Towards the Improvement of Alignment between Heterogeneous Domain Ontologies

Mariela Rico[1], Ma. Laura Caliusco[1],
Omar Chiotti[1,2], and Ma. Rosa Galli[1,2]

[1] CIDISI Research Center, UTN-FRSF, Lavaise 610, S3004EWB Santa Fe, Argentina
{mrico,mlcaliusco}@santafe-conicet.gov.ar
[2] INGAR-UTN-CONICET, Avellaneda 3657, S3002GJC Santa Fe, Argentina
{chiotti,mrgalli}@santafe-conicet.gov.ar
http://cidisi.frsf.utn.edu.ar/index.htm

Abstract. In open or evolving systems, different parties could adopt different ontologies. This fact, instead of reducing the semantic heterogeneity, moves the heterogeneity problems to a higher level. A plausible solution to this problem is the ontology-matching process, which aims at finding correspondences between semantically related entities of different ontologies. In this paper, an approach for improving this set of correspondences, called alignment, is proposed. This approach is based on the idea of making explicit features generally implicit in the ontologies.

1 Introduction

The evolution of Information Systems and Communication Technologies, particularly those related to the Internet, is leading the implementation of connections among heterogeneous information systems. In this scenario, large amount of data are available in different formats and platforms. Data repositories vary from structured database to unstructured or semi-structured files. The lack of agreement on data representation (syntax and semantics) across heterogeneous systems makes the interoperability problem very complex.

In order to solve this problem, standards and tools to facilitate information integration at both syntactic and semantic levels have been developed. As regards syntactic level, the eXtensible Markup Language (XML) has gain the attention as a standard for data syntactic representation. As regards semantic interoperability, ontologies have been proposed as an artifact to represent the data semantics. With the implementation of this new paradigm, the interoperability problem has moved to another level. In open or evolving systems, different parties could adopt different ontologies. Thus, merely using ontologies does not solve the interoperability problems: they are raised at a higher level.

Ontology-matching is a plausible solution to allow the interoperability between heterogeneous domain ontologies [6]. Ontology-matching aims at finding correspondences between semantically related entities of different ontologies. Ontology-matching results, called alignments, can express with various degrees

XXXIV Conferencia Latinoamericana de Informática (CLEI 2008), pp. 579-588, 2008.

of precision the relations between the ontologies under consideration [7]. Alignments can be used for various tasks such as ontology merging, query answering, or data translation. Particularly in the last years, the data translation process has gained the attention of many researchers.

The data translation process requires managing conversion rules to translate the instances of an ontology into the instances of another one. The conversion functions defined in the ECOIN project [8] are an example of these conversion rules. These conversion functions were defined to deal with a single ontology with multiple contexts which express the specific specializations of a shared domain model. As a consequence, some of the situations presented in ontology-matching between heterogeneous domain ontologies are managed, while others are not.

This paper assumes that the generation of conversion rules will be improved if the alignment between ontologies is improved. To this aim, an approach for improving the alignment between heterogeneous domain ontologies is presented. This approach is based on the idea of disaggregating the ontology entities by making the domain features affecting their semantics explicit. The remainder of the paper is organized as follows. Section 2 introduces the main concepts around the interoperability between heterogeneous ontologies. Section 3 presents the approach and preliminary results. Finally, Section 4 is devoted to the conclusions.

2 Interoperability between Heterogeneous Ontologies

In this section, an ontology definition is given and, in accordance with this definition, the ontology heterogeneity is characterized. Finally, a discussion about how to face the ontology heterogeneity problem is presented.

2.1 Characterizing Ontology Heterogeneity

A *domain ontology* is a representational artifact that represents the semantics of a given domain of discourse. A *domain* is a portion of the reality that forms the subject-matter of a single science or technology of mode of study [13]. The representational units of a domain ontology are: terms, relations between them, axioms and instances [2]. For simplicity, from now onwards the word *ontology* will be used to refer to the expression *domain ontology*. A *term* is a word or group of words representing an entity from the domain of discourse. These entities are first order entities in reality [13] such as organizations, persons, products, among others. *Relations* represent a type of association between entities of the domain. They can be divided into hierarchical (is-a), mereological (part-of), conceptual (e.g., synonym and antonym), and particular (defined by the ontology designer), among others [3]. *Axioms* serve to represent sentences that are always true in a domain. They are normally used to represent knowledge that cannot be formally defined by the other representational units. Finally, an *instance* is a certain individual of a corresponding entity in the domain of discourse. For example, a particular commercialized product is an instance of the entity product. A term representing an entity and its instances are related by the association instance-of.

In literature, different classifications of ontology heterogeneity conflicts can be found [7, 5, 1, 9]. In this paper, the ontology heterogeneity conflicts are classified in four categories considering the previous ontology definition:

- *Instance conflicts* are discrepancies in the representation or interpretation of instantiated data values, which can differ in their measurement unit, precision and spelling.
- *Terminological conflicts* are differences in names due to alternatives to depict the same reality such as using distinct terms for the same entities of reality. These can be caused by the use of different natural languages.
- *Data versus metadata conflicts* are disagreements about what is data and metadata; e.g., an instance of an ontology can be represented as a term in another ontology.
- *Structural conflicts* result from the use of different structures for representing the same entity of reality.

2.2 Addressing Ontology Interoperability

In open or evolving systems such as the Semantic Web, different parties could, in general, adopt different ontologies. Thus, merely using ontologies does not reduce heterogeneity: it raises heterogeneity problems at a higher level [7].

Ontology-matching is a plausible solution to the semantic heterogeneity problem [7]. Ontology-matching aims at finding correspondences between semantically related entities of different ontologies. These correspondences may stand for equivalence as well as other relations, such as consequence, subsumption, or disjointness, between ontology entities. Ontology-matching results, called alignments, can thus express with various degrees of precision the relations between the ontologies under consideration [7].

There are different algorithms for implementing the match process, which can be classified in schema-based and instance-based matching. A schema-based matcher considers the structural conflicts and uses some similarity measure to determine correspondence. An example of these matchers is the H-Match algorithm, which dynamically performs ontology matching at different levels of depth [4]. On the other hand, an instance-based matcher considers the instances conflicts and terminological conflicts. An example of these matchers is the FCA-Merge [14]. These matchers can also be combined like in the IF-Map algorithm [12]. A complete classification of these algorithms can be found in [7].

3 Improving Ontology Alignment

3.1 Ontology Alignment in Data Translation Process

A data integration process is intended to integrate information coming from multiple sources without using a central entity [11]. Each source may use different terminologies and metadata model to represent their data. In order to establish an exchange of (meaningful) information between sources, the following steps have to be followed: (1) Identify, characterize, and establish correspondences between the entities of reality as represented in ontologies; (2) Define the conversion rules for translating the instances of one ontology into the instances of another one and, (3) Execute conversion rules. The first step is known as matching process [7] and the others are the typical activities of a data translation process. Fig. 1 outlines the data integration process: conversion rules (CR) are generated from the alignment (A), which is the result of the matching process.

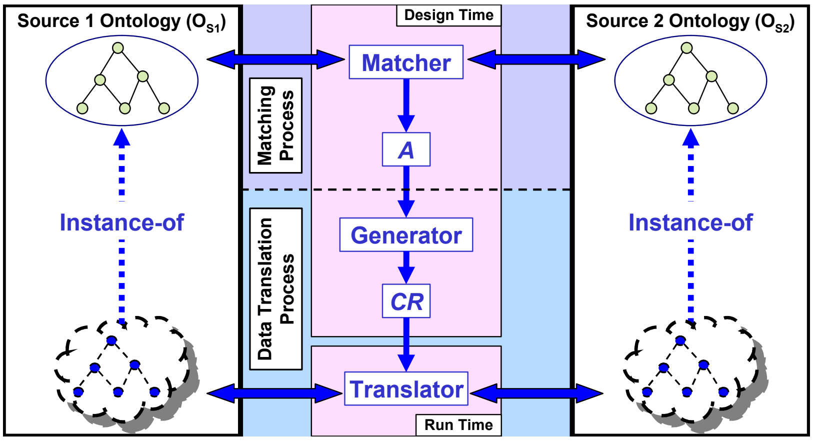

Fig. 1. Data translation scenario with matching.

The alignment is a set of correspondences between two or more heterogeneous ontologies. The improvement of this alignment is crucial to facilitate the generation of conversion rules. To this aim, it is proposed to improve the input to the matching process.

3.2 Exploiting Domain Contextual Features for Alignment

In order to provide better input to the matching process, it would be convenient representing the entities of reality, whose instances must be translated, with the necessary degree of detail. To this aim, it is proposed to make implicit features explicit in the ontologies. These features, called the context, are those that characterize a domain and whose meaning is dependent on that domain. The ontologies that represent the semantics of the sources are called *original ontologies*, while the result of making the context explicit is called *extended ontology*. In extended ontologies, representational units can be divided in two categories: representational units of content and representational units of context. The first ones are those that already existed in the original ontologies, whereas the last ones are those that are added in the extended ontologies.

With the aim of showing the approach, the integration of data between two data sources is considered. The Source 1 contains the information of the products manufactured by an enterprise, and the Source 2 contains the information of the products that are sold by a retailer. Both enterprises interchange information between them. In order to avoid misunderstanding, the data sources are semantically described by the ontologies shown in Fig. 2.

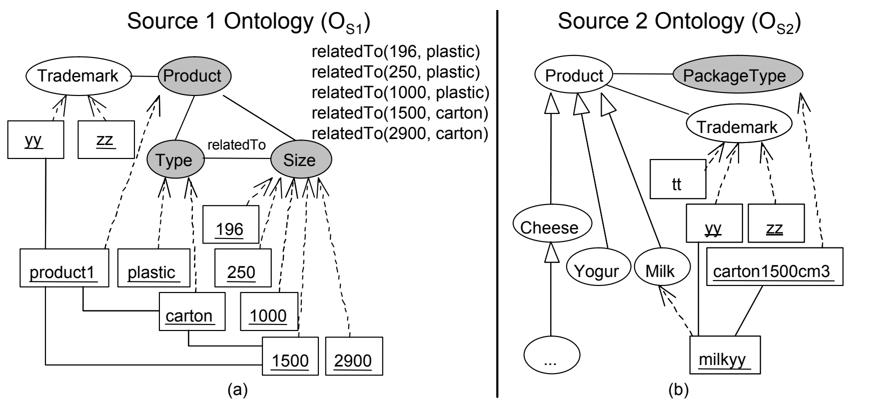

Fig. 2. Fragments of two ontologies to be matched.

The entity of reality "package" represented by PackageType term in O_{S2} (Fig. 2.b) has implicit an association with a representation dimension. In this case, the dimension is not metric, but an enumeration of possible values. In order to make this dimension explicit, a term representing it has to be added to the original ontology. Fig. 3 shows how the portion of interest of the extended ontology would be like (shaded terms belong to the set of representational units of context). Additionally, this figure shows how the carton1500cm3 instance would be represented. As can be seen, the carton1500cm3 instance (Fig. 2.b) has been divided in two parts in the extended ontology, packagetype1 instanceOf PackageType and carton1500cm3 instanceOf PackageDimension; these two new instances are related by the associatedWith relation.

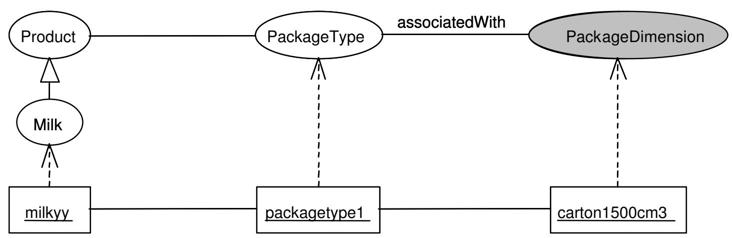

Fig. 3. A portion of an extended O_{S2}, and the carton1500cm3 instance representation.

Similarly, the entity of reality "package" represented by the Product, Type and Size terms plus their relations in O_{S1} (Fig. 2.a) has implicit an association with a set of representation dimensions, called multi-dimension. These dimensions are qualities that cannot be assigned a value on one dimension without giving it a value on the other; they are called integral dimensions in the sense of [10]. In Fig. 4, TypeDimension and SizeDimension terms represent integral dimensions that are used to define the Product Multi-Dimension, represented by the term ProductMultiDimension. In its definition, this Product Multi-Dimension has a set of rules constraining the relations between its constituting dimensions showed by a note related to the ProductMultiDimension term.

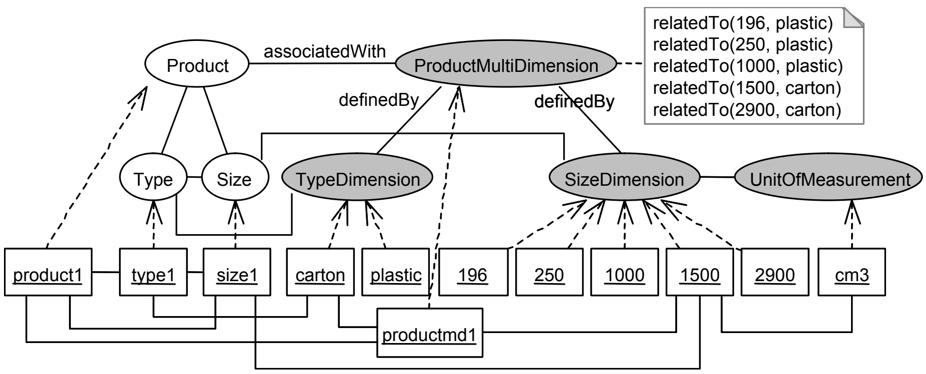

Fig. 4. A portion of an extended O_{S1}, and the equivalent of the carton1500cm3 instance representation

The Type Dimension is an enumeration of possible values (such as "carton" and "plastic"), and the Size Dimension is metric, i.e. its possible values are in the set of nonnegative numbers (196, 250, etc.). Additionally, Size Dimension has an associated metric unit, represented by the term UnitOfMeasurement, indicating in this way that all the possible values of Size Dimension are expressed in that metric unit, cm^3 in this case. The Type term in O_{S1} is implicitly associated with the Type Dimension, and the Size term in O_{S1} is implicitly associated with the Size Dimension. The lower part of Fig. 4 shows how the equivalent of the carton1500cm3 instance would be represented in the extended O_{S1} by defining the instance productmd1. This instance is defined by two instances of integral dimensions: carton and 1500. carton is an instance of the TypeDimension term and is the value of the type1 instance in the manufacturing enterprise domain, i.e. the value of the type of product. 1500 is an instance of the SizeDimension term and represents the value of the size1 instance in the the manufacturing enterprise domain, i.e. the value of the size of product. Finally, cm3 indicates that 1500 are cm3, i.e. the value of the size of product is expressed in cm^3.

The features inherent to entities of reality, whose representation and interpretation depends on the domain in which they are considered, can be observed in Fig. 3 and 4. In the manufacturing enterprise domain, the entity of reality "package" requires a representation so simple that it can be represented by a dimension, whereas in the domain of the retailer it requires of a set of dimensions.

In order to facilitate the translation process, it is proposed to represent the entities of reality, whose instances must be translated, with the necessary degree of detail. However, this does not guarantee that the translation of instances of an entity of reality from a domain to another one can be done by itself. The translation process needs to be able to identify correctly the set of representational units that represents an entity of reality and its semantics in a given domain. For example, in the extended O_{S2} (Fig. 3) the absence of a representational unit that designates the semantics of the entity of reality represented by the PackageType term can be noticed. The same happens in the extended O_{S1} (Fig. 4) with the Product, Type, and Size terms plus their relations. The missing representational unit in each of these figures is the term Package. Fig. 5 shows the two resulting extended ontologies, where the term ProductMultiDimension from the extended O_{S1} was renamed PackageMultiDimension with the aim of properly refer to the semantics of the entity of reality.

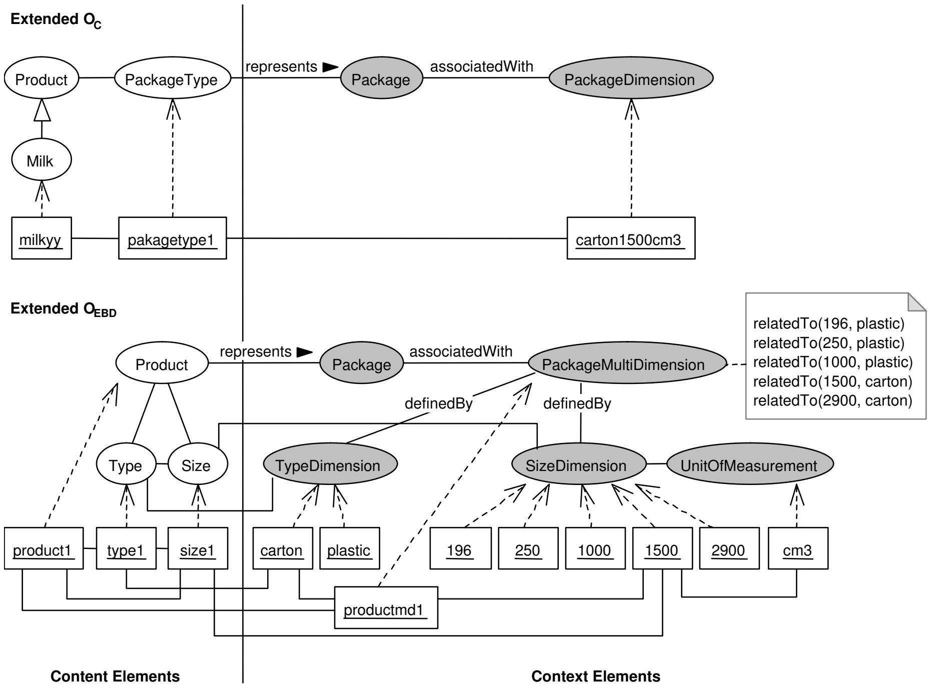

Fig. 5. Extended O_{S2} and extended O_{S1} with instances.

The Package term refers to the semantics of an entity of reality, whose representation and interpretation depend on the domain in which it is considered. In the domain of the manufacturing enterprise, the Package term is associated with a dimension, whereas in the domain of the retailer it is associated with a multi-dimension. Thus, Package term represents a common feature to both domains, whose instances of ontologies have to be translated to each other. This kind of terms will be used to facilitate the generation of the conversion rules which will be automatically executed by the translation process to allow semantic interoperability between both enterprises. Additionally, considering that the conversion rules are generated from an alignment, entities of reality represented in this way are a step toward the complete identification of the elements to translate and its features by the matching process.

3.3 Experimental Results and Discussion

In order to evaluate the effects on the alignment because of a better input to the matching process, the HMatch Protégé Plugin 1.5[3] was used [4]. This plugin implements the H-Match algorithm that provides a ranking of similarity between the concepts of two ontologies considering linguistic and contextual features. The first ones refer to the names of ontology elements and their meaning. The second ones refer to the properties and concepts directly related to a given concept.

Fig. 6 shows the alignment (A_1) between the O_{S1} and O_{S2} ontologies ("Source concepts" column and "Comparison concepts" column respectively), and the measure of semantic affinity of two concepts in the range [0, 1] ("Matching value"). In order to perform the matching, the deep matching model was used, which considers concept names and the whole context of concepts. The minimum level of semantic affinity required to consider two concepts as matching concepts was set at 0.6, and the one-to-many (1:n) strategy was used for defining a set of mappings for each concept of O_{S1}. In addition, it was set that the impact of the linguistic affinity is equal to the impact of the contextual affinity.

Source concepts	Comparison concepts	Matching value
Product	Product	0.815
Trademark	Trademark	1.0
Type	PackageType	0.8

Fig. 6. Matching results between the O_{S1} and O_{S2} ontologies with H-Match.

The entity of reality "package" was represented by the Product, Type and Size terms plus their relations in O_{S1} (Fig. 2.a), and by the PackageType term in O_{S2} (Fig. 2.b). The expected correspondences were:

$$\begin{array}{ll} \text{Product} = \text{PackageType}, & \text{Type} = \text{PackageType} \qquad (E_1) \\ \text{Size} = \text{PackageType}, & \text{Trademark} = \text{Trademark} \end{array}$$

[3] `http://islab.dico.unimi.it/hmatch/downloads.php?cat_id=1`

The alignment (A_2) between the extended ontologies (Fig. 5) is shown in Fig. 7. These results correspond to the same option set for obtaining A_1.

Source concepts	Comparison concepts	Matching value
Package	Package	0.71033
PackageMultiDimension	PackageDimension	0.64
Product	Product	0.61571
SizeDimension	PackageDimension	0.64
Trademark	Trademark	1.0
Type	PackageType	0.8
TypeDimension	PackageDimension	0.64

Fig. 7. Matching results between the extended O_{S1} and the extended O_{S2} ontologies.

In this case, the entity of reality "package" was represented by the PackageMultiDimension, TypeDimension, and SizeDimension terms plus their relations in O_{S1} (Fig. 5), and by the PackageDimension in O_{S2} (Fig. 5). The expected alignment was:

Package = Package,	Product = PackageType	
PackageMultiDimension = PackageDimension,	Type = PackageType	(E_2)
TypeDimension = PackageDimension,	Size = PackageType	
SizeDimension = PackageDimension,	Trademark = Trademark	

Evaluation Since *precision* (P) and *recall* (R) are well understood and widely accepted, they were used to evaluate the results previously obtained. Precision and recall are based on the comparison of the resulting alignment A with the expected alignment E, effectively comparing which correspondences are discovered and which are not. Precision measures the ratio of correctly found correspondences (true positives - $E \cap A$) over the total number of returned correspondences (A). Recall measures the ratio of correctly found correspondences over the total number of expected correspondences (E). The values obtained for these measurements are:

$$P(A_1, E_1) = E_1 \cap A_1/A_1 = 2/3 = 0,66 \quad R(A_1, E_1) = E_1 \cap A_1/E_1 = 2/4 = 0,5$$
$$P(A_1, E_2) = E_2 \cap A_1/A_1 = 6/7 = 0,85 \quad R(A_1, E_2) = E_2 \cap A_1/E_2 = 6/8 = 0,75$$

Judging by these preliminaries results, it can be inferred that both criteria are better when the matching process is applied to the extended ontologies.

4 Conclusions

In order to improve the alignment between heterogeneous domain ontologies, the development of extended ontologies as a previous step to the matching process execution was proposed.

Despite the disadvantage of the required time to develop the extended ontologies, this approach improves significantly the output of the matching process, and facilitates the generation of conversion rules that are the core of the data translation process.

Acknowledgments. The authors are grateful to "Universidad Tecnológica Nacional (UTN)", "Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)", and "Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT)" for their financial support.

References

1. Bouquet, P., Ehrig, M., Euzenat, J., Franconi, E., Hitzler, P., Krtzsch, M., Serafini, L., Stamou, G., Sure, Y., Tessaris, S.: Specification of a Common Framework for Characterizing Alignment. Deliverable D2.2.1. Knowledge Web NoE. (2004)
2. Brusa, G., Caliusco, M.L., Chiotti, O.: Towards Ontological Engineering: A Process for Building a Domain Ontology from Scratch. Expert Systems - The Journal of Knowledge Engineering. In Press.
3. Caliusco, M.L.: A Semantic Definition Support of Electronic Business Documents in e-Colaboration. PhD Thesis, Universidad Tecnolgica Nacional, S.Fe, AR (2005)
4. Castano, S., Ferrara, A., Montanelli, S., Racca, G.: Semantic Information Interoperability in Open Networked Systems. In: Int. Conference on Semantics of a Networked World (ICSNW), pp. 215–230. Paris, France (2004)
5. Corcho, O.: A Declarative Approach to Ontology Translation with Knowledge Preservation. PhD Thesis, Universidad Politcnica de Madrid, Madrid, ES (2004)
6. Davies, J., Studer, R., Warren, P.: Semantic Web Technologies: Trends and Research in Ontology-Based Systems. John Wiley, London (2007)
7. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, London (2007)
8. Firat, A., Madnick, S., Manola, F.: Multi-dimensional Ontology Views via Contexts in the ECOIN Semantic Interoperability Framework. In: Shvaiko, P., Euzenat, J., Leger, A., McGuinness, D.L., Wache, H. (eds.) Contexts and Ontologies: Theory, Practice and Applications: Papers from the 2005 AAAI Workshop. pp. 1–8. AAAI Press (2005)
9. Ghidini, C., Giunchiglia, F.: A Semantics for Abstraction. In: 15th European Conference on Artificial Intelligence (ECAI), pp. 343–347. Valencia, ES (2004)
10. Guizzardi, G.: Ontological foundations for structural conceptual models. PhD Thesis. Telematica Instituut Fundamental Research Series, vol. 015. Enschede (the Netherlands): Telematica Instituut (2005)
11. Halevy, A., Ashish, N., Bitton, D., Carey, M., Draper, D., Pollock, J., Rosenthal, A., Sikka, V.: Enterprise Information Integration: Successes, Challenge and Controversies. In: The ACM SIGMOD Conference. Baltimore MD, USA (2005)
12. Kalfaglou, Y., Schorlemmer, M.: IF-Map: an Ontology Mapping Method Based on Information Flow Theory. Journal on Data Semantics. 98–127 (2003)
13. Smith, B., Kusnierczyk, W., Schober, D., Ceusters, W.: Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain. In: KR-MED 2006, Biomedical Ontology in Action. Baltimore MD, USA (2006)
14. Stumme, G., Medche, A.: FCA-Merge: Bottom-up Merging of Ontologies. In: 17th Int. Joint Conference on Artificial Intelligence, pp 225–234. Seattle, USA (2001)