Exploring a New Approach to the Alignment of Ontologies[*]

Isabel F. Cruz[†] and **Afsheen Rajendran**
Department of Computer Science
University of Illinois at Chicago
851 South Morgan Street (M/C 152)
Chicago, Illinois 60607-7053

Abstract

We address the issue of data integration over semantically-heterogeneous data sources using an ontology-based integration mechanism. The semi-automatic techniques that we explore are directed to helping experts establish mappings between ontologies using a combination of manual methods, which are needed for the accuracy of the mappings, and automatic methods, which facilitate the experts' tasks. We illustrate our approach using examples from a geospatial application for querying land use patterns in the State of Wisconsin.

Introduction

Ontology-driven approaches to data integration minimize maintenance and scalability problems, as the autonomous distributed data sources can be added or removed from the integrated system as necessary. Such approaches have been proposed in the geospatial and biological domains (Fonseca & Egenhofer 1999; Ludaescher, Gupta, & Martone April 2001).

We assume the existence of a *central ontology*, which describes the domain of interest, and we will be referring to the sources (and the ontologies associated with them) either as *distributed*, as seen from a central integration site, or as *local*, as seen from the site of that source.

An application that needs to use the data from the heterogeneous sources expresses its information requests in terms of the entities in the central ontology thus giving users the appearance of a single homogeneous data source. In order to integrate the heterogeneous data, mappings between entities in the central ontology and those in the distributed ontologies have to be determined. Such mappings can be used in *ontology merging*, where information from the local ontologies are to be included in a coherent, single ontology or in *ontology alignment*, where the ontologies are to be made coherent with one another, but kept separately (Noy & Musen 2000).

There are different types of heterogeneities. Syntactic heterogeneities arise due to differences in the representation of the same conceptual model (e.g., relational or object-oriented models). Schematic heterogeneities arise due to structural differences within the same data model. Even when the representation and structure are the same, naming and cognitive heterogeneity might exist—the former type corresponding to entities that are the same but have different names, and the latter type corresponding to entities that perform multiple roles in different contexts. For example, an agricultural expert perceives water ways to be a source of irrigation, while a transportation expert perceives them as a mode of transportation. Naming and cognitive differences of the entities to integrate lead to semantic heterogeneities (Kashyap & Sheth 1998), which have been recognized as the hardest ones to solve (Bishr 1998).

In this paper, we focus on the mappings between the central ontology and a local ontology and take a new view at how to determine these mappings so as to allow for the partial automation of the mapping process. We look at several examples that are provided by the Wisconsin Land Information System (WLIS). In particular, we focus on land usage and on the semantic heterogeneity that is evidenced in this particular domain (Wiegand *et al.* 2002). The land use database system that we consider stores information about land parcels in XML format. Sample XML data about a land parcel contains an identification number for the parcel, the category of land usage under which it is classified, the name of the file that contains the pertinent shape information, and information about the owner of the parcel.

In WLIS, semantic heterogeneities manifest themselves in that not only the attribute names for the land use code vary, but also the classification codes themselves vary from county to county and even within the same county. For example, the land use classification scheme for the city of Madison is different from the one that is used in the rest of Dane county, which contains Madison. Table 1 illustrates such heterogeneities.

What differentiates our automatic methods from related approaches (Hovy 1998; McGuinness *et al.* 2000; Noy & Musen 2000; Gennari & Musen 1998; Bergamaschi, Guerra, & Vincini 2002; Corcho & Gomez-Perez

[*] This was work was partially supported by NSF Awards EIA-0091489 and ITR IIS-0326284.

[†] Corresponding author. Email address: ifc@cs.uic.edu.

Planning Authority	**Attribute**	**Code**	**Description**
Dane County	Lucode	91	Cropland Pasture
Racine County	Tag	811	Cropland
		815	Pasture and Other Agriculture
Eau Claire County	Lu1	AA	General Agriculture
City of Madison	Lu_4_4	8110	Farms

Table 1: Examples of heterogeneity of attribute names and values in WLIS.

2001) are the deduction operations that can be propagated along the ontologies. We identify the cases where such deduction operations can be performed automatically or where the user has to intervene manually. We anticipate that our approach could be used in conjunction with some of the tools and techniques of the related approaches.

In the ontologies that we consider (for land use management), an ontology is a hierarchy where entities refer to the codes (the vertices in the hierarchy) and relationships are established between a parent code and a child (the edges between the corresponding vertices). Such relationships represent generalization/specialization between the codes. In our case, there are no explicitly represented properties or attributes associated with the codes. Therefore, we have a simpler structure than that found in other systems, such as frame-based systems (Noy & Musen 2000; McGuinness *et al.* 2000), and the decision of whether two entities match has to be solely based on the codes.

The rest of the paper is organized as follows. After a short discussion on data integration, we introduce the mapping types that we consider, followed by the deduction operations for the semi-automatic alignment process. After that, we discuss some aspects of ontology merging and we conclude with directions for future research.

Data Integration

There are two approaches followed for specifying the mappings between the global schema (or ontology) and the local data sources. In the first approach: the *global-as-view (GAV)* and the *local-as-view (LAV)* approach. In the LAV approach, the system can be easily maintained and extended. When a new source is added, only its definition is provided, without entailing changes in the global schema. In the GAV approach, system maintenance is more difficult as adding a new source may require changing the definition of the entities in the global schema. On the other hand, query processing techniques needed in the LAV approach are known to be more sophisticated than in the GAV approach (Lenzerini 2001).

In a fluid network of data sources, such as in WLIS, a source-centric approach is preferable, so as to have the capability to add or update new local data sources with minimal effort, as such data sources become available in electronic form. In some geographic applications, the above mentioned difficulties in using the LAV approach are alleviated in that we do not need to combine answers from different data repositories in order to produce results for a particular region (Cruz & Rajendran 2003).

However, in real-world applications such as WLIS, pure LAV or GAV approaches might not be appropriate. The data collected by an agency may contain additional levels of classification to those already in the ontology, reflecting the primary area of interest for the agency. For example, a county in which agriculture is the main occupation may have more categories of agricultural land usage than the global schema drawn up for the state. Such differences in the resolution of the data can be handled by storing that information in the ontology. Such an approach departs from a pure LAV approach, as, upon request from a distributed source, the team of experts that create and maintain the ontology will now integrate information from that source into the ontology, so as to improve the resolution of the answer to the query. This is a situation where ontology merging can be used.

Mapping Types

In the examples of this paper, which are taken from WLIS, we represent the ontologies as trees. In the figures, the tree on the left represents the central ontology and the tree on the right represents the local ontology. The vertices of the trees correspond either to existing entities in the ontology (*real vertices*) or to entities created with the end of semantically grouping entities (*virtual vertices*). The former vertices are represented using a solid line and the latter are represented using a dashed line. Entities corresponding to virtual vertices do not explicitly appear in the underlying data instances and are only added to the local ontologies for the purpose of establishing mappings from the central ontology to the distributed source. For each distributed source, a local expert establishes these mappings.

In Figure 1, the codes *Agriculture–Woodlands–Forests* and *Agriculture–Woodlands–Non-forests* in the central ontology are respectively mapped to the land use codes *Forestry* and *Non-forest woodlands* in the local ontology (used by Dane County). There is no local land use code corresponding to *Agriculture–Woodlands*. To better align the local ontology with the central ontology, a virtual vertex is introduced corresponding to *Agriculture–Woodlands*.

The codes *Agriculture Woodlands–Forests*, *Agriculture Woodlands–Non-forests*, *Forestry* and *Non-forest*

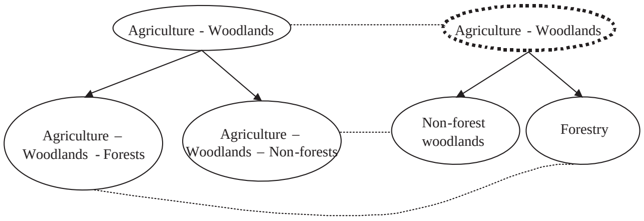


Figure 1: Real and virtual vertices.

woodlands are semantically at the same level of detail or *resolution* in the two ontologies. Similarly, the two vertices corresponding to *Agriculture–Woodlands* are also at the same level. We say that such entities are *aligned*. Initially, the information as to which entities in the different ontologies are aligned must be provided by the local expert. Once two entities are known to be aligned, the nature of the relation between them can be characterized using the following mapping types: *exact*, the connected vertices are semantically equivalent, *approximate*, the connected vertices are semantically approximate, *null*, the vertex in the central ontology does not have a semantically related vertex in the local ontology, *superset*, the vertex in the central ontology is semantically a superset of the vertex in the local ontology, and *subset*, the vertex in central the ontology is semantically a subset of the vertex in the local ontology. Technically, an *exact* mapping is equivalent to both a *subset* and a *superset* mapping. However, we will reserve the terms *subset* and *superset* for proper inclusion and containment, respectively.

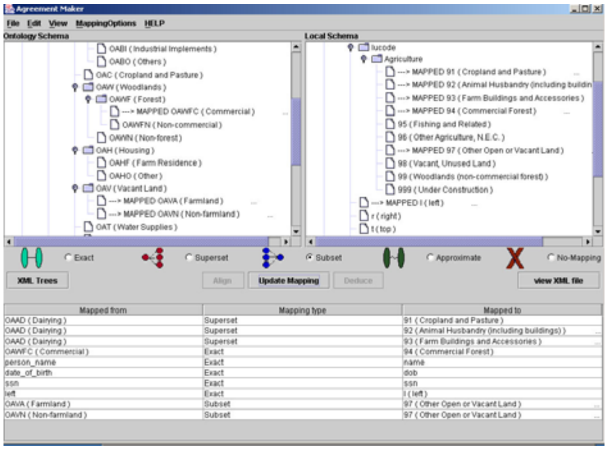
Difficulties in establishing the mappings occur in several circumstances. For example, the semantic equivalent of an entity in the central ontology could be distributed over several vertices or parts of a vertex in the local ontology and vice versa. Therefore, a mapping can establish the connection between vertices in their entirety or between parts of a vertex. Another difficulty is that entities in the central ontology may have no correspondence with an entity in the local ontology and conversely, entities in the local ontology may have no correspondence with entities in the central ontology.

Figure 2 illustrates several mappings between vertices in two ontologies for land use patterns. The vertices corresponding to *Industry, Mining* and *Manufacturing* in the central ontology can be mapped to those corresponding to *Industrial Sector, Mining* and *Mfg.* in the local ontology. In the central ontology, the vertex *Plastic wares* denotes entities that are made of plastic or glass. However, in the local ontology, there is a vertex *Plastics* and another vertex *Rubber and Glass*, which denotes manufactured objects made of rubber or glass.

The *Manufacturing* and the *Mfg.* vertices are aligned. Similarly, the two *Mining* vertices are also aligned. *Manufacturing* is semantically equivalent to *Mfg.*, as both denote a collection of industries producing plastics, glass, and rubber products. Hence, this mapping is of type *exact* as denoted in the mapping from the *Manufacturing* vertex to the *Mfg.* vertex. *Plastic wares* is semantically a *superset* of the *Plastics* vertex and *Rubber* is semantically a *subset* of the *Rubber and Glass* vertex.

Currently, the local expert establishes the mappings manually with the user interface shown in Figure 3. The central ontology and the local ontology are shown in the top left and right panes respectively. The current set of mappings is shown in the bottom pane and helps the user in keeping track of entities that have already been mapped and of those which are yet to be mapped. The mapping options are shown in the center of the application window and can be chosen while specifying the mappings.

The user selects an entity or a collection of entities in the central ontology, the equivalent entity or collection of entities in the local ontology, and one of the possible mapping options and then asks the system to update the mappings. As the user specifies mappings for the entities in the central ontology, the current set of mappings displayed in the bottom pane is updated to reflect the changes. The user can change the mapping of an entity in the ontology by selecting it and the equivalent entity or entities in the local ontology and then specifying the new mapping option. Once all the entities in the ontology have been mapped, the user asks the system to create the agreement file, which is an XML document that contains all the mappings (Cruz & Rajendran 2003).



Ontology Schema

OABI (Industrial Implements)
OABO (Others)
OAC (Cropland and Pasture)
OAW (Woodlands)
OAWF (Forest)
---> MAPPED OAWFC (Commercial)
OAWFN (Non-commercial)
OAWN (Non-forest)
OAH (Housing)
OAHF (Farm Residence)
OAHO (Other)
OAV (Vacant Land)
---> MAPPED OAVA (Farmland)
---> MAPPED OAVN (Non-farmland)
OAT (Water Supplies)

Local Schema

lucode
Agriculture
---> MAPPED 91 (Cropland and Pasture)
---> MAPPED 92 (Animal Husbandry (including buildin
---> MAPPED 93 (Farm Buildings and Accessories)
---> MAPPED 94 (Commercial Forest)
95 (Fishing and Related)
96 (Other Agriculture, N.E.C.)
---> MAPPED 97 (Other Open or Vacant Land)
98 (Vacant, Unused Land)
99 (Woodlands (non-commercial forest))
999 (Under Construction)
---> MAPPED l (left)
r (right)
t (top)

Mapped from	Mapping type	Mapped to
OAAD (Dairying)	Superset	91 (Cropland and Pasture)
OAAD (Dairying)	Superset	92 (Animal Husbandry (including buildings))
OAAD (Dairying)	Superset	93 (Farm Buildings and Accessories)
OAWFC (Commercial)	Exact	94 (Commercial Forest)
person_name	Exact	name
date_of_birth	Exact	dob
ssn	Exact	ssn
left	Exact	l (left)
OAVA (Farmland)	Subset	97 (Other Open or Vacant Land)
OAVN (Non-farmland)	Subset	97 (Other Open or Vacant Land)

Figure 3: User interface for establishing the mappings.

Semi-automatic Alignment

The mappings that we define can be integrated in a semi-automatic alignment methodology to simplify the task of aligning ontologies. The user initially identifies the hierarchy levels that are aligned in the two ontologies. Then the alignment component propagates. When ambiguities or inconsistencies are encountered, or the algorithm can not propagate values any further, those vertices are singled out. As in other approaches,

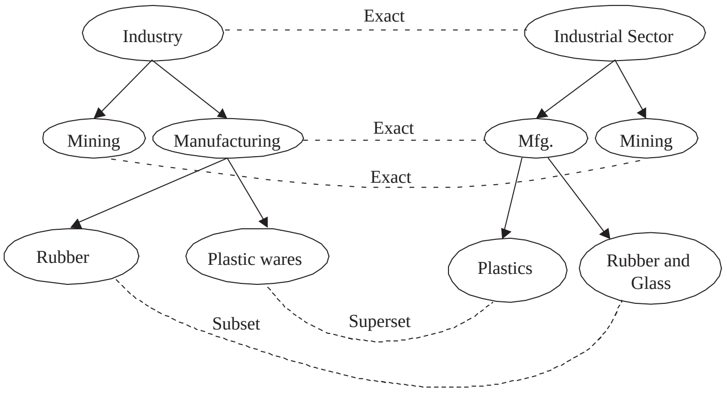


Figure 2: Mapping types.

the user can then manually assist the algorithm by mapping entities manually (McGuinness *et al.* 2000; Noy & Musen 2000).

In Figure 4, vertices *b* and *c* in the central ontology are mapped using mapping types *exact* and *superset* to vertices *e* and *f* in the local ontology. The mapping type between their parents *a* and *d* can be deduced to be *superset* based on the mapping between the children, because we consider that the semantic content of the parent is the generalization of the semantic contents of its children. For all the children of *d* there is a child of *a* that has been mapped to it. This is the *Fully Mapped (FM)* case.

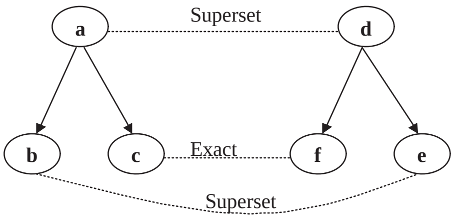


Figure 4: Fully mapped deduction operation.

The *Partially Mapped (PM)* case occurs if there are some children in the local ontology to which no children in the central ontology have been mapped. For example, in Figure 5, vertices *b* and *c* in the central ontology are mapped to vertices *e* and *f* using mapping type *exact*. But vertex *g* has no corresponding vertices in the central ontology. As a result, vertex *a* is mapped to vertex *d* using a mapping of type *subset*.

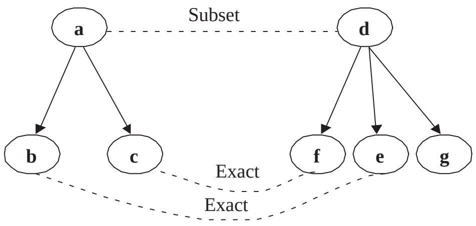


Figure 5: Partially mapped deduction operation.

Table 2 lists the different possible combinations of vertex mappings and the resulting mappings for their parents. The *Fully Mapped (FM)* and *Partially Mapped (PM)* cases are shown respectively in columns 3 and 4 of the table. A *User-defined* entry in the table (abbreviated to *User-def*) indicates that the parent's mapping type cannot be automatically deduced and the user has to provide the appropriate mapping type manually. These deduction operations can easily scale up to include the cases where a vertex has more than two children. They will be performed recursively, starting from the vertices that are aligned and traveling up the central ontological tree, to deduce the mapping types between the central ontology and the local ontologies. As previously mentioned, all combination results can be overridden by the user to accommodate intricate cases.

For example, Figure 2 illustrates a challenge to the full automation of the alignment process using the deduction operations. The mappings from the children of *Manufacturing* to the children of *Mfg.* are of types *subset* (*Rubber* entities are a subset of the entities that belong to *Rubber and Glass*) and *superset* (*Plastic wares* contain entities that are made of plastic or glass, therefore being a *superset* of the entities that belong to *Plastics*). Therefore, according to Table 2, user intervention is required. Ideally, one would want the propagation process to recognize that the union of the children of *Manufacturing* and the union of the children of *Mfg.* represent the same type of manufactured objects, thus leading to an *exact* mapping between those two entities. This example shows the need for extending our current framework.

Child 1	Child 2	FM	PM
Exact	Exact	Exact	Subset
Exact	Approx	Approx	Subset
Exact	Superset	Superset	User-def
Exact	Subset	Subset	Subset
Exact	Null	Superset	User-def
Approx	Approx	Approx	Subset
Approx	Superset	Superset	User-def
Approx	Subset	Subset	Subset
Approx	Null	Superset	User-def
Superset	Superset	Superset	User-def
Superset	Subset	User-def	User-def
Superset	Null	Superset	User-def
Subset	Subset	Subset	User-def
Subset	Null	User-def	User-def
Null	Null	User-def	User-def

Table 2: Automatic mapping deduction operations.

Ontology Merging

Each local ontology might have a different organization of the entities based on the primary function of the agency maintaining it. For example, a county in which agriculture is the main occupation may have more categories of agricultural land usage than the central ontology. When such a local ontology is aligned to the central ontology, there might be several places where the mapping type is *null*, whereas there are vertices in the local ontology to which no vertices in the central ontology have mapped. This can indicate that a particular criterion of classification is missing in the central ontology thus leading to loss of resolution of the data when local ontologies using that classification technique are aligned. In such cases, the expert in charge of maintaining the central ontology can add the missing classification. This can be viewed as merging entities from local ontologies into the central ontology.

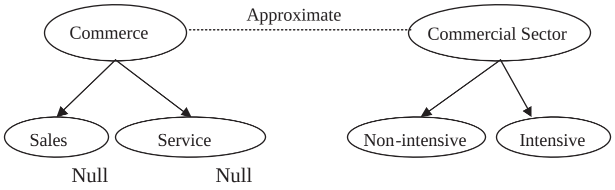


Figure 6: Ontology alignment before the deduction process.

For example, in the central ontology of Figure 6, commercial land usage is classified as *Sales* and *Service* (based on the primary function of the commercial establishment). In the local ontology, commercial land usage is sub-classified as *Commercial Intensive* and *Commercial non-intensive* (based on the size of the operations). The two parent vertices are considered aligned, because they have similar resolution. As shown in Figure 6, vertices *Sales* and *Service* cannot be mapped to any of the vertices in the local ontology and hence have their mapping type as *null*. Therefore, the mapping type between *Commerce* and *Commercial Sector* cannot be automatically deduced and is specified as *approximate* by the user.

The classification of commercial land usage, based on the scale of operations, is missing from the central ontology and could be introduced to better align local ontologies that use that classification scheme. The alignment of the ontologies after the additional level of classification is introduced is shown in Figure 7. Notice that here the mapping *superset* from *Commerce* to *Commercial Sector* was correctly deduced using the automatic method, therefore illustrating one of the possible ways in which our approach could be used to assist in the merging of ontologies or to measure the adequacy of the merging.

Future Work

Clearly, there are several issues that still need to be investigated, pertaining to the modeling and implementation of the alignment process. While our types of mappings appear to be "adequate" to express the mappings in our current application, we would like to characterize the notion of *adequacy* for such mappings, especially as integrated with the alignment process. Also, we have made some simplifying assumptions concerning the alignment process. For example, we did not take into account inversions in the order of the ontologies (called "bowties") (Hovy 1998).

Our assumptions may or may not hold in other applications or for other types of ontologies. An interesting question is whether our approach will work well for ontologies that are very dissimilar. Also, the fact that currently the global ontology is a tree without attributes, simplifies the implementation of the mappings. However, more powerful methods (albeit computationally less efficient), such as those used in (McGuinness *et al.* 2000), are currently not possible.

We have identified automatic mapping deduction operations but need to further explore an algorithm that incorporates such automatic steps and takes into consideration the intricacies we have exemplified. The integration of our mappings into the querying process, extending our previous results (Cruz *et al.* 2002; Cruz & Rajendran 2003) will need to be investigated.

Finally, our focus has been on integrating data from a single theme, specifically the theme of land use within the geospatial domain. Further research efforts will concentrate on providing mechanisms for integrating data from multiple themes and therefore using multiple central ontologies.

Acknowledgments

We would like to thank Nancy Wiegand and Steve Ventura, from the Land Information & Computer Graphics Facility at the University of Wisconsin-Madison, for the discussions on land use problems, and William Sunna at

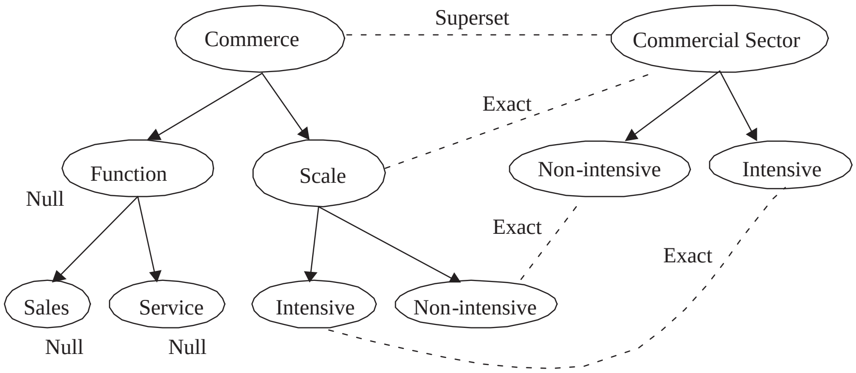


Figure 7: Ontology alignment after deduction.

UIC, for his help with the implementation of the alignment interface.

We are indebted to the referees for their insightful comments and suggestions.

References

Bergamaschi, S.; Guerra, F.; and Vincini, M. 2002. A Data Integration Framework for E-Commerce Product Classification. In *1st International Semantic Web Conference (ISWC)*, 379–393.

Bishr, Y. 1998. Overcoming the Semantic and Other Barriers to GIS Interoperability. *International Journal of Geographical Information Science* 12:299–314.

Corcho, O., and Gomez-Perez, A. 2001. Solving Integration Problems of E-Commerce Standards and Initiatives through Ontological Mappings. In *IJCAI'01 Workshop on Ontologies and Information Sharing*.

Cruz, I. F., and Rajendran, A. 2003. Semantic Data Integration in Hierarchical Domains. *IEEE Intelligent Systems* March-April:66–73.

Cruz, I. F.; Rajendran, A.; Sunna, W.; and Wiegand, N. 2002. Handling Semantic Heterogeneities using Declarative Agreements. In *International ACM GIS Symposium*, 168–174.

Fonseca, F., and Egenhofer, M. 1999. Ontology-driven Geographic Information Systems. In *7th ACM Symposium on Advances in Geographic Information Systems*, 14–19.

Gennari, P., and Musen, J. 1998. Mappings for Reuse in Knowledge-based Systems. In *11th Workshop on Knowledge Acquisition, Modelling and Management, KAW 98*.

Hovy, E. 1998. Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses. In *First International Conference on Languages Resources and Evaluation (LREC)*.

Kashyap, V., and Sheth, A. 1998. Semantic Heterogeneity in Global Information System: The Role of Metadata, Context and Ontologies. In Papazaglou, M., and Schlageter, G., eds., *Cooperative Information Systems: Current Trends and Directions*. Academic Press. 139–178.

Lenzerini, M. 2001. Data Integration is Harder than You Thought. In *9th International Conference on Cooperative Information Systems CoopIS*, 22–26. LNCS 2172, Springer Verlag.

Ludaescher, B.; Gupta, A.; and Martone, M. E. April 2001. Model-based mediator system with domain maps. In *17th Intl. Conference on Data Engineering (ICDE)*, 81–90.

McGuinness, D. L.; Fikes, R.; Rice, J.; and Wilder, S. 2000. An Environment for Merging and Testing Large Ontologies. In *Seventeenth International Conference on Principles of Knowledge Representation and Reasoning (KR-2000)*, 483–493.

Noy, N. F., and Musen, M. A. 2000. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. In *The Sixteenth National Conference on Artificial Intelligence (AAAI)*, 450–455.

Wiegand, N.; Patterson, D.; Zhou, N.; Ventura, S.; and Cruz, I. F. 2002. Querying Heterogeneous Land Use Data: Problems and Potential. In *National Conference for Digital Government Research*, 115–121.