Mapping Regulations to Industry-Specific Taxonomies

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

Gloria T. Lau
Stanford University
Dept. of Civil & Environmental Eng.
Stanford, CA 94305-4020
glau@stanford.edu

Kincho H. Law
Stanford University
Dept. of Civil & Environmental Eng.
Stanford, CA 94305-4020
law@stanford.edu

ABSTRACT

For each industry, there exist many taxonomies that are intended for various applications. There are also multiple sources of regulations from different government agencies. Industry practitioners, unlike legal practitioners, are familiar with one or more industry-specific taxonomies but not necessarily regulatory organization systems. To help browsing of regulations by industry practitioners, we propose to map regulations to existing industry-specific taxonomies.

A mapping from a single taxonomy to a single regulation is a trivial keyword matching task. From there, we examine techniques to map a single taxonomy to multiple regulations, as well as to map multiple taxonomies to a single regulation. Cosine similarity, Jaccard coefficient and market-basket analysis are tested to model the similarity metric between concepts from different taxonomies. Preliminary evaluations of the three metrics are performed. Examples from the building industry are drawn to illustrate the betterment of regulatory usage 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*, I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems – *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, 24], analysis of regulations [15, 16], and compliance guidance for regulations [12, 13]. Methodologies and tools that enable the *browsing* of regulations according to industry-specific taxonomies are relatively lacking.

ICAIL '07, June 4-8, 2007, Palo Alto, CA USA.

Regulations, like most government information, are organized according to the classification system of the agency rather than the mental models of users [6]. There is a clear need and benefit of traversing regulations using existing industry taxonomies. For instance, in the architectural, engineering and construction (AEC) domain, there are a few ontologies that describe the semantics of building models, such as the CIMsteel Integration Standards (CIS/2) [7] , the Industry Foundation Classes (IFC) [10], and the OmniClass construction classification system (OmniClass, see Figure 1) [22]. These ontologies are all targeted towards the same user group, namely the AEC practitioners, but the structures, vocabularies and coverage differ depending on the application. Most AEC practitioners are familiar with the terms and vocabulary in these ontologies - for them to browse through regulations for compliance requirements, adhering to an existing taxonomy that they 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.

Figure 1: Excerpt from OmniClass Construction Classification System

In this paper, we present a systematic approach to mapping 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, where we reuse the relatedness analysis core from [15] to compute relevancy between 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 map them to one regulation. The natural next step, mapping

multiple taxonomies to multiple regulations, is proposed as a future task.

2. ONE TAXONOMY TO ONE REGULATION

Mapping one taxonomy to one regulation is a simple keyword latching task. We work with taxonomies from both the AEC industry and the environmental protection group to experiment with different types of taxonomies. Figure 2 shows the International Building Codes (IBC) [11] latched with the OmniClass. 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 as a unit of interest. Users can then traverse the taxonomy and browse relevant sections of the regulation.

Extending this mapping from one taxonomy to multiple regulations unfortunately leads to a classic problem of information overload. It results in a Google-like user interface per taxonomy node, where sections from different regulations are interlaced. For instance, the concept "chlorine" maps to over 30 sections each in the Alabama and Arizona drinking water standards. For web content, users quickly become frustrated with information overload, and intelligent retrieval and presentation of web results become the key issue for search [4]. Fortunately, regulatory documents are much more organized 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.

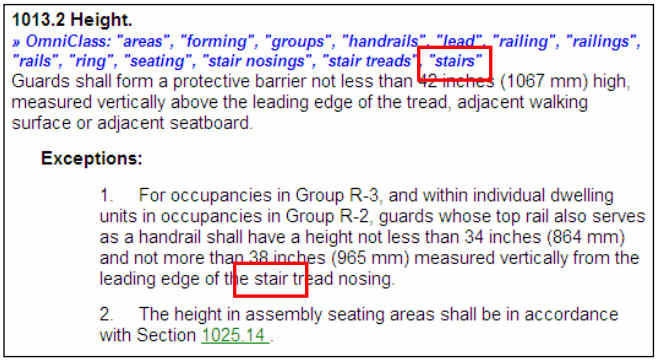

Figure 2: Regulation Latched with Taxonomy Concepts

3. ONE TAXONOMY TO MULTIPLE REGULATIONS

Traversing multiple regulation trees simultaneously using one taxonomy is a challenging problem. It is not uncommon for industry practitioners to be familiar with one particular regulation but not others. For example, architects might be familiar with California state code but not Federal code; nonetheless, some projects might require understanding of both [8]. In this scenario, it is beneficial to map the taxonomy to California code first, and then branch out to recommend related sections from the Federal code. In general, focusing on one regulation as the base for recommendations of further readings from other regulations significantly reduces information overload.

Figure 3 shows an example of linking multiple regulations. After browsing down the taxonomy tree to the concept "chlorine", users are shown a list of matched sections from the Alabama regulation. As illustrated in Section 2, 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 stay focused on the regulation of their choice, and at the same time acquire relevant sections from other regulations as needed. Part of the challenges to developing such a system is the desirable user interface, which is beyond the scope of this work. The remaining challenge lies in the methodologies for making recommendations based on relevancies between sections from different regulations. For this task, we reuse the relatedness analysis core from [15, 16], which compares sections from different regulations based on shared features using a cosine similarity measure. The hierarchical and referential information are taken into account in the comparative analysis as well.

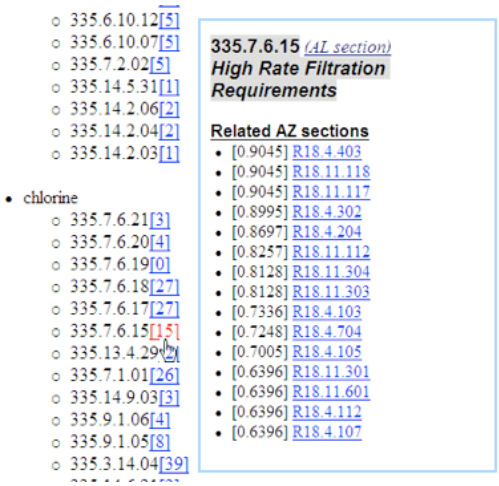

Figure 3: Chlorine mapped to Section 335.7.6.15 in AL code, which have 15 related sections in AZ code

4. 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, there exist multiple taxonomies per industry for different applications. Organizations are interested in translating from one taxonomy to another for various applications [2, 17]. Mapping regulations to a single taxonomy has limited usability of the system. However, traversing regulations using multiple taxonomy trees pose a non-trivial problem. There are much research effort on ontology merging [21, 25], which provides a solution for data interoperability but not as a front-end representation format. Users would need to learn the newly merged ontology in order to browse regulations, which defeats the original intent of using existing taxonomies to help locate regulatory provisions. Using the same argument from Section 3, we believe that focusing on one taxonomy that users are familiar with is a good 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.

Error! Reference source not found. illustrates the proposed system with two taxonomies, the OmniClass [22] and the IFC [10], mapped to the International Building Code [11]. The OmniClass is altered from its original representation, shown in **Error! Reference source not found.**, 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. For example, if the user is interested in "steel decking", the system can help to locate a list of IBC sections that are related to "steel decking", sorted in order of relevance, followed by 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 [18, 19]. It is difficult to interoperate among heterogeneous ontologies for generic web services; however, our problem is slightly more manageable since our ontologies are industry specific and are targeted towards the same group of users. Similar to the techniques presented in Section 3, 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 are meticulously drafted and reviewed for accuracy. Unlike web content, regulations are unlikely to have random co-occurrences of phrases in the same provision.

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 matched in section *k*. In subsequent sections, we will discuss three metrics to compute the similarity score among concepts. In our example shown in **Error! Reference source not found.**, 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".

4.1 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 [14, 20]. 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:

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

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

4.2 Jaccard Similarity Coefficient

Jaccard similarity coefficient [20, 23] is a statistical measure of the extent of overlapping between two vectors. It is defined as the size of the intersection divided by the size of the union of the vector dimension sets:

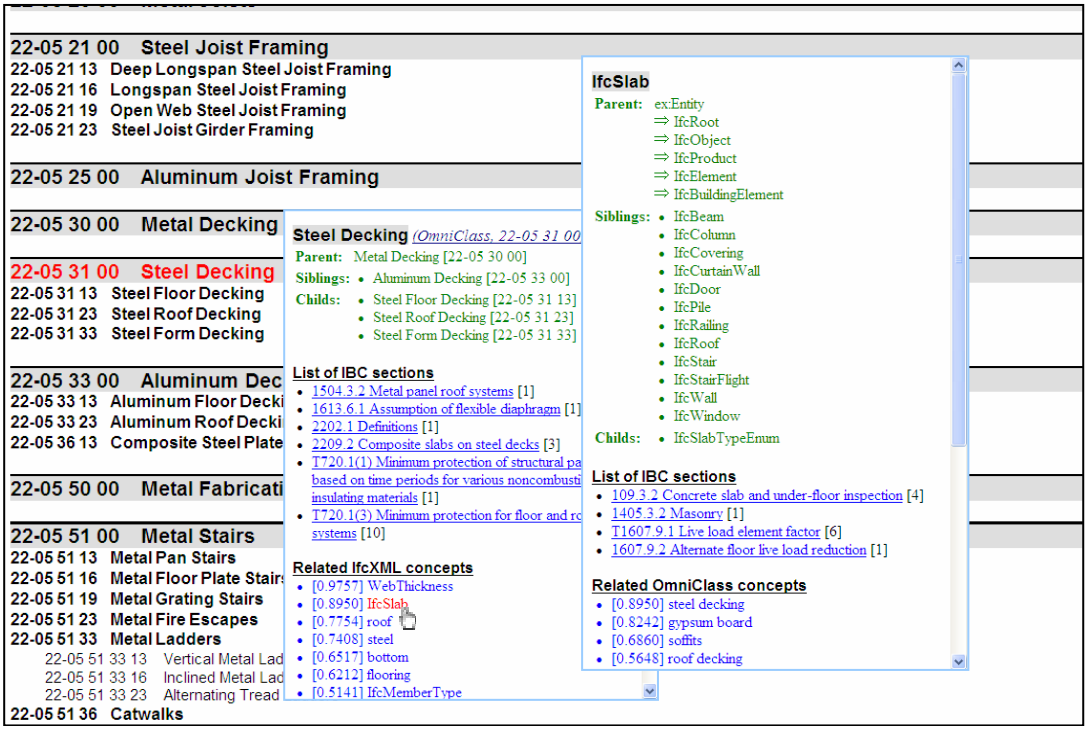

Figure 4: Traversing the IBC using OmniClass Taxonomy with Relevant Concepts from the IFC Taxonomy

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

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.

4.3 Market-Basket Model

Market-basket model is a probabilistic data-mining technique to find item-item correlation [9]. 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. Using the same notations of N_{11}, N_{10}, N_{01} and N_{00} as in Section 4.2, the probability of concept *j* is computed as

$$\Pr(j) = \frac{N_{11} + N_{01}}{N_{11} + N_{10} + N_{01} + N_{00}}$$

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

$$Sim(i, j) = \frac{N_{11}}{N_{11} + N_{01}} - \frac{N_{11} + N_{01}}{N_{11} + N_{10} + N_{01} + N_{00}}$$

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 extreme case where the occurrence of concept *j* is not significant in the corpus, but it appears in every section that concept *i* appears.

4.4 Evaluations of the Metrics

The results of concept matching using the three metrics are compared with the result from domain experts. Twenty concepts are randomly selected from the OmniClass and the IFC hierarchies respectively, and pairwise similarity scores are computed using the three metrics. Root mean square errors (RMSEs) are used to measure the difference between the predicted values and the true values. Precision and recall values are computed to evaluate the accuracy of predictions and the coverage of accurate pairs. Precision measures the fraction of predicted matches that are correct whereas recall measures the fraction of correct matches that are predicted.

Table 1 shows the results of the three metrics compared using the RMSE, precision and recall measures with a similarity score threshold of 0.4. Jaccard similarity is not preferred due to its unacceptably low recall despite a perfect precision. Cosine similarity appears to be average among the three metrics. The market-basket model outperforms the other two metrics in terms of RMSE, and it also produced the highest recall with satisfactory precision.

	Cosine	Jaccard	Market Basket
RMSE	0.1000	0.1300	0.0825
Precision	0.9130	1.0000	0.7955
Recall	0.3559	0.1186	0.5932

Table 1: Evaluation Results of the Three Metrics

5. CONCLUSIONS & FUTURE TASKS

Regulatory documents are written by government agencies who organize the material to suit the needs of the government as well as legal practitioners. From industry practitioners' standpoint, the original hierarchy might not be the easiest retrieval model for regulations. In this work, we propose to map industry-specific taxonomies to regulations to increase usability of regulations 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.

The 1-1, 1-n, and 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 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. Formal evaluations of the similarity metrics and the usability of the system are much needed.

6. 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.

7. REFERENCES

[1] K. Al-Kofahi, A. Tyrrell, A. Vachher and P. Jackson. "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, pp. 88-93, 2001.

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

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

[4] N. Bonnel, V. Lemaire, A. Cotarmanac'h and A. Morin. "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, pp. 209-216, 2006.

[5] S. Brüninghaus and K.D. Ashley. "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, pp. 42-51, 2001.

[6] J.E. Fountain. *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, 2002.

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

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

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

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

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

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

[13] S. Kerrigan and K. Law. "Logic-Based Regulation Compliance-Assistance," In *Proceedings of the 9th International Conference on Artificial Intelligence and Law (ICAIL 2003)*, Edinburgh, Scotland, pp. 126-135, Jun 24-28, 2003.

[14] B. Larsen and C. Aone. "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, pp. 16-22, 1999.

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

[16] G. Lau, K. Law and G. Wiederhold. "Legal Information Retrieval and Application to E-Rulemaking," In *Proceedings of the 10th International Conference on Artificial Intelligence and Law (ICAIL 2005)*, Bologna, Italy, pp. 146-154, Jun 6-11, 2005.

[17] R. Lipman. "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, pp. 3087-3096, Jun 14-16, 2006, 2006.

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

[19] P. Mitra and G. Wiederhold. "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, pp. 45-50, 2002.

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

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

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

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

[24] E. Schweighofer, A. Rauber and M. Dittenbach. "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, pp. 78-87, 2001.

[25] G. Stumme and A. Maedche. "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, pp. 16-18, 2001.