Relating Taxonomies with Regulations

Chin Pang Cheng, Jiayi Pan
Stanford University
Dept. of Civil & Environmental Eng.
Stanford, CA 94305-4020
(cpcheng, pjy@stanford.edu)

Gloria T. Lau, Kincho H. Law
Stanford University
Dept. of Civil & Environmental Eng.
Stanford, CA 94305-4020
(glau, law@stanford.edu)

Albert Jones
Enterprise Systems Group
NIST
Gaithersburg, MD 20899-0001
(albert.jones@nist.gov)

ABSTRACT

Increasingly, taxonomies are being developed for a wide variety of industrial domains and specific applications within those domains. These industry or application specific taxonomies attempt to represent the vocabularies commonly used by the practitioners. These formal representations have the potential to automate information retrieval, facilitate interoperability and improve decision making. Decisions made must comply with existing government regulations and codes of practices, which are not always known to the industry practitioners. Although regulations and codes are now in digital forms and are often available online, it remains difficult to search for relevant regulatory information that are applicable to particular decisions. As industry practitioners, unlike legal practitioners, are familiar with one or more industry-specific taxonomies but not necessarily regulatory organization systems, it would be desirable to relate regulations with existing industry-specific taxonomies.

The mapping from a single taxonomy to a single regulation is a trivial keyword matching task. In this paper, we examine techniques to map a single taxonomy to multiple regulations, as well as to map multiple taxonomies to a single regulation. Those techniques include cosine similarity, Jaccard coefficient and market-basket analysis. These techniques provide a metric that measures the similarity between concepts from different taxonomies. Preliminary evaluations of the three metrics are performed using examples from the building industry. These examples illustrate the potential regulatory benefits from the mapping between various taxonomies and regulations.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models*, J.1 [**Administrative Data Processing**]: *law.*

Keywords

Heterogeneous Ontologies, Taxonomy Interoperability, Relatedness Analysis, Regulation Retrieval.

1. INTRODUCTION

Government regulations are an important asset of the society. They extend the laws governing the country with specific guidance for corporate and public actions. Ideally regulations should be readily retrievable by interested individuals. To aid understanding of the law, much prior research focused on the abstraction and retrieval of case law [1, 3, 5, 30, 38, 40], analysis of regulations [19, 20], and compliance guidance for regulations [15, 16]. Relatively little research, however, has been devoted to methodologies and tools that allow practitioners to *intelligently browse and retrieve* relevant regulations utilizing familiar terms and vocabularies (for example, according to industry-specific taxonomies).

Increasingly, taxonomies are being developed to capture and represent those terms and vocabularies formally for a wide variety of industrial domains. These taxonomies can facilitate information interoperation and regulation retrieval. Interoperability is important because it allows practitioners to access, relate, and combine information from multiple, heterogeneous sources and therefore increases the value of information. Recent studies by the National Institute of Standards and Technology (NIST) have reported that inefficient interoperability and integration led to significant costs to the construction as well as the automotive industries [7, 14].

Ontologies have been proposed as a way to remove these inefficiencies. One recent forecast estimates that "By 2010, ontologies ….will be the basis for 80 percent of application integration projects" [32]. Ontologies serve as a means for information sharing and capture the semantics of domain-specific information in a formal and computer interpretable form. They have the potential to automate much of the integration process, thereby reducing cost and time significantly.

Building a single ontology for an entire domain, however, has proven to be neither efficient nor practical. Rather, small communities that need to exchange information frequently tend to build their own distinct ontologies [36]. In practice, multiple terminology classifications or data model structures exist. For instance, the architectural, engineering and construction (AEC) community has built several ontologies that describe the semantics of building models. Even though these ontologies are all targeted towards the same user group, the structures, vocabularies and coverage differ depending on the application.

Government regulations, on the other hand, are often organized and classified by the needs of the agency that enforces them, rather than the needs of the communities that use them [6]. Consequently, there is a clear need and benefit of bridging these two distinct needs. One way to build such a bridge is to develop methods and tools that enable practitioners to browse and retrieve government regulations using their own terms and vocabularies, for example, via existing industry taxonomies. For instance, to browse through regulations for compliance requirements, adhering to an existing taxonomy that the users are familiar with minimizes learning of new classification and vocabularies. Their mental models are better represented using existing taxonomies rather than agency's classification for regulations.

In this paper, we present a systematic approach to map regulations to industry-specific taxonomies. We begin with linking one taxonomy to one regulation which is a trivial keyword extraction

task. Extending one taxonomy to multiple regulations requires clustering of relevant sections from different regulations. To do this, we reuse the relatedness analysis core from [19] to compute relevancy between those sections. We then discuss the need and the challenges of mapping multiple taxonomies to a single regulation. Three different methodologies are investigated to cluster relevant concepts from different taxonomies in order to compute relevancy between those concepts. Cosine similarity and Jaccard coefficient, two vector-based similarity measures commonly used in the field of information retrieval are adopted to compare semantic similarity between ontologies. The market basket model, a popular technique in data mining, is modified as a relatedness analysis measure for ontology mapping. Our preliminary evaluations of the three metrics are discussed. We conclude with our proposal to address methods for mapping multiple taxonomies to multiple regulations.

2. Illustrative Ontology Standards and Regulatory Corpus

We work with taxonomies and regulatory corpus from both the building industry and the environmental protection industry [15, 16, 19, 20]. To illustrate their organization and structure, we present briefly the ontology standards and classification systems that are commonly used in the building industry. For the AEC industry, there are a few ontologies that describe the semantics of building models, such as the CIMsteel Integration Standards (CIS/2) for the steel building and fabrication industry [8], the Industry Foundation Classes (IFC) initiated by the CAD vendors for design description of building components [12], and the OmniClass construction classification system (OmniClass) for the construction specification, materials and product components [34].

Figures 1 and 2 show excerpted examples of the *OmniClass* and *IfcXML* standards. Typical of ontology standards, both are organized hierarchically with implicit "is-a" type relationships defined accordingly. OmniClass consists of 15 tables, each of which represents a different facet of construction information. Each term is associated with a unique ID. For example, the term "Sound and Signal Devices" is associated with the ID "23-85 10 11 11". For the IfcXML, the Industry Foundation Class objects are expressed in an XML structure that defines the hierarchical relationship between elements and entities. To extract the object terms for mapping purposes, the two standards are preprocessed to eliminate the miscellaneous information, such as the IDs in the OmniClass and the element names, group names and type names in the IfcXML, as well as the duplicated terms.

Regulations are voluminous and cover a broad range of scopes and topics. Increasingly, regulatory documents are available online and organized in XML structure. The International Building Code (IBC) [13], which represents the code of practice in the building industry, is employed as one of the regulatory document corpuses. Figure 3 shows a provision in IBC and its representation in XML structure. One notable feature of regulations is that they are typically organized into sections and sub-sections, each of which contains contents with a specific topic or scope. The tree hierarchy of regulations provides useful information that can be explored, for example, to locate similar sections and to build an e-government system [19, 20].

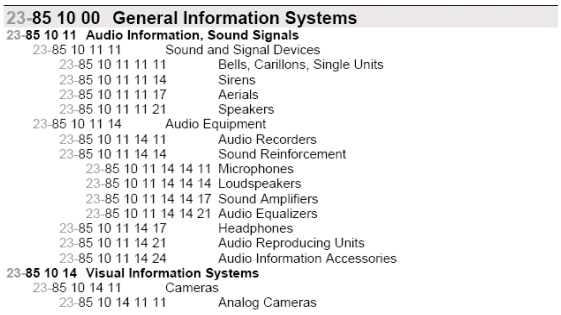


Figure 1: Excerpt from OmniClass Construction Classification System

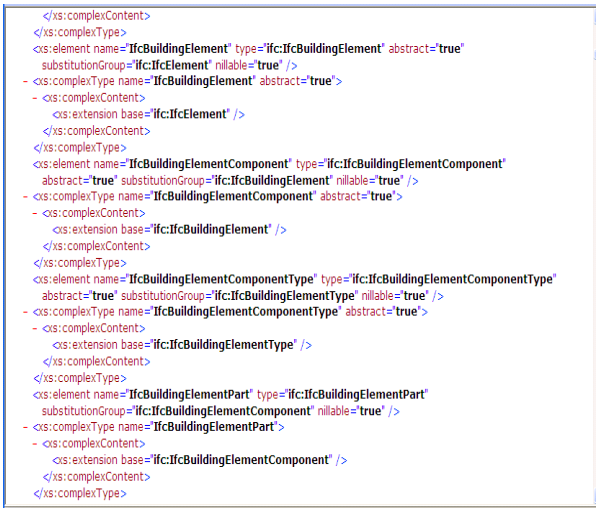
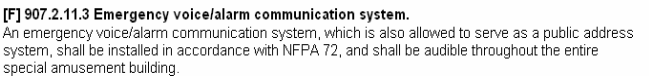


Figure 2: Organization of IfcXML


[F] 907.2.11.3 Emergency voice/alarm communication system.
An emergency voice/alarm communication system, which is also allowed to serve as a public address system, shall be installed in accordance with NFPA 72, and shall be audible throughout the entire special amusement building.

```
<LEVEL level-depth="8" style-id="0-0-0-304" style-name="Section3" style-name-escaped="Section3" toc-section="true">
<RECORD id="0-0-0-5529" number="5529" version="3">
<HEADING>
[F] 907.2.11.3 Emergency voice/alarm communication system.
</HEADING>
<PARA>
<DESTINATION id="0-0-0-3521" name="IBC2006907.2.11.3"/>
<CHARFORMAT bold="1" hidden="0" italic="0" strike-out="0" underline="0">[F] 907.2.11.3 Emergency voice/alarm communication system. </CHARFORMAT>
</PARA>
</RECORD>
<LEVEL level-depth="0" style-id="0-0-0-0" style-name="Normal Level" style-name-escaped="Normal-Level" toc-section="false">
<RECORD id="0-0-0-5530" number="5530" version="3">
<PARA style-id="0-0-0-15" style-name="Body3" style-name-escaped="Body3">An emergency voice/alarm communication system, which is also allowed to serve as a public address system, shall be installed in accordance with NFPA 72, and shall be audible throughout the entire special amusement building.</PARA>
</RECORD>
</LEVEL>
</LEVEL>
```

Figure 3: An IBC Provision and XML Structure

3. ONE TAXONOMY TO ONE REGULATION

Mapping one taxonomy to one regulation is a simple keyword latching task. There are many commercial tools available to latch keywords from documents into a taxonomy. Industry taxonomies are hierarchical classification systems which are generally less than 10 levels deep. Node labels in the taxonomy tree are treated as concept keywords, and they are mapped to sections in the regulation where they appear. As regulations tend to be voluminous, we use a section or subsection as a unit of interest. Figure 4 shows the International Building Codes (IBC) [13] latched with the OmniClass. Users can then traverse the taxonomy and browse relevant sections of the regulation.

Extending the mapping from one taxonomy to multiple regulations unfortunately leads to the classic problem of information overload. For instance, suppose we want to search the Web to find state regulations governing chlorine levels in drinking water. If we search the drinking-water regulations in Alabama and Arizona for the concept "chlorine", we would find over 30 sections in each. The actual relevancy of these 60 sections to chlorine levels is not known. The problem is that Web content ignores the actual structure of the documents. Consequently, search engines cannot take that structure into account when computing relevancy. The result is that users quickly become frustrated with information overload; intelligent retrieval and presentation of web results has become the holy grail of search [4]. Fortunately, regulatory documents are much more organized and structured than web content, and we propose to solve the problem of information overload by clustering relevant sections from different regulations and pivoting on one regulation that the user is most familiar with. We discuss our approach in the following section.

4. ONE TAXONOMY TO MULTIPLE REGULATIONS

Simultaneous traversal of multiple regulation trees using one taxonomy is a challenging but real problem. It is not uncommon for industry practitioners to be familiar with one particular regulation but not the others. For example, an engineer from Montgomery might be familiar with Alabama state code, but not Arizona state code. Nonetheless, if a water distribution system is to be designed that provides water to Phoenix from lakes near Montgomery, the engineer would need an understanding of both [9]. In this scenario, finding the relevant Arizona regulations on chlorine levels might pose a serious problem. We believe that it is beneficial to map the taxonomy to Alabama code first, and then branch out to recommend related sections from the Arizona code. In general, focusing on one regulation as the basis for finding relevant sections from other regulations significantly reduces information overload.

Figure 5 shows a simple user interface that shows a scenario of finding related provisions between regulations from the two states. After browsing down the taxonomy tree to the concept "chlorine", users are shown a list of matched sections from the Alabama regulation. As discussed in Section 3, matching sections to taxonomy concept is simply keyword latching. Selecting Section 335.7.6.15 of the AL code shows that there are 15 recommended sections from the Arizona regulation. A user can

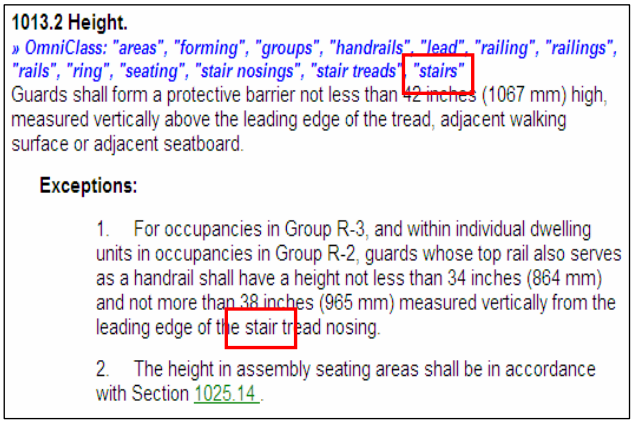


Figure 4: Regulation Latched with Taxonomy Concepts

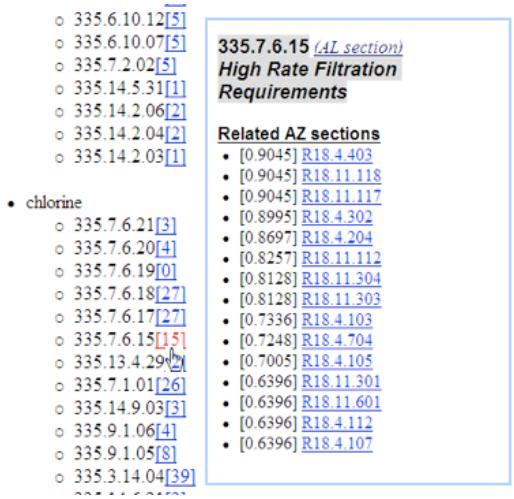


Figure 5: Chlorine mapped to Section 335.7.6.15 in AL code, which have 15 related sections in AZ code

stay focused on the regulation of their choice, and at the same time acquire relevant sections from other regulations as needed. There are two major challenges to developing such a system: a suitable user interface and a methodology for determining relevant regulations. In this paper, we focus on methodologies for making recommendations based on relevancies between sections from different regulations.

To identify related provisions from different regulations, we reuse the relatedness analysis core from [19, 20], which compares sections from different regulations based on shared features using a cosine similarity measure (see Section 5.1) [18, 31]. The goal is to identify the most strongly related provisions using not only a traditional term match but also a combination of feature matches, and not only content comparison but also structural analysis. Regulations are first compared based on conceptual information as well as domain knowledge through a combination of feature matching. Regulations also possess specific structures, such as a tree hierarchy of provisions and the referential structure. These structures represent useful information for locating related provisions and are, therefore, used in the analysis as well. For the detailed discussion on the evaluations of results from the relatedness analysis of provisions, see [19].

5. MULTIPLE TAXONOMIES TO ONE REGULATION

Apart from mapping one taxonomy to many regulations, we also attempt to map many taxonomies to one regulation. As suggested in the Introduction section, multiple taxonomies have been developed for different applications within the same industry domain. Most industry practitioners are familiar with at least one of the taxonomies; but, they frequently need to deal with others for various applications [2, 23]. Therefore, mapping regulations to a single taxonomy is limiting usability of the system. However, traversing regulations using multiple taxonomy trees pose a non-trivial problem. There are much research efforts on ontology merging [33, 39]. These efforts produced a merged ontology that can be used for data interoperability but not as a front-end representation format. Since users would need to learn the newly merged ontology in order to browse regulations, this would defeat the original intent of using existing taxonomies to help locate regulatory provisions. Using the same argument from Section 4, we believe that focusing on one taxonomy that users are familiar with is the right starting point to traverse regulations. Once users reach a taxonomy node of interest, related concepts from other taxonomies can be suggested and users can switch their focal point from one taxonomy to another.

Figure 6 illustrates the proposed approach using the OmniClass [34] and the IFC [12] taxonomies, and the International Building Code (IBC) regulations [13]. The OmniClass is altered from its original representation, shown in Figure 6, to display a widget upon mouse-over that includes an ordered list of matching IBC sections and recommended relevant IFC concepts. In this scenario, the user is more familiar with the OmniClass hierarchy, and thus starts browsing the IBC using this taxonomy. The user uses the term "steel decking" from OmniClass to find an ordered list of matching IBC sections and relevant IFC concepts. Upon locating a list of IBC sections that are related to "steel decking", sorted in order of relevance, the user also sees a list of related IFC concepts including "slab". Mousing-over the IFC concept "slab" brings the focal point to the IFC hierarchy, where the user is presented with the same analysis – namely the IFC elements around this concept "slab", a ranked list of matching IBC sections, and a ranked list of relevant OmniClass concepts.

As opposed to locating related sections from multiple regulations, the task here is to identify similar or related concepts from multiple taxonomies. Ontology mapping has been an active research area since the semantic web movement [28, 29]. It is difficult to develop mappings between two arbitrary ontologies. In our case, however, the problem is slightly more manageable since our ontologies are very industry specific and are targeted towards the same group of users. Similar to the techniques presented in Section 4, the relevance among concepts from different ontologies is computed using a vector comparison approach. A document corpus is used to relate concepts by computing their co-occurrence frequencies. This training corpus must be carefully selected as it represents the relevancy among concepts from different taxonomies. Conveniently, we have a corpus of regulatory documents that have been meticulously drafted and reviewed for accuracy. Unlike web content, regulations do not have random co-occurrences of phrases in the same provision. This dramatically increases the likelihood of finding real matches.

Consider a pool of m concepts and a corpus of n regulation sections. A frequency vector $\vec{C}_i$ is an n-by-1 vector storing the occurrence frequencies of concept i among the n documents. That is, the k-th element of $\vec{C}_i$ equals the number of times concept i is

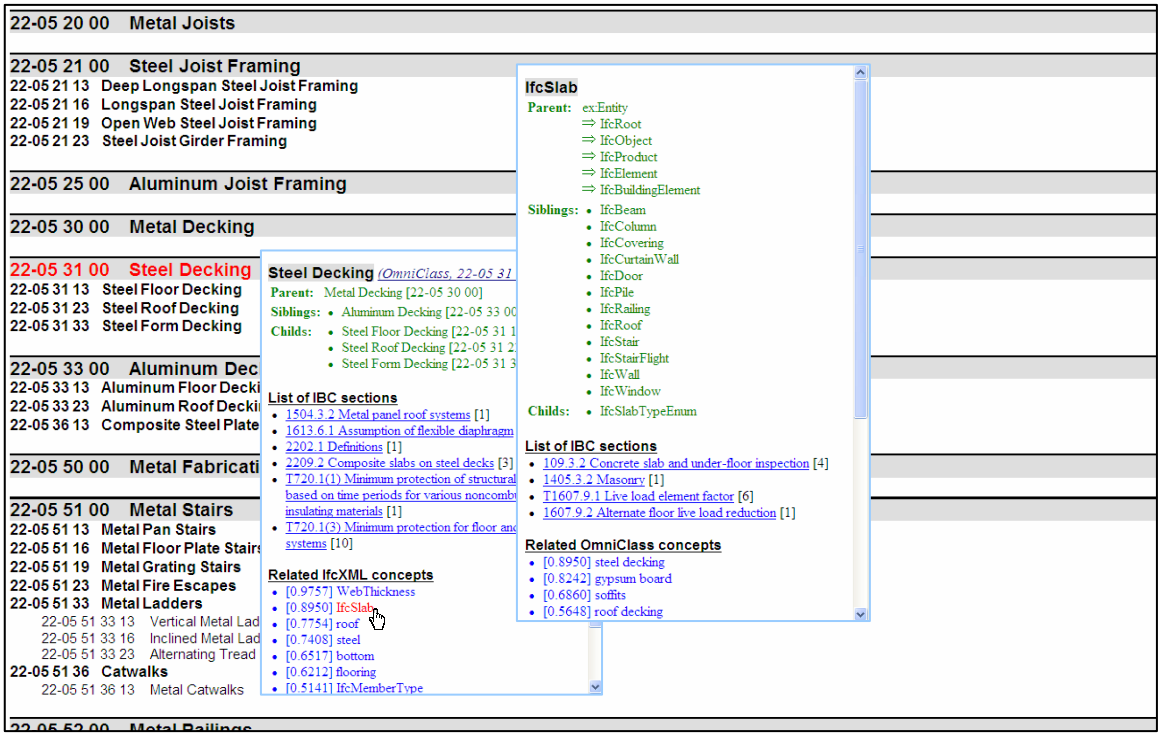


Figure 6: Traversing the IBC using OmniClass Taxonomy with Relevant Concepts from the IFC Taxonomy

matched in section *k*. In subsequent sections, we will discuss three metrics to compute the similarity score among concepts. In our example shown in Figure 6, to relate "steel decking" from the OmniClass to "slab" from the IFC, we compute their similarity score based on the defined metrics. As shown in the figure, their cosine similarity score is 0.895, which ranks second among all IFC concepts that are relevant to "steel decking".

5.1 Cosine Similarity

Cosine similarity is a non-Euclidean distance measure between two vectors. It is a common approach to compare documents in the field of text mining [18, 31]. Given two frequency vectors $\vec{c}_i$ and $\vec{c}_j$, the similarity score between concepts *i* and *j* is represented using the dot product:

$$Sim(i, j) = \frac{\vec{c}_i \cdot \vec{c}_j}{\left|\vec{c}_i\right| \times \left|\vec{c}_j\right|}$$

The resulting score is in the range of [0, 1] with 1 as the highest relatedness between concepts *i* and *j*.

5.2 Jaccard Similarity Coefficient

Jaccard similarity coefficient [31, 37] is a statistical measure of the extent of overlap between two vectors. It is defined as the size of the intersection divided by the size of the union of the vector dimension sets:

$$Jaccard(i, j) = \frac{\left|\vec{c}_i \cap \vec{c}_j\right|}{\left|\vec{c}_i \cup \vec{c}_j\right|}$$

Jaccard similarity coefficient is a popular similarity analysis measure of term-term similarity due to its simplicity and retrieval effectiveness [17]. Two concepts are considered similar if there is a high probability for both concepts to appear in the same sections. To illustrate the application to our problem, let N_{11} be the number of sections both concept *i* and *j* are matched to, N_{10} be the number of sections concept *i* is matched to but not concept *j*, N_{01} be the number of sections concept *j* is matched to but not concept *i*, and N_{00} be the number of sections that both concept *i* and *j* are not matched to. The similarity between both concepts is then computed as

$$Sim(i, j) = \frac{N_{11}}{N_{11} + N_{10} + N_{01}}$$

Since the size of intersection cannot be larger than the size of union, the resulting similarity score is between 0 and 1.

5.3 Market-Basket Model

Market-basket model is a probabilistic data-mining technique to find item-item correlation [11]. The task is to find the items that frequent the same baskets. The *support* of each itemset *I* is defined as the number of baskets containing all items in *I*. Sets of items that appear in *s* or more baskets, where *s* is the support threshold, are the *frequent itemsets.*

Market-basket analysis is primarily used to uncover association rules between item and itemsets. The *confidence* of an association rule $\{i_1, i_2, ..., i_k\} \rightarrow j$ is defined as the conditional probability of *j* given itemset $\{i_1, i_2, ..., i_k\}$. The *interest* of an association rule is defined as the absolute value of the difference between the confidence of the rule and the probability of item *j*. To compute the similarities among concepts, our goal is to find concepts *i* and *j* where either association rule $i \rightarrow j$ or $j \rightarrow i$ is high-interest.

Consider a corpus of *n* documents. Let N_{11} be the number of sections both concept *i* and *j* are matched to, N_{10} be the number of sections concept *i* is matched to but not concept *j*, and N_{01} be the number of sections concept *j* is matched to but not concept *i*. The probability of concept *j* is computed as

$$\Pr(j) = \frac{N_{11} + N_{01}}{n}$$

and the confidence of the association rule $i \rightarrow j$ is

$$Conf(i \rightarrow j) = \frac{N_{11}}{N_{11} + N_{01}}$$

The forward similarity of the concepts *i* and *j*, which is the interest of the association rule $i \rightarrow j$ without absolute notation, is expressed as

$$Sim(i, j) = \frac{N_{11}}{N_{11} + N_{01}} - \frac{N_{11} + N_{01}}{n}$$

The value ranges from -1 to 1. The value of -1 means that concept *j* appears in every section while concept *i* does not co-occur in any of these sections. The value of 1 is unattainable because (N_{11} + N_{01}) cannot be zero while confidence equals one. Conceptually, it represents the boundary case where the occurrence of concept *j* is not significant in the corpus, but it appears in every section that concept *i* appears.

Table 1 summarizes the similarities and differences among the three metrics, namely the cosine similarity, Jaccard similarity and the market basket model. All are statistical analysis measures that leverage the co-occurrence frequencies of concepts in the regulatory corpus.

5.4 Use of Regulation Hierarchy Structural Information

Many related concepts can be uncovered by treating each section in a regulation as an independent document. In this model, a concept-document matrix is generated to compute concept co-occurrence in documents, which are really regulatory sections. This model is generally sufficient in revealing most related concepts, but some related concepts rarely co-occur in the same sections. For example, if two concepts contain an Is-A relationship, like door furniture and door hardware, they may be used in the same regulation interchangeably but in different sections.

Table 1: Similarities and differences among the three statistical analysis measures

	Cosine Similarity	Jaccard Similarity	Market basket Model
Non-Euclidean	Yes	Yes	Yes
Vector-based	Yes	Yes	No
Underlying methodology	Vector space model	Set theory (Intersections and unions)	Probability theory and association rule
Symmetrical forward and backward scores	Yes	Yes	No
Range of scores	[0, 1]	[0, 1]	[-1, 1)
Usage	Often used as the baseline metric	Computationally effective	To discover potential item-item correlation

Is-A-related concepts are also hard to find if each section is treated as if it were an independent document. The relationship between *Is-A*-related concepts, such as "building materials" and "concrete" as shown in Figure 7, are sometimes implicit from the structures of sections. For example, the descriptions of "building materials" and those of "concrete" may not appear in the same section. Instead, the sections describing "concrete" are usually the subsections of the sections describing "building materials." If we consider subsections and its parent section in the computation of the similarity score between "building materials" and "concrete," the implicit relationship between building materials and concrete might become more obvious. Therefore, the hierarchical structure of sections needs to be considered to extract non-trivial related concepts.

Regulations contain well-organized hierarchical structures with sections and sub-sections of specific topic or scope. There are organizational and referential structures explicitly defined in regulations. Section 4 briefly discussed the usage of the regulation hierarchical information to locate related sections from different regulation trees. The results [20] show that regulatory structure sometimes helps to reduce prediction error of related provisions [19]. Here, regulation hierarchy will be considered to uncover semantic relationships between concepts from different taxonomies in the same manner.

Well-structured regulations could be represented as a hierarchical tree, where each section corresponds to a discrete node. As illustrated in Figure 8, each section has a parent section, a set of sibling sections and a set of child sections. In general, for a section with a particular topic, the parent section describes a broader topic, the sibling sections describe parallel topics, and the child sections describe more specific topics. In our computation, we will consider the co-occurrence of concepts in a broader scope, namely the parent, sibling and child sections.

The frequency matrix C is modified to take the parent section, sibling sections and child sections into consideration. To include the parent section, the weighted numbers of occurrence for all the concepts in the parent section are added to the numbers of occurrence in the self section. Similarly, the sibling sections and child sections are then included with a discounted weight. In our formulation below, we will denote *Par(k)*, *Sib(k)* and *Child(k)* as the parent section, set of sibling sections and set of child sections of Section k. The *k*-th element of frequency vector $\vec{c}_i$, i.e., the number of times concept *i* is matched to Section *k*, is updated as

$$c_i(k) := c_i(k) + w_p c_i(Par(k)) + w_s \sum_{u \in Sib(k)} c_i(u) + w_c \sum_{v \in Child(k)} c_i(v)$$

where w_p, w_s and w_c are the weights of the parent, sibling and child sections respectively.

5.5 Evaluations of the Measures

Twenty concepts are randomly selected from the OmniClass and the IFC hierarchies respectively, and pairwise similarity scores are computed using the three relatedness analysis measures described above. The results of concept matching performed by domain experts are treated as the true matches. Root mean square error (RMSE), precision, recall and F-measure are used as performance metrics to evaluate and compare the three measures and the use of regulation structural information. A baseline ontology matcher is compared to the three measures using precision and recall as the evaluation metrics.

Three domain experts are asked to identify the related concept pairs among a total of 400 possible pairs. Related concept pairs are assigned a true value of one; all other pairs are assigned a true value of zero. As for the predicted values, two concepts are predicted as similar or related if the computed similarity score is larger than certain threshold score. Based on each of the three manual and independent mappings, we computed values of RMSE, precision, recall and F-measure for the three measures, the baseline ontology matcher, and different regulation structural information inclusions. The averages of the results from the three true answers are then taken as the final results. Details are given in the following sections.

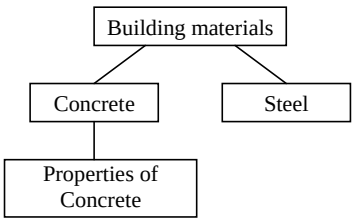


Figure 7: Example of related but rarely co-occurring concepts

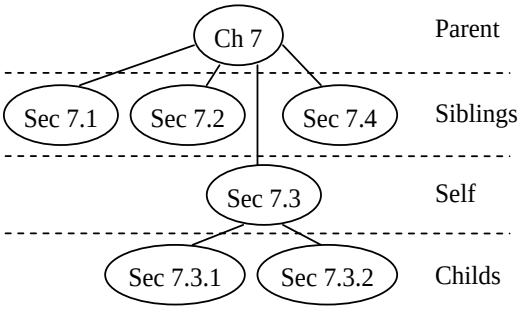


Figure 8: Tree hierarchy of sections in regulations

5.5.1 Root Mean Square Errors (RMSE) among the Three Measures

Root mean square error (RMSE) is a metric to compute the difference between the predicted values and the true values so as to evaluate the accuracy of the prediction. Comparison between ontology of *m* concept terms and ontology of *n* concept terms involves *m* by *n* concept-concept pairs. Therefore the RMSE is calculated as

$$RMSE = \sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left|true_{i,j} - predicted_{i,j}\right|}$$

Figure 9 shows the results of the three measures compared using RMSE for threshold similarity scores ranging from 0.15 to 0.9. No regulation hierarchy structural information is considered. As illustrated in Figure 9, we conclude that the market-basket model results in the lowest RMSE for most threshold similarity scores. This means that the market-basket model outperforms the other two measures in locating related concept pairs from different ontologies, using provisions from regulations as independent documents in the co-occurrence computation. Cosine similarity appears to be average among the three measures.

5.5.2 Precision, Recall and F-measure among the Three Measures

We use precision, recall and F-measure values to compare the three similarity analysis measures and the use of regulation hierarchy structural information. While RMSE takes both correctness and incorrectness of prediction into consideration, precision and recall emphasize correctness only. Precision and recall evaluates the accuracy of predictions and the coverage of accurate pairs. Precision measures the fraction of predicted matches that are correct, i.e., the number of true positives over the number of pairs predicted as matched. Recall measures the fraction of correct matches that are predicted, i.e., the number of true positives over the number of pairs that are actually matched. They are computed as

$$Precision = \frac{\left|True\ Matches \cap Predicted\ Matches\right|}{\left|Predicted\ Matches\right|}$$

$$Recall = \frac{\left|True\ Matches \cap Predicted\ Matches\right|}{\left|\text{True}\ Matches\right|}$$

There is always a tradeoff between precision and recall. F-measure is therefore leveraged to combine both metrics. It is a weighed harmonic mean of precision and recall. In other words, it is the weighed reciprocal of the arithmetic mean of the reciprocals of precision and recall. It is computed as

$$F - Measure = \frac{2 \cdot (Precision \times Recall)}{Precision + Recall}$$

Figure 10 shows the results for the three relatedness analysis measures using F-measure, a combination of precision and recall. The market-basket model shows the highest F-measure values in all cases, again, consistent with the RMSE results. In fact, market-basket model achieves the highest recall rate with relatively high precision in all cases. Jaccard similarity is not preferred due to its low F-measure values, resulted from its very low recall rates.

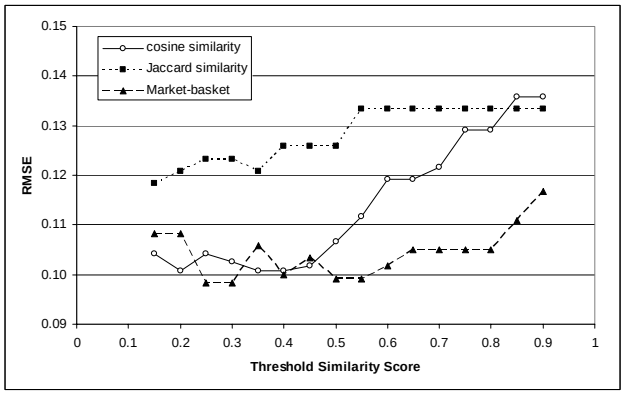


Figure 9: Evaluation results of the three measures using RMSE

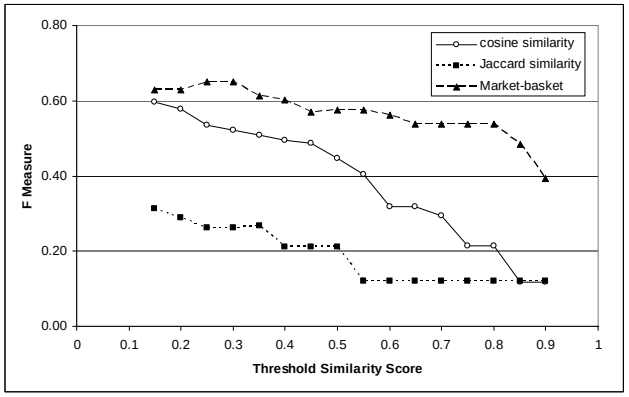


Figure 10: Evaluation results of the three measures using F-measure

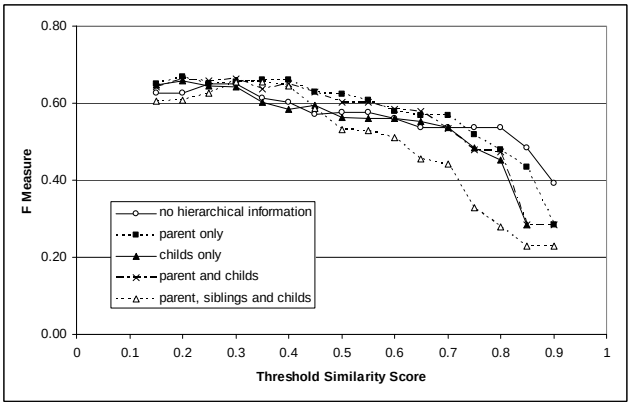


Figure 11: Evaluation results of market-basket model using F-measure

Cosine similarity appears to be average among the three measures, consistent with the RMSE results.

As the market-basket model outperforms cosine and Jaccard similarities using both RMSE and the F-measure, we will evaluate the impact of regulation hierarchy using the market-basket model as the similarity measure of choice. As shown in Figure 11, the effect of including regulatory structure in the analysis is inconclusive. In general, it increases recall rate and reduces

precision, as more regulatory nodes are considered to locate related concepts. The inclusion of parent section produced a slightly higher F-measure in most threshold scores, likely due to the fact that parent relationship is one to one which minimizes the impact on precision. Other relationships, such as sibling and child, are not one to one; the number of such relationships, therefore, could heavily tax precision with only minor increase in recall.

5.5.3 Comparison of the Domain-based Model with the Lexicon-based Model

In addition to comparing the three measures with one another, we also evaluated our domain-based approach to a traditional lexicon-based approach. Ontology mapping is an active research topic, and common mapping methods discover the semantic similarity between ontology elements using rule-based [22, 28], lexicon-based [24, 35] and structure-based methods [25, 27]. Our approach is comparable to a lexicon-based approach, where dictionary and thesaurus are used to enumerate synonyms, homonyms, and etc. In our analysis, we use a domain-specific corpus, i.e., a domain-appropriate regulation, to uncover such semantic relationships.

A thesaurus is necessary to compare our model to a lexicon-based one. A common thesaurus is the WordNet [26], which is a well-known lexical resource for the English language. Synonyms in WordNet are interlinked by means of conceptual-semantic and lexical relations. It is one of the most widely adopted synonym sources for ontology matching techniques including CUPID [24], Learned Ontology Model (LOM) [21], and Version Matching Approach (VMA) [41]. Table 2 shows the result for comparing domain-based ontology mapping method with a lexicon-based element matcher using WordNet.

Table 2 shows that our domain-based approach outperforms the lexicon-based matcher in terms of precision and recall. Some examples of matches that are found by our domain-based matcher but not by the lexicon-based matcher are: (sound and signal devices, IfcSwitchingDeviceType), (door hardware, IfcBuildingElementComponent), (steel decking, IfcSlab), and (sound and signal devices, IfcAlarmType). The reliability of lexicon-based matchers is not guaranteed because their use of stemmers to reduce derived words to their root form, e.g., from piling to pile, is not always appropriate for the domain [10]. In addition, many concepts have different meanings when used in different domains, so that their synonyms and definitions could be different.

We should note that WordNet is a generic linguistic thesaurus rather than an industry-specific taxonomy. As a result, it contains little and imprecise information of the terminology used by the OmniClass and IfcXML. The result shows that domain-related corpora, such as regulations and technical specifications, are useful in discovering the semantic relationships across multiple ontologies.

Table 2: Precision and recall comparisons of domain-based ontology mapping to lexicon-based ontology mapping

Score threshold	Approaches	Cosine Similarity		Jaccard Similarity		Market-basket Model	
		P	R	P	R	P	R
0.2	Lexicon-based Matcher	0.50	0.03	0.00	0.00	0.00	0.00
	Domain-based Matcher	0.79	0.53	0.91	0.17	0.70	0.71
0.3	Lexicon-based Matcher	0.50	0.03	1.00	0.03	0.50	0.03
	Domain-based Matcher	0.83	0.41	0.90	0.15	0.75	0.71
0.4	Lexicon-based Matcher	1.00	0.03	1.00	0.03	0.50	0.03
	Domain-based Matcher	0.91	0.36	1.00	0.12	0.80	0.59
0.5	Lexicon-based Matcher	1.00	0.03	1.00	0.03	1.00	0.03
	Domain-based Matcher	0.90	0.31	1.00	0.11	0.81	0.51
0.6	Lexicon-based Matcher	1.00	0.03	1.00	0.03	1.00	0.03
	Domain-based Matcher	0.92	0.20	1.00	0.07	0.81	0.49

6. CONCLUSIONS & FUTURE TASKS

Regulatory documents are written by government agencies who organize the material to suit their own needs. From industry practitioners' standpoint, the original hierarchy might not be the easiest retrieval model for regulations. In this paper, we proposed a system to map concepts from industry-specific taxonomies to similar concepts in those regulations to increase their usability by industry practitioners. A running example from the AEC industry is shown to illustrate the need, the usage and the benefit of the mapping system.

A 1-1, 1-n, n-1 mapping between taxonomies and regulations are demonstrated. We plan to implement an n-n concept-section mapping in the future, by combining the techniques of concept comparisons and section comparisons. In section comparisons, the hierarchical information of regulations is used to enhance the analysis; we also plan to incorporate the hierarchical information of taxonomies into concept comparisons. In concept comparisons, three similarity metrics are tested, whereas only cosine similarity is implemented for regulatory comparisons which are due for more testing. Among the three metrics, we have shown in this paper that the market-basket model performs the best in terms of RMSE and F-measure, which is a combination of precision and recall. The comparison between the lexicon-based approach and our domain-based approach shows that our approach results in higher precision and recall. A few examples are given to show matches that are found by our domain-based approach but not the lexicon-based approach.

In the future, we plan to engage potential users to help perform formal evaluations of the similarity metrics and the usability of the system. To improve usability, a better user interface is much needed, and we plan to investigate the need to implement or adopt such visualization tool. An ideal user interface should facilitate access to the mapping of multiple taxonomies and the browsing of

regulations by industry practitioners, rulemakers and domain experts.

7. ACKNOWLEDGMENTS

The authors would like to thank the International Code Council for providing the XML version of the International Building Code (2006). The authors would also like to acknowledge the supports by the National Science Foundation, Grant No. CMS-0601167, the Center for Integrated Facility Engineering (CIFE) at Stanford University and the Enterprise Systems Group at the National Institute of Standards and Technology (NIST). Any opinions and findings are those of the authors, and do not necessarily reflect the views of NSF, CIFE and NIST.

8. REFERENCES

[1] Al-Kofahi, K., Tyrrell, A., Vachher, A., and Jackson, P. 2001. A Machine Learning Approach to Prior Case Retrieval. In Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001), St. Louis, Missouri, 88-93.

[2] Begley, E.F., Palmer, M.E., and Reed, K.A. 2005. Semantic Mapping Between IAI ifcXML and FIATECH AEX Models for Centrifugal Pumps, Technical.

[3] Bench-Capon, T.J.M. 1991. Knowledge Based Systems and Legal Applications, Academic Press Professional, Inc., San Diego, CA.

[4] Bonnel, N., Lemaire, V., Cotarmanac'h, A., and Morin, A. 2006. Effective Organization and Visualization of Web Search Results. In Proceedings of the 24th IASTED International Conference on Internet and Multimedia Systems and Applications, Innsbruck, Austria, 209-216.

[5] Brüninghaus, S., and Ashley, K.D. 2001. Improving the Representation of Legal Case Texts with Information Extraction Methods. In Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001), St. Louis, Missouri, 42-51.

[6] Fountain, J.E. 2002. Information Institutions and Governance: Advancing a Basic Social Science Research Program for Digital Government, Technical Report, National Center for Digital Government, John F. Kennedy School of Government, Harvard University.

[7] Gallaher, M., O'Connor, A., Jr., J.B., and Gilday, L. 2004. Cost Analysis of Inadequate Interoperability in the US Capital Facilities Industry, Technical Report, GCR 04-867, NIST.

[8] Garas, F., and Hunter, I. 1998. CIMSteel (Computer Integrated Manufacturing in Constructional Steelwork) - Delivering the Promise. Structural Engineering, 76 (3), 43-45.

[9] Gibbens, M.P. 2000. CalDAG 2000: California Disabled Accessibility Guidebook, Builder's Book, Canoga Park, CA.

[10] Grabar, N., and Zweigenbaum, P. 2000. Automatic Acquisition of Domain-Specific Morphological Resources from Thesauri. In Proceedings of RIAO 2000: Content-Based Multimedia Information Access, Paris, France, 765-784.

[11] Hastie, T., Tibshirani, R., and Friedman, J.H. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, NY.

[12] Industry Foundation Classes (IFC), 1997. International Alliance for Interoperability (IAI).

[13] International Building Code 2000, International Conference of Building Officials (ICBO), Whittier, CA, 2000.

[14] Jacobs, J., and Linden, A. 2002. Semantic Web Technologies Take Middleware to the Next Level, Technical Report, T-17-5338, Gartner Group, http://www.gartner.com/DisplayDocument?doc_cd=109295.

[15] Kerrigan, S. 2003. A Software Infrastructure for Regulatory Information Management and Compliance Assistance, Ph.D. Thesis, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA.

[16] Kerrigan, S., and Law, K. 2003. Logic-Based Regulation Compliance-Assistance. In Proceedings of the 9th International Conference on Artificial Intelligence and Law (ICAIL 2003), Edinburgh, Scotland, 126-135.

[17] Kim, M.-C., and Choi, K.-S. 1999. A Comparison of Collocation-based Similarity Measures in Query Expansion. Information Processing and Management: an International Journal, 35 (1), 19-30.

[18] Larsen, B., and Aone, C. 1999. Fast and Effective Text Mining Using Linear-Time Document Clustering. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, 16-22.

[19] Lau, G. 2004. A Comparative Analysis Framework for Semi-Structured Documents, with Applications to Government Regulations, Ph.D. Thesis, Civil and Environmental Engineering, Stanford University, Stanford, CA.

[20] Lau, G., Law, K., and Wiederhold, G. 2005. Legal Information Retrieval and Application to E-Rulemaking. In Proceedings of the 10th International Conference on Artificial Intelligence and Law (ICAIL 2005), Bologna, Italy, 146-154.

[21] Li, J. 2004. LOM: A Lexicon-based Ontology Mapping Tool. In Proceedings of the Information Interpretation and Integration Conference (I3CON) and the Performance Metrics for Intelligent Systems (PerMIS) Workshop, Gaithersburg, MD

[22] Li, W., Clifton, C., and Liu, S. 2000. Database Integration using Neural Network: Implementation and Experiences. Knowledge and Information Systems, 2 (1), 73-96.

[23] Lipman, R. 2006. Mapping Between the CIMsteel Integration Standards (CIS/2) and Industry Foundation Classes (IFC) Product Model for Structural Steel. In Proceedings of the Conference on Computing in Civil and Building Engineering, Montreal, Canada, 3087-3096.

[24] Madhavan, J., Bernstein, P.A., and Rahm, E. 2001. Generic Schema Matching with Cupid. In Proceedings of the 27th International Conference on Very Large Data Bases (VLDB), Rome, Italy, 49-58.

[25] Melnik, S., Garcia-Molina, H., and Rahm, E. 2002. Similarity Flooding: A Versatile Graph Matching Algorithm.

In Proceedings of the 18th International Conference on Data Engineering (ICDE), San Jose, CA, 117-128.

[26] Miller, G.A., Beckwith, R., Fellbaun, C., Gross, D., and Miller, K. 1993. Five Papers on WordNet, Technical Report, Cognitive Science Laboratory, Princeton, NJ.

[27] Milo, T., and Zohar, S. 1998. Using Schema Matching to Simplify Heterogeneous Data Translation. In Proceedings of the 24th International Conference On Very Large Data Bases, New York, NY, 122-133.

[28] Mitra, P. 2003. An Algebraic Framework for the Interoperation of Ontologies, Ph.D. Thesis, Computer Science Department, Stanford University, Stanford, CA.

[29] Mitra, P., and Wiederhold, G. 2002. Resolving Terminological Heterogeneity in Ontologies. In Proceedings of Workshop on Ontologies and Semantic Interoperability at the 15th European Conference on Artificial Intelligence (ECAI), Lyon, France, 45-50.

[30] Moens, M.-F., Uyttendaele, C., and Dumortier, J. 1997. Abstracting of Legal Cases: The SALOMON Experience. In Proceedings of the 6th International Conference on Artificial Intelligence and Law (ICAIL 1997), Melbourne, Australia, 114-122.

[31] Nahm, U.Y., Bilenko, M., and Mooney, R.J. 2002. Two Approaches to Handling Noisy Variation in Text Mining. In Proceedings of the ICML-2002 Workshop on Text Learning, Sydney, Australia, 18-27.

[32] NIST. 1999. Interoperability Cost Analysis of the US Automotive Supply Chain, Technical Report, #99-1, http://www.nist.gov/director/prog-ofc/report99-1.pdfT, NIST Strategic Planning and Economic Assessment Office.

[33] Noy, N.F. 2003. "Tools for Mapping and Merging Ontologies," In S. Staab and R. Stude (Eds.), Handbook on Ontologies, Springer-Verlag, pp. 365-384.

[34] OmniClass Construction Classification System, Edition 1.0, 2006. Construction Specifications Institute (CSI). http://www.omniclass.org.

[35] Palopoli, L., Sacca, D., Terracina, G., and Ursino, D. 1999. A Unified Graph-based Framework for Deriving Nominal Interscheme Properties, Type Conflicts and Object Cluster Similarities. In Proceedings of the 4th IFCIS International Conference On Cooperative Information Systems (CoopIS), Edinburgh, Scotland, 34-45.

[36] Ray, S. 2002. Interoperability Standards in the Semantic Web. Journal of Computing and Information Science in Engineering, 2, 65-69.

[37] Roussinov, D., and Zhao, J.L. 2003. Automatic Discovery of Similarity Relationships Through Web Mining. Decision Support Systems, 25, 149-166.

[38] Schweighofer, E., Rauber, A., and Dittenbach, M. 2001. Automatic Text Representation, Classification and Labeling in European Law. In Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001), St. Louis, Missouri, 78-87.

[39] Stumme, G., and Maedche, A. 2001. Ontology Merging for Federated Ontologies on the Semantic Web. In Proceedings of the International Workshop on Foundations of Models for Information Integration (FMII 2001), Seattle, WA, 16-18.

[40] Thompson, P. 2001. Automatic Categorization of Case Law. In Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001), St. Louis, Missouri, 70-77.

[41] Wang, H., Akinci, B., and Garrett, J. 2007. A Formalism for Detecting Version Differences in Data Models. Journal of Computing in Civil Engineering, 21 (5), 321-330.