Semantic Association of Taxonomy-based Standards Using Ontology

Hung-Ju Chu, Randy Y. C. Chow, Su-Shing Chen
Computer and Information Science and Engineering, University of Florida
Gainesville, FL, U.S.A.
{hchu, chow, suchen}@cise.ufl.edu
Raja R.A. Issa, Ivan Mutis
Rinker School of Building Construction, University of Florida.
Gainesville, FL, U.S.A.
{raymond-issa, imutis}@ufl.edu

ABSTRACT

The vision of *semantic interoperability*, the fluid sharing of digitalized knowledge, has led much research on *ontology/schema mapping/aligning*. Although this line of research is fundamental and has brought valuable contributions to this endeavor, it does not represent a solution to the challenge, *semantic heterogeneity*, since the performance of proposed approaches significantly relies on the degree of uniformity, formalization and sufficiency of data representations but most of today's independently developed information systems seldom have common knowledge modeling frameworks and their data are often not formally and adequately specified. Consequently, a workable solution usually requires interventions of domain experts.

In human society, *hierarchically structured standards* (or *taxonomies*) for characterizing complex application processes and objects used in the processes are often used as a common and effective way to achieve some *semantic agreements* among stakeholders within a domain. This research hypothesizes that the establishment and the use of such standards can serve as a framework that can effectively facilitate the reconciliation of *semantic heterogeneity* in complex application domains. However, the reality shows that a comprehensive priori consensus is extremely difficult, if not impossible, to reach. Consequently, various complementary and competing standards are often created and their constant-changing nature yields another level of challenge in achieving the hypothesis.

This paper focuses on the development of methodology for bridging *complementary* standards within an application domain. It exemplifies such standards in building construction industry where interoperability problems are prevalent and human interactions are commonplace. It proposes a semi-automatic approach for *semantically associating* the standards to reduce costly human intervention in a workflow. The approach formalizes standards by using ontology and discovers their affinity (to what degree they are related with respect to their usage) from automated project document processing and semi-automatic domain expert inputs. A high-level architecture of an integration framework in web environment is suggested for depicting the role of the semantic association approach in the system.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing - *Indexing methods, Linguistic process*; I.2.4 [Artificial Intelligence]: I.2.1 Applications and Expert Systems - *Industrial automation*; I.2.4 Knowledge Representation Formalisms and Methods; I.2.6 [Artificial Intelligence]: Learning - Knowledge Acquisition

Keywords
taxonomy and standards, semantic interoperability, ontology-based knowledge extraction, semantic mapping.

1. INTRODUCTION

The vision of *semantic interoperability*, the fluid sharing of digitalized knowledge, has led much research on *ontology* (formal specification of conceptualization) and its languages, such as Web Ontology Language (OWL) [8]. The language provides primitives for specifying concepts, properties, explicit semantic relationships, and logical constraints on those objects. However, it does not address the issue of *semantic heterogeneity* between two independently developed ontologies. For example, a program that reads an ontology in OWL does not understand another ontology in the same language unless there is an explicit mapping between them. This difficulty has led much research on *ontology/schema mapping/alignment* [4], [5], [6], [11], [12], [13], and [14] and various *matching* technologies have been developed based on the attributes of objects and their associated data. Although this line of research is fundamental and has brought valuable contributions to this endeavor, it does not represent a solution to the challenge as we see. The performance of proposed approaches significantly relies on the degree of uniformity, formalization and sufficiency of data representations. Unfortunately, the concept of unified, formal, and sufficient specification is often an after-thought and most of today's independently developed

information systems seldom have common knowledge modeling frameworks and their data are often not formally and adequately specified. Consequently a workable solution usually requires interventions of domain experts.

In human society, *hierarchically structured standards* (or *taxonomies*) for characterizing complex application processes and objects used in the processes are often used as a common and effective way to achieve some *semantic agreements* among stakeholders within a domain. This research hypothesizes that the establishment and the use of such standards can serve as a framework that can effectively facilitate the reconciliation of *semantic heterogeneity* in complex application domains. However, the reality shows that a comprehensive priori consensus is extremely difficult, if not impossible, to reach. Consequently, various complementary and competing standards are often created and their constant-changing nature yields another level of challenge in achieving the hypothesis.

This paper focuses on the development of methodology for bridging complementary standards within an application domain. We have chosen a target application in the building construction domain, where interoperability problems are prevalent and human interactions are commonplace. In that domain, a variety of taxonomy-based standards have been established but still lack a uniform and systematic way for supporting efficient collaboration among project participants using different standards. This problem is further compounded by the complexity and the dynamics of business applications, which often require changes of the well-known standards. The interoperability cost in such environment is tremendous. For example, based on a recent National Institute of Standards and Technology (NIST) report [3], a conservative figure of $15.8 billion was determined to be annual costs due to a lack of interoperability in the capital facilities industry in 2002.

Two mainstream complementary standards, MasterFormat and UniformatII, in that domain are considered in our research. MasterFormat [1] is a specification standard established by the Construction Specification Institute (CSI) for most nonresidential building construction projects in North America. UniformatII is a newer American Society of Testing and Materials (ASTM) standard aiming at providing a consistent reference for the description, economic analysis, and management of buildings during all phases of their life cycles [2]. These standards were created by different stakeholders with different perspectives for different purposes. For instance, an architect is interested in the design and structure of a building, a contractor wants to know what materials are used and how much they cost, and a building inspector is concerned about building code compliance issues. MasterFormat classifies items primarily based on the specification of products and materials used in construction, so it is based on a conceptual view of a contractor. Complementarily, the taxonomical classification in Uniformat II is primarily based on the attributes and location of structural building components, such as foundations and exterior walls, which reflects the architect's view of a construction project. Although their views are different but both address the same building object. In other words, the taxonomies of the standards classify the same set of objects but on different attributes. From here one can easily infer that cross-referencing or document conversion between the standards is inevitable for interaction among project participants in applications such as cost estimation and code compliance checking. For example, a wall (interial or exterial) in UniformatII needs to be associated with the material (metal, wood or fiberglass) in MaterFormat and conformed to its intended usage (hurricane or fire proof) according to building code regulations (standards yet to be formalized by the industry). In general, UniformatII by design is more suitable as a participant communication/interaction framework than MasterFormat during the earlier phases of the life cycle. On the other hand, Masterformat has been used for years and has gained the majority of the construction industrial support for specifying detailed project documents. To facilitate more efficient collaboration among project participants, it is a common practice to supplement UniformatII with Preliminary Project Descriptions (PPDs) or schematic design in earlier phases, and convert them to construction documents in Masterformat during later phases. In addition, the conversion is also necessary for cost calculation since most databases of building materials suppliers are based on MasterFormat. It is desirable to transform pre-bid elemental estimates to MasterFormat, and from there to the trade costs of the project [2]. This process is often tedious and requires cross-area knowledge. Currently, it is done manually by domain experts and it is considered a major cause that hampers interoperability in the construction domain. Bridging the two standards is a key enabler for enhancing the interoperability.

Directly matching approaches based on attributes of the entities of the standards are expected to be inefficient due to the heterogeneous nature of complementary standards. This paper proposes a practical compromise by redefining the notion of mapping with a semi-automatic semantic extraction framework to assist domain experts in achieving interoperability. The mapping is termed as semantic association for relating elements between standards, and is dependent on the intended use such as cross-referencing of elements or specification semantic mapping. The semantic relationship can be characterized in two measurements: similarity (how closely objects resemble each other in their representation) and affinity (to what degree they are coupled in their usage). In some sense similarity is more static while affinity is more dynamic and general. For example, a bicycle is similar to a car due to their physical structures and properties. However, gasoline is more affinitive to a car although they do not resemble each other. Exploiting affinity in addi-

tion to similarity through semantic association is the focus of this research.

The approach consists of three components: formalization of taxonomies, ontology-based semantic extraction and measurement of affinity. The first component is a simple and yet novel approach for annotating a standard in primitive descriptive statements constructed by a set of necessary and sufficient orthogonal relations. They are then normalized and generalized into ontology. The second component shows how the ontology can be used for the extraction of relevant information from the instances in other standards for semantic association. The third component quantifies the affinity for ranking the extracted metadata to identify optimal association. The following sections detail the three components and outline an overall architecture of an integration framework depicting the relationship between the proposed approach and other related technologies and systems.

2. FORMALIZATION OF TAXONOMY

Taxonomies are initially designed for human consumption therefore some domain knowledge that is obvious and assumed by stakeholders is often omitted in their specifications. Moreover, taxonomies classifying large and complex items usually have the following characteristics:

1. The entities being classified and the attributes upon which the classification is based, are themselves complex concepts.
2. Multiple attributes (different concepts) might be used to classify entities at the same level.
3. Attributes are not orthogonal and might result in overlapping concepts in low-level entities (an object can fit into multiple categories).

There is a need for a systematic approach for annotating assumed semantics, clarifying complex concepts, and transforming them into formal representation before taxonomies can be effectively used for semantic association.

Semantic depends on context and context depends on applications. In other words, the semantic of a standard is open depending on how they are used. To avoid a standard being bound to specific applications, the intrinsic semantic of a standard without context should include the following:

1. the attributes being used for classification under the general perception in the application domain and
2. the entities under the inheritance of the taxonomy and the attributes.

To model the intrinsic semantics, ontology is considered in this research. The following subsection describes a systematic approach for transforming taxonomy into ontology.

Ontology Development from Taxonomy

The term, ontology, has been widely used in several disciplines, such as philosophy, epistemology, and computer science. There is much confusion in its definition. For example, in philosophy it refers to the subject of existence while in epistemology it is about knowledge and knowing. In computer science, many people use Gruber's definition [10] – an explicit specification of a conceptualization. In the context of our research, we interpret it as a description of the concepts/terms and relationships that can exist in an application domain. Centered on terms and relations, the transformation of taxonomy into ontology is described in the following steps.

Step 1: relation set identification

The goal of this step is to identify a sufficient and necessary set of orthogonal relations for a given taxonomy/standard so that assumed domain knowledge and complex concepts can be formally specified. This step should be manually done by standard committees who know best about the original intended use of the standards. The set should be constructed from two types of relations: primitive and derived.

Primitive relations are those that are unambiguously understood by the general public and the relationship between concepts connected by them does not change over time. Moreover, they reflect the intrinsic properties of objects or describe time and space and the intention of users when the objects are used. In addition, their definitions should include set relationship, such as instance-instance, instance-class, and class-class, to avoid ambiguity. For example, *part_of* is ambiguous since it could mean a subcomponent of an object or the membership of an object in a class. Its meaning can be identified as the first explanation if instance-instance is specified.

Derived relations are those that can be composed/modeled from primitive relations.

To elaborate this step, a small portion of the top three levels in MasterFormat taxonomy, Division 5 (D5) Metals and Division 6 (D6) Wood and Plastic rooted from Material, is exemplified as follows:

Division 5- Metals

- 05100 Structural Metal Framing
  - 05120 Structural steel
  - 05140 Structural aluminum
  - 05160 Metal framing systems
- 05400 Cold formed metal framing
  - 05410 Load bearing metal studs
  - 05420 Cold formed metal joists
  - 05430 Slotted channel framing

Division 6 - Wood and Plastics

- 06100 Rough carpentry

06110 Wood framing

06400 Architectural woodwork

06460 Wood frames

The following relations are identified for formalizing the above example:

1. *used_for* (class-class, human intention): purpose
2. *kind_of* (class-class, intrinsic): containment relation of attributes of instances.
3. *instance_of* (instance-class, intrinsic): membership
4. *made_of* (class-class, intrinsic): material component

Table 1 shows the mathematical properties of these relations that are used in the subsequent step for data normalization. They are also used for reasoning in knowledge extraction.

Table 1. Mathematical Properties of the relations

Relations	Transitive	reflexive	antisymmetric
used_for	-	-	-
kind_of	+	+	+
instance_of	+	+	+
made_of	+	-	-

Step 2: relation statements construction

This step is to construct simple statements using the relations defined in step one and all keywords in the taxonomy. The statements are then processed in subsequent steps for constructing ontology. There are two advantages using this bottom-up approach for formalizing taxonomies. One is that it can better address the dynamic nature of standards by enabling incremental updates and modifications of the statements and their resulting ontology. The other advantage is that domain experts who are not familiar with ontology can directly express their knowledge in the simple statements without communication overhead with knowledge modeling experts.

The following are examples of relation statements that partially describe the example shown in previous step.

1. Metals (D5), Wood (D6), Plastics (D6_1) are *instance_of* Material (root) → (D5_root, D6_root, D6_1_root)
2. Metals (D5) are *used_for* framing → 05100_1
3. Structural is a *kind_of* "metal framing" (05100_1) → 05100
4. Cold formed is a *kind_of* "metal framing" (05100_1) → 05400
5. Studes are *made_of* Metals (D5) → (05410_1)
6. "Load bearing metal studs" are *kind_of* Metal studs (05410_1) → 05410
7. 05410 is *used_for* 05400 → (05400_05410)

Note that each statement is given a unique identifier (following →) derived from the original identifier of a taxonomy entity.

Step 3: normalization

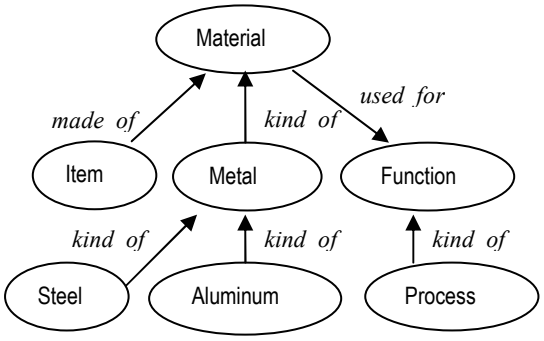
It is likely that redundant or conflict statements are generated along the way when domain experts annotate their taxonomies in the above steps. Based on the mathematical properties of the relations, this step normalizes the statements by:

1. redundancy elimination (removing same or equivalent statements)
2. conflict detection (for example: A-r1-B, and B-r1-A statements are conflict if r1 has asymmetric property)
3. implication detection (for example, A-r1-B, and B-r1 C statements imply A-r1-C through transitive property).

Step 4: semi-automatic generalization

This step is to generalize the resulting statements from step 3 into higher-level concepts connected by the same set of relations. Human being intervention is required in this step due to the complexity of the process. For example, if there exist A-r1-C, A-r1-D, B-r1-C, and B-r1-D, they can be generalized to concept1{A,B}-r1-concept2{C,D} by union. However, it becomes difficult when the above example is extended to include concept1{A,B}-r1-E and concept2{C,D}-r2-F. One cannot conclude concept1{A,B}-r1-concept2{C,D,E} unless an exception indicating no E-r2-F is added. Alternatively, it can be generalized to concept1{A,B}-r1-concept3{E,concept2{C,D}}. The system interacts with users by prompting the dilemmas for resolutions along the process of a whole taxonomy.

Figure 1 shown below depicts the generalized view or ontology of the relation statements shown in previous steps.



{metals, wood, plastics ..} are *instance_of* Material

{stud, joist ..}are *instance_of* Item

{framing, ..}are *instance_of* Function

{cold formed, structural ..} are *instance_of* Process

Figure 1. Ontology Example

4. ONTOLOGY-BASED SEMANTIC EXTRACTION

The task of the previous module, standard formalization, is usually a one-time effort (though it is an iterative process) and it needs significant domain experts' involvement. This module is different in that it is used in every workflow/task and extracted semantics can be accumulated in repository and used for improving future semantic association performance. Also, it can be relatively automated by using general linguistic processing technologies.

Standards, such as UniformatII and MasterFormat, addressed in this paper are functionally complementary to each other in an application domain and they are costly cross-referenced by domain experts in workflows due to their complexity (vast many-to-many mappings). This module basically is to automat the process by mimicking a domain expert doing cross-referencing from the context of a standard-compliant project specification, a script representation indexed of the standard, which defines intentionality. For example, the following text is quoted from a PPD [7] under entity B2010 in UniformatII taxonomy:

B SHELL

B20 EXTERIOR CLOSURE

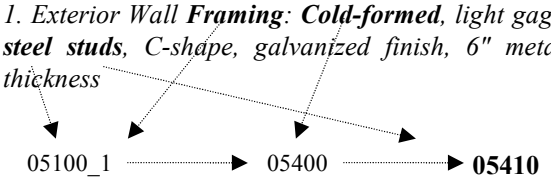
B2010 EXTERIOR WALLS

1. *Exterior Wall **Framing**: **Cold-formed**, light gage **steel studs**, C-shape, galvanized finish, 6" metal thickness as designed by manufacturer according to American Iron and Steel Institute (AISI) Specification for the Design of Cold Formed Steel Structural Members, for L/240 deflection. Downside: specifications often contain note-style sentences.*

Supposedly, the PPD is written by an architect and a contractor wants to estimate cost for exterior walls. He might comprehend that the wall framing will be made of cold-formed steel studs (semantic). Based on his expertise, he identifies that its corresponding entity in Masterformat is 05410 Load bearing metal studs (association). The following paragraph shows how the ontology/relation statements being used for discovering the semantic under the context of entity B2010 that links the entity to MasterFormat entity 05410 (semantic association):

B2010 Exterior Wall:

*1. Exterior Wall **Framing**: **Cold-formed**, light gage **steel studs**, C-shape, galvanized finish, 6" metal thickness*



In the diagram, "steel" and "framing" match the statement 05100_1 (one of the identifiers of the relation statements exemplified in previous subsection) which is Metals (D5) *used_for* framing. The "steel" matches "Metals" through the transitive property of the relation, *kind_of*. The match is extended to statement 05400, which includes "cold-formed". Finally "studs" is added to the match of statement 05410, through statement (05400_05410). Indeed the entity B2010 Exterior Wall in UniformatII has a semantic relationship with 05410 Load bearing metal studs in MasterFormat and the semantic can be described by the relation *made_of*.

One characteristic worthy of mentioning is that the entity B2010 Exterior Wall in the taxonomy provides a good context for helping refining the association. For instance, the above matching, even without the "framing" keyword, is still possible since the inherited semantic of the hierarchy, shell, closure, and exterior walls, has very close meaning as framing.

As shown in the above example, the documents or specifications that this research addresses have following characteristics:

1. Content has limited scope. It often details what, where, how, and when objects and activities being involved in a domain application. It usually contains rich semantics (author's intention for communicating with other stakeholders) related to standards (due to the agreement among stakeholders) that coordinate objects and activities in the domain.
2. Content are categorized according to taxonomy. In other words, text in a document has some assumption or context, which is inherited along the taxonomy hierarchy.
3. Terminologies are relatively unified and unambiguous.
4. Sentences are relatively free styled, such as note-styled or template-styled due to writing convention or standards.

These characteristics distinguish this research from others, such as [9] and [15] which extract shallow information from general or web documents.

In addition to the intrinsic semantics of standards, this module also explores their application or context semantics in order to achieve more effective semantic extraction. The application semantics depend on the stakeholders' view or interests, such as information they intent for. For example, a cost estimator might look for MasterFormat items and some numerical information so that they can link them to their MasterFormat-based cost databases. On the other hand, an inspector might be interested in the same information but in different view points that yield to different semantics. For example, to a cost estimator, "*6" metal thickness*" in the PPD means how much the studs with such thickness cost. But for an inspector, it means 6" thickness compliance to associated code.

In summary, this module extracts semantics from the instances (specifications) of multiple standards based on three kinds of ontologies: the ontology of the source standard, the

ontology of target standard, and the application ontology based on the stakeholders' views. The extracted semantics are evidences of semantic association of entities between source and target standards.

5. MEASUREMENT OF AFFINITY

The ontology-based semantic extraction module can be implemented via a matching process between relation statements and text. The goal is to identify a set of matched relation statements of related entities with respect to their standards. For a given entity, its associated relation statements carry different weights depending on their positions in the taxonomy and the information content [16] of their keywords. The measurement of affinity is to quantify the weights so that the degree of the closeness between matched relation statements and their associated entity can be determined. Based on the measurement, a ranking scheme can be devised to identify optimal semantic associations among all matches. The ranking scheme can be modeled as a function of the following factors:

1. Number of relation statements matched.
2. Number of keywords matched.
3. Quality of the matches. The measurement of the quality is an open question. Basically the more specific the matches are, the higher quality they represent. One effective way to model the quality is by their positions in the taxonomy (higher level means less specific and thus carries less weight) and by the information content of their keywords. The information content can be quantified by their inverse document frequency (IDF) [17] combined with their counts in the taxonomy (appearing more times means less specific and thus carries less weight)

For instance, in the given example, several entities in MasterFormat contain "framing" and "Metals", which are all candidates for semantic association. The entity 05410 is considered as the optimal one because it matches more keywords along its taxonomy hierarchy and some of them, such as studs, are very specific with respect to both position and IDF.

6. ARCHITECTURE

The major thrust of the research is to develop an integration framework that facilitates exploitation of semantics from taxonomy-based standards and instantiations of the standards to achieve higher interoperability between domain participants and their information systems. To demonstrate the applicability of the proposed approach toward the goal, this section shows an overall architecture depicting one possible implementation and its relationship with other related technologies.

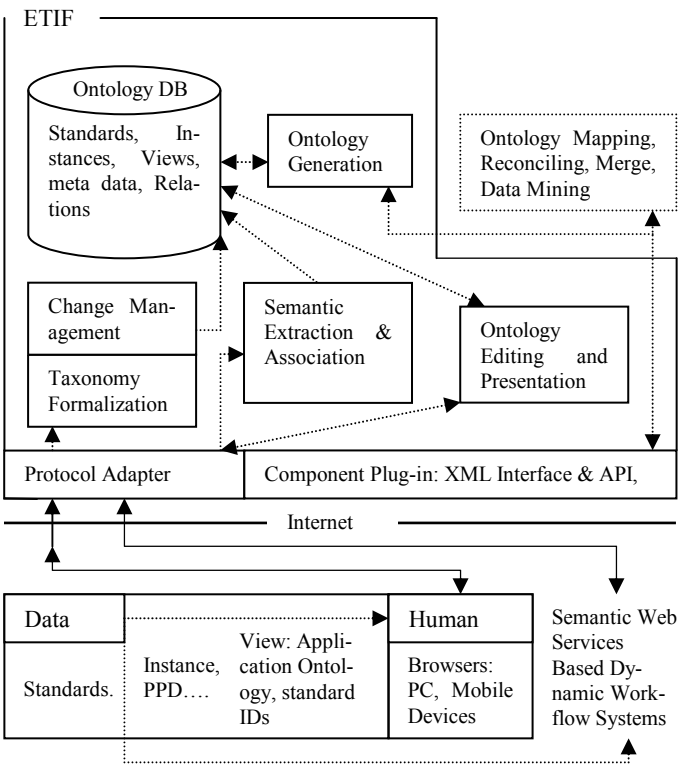


Figure 2. Extensible Taxonomy-based Integration Framework (ETIF)

In the framework shown in Figure 2, relations and relation statements of various versions of standards written in natural languages are developed and uploaded via web-based tools to the system by stakeholders in the application domain. The taxonomy formalization along with the change management modules process them through parsing, normalization, generalization, linguistic processing (such as inflection, derivation, compounds, and synonyms), and indexing for incremental update in the ontology database. For a particular application, the stakeholders upload instances of the source standard (e.g., PPDs), target standard, and its application ontology. After processing the free text of PPD instances through linguistic techniques such as tokenization, chunk parsing, and grammatical function recognition [9], the system applies the semantic extraction and ranking algorithms, and returns/deposits extracted metadata and semantic association to the ontology database and also to the users or clients, if applicable, for feedback.

The integration of competing and complementary standards is a critical step for enhancing interoperability among heterogeneous systems using the standards. The proposed semantic association is only one aspect in this effort. It should be supplemented with other technologies such as ontology mapping, reconciling, and merging to provide a practical and complete solution. The framework includes a plug-in mechanism via XML-based interfaces and API for external software component integration.

The formalized standards, their instances, users' application ontologies, and extracted metadata form a semantic rich ontology repository. Integrating the repository with other ontology techniques through the plug-in mechanism allows the effective construction of application domain ontology.

Web services enriched with the vision of the semantic web have emerged as a mainstream solution to system integration over the Internet. Following the same trend, the implementation of the proposed framework adopts the Web Ontology Language (OWL) [8] with the intention of integrating building construction workflow systems via semantic web services.

7. CONCLUSION AND FUTURE WORKS

This paper demonstrates the effective use of taxonomy for ontology developments and the semantic association of ontology for interoperability in a workflow system with building construction as the target example. It illustrates a systematic approach to semantic association through taxonomy formalization and ontology-based semantic extraction. The overall system implementation in web environment is also proposed. Current activities of the research project include the complete ontological formalization of the MaterFormat and UniformatII standards, refinement of the affinity measure for general taxonomy, and the integration of the algorithms with dynamic workflow systems through semantic web services.

8. ACKNOWLEDGMENTS

This work is partially supported by an NSF research grant ITR-0404113.

REFERENCES

[1] Construction Specifications Institute. MasterFormat 95™ : *Alexandria, VA: The Construction Specifications Institute, 1995 edition.*

[2] Charette, R. P. and Marshall, H. E.: UNIFORMAT II Elemental Classification for Building Specifications, Cost Estimating, and Cost Analysis, *NISTIR 6389, Gaithersburg, MD: National Institute of Standards and Technology, October, 1999*

[3] Gallaher, M. P.; O'Connor, A. C.; Dettbarn, J. L., Jr.; Gilday, L. T.: Cost Analysis of Inadequate Interoperability in the U.S. Capital Facilities Industry, *NIST GCR 04-867, Gaithersburg, MD: National Institute of Standards and Technology, August, 2004.*

[4] Jayant Madhavan, Philip A. Bernstein, and Erhard Rahm: Generic Schema Matching with Cupid, *at the Twenty Seventh International Conference on Very Large Databases (VLDB'2001), Roma, Italy.*

[5] N.F. Noy and M.A. Musen. The prompt suite: Interactive tools for ontology merging and mapping. *Journal of Human-Computer Studies, 59(6):983--1024, 2003*.

[6] M. Paolucci, T. Kawamura, T. Payne, and K. Sycara. Semantic matching of web services capabilities. *In The First International Semantic Web Conference (ISWC), 2002.*

[7] Rosen, Harold J. : Construction specifications writing : principles and procedures 5th edition, *Hoboken, N.J. : J. Wiley, c2005*.

[8] Mike Dean and Guus Schreiber: Editors OWL Web Ontology Language Reference, *W3C Recommendation, http://www.w3.org/TR/2004/REC-owl-ref-20040210, 10 February 2004.*

[9] Maedche, A., Neumann, G., Staab, S.: Bootstrapping an Ontology-Based Information Extraction System, *Intelligent Exploration of the Web, Springer 2002*.

[10] Gruber, T.R., A Translation Approach to Portable Ontology Specification: *Knowledge Acquisition 5: 199-220, 1993.*

[11] Rahm, E and Bernstein, P. A. "A Survey of Approaches to Automatic Schema Matching." *The VLDB Journal, Vol. 10, pp. 334-350, 2001.*

[12] Do, H., Melnik, S. and Rahm, E. "Comparison of Schema Matching Evaluations." In Proceedings of the 2nd Int. *Workshop on Web Databases (German Informatics Society), 2002*.

[13] Aberber, K., Cudré-Mauroux, P. and Hauswirth, M. "The Chatty Web: Element Semantics through Gossiping." *The Proceedings of the 20th International World Wide Web Conference, pp. 197 – 206, 2003*.

[14] Doan, A., Madhavan, J., Domingos, P. and Halevy, A. "Learning to Map between Ontologies on the Semantic Web." *The VLDB Journal, Vol. 12, pp. 303-319, 2003.*

[15] David W. Embley , Douglas M. Campbell , Randy D. Smith , Stephen W. Liddle.: Ontology-based extraction and structuring of information from data-rich unstructured documents, *Proceedings of the seventh international conference on Information and knowledge management, p.52-59, November 02-07, 1998, Bethesda, Maryland, United States*

[16] Ross, S.: A First Course in Probability. *Macmillan Publishing, 1976.*

[17] Church, K. W. and Gale, W. A. : Inverse document frequency (IDF): A measure of deviations from Poisson. In Yarowsky, D. and Church, K., editors, *Proceedings of the Third Workshop on Very Large Corpora, pages 121--130. Association for Computational Linguistics. 1995.*