entity A negotiate the mappings.

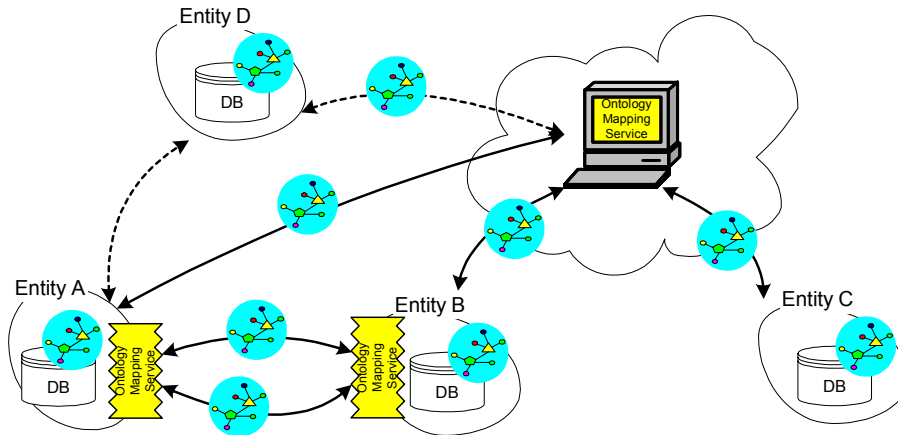


Figure 2 - Mapping service models according to entities relations

Application of each of this models depends both on the capabilities of the entities to derive the mappings, but also on the level of trust between entities. In concurrent/cooperative mapping process, entities share private knowledge, which in many application scenarios its is not acceptable. But the same holds in requesting an external service, since the entities must share their private knowledge with that service. Mechanisms to provide confidentiality and security between entities are out of scope here. Additionally, it is clear that trust between entities and entities capabilities to individually derive mappings are not sufficient conditions if the goal is to cooperatively derive mappings. In that sense, entities involved in the process must have high level negotiation and arguing capabilities. This activity, called ontology negotiation is still in early research stages (Bailin and Truszkowski, 2001) but we believe that it will be of major importance in near future.

4. MAPPING PROCESS

According to previous considerations we know introduce and describe the identified phases in the ontology mapping process. It is important to notice that some of them would not be present in every mapping system:

Normalization is the process whereby ontologies are translated to a unifying model and language in order to minimize reasoning incoherencies and maximizing processing abstraction (Omelayenko and Fensel, 2002) even if in such translation some semantic information is lost. Additionally, this phase perform syntactical unification (e.g. abbreviations, acronyms). This phase is not mandatory but it would be frequently necessary, not only because of the representation language (XML, RDF/S, DAML+OIL, Ontolingua, OWL, etc) but also because of the specificities of the represented domain. It is important to notice that this process should not make any semantic change, besides the normalization concerned with representation primitives. As an example for the syntactical normalization, imagine one ontology referring to *id* and another one referring to *identification*. Translating *id* for *identification* does not (at least apparently) make any semantic inference and transformation, but if the terms were *name* and *identification*, some semantic transformation occur, so it should not be done in this phase. Syntactical normalization operations includes but are not limited to standardization in acronyms (expansion), abbreviations (expansion), tokenization, prefixes, conjunctions, prepositions, articles (elimination).

Similarity measuring phase is the set of processes that specifies similarities between elements from both ontologies. We understand ontology elements as ontology constituents that share certain characteristics. For example, concepts can be understood as the aggregation of properties and attributes; properties may be specified as relations between two or more concepts; and attributes may be seen as special kind of properties that relate a single concept and an atomic value. Also, a cluster of concepts can be considered an element that can be compared. Four similarity measuring approaches exists:

- Syntactical analysis measure similarities between ontology terminology considering elements as set of characters. Simple techniques, like edit distance proposed by Levenshtein, are normally applied;
- Lexical analysis measure similarity according to the terminology semantic meanings and lexical relations. Typical approaches use dictionaries and other lexical tools like WordNet;
- Structural analysis consider the relations between elements in the ontology, focusing in its hierarchical relations. Clustering techniques, based on both syntactical and lexical similarities are typically used in this analysis;
- Extensional analysis defines similarities between instances. This technique has been widely used in semi-structured data sources like HTML documents, where a formal description was not available. With the increased use of ontologies this analysis had been somehow disregarded in research but it is increasingly assuming once again special relevance. In fact, its arising relevance is specially related to the transformation phase where unique identification and data transformation refinement are two examples of relevant application. Figure 3 presents a simple example of such situation: at ontology level would not be possible to define mapping between instance values, thus the value of Course attribute would not be filled if extensional analysis would not be performed.

Bridging phase corresponds to two substantially different tasks, although closely related:

- Association sub-phase take into account the similarity measures found in previous phase to define semantic associations between elements from both ontologies. Associations are categorized according to following dimensions:

- Cardinality, concerns to the number of objects intervening in both sides of the mapping, ranging theoretically from 1:1 to m:n. However it is not convenient to allow m:n mappings, but these can be transformed into 1:n and n:1 mappings;
- Scope dimension relates to the mapping specification much like the OO modeling perspective (abstraction, hierarchy, encapsulation, etc.). No research is done in this domain;
- Element dimension respects to the type of elements being associated. In most common type of ontologies three kind of objects are considered: Concepts, Attributes and Properties. A bridge always relate two or more objects of the same type. For example, in a 1:n-concept bridge, only concepts are related, in this case 1 concept in the source ontology and n in the target one.
- Definition of functions. This sub-phase intends to specify the functions necessary to transform instances from the data source, described according to source ontology, to the target ontology representation. In simplest cases, the necessary function would be just a copy of values, but in complex cases, cross concept/relations procedures could be necessary. This is probably the phase that less possibilities has to be automated, and thus, more user/expert intervention is required.

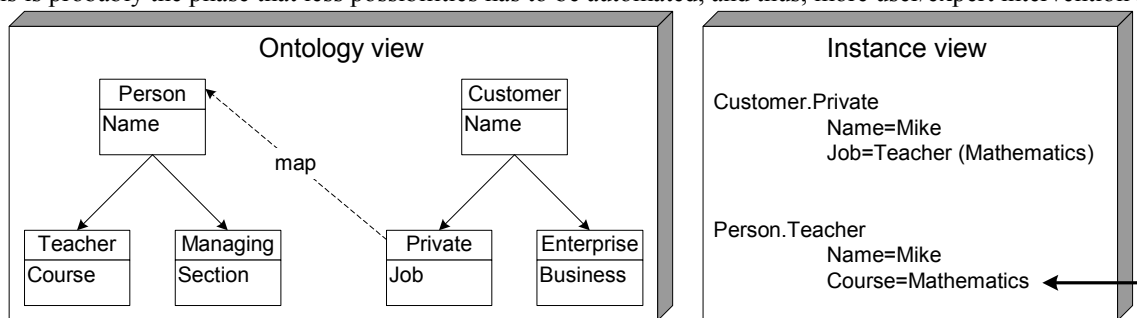


Figure 3 – Ontology view Vs. Instance view: data may complement ontology mapping

Representation phase intends to represent the previously defined bridges according to an external representation language. This representation is necessary when the execution phase is performed by other entity than that who specified the bridges. In this sense, this phase is not mandatory if the bridging and execution phases are carried by the same entity. Nevertheless, in order to reuse those mappings, it is recommended to represent them in an external, eventually widely accepted standard language. However, such language, besides a minimal proposal (Omelayenko, 2002), does not currently exist.

Transformation phase intends to effectively translate instances from one data source to the other, according to the semantic bridges specified in the bridging phase. Bridging and transformation phases are independent and we believe that the external and internal services will be used interchangeably, which substantiate the requirements for powerful formalized representation mechanisms. We suggest the usage of standard transformation engines like TRIPLE or those based on XSLT. TRIPLE is very general which permits to process any formalized language, but in contrast it does not scale up. XSLT is very verbose, hard to code, which does not fit our dynamicity requirements. However it is a widely accepted standard, worthing to carry out some research in dynamically represent the bridges over it.

Negotiation phase concerns with the process whereby the intervening entities try to achieve a consensus about the semantic bridges. The negotiation process depends extensively in two components: the negotiation protocol and the arguing capabilities of the entities. There is little research on this domain and not specifically applied to ontology mapping. Also, no argumentation is considered. We suggest to apply the information found in previous phases to argue and justify the decisions, which may assume special relevance in arguing about the proposed associations. The negotiation process may assume two distinct models according to the moment the negotiation occurs:

- Bridge by bridge: for each derived bridge a negotiation procedure occurs. Its advantages respect the possibility to drive the mapping process according to both entities interests;
- At the end of the bridging phase: both entities complete the bridging phase and then execute the negotiation. Advantages of this approach concern the possibility to perform a top-down argumentation, justifying from generic mappings through more specific ones.

It is obvious the advantages of one approach represent the disadvantages of the other. However, it is perfectly possible to combine both approaches, which allows to cope with each other disadvantages while maintaining the advantages.

5. SUMMARY AND FUTURE WORK

This paper described several dimensions of the ontology mapping problem in the context of the Semantic Web, and specially in respecting very dynamic application scenarios like marketplaces and information retrieval. Different strategic approaches were identified and described, which allowed to systematize the mapping process in 6 different dimensions and phases. Currently we are focusing on the bridging questions, a phase where some research has been done but less formalization (e.g. cardinality and scope) exists. This will be consubstantiated in a mapping meta-ontology that will permit to formalize mapping concepts in a first stage, and further represent the mappings.

One of the less analyzed dimensions of the problem is Evolution, which limits the specification of a conceptual architecture encompassing the overall mapping life-cycle.

Considering that mapping is an inherently subjective process, no sufficiently accurate mappings are achievable without user intervention. In that sense we are researching the inclusion of machine learning strategies.

REFERENCES

- Bailin and Truszkowski, 2001; Sidney C. Bailin, Walt Truszkowski; "Ontology Negotiation between Agents Supporting Intelligent Information Management"; Workshop on Ontologies in Agent Systems at the 5th International Conference on Autonomous Agents; Montreal, Canada; May 2001.
- Bergamaschi et al., 2001; S. Bergamaschi, S. Castano, D. Beneventano e M. Vincini; "Semantic Integration of Heterogeneous Information Sources", Special Issue on Intelligent Information Integration, Data & Knowledge Engineering, Vol. 36, Num. 1, Pages 215-249, Elsevier Science B.V.; 2001.
- Berners-Lee, 1999; Tim Berners-Lee, Mark Fischetti (Contributor), Michael L. Dertouzos; "Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web"; 1999.
- Critchlow, Ganesh and Musick, 1998, T. Critchlow, M. Ganesh, R. Musick; "Automatic Generation of Warehouse Mediators Using an Ontology Engine"; in Proceedings of the 5th International Workshop on Knowledge Representation meets Databases (KRDB'98); May 1998.
- Doan et al., 2002; AnHai Doan, Jyant Madhavan, Pedro Domingos, Alon Halevy; "Learning to Map between Ontologies on the Semantic Web"; in Proceedings of the World-Wide Web Conference (WWW-2002); 2002.
- Gruber, 1993; T.R. Gruber; "A Translation Approach to Portable Ontology Specifications"; Knowledge Acquisition 5, 2, 199-221; 1993.
- Guarino, 1994; N. Guarino; "The Ontological Level"; Invited paper Presented at IV Wittgenstein Symposium; Kirchberg, Austria; in R. Casati, B. Smith and G. White (eds.), Philosophy and the Cognitive Sciences; Vienna, Austria; 1994.
- Madhavan, Bernstein, and Rahm, 2001; Jayant Madhavan, Philip A. Bernstein, and Erhard Rahm; "Generic schema matching with Cupid"; in Proceedings of the 27th International Conferences on Very Large Databases, pages 49-58; 2001.
- Melnik and Decker, 2000; S. Melnik and S. Decker; "A Layered Approach to Information Modeling and Interoperability on the Web"; in Proceedings of ECDL'00 Workshop on the Semantic Web; Lisbon, Portugal; September 2000.
- Mitra et al., 2000; Prasenjit Mitra, Gio Wiederhold and Martin Kersten; "A Graph-Oriented Model for Articulation of Ontology Interdependencies"; in Proceedings of Conference on Extending Database Technology, (EDBT 2000); Konstanz, Germany; March 2000.
- Neumann et al., 2001; H. Neumann, G. Schuster, H. Stuckenschmidt, U. Visser, T. Vögele and H. Wache; "Intelligent Brokering of Environmental Information with the BUSTER System"; in Proceedings of the International Symposium Informatics for Environmental Protection, pages 505-512; Zürich, Switzerland; October 2001.
- Omelayenko, 2002; Omelayenko B.; "Integrating Vocabularies: Discovering, Representing and Compiling Vocabulary Maps"; in Proceedings of the First International Semantic Web Conference (ISWC-2002); Sardinia, Italy; June 9-12, 2002.
- Omelayenko and Fensel, 2002; Omelayenko B. and Fensel D.; "Scalable Document Integration for B2B Electronic Commerce"; submitted; 2002.
- Park, Gennari and Musen, 1997; J. Y. Park, J. H. Gennari, & M. A. Musen; "Mappings for Reuse in Knowledge-based Systems"; SMI Report Number: SMI-97-0697; 1997.
- Pinto, Gomez-Perez and Martins, 1999; H.S. Pinto, A. Gomez-Perez, and J.P. Martins; "Some issues on ontology integration"; in Proceedings of the IJCAI'99 Workshop on Ontology and Problem-Solving Methods: Lesson learned and Future Trends; volume 18, pp. 7.1--7.11; Stockholm, Sweden; August 1999.
- Rahm and Bernstein, 2001; E. Rahm and P. A. Bernstein; "A survey of approaches to automatic schema matching"; in The VLDB Journal 10: 334-350; 2001.
- Ushold and Gruninger, 1996; Ushold, M. and Gruninger, M.; "Ontologies: Principles, Methods and Applications"; The Knowledge Engineering Review, 11(2): 93-136; 1996.
- Visser and Tamma, 1999; Visser, P.R.S & Tamma, V.A.M.; "An Experience with Ontology-based Agent Clustering"; in Proceedings of the IJCAI 99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends, pp.12-1-12-13; Stockholm, Sweden; August 1999.