Regnet: An Infrastructure for Regulatory Information Management and Compliance Assistance

Shawn Kerrigan, Charles Heenan, and Kincho H. Law
Department of Civil and Environmental Engineering
Stanford University
Stanford, CA 94305-4020
Email Contact: law@ce.stanford.edu

1 Abstract

The Regnet Project aims to develop a formal information infrastructure for regulatory information management and compliance assistance. This paper discusses three basic milestones of current research and development efforts. The first is the creation of a document repository containing federal and state regulations and supplemental documents. This repository includes a suite of concept hierarchies that enable users to browse documents according to the terms they contain. The second is an XML framework for representing regulations and associated metadata. The XML framework enables the augmentation of regulation text with tools and information that will help users understand and comply with the regulation. The third milestone is the creation of a compliance assistance system built upon the XML framework. The prototype effort for the document repository has been focused on environmental regulations and related documents. The compliance assistance system is illustrated in the domain of used oil management.

2 Document Repository

A primary objective for the Regnet information infrastructure has been the development of a document repository for environmental regulations. Currently, this document repository contains 80 megabytes of text-based content downloaded for research purposes from the Environmental Protection Agency's (EPA) website and the Westlaw online legal information system. A complete environmental regulation-oriented document repository must contain not only regulation text but also the supporting documents for all sections of that text. Supplemental documents are important because they often contain information that is necessary for the accurate interpretation of the regulation(s) to which they refer. Since the compliance assistance infrastructure uses the "used oil" regulations as its demonstration domain, the document repository contains supplemental documents dealing with used oil.

In the repository, more than 80% of the text consists of the Code of Federal Regulations, Title 40: Protection of the Environment (40 CFR). The remaining 20% consists of supplemental documents related to the management of used oil. These supporting document types include the preamble to the regulation text found in 40 CFR 261 and 279, administrative decisions, guidance documents, federal cases, letters from the general counsel of the EPA, as well as letters of interpretation from the EPA.

The contents are available through the mediation of one or more searchable concept hierarchies. A commercial software package from Semio Corporation is used to populate these concept hierarchies with documents according to the terms they contain. Documents receive topical XML metatags based upon which categories they inhabit. To date, the repository text has been categorized according to an alphabetical index of terms, according to the EPA's list of extremely hazardous substances, and according to geography (U.S. States). Through the Regnet Project website, visitors can use these categorization structures to search the Regnet document repository for texts that address a common topic even when those texts do not explicitly reference one another (See Figure 1).

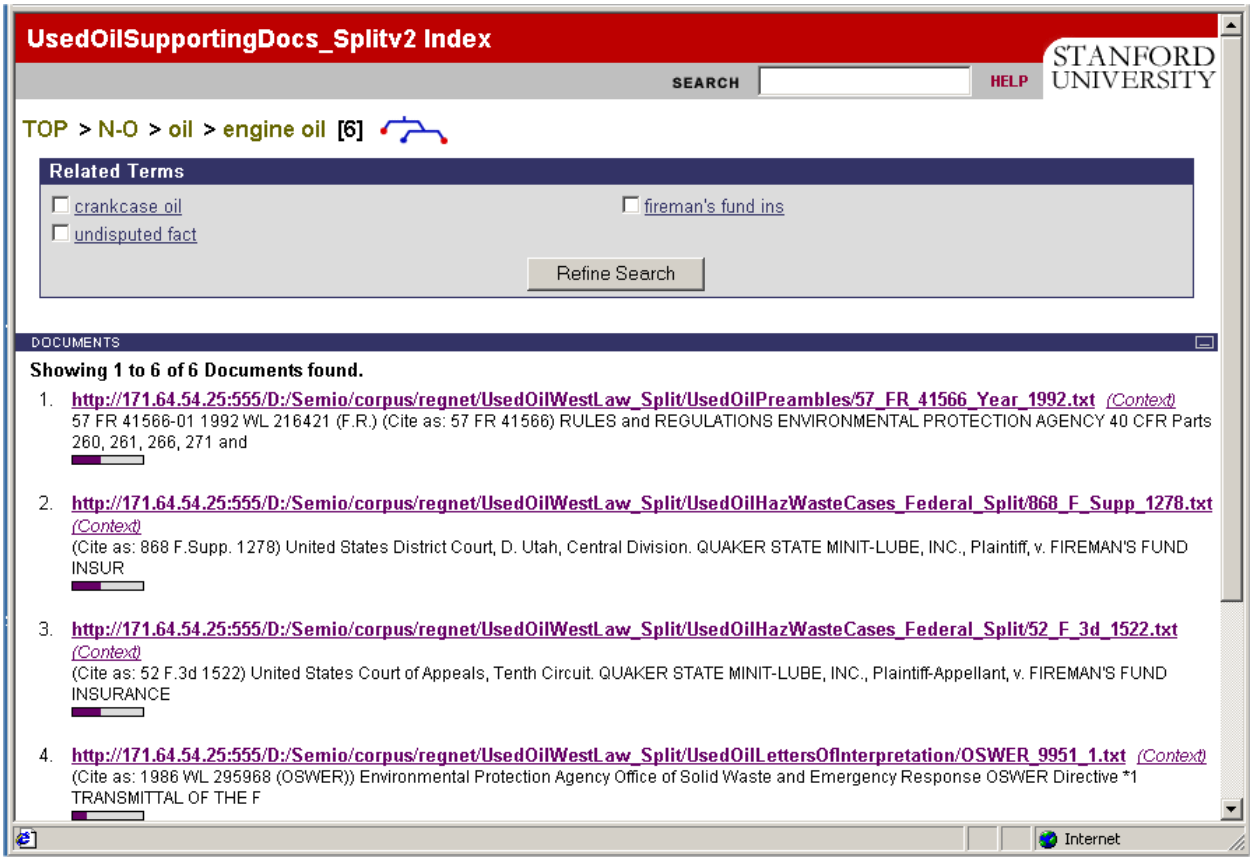

Figure 1: Results of browsing the term "engine oil" in a concept hierarchy.

3 XML Framework and Metadata

The second milestone for the Regnet Project is the development of an XML framework for environmental regulations. This framework enables the augmentation of a regulation with various types of regulation-specific metadata. This section discusses why XML was chosen for the regulation framework, how the framework is structured, and outlines each type of metadata that has been developed.

3.1 Why XML?

XML was chosen as the representational formalism with which to store and annotate regulations. The reason is that XML is an open standard that allows designers to create any element needed to structure information in a given file. With XML, it is possible to structure the information in a document according to its conceptual meaning rather than simply according to how it should be displayed. Since XML offers the expressive capabilities of HTML along with the ability to specify information about conceptual content (Navarro, 2000), it is well suited to the task of structuring and augmenting regulation text with content-related metadata.

The Regnet XML-based regulation framework includes XML tags for each level of regulation text, such as subpart, section or subsection. This structure mirrors the standard structure of federal environmental regulations. A parsing system was built to transform federal environmental regulations from Portable Document Format (PDF) into the Regnet XML format. The parser creates the core XML structure and populates each level of the framework with the appropriate level of regulation text.

3.2 XML-Metadata

Once the Regnet XML-based regulation framework has been populated with regulation text, it is possible to augment the regulation with metadata about the regulation provisions by inserting new XML elements into the document. Four such metadata types that have been added to the Regnet regulation framework are concept tags, reference tags, definition tags, and logic and control processing tags.

One form of metadata applied to the Regnet XML framework is concept metadata. A program was written to automatically add concept elements to a regulation using a two-step process. First, using the Semio software package, terms in a given provision are identified by comparing all regulation text against an alphabetical index as the controlling hierarchy. Second, XML concept elements for each term are inserted at each level of the regulation framework in which the concepts appear.

The addition of concept tags to the Regnet XML framework allows for the development of viewing systems that can dynamically generate links to related supporting documents in the document repository. This is useful because supporting documents and regulations may not directly reference each other even when they address the same topic. The automatic application of concept tags to the XML framework means that as new supporting documents are added to the document repository, regulations stored in the framework can automatically be linked to them via the terms that they share in common.

A second type of metadata applied to the Regnet regulation framework concerns references embedded in the regulation text. Regulation provisions tend to contain a large number of casual English references to other provisions. These references are cumbersome to look up manually. Moreover, they reduce the readability of the regulation text itself. Ideally, these references should be automatically linked in a manner similar to hyperlinks so that readers of the regulation documents can follow the regulations more easily.

Regulation references range in complexity from relatively straightforward to complex. An example of a straightforward casual English reference is the text “as stated in 40 CFR section 262.14(a)(2).” An example of a more complex reference is the text “the requirements in subparts G through I of this part” (where the current part is part 265). This latter example could be converted manually into the following list of complete references: ref.40.cfr.265.G, ref.40.cfr.265.H, and ref.40.cfr.265.I. However, given the large volume of federal and state environmental regulations, such manual translation of references is too time consuming to be practical.

To solve this problem, a parsing system was developed using a context-free grammar and a semantic representation/interpretation system that is capable of tagging regulation provisions with the list of references they contain. These references are not tagged as hyperlinks, which tie the reference to a particular source for the referred document. Rather, the reference tags provide a complete specification for *what* regulation provision is referenced. *Where* the regulation is located is not specified so that an XML regulation viewing system may select any document repository of regulations from which to retrieve the referenced provision.

A third type of metadata added to the Regnet XML regulation framework makes word and acronym definitions available to a regulation viewing system. The large number of domain-specific terms and acronyms that appear in regulations can make regulation text difficult for novices to understand. The ability to add definition metadata to the Regnet regulation framework allows a regulation viewing system to incorporate explicit definitions of terms and acronyms into its user interface.

Logic and control processing metadata is the fourth type of metadata that has been added to the Regnet regulation framework. Logic metadata comes in two variations. There is only one form of control processing metadata. Regulation logic metadata represents a rule or concept from a regulation using First Order Predicate Calculus (FOPC) logic sentences. These logic sentences are used to represent the rules that must be followed for an entity to be in compliance with the regulations. User interface logic metadata uses FOPC logic sentences to represent compliance questions and a list of possible user answers

to those questions. Control processing metadata provides information about what provisions of a regulation need to be checked for compliance. Each type of logic or control processing metadata can be associated with any regulation provision in the document. In the Regnet framework, these three types of metadata are necessary for the system to be able to verify compliance with a regulation. However, they must be specified by a domain expert as they cannot be generated automatically. For the purposes of demonstration, some used oil provisions of 40 CFR (40 CFR 279) were manually tagged with regulation logic metadata, with user-interface logic metadata, and with control processing metadata.

4 Compliance Assistance System

The third milestone for the Regnet Project is the development of a web-based compliance assistance system that is built upon the XML regulation framework and that takes advantage of the regulation metadata described in the previous section.

One feature of this web-based compliance assistance system is that it helps guide the user through the regulations covering the management of used oil (40 CFR 279). The user selects a regulation provision to check against for compliance. At each step, the system asks the user for input that helps the system clarify whether the user is in violation of the provision or whether the user is probably in compliance. When the system completes the check against the regulation provisions or detects a conflict between the user's answers and the regulation, it displays a summary of the question-and-answer history as well as the results of the compliance check. This functionality is implemented through a web interface implemented on top of a compliance checking component (RCCsession).

The RCCsession component controls the process used to check for violations. It begins by parsing the XML-structured regulation to extract the information necessary to run a compliance check against the document. This information includes the logic metadata as well as the control processing metadata. The RCCsession follows the control processing metadata in the XML-regulation structure and manages lists of regulation rules and the associated user responses. Requests for user input passed back to the web interface. A user response is converted into the corresponding FOPC user-interface logic sentence for that response. Whenever the RCCsession component has a set of logic sentences that it needs to check for contradictions, it writes an input file containing those logic sentences and passes that file to Otter, an automated-deduction program (Wos, 1999).

The results produced by Otter are stored in an output file. The RCCsession reads this output file to see if it contains a proof and then takes whatever actions necessary to continue the compliance checking procedure. In this manner, the RCCsession controls the flow of processing while using Otter to check for logic contradictions in the background. The overall system design is such that any FOPC theorem prover can be used to perform the logic checks for the RCCsession component. Figure 2 shows the organization of the compliance assistance system, which includes a web interface (RASweb), the RCCsession component, and the Otter automated-deduction program.

An example of the user interface to the compliance assistance system is shown in Figure 3. In order to facilitate greater understanding of the regulations, the system makes available a number of enhancements while guiding the user through a compliance check. The system can automatically insert hyperlinks to any referenced regulation provisions. If selected, these provisions will open in a new window. The system can display in green text any terms that have associated definition metadata, with "mouse-over" support for popup windows that explain the terms. Further, key conceptual phrases for the provision can be displayed and linked, enabling instant access to repository documents related to the provision being viewed. Finally, the user can choose to view the regulation text at a range of levels, so as to gain an understanding of the context for the current compliance question being asked. For example, the user may view the regulation text at the subpart, section, or current provision level.

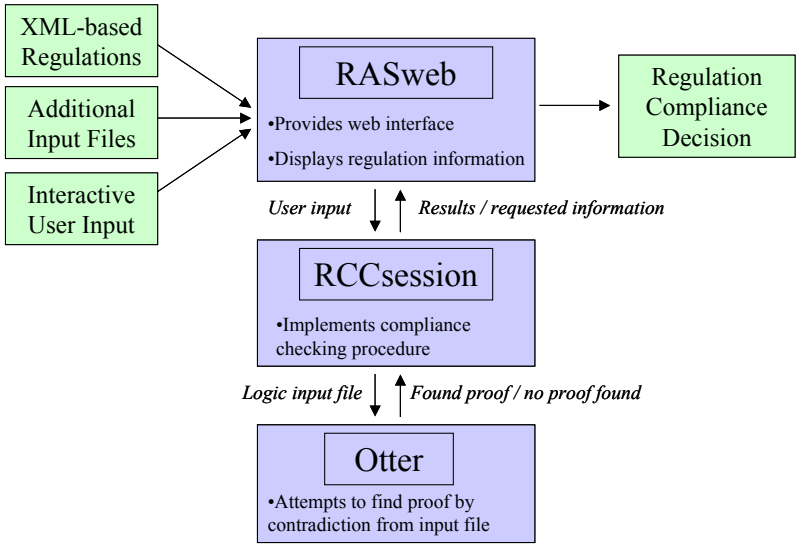

Figure 2: Components of the compliance assistance system.

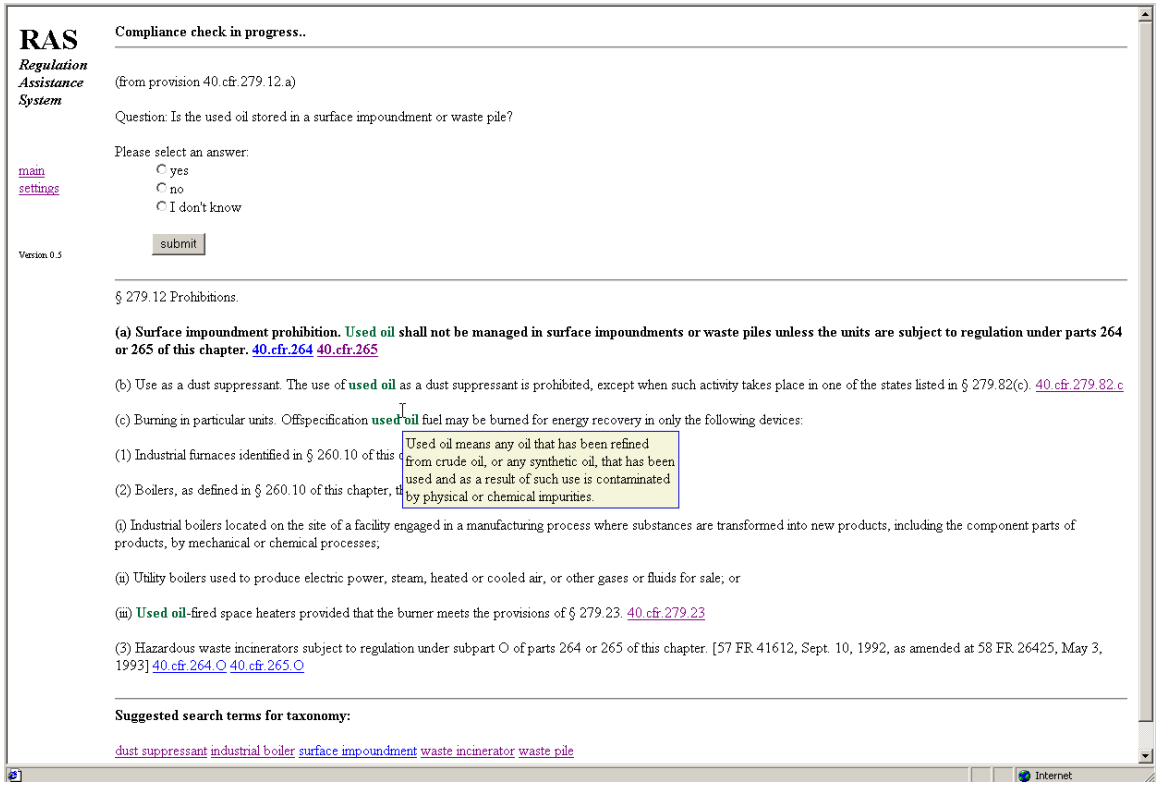

Figure 3: User interface to the compliance assistance system.

5 Related Work

There has been a great deal of work done on building expert systems for law. In the past thirty years great strides have been achieved in advancing expert system (Wahlgren, 1992; Zeleznikow, 1994) and theorem-proving technology (Wos, 1999). Research for new formalisms and specialized logics

(Greiner, 2001; Shanahan, 1997) continue to improve reasoning speed and non-monotonic reasoning capabilities. Much of the earlier work in IT and law focused on building systems to optimize decisions with respect to laws, particularly tax law (McCarty, 1977). Some of the recent work has focused on investigations into case-based reasoning and information retrieval (Brüninghaus, 1997; Stranierim, 1999). The work most related to this research is that of Christopher Royles at University of Liverpool (Royles, 1998).

6 Summary

The goal of the Regnet Project is to develop an information infrastructure for regulatory information management and compliance assistance. This paper briefly describes some of the work that has been done to date on creating a regulation document repository, on developing an XML-based framework for representing regulations and associated metadata, and on creating a compliance assistance system built upon the Regnet XML framework.

7 Acknowledgements

This research project is sponsored by the National Science Foundation, Grant Numbers EIA-998368 and EIA-0085998. The authors would like to acknowledge the support by Semio Corporation in providing software for this research. The authors would like to thank professors Gio Wiederhold, Barton Thompson and Jim Leckie for their valuable suggestions in this project.

8 References

Brüninghaus, S., Ashley, K., 1997. Finding Factors: Learning to Classify Case Opinions under Abstract Fact Categories, Proceedings of the Sixth International Conference on Artificial Intelligence and Law, 123-131.

Greiner, R., Darken, C., Santaso, N. I., 2001. Efficient Reasoning, ACM Computing Surveys, ACM 33, 1, 1-30.

McCarty, T., 1977. Reflections on Taxman: An Experiment in Artificial Intelligence and Legal Reasoning, Harvard Law Review.

Navarro, A., White, C., Burman, L., 2000. Mastering XML, SYBEX Inc.

Royles, C. A., Bench-Capon, T. J. M., 1998. Dynamic Tailoring of Law Related Documents To User Needs. DEXA Workshop 1998: 609-613.

Shanahan, M., 1997. Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia, MIT Press.

Stranierim, A., Zeleznikow, J., 1999. The evaluation of legal knowledge based systems, Proceedings of the Seventh International Conference on Artificial Intelligence and Law, 18-24.

Wahlgren, P., 1992. Automation of Legal Reasoning, Kluwer Law and Taxation Publishers.

Wos, L., Pieper, G., 1999. A Fascinating Country in the World of Computing, World Scientific Publishing Co. Pte. Ltd.

Zeleznikow, J., Hunter, D., 1994. Building Intelligent Legal Information Systems: Representation and Reasoning in Law, Kluwer Law and Taxation Publishers.