A SOFTWARE INFRASTRUCTURE FOR REGULATORY INFORMATION MANAGEMENT AND COMPLIANCE ASSISTANCE

A DISSERTATION
SUBMITTED TO
THE DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Shawn L. Kerrigan

August 2003

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Kincho H. Law
(Principal Adviser)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

James O. Leckie

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Barton H. Thompson, Jr.

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Gio Wiederhold

Approved for the University Committee on Graduate Studies.

Abstract

There is a great deal of information available online regarding environmental regulations, as well as supplementary documents associated with the regulations. The sheer volume and complexity of this information, coupled with its scattered distribution across many different sources, makes any attempt to understand and interpret the information a daunting task. Other factors, such as the high density of cross-referencing between regulatory documents and the heavy reliance on acronyms, also contribute to reducing the readability of the documents. Since environmental regulations have the force of law, it is important that the regulated community be able to locate, understand, and comply with them. It is also advantageous for society to make these regulations as easy to locate and understand as possible so that the environment is protected to the extent provided by the law.

Currently, environmental regulation compliance checking is largely a paper-based process. Where modern information technology has been utilized, it has generally been used simply to make available online versions of the paper-based guides and forms. Our vision for the regulation compliance process is to have organized and up-to-date regulatory information and compliance assistance procedures available over the Internet. Towards that end, we seek to develop information management frameworks that can facilitate public access to regulations and that can also facilitate the compliance process. This will help improve the completeness of regulatory documentation available to interested parties, and will also help resolve the issue of knowing when one's research on a regulatory topic is complete. Information management frameworks may also improve the transparency of compliance requirements through the use of clear presentation and linking. Transitioning the information technology used in environmental regulatory environments from the current state of online forms and scattered documentation to a state where interactive systems and organized documentation are available online could potentially have a significant positive effect on the rate of compliance among businesses.

This thesis addresses the problem of regulation compliance by developing a formal information infrastructure for regulatory information management and compliance assistance. There are three main contributions made in this thesis. First, a document repository containing regulations and supplemental documents is designed to facilitate gathering, storing, and categorizing these regulatory documents in order to make them more accessible. This repository includes a suite of concept hierarchies that enable users to browse documents according to the terms they contain. Second, an XML framework is proposed to structure the representation of regulations and the 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. Third, an Internet-enabled regulation assistance system is built that can guide users through regulation requirements to help them determine if they are in compliance, and also identify relevant supplementary documents. In addition, it is shown that the system can be used as a component in online industry-specific compliance guides.

Acknowledgments

The debts that I have accumulated during my five years at Stanford are numerous. I would like to thank some of the people who have provided me with assistance over the years. First, I would like to thank my family. Without their support and encouragement I never would have made it to Stanford. Their encouragement over the past several years helped sustain me through the ups and downs of conducting research work. I feel very lucky to have such a wonderfully supportive family.

My deepest thanks go to my principal thesis advisor, Professor Kincho Law, for his guidance and support throughout my graduate career at Stanford. His dedication to helping students identify and pursue their research interests has made this thesis possible. Over the past five years I have learned a tremendous amount from him about both research and life, and I am grateful to have had the opportunity to work with him.

I would like to thank Professors James Leckie, Barton H. Thompson, Jr., and Gio Wiederhold for their support and advice throughout this research project. The research presented in this thesis is an interdisciplinary work, and I have needed to learn a great deal in the areas of environmental engineering, law, and computer science to complete this research. Each of Professors James Leckie, Barton H. Thompson, Jr., and Gio Wiederhold provided significant support in their respective areas of expertise that helped the research presented in this thesis come together. In addition, I would like to thank

Professor Hector Garcia-Molina for chairing my thesis defense committee on short notice.

I would also like to thank the other members of Professor Kincho Law's Engineering Informatics Group (EIG) for their support as fellow researchers and friends. I am particularly indebted to the EIG members with whom I worked most closely on the research work presented in this thesis: Charles Heenan, Gloria Lau, Pooja Trivedi, Liang Zhou, and Haoyi Wang. All the members of the Engineering Informatics Group have contributed in some way to my research work at Stanford, and I would also like to thank them all for their support: Jun Peng, David W. Liu, Jerome P. Lynch, Chuck Han, Jie Wang, Jinxing Cheng, Bill Labiosa, Yang Wang, Xiaoshan Pan, and Arvind Sundararajan. Working with this talented group of researchers truly enriched my experience at Stanford, and I am grateful for having had the opportunity to get to know all these wonderful people.

I am also indebted to the numerous members of the regulatory and regulated communities who took time out of their busy schedules to meet with me and provide feedback on my research work. Some of the people I owe a special thanks to are Cheryl Nelson, Robert Parkhurst, Phil Bobel, Rick Ferguson, Gordon Blancher, Ken Torke, Larry Gibbs, Ole Christensen, and Ned Black.

This research is sponsored by the National Science Foundation, Grant Numbers EIA-9983368 and EIA-0085998. I would also like to acknowledge an equipment grant from Intel Corporation and software support from Semio Corporation. Finally, I would like to thank the Stanford Graduate Fellowship program for showing confidence in my abilities as a researcher by providing me with three years of financial support for graduate studies when I initially started at Stanford.

Table of Contents

List of Tables

List of Figures

Number *Page*

Chapter 1

Introduction

1.1 Motivation

There is a great deal of information available online regarding environmental regulations, as well as supplementary documents associated with the regulations. The sheer volume and complexity of this information, coupled with its scattered distribution across many different sources, makes any attempt to understand and interpret the information a daunting task. Other factors, such as the high density of cross-referencing between regulatory documents and the heavy reliance on acronyms, contribute to reducing the readability of the documents that can be located. Since environmental regulations have the force of law, it is important that companies be able to locate, understand, and comply with them. It is also advantageous for society to make these regulations as easy to locate and understand as possible so that the environment is protected to the extent provided by the laws in place.

The burden of complying with environmental regulations can fall disproportionately on small businesses, since these businesses may not have the expertise or resources to keep track of regulations and their requirements [79]. That the requirements of these complex regulations change over time further compounds the problem [93]. As noted in the Washington Post, "Deciphering and complying with federal regulations is a legal and paperwork nightmare for many businesses. To keep pace, some hire consultants – sort of regulatory accountants – to keep track of the applicable health, safety, environmental and equal-opportunity rules" [91]. This burden has been recognized and targeted by legislation designed to address the problem. Through the Regulatory Flexibility Act (RFA) [80], amended by the 1996 Small Business Regulatory Enforcement Fairness Act (SBREFA) [92], the United States Environmental Protection Agency (EPA) has a commitment to take into account the burden environmental regulation can place on small businesses. Among many other requirements, SBREFA requires the EPA to publish Small Entity Compliance Guides that are written in plain language, support the rights of small entities in enforcement actions (e.g., reducing civil penalties for violations), and provide Congress and the General Accounting Office with copies of all final rules and supporting analyses [81]. This act clearly recognizes the information problem facing businesses, particularly small businesses, that must comply with environmental regulations.

The United States Environmental Protection Agency was formed in 1970 to assume management of a variety of federal programs targeting the environment. At the time, the nation was faced with major environmental issues on a number of fronts – air, water, and land. The EPA merged 15 different agencies, or parts of agencies, into one entity to address the environmental issues. In the early days, the EPA focused on enforcement actions to reduce pollution in major cities and industries [84]. More recently, the EPA has placed an increased emphasis on compliance assistance, rather than enforcement actions, to increase the rate of compliance with environmental regulations.

One of the EPA's primary tasks is to develop regulations that implement statutes passed by Congress, which govern the regulated community and protect the environment. Over time, the regulations have become increasingly complex and difficult to comprehend. As

Dawson and Davies noted in an environmental law book review, “Complex, ever-growing, and oft-adapting to the social, political, biophysical, and economic influences it faces, American environmental law in 2000 is a giant leap away from its beginnings of the late-1960s and early-1970s. … With such breadth, depth, and complexity, understanding environmental law is becoming more challenging for practitioners and the judiciary alike.” [30].

Some of the reasons why the current regulatory system has evolved and how the current regulatory system has a number of drawbacks were discussed by Richard Stewart in a recent law review article. Two paragraphs from this article illustrate why new information tools for working with regulations are becoming a necessity [95]:

> “The U.S. environmental regulatory system has contributed substantially to reducing or limiting increases in air and water pollution and toxic waste problems, and has also furthered natural resource protection and preservation. … Despite its accomplishments, however, the U.S. environmental regulatory system suffers from a number of well-known shortcomings, including fragmentation, rigidity, complexity, and high compliance and administrative costs. These deficiencies were of less importance in the early stages of environmental regulation, when it was imperative to halt and reverse rising levels of pollution and hazardous waste, clean up extremely hazardous waste dumps, and halt highly destructive ecosystem alteration. It was concluded that only the federal government could ensure that these urgent needs would be met. … A series of centralized command-and-control regulatory programs aimed at particular types of environmental problems were established through separate statutes enacted by Congress in piecemeal fashion. Command regulation targeted on major facilities and development projects promised and often delivered effective action. The inherent inefficiencies of the command system were not apparent or of much concern because the means of reducing pollution and waste were obvious and controls were relatively cheap to implement. Different statutes were enacted for the control of pollutants and

> wastes discharged into different media and each such statute contained a variety of separate provisions aimed at different types of sources or problems with little or no attempt at overall consistency or coordination. The resulting fragmentation and lack of coordination in the overall regulatory effort were of little concern because it was thought important to target controls on the most obvious and accessible environmental problems quickly rather than devote the time and effort necessary to construct an integrated regulatory system.
>
> The situation that we face now and for the future is in many important respects quite different. Major sources of pollution and wastes are already tightly controlled. Further reductions will be quite costly and require significant advances in technologies and the organization of production, distribution, consumption, and treatment of post-consumer residuals. In order to sustain further environmental progress in the face of continued economic growth, not only must additional reductions be obtained from major sources, but discharges from small, non-point or area sources must be significantly curtailed, including those in the consumer, services, and agricultural sectors."

The shift in target of modern regulation towards smaller sources of environmental pollution, coupled with the lack of an easily comprehensible regulatory structure sets the stage for the research work presented in this thesis. Smaller entities, which often lack the expertise to research environmental regulatory requirements using current tools, will benefit from the application of modern information technology to develop tools that help them identify what they need to do in order to be in compliance with all applicable regulations.

With more tools to assist businesses in understanding compliance requirements, greater compliance with environmental regulations could be achieved. In a law review article on the foundations of environmental law, David Spence argued that, "… there is ample reason to believe that most businesspeople are guided by the desire to comply with environmental laws because (1) they consider themselves law-abiding citizens, (2) they value environmental protection, (3) they seek to avoid the public condemnation that accompanies non-compliance, or (4) all of the above" [93]. He further argued that "there is ample reason to believe that a growing percentage of environmental violations result from a misunderstanding of regulatory requirements or are otherwise unintended." Information technology can potentially and beneficially be applied to address these issues, helping people to better understand regulatory requirements, and therefore reducing environmental regulation violations. This thesis describes research work that investigates how information technology can be used to help people better understand regulatory requirements.

1.2 Current Compliance-Assistance and Vision for the Future

The current state of environmental regulation compliance checking is largely a paper-based process. Where modern information technology has been utilized, it has generally been used simply to make the paper-based guides and forms available online. The compliance process varies significantly between industries. Even within an industry, the process can vary from company to company depending upon its size, product line or service offerings, and the type of expertise it maintains on staff. However, the significant drain on resources that results from spending the time and money necessary to find and interpret environmental regulations has been a common theme and concern among the businesses interviewed for this research work.

Earlier research by Jie Wang that laid the groundwork for this project has identified a clear need for greater use of information technology in environmental regulation [106]. Through interviews with semiconductor companies, waste treatment storage and disposal facilities (TSDF), and law firms, it was found by Wang that, "All parties agree that a new automated information management system capable of conducting hazardous regulation compliance checking over the Internet is necessary" [106].

In terms of process, a large computer manufacturing company that maintains staff trained as environmental compliance experts faces many of the same problems as a small auto repair shop. Both must deal with the difficult task of tracking down all the necessary regulations that apply to their business. Both must identify any additional documents and information associated with the regulations that might affect the interpretation of the regulation. Both must examine carefully the regulations, trying to understand the terminology and develop an interpretation of the regulation requirements. The larger companies have more resources – person-hours, tools, money, and expertise – to dedicate to these tasks, but many of the problems they must deal with are the same as those faced by the small businesses.

Businesses often resort simply to searching the Internet for documents that might help them understand what regulations apply to them, and what the requirements of these regulations are [26]. Businesses that were interviewed for this research work often keep a set of agency websites that they would comb through looking for relevant information, and they would augment this set of websites with searches on web search engines such as Google. A better-equipped company might also have access to specialized legal information retrieval tools, such as Westlaw or LexisNexis, to help identify relevant documents.

Both large and small companies have the problem of not knowing when their regulatory research is complete, or if the material they are working with is the most up-to-date. They also have difficulties in determining if their interpretation of the regulation requirements is correct. Most of the companies interviewed for this project would make a good-faith effort to track down relevant regulatory information, and would then make a good faith effort at interpreting the requirements of the regulation. All of the companies that were interviewed felt that locating and interpreting the relevant regulations is a complex and difficult task, even with the help of supplementary information.

Our vision for the future of the regulation compliance process is to have organized and up-to-date regulatory information and compliance-checking assistance procedures available online. Towards that end, we seek to develop information management frameworks that can facilitate public access to regulations and that can also facilitate the compliance process. This will help improve the completeness of regulatory documentation available to interested parties, and will also help resolve the issue of knowing when one's research on a regulatory topic is finished. Information management frameworks may also improve the transparency of a regulation's compliance-requirements through the use of clear presentation and linking. Transitioning the information technology used in environmental regulatory environments from the current state of online forms and scattered documentation to a state where interactive systems and organized documentation are available online could have a significant positive effect on the rate of compliance among businesses.

1.3 Current State of E-Government

Digital government, electronic government, or e-Government, are synonyms that refer to using information technology, particularly Internet-based technology, to improve government services. This section discusses the current state of information systems for improving e-government. The growing need to address information problems in the environmental regulatory domain has coincided with new research on the application of information technology to the functioning of the government [76-78], so there is a spectrum of work that could be addressed. This section will focus on some of the most relevant work that has been done.

1.3.1 Practice in Government

This section describes the current state of practice for applying information technology to the communication of government policies and laws. The application of information technology can be focused on a number of different types of interactions, including government to citizen, government to business, or government to government. Typically these applications are in the form of web portals to organize information available online. Some of these portals contain custom-built expert systems to help users find relevant information. A selection of these portals and systems are described in this section.

1.3.1.1 Government to Citizen

There are a number of web portals dedicated to improving the interaction between the government and citizens. GovBenefits.gov[1] aims to assist citizens in identifying all government benefits and services for which they may be eligible. The GovBenefits.gov project is a partnership of many government agencies and organizations, managed by the Department of Labor. The site uses a series of increasingly narrow questions to determine a list of programs for which a user may be eligible. While the site currently focuses on federal benefits, there are plans to expand coverage to state benefits as well.

The Recreation One-Stop project is a data sharing initiative by the Department of the Interior to build a portal, Recreation.gov[2], for citizens interested in information on government recreation sites. This portal provides information on a variety of government-managed recreation sites, along with links to any associated websites. The database upon which the portal operates is freely available for download as an XML document, so as to establish a data standard and facilitate data sharing.

[1] The web address for this portal is http://www.govbenefits.gov.

[2] The web address for this portal is http://www.recreation.gov.

FirstGov.gov[3] is an initiative to build a gateway to all government information and services available electronically. Since this U.S. Government portal went online on September 22nd, 2000, it has grown to include searchable access to over 186 million web pages from a number of government sources: federal and state governments, the District of Columbia, and U.S. territories. The portal links users to a number of gateways for citizens, businesses, and government employees. This portal addresses the problem of fragmented information being offered by a multitude of government agencies. The services offered to citizens cover a wide range, including finding government benefits, applying for government jobs, renewing driver's licenses, and applying for social security.

EPA.gov[4] is the primary EPA portal on the web. In addition to providing regulatory information to businesses, this website provides a number of features for citizens online. For example, information on a variety of emissions and environmental data is available from dynamically generated maps through the EnviroMapper system. The EPA website provides a broad coverage of environmental and regulatory information that citizens can use to learn more about environmental issues and what the government is doing to address them. In addition, the website provides a sub-portal for kids and teachers that provides environmental information in a less-technical form designed to be more accessible for children.

1.3.1.2 Government to Business

Some e-government projects that target government to citizen interactions also target government to business interaction. Among the web portals of note that cover both of these areas are the EPA.gov and FirstGov.gov websites.

[3] The web address for this portal is http://www.firstgov.gov.

[4] The web address for this portal is http://www.epa.gov.

Much of the information available on the EPA.gov website that benefits citizens may also be useful to businesses. In addition, a large component of the EPA website addresses businesses in particular. The website provides access to a great deal of information on statutes, regulations, proposed rules, and guidance documents. There is also a directory containing over 800 different topics, and numerous links to partner websites that provide additional services such as compliance assistance information.

Another project managed by the EPA is the E-Rulemaking project Regulations.gov[5]. This project is designed to increase and improve participation in federal rulemaking by citizens, businesses, and other entities. In addition, the EPA hopes the system will streamline rulemaking processes and improve efficiency of the agency's internal processes. Through Regulations.gov, it is possible to search for proposed regulations, review the relevant documents, and submit comments on the proposed rules. In a recent E-Government Strategy publication, the Bush administration has gone so far as to state, "… with the implementation of the E-Rulemaking initiative, businesses will no longer need the assistance of a lawyer or lobbyist to participate in the regulatory process" [20].

The e-CFR[6] project, managed by the National Archives and Records Administration's (NARA) Office of the Federal Register (OFR) and the Government Printing Office (GPO), is a prototype project to provide the Code of Federal Regulations in an electronic format. The e-CFR allows users to view an up-to-date version of the frequently updated Code of Federal Regulations. Since printed regulations rapidly become out of date, providing current regulations online is a significant improvement.

FirstGov.gov provides a gateway for businesses and non-profits. The business section of the FirstGov.gov portal offers a range of information and services. Information provided by FirstGov.gov includes business data and statistics, how to buy and sell to the government, and financial assistance and tax information. Services provided online

[5] The web address for this portal is http://www.regulations.gov.

[6] The web address for this portal is http://www.access.gpo.gov/ecfr/.

include filing wage reports, getting an exporters license, or registering for an Employer ID Number.

Export.gov[7], managed by the Department of Commerce, is a web portal designed to provide access to all export-related assistance and marketing information provided by the U.S. Federal government. It is particularly targeted at small and medium-size businesses interested in conducting businesses abroad. The Export.gov portal provides information on all stages of the export process, along with hyperlinks to related information (e.g., agricultural or other export statistics) and agencies.

BusinessLaw.gov[8], managed by the United States Small Business Administration, is a web portal for businesses, particularly small businesses, designed to provide an online guide to legal and regulatory information. The portal is a gateway to federal, state and local information relevant for businesses. In addition to "plain English guides" and standard search tools, BusinessLaw.gov provides a number of interactive tools and decision trees to help users find information.

1.3.1.3 Government to Government

E-government initiatives targeting government to government interactions generally aim to improve coordination between various government agencies or levels of government (i.e., federal, state or local governments). Similar to the initiatives described in the government to citizen, or government to business sections, the initiatives described here address more than one audience, but the primary target is government to government interactions.

[7] The web address for this portal is http://www.export.gov.

[8] The web address for this portal is http://www.businesslaw.gov.

The GeoSpatial One-Stop initiative, managed by the Department of Interior, is developing the geodata.gov[9] portal to simplify the ability of all levels of government to access and share map-related data. This portal may increase government efficiency by consolidating redundant data and improving inter-government coordination. Users can publish geospatial data, search for geospatial data, browse maps, and sign up for email notifications when new data is added that satisfies user-specified criteria. This service is useful for a wide variety of government activities such as disaster management, environmental protection, planning and decision-making.

The DisasterHelp.gov portal[10], part of the Disaster Management e-Government initiative, is designed to provide federal, state and local emergency mangers with online access to disaster management information. This includes continually updated information on current disaster situations, as well as tools for planning and managing a disaster response.

The Department of Health and Human Services is leading the Grants.gov[11] e-Government initiative, which aims to simplify the grant management process. Grants.gov plans to provide a centralized online system for finding grants from any federal agency. This will not only make things simpler for grant applicants, but also will simplify the management of grants and eliminate redundancies by standardizing federal grant management activities among agencies.

1.3.1.4 Summary of Government Portals

All of the government portal initiatives described above are significant improvements in the delivery of government services. The primary drawback of the systems is that they require manual linkages, and therefore must be manually updated. The manual updating of these systems requires a heavy maintenance load in order to keep them current. The

[9] The web address for this portal is http://www.geodata.gov.

[10] The web address for this portal is http://www.disasterhelp.gov.

[11] The web address for this portal is http://www.grants.gov.

work presented in this thesis will demonstrate how a regulatory support system can be constructed such that it requires less manual maintenance.

1.3.2 Expert Systems

Software systems that encode knowledge from a human expert and use that information to provide advice to users are generally called “expert systems.” Various types of decision support systems and expert system technologies have been around since the 1960’s. Early examples include the DENDRAL project [54] that analyzed chemical structures, the MYCIN project [90] that was a medical treatment advisor, and Digital Equipment Corporation’s XCON system [7] for computer system configuration. Many expert systems have been developed for both research and commercial purposes since then. Some of the portals described earlier also make available expert systems or decision tree systems to assist users in understanding certain legal requirements.

There are a number of expert systems available through the government portal sites. A few of these systems are described below:

- The Auto Dismantler & Recycler Environmental Audit Advisor is designed to help users understand environmental regulations that apply to their work, as well as help users figure out what they need to do to comply with the regulations.[12]

- The Motor Vehicle (Class V) Waste Disposal Wells Advisor is designed to help users understand an EPA rule governing motor vehicle waste disposal wells, in addition to helping users figure out what they need to do to comply with the rule.[13]

[12] The Auto Dismantler & Recycler Environmental Audit Advisor is located at the web address http://www.smallbiz-enviroweb.org/auto/autointro.html.

[13] The Motor Vehicle (Class V) Waste Disposal Wells Advisor is located at the web address http://www.smallbiz-enviroweb.org/classv/intro.html.

- The OSHA Fire Safety Advisor is a system available online as one of the Department of Labor's elaws Advisors systems, which are designed to help employees and employers learn about federal employment laws. The Fire Safety Advisor targets employers' responsibilities and asks a series of questions so that it can identify potential fire safety violations. The system provides excerpts of regulation provisions that the employer may be in violation of as results.[14]

- The I-9 Employment Eligibility Verification Form Wizard is available through BusinessLaw.gov. This system is designed to help business owners determine if they need to fill out an I-9 Employment Eligibility Verification form for an employee by asking a series of questions. If the user needs to fill out the form, the system can help the user fill out the I-9 form on-line by asking another set of questions required to complete the form.[15]

To the best of our knowledge, these systems were all manually created with expert system shells that use "If…Then" rules.[16] Systems that are largely built manually using this approach may have a high initial reliability, but they tend to be brittle if changes to the rules need to be made. These types of systems also tend to be expensive to create, thus hindering them from achieving wide spread adoption.

Perhaps the most significant limitation of the systems currently available online is that they do not directly map to the regulations or legal documents that they represent. The failure to map to the source documents creates four significant disadvantages. First, because users do not see the regulation text as they interact with the system, users may have difficulty understanding the results produced by the system. Second, since users do not see the regulations during processing they may have trouble learning how the

[14] The OSHA Fire Safety Advisor is located at the web address http://www.dol.gov/elaws/fire.htm.

[15] The I-9 Employment Eligibility Verification Form Wizard is located at the web address http://sba.activenav.com/sba/index_main.htm.

[16] Per email correspondence with a researcher at CONSAD Research Corporation, which developed many of the systems described in this section.

regulation works, and may have difficulty re-tracing the results of the system on paper for validation purposes. Third, since users cannot track how the system is proceeding with its analysis, they will have trouble investigating background information on issues or questions the system raises. This third point means that the system is frozen in time, since any information not encapsulated in the system at the end of its development cannot be conveyed. It also means the system loses value as a learning tool since it is difficult for users to use the system as a point of departure from which to look for more information. Fourth, updating the system as the regulation changes is difficult, since without a mapping between the regulation and the rules in the system it may not be clear what parts of the system need to be changed when the regulation is altered. The research work described in this thesis aims to provide an information management framework that is capable of providing guidance similar to that provided by expert systems, while alleviating the four drawbacks identified above.

1.3.3 Legal Information Systems

Digital technology has significantly impacted the way legal practitioners identify appropriate legal material. Legal information providers like Westlaw and LexisNexis make available massive databases that can be searched using keywords or natural language to identify relevant legal documents. Westlaw and LexisNexis currently dominate the legal information provider marketplace in the United States. The vast document collections provided by these companies include a large number of topics from federal and state case law, a range of legal and news publications, and a variety of public records and filings. The value of these databases is such that they have become essential tools for many legal professionals [70].

These professional legal information services can be prohibitively expensive [107]. In addition, the organization and annotation of the information available through these services is tailored to legal professionals. The high cost of Westlaw and LexisNexis and their focus on the legal professional audience make them an inaccessible option for many companies, particularly small companies, that are interested in investigating regulatory requirements.

With the advent of the Internet, a number of smaller legal information providers emerged. FindLaw.com[17] was among the more popular and successful of these services. FindLaw started in 1995 as a list of Internet websites that was assembled for a workshop of the Northern California Law Librarians, and was launched as a commercial service in early 1996. This free service quickly grew in popularity, and in 2001 it was acquired by West Group (the parent company of Westlaw). While the FindLaw success story illustrates the demand for more affordable alternatives to Westlaw and LexisNexis for legal information, it does not address the issue of providing regulatory (legal) information for the non-legal professionals.

1.4 Regulatory Information Infrastructure

To better understand the context for the work described in this thesis it is useful to examine the broader issue of what a complete environmental regulatory infrastructure might contain. A complete infrastructure should provide a formal and practical way to enhance the access and retrieval of government regulations as well as to provide support for the users, framers and critics of the regulations. The infrastructure must have a formal basis in order to scale to the size and range of regulatory problems that it is designed to address. The development of ad hoc solutions will result in a patchwork of partial solutions that cannot scale to address the range of issues presented by the environmental regulatory domain. The infrastructure must be practical, since it is designed to address a practical problem with real world ramifications. There are five key

[17] The web address for FindLaw is http://www.findlaw.com.

areas in which work needs to be done to provide an environmental regulatory infrastructure. These areas are:

- Development of repositories and access tools: Gathering repositories of regulation documents allows for more effective searching and organization techniques to manage regulatory information. Repositories are a core building block for the development of many other tools. Access tools are necessary to make use of the information in the repository. In order to achieve broad accessibility for the information the repository must be accessible via the Internet. A variety of approaches are possible for providing this access.

- Development of ontologies: Ontology development entails the development of a formal vocabulary and structural thesauri to describe terms in the regulatory domain. These ontologies will enable, among other things, greater interoperation between regulatory domains by addressing the language barrier created by different terminologies.

- Development of representational formalisms: Structuring the representation of regulations and other important regulatory documents will allow more effective software-based processing of these documents. Documents should be structured to allow for annotation with meta-data that can be used by a variety of software tools.

- Development of analysis tools: Analysis tools facilitate a number of undertakings such as the identification of conflicting regulations, the investigation of the impact of a new regulation, or the investigation of the impact of repealing a regulation.

- Development of online compliance checking tools: Online compliance checking will make it easier for entities to determine if they are in compliance with regulations.

The research work described in this thesis addresses three of these five research areas: development of repositories and access tools, development of representational formalisms, and the development of an online compliance checking system.

1.5 Research Goals

The primary goal of this work is to explore and present solutions to the information problem posed by the environmental regulatory system. Through the application of information retrieval techniques and development of new methodologies, this work aims to develop a formal information infrastructure for regulatory information management and compliance assistance. The experimental scope of this work covers Code of Federal Regulations (CFR) Title 40: Protection of the Environment. Implementation examples focus on the regulations covering hazardous waste and the management of used oil.

There are three main elements of research and development addressed in this thesis. 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 element is the creation of a regulation assistance system (RAS) built upon the XML framework and document repository. The relationship between these elements is shown in Figure 1.1. The document repository lays part of the foundation for the regulation assistance system. The XML regulation structure provides the core data structure for the regulation assistance system. The regulation assistance system builds upon and extends work done for both the regulation repository and the XML regulations.

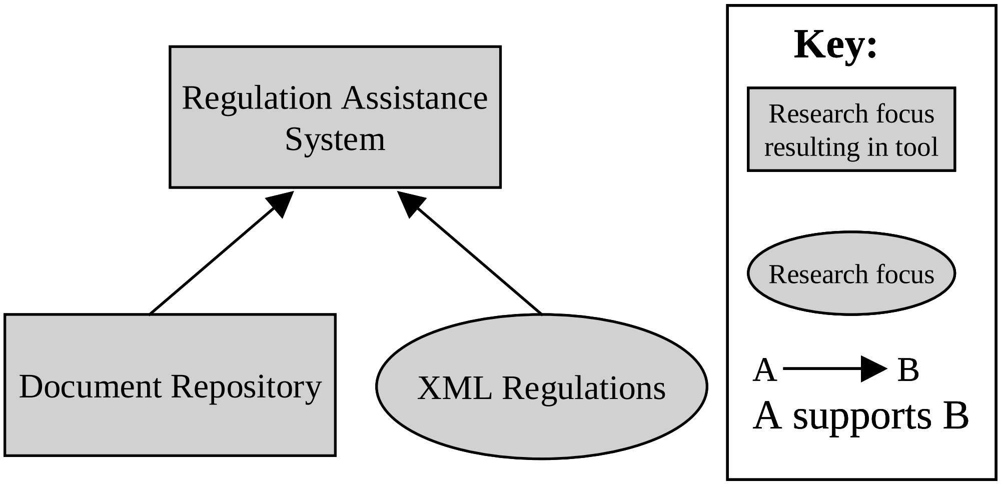


Figure 1.1 Relationship between RAS, document repository and XML regulations

The major research questions this research work seeks to address are:

- How can we more effectively organize regulatory information?
- How can we make the information and rules encoded in regulations more accessible to the people who need to understand it?
- How can we represent the information and rules in environmental regulations in a computer interpretable format? In other words, what can be done to represent the information and rules in a way that software systems can better work with them?
- If we can represent a regulation's information and rules in a computer interpretable format, how can we structure this information to assist with regulation compliance checking?

We have consulted with a number of individuals from industry, government and academia throughout the course of this research work. Individuals from government and industry whom we have consulted with include: a senior regulatory advisor from EPA Region 9; a superfund project manager from EPA Region 9; an environmental compliance manager from a major computer manufacturer; a lawyer with extensive experience in environmental law; an environmental health and safety specialist for a used oil recycler; an owner of a small auto repair company, a number of regulatory compliance officials from nearby local governments; and representatives from Stanford's environmental health and safety department. In addition to meeting with these contacts individually, we have organized a small workshop at Stanford attended by many of these individuals. We have also forwarded our work to a number of agencies for comments and suggestions. This group of diverse individuals from industry and government plays an important part in this project due to the applied nature of this research. We value the feedback from a range of perspectives because it helps us focus on practical solutions.

1.6 Thesis Outline

The objective of this research is to develop an Internet-enabled software framework that facilitates the identification and understanding of relevant environmental regulations by taking advantage of modern computer science techniques. The framework is designed to provide users, developers, and public critics of regulations with a tool that helps them work with regulatory documents.

The rest of this thesis is organized into the following five chapters:

- Chapter 2 covers the document repository and access tools developed in this research work. The chapter introduces background information on supplemental regulatory documents and why these documents are important. Some terminology from the information management field will also be introduced. This will be followed by a discussion of some different types of categorization systems. The approach for developing classification structures in this work will then be introduced, and examples of the classification structures that have been created in the demonstration system will be provided. Finally, some of the more promising extensions possible for the document repository will be presented.

- Chapter 3 describes an XML framework for the representation of regulations. First, a brief review of the various common approaches for the representation of documents and regulations will be given. Second, the XML regulation structure developed in this thesis will be introduced, and the parsers for converting regulations into the XML structure are described. Third, the meta-data elements added to the XML regulation, such as concepts, definitions, references and legal interpretations, will be described in detail. The remainder of the chapter will describe the XML structuring of regulations and the annotation of these XML regulations with meta-data.

- Chapter 4 discusses the development of a regulation assistance system (RAS). This system guides a user through regulation requirements to assist them in determining if they are in compliance with the regulation. The required extensions to the XML regulation standard from Chapter 3 are also examined in detail in this chapter. First, an overview of how the regulation assistance system works, and the motivation for it, will be provided as a point of reference. Second, propositional and predicate logic will be briefly introduced as a form of metadata. Third, the other types of metadata added to the XML regulations to enable a logic-based compliance-assistance system will be discussed. Fourth, the algorithms used for compliance checking will be examined. Fifth, the use of the regulation assistance system will be illustrated with examples, including an example of how the system can help the users who are unsure of how to best comply with a regulation. Finally, related research work in this area will be reviewed.

- Chapter 5 discusses the regulation assistance system in the context of the overall compliance problem. An example of how other software tools can build upon the compliance assistance framework is also provided. This example will address the issue of identifying relevant regulation provisions with which one needs to comply.

- Chapter 6 summarizes the contributions of this thesis and examines areas important for future research work. This chapter also discusses how the research work described in this thesis fits into the wider context of legal issues, regulatory issues, and security and privacy issues.

Chapter 2

Document Repository

2.1 Introduction

One of the primary objectives for this research work is to develop a formal process for building a document repository for environmental regulations. By document repository, we mean a place to store, organize, and make regulatory documents available for retrieval. For research purposes, the prototype document repository contains 125 megabytes of text-based content downloaded from the federal Environmental Protection Agency's (EPA) website, the Westlaw online legal information system, and other online sources.[18] We recognize the range and volume of supplementary documents for environment regulations, and our current document repository only represents a small fraction of all candidate documents. A complete environmental regulation-oriented document repository should contain the supporting documents for all sections of all environmental regulations. The number of documents gathered in this project was for

[18] The documents used in the document repository are on file at the Regnet project website, http://eil.stanford.edu/regnet/.

demonstration purposes only, and is sufficient for the research work described in this thesis.

Supplemental documents are important because they often contain information that is necessary for the accurate interpretation of the regulations to which they refer. Since the compliance assistance system uses the “used oil” regulations as its demonstration domain, the document repository contains a large number of supplemental documents dealing with used oil. These supporting document types include regulation preambles, 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 of the document repository are available through a search interface or the mediation of one or more searchable concept hierarchies. A commercial software package from Semio Corporation[19] has been employed 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 in a number of ways, including an alphabetical index of terms, according to the EPA’s list of extremely hazardous substances, and according to topic areas of regulation and pollution. These categorization structures are used to search the document repository for texts that address a common topic even when those texts do not explicitly reference one another.

This chapter will first introduce the background information on supplemental regulatory documents and why these documents are important. Some terminology from the information management field will also be introduced. This will be followed by a discussion of some different types of categorization systems. The approach for developing classification structures in this work will then be introduced, and examples of the classification structures that have been created for the demonstration system will be

[19] Semio Corporation has since been acquired by Entrieva. Information on the SemioTagger software package used in this research is available at the web address http://www.entrieva.com/.

provided. Finally, some of the more promising extensions possible for the document repository will be discussed.

2.2 Environmental Regulatory Documents

2.2.1 Federal, State, and Local Regulations

The EPA develops federal environmental regulations to implement statutes passed by the U.S. Congress. The regulations developed by the EPA provide the detailed requirements and procedures necessary to satisfy the objectives of the statutes. The Administrative Procedures Act (APA) governs the rule-making procedure used by the EPA to promulgate new regulations [1]. The APA provides for some different approaches to developing regulations, but the informal rule making approach is the most commonly used method [102]. In the informal rule making approach, the EPA must publish a notice of the proposed regulation in the Federal Register. The Federal Register is the official daily publication that federal agencies and organizations use to publish their regulations, proposed regulations, and other official notices. Interested parties then have a period of time, usually at least 30 days, to submit comments on the proposed regulation. After considering the comments, the EPA may then publish the final regulation in the Federal Register, and after another 30 days the agency can begin to implement the new rule [69]. The new rule is added to the Code of Federal Regulations (CFR), which is a codification of the general and permanent regulations published in the Federal Register.

The structure of environmental agencies and programs varies from state to state, so environmental regulations can also be issued by various state agencies responsible for the environment. All states are governed by federal regulations, with many states having authorization from the EPA to administer federal environmental policy in their state. In this case the state will have environmental regulations that meet the federal standards and comply with the federal regulations. In states that do not have authorization to administer federal environmental policy, both federal and state regulations will be in force. In this case, federal regulations will always need to be satisfied, but state regulations may add additional or stricter regulations that must also be satisfied.

The power of local government to enact environmental laws varies from state to state, according to the power granted by the state to the local governments. The processes and laws that occur at this level of government vary too much to summarize here, but regulation at the local level has become increasingly significant in recent years. In a law review article on the advent of local environmental law, John Nolon notes, "… there has been a remarkable and unnoticed trend among local governments to adopt laws that protect natural resources. These local environmental laws take on a number of forms. They include local comprehensive plans expressing environmental values, zoning districts created to protect watershed areas, environmental standards contained in subdivision and site plan regulations, and stand-alone environmental laws adopted to protect particular natural resources such as ridgelines, wetlands, floodplains, stream banks, existing vegetative cover, and forests" [68].

2.2.2 Supporting Documents

Supporting documentation for regulations is available from a large number of sources. The dispersed nature of this material makes it difficult for individuals to find all the relevant documents for regulation provisions. This research addresses this issue by building a repository in which it is be possible to store all regulation material. The supporting information may exist in a variety of different document types, such as regulation preambles, guidance documents, letters of interpretation, letters from the general counsel of the EPA, memos, court cases, and administrative decisions.

A regulation preamble provides the justification for a regulation's rules and is published along with the regulation. Black's Law Dictionary (7th ed. 1999) defines preamble as "an introductory statement in a constitution, statute, or other document explaining the document's basis and objective; esp., a statutory recital of the inconveniences for which the statute is designed to provide a remedy.” Although a preamble is not considered part of the official regulation, it provides some background for why the rule was developed. This background often includes why the EPA chose the regulatory structure defined in the regulation. Preambles generally cover topics such as: whom the regulation will affect, what studies were performed that have indicated the need for the regulation, and how the regulation addresses selected public comments received by the EPA on draft regulations. Preambles can provide valuable guiding principles for how the EPA interprets the text of the regulation.

Guidance documents, also known as non-binding guidance or interpretive documents, are informal policy documents that elaborate on how the EPA interprets or implements regulation requirements. These documents do not have the force of law. Guidance documents do not go through formal rulemaking procedures, and they are not legally binding on the regulated community. Guidance documents can, however, play a very significant role in determining how to comply with a regulation.

Letters of interpretation, memos, and administrative decisions are all forms of supplementary documents that, while less formal than guidance documents, may provide information similar to a guidance document on how the EPA interprets or implements regulation requirements. While these documents are not themselves legally binding, similar to guidance documents they provide important information on how the EPA may treat legally binding regulation requirements.

Court cases can shape environmental regulations through enforcement decisions or the application of judicial review. Enforcement decisions are decisions by courts on cases about the enforcement of environmental laws. The results of enforcement decisions by the courts can affect the compliance process through the clarification of ambiguous laws. For example, in the U.S. v Plaza Health Laboratories case the court decided that in the application of the Clean Water Act a person could not be considered a “point source” for pollution, overturning previous convictions against an individual for polluting the Hudson River [100]. The clarification of the interpretation for point sources, which are an important category for determining regulatory requirements, is an example of how courts may affect the regulatory system through decisions in enforcement proceedings.

Judicial review is a court’s authority to “determine the legality and constitutionality of an action of a government official, agency, or legislative body” [42]. Fred Anderson, in an article in the Duke Environmental Law & Policy Forum, notes that judicial review is an important part of the regulatory process [3]:

> “Judicial review of agency action plays a critical and distinctive role in the oversight of agency rulemaking. In contrast to legislative oversight, which focuses on an agency's budget and effectiveness, or executive supervision through the selection of agency leadership and policy, review of agency conduct by a court focuses on legality and reasoned decision-making. In this way, judicial oversight supplements political controls on administration by checking whether the agency is acting in accordance with the will of the political branches as expressed in its enabling legislation. Judicial oversight can also serve as a second look at the exercise of reasoning and expertise by agencies. As an independent check on the validity of administrative decisions, judicial review also contributes to the political legitimacy of regulation.
>
> Specifically, judicial review seeks to ensure that agency action violates no constitutional command (e.g., due process), is authorized by statute or other law, observes procedural requirements, and has a substantial basis in fact and reason.”

Since courts may strike down or change the interpretation of regulations as a result of judicial review, court cases may provide insight into how a regulation will be interpreted. It is therefore important that information on court cases be available in the document repository.

2.2.3 Why Supplementary Documents are Important

While generally not legally binding, supplementary documents can be important sources of information, without which it may be difficult to fully comply with environmental regulations. This section provides some background on the importance of supplementary documents.

The Beazer East, Inc. v. U.S. EPA case provides an example of the importance of supplementary regulation documents [9]. In this case the EPA brought action against Beazer East for violating RCRA groundwater-monitoring regulations with regard to its plant’s aeration basins. Beazer claimed the aeration basins were “tanks”, not “surface impoundments”, and were therefore not subject to RCRA groundwater-monitoring regulations. The primary issue the court resolved in this case was whether or not, under RCRA regulations, Beazer’s aerations basins qualified as “tanks” or if they should be treated as “surface impoundments.” Beazer’s aeration basins consisted of six-inch thick reinforced concrete basins approximately 15ft deep and 80ft across, built into the ground. In 40 CFR, Section 260.10, a “tank” is defined as:

> “Tank means a stationary device, designed to contain an accumulation of hazardous waste which is constructed primarily of non-earthen materials (e.g., wood, concrete, steel, plastic) which provide structural support.”

The regulation also defines “surface impoundment” as:

> “Surface impoundment or impoundment means a facility or part of a facility which is a natural topographic depression, man-made excavation, or diked area formed primarily of earthen materials (although it may be lined with man-made materials), which is designed to hold an accumulation of liquid wastes or wastes containing free liquids, and which is not an injection well. Examples of surface impoundments are holding, storage, settling, and aeration pits, ponds, and lagoons.”

An internal EPA memo, known as the “Weddle memorandum”, clarified the tank and surface impoundment definitions by using “structural support” as the deciding factor. The memo stated that in order for something to be classified as a tank, it must have sufficient structural support to maintain structural integrity even if removed from the ground and free standing. Beazer conceded that its aeration basins did not meet the standards laid out in the Weddle memorandum. Beazer claimed it had been deprived of adequate notice, however, since the standard described in the Weddle memorandum amounted to a new “rule” that had been adopted without the notice and comment procedures required by Section 553 of the Administrative Procedures Act (APA) [1].

The court found in favor of the EPA, ruling that the Weddle memoradum was an “interpretive rule”, not a “legislative rule”. Interpretive rules are exempted from the notice and comment requirements of the APA. Legislative rules, which must be promulgated in accordance with the APA, are rules that impose new duties upon regulated parties and have the force of law. Interpretive rules interpret language already existing in regulations, simply stating what an agency thinks a regulation means. Since the EPA’s interpretation of the tank requirements was reasonable and consistent with the regulation, the court ruled against Beazer in the case.

Supplementary information has taken on such importance in recent years that the EPA has come under criticism for possibly attempting to avoid the due process requirements of the APA by using informal documents to set policy [96]. A description of how supplementary documents can be used to set important aspects of policy was provided by the U.S. Court of Appeals for the D.C. Circuit in Appalachian Power Company v. EPA [4]. The court criticized the EPA for using supplementary documents to circumvent the requirements of the APA and avoid judicial review of its lawmaking:

> “The phenomenon we see in this case is familiar. Congress passes a broadly worded statute. The agency follows with regulations containing broad language, open-ended phrases, ambiguous standards and the like. Then as years pass, the agency issues circulars or guidance or memoranda, explaining, interpreting,

> defining and often expanding the commands in the regulations. One guidance document may yield another and then another and so on. Several words in a regulation may spawn hundreds of pages of text as the agency offers more and more detail regarding what its regulations demand of regulated entities. Law is made, without notice and comment, without public participation, and without publication in the Federal Register or the Code of Federal Regulations. With the advent of the Internet, the agency does not need these official publications to ensure widespread circulation; it can inform those affected simply by posting its new guidance or memoranda or policy statement on its web site. An agency operating in this way gains a large advantage. "It can issue or amend its real rules, i.e., its interpretative rules and policy statements, quickly and inexpensively without following any statutorily prescribed procedures." Richard J. Pierce, Jr., Seven Ways to Deossify Agency Rulemaking, 47 ADMIN. L.REV. 59, 85 (1995). The agency may also think there is another advantage--immunizing its lawmaking from judicial review."

In some cases, courts have struck down supplementary documents for overstepping their bounds as interpretive documents. For example, in the case Appalachian Power Company v. EPA [4] the court found that the EPA issued a guidance document that "broadened underlying EPA rule and its promulgation was thus improper absent compliance with formal rulemaking procedures". The court then set aside the guidance document in its entirety.[20]

Supplementary documents play a very important role in environmental regulations. For this reason, it is critical that individuals writing, enforcing, or seeking to comply with regulations have access to all relevant and available supplementary documentation. The

[20] Due to its demonstrative nature, the document repository is not currently being audited to remove supplemental documents that have been invalidated (such as documents that have been struck down by the courts).

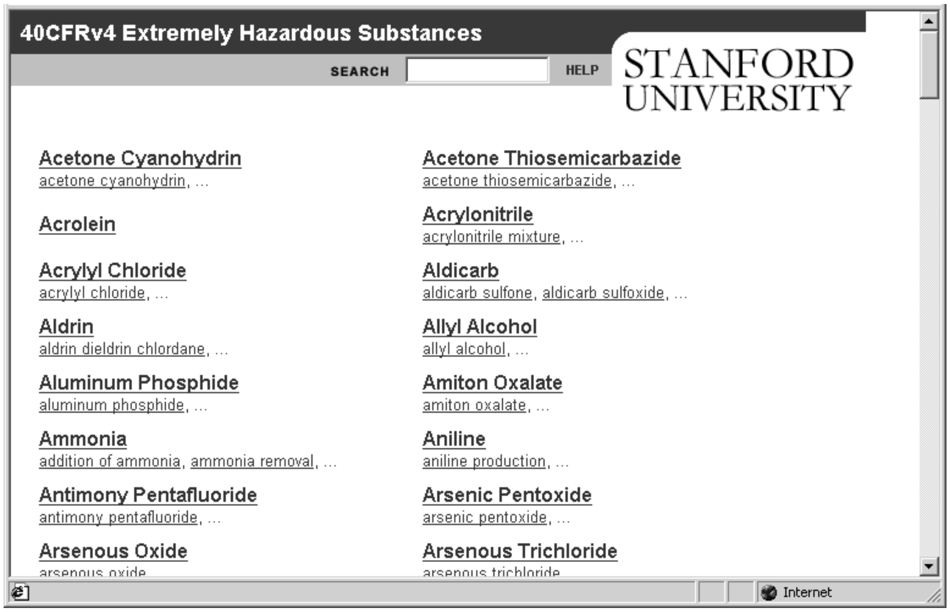
document repository described in this chapter is designed to make these important documents more accessible.

2.3 Categorization of Documents

This section explores the use of categorization systems as an organizing principle for the document repository. Categorization systems allow us to classify or categorize (the two terms are used interchangeably here) documents into modular hierarchical structures so that they are easier to browse and search. Browsing documents according to a categorization structure is a natural way to navigate documents. This is a common approach popular among web directories for web page searching. An example from the document repository is shown in Figure 2.1. First, the nature of categorization will be discussed. Categorization is a natural, yet challenging, approach to use for organizing information. Second, some information retrieval issues and terminology will be discussed. Third, the remainder of this section will focus on approaches for constructing categorization systems.

2.3.1 Categorization

There is a wide variety of terminology in use for classification work. For purposes of clarity, we will use the same terminology as used by Charles Heenan in his manuscript on classification [41]. Because the terms “taxonomy” and “ontology” take on a variety of definitions in the literature, we will use the terms “classification hierarchy,” “categorization hierarchy,” “classification structure,” and “categorization structure” instead of “taxonomy” or “ontology”. For a detailed history and more general treatment of classification the reader is referred to [41].



40CFRv4 Extremely Hazardous Substances

Acetone Cyanohydrin
acetone cyanohydrin, …

Acetone Thiosemicarbazide
acetone thiosemicarbazide, …

Acrolein

Acrylonitrile
acrylonitrile mixture, …

Acrylyl Chloride
acrylyl chloride, …

Aldicarb
aldicarb sulfone, aldicarb sulfoxide, …

Aldrin
aldrin dieldrin chlordane, …

Allyl Alcohol
allyl alcohol, …

Aluminum Phosphide
aluminum phosphide, …

Amiton Oxalate
amiton oxalate, …

Ammonia
addition of ammonia, ammonia removal, …

Aniline
aniline production, …

Antimony Pentafluoride
antimony pentafluoride, …

Arsenic Pentoxide
arsenic pentoxide, …

Arsenous Oxide
arsenous oxide

Arsenous Trichloride
arsenous trichloride

Figure 2.1 Example categorization of the document repository

A paper by Kenneth Bailey [6] contains an excellent introduction to classification. The following excerpt from Bailey's work also leads to why classification structures are an excellent fit for a regulation repository:

> "In its simplest form, classification is merely defined as the ordering of entities into groups or classes on the basis of their similarity. Statistically speaking, we generally seek to minimize within-group variance, while maximizing between-group variance. This means that we arrange a set of entities into groups, so that each group is as different as possible from all other groups, but each group is internally as homogenous as possible. By maximizing both within-group homogeneity and between-group heterogeneity, we make groups that are as distinct (nonoverlapping) as possible, with all members of a group being as alike as possible. …
>
> Almost everything is classified to some degree in everyday life, from chewing gum (bubble and nonbubble), to people (men and women), to animals, to

> vegetables, to minerals. Grouping objects by similarity, however, is not quite as simple as it sounds. Imagine that we throw a mixture of 30 knives, forks, and spoons into a pile on a table and ask three people to group them by “similarity.” Imagine our surprise when three different classifications result. One person classifies into two groups of utensils, the long and the short. Another classifies into three classes,—plastic, wooden, and silver. The third person classifies into three groups,—knives, forks, and spoons. Whose classification is ‘best’? …
>
> The lesson here should be obvious—a classification is no better than the dimensions or variables on which it is based. If you follow the rules of classification perfectly but classify on trivial dimensions, you will produce a trivial classification. As a case in point, a classification that they have four legs or two legs may produce a four-legged group consisting of a giraffe, a dining-room table, and a dancing couple. Is this what we really want?
>
> One basic secret to successful classification, then, is the ability to ascertain the key or fundamental characteristics on which the classification is to be based. A person who classifies mixtures of lead and gold on the basis of weight alone will probably be sadder but wiser. It is crucial that the fundamental or defining characteristics of the phenomena be identified. Unfortunately, there is no specific formula for identifying key characteristics, whether the task is theory construction, classification, or statistical analysis. In all of these diverse cases, prior knowledge and theoretical guidance are required in order to make the right decisions.”

Bailey’s example of people organizing the same objects into three different classifications, each perfectly valid, illustrates precisely why categorization structures are a good fit for organizing information in a document repository. Different people and organizations have different ways of structuring the same data so as to best fit their needs. For example, large semiconductor manufacturers, small auto body repair shops, and authors of new regulations will probably each have different ways of organizing regulation documents. There is no inherently “correct” way to structure all the information in the document repository, and categorization structures allow us to provide a number of different perspectives. Recognizing that different entities will want information classified in different ways is key to taking advantage of the new possibilities provided by current information technology. Jane Fountain, in a report on information, institutions and governance, notes that, “Currently, most government information on the web is organized according to the classification systems of agencies rather than the mental models of users. … Hence, although the Internet and web, in theory, make government information more accessible to the public, organization online often replicates paper-based classification schemes and therefore merely automates the status quo” [33]. Developing approaches for managing regulatory information that take into account that there will be many different ways of organizing information may alleviate the problem Fountain points out, and that is what the document repository discussed in this chapter is designed to do.

Figure 2.2 illustrates how different categorization structures can offer different perspectives on the same set of data. Note that documents grouped together in one categorization structure may not be grouped together in a different structure (e.g., documents D and E). In addition, documents that are covered in one structure may not be available in a different structure (e.g., documents A and E).

In order to be scalable, a classification hierarchy should have a clear logical relationship between categories and subcategories, and the overall structure should coherently subdivide the documents it is designed to cover. The most common logical relationship used in classification hierarchies is the generalization of the “is a kind of” relationship. For example, “Oak tree” could be a child node of “tree” if the “is a kind of” relationship were the organizing principle for the hierarchy. Other popular organizing principles are “part-to-the-whole” (i.e., “leaf” could be a child of “tree”) and “is-a-process-of” (i.e., “photosynthesis” could be a child of “tree”). The development of scalable classification

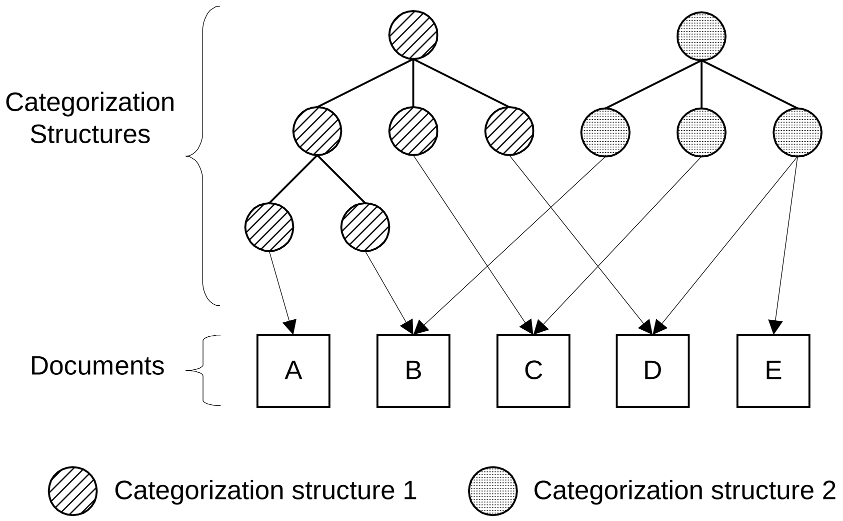


Figure 2.2 Illustration of multiple categorization structures over one set of documents

hierarchies is more of an art than science, but using clear logical relationships consistently throughout the structure is key to developing a successful structure. Approaches to developing good categorization structures will be discussed further in Section 2.3.3. In designing the document repository for environmental regulations, we seek to identify some of the defining characteristics along whose dimensions different categorization structures may be developed for environmental regulations and their related supplementary documents.

2.3.2 Information Retrieval

One of the primary goals in developing the document repository is to make documents relevant for interpreting regulations more accessible, or easier to find. This area of study is generally known as information retrieval. This section describes briefly some of the issues in information retrieval that are relevant to our research.

2.3.2.1 Precision and Recall

Two common metrics for evaluating information retrieval systems are precision and recall. Precision is the ratio of correctly returned documents to the total number of returned documents. This ratio reflects how likely it is that a returned document is relevant. Recall is the ratio of correctly returned documents to the total number of relevant documents in the system. This ratio reflects how complete a result set from a search may be. The quantities used to calculate precision and recall are illustrated in Figure 2.3. Figure 2.4 shows the equations for calculating precision and recall.

It is desirable for most information retrieval systems to have a balance between these two indicators. If one has a very high value and the other is very low the resulting system may not be very useful. For example, a system might attain high precision by simply returning a single document with the highest relevance score for the given query. Since that single document would very likely be relevant, the system would have a high precision rating. This, however, will not be a very useful system for most users since they probably want to see a reasonably sized set of highly relevant documents returned as the result (i.e., users would like greater recall).

A system designed to maximize recall could simply return the set of all documents it contains in response to any query. The results would have a low precision rating, but no relevant documents would be missed by this approach so it would have very high recall. Users will probably have difficulties using this system because of the information overload, so high recall with low precision is also not useful.

2.3.2.2 Polysemy and Synonymy

Polysemy refers to the fact that a word can have multiple meanings. For example, the word “bank” might refer to a riverbank or a financial institution. This creates an information retrieval problem, since it may not be obvious what sense of a word a user

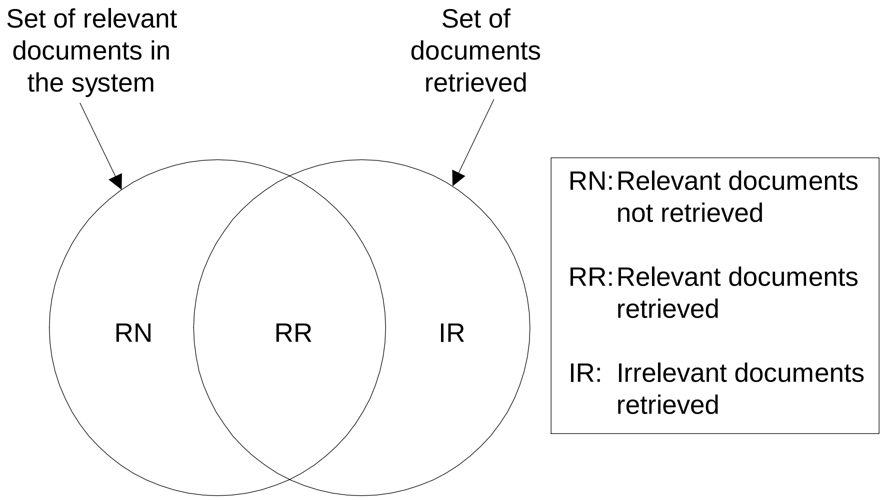

Set of relevant documents in the system
Set of documents retrieved
RN
RR
IR
RN: Relevant documents not retrieved
RR: Relevant documents retrieved
IR: Irrelevant documents retrieved

Figure 2.3 Illustration of quantities used to calculate precision and recall

$$precision = \frac{\#RR}{\#RR + \#IR} \; ; \qquad recall = \frac{\#RR}{\#RR + \#RN}$$

Figure 2.4 Precision and recall equations

intends when constructing a database query. Similarly, it may not be obvious what sense of a word is used in a particular document. The precision of search results suffers due to polysemy because unwanted documents will be selected.

An approach for addressing the polysemy problem is to use word sense disambiguation techniques to determine what sense of the word is intended. There has been significant work done to address this problem [25, 47, 103]. However, polysemy continues to be a difficult information retrieval issue. In context there is less polysemy, so with a categorization structure the problem is minimized.

Synonymy refers to the fact that there can be multiple words or phrases that express the same concept. This is a problem for information retrieval tasks because it may result in low recall for searches. A search for a word or phrase will miss documents that contain synonyms for the search string, thus decreasing the recall of an information retrieval system.

A straightforward approach to this problem is to apply dictionaries and thesauri to identify words and phrases that express the same concept. This approach can result in a rapid increase of the polysemy problem and consequent decrease in the precision of search results. More advanced approaches have been developed that use conceptual features to address the synonymy problem more successfully, with less of a degrading effect on precision [58].

2.3.3 Categorization Systems

There are many possible approaches for developing a classification system for documents. The wide range of methods can be grouped into three broad categories: manually constructed, automatically constructed, or semi-automatically constructed categorization systems [41]. Each of these approaches has advantages and disadvantages. The most relevant issues with respect to building a regulation document repository will be discussed next.

2.3.3.1 Classification Automation

Manually constructed categorization hierarchies are a straightforward solution for organizing small sets of documents. In the manual approach a categorization hierarchy is manually constructed, and then documents are manually added to the appropriate category in the structure. The benefits of this approach are that it is possible to build a high quality categorization hierarchy with clear relationships between categories, and the accuracy of documents within the categories may also be very high. There are a number of drawbacks, however. First, manual categorization is extremely time-consuming. This is particularly true for large categorization efforts involving a team of people organizing documents, where detailed guidelines and strict attention to details are necessary. Second, it is very difficult to ensure consistent categorization in the manual categorization efforts. Even when a single individual is responsible for categorizing all documents, that person may put the same document in different categories at different points in time. When the categorization is a team project, the problem is multiplied since different people may put the same document in different categories even though they use the same set of elaborate guidelines. Despite these drawbacks, it is possible to build high-quality categorization hierarchies manually. Some great successes in this area include the MeSH[21], the Yahoo directory[22], and the open directory project.[23]

In the automatic categorization approach, a categorization hierarchy is automatically derived from the document set, and documents are automatically added to this categorization hierarchy. There are many methods for accomplishing this goal, and it continues to be an active area of research [27, 50]. In some cases, the categorization hierarchy is automatically extracted from an uncategorized set of sample documents. This methodology has a drawback illustrated by the Bailey quotation in Section 2.3.1; that is, there are many possible dimensions along which to organize a categorization hierarchy, and not all of them will be useful. Other methods for automatically creating a categorization hierarchy involve the use of training sets that have already been properly categorized. The system can then use the training set to automatically generate rules for populating a categorization hierarchy with documents. The use of training sets is generally quite effective when the training sets are reasonably large, but any manual

[21] MeSH, the Medical Subject Headings controlled vocabulary, is used for indexing articles, for cataloging books, and for searching MeSH-indexed databases. The MeSH vocabulary facilitates the retrieval of information that may span different terminologies. MeSH is managed by the National Library of Medicine, and is available on the Internet at the web address http://www.nlm.nih.gov/mesh.

[22] The Yahoo directory is a directory of websites developed by Yahoo! Inc. The web address for this directory is http://dir.yahoo.com/. A staff of editors at Yahoo categorizes web pages into the manually developed classification hierarchy, which was one of the first to popularize this approach to organizing the World Wide Web.

[23] The open directory project is a directory of websites maintained by a community of volunteer editors. The web address for this directory is http://dmoz.org/. Editors volunteer to maintain a small portion of the complete classification hierarchy. The open directory project forms the core web directory for a number of search engines, such as Netscape Search, AOL Search, Google, and Lycos.

categorization errors in developing training sets may be magnified by the automatic classification system in the end results. There are two main drawbacks to a fully automated classification system. First, the logical relationships within the categorization hierarchy may not be explicitly clear. Since the primary reason for using categorization structures is that they tend to be intuitively clear and easy for people to work with; constructing categorization hierarchies that are not intuitive to work with reduces their value. Second, since the logical transparency of the structures may be low, it may be difficult to audit the classification structures for quality. Thus, the quality of the categorization structure could be reduced. Despite these drawbacks, fully automated categorization systems may be quite useful when it is necessary to quickly and inexpensively categorize large document sets, particularly when quality is not a primary concern.

Partially automated classification systems seek to blend the advantages of both manual and automatic categorization. There are many possible combinations of manual and automatic categorization, so the discussion here will focus on the most common method. In this approach, the categorization structure is designed manually, perhaps with the assistance of software tools, and the categorization hierarchy is automatically populated with documents. Constructing the categorization hierarchy manually allows the use of human judgment to develop useful logical relationships within the hierarchy. Automatically populating the hierarchy with documents according to pre-specified categorization rules reduces the two main drawbacks to manual categorization. First, populating the categorization hierarchy is fast and efficient, much less time-consuming and less expensive than doing it manually. Second, automatically populating the categorization hierarchy ensures that the application of categorization rules is consistent. This form of partially automated classification must deal with the problems of developing an effective categorization hierarchy and specifying good classification rules. Extensive experimentation and iterations are necessary for building a good classification hierarchy using a partially automated classification approach.

There are many factors to consider when deciding which of the three categorization approaches to use for organizing documents: manual, automatic, or partially automated. The most salient feature, however, is the trade-off between the error rates of fully automated approaches, and the time and cost of more manual approaches. In building the regulatory document repository, we use a partially automated approach to categorizing documents. Given the large volume of environmental regulatory related documents and limited resources available from government or industry to organize them, a manual categorization approach would be impractical. This is particularly true when one considers that there are a variety of different perspectives that industry and government groups would like to see, thus splitting these limited resources into a multitude of categorization efforts. A fully automated approach to organizing environmental regulatory information would not be a good fit either, since having clear logical structures and low error rates is very important. Locating relevant environmental regulatory documents is extremely important, so logically incoherent categorization structures, or a high rate of incorrectly categorized documents would not be acceptable. The problem of incorrectly categorized documents is particularly acute for environmental regulatory information, since the proper category for a document sometimes depends upon minor conceptual issues.

2.3.3.2 Approaches to Developing a Classification Hierarchy

An essential component of partially automated classification is to develop the classification hierarchies. This section addresses several approaches for developing these hierarchies. Categorization hierarchies can be developed from a top-down perspective, bottom-up perspective, or a hybrid combination of these two methods.

A top-down approach to developing a classification hierarchy refers to the approach of conceptualizing a meaningful way to break down documents into a set of categories, and expanding these categories into subcategories to whatever depth seems appropriate. The entire process is done without examining representative documents from the set of documents to be categorized. Rules can then be developed to filter documents into appropriate categories within the classification hierarchy. While very clean logical structures can result from this type of approach, there are several weaknesses to this method. First, a set of documents may not map well to an abstractly created classification hierarchy. Some categories may be empty, or nearly empty. Other categories may be populated with so many documents that it may be difficult to identify those of interest. Second, there may be many documents in the set of input documents that do not fit into any of the categories that were developed with the top-down approach. These documents will be incorrectly classified, or not classified at all, thus making them inaccessible.

A bottom-up approach is basically the method of browsing through a set of input documents and developing a classification hierarchy based upon the terms and concepts that seem to stand out in the document collection. This approach can be very effective for a static document collection. However, if the document collection grows or changes over time, it can be difficult to adapt the classification hierarchy to the new data. In addition, a bottom-up classification hierarchy will not generalize well if applied to other document sets. This is because the prominent terms, concepts, and depth of topic coverage will be very specific to the particular document set for which the classification hierarchy is developed.

A combination of top-down and bottom-up approaches, called a hybrid approach, balances the strengths and weaknesses of the two methods. When using a hybrid approach to developing a classification structure, a top-down conceptualization of the classification hierarchy is iteratively refined using the data from a bottom-up perspective. For example, the top levels of a classification hierarchy might be developed using a top-down approach. Basic classification rules for adding documents to the respective categories could then be developed, and an automated system could populate the structure with documents. The designer could then survey the results, investigating how well the categories break down the documents into manageable units, and what types of documents failed to match any categories within the classification structure. Using an iterative approach, subcategories can then be added to the initial classification structure until the designer is satisfied with the distribution and coverage of the documents included in the classification hierarchy. This hybrid approach to designing a classification structure should mitigate some of the problems associated with a top-down method, while improving the generalization of a bottom-up approach.

2.4 Document Repository Features

As mentioned in the previous section, a semiautomatic approach to developing classification hierarchies is used for the development of the document repository for environmental regulations. In this section we will discuss the process used for designing and refining categorization hierarchies. A software package from Semio Corporation was used for purposes of building the classification hierarchies. There are a number of software programs available from companies, research entities, or the open source software community that provide categorization tools. The use of a commercial software package from Semio Corporation provides many useful features, such as a graphical user interface, noun phrase extraction services, and other tools that greatly facilitate this research work. Nevertheless, the issues discussed in this section are applicable to designing and building classification hierarchies to organize sets of documents in general.

We will illustrate the process for building a categorization hierarchy using a hybrid top-down, bottom-up strategy with the Semio software package. Once one is familiar with the set of documents to be organized, the first step is to develop an initial high-level categorization hierarchy. With the software tools used in this research project, this entailed developing a small text file with a few high-level categories, and “latching” noun phrases that help assign documents to those categories.

When a document is being processed, the software automatically extracts noun phrases from the document that are characteristic of the topics the document is related to. For convenience these noun phrases are termed “concepts”. Concepts are useful when developing categorization hierarchies because they can be used to assign documents that contain specific concepts to particular locations within the categorization hierarchy.

In Semio, a text file containing an initial categorization hierarchy has the form shown in Figure 2.5. Category names are denoted by a word or a phrase preceded by an exclamation point. An indented list of words and phrases preceded by plus or minus characters indicate the latching concepts for that particular category. Concepts preceded by a plus character indicate that documents containing those concepts should be placed under the related category. For example, documents containing the concept “amendment” should be placed under the “On the Topic of Regulation” category in Figure 2.5. Concepts preceded by minus character indicates that documents containing those concepts should not be placed under that particular category even if they contain other latching concepts for the category. For example, “penalty” and “sanction” are exclusionary concepts under the “On the Topic of Regulation” category, and they prevent documents containing these concepts from latching into this category. An indented line that is started by an exclamation point indicates another category within the categorization hierarchy. The tabular depth of the category name indicates the depth within the categorization hierarchy. For example, “Permits” and “Penalties and Sanctions” are both subcategories of “On the Topic of Regulation” in Figure 2.5.

Once an initial specification file for the categorization hierarchy is created, the software package can be used to assign documents to populate the classification structure. When the classification structure is populated with documents it is possible to get statistics indicating how well the classification structure represents the content of the document corpus. For example, the percent of documents in the document corpus that are matched

```
!On the Topic of Regulation
        +agreement
        +amendment
        +authority
        +jurisdiction
        -penalty
        -sanction
        !Permits
                +permit
                +permittee
        !Penalties and Sanctions
                +penalty
                +sanction
```

Figure 2.5 Categorization hierarchy specification file

into the classification structure is one coarse but useful indicator. If a large percentage of the documents fail to latch into the categorization hierarchy, it is clear that some important topic areas are missing from the categorization structure.

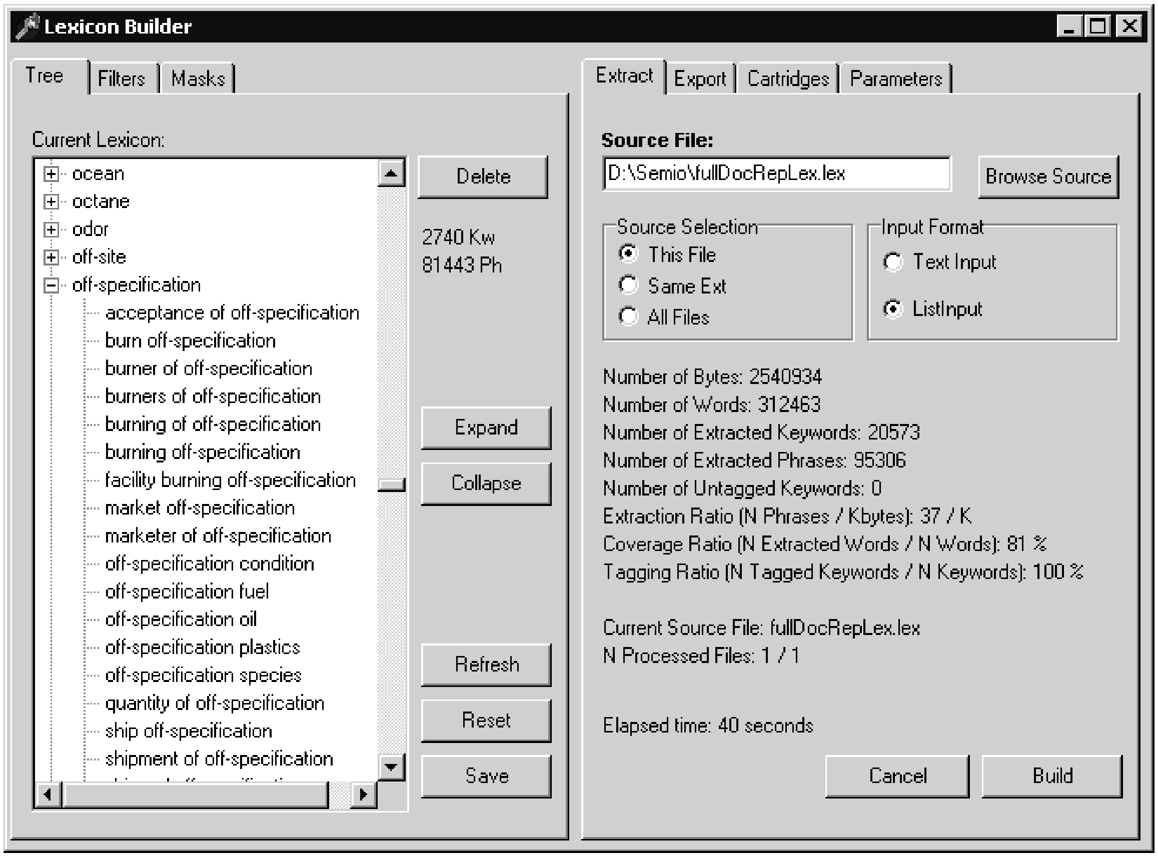
After populating the classification structure, it is possible to browse through the results of the classification hierarchy. This manual inspection of how documents are assigned to categories allows one to further refine the classification hierarchy specification file. If some categories appear to be heavily overpopulated with documents, it is useful to inspect the contents of documents in the category and consider refining the category into several subcategories. If some types of documents have been assigned to an incorrect category, it is useful to specify exclusionary concepts to prevent the documents from latching into the improper category. One can also add additional concepts to the correct category so that the documents will latch to the proper location.

This iterative process of modifying the classification hierarchy specification file, reclassifying documents, investigating the results, and then further refining the classification hierarchy specification can be used to develop a stable, inclusive, and logically clear categorization hierarchy.

In addition to the general steps described above, other lexical tools may be useful when developing categorization hierarchies. For example, it is possible to have Semio export a list of all the concepts it identified in all documents in the corpus. The file of exported concepts can be scanned to determine what types of concepts Semio has extracted, and can further be used in either latching or exclusionary rules. Semio also provides an additional program called Lexbuilder to aid in the investigation of the extracted concepts list. Many concepts share words in common, and the Lexbuilder tool aids in the investigation of how these concepts are related. Figure 2.6 illustrates how Lexbuilder aids in the investigation of concept noun phrases in the document corpus.

The software package used in this research work is partially able to deal with the synonymy problem. By inspecting the extracted concepts list or working with the Lexbuilder tool, it is possible to identify synonymous concepts that were being extracted by the system. It is then possible to specify that these synonymous concepts should all map to a single concept, so that they will be properly categorized. One drawback is that this solution to the synonymy problem is largely manual. Without proactive support from a software package to identify synonymous concepts in the document corpus, it is likely that many synonymous concepts will be missed when dealing with a very large document corpus.

In this work, we have not addressed in detail the polysemy problem since the tools we employed are not equipped for this problem. Polysemy does create some problems for some of the categorization hierarchies we have developed, because occasionally different senses of the same concept phrase could be used for latching rules in different categories. One way to address this issue is to add a number of exclusionary concept phrases targeting the problem documents. When these exclusionary concept phrases are added to the categories for which polysemy is an issue the incorrect latching of documents is reduced.



Current Lexicon:

- ocean
- octane
- odor
- off-site
- off-specification
  - acceptance of off-specification
  - burn off-specification
  - burner of off-specification
  - burners of off-specification
  - burning of off-specification
  - burning off-specification
  - facility burning off-specification
  - market off-specification
  - marketer of off-specification
  - off-specification condition
  - off-specification fuel
  - off-specification oil
  - off-specification plastics
  - off-specification species
  - quantity of off-specification
  - ship off-specification
  - shipment of off-specification

2740 Kw
81443 Ph

Source File: D:\Semio\fullDocRepLex.lex

Source Selection: ● This File ○ Same Ext ○ All Files

Input Format: ○ Text Input ● ListInput

Number of Bytes: 2540934
Number of Words: 312463
Number of Extracted Keywords: 20573
Number of Extracted Phrases: 95306
Number of Untagged Keywords: 0
Extraction Ratio (N Phrases / Kbytes): 37 / K
Coverage Ratio (N Extracted Words / N Words): 81 %
Tagging Ratio (N Tagged Keywords / N Keywords): 100 %

Current Source File: fullDocRepLex.lex
N Processed Files: 1 / 1

Elapsed time: 40 seconds

Figure 2.6 Lexbuilder tool for working with extracted concepts

2.4.1 Categorization Hierarchies Developed[24]

A number of categorization hierarchies have been developed using either a top-down, or a hybrid top-down and bottom-up strategy. One of the top-down classification structures designed classifies documents according to the EPA's list of extremely hazardous substances[25]. This classification structure can be useful for individuals interested in how

[24] The development of these classification structures is a joint effort with Mr. Charles Heenan. The classification structures are presented here to provide a complete treatment of building the document repository.

[25] The EPA’s list of extremely hazardous substances was downloaded from the EPA website, http://www.epa.gov.

particular chemical compounds are regulated. Because this is a strict top-down approach to classification, one would expect that many documents in the repository would not fall anywhere within categorization hierarchy. That is indeed the case, with less than 25% of the documents in the document repository being attached to the hierarchy. As expected, this classification hierarchy fails to capture a significant percentage of the documents within the document repository. This does not mean that the categorization hierarchy is not useful, however, since it was intentionally designed to address a very specific issue and broad coverage is not necessary.

One of the hybrid classification structures developed is intended to address the issues of regulation, pollution, and waste. In addition to the iterative process described above for constructing and refining a classification structure, this classification structure has also been reviewed by legal experts and individuals from industry. When this classification structure is populated with the documents from the document repository, nearly 95% of the documents in the document repository are attached to the hierarchy. As expected, the hybrid approach to developing categorization structures achieves a good coverage of the document repository.

2.4.2 Browsing

The document repository uses a standard classification hierarchy interface for accessing documents. The interface allows the users to navigate through the classification structure as a menu of category to sub-category links along with hyperlinks to documents that are located in the various classifications.

To illustrate how categorization hierarchies can be used to organize documents, the next several figures illustrate browsing through the classification structure on the topics of regulation. Figure 2.7 shows a top-level view of how one might examine the documents from the perspective of regulations (i.e., violations, penalties, etc.) or the topic of pollution and waste (i.e., pollution, liquid waste, solid waste). Figure 2.8 shows the results if we select the pollution and waste category. In this case one is presented with a list of subcategories such as solid waste, waste management, emissions, etc. Next to each subcategory is a number in parenthesis indicating how many documents are located within that particular subcategory. In addition to subcategories, a list of concepts is presented to indicate the concepts that latch the documents into the classification hierarchy below this category. At any point within the classification hierarchy, one can view all documents that are located within a classification category. This is shown in Figure 2.9 as a set of document links. One can also find all documents that contain a particular concept, in which case the context within which the concept appears in the documents can be viewed. This is shown in Figure 2.10.

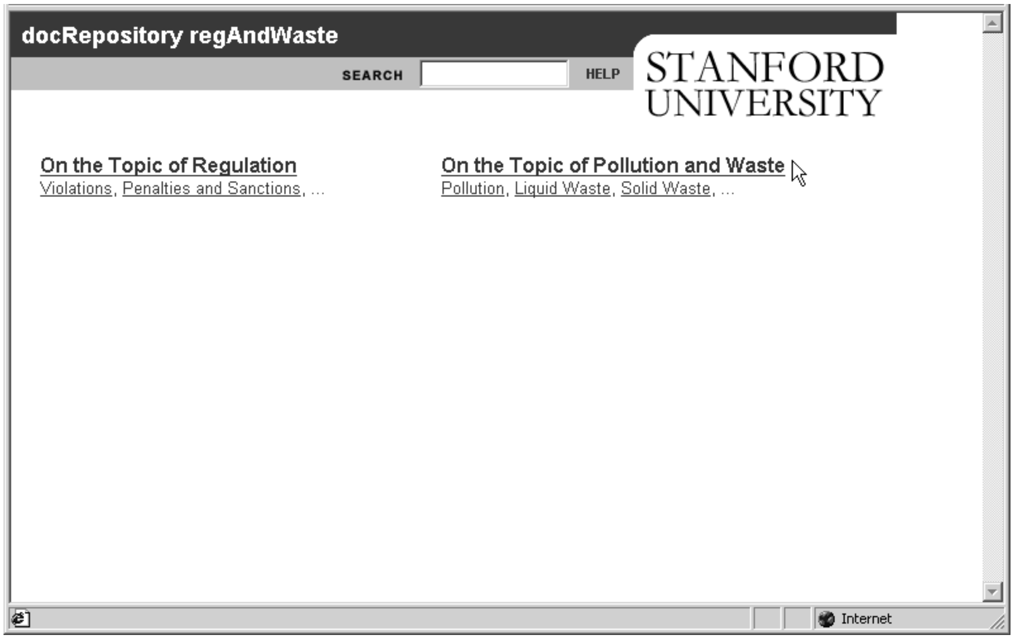
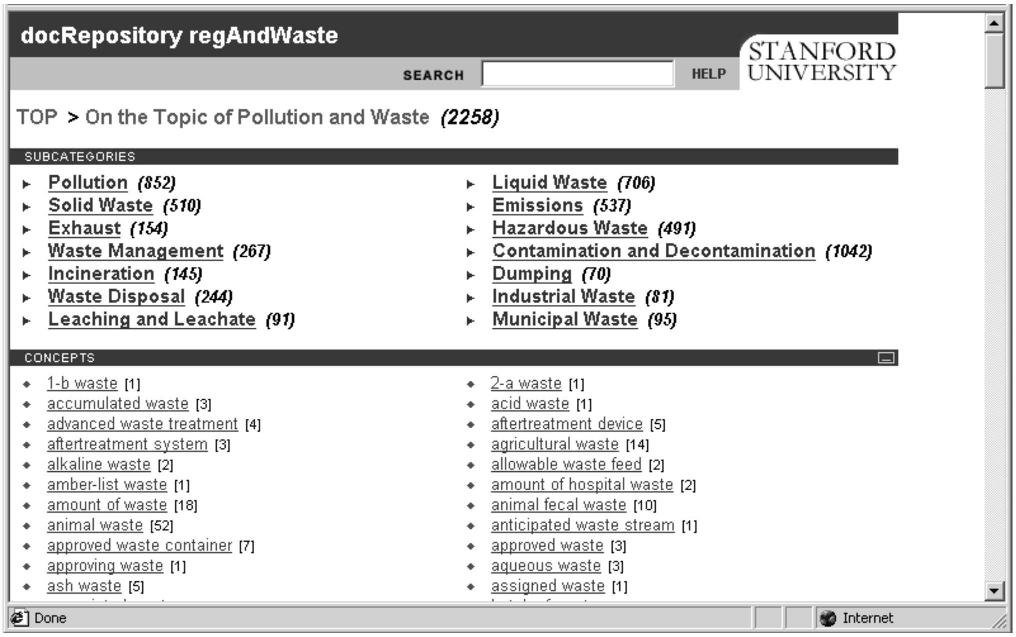

docRepository regAndWaste
SEARCH
HELP
STANFORD UNIVERSITY
On the Topic of Regulation
Violations, Penalties and Sanctions, ...
On the Topic of Pollution and Waste
Pollution, Liquid Waste, Solid Waste, ...
Internet

Figure 2.7 Top level view of regulation, pollution and waste categorization hierarchy



docRepository regAndWaste

TOP > On the Topic of Pollution and Waste *(2258)*

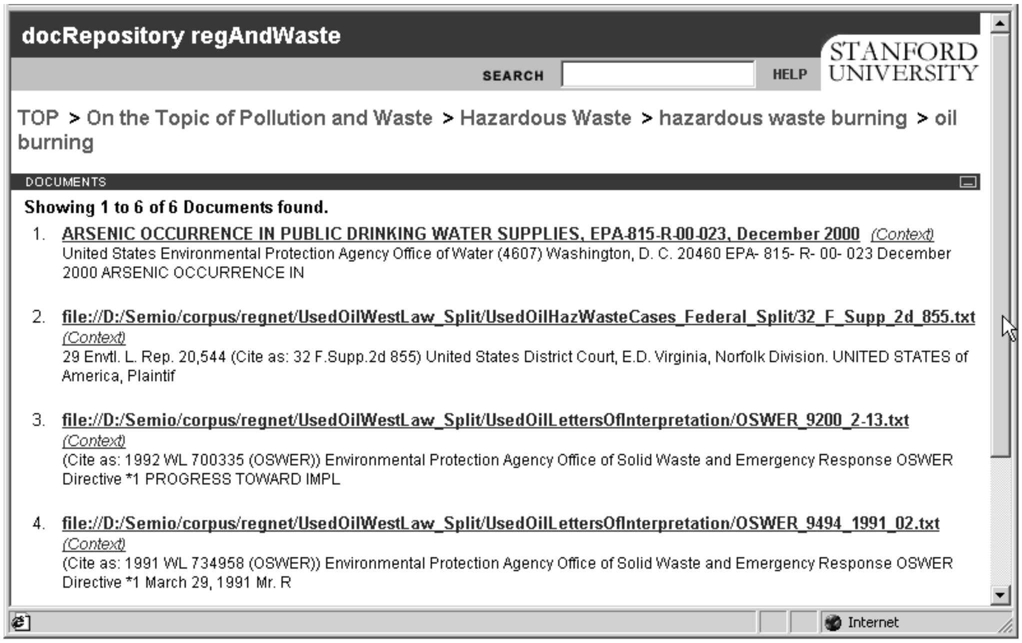
SUBCATEGORIES

- Pollution *(852)*
- Solid Waste *(510)*
- Exhaust *(154)*
- Waste Management *(267)*
- Incineration *(145)*
- Waste Disposal *(244)*
- Leaching and Leachate *(91)*
- Liquid Waste *(706)*
- Emissions *(537)*
- Hazardous Waste *(491)*
- Contamination and Decontamination *(1042)*
- Dumping *(70)*
- Industrial Waste *(81)*
- Municipal Waste *(95)*

CONCEPTS

- 1-b waste [1]
- accumulated waste [3]
- advanced waste treatment [4]
- aftertreatment system [3]
- alkaline waste [2]
- amber-list waste [1]
- amount of waste [18]
- animal waste [52]
- approved waste container [7]
- approving waste [1]
- ash waste [5]
- 2-a waste [1]
- acid waste [1]
- aftertreatment device [5]
- agricultural waste [14]
- allowable waste feed [2]
- amount of hospital waste [2]
- animal fecal waste [10]
- anticipated waste stream [1]
- approved waste [3]
- aqueous waste [3]
- assigned waste [1]

Figure 2.8 View of subcategories and concepts



docRepository regAndWaste

TOP > On the Topic of Pollution and Waste > Hazardous Waste > hazardous waste burning > oil burning

DOCUMENTS

Showing 1 to 6 of 6 Documents found.

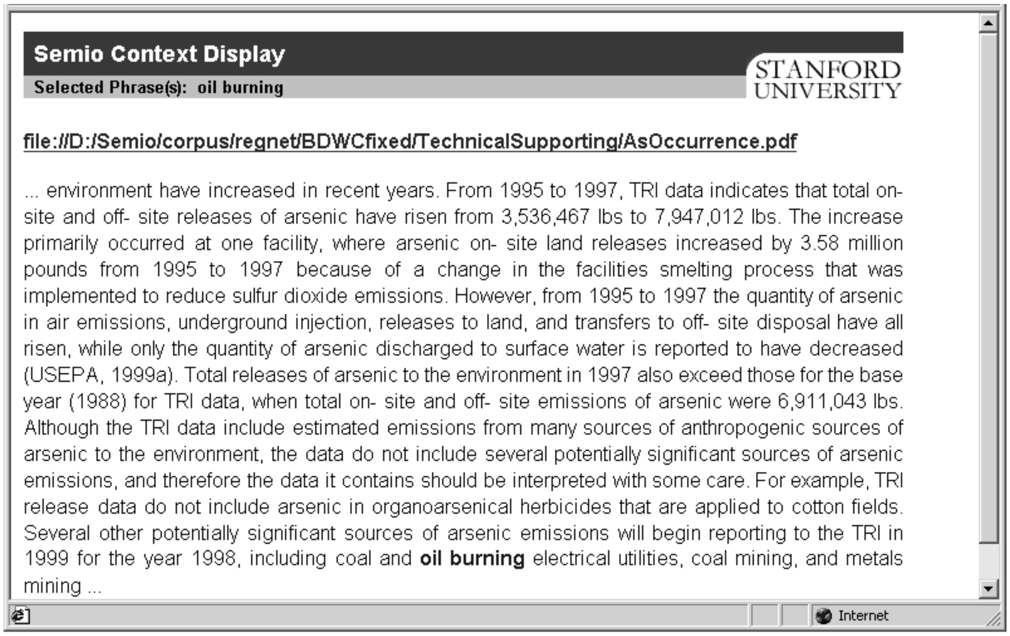
1. **ARSENIC OCCURRENCE IN PUBLIC DRINKING WATER SUPPLIES, EPA-815-R-00-023, December 2000** *(Context)*
   United States Environmental Protection Agency Office of Water (4607) Washington, D. C. 20460 EPA- 815- R- 00- 023 December 2000 ARSENIC OCCURRENCE IN

2. **file://D:/Semio/corpus/regnet/UsedOilWestLaw_Split/UsedOilHazWasteCases_Federal_Split/32_F_Supp_2d_855.txt** *(Context)*
   29 Envtl. L. Rep. 20,544 (Cite as: 32 F.Supp.2d 855) United States District Court, E.D. Virginia, Norfolk Division. UNITED STATES of America, Plaintif

3. **file://D:/Semio/corpus/regnet/UsedOilWestLaw_Split/UsedOilLettersOfInterpretation/OSWER_9200_2-13.txt** *(Context)*
   (Cite as: 1992 WL 700335 (OSWER)) Environmental Protection Agency Office of Solid Waste and Emergency Response OSWER Directive *1 PROGRESS TOWARD IMPL

4. **file://D:/Semio/corpus/regnet/UsedOilWestLaw_Split/UsedOilLettersOfInterpretation/OSWER_9494_1991_02.txt** *(Context)*
   (Cite as: 1991 WL 734958 (OSWER)) Environmental Protection Agency Office of Solid Waste and Emergency Response OSWER Directive *1 March 29, 1991 Mr. R

Figure 2.9 Links to documents



Semio Context Display

Selected Phrase(s): oil burning

file://D:/Semio/corpus/regnet/BDWCfixed/TechnicalSupporting/AsOccurrence.pdf

... environment have increased in recent years. From 1995 to 1997, TRI data indicates that total on-site and off- site releases of arsenic have risen from 3,536,467 lbs to 7,947,012 lbs. The increase primarily occurred at one facility, where arsenic on- site land releases increased by 3.58 million pounds from 1995 to 1997 because of a change in the facilities smelting process that was implemented to reduce sulfur dioxide emissions. However, from 1995 to 1997 the quantity of arsenic in air emissions, underground injection, releases to land, and transfers to off- site disposal have all risen, while only the quantity of arsenic discharged to surface water is reported to have decreased (USEPA, 1999a). Total releases of arsenic to the environment in 1997 also exceed those for the base year (1988) for TRI data, when total on- site and off- site emissions of arsenic were 6,911,043 lbs. Although the TRI data include estimated emissions from many sources of anthropogenic sources of arsenic to the environment, the data do not include several potentially significant sources of arsenic emissions, and therefore the data it contains should be interpreted with some care. For example, TRI release data do not include arsenic in organoarsenical herbicides that are applied to cotton fields. Several other potentially significant sources of arsenic emissions will begin reporting to the TRI in 1999 for the year 1998, including coal and **oil burning** electrical utilities, coal mining, and metals mining ...

Figure 2.10 Context for terms of interest

2.5 Related Research and Future Extensions

Information retrieval (IR) is a broad field that covers a range of tasks such as representing, storing, organizing, and retrieving information [5]. There has been a great deal of research done on information retrieval over the years, with an extensive body of literature available on this topic. Some of the topic areas that are most relevant to the research presented in this thesis are the areas of categorization systems [6, 27, 41, 50], visualizations [38, 48, 61, 71, 89], synonymy and polysemy [25, 47, 58, 103]. Many of the works cited in each of these areas have been discussed earlier in this chapter.

Much of our work has been focused on building a document repository as a demonstration of what can be done with modern information organization and retrieval tools. The core work introduced in this chapter builds the foundation for the regulation assistance system that will be discussed in Chapter 4. Our work on applying information retrieval tools to building a document repository has been focused on using categorization systems to create a core document repository with demonstrable usefulness. Information management is a very rich field, with a great deal of active research. Much of this research work could be applied to extend the work we have done on the document repository. This section will discuss some possible extensions.

- Visualizations: User interface design is an active area of research [38, 48, 61, 89]. There are many directions in user interface research for information visualization that might be interesting to apply to the document repository. In particular, the star tree interface, shown in Figure 2.11, provides an innovative way to navigate categorization hierarchies. This interface allows users to see multiple levels of the categorization hierarchy as well as separate branches of the hierarchy as they navigate the structure. Research work has shown that for problems with high information scent, or good "proximal cues to the value of distal information," the star tree interface can be very efficient for finding information [71].

- Navigation between documents: The repository could use tools to facilitate navigation between documents. First, the repository could attempt to automatically make available through hyperlinks any references appearing in a document being viewed. Second, the repository could automatically identify documents that appear similar to a document being viewed. A possible interface for navigating between documents is shown in Figure 2.12.

- Search tools: More powerful search tools can further enhance the usefulness of the document repository. For example, a more robust query system would be useful for querying the document repository with a number of parameters such as author, date, title or other structured parameter. The ability to include boolean or natural language queries would also be extremely useful. Searching for keywords

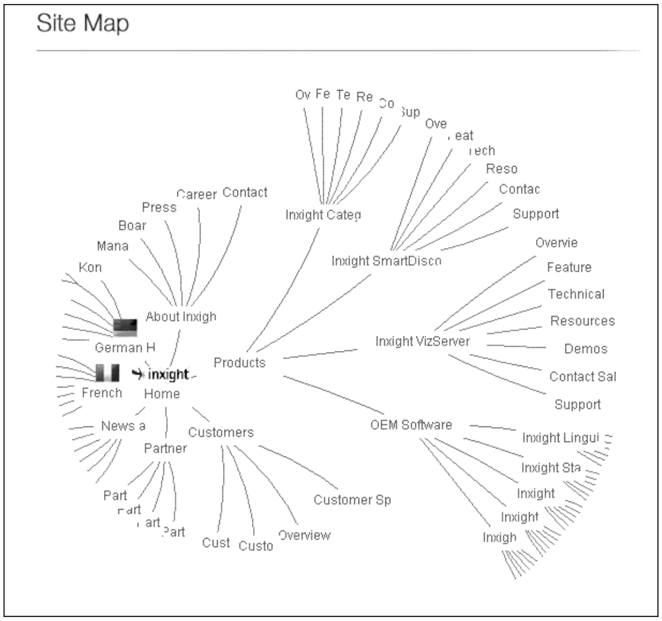
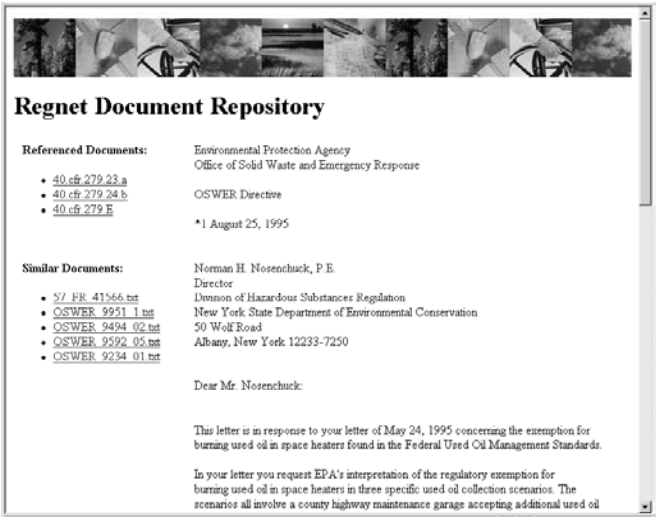


Figure 2.11 Inxight Star Tree[26]



Regnet Document Repository

Referenced Documents:

- 40.cfr.279.23.a
- 40.cfr.279.24.b
- 40.cfr.279.E

Similar Documents:

- 57_FR_41566.txt
- OSWER_9951_1.txt
- OSWER_9494_02.txt
- OSWER_9592_05.txt
- OSWER_9234_01.txt

Environmental Protection Agency
Office of Solid Waste and Emergency Response

OSWER Directive

*1 August 25, 1995

Norman H. Nosenchuck, P.E.
Director
Division of Hazardous Substances Regulation
New York State Department of Environmental Conservation
50 Wolf Road
Albany, New York 12233-7250

Dear Mr. Nosenchuck:

This letter is in response to your letter of May 24, 1995 concerning the exemption for burning used oil in space heaters found in the Federal Used Oil Management Standards.

In your letter you request EPA's interpretation of the regulatory exemption for burning used oil in space heaters in three specific used oil collection scenarios. The scenarios all involve a county highway maintenance garage accepting additional used oil

Figure 2.12 Possible interface extension for viewing documents

[26] The Inxight Star Tree, developed by Inxight Software, is available at the web address http://www.inxight.com/.

or phrases could be enhanced by index-mediated/thesaurus-assisted search techniques, along with other techniques for addressing the polysemy and synonymy problems.

- References: Supplementary documents tend to contain a significant number of references to regulations, sections of the United States Code (U.S.C), and other documents. These references are cumbersome to look up manually and reduce the readability of the documents. These references could be automatically linked by hyperlinks so that readers of the documents could locate the referenced documents more easily. Since many the references in the supplementary documents are references to regulations and sections of the U.S.C., a parsing system could be developed to assist with quickly identifying references. A parsing system for Code of Federal Regulations (CFR) references will be discussed in Chapter 3, and this system can be adapted to parse supplementary documents as well.

- Identifying similar documents: In addition to linking documents by reference, the repository could make available links to documents similar to the one being viewed. This would enable fast identification of possibly related documents. For example, documents created by federal entities may not explicitly reference state documents, but it is possible to identify related state documents by the similar concepts they contain. One approach would be to combine the noun-phrase extraction systems used in the categorization research work with a document similarity model to identify similar documents. The noun-phrase extraction system attempts to identify the “concepts”, or important noun-phrases and ideas that are discussed in the text. These concepts could then be used to determine the similarity between documents. There are a variety of methods for calculating document similarity, such as the vector model [5], that could use the concepts to identify similar documents. Such a document similarity model would enable the repository viewing system to make the related documents conveniently available using hyperlinks similar to the method described for references in the documents.

- Update notification: As new documents are created and added to the document repository, users may not learn of them unless they continuously recheck the document repository for new information. The repository could leverage its categorization structures to provide notification services for users when new supplementary information on regulations becomes available. Users could "track" the contents of the various categories in a categorization structure. Any time new documents were added to the "tracked" category, users could receive an email update regarding the updated contents of the category. That would enable users of the document repository to keep current on information in areas that they are concerned with. This feature would allow users to take advantage of the different perspectives offered by having different categorization structures covering the same set of documents. The categorization hierarchies described earlier in this paper are flexible units around which to build this document notification system, since multiple categorization hierarchies, each tailored to different audiences, can cover the same set of documents. This would allow for a finer granularity of notification than if a single categorization structure were used, since different users might track different sets of overlapping documents.

2.6 Summary

This chapter has presented the approach to building a document repository for environmental regulation. First, this chapter introduced the background information on supplemental regulatory documents and why these documents are important. Some terminologies from the information management field were also introduced. This was followed by a discussion of different types of categorization systems. The approach for developing classification structures was then introduced, and examples of classification structures that were created using this approach were provided. Finally, some of the more promising extensions possible for the document repository were discussed.

The document repository functions as a stand-alone resource for users who need to locate environmental regulatory information. The repository simplifies the task of identifying relevant regulatory information by assembling the information in one location and building useful categorization structures on top of it. Not only is the document repository a stand-alone resource, it provides a building block upon which other environmental regulatory software systems can be built, such as a compliance assistance system that will be discussed in later chapters.

Chapter 3

XML Representation of Regulations

3.1 Introduction

This chapter discusses the development of a formal framework to structure regulatory information such that it will be more amenable to software processing. Specifically, the eXtensible Markup Language (XML), which has emerged as a standard for data representation [65], is used to structure regulations and tag them with metadata. The XML structured framework lays the foundation for the regulation assistance system (RAS) discussed in the next chapter. The RAS makes use of the regulations annotated with meta-data to provide compliance assistance services. The development of the RAS system motivates the design of the XML structure and associated meta-data described in this chapter.

This chapter is organized as follows: First, various common approaches for the representation of documents and regulations are discussed. Second, the XML regulation structure developed in this thesis is introduced, and the parsers for converting regulations into the XML structure are described. Third, the meta-data elements added to the XML regulation, such as concepts, definitions, references and legal interpretations, are described in detail. The remainder of this chapter describes the XML structure for regulations and the annotation of this XML structure with meta-data.

3.2 Document Structures

Regulations currently come in many different document formats. Paper-based versions are the most commonly used document format for regulations. Electronic versions of regulations usually come in one of two different formats, either Portable Document Format (PDF) or Hypertext Markup Language (HTML). When we began this research work in August of 2000, PDF was the most popular way of distributing software copies of federal and state environmental regulations on compact discs (CD's) or through federal and state EPA websites. Since the start of this project, some agencies have begun to distribute regulations using HTML via the web.

There are a number of approaches available for storing and viewing textual information on computers. Some of the more common media are plain text, PDF, HTML and XML. Plain text, PDF, and HTML have important limitations in that they limit further processing with compliance-assistance software tools, so these formats are not the best formats to use for representing regulations.

Plain text documents are made up of unstructured text. This format is very simple and easy for people to work with and read, but offers very few features. Images, sounds, and other non-text elements cannot be incorporated into the document. Because plain text files are unstructured text, they are difficult for machines to process. Humans have no difficulties viewing the various components that make up the contents of a text file, but extracting the structures and semantic information from a text file is a challenging task for machines. Plain text files are not very useful for the representation of regulations if the regulations are to be further processed by software tools providing regulation compliance assistance services.

PDF files are more expressive than plain unstructured text files in that they allow images and some other non-text elements. PDF is a proprietary format designed by Adobe System Inc., so the standard is not available for customization to the regulation representation problem. PDF files only improve the visual appearance of information, they do not add structure or meaning to the information they contain. Because most regulations in PDF format are simply display-enhanced versions of text files, where the content of the regulation is not structured or tagged with semantic information, processing PDF files poses many of the same information extraction problems as unstructured text. The proprietary nature of the PDF format increases the complexity of the information extraction problem in that it is more difficult to perform information extraction tasks with PDF files than with plain text files. One recent paper on extracting references from research papers in PDF has concluded that, “analyzing PDF source is really hard” [12].

HTML is an open standard that allows the incorporation of display formatting, images, sounds and video [86]. Like PDF files and plain text, however, HTML does not provide structure or meaning to the document. It is possible to add some meaning through the use of “meta” or non-standard tags in HTML. For example, some commercial HTML programs use these tags to store editing information. However, adding “meta” or non-standard tags does not provide a clean, extendable structure. HTML is primarily a method for describing how data should be displayed. It does not effectively represent the conceptual structure or meaning of data.

XML is an open standard that allows designers to create any element needed to structure information in a file. Using XML, it is possible to structure the information in a document according to its conceptual meaning, as opposed to simply according to how it should be displayed. It is also possible to incorporate tags that may add images, sounds, and videos. In essence, XML offers the expressive capabilities of HTML, along with the ability to specify meta-information about the conceptual content of a document [65].

An XML document is made up of elements. All elements have start and end tags, and between these start and end tags an element may contain additional elements of the same or different type. In this way an XML document is organized hierarchically, as a tree structure, since each element has only one parent and tree branches cannot intersect. XML element start tags are of the form “<elementName>” and end tags are of the form “</elementName>”, where “elementName” is simply a placeholder for the name of the element. XML elements may also be represented with a single tag combining the start and end tag by using the syntax “<elementName/>”. The start tag for an element may contain attributes, which provide additional information specific to the element. Attributes are written in the form attributeName=“attributeValue”. For an accessible introduction to XML and examples of what it is generally being used for, please see Usdin and Graham’s article on XML [101].

In this research work, XML is chosen as the representational format for regulations, since it is well suited to the task of structuring and augmenting regulation text with content-related metadata. This is particularly true since regulations generally have a hierarchical nature. For example, we will later discuss tagging regulations with key conceptual phrases to enhance search or browsing, internal reference linking for easier navigation, definitions to clarify ambiguous terms, legal interpretations to clarify ambiguous provisions, and logical translations to facilitate automated compliance checking. An XML structure makes it possible to add all of these pieces of meta-data at the most appropriate location in the document. For example, some meta-data may apply to the entire document, while other meta-data may only apply to a single regulation provision. The XML structure described next allows regulations to be tagged and enhanced with the various features described above.

3.3 An XML Structure for Regulations

3.3.1 Overview

The XML framework enables the augmentation of a regulation with various types of regulation-specific metadata. This section discusses why XML is chosen for the regulation framework, how the framework is structured, and outlines each type of metadata that has been developed.

The XML-based regulation framework uses a nested structure of XML elements to represent each level of regulation text, such as subpart, section or subsection. This hierarchical structure mirrors the standard structure of federal and state environmental regulations. Parsing systems have been built to transform federal environmental regulations from Portable Document Format (PDF) and HTML into the XML format. The parser creates the core XML structure and populates each level of the framework with the appropriate level of regulation text.

Once the 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 discussed in this chapter are concept elements, reference elements, definition elements, and legal interpretation elements. These types of metadata are introduced below and later described in detail in Section 3.4. Additional types of meta-data elements for regulation logic and control processing that are used for compliance assistance are discussed in Chapter 4.

3.3.2 Base XML Structure for Regulations

An XML Document Type Definition (DTD) file provides grammar for an XML document. A DTD specifies the structure of an XML document by defining the elements of the document, and how those elements can be nested.

A standardized XML structure has been developed for representing regulations and can be validated against a DTD (the full specification of the DTD is shown in Appendix A). This DTD is designed to be applicable to all regulations, with a focus on federal, state and local environmental regulations.[27]

The XML markup used to tag regulations is at its core a nested structure that reflects the hierarchical structure of the regulations. Figure 3.1 shows an abbreviated XML regulation for 40 CFR Part 279, a regulation governing the standards for the management of used oil. The entire regulation is stored within a single “regulation” element. The “regulation” element has the attributes id, name, type, versionDate, and source. The attribute id identifies the reference to the represented regulation provision, in this case 40 CFR 279. All regulation references are transformed into a standard reference format with the “.” symbol separating components of the reference. For example, the reference 40 CFR 279.12(c)(1) is stored as “40.cfr.279.12.c.1”. The attribute “name” is the title of the regulation provision. The attribute type indicates the type of regulation; for example U.S. Federal, Illinois, California, etc. The attribute versionDate indicates the date that the source regulation, from which the XML regulation is modeled, was created. Regulations are documents that are modified over time, so including the date in the XML regulation is critical. Without the date information, it is difficult to tell if the XML regulation is current. The attribute source refers to the source for the regulation that has been used to create the XML regulation; the source might be a website, an ftp site, a printed-paper version, or a variety of other sources.

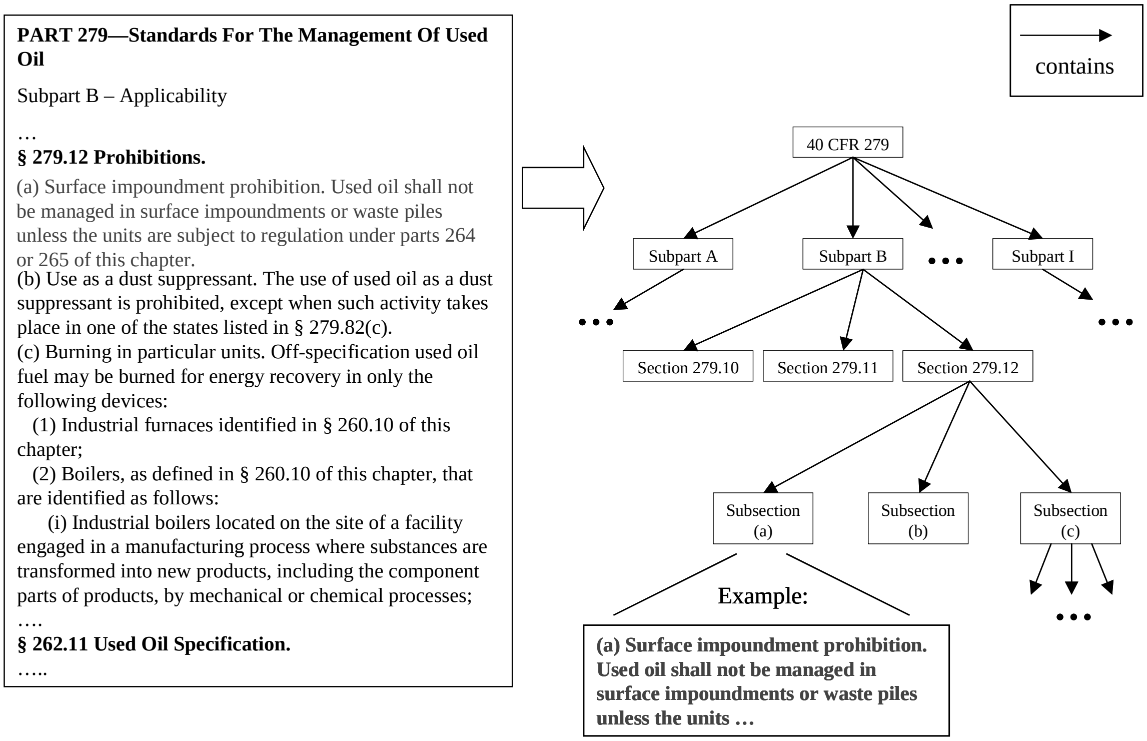
[27] The definition of the XML structure is a collaborative effort with the other members of the research team, Ms. Gloria Lau and Mr. Haoyi Wang, who work with building code regulations.

```
<regulation id="40.cfr.279" name="Standards For The Management Of Used Oil" type="US Federal"
  versionDate="January 24, 2003" source="http://www.access.gpo.gov/ecfr/" >
  …
  <regElement id="40.cfr.279.B" name="Subpart B">
    …
    < regElement id="40.cfr.279.12" name="Prohibitions">
      < regElement id="40.cfr.279.12.a" name="Surface Impoundment prohibition">
        <regText>
          <paragraph>
          Used oil shall not be managed in surface impoundments or waste piles…
          </paragraph>
        </regText>
      </regElement>
      <regElement id="40.cfr.279.12.b" name="Use as a dust suppressant">
      …
      </regElement>
    …
    </regElement>
    …
  </regElement>
  …
</regulation>
```

Figure 3.1 Abbreviated representation of a regulation provision

Within the “regulation” element, the regulation provisions are stored in nested “regElement” elements that specify the reference id and provision title within the hierarchy for all elements within the current element. The id is an identifier for reference to the provision. For example, one of the nested regElements in Figure 3.1 has the id 40.cfr.279.12. The name attribute for the element is the title for the provision, if one exists. For example, the title for the 40.cfr.279.12 element is “Prohibitions”. The first embedded “regElement” element shown in Figure 3.1 contains Subpart B of 40 CFR Part 279. Elements in the lower levels of the nested tree structure contain more specific provisions. This structure is shown graphically in Figure 3.2.

The modeling of all the structural elements within the regulation element as regElements rather than using different elements for parts, subpart, sections, etc. is a design decision to provide flexibility. If we were only interested in federal legislation, this flexibility would not be necessary. “Federal legislation is highly structured. The basic unit of legislation is the section. Sections can contain seven (7) levels of hierarchy within them (subsection, paragraph, subparagraph, clause, subclause, item, and subitem). Sections can also be within seven (7) higher levels (division, title, subtitle, chapter, subchapter, part, and subpart)”.[28] Our approach is designed to be generally applicable to a variety of regulations, some of which might not use the federal legislative format. As such, the XML structure proposed here can represent the tree structure of a regulation without forcing a naming hierarchy.



PART 279—Standards For The Management Of Used Oil

Subpart B – Applicability

…

§ 279.12 Prohibitions.

(a) Surface impoundment prohibition. Used oil shall not be managed in surface impoundments or waste piles unless the units are subject to regulation under parts 264 or 265 of this chapter.

(b) Use as a dust suppressant. The use of used oil as a dust suppressant is prohibited, except when such activity takes place in one of the states listed in § 279.82(c).

(c) Burning in particular units. Off-specification used oil fuel may be burned for energy recovery in only the following devices:

(1) Industrial furnaces identified in § 260.10 of this chapter;

(2) Boilers, as defined in § 260.10 of this chapter, that are identified as follows:

(i) Industrial boilers located on the site of a facility engaged in a manufacturing process where substances are transformed into new products, including the component parts of products, by mechanical or chemical processes;

….

§ 262.11 Used Oil Specification.

…..

(a) Surface impoundment prohibition. Used oil shall not be managed in surface impoundments or waste piles unless the units …

Figure 3.2 Diagram of how regulations are structured

[28] This quotation is from an online paper entitled, “Drafting Legislation Using XML at the U.S. House of Representatives.” This document is available at the web address http://xml.house.gov/drafting.htm, and was accessed on May 7th, 2003.

Within any provision specified by “regElement” elements, “regText” elements are used to store the actual text of the regulation provision. In Figure 3.1, part of the text for 40 CFR 279.12 (a) is shown to illustrate the usage of “regText” elements. This core XML markup provides a tree structure that enables regulation provisions to be tagged with meta-data at the appropriate level within the document. Once the type of structure shown in Figure 3.1 is established, it is possible to add other elements that tag the document with meta-data at any point in the XML tree structure.

The “regText” elements may contain several formatting elements: paragraph, table and pre elements.

- The paragraph element is used to denote paragraphs, and is similar to the <p> element in HTML. Paragraph elements may contain table, pre and paragraph elements just like regText elements.

- The pre element serves the same purpose in the XML regulation as the pre tag in HTML, denoting text that should be rendered verbatim; without any changes in spacing.[29]

- Tables are represented in the XML regulations in the same manner they are represented in HTML code. Table elements may be placed within regText and paragraph elements. Table elements may contain “tr” elements to designate rows of the table. Within the tr elements, “td” elements are used to represent individual cells of the table. Regular text, paragraph elements, pre elements or table elements may be placed within the td elements.

- Figures cannot be stored within the XML regulation structure, but references to figures may be stored in the XML structure. Figure references are stored in “img” elements, with a “source” attribute to provide the path to the figure. The source

[29] The HTML pre element is described by the W3C on a webpage located at http://www.w3.org/TR/REC-html40/struct/text.html#h-9.3.4

path can be a local directory path or a URL. Img elements may be used within regText, paragraph and tr elements. A DTD for all of these elements is shown in Figure 3.3.

3.3.3 Conversion of Regulations into the XML Structure

United States federal and state environmental regulations are used as a case study to investigate the usefulness of the XML structure. Two approaches have been taken in converting regulations to XML for this purpose. First, a parsing system has been built to convert PDF regulations to the XML regulation structure. Second, a parsing system has been developed to convert HTML regulations to the XML regulation structure. A discussion of how each of these systems operates, along with a comparison of the effectiveness of the two systems, is described in the following sections.

3.3.3.1 Converting PDF Regulations into XML Structure

In this work, a number of different environmental regulations were gathered in PDF format. For example, regulations composing CFR Title 40, Protection of the Environment, were downloaded from the federal Environmental Protection Agency's website[30] in PDF format and formed the core of the project's regulation corpus. In addition, sample PDF environmental regulations were downloaded from the websites of state agencies in Illinois, Florida, and New York. While regulations from each of these sources differ slightly in terms of parsing issues that need to be addressed, there are also strong commonalities that enable the development of a core parsing system that could then be specialized for a particular regulation source. The rest of this section discusses the operation of the core parser, which is specialized to focus on U.S. Federal EPA regulations (40 CFR).

[30] The EPA's website is located at http://www.epa.gov.

```
<!ELEMENT regulation (regElement+)>
<!ATTLIST regulation id ID #REQUIRED
                name CDATA #REQUIRED
                type CDATA #REQUIRED
                versionDate CDATA #REQUIRED
                source CDATA #REQUIRED>
<!ELEMENT regElement (regText?, regElement*)>
<!ATTLIST regElement id ID #REQUIRED
                name CDATA #REQUIRED>
<!ELEMENT regText (#PCDATA | paragraph | table | pre | img)*>
<!ELEMENT paragraph (#PCDATA | paragraph | table | pre | img )*>
<!ELEMENT table (td)*>
<!ELEMENT td (tr)*>
<!ELEMENT tr (#PCDATA | paragraph | table | pre | img )*>
<!ELEMENT pre (#PCDATA)>
<!ELEMENT img – O Empty>
<!ATTLIST source CDATA #REQUIRED>
```

Figure 3.3 DTD for structuring regulation text

In the first year of this project, federal EPA regulation documents were available on the web in a two-column PDF format and contained many words split across lines. An example of the two-column format is shown in Figure 3.4 with arrows indicating the words split across lines. A parsing system has been built to transform the original regulations in PDF format into XML-structured regulations. The parsing system consists of four basic steps, as shown in Figure 3.5.

In the first step the PDF regulations are converted into plain text documents using a PDF-to-text conversion tool called pdftotext, which is a free software program distributed under the GNU General Public License (GPL)[31]. This conversion results in a text file with several remaining issues that need to be resolved before it can be translated into the XML regulation format. The most significant issues are that the converted text regulations have multiple columns of text with a large number of words that are

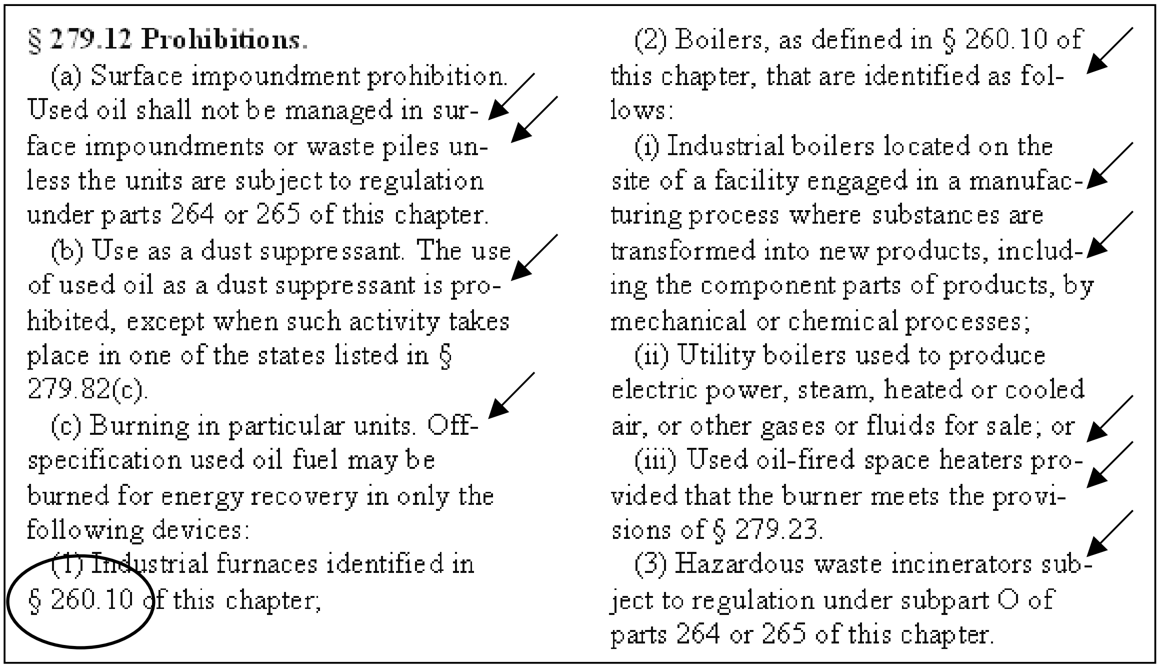
[31] The pdftotext program is available at the web address http://www.foolabs.com/xpdf/.

hyphenated and split across columns, and there are no clear identifiers for the section headings.

Step two of the XML conversion process unifies the multicolumn format and reassembles split words with the help of an online Webster dictionary. First the page breaks in the document are identified by the white space between them. Next, the columns of text are identified by the white space between them. Then the right column of text is concatenated below the left column of text. During this process, words that had been hyphenated and split across lines are reconstructed with the assistance of a Webster dictionary, which is used to verify that reconstructed words are in fact valid words.

Step three of the conversion process is to parse the table of contents, which lists the sections contained in the document, to increase the accuracy of the section identification process. Frequently section headers in the text document are difficult to identify. References appear throughout the regulation, so it is difficult to differentiate between section numbers appearing in the body of the document that have been wrapped to a new line and section numbers that identify the start of a new section. This is shown in Figure 3.4 by the reference to §260.10, which appears at the bottom of the left column. By parsing the table of contents, the conversion process can gain information on what regulation sections need to be identified.

The fourth step in the conversion process parses through the cleaned text-based regulations and outputs the regulations with XML tags that define the document structure. Identification of sections takes advantage of the information provided by the table of contents to improve matching accuracy. Identification of subsections and deeper provisions is generally easier, and the parsing system attains high accuracy for this task despite not having a table of contents to assist it. As the parser moves through the text document, it writes an XML file that contains XML tags to indicate the structure of the document along with the original text.



§ 279.12 Prohibitions.

(a) Surface impoundment prohibition. Used oil shall not be managed in surface impoundments or waste piles unless the units are subject to regulation under parts 264 or 265 of this chapter.

(b) Use as a dust suppressant. The use of used oil as a dust suppressant is prohibited, except when such activity takes place in one of the states listed in § 279.82(c).

(c) Burning in particular units. Off-specification used oil fuel may be burned for energy recovery in only the following devices:

(1) Industrial furnaces identified in § 260.10 of this chapter;

(2) Boilers, as defined in § 260.10 of this chapter, that are identified as follows:

(i) Industrial boilers located on the site of a facility engaged in a manufacturing process where substances are transformed into new products, including the component parts of products, by mechanical or chemical processes;

(ii) Utility boilers used to produce electric power, steam, heated or cooled air, or other gases or fluids for sale; or

(iii) Used oil-fired space heaters provided that the burner meets the provisions of § 279.23.

(3) Hazardous waste incinerators subject to regulation under subpart O of parts 264 or 265 of this chapter.

Figure 3.4 Double-column regulation provision with words split across lines

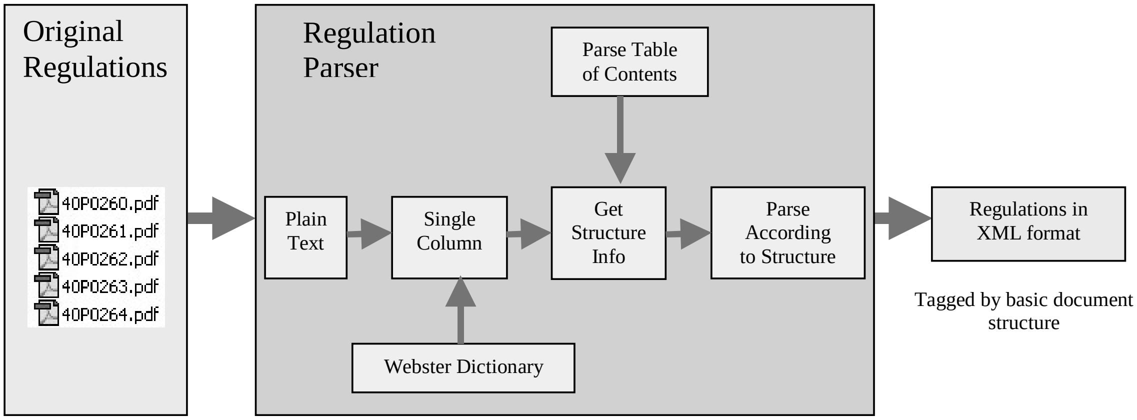


Figure 3.5 Conversion of plain text regulations to XML format

The XML regulations resulting from this process are generally imperfect and require manual modification to accurately represent the regulations. The PDF to XML conversion process has three primary drawbacks. First, the initial PDF-to-text conversion results in a loss of formatting information, which introduces a number of problems. Loss of formatting information means that specially formatted identifiers, such as section headings being rendered in bold, cannot be utilized in the subsequent parsing steps. In addition, tables tend to become scrambled beyond recovery by the PDF-to-text conversion. Second, the information extraction task is difficult since there are only weak indications of the regulation's structure in the text documents. This is complicated by the fact that the PDF-to-text conversion introduces noise into the document in the form of irregular spacing and occasionally scrambled text. Third, figures and other non-text components of the regulations are lost. Significant manual intervention is required to deal with all the problems mentioned above, resulting in a time-consuming conversion process.

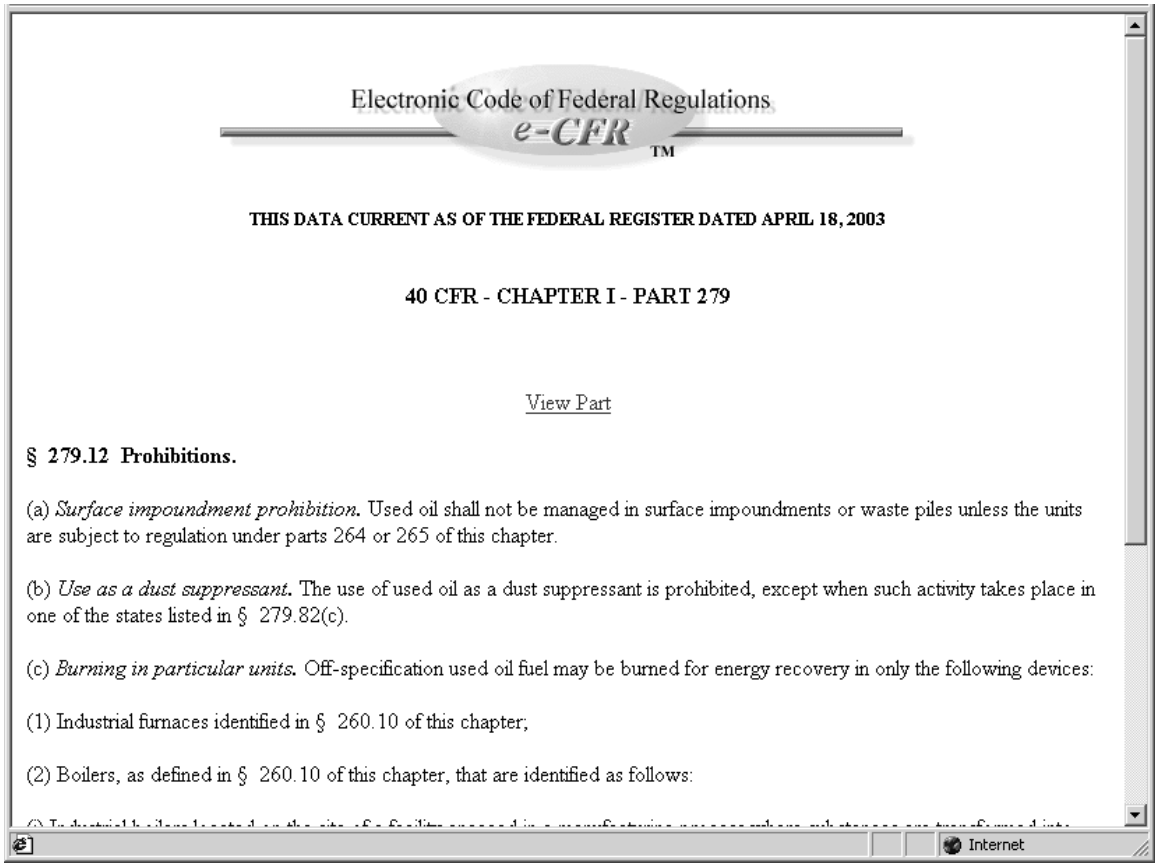
3.3.3.2 HTML to XML Conversion

Recently, government agencies have begun to make available regulatory information on the web in HTML format. After HTML regulations became available, we developed a parser to convert HTML regulations downloaded from the Electronic Code of Federal Regulations (e-CFR) website into XML.[32] Figure 3.6 shows a sample HTML regulation from the e-CFR website. Two hundred and eighty-seven regulation Parts, all of the regulation Parts within 40 CFR that were available through the e-CFR website, were downloaded in January of 2003 for the development and testing of a HTML to XML regulation parser.

As shown in Figure 3.7, the HTML to XML conversion process is a three-step process. This process has significant reliability improvements over the PDF to XML process. The

[32] The e-CFR website is maintained by the U.S. National Archives and Records Administration at the web address http://www.access.gpo.gov/ecfr/. All regulations were downloaded in January of 2003.

first step is to trim down the unnecessary information in the HTML file and begin preparing the file for further processing into an XML structure. This step removes the HTML tags that will not be used in the conversion process. For example, the HTML tag “font” is not necessary for the conversion, so it can be removed. Any characters that are illegal in XML are also removed or substituted with legal representations at this point. For example, the “&” character is replaced with the legal XML entity representation “&amp;”. This step also removes regulation content that is not needed for the final XML document, such as the table of contents for the regulation Part.



THIS DATA CURRENT AS OF THE FEDERAL REGISTER DATED APRIL 18, 2003

40 CFR - CHAPTER I - PART 279

§ 279.12 Prohibitions.

(a) *Surface impoundment prohibition.* Used oil shall not be managed in surface impoundments or waste piles unless the units are subject to regulation under parts 264 or 265 of this chapter.

(b) *Use as a dust suppressant.* The use of used oil as a dust suppressant is prohibited, except when such activity takes place in one of the states listed in § 279.82(c).

(c) *Burning in particular units.* Off-specification used oil fuel may be burned for energy recovery in only the following devices:

(1) Industrial furnaces identified in § 260.10 of this chapter;

(2) Boilers, as defined in § 260.10 of this chapter, that are identified as follows:

(i) Industrial boilers located on the site of a facility engaged in a manufacturing process where substances are transformed into

Figure 3.6 Initial HTML regulation from e-CFR

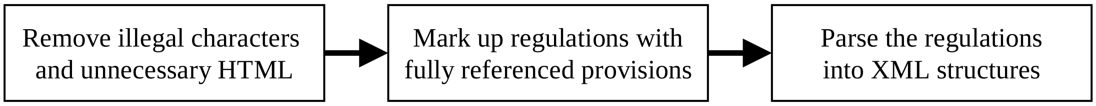

Remove illegal characters and unnecessary HTML
Mark up regulations with fully referenced provisions
Parse the regulations into XML structures

Figure 3.7 Process for converting HTML regulation to XML

The second step in the conversion process involves detecting the structure in the regulation file, and adding information to the file to make the regulation structure more explicit. The pattern matching capabilities of Perl, a programming language, are well suited for this type of text processing [105]. Pattern matching is used to identify the hierarchical structure of the regulation, assisted by the HTML formatting tags used in the e-CFR regulations. For example, HTML tags for displaying text in bold (<b>) mark section headings, resolving the problem of identifying section references that have wrapped to a new line (the circled text "§260.10 of this chapter" illustrates this issue in Figure 3.4). As each component of the outline structure is identified in the file, a full reference to the provision is inserted at the start of the line to be used in the final conversion step. This second step produces a regulation file with each provision of the regulation text tagged with a complete reference to its location within the regulation tree structure.

The third step in the conversion process involves transforming the regulation file into the XML structure. This process is facilitated by the tagging of each regulation provision with its "id", or full reference path, that was done in step two of the conversion process. The parser still makes use of some remaining HTML tags at this stage to ensure a clean transformation of the regulation. For example, the parser is able to distinguish provision titles from provision text because the former is identified with italics tags from the original e-CFR HTML regulation. This enables the proper "name" data to be inserted in the XML regulation. Figure 3.6 shows the formatting of the initial HTML regulation. From all of this information the parser is able to convert the regulations into the XML structure.

The HTML to XML conversion process reduces most of the PDF to XML conversion problems. First, the HTML parser can take full advantage of formatting information that is part of the HTML regulation. Table information in particular is straightforward to process from HTML, so this information is not lost. Second, the information extraction process is somewhat easier than in the case of PDF conversion because HTML tags help delineate the document's structure. For example, the start of sections may have tags for paragraphs or other formatting dividers. Third, figures and other non-text components of the regulations are straightforward to preserve from the HTML. Since all of these issues are improved by means of the HTML to XML conversion process, significantly less manual intervention is required to convert HTML regulations to XML than is required to convert PDF regulations. Therefore, the HTML to XML conversion is both less time-consuming and less error-prone than the PDF to XML conversion. We were able to transform all 287 regulation Parts within 40 CFR that we downloaded without much manual intervention using the HTML to XML process described above.[33]

3.4 Adding Metadata to XML-Structured Regulations

3.4.1 Overview

A key design paradigm behind this research work is that by bringing all the meta-data directly to the regulations a highly portable multi-use document can be constructed, and the usefulness of the complete and integrated document will be greater than the individual parts. For example, incorporating concept phrase elements directly into the document allows processing systems to provide a number of features such as automated

[33] This conversion process took about 3 hours on a Pentium III Linux computer.

document linking or similarity analysis. This design paradigm should facilitate the development of regulation documents that are rich in data content along with processing engines that operate from a common data standard. This section describes some of the key pieces of meta-data that are added to the XML regulations.

3.4.2 Concepts

One type of metadata added to the regulations is an XML element to denote the conceptual content of a regulation provision. The word “concept” in this context refers to noun-phrases occurring in the regulation that are indicative of the topic being covered in the regulation. For example, the concept phrases “waste pile” and “surface impoundment” provide reasonable indications of what issues a regulatory provision covers.

There are many tools available that may be used to identify key concept noun phrases in a document. The techniques commonly used by these document analysis systems to identify key concept noun phrases include techniques such as sentence structure analysis, word frequency analysis, and word collocation analysis. The research work described in this thesis makes use of a commercial software product, Semio Tagger, to automatically identify concepts in the regulations. This is the same software package used for the categorization purposes described in Chapter 2.

We use Semio Tagger to extract a list of noun phrases from our document repository text corpus. The text corpus consists of all the regulations in 40 CFR, plus court cases and other supplementary documents related to these regulations. Semio Tagger originally extracted 67,106 concepts from the repository. These concepts were then manually scanned to remove noun phrases that were clearly not useful, such as “figure b114-1” and “subparts c-d”, such that the list was reduced to 65,857 noun phrases.

We have developed a program to read through the regulations and apply the Apache Jakarta Lucene word stemmer [24]. Word stemming is the process of reducing words down to their word stems. Reducing words to their stems allows software systems to better match words that would not otherwise match. For example, using word stemming we can match “waste” and “wastes”, or “disposal” and “dispose”. Information retrieval systems often use word stemming as part of their document retrieval process. Since we envision the concepts we are adding will be useful for retrieving documents and performing similarity analysis, word stemming is an important tool for adding concept metadata to the regulations. The Lucene PorterStemmer implements the simple and efficient Porter algorithm [72] to stem words. The Porter algorithm uses a series of rewrite rules for a word to reduce the word down to its word stem.

The basic procedure in adding concepts into the XML document is as follows. First, both the words in the regulation text and the concept list of noun phrases are stemmed using the Porter stemming algorithm. In places where the stemmed regulation text matched the stemmed concept phrase, the unstemmed form of the concept is then inserted in the XML regulation to annotate the matching regulation text. The annotation is done by adding an XML element containing the concept within the regElement for the regulation provision. An example of XML concept elements added for this purpose is shown in Figure 3.8.

```
<regElement id="40.cfr.279.12.b" name="Use as a dust suppressant">
    <regText>
    Used oil shall not be managed in surface impoundments or waste piles unless the units are subject
    to regulation under parts 264 or 265 of this chapter.
    </regText>
    <concept name="waste pile" />
    <concept name="surface impoundment" />
</regElement>
```

Figure 3.8 Example of concept XML element

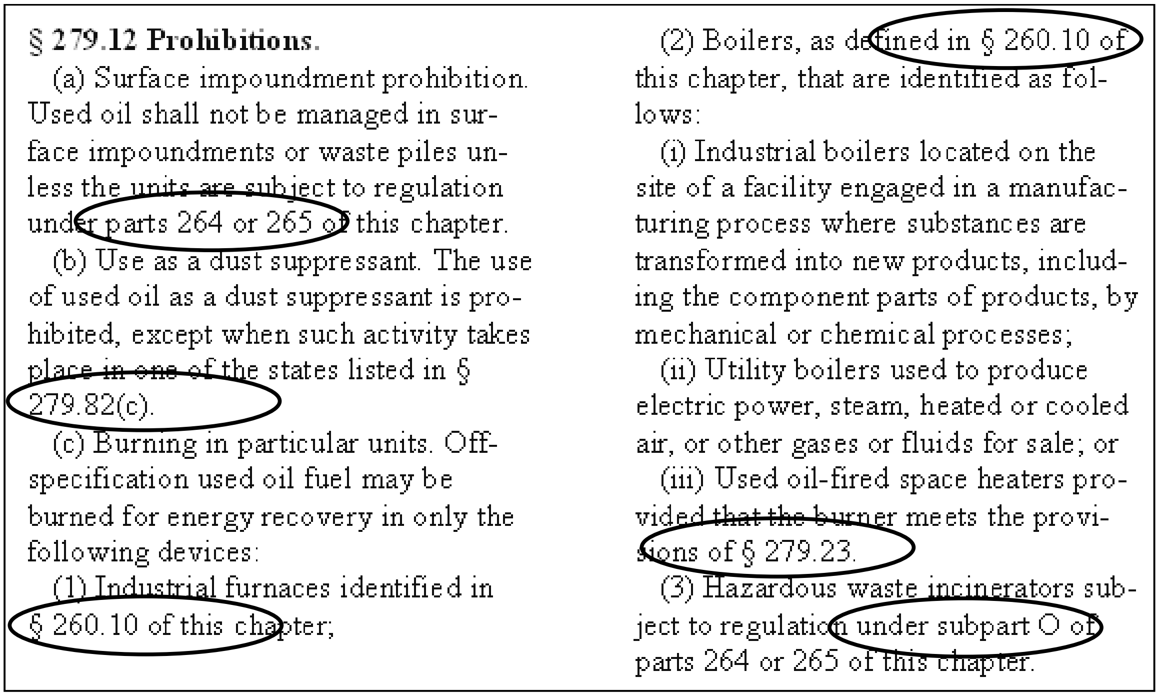
Tagging regulations with key conceptual phrases enables two important usages. First, if all the provisions within a regulation are tagged with conceptual phrases, it becomes possible to identify similar or related regulation provisions that do not explicitly reference each other. For example, federal regulations generally do not reference state regulations, yet someone reading a federal environmental regulation might want to compare it with a California environmental regulation. If the California regulations had a different structure or were composed of separate regulation components issued by different governing bodies, comparing the California regulations with the federal regulation will be a difficult task. If all the relevant regulations are tagged at the provision level with their key conceptual phrases, reconciling the myriad of regulation content structures and organizational origins may become a tractable problem. Even though some terminology differs, in general similar regulation provisions should have a number of overlapping key conceptual phrases, so that the related federal and California regulation provisions should be identifiable.

Second, tagging regulations with key conceptual phrases enables a tight integration with the document repository. The concepts can be considered predefined search terms. Supplementary documents in the document repository that also share a concept or multiple concepts with a regulatory provision may be strongly related to that regulation. Using this idea of predefined search terms means that the XML regulation does not need to explicitly reference every related document individually. Instead, one only needs to ensure that the important supplementary documents share concepts with the related provision. This approach also has the effect that as documents are added to the document repository they immediately become implicitly “linked” from any regulation with which they share concepts.

3.4.3 References

A second type of metadata applied to the XML framework addresses the references embedded in the regulation text. Regulation provisions tend to contain a large number of casual English references to other provisions. The density of cross-referencing is illustrated in Figure 3.9, which shows a sample regulation provision with all cross-references circled. These references are cumbersome to look up manually, reducing the readability of the regulation text itself. Moreover, these natural language references are difficult for software to make use of; the reference information is not very convenient for software programs to interpret. If references are explicitly marked throughout the XML regulations using a standard format, tools making use of reference data can be more easily constructed.

Tools that can make use of regulation references include regulation viewing systems that link the references with hyperlinks [66], similarity analysis systems that use references in their algorithms [17, 52], and regulation analysis systems that investigate the structure of



§ 279.12 Prohibitions.

(a) Surface impoundment prohibition. Used oil shall not be managed in surface impoundments or waste piles unless the units are subject to regulation under parts 264 or 265 of this chapter.

(b) Use as a dust suppressant. The use of used oil as a dust suppressant is prohibited, except when such activity takes place in one of the states listed in § 279.82(c).

(c) Burning in particular units. Off-specification used oil fuel may be burned for energy recovery in only the following devices:

(1) Industrial furnaces identified in § 260.10 of this chapter;

(2) Boilers, as defined in § 260.10 of this chapter, that are identified as follows:

(i) Industrial boilers located on the site of a facility engaged in a manufacturing process where substances are transformed into new products, including the component parts of products, by mechanical or chemical processes;

(ii) Utility boilers used to produce electric power, steam, heated or cooled air, or other gases or fluids for sale; or

(iii) Used oil-fired space heaters provided that the burner meets the provisions of § 279.23.

(3) Hazardous waste incinerators subject to regulation under subpart O of parts 264 or 265 of this chapter.

Figure 3.9 Illustration of the density of cross referencing within 40 CFR

the regulations. If regulations are not annotated with references in a standardized format in advance, natural language reference tracking capabilities would be needed. As the number and type of regulations increase, extracting the references can be even more complicated since different regulations will have slightly different referencing styles. Adding reference data at the time an XML regulation is created reduces the complexity for the development of other processing systems which need to use the references.

The complexity of regulation references ranges 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 can be converted manually into the following list of complete references: 40.cfr.265.G, 40.cfr.265.H, and 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 for existing regulations. The same problem of dealing with a huge number of natural language references has been faced by at least one other researcher, Justin Needle, when he was working with JUSTIS, a legal research data provider[34]. In an article on the automatic linking of legal citations, Justin Needle writes [66]:

> "The conventional method of creating hypertext links between documents involves manually editing each document and inserting fixed links at the database production stage. Unfortunately, there is a major problem. The JUSTIS databases contain millions of citations which, in order to achieve the required functionality, need to be converted into millions of corresponding hypertext links. The manual creation of links on this scale is not really an option since link creation is a laborious process, requiring the services of skilled, and expensive, editors. Even if an editor is able to identify and process ten links per hour, which is optimistic, then the human effort required will be approximately a hundred thousand hours

[34] JUSTIS is available at the web address http://www.justis.com.

> per million links created. … Clearly, there ought to be a better way of tackling the problem.”

For the research work presented in this thesis, 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 that they contain. The parsing system consists of two phases. First, a context-free grammar parser scans through the regulation text, constructing parse trees as shown in Figure 3.10. The original reference for the parse tree in Figure 3.10 was, “… Subpart O of part 264 or 265”. Then a secondary parser converts the parse tree into lists of fully specified references. These references are inserted into the XML regulation as new child elements of the appropriate regElement XML element.

Note that the references are not tagged as hyperlinks, which would tie the reference to a particular source for the referred document. Rather, the reference tags simply 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. Examples demonstrating the usage of the XML element for regulation references are shown in Figure 3.11. The next section discusses the development of the reference parser in detail.

3.4.3.1 Development of a Reference Parser

3.4.3.1.1 Simple Tabular Parsing

Parsing can be viewed as a search problem of matching a particular grammar and lexicon to a set of input tokens by constructing all possible parse trees [46]. The grammar defines a set of categories and how the categories can be manipulated. The lexicon

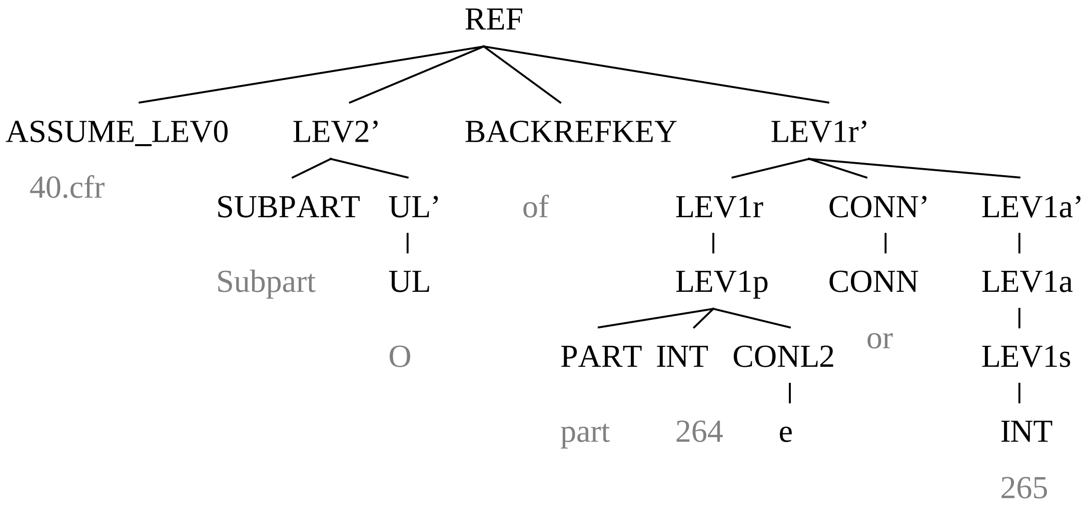


Figure 3.10 Example parse tree for identifying regulation references

```
<regElement id=“40.cfr.279.12.a” name=“Surface impoundment prohibition” >
      <regText>
      Used oil shall not be managed in surface impoundments or waste piles unless the units are subject to
      regulation under parts 264 or 265 of this chapter.
      </regText>
      <reference id=“40.cfr.264” />
      <reference id=“40.cfr.265” />
</regElement>
```

Figure 3.11 Example of a reference XML element

defines to what categories the input tokens belong. The search problem is associated with manipulating the grammar to find all possible matches with the input tokens. A simple top-down, left-to-right parser is described in this section.

Suppose we start with a very simple model of English grammar. In this grammar we might say that all sentences are composed of a noun plus a verb phrase. Verb phrases can be a verb plus a noun, or simply a single verb. This grammar can be represented as shown in Figure 3.12. We can then create a small lexicon containing the words “cars”, "oil", and "use", in which we define what categories these words may match. This simple lexicon is shown in Figure 3.13.

We can use the simple grammar and lexicon to parse the sentence "cars use oil". We can model the parsing process with a category stack, an input stack, and a set of operations for manipulating these stacks. We start the parsing by adding the "S", or a sentence start symbol, to the category stack and the input tokens to the input stack. We can then use two operations to parse the input. The expand operation is used to expand the top category on the category stack using one of the grammar rules in Figure 3.12. The match operation is used to match the top category in the category stack with the top token in the input stack according to the lexicon rules in Figure 3.13. A parse is considered successful when both the category stack and the input stack are empty. Table 3.1 shows the successful parsing of the sentence "cars use oil", and Figure 3.14 shows the corresponding parse tree.

In this simple tabular parsing strategy, it is necessary to try all possible expansions of the grammar categories. For example, the "VP" category in Table 3.1 could also have been expanded to be a "V". Since this expansion would not have resulted in a successful parse the expansion was not used in the example in Table 3.1. When a program is searching for a parse for an input stack, it will not know in advance which expansions will result in a successful parse. Therefore it must perform all possible expansions. The general procedure used, however, is the same as that illustrated in Table 3.1. It is possible to have multiple parses for a single set of input tokens. It should also be noted that the lexicon may map the same input token to multiple categories. The parser design described here is known as left-recursive, and if a grammar rule is left recursive the parsing algorithm will not terminate [46]. This is because a rule like "VP → VP N" can be expanded an infinite number of times. In our work, left recursive grammar rules are not permitted.

Table 3.1 Simple parsing example

Category Stack	**Input Stack**	**Operation**
S	Cars use oil	Start
N VP	Cars use oil	Expand
VP	use oil	Match
V N	use oil	Expand
N	Oil	Match
		Finish

S → N VP
VP → V N
VP → V

Figure 3.12 Simple grammar

N → cars
N → oil
V → use
N → use

Figure 3.13 Simple lexicon

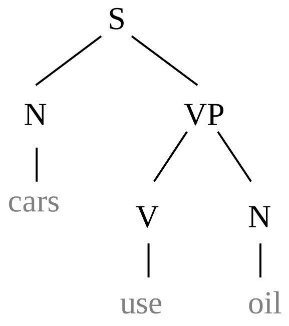


Figure 3.14 Simple parse tree

3.4.3.1.2 Constructing a Reference Parse Tree

As mentioned above, the reference parsing system is based on a simple tabular parser. This simple tabular parser must be adapted to the reference identification problem. The simple tabular parser knows the start and end of the sentence in advance, and incorporates this information into the algorithm. The reference identification problem is different from general sentence parsing in that the start and end of the reference are not known in advance. The termination conditions for the reference parser are changed so that the parse is considered complete if the category stack is empty. In other words, the input stack does not need to be empty for the parse to be complete. The parser was also modified to recognize a number of special category tokens in addition to the lexicon. These special categories, which can be used in addition to a vocabulary specified in the lexicon, are shown in Table 3.2. In addition, grammar specifications can use special categories such as "txt(abc)" to match "abc" input, which makes some patterns more transparent by bypassing the lexicon.

A third type of grammar category was of the form "ASSUME_LEV1", where LEV1 represents the level within the reference hierarchy. When the parser produces a category that starts with "ASSUME_" during a parse, it matches the assumed reference level (whatever follows the underscore after ASSUME, in this case "LEV1") with the identifier for that level within the XML tree[35]. This is useful when a natural language reference does not fully denote the reference. For example, in Figure 3.10 the category "ASSUME_LEV0" was automatically assumed to be "40.cfr", because the reference "… Subpart O of part 264 or 265" did not explicitly state that parts 264 and 265 are in 40 CFR.

A WordQueue object is used to tokenize and buffer the input. Regulation provisions are read individually and added into the WordQueue. The tokenized regulation provision is

[35] This is similar to the standard treatment of "empty" rules, except an assumed value is matched.

Table 3.2 Special reference parsing grammar categories

Category	**Matches**
INT	Integers
DEC	Decimal numbers
NUM	Integers or decimal numbers
UL	Uppercase letters
LL	Lowercase letters
ROM	Roman numerals
BRAC_INT	Integers enclosed in ()
BRAC_UL	Uppercase letters enclosed in ()
BRAC_LL	Lowercase letters enclosed in ()
BRAC_ROM	Roman numerals enclosed in ()

then passed to the parser to look for a reference. If a reference is found, the tokens that constitute the reference are removed from the queue. Otherwise, the first token in the queue is removed and the input is returned to the parser. Once the input queue is empty, the next regulation provision is read.

Many experiments have been conducted to develop the algorithm for tokenizing the regulation provision text input. Splitting first on white space and then splitting off any trailing punctuation will not work. Some of the text includes lines like, “oil leaks (as in §§279.14(d)(1)). Materials…”. The tokenized version of this line (using a space delimiter) should look like, “oil leaks (as in § § 279.14 (d) (1)) . Materials”. The algorithm cannot split on all “.” characters because some may occur as part of a number. It cannot split on all the parenthesis characters because some may be part of an identifier “(d)” that should be preserved. The solution is to first split the input on white space, and then perform a second pass on each individual token. This second pass involves splitting the token into smaller tokens until no more splits are possible. The process splits off starting “§” symbols, trailing punctuation, unbalanced opening or closing parenthesis[36], and groups of characters enclosed in parenthesis.

An iterative process is used to develop the reference parsing grammar. First, a core grammar and lexicon are created by manually reading through the regulations and developing a grammar and lexicon to parse the manually identified references. Next a reference prediction system, discussed in section 3.4.3.2, is run on a large regulation text corpus to produce lists of text with high probability of containing a reference that the system could not parse. Actual references are then manually identified from this list and the parsing grammar and lexicon updated to capture the references. This process is repeated until the reference prediction system fails to find any real references that could not be parsed. Figure 3.15 shows the basic grammar and Figure 3.16 shows the basic lexicon.

3.4.3.1.3 Interpreting the Reference Parse Tree

Once a parse tree is created using the parsing algorithm described above, the problem remains of interpreting this parse tree so that references can be listed in a standard format. This section describes the process used to convert parse trees into lists of references.

The semantic parsing system is built on top of a simple tabular parser that does a modified depth-first processing of the parse tree. Each node in the tree is treated as an input token. The processing deviates from strict depth-first processing when special control categories are encountered. Grammar and lexicon files provide control information to the semantic interpreter. The parsing algorithm differs from a simple tabular parser in that when a category label is found, it is not removed from the category search stack. Instead, the category found is marked “found” and remains on top of the

[36] This means splitting off unequal numbers of “(“ or “)” on a token. For example, “(d))” is split into “(d)” and “)”.

```
REF --> LEV0'
REF --> ASSUME_LEV0 LEV2' BackRefKey LEV1r'
LEV0' --> LEV0
LEV0' --> LEV0 CONN' LEV0'
LEV0 --> INT CFR LEV1a'
LEV1a' --> LEV1a CONN' LEV1a'
LEV1a' --> LEV1a
LEV1a --> LEV1p
LEV1a --> LEV1s
LEV1p --> PART INT CONL2
LEV1r' --> LEV1r CONN' LEV1a'
LEV1r --> LEV1p
LEV1s --> INT
CONN' --> CONN
CONL2 --> txt(,) LEV2'
CONL2 --> e
LEV2' --> SUBPART UL'
UL' --> UL
UL' --> UL CONN' UL'
```

Figure 3.15 Partial grammar for the reference parsing system

stack. The next matching category can be the “found” category or the second category in the stack. If the second category in the stack is matched, it is marked found and the top category is removed.

The grammar file is essentially a list of templates that specify what type of reference is well formed. All grammar rules for the parser that interprets the reference parse trees must start with “REF --> “. The grammar used for interpreting the parse trees is shown in Figure 3.17.

The two grammar rules in Figure 3.17 correspond to the two types of references that appear in 40 CFR regulations: 40 CFR 262 Subpart F (which refers to Chapter 40, Part 262, Subpart F), and 40 CFR 262.12(a)(13)(iv) (which refers to Chapter 40, Part 262, Section 12, subsection a, paragraph 13, subparagraph iv)

```
CONN --> and
CONN --> or
CONN --> ,

PART --> part
PART --> parts
PART --> Part
PART --> Parts

SUBPART --> subpart
SUBPART --> subparts
SUBPART --> Subpart
SUBPART --> Subparts

BackRefKey --> of
BackRefKey --> in

CFR --> CFR
CFR --> cfr
```

Figure 3.16 Partial lexicon for the reference parser

```
REF --> LEV0 LEV1 LEV2
REF --> LEV0 LEV3 LEV4 LEV5 LEV6 LEV7
```

Figure 3.17 Reference interpretation grammar

The lexicon file specifies how to treat different parsing categories. A shortened version of the interpretation lexicon is shown in Figure 3.18. The complete lexicon appears in Appendix B. As shown in Table 3.3, there are five semantic interpretation categories that can be used in the lexicon. These categories are used to classify the categories used by the reference parser when constructing the parse tree.

Table 3.3 Lexicon categories

Category	**Meaning**
PTERM	Indicates the node is a printing terminal string (to be added the reference string currently being built)
NPTERM	Indicates the node is a non-printing terminal string (the node is ignored)
SKIPNEXT	Indicates the next child node of parent should be ignored and not processed
REFBREAK	Indicates the current reference string is complete, and a new reference string should be started
INTERPOLATE	Indicates that a list of references should be generated to make a continuous list between the previous child node and the next child node. (If the child node sequence was “262, INTERPOLATE, 265”, this would generate the list “263, 264”)

```
PTERM --> INT
PTERM --> CFR
PTERM --> UL

NPTERM --> PART
NPTERM --> SUBPART
NPTERM --> e

SKIPNEXT --> BackRefKey

REFBREAK --> CONN
REFBREAK --> CONN'
```

Figure 3.18 Partial lexicon for the parse tree interpreter

The semantic parser works by attempting to match the category stack to the nodes in the tree. The parser maintains a “current reference” string that is updated as nodes in the parse tree are encountered. References are added to a list of complete references when the parser encounters “REFBREAK” or “INTERPOLATE” nodes, or completes a full parse of the tree. Two examples follow that explain this process in detail.

Figure 3.19 shows a parse tree where the original reference is “40 CFR parts 264 and 265”. The semantic interpretation parser transforms this reference into two complete references: 40.cfr.264, and 40.cfr.265. Figure 3.19 is an example of a simple parse tree that can be interpreted. The parser starts by expanding the REF category in its search list to “LEV0 LEV1 LEV2”. It then starts a depth-first parse down the tree, starting at REF. The LEV0’ node matches LEV0, so this category is marked as found. The LEV0 node also matches the LEV0 search category

Next the children of LEV0 are processed from left to right. Looking up INT in the interpreter lexicon (Figure 3.18) shows it is a PTERM, so the current reference string is updated to be “40”. Looking CFR up in the interpreter lexicon shows that it is also a PTERM, so the leaf’s value is appended to the current reference string to form “40.cfr”. Next, LEV1a’ is processed, and a note is made that the incoming current reference string was “40.cfr”. LEV1a’ matches LEV1, so the top LEV0 search category is discarded and the LEV1 category is marked as found. Processing continues down the LEV1a branch of the tree to the LEV1p node. The PART child node is found to be a NPTERM in the lexicon, so the content of the PART leaf node is not appended to the current reference string. INT is found to be a PTERM, so the content of this leaf node is concatenated to the search string. Since CONL2 is also a NPTERM, the algorithm traverses back up to LEV1a’. The next child node to be processed is CONN’, which is found to be a REFBREAK in the lexicon. This means that the current reference is complete, so “40.cfr.264” is added to the list of references and the current reference is reset to “40.cfr”, the value it had when the LEV1a’ parent node was first reached. Processing then continues down from LEV1a’ to the right-most leaf of the tree. At this point the current reference is updated to “40.cfr.265” and a note is made that the entire tree has been traversed, so “40.cfr.265” is added to the list of identified references. Next the

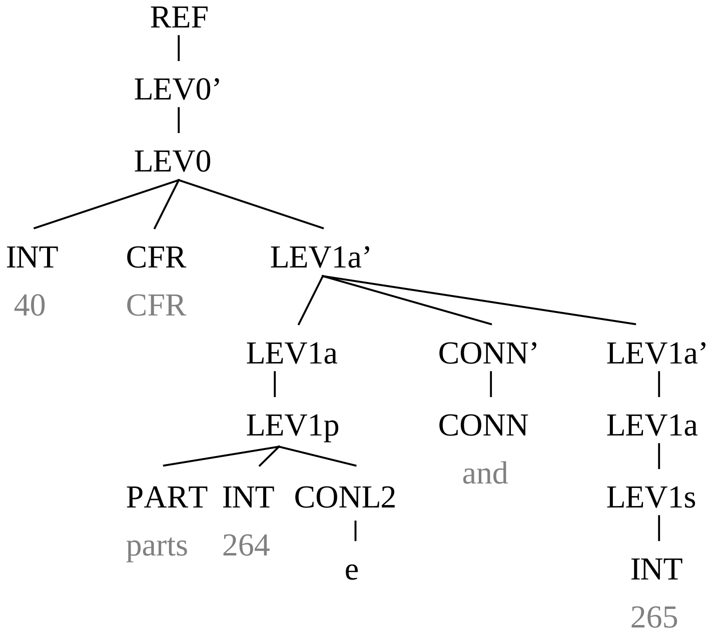


Figure 3.19 Example of a simple parse tree

parser would try the other expansion of REF as "LEV0 LEV3 LEV4 LEV5", but since it would be unable to match LEV3 this attempt would fail. The final list of parsed references thus contains 40.cfr.264 and 40.cfr.265.

The basic approach described above can be extended to handle references where the components of the reference do not appear in order. For example, the parser might encounter the reference "paragraph (d) of section 262.14". A proper ordering of this reference would be "section 262.14, paragraph (d)". To handle these cases, if the top of the category search stack cannot be matched to a node in the tree, the remainder of the parse tree is scanned to see if the missing category appears elsewhere in the tree (a "back-reference"). If the category is found, it is processed and appended to the current reference before the algorithm returns to the original part of the parse tree. If multiple references are found during the back-reference call, the order needs to be reversed to maintain correctness. This allows parsing an interpretation from complex parse trees as shown in Figure 3.20.

The parse tree shown in Figure 3.20 originates from the reference, "Subpart O of part 264 or 265". The semantic interpretation parser transforms the reference into two complete references: 40.cfr.264.O, and 40.cfr.265. In cases of ambiguous meaning, the parser maximizes the scope of ambiguous references. For example, the parse tree in Figure 3.20 could also be interpreted as 40.cfr.264.O and 40.cfr.265.O, but this might be incorrect if 40.cfr.264.O and 40.cfr.265 were actually intended.

Figure 3.20 is an example of a complex parse tree that can be interpreted. A brief explanation of this parse tree follows. In this example, the semantic parser first expands the starting REF category to be "LEV0 LEV1 LEV2". LEV0 matches the ASSUME_LEV0 leaf, and the current reference string is updated to be "40.cfr". Next, the parser encounters LEV2', which does not match LEV0 or LEV1. The parser then searches for a possible "back-reference" (a level of the reference that is out of order,

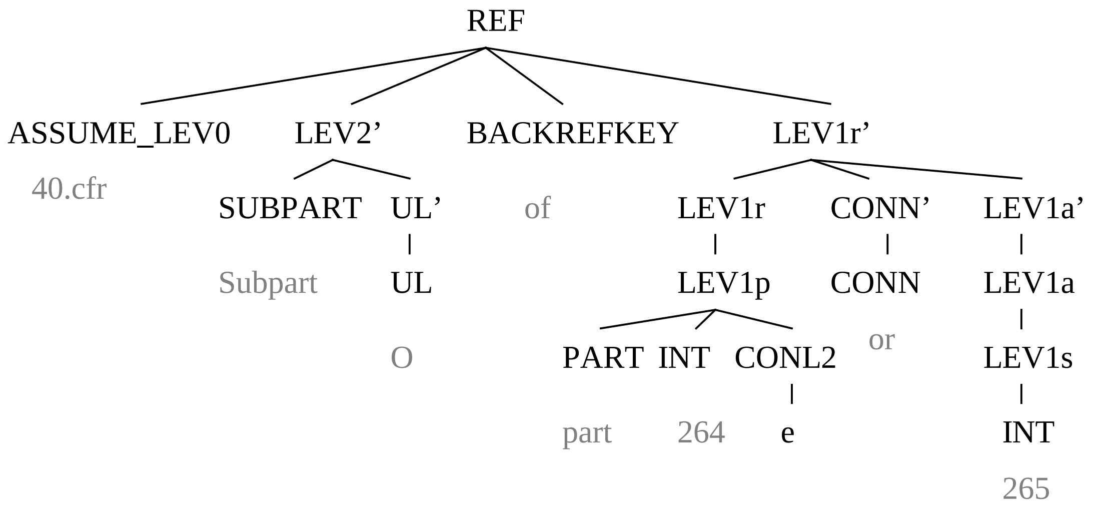


Figure 3.20 Complex parse tree

referring back to a lower level), which it finds as LEV1r'. The parser processes this part of the tree, concatenating the INT under LEV1p to the reference string. It also notes the reference string is complete upon encountering the CONN' (a REFBREAK), so a new reference string is started with "265" and a note is made that the rightmost leaf of the tree has been found. The parser uses back-reference calls to effectively re-order the depth-first parsing process such that reference components are parsed in descending order. Upon returning from the back-reference function call, it is noted that multiple references have been encountered, so a reconciliation procedure is run to swap "40.cfr.264" with "40.cfr.265" in the complete reference list and to set "40.cfr.264" as the current reference string. Now the parser can match the LEV2' category and update the current reference list to be "40.cfr.264.O". Next the parser encounters the BACKREFKEY category, which the lexicon identifies as type SKIPNEXT, so the parser can skip the next child node. Skipping the next child node brings the parser to the end of the tree. Since the parser noted earlier that it had processed the right-most leaf, which indicates a successful semantic parsing attempt, the parser adds the "40.cfr.264.O" to the lists of references found in the parse tree. The subsequent attempt to parse the tree using "LEV0 LEV3 LEV4…" will fail to reach the rightmost leaf, so no more parses will be recorded. Thus, the final reference list is 40.cfr.264.O and 40.cfr.265.

The parsing system developed in this research work, along with the semantic interpreter for the parse trees, should be simple to reconfigure to parse and interpret a variety of different referencing systems or text patterns. Using a grammar and lexicon to specify how to treat categories from a parsed reference provides a great deal of flexibility for the system. New grammar and lexicon files can be introduced to change the system for new types of references. The main limitation of the system is that grammar rules cannot be left-recursive.

3.4.3.2 Statistically-Based Reference Parser

In this research, an n-gram model is employed to make the parsing process more efficient by skimming over text that was not predicted to contain a reference. An n-gram model is a probabilistic model for sets of n sequential words [56]. For example, one might use unigrams, bigrams or trigrams in a model. A unigram is a single word, a bigram is a pair of words, and a trigram is a sequence of three consecutive words. These n-grams can be used to predict where a reference occurs in a regulation by how frequently each n-gram precedes a reference string.

To develop an n-gram model, a regulation corpus of about 650,000 words was assembled. The parser found 8,503 references after training on this corpus. These 8,503 references are preceded by 184 unique unigrams, 1,136 unique bigrams, and 2,276 unique trigrams. For these n-grams to be good predictors of a reference, they should occur frequently enough to be useful predictors, but they should not occur so frequently in the general corpus that their reference prediction value is low.

For the unigrams, it is interesting to note that 18 of the most "certain" predictors are identified as highly "certain" because they are only seen once in the entire corpus. Some other unigrams that one might intuitively expect to be good predictors actually are weak predictors for references. For example, "in" has a 5% prediction value. This is because the 2,626 references that are preceded by "in" are so heavily outweighed by the 49,325 total occurrences of "in" in the corpus. These two factors make the unigram model a weak one, since words with high certainty tend to be those that are rarely seen, and words that preceded many references tend to be common words that also appear often throughout the corpus. One exception to this is the word "under", which precedes 1,135 references and only appears 2,403 times in the corpus (a 47% prediction rate).

The bigram model is a good predictor of references. While over 200 (18%) of the bigrams only occur once in the corpus, the significance of bigrams that precede a reference is *not* diminished by an even larger number of occurrences in the corpus (as is the case for the “in” unigram). For example, “requirements of”, which precedes 1,059 references is seen 1,585 times total in the regulation corpus.

The trigram model helps refine some of the bigram predictors. For example, “described in” with a 61% prediction rate is refined into 35 trigrams with prediction probabilities ranging from 11% to 100%. In general however, the trigram model appears to split things too far, since about 1/3 of the trigrams only appear a single time in the entire corpus.

Before attempting a parse on the input, the three n-gram models are used together by calculating a weighted sum of unigram (U), bigram (B) and trigrams (T) using the following equation: $\lambda_1 U + \lambda_2 B + \lambda_3 T \geq 1$. In this equation, a threshold of 1 is used to determine if the parse should be carried out. By changing the λ weightings, different parts of the text are selected for parse attempts.

While the n-gram model is effective for speeding up parsing, there is a tradeoff between parsing speed and recall. To study this tradeoff, the n-gram model was trained on the 650,000-word corpus and then tested on a 36,600-word corpus. There were 569 references in the test corpus. To experiment with the possible λ parameter values, a brute-force search was done through a range of values (λ_1 = 1-20,000, λ_2 = 1-10,000, λ_3 = 1-640). There were over 10,000 passes through the test file completed during this experiment. The number of reference parse attempts and successful reference parses were recorded. Examples with the lowest number of parse attempts for a given level of recall were selected from the test runs. This process provides an efficient frontier that shows the best efficiency (successful parses / total parse attempts) for a given level of recall. These results are shown in Figure 3.21.

The x-axis in Figure 3.21 shows the level of recall for the pass through the test file. So as to provide an indicator of the extra work by the parser, the y-axis shows the total number of parse attempts divided by the total number of references in the document. As can be seen from Figure 3.21, there is clearly a change in the difficulty of predicting references as the recall level goes above 90%. For recall levels between 0 and 90% the amount of work for increasing the level of recall is relatively low. For recall levels above 90%, however, any additional increase in recall will come at a very significant increase in the number of parse attempts. The usefulness of the bigrams and trigrams is exhausted around 90% recall, most likely due to a sparseness problem in the training data. The only way to increase the number of reference predictions beyond the 90% recall level is to shift the focus to the unigram model – which was noted earlier to have much lower accuracy than the bigram or trigram models. This accounts for a significant increase in the tradeoff between recall and parse attempts.

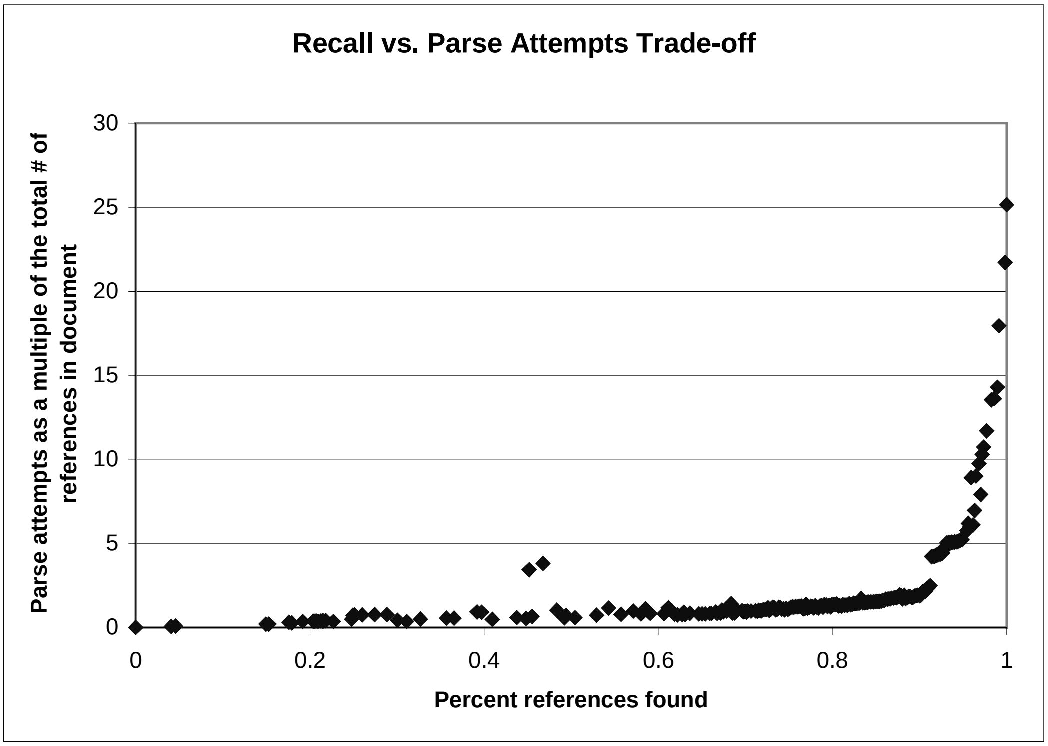


Figure 3.21 Trade-off between recall and required number of parse attempts

It was surprising to see that the prediction system is able to achieve 100% recall in the test file for our experiment, since the test file contains previously unseen data. The 100% recall should not be achievable in general because there may exist a word that precedes a reference that has not been seen before in the training data. The total number of parse attempts to achieve 100% recall on the test file was only 14,310. This compares quite well to the 37,132 parse attempts required to check the document for references without using the n-gram model (i.e., by attempting all possible parses), since it reduces the number of parsing attempts necessary by more than a factor of two.

This reference parsing research addresses three questions for regulation reference extraction. First, it is shown that an effective parser can be built to recognize and transform environmental regulation references into a standard format. Second, it is also shown that an n-gram model can be used to help the parser "skim" through a document quickly without missing many references, and the time/recall tradeoff has been explored as shown in Figure 3.21. Third, it is found, qualitatively at least, that an n-gram reference-prediction model is a useful tool for grammar development when attempting to build a parser for sections of text.

It is possible to further refine and design a probabilistic guided reference parser that efficiently scans a document for references. Since our main interest in this research project is to identify as many references as possible and store them in the document, a fast parsing system is not the objective of our work and has not been pursued further.

3.4.4 Definitions

A third type of metadata added to the 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 regulation framework allows a regulation viewing system to incorporate explicit definitions of terms and acronyms into its user interface.

A regulation generally provides specific definitions for many terms appearing throughout the regulation. By making these definitions explicitly clear, software systems can assist regulation readers in understanding the often definition-intensive documents. An example of how definitions may be used is described in Section 4.4.1. Figure 3.22 illustrates a definition element in XML.

Environmental regulations include many terms and acronyms that are specialized to the entire environmental regulatory domain, and are therefore not defined in each regulation itself. There are many other sources of definitions aside from the regulation itself. For example, a quick search on the Internet for environmental regulation glossaries reveals hundreds of sites. Because there are so many sources for definitions and acronyms, and often so many different definitions and acronym expansions for individual terms, only terms defined in the regulation itself are tagged with XML elements. The addition of definition elements to the XML regulations is currently done manually, though a more sophisticated natural language parser might be effective for automatically inserting definition elements in an XML regulation.

```
<definition>
  <term>
    used oil
  </term>
  <definedAs>
    Used oil means any oil that has been refined from crude oil, or any synthetic oil,
    that has been used and as a result of such use is contaminated by physical or
    chemical impurities.
  </definedAs>
</definition>
```

Figure 3.22 A definition XML element

3.4.5 Legal Interpretations

Some environmental regulation provisions can be ambiguous and hard to interpret. The meaning of the provisions may be slightly different than a straightforward reading of the provisions would imply. This problem is common to many forms of primary legal sources, for which raw documents without appropriate annotation can be misleading [107]. The regulation provision may have acquired some important nuances through the results of court cases or guidance documents issued by a regulatory agency. These important nuances may not be well conveyed in the regulation text, and without some type of annotation to make these finer points clear the regulation may be misleading.

For example, 40 CFR 261.4(b)(1) discusses solid wastes that are not to be considered hazardous wastes under the regulations. Household wastes are among the wastes that the regulation exempts from hazardous waste regulation. This includes household waste that is collected, transported, stored, treated, disposed, recovered, or reused. The regulation states that [29]:

> “A resource recovery facility managing municipal solid waste shall not be deemed to be treating, storing, disposing of, or otherwise managing hazardous wastes for the purposes of regulation under this subtitle, if such facility:
>
> (i) Receives and burns only
>
> (A) Household waste (from single and multiple dwellings, hotels, motels, and other residential sources) and
>
> (B) Solid waste from commercial or industrial sources that does not contain hazardous waste; and
>
> (ii) Such facility does not accept hazardous wastes and the owner or operator of such facility has established contractual requirements or other appropriate

> notification or inspection procedures to assure that hazardous wastes are not received at or burned in such facility.”

A casual reading of this provision might lead one to conclude that a municipal waste incinerator that accepts household waste need not be concerned with hazardous waste regulations. This was how most owners and operators of incinerators that accept household waste treated the provision until 1994. In 1994 the U.S. Supreme Court [28] ruled that such incineration facilities are not subject to RCRA Subtitle C as a hazardous waste treatment, storage or disposal facility. The Court also ruled that although such facilities were exempted through the household waste exemption, the ash produced by the facilities was not exempt. Therefore ash produced by incinerating household waste can be regulated as a hazardous waste if it has hazardous characteristics, and incineration facilities that burn household waste can be considered generators of hazardous waste. Identifying this interpretation of the regulation provision is of great significance for accurately complying with the regulation, and identifying this interpretation might be difficult without an annotation that explains the correct interpretation.

To resolve this problem, it is important to enable the tagging of regulation provisions with legal interpretations written by experts familiar with the regulation. A simple XML element to encapsulate legal interpretations has been added to the XML regulation structure to enable this important annotation capability. The addition of legal interpretation elements to the XML regulations is done manually, since the creation of a legal interpretation requires a legal expert to read and interpret the regulation provision. Figure 3.23 illustrates how the legalInterpretation element may be used to annotate a regulation with important notes about how a regulation should be interpreted.

```
<regElement id="40.cfr.261.4.b.1" name="Solid wastes which are not hazardous
wastes”>
   <regText>
      The following solid wastes are not hazardous wastes:
      (1)  Household waste, including household waste that has been collected,
      transported, stored, treated, disposed, recovered (e.g., refuse-derived fuel) or
      reused…
   </regText>
   <legalInterpretation>
      This provision has been upheld, but narrowed in scope by the U.S. Supreme
      Court.  Household waste is generally not considered a hazardous waste.  The court
      narrowed this provision when it decided ash produced by incinerating household
      waste is regulated as a hazardous waste if it has hazardous characteristics.  Thus, if
      an incineration facility burns household waste, it can be considered a generator of
      hazardous waste.
   </legalInterpretation>
</regElement>
```

Figure 3.23 Illustration of the legalInterpretation element

3.5 Related Research

There has been significant work done in the area of representing legal documents in an XML format. Much of this work has been done in the European research community. Boer et al. [14, 15] proposed a Legal XML standard in 2002, the MetaLex standard[37]. The standard is particularly notable for its design as a language independent legal standard and its aspirations to facilitate the use of XML for more than search and presentation services. It is aimed to standardize legal documents for purposes such as filtering, presentation, document management, knowledge representation, search, code generation, rule generation, classification and verification. The legal applications of the MetaLex standard include legislation, case law, written public decisions, business

[37] The MetaLex Project is located at the web address http://www.metalex.nl.

regulations and contracts. The work described in [14] focuses on Dutch legislation, but the goal is to develop a standard that will cover all legal sources.

Marchetti et al. [57] has worked on developing data standards for the representation of Italian legislation and tools for accessing this legislation. This group of researchers has made strong statements about the positive effects that XML may have on the legislative process: "We dare say that markup languages, and in particular XML, can provide interesting results at both ends of the legislative process: at the drafting stage, enforcing some or all the drafting rules defined for our norms [rules for drafting normative documents]; at the accessibility stage, fostering easy and sophisticated searching and rendering tools for the public at large. Furthermore, XML may constitute a great influence on several other aspects of the legislative process, providing support for the consolidation of laws, rationalising the legislative process, improving the referencing and connections among the norms, etc."

A number of associations have also worked on XML standards for legal documents. LegalXML.org is one group working towards developing XML standards for a number of different legal documents, such as court filings. These XML standards are designed to be specifications to support eContracts, eNotarization, integrated justice, lawful intercept, legislative documents, online dispute resolution, and legal transcripts. In March of 2002, LegalXML.org joined with OASIS, which is a not-for-profit global consortium that is developing a wide variety of SGML/XML standards for e-business[38]. OASIS also manages XML.org[39], a clearinghouse for XML related schemas and other documents. As of this writing, XML.org is planning a new focus on XML developments for E-Government.

The United States federal government has also begun working with XML in recent years. The U.S. Congress began looking into the use of SGML for drafting and exchanging

[38] OASIS is located at the web address http://www.oasis-open.org/.

[39] XML.org is located at the web address http://www.xml.org.

legislative documents in 1997. This investigation of new drafting tools and data standards later shifted to XML, and by December 2000 XML was adopted as the standard for exchanging legislative documents between the House, Senate and other legislative branch agencies. Along with this new data standard, XML tools aiding the drafting of documents have been successful in solving a number of drafting problems that previously plagued document drafters. These include addressing numerous issues related to the formatting and cross-referencing needs of hierarchical legislation. The adoption of the new XML standard has been going on for just over two years now. The Office of Legislative Counsel of the U.S. House of Representatives, which provides drafting assistance to the House of Representatives, commented on these efforts: "The transition to XML for the drafting of legislation has been both challenging and highly rewarding for the House. At this point, the House has been using the XML authoring environment for House-only resolutions since January 2001 and began drafting bills in XML in September 2002. The House plans to draft over 95% of introduced bills in XML by January 2003."[40]

XML continues to be an active research topic, with work continuing in areas ranging from knowledge management [23] to the secure and selective dissemination of documents [13]. All of these areas of work will contribute to the increasing benefits of adopting XML as a primary legal document standard.

The importance of adding reference metadata to XML regulations is clear from the work in recent years focusing on exploiting the power of reference data. For example, work by Brin and Page demonstrated that hyperlinks in HTML documents can be used to build a more effective search engine [17]. This line of research work has culminated in the popular search engine Google[41]. Work has also been done to automatically link together scholarly work on the Internet using references [11]. Citeseer, an autonomous citation indexing system for academic literature, is described by Giles et al. in [36]. The Citeseer

[40] This quotation is from an online paper entitled, "Drafting Legislation Using XML at the U.S. House of Representatives." This document is available at the web address http://xml.house.gov/drafting.htm, and was accessed on May 7th, 2003.

[41] Google is located at the web address http://www.google.com.

system uses references in academic literature to build a network of related documents, which facilitates searching for related articles. It also allows users to view the context in which citations are used, allowing researchers to see what authors say about particular papers. The importance of automated linking of legal citations has been noted by Needle [66]. Needle was focused on the hypertextual linking of documents in the JUSTIS legal databases, and used the commercial system Syntalex towards this end. References within regulations have also been investigated as features for performing similarity analysis by Lau et al. [52]. It was found that using regulation references can facilitate the identification of relevant provisions for regulation readers, and can also reveal hidden similarities between regulation provisions.

The work described in this chapter to extract and transform regulation references from the regulation text is a variant of the general information extraction problem. The information extraction problem is one that has been studied with increasing interest since the advent of the Internet in the late 1990's. Many researchers have developed a variety of techniques to address the information extraction problem, examples include work by Kushmerick et al. [49], Hsu et al. [43], and Muslea et al. [63]. Muslea has provided a survey of much of the current work [62].

3.6 Summary

This chapter has discussed the development of an XML structured regulation and framework. This framework is important because it provides a formal way to structure regulatory information such that it will be more amenable to software processing for regulation assistance services. It also lays the foundation for the regulation assistance system (RAS) discussed in the next chapter. The web-based regulation assistance system makes use of the XML regulations annotated with meta-data to provide compliance assistance services. The development of the RAS system motivates the design of the XML structure and the associated metadata described in this chapter.

This chapter has described how environmental regulations can be developed and stored as structured XML documents. First, various types of documents that can be used to represent regulations were discussed. Second, the XML regulation structure developed in this project was introduced, and parsers for converting PDF and HTML regulations into the XML structure were described. Third, meta-data elements added to the XML regulation, such as concepts, definitions, references and legal interpretations, were described in detail. The next chapter will describe how these structured XML regulations tagged with metadata can be used to provide compliance assistance services.

Chapter 4

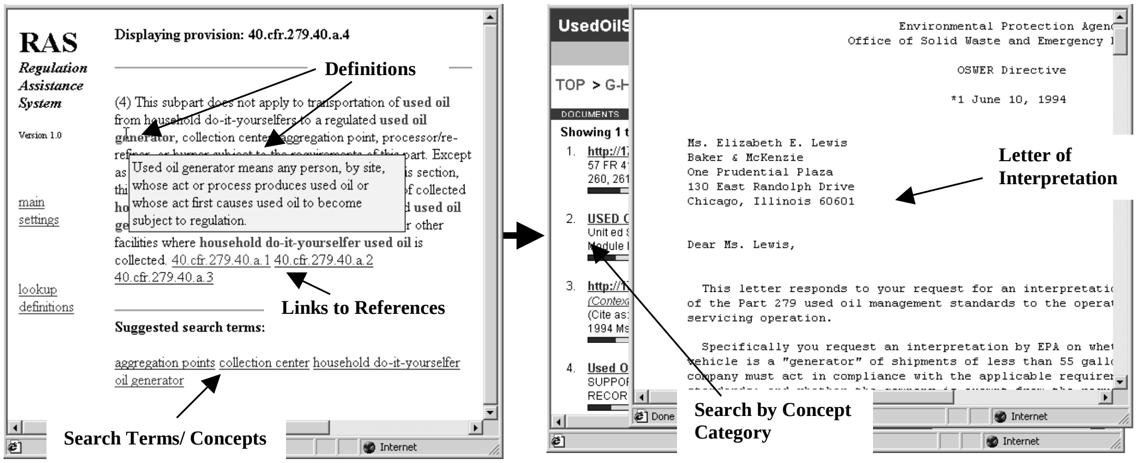
Building A Compliance Assistance System

4.1 Introduction

There has been a push in the United States by the executive office for government agencies to put more emphasis on compliance assistance in lieu of enforcement to encourage companies to comply with regulations [21, 64]. Towards this end, specialized programs using expert system technologies have been built to assist users in understanding regulation requirements for particular circumstances [16]. As noted in Chapter 1, since these systems are not explicitly grounded in the regulations they have a number of important limitations. This chapter describes our research on developing a compliance assistance infrastructure that is capable of assisting users in determining regulation requirements. In addition, the compliance assistance infrastructure builds upon the XML regulation framework and takes advantage of the regulation metadata described in Chapter 3. Besides the concept, reference and definition tags discussed in Chapter 3, we add logic and control processing metadata to the XML regulation framework. This approach allows the construction of a regulation assistance system with clear linkages to the regulation text, thus overcoming many of the limitations of the systems currently in use.

Before we discuss in detail the development of the compliance assistance system, an overview of how the regulation assistance system works, and the motivation for it, is provided as a point of reference. The primary objective of this web-based regulation assistance system is to assist users in determining regulation requirements, and help them determine if they meet those requirements. The system gathers information from a user and informs them if it identifies a conflict between information they have provided and the regulation rules. For example, a user may start by selecting a regulation provision to check against for compliance. The RAS system will then step through the regulation provision by provision. At each step, the system will ask 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. 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, utilizing the metadata with which the regulations are tagged. The system provides viewing and search tools using definitions, key conceptual phrases, and references, all of which will be described in this chapter. Some of the uses for this metadata is shown in Figure 4.1.

A web interface asks users questions based on information in the XML logic metadata. Users may select a response from a menu of possible responses, including an "I don't know" option that forks the compliance-checking process along all possible answers. The ability to allow users to fork the compliance process along all possible paths at any time is useful for exploring different scenarios. When the system completes a check



Displaying provision: 40.cfr.279.40.a.4

(4) This subpart does not apply to transportation of **used oil** from household do-it-yourselfers to a regulated **used oil generator**, collection center, aggregation point, processor/re-refiner, or burner subject to the requirements of this part. Except as

Used oil generator means any person, by site, whose act or process produces used oil or whose act first causes used oil to become subject to regulation.

facilities where **household do-it-yourselfer used oil** is collected. 40.cfr.279.40.a.1 40.cfr.279.40.a.2 40.cfr.279.40.a.3

Suggested search terms:

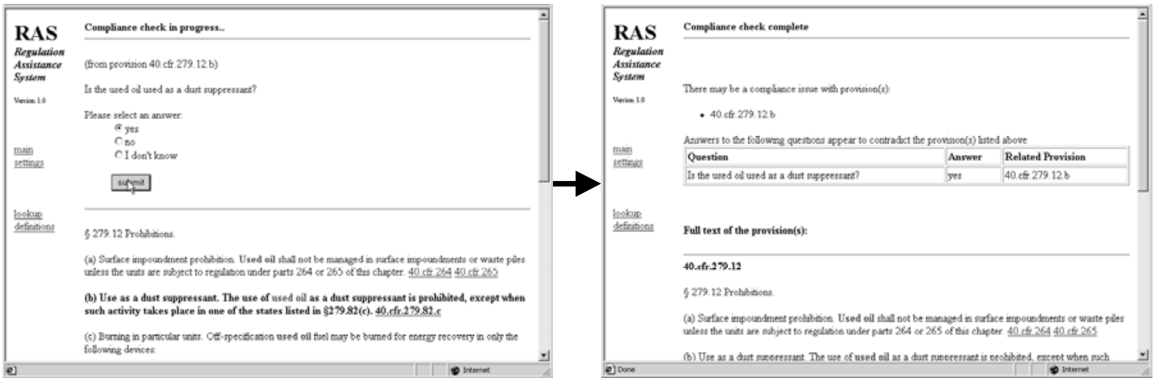
aggregation points collection center household do-it-yourselfer oil generator

```
                              Environmental Protection Agenc
                     Office of Solid Waste and Emergency I

                                       OSWER Directive

                                      *1 June 10, 1994


Ms. Elizabeth E. Lewis
Baker & McKenzie
One Prudential Plaza
130 East Randolph Drive
Chicago, Illinois 60601


Dear Ms. Lewis,


  This letter responds to your request for an interpretatic
of the Part 279 used oil management standards to the operat
servicing operation.

  Specifically you request an interpretation by EPA on whet
vehicle is a "generator" of shipments of less than 55 gallo
company must act in compliance with the applicable requirem
```

Figure 4.1 Definition, reference and concept usage

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. The use of the system and the results produced by it are illustrated in Figure 4.2. Downloadable logs of completed compliance checks allow users to maintain detailed records of their compliance checks. The logs of compliance checks can also be uploaded and edited for future compliance checks against the same or updated regulations.

In addition to guiding users through regulation requirements, the regulation assistance system demonstrates how the regulation meta-data can be used. The RAS functionality is implemented with a web interface that communicates with a compliance checking system. The compliance checking system interacts with a theorem prover component. The compliance checking system controls the process used to check for violations. First, it parses the XML-structured regulation to extract the information necessary to run a compliance check. The XML structure allows the system to properly scope the meta-data and reduce the amount of extraneous data passed to the reasoning system. Only the logic and control processing metadata necessary for the compliance checking are acquired and



Compliance check in progress..

(from provision 40.cfr.279.12.b)

Is the used oil used as a dust suppressant?

Please select an answer:

● yes
○ no
○ I don't know

§ 279.12 Prohibitions.

(a) Surface impoundment prohibition. Used oil shall not be managed in surface impoundments or waste piles unless the units are subject to regulation under parts 264 or 265 of this chapter. 40.cfr.264 40.cfr.265

(b) Use as a dust suppressant. The use of used oil as a dust suppressant is prohibited, except when such activity takes place in one of the states listed in §279.82(c). 40.cfr.279.82.c

(c) Burning in particular units. Off-specification used oil fuel may be burned for energy recovery in only the following devices:

Compliance check complete

There may be a compliance issue with provision(s):

- 40.cfr.279.12.b

Answers to the following questions appear to contradict the provision(s) listed above

Question	Answer	Related Provision
Is the used oil used as a dust suppressant?	yes	40.cfr.279.12.b

Full text of the provision(s):

40.cfr.279.12

§ 279.12 Prohibitions.

(a) Surface impoundment prohibition. Used oil shall not be managed in surface impoundments or waste piles unless the units are subject to regulation under parts 264 or 265 of this chapter. 40.cfr.264 40.cfr.265

(b) Use as a dust suppressant. The use of used oil as a dust suppressant is prohibited, except when such

Figure 4.2 Example compliance-checking session

dynamically loaded into the reasoning system. This is important because the performance of FOPC theorem provers decreases rapidly as the number of logic sentences used for reasoning increases. The system design is such that any FOPC theorem prover can be used to perform the logic checks. Presently, we employ Otter, a publicly available theorem prover developed at the Argonne National Laboratory [60].

This chapter will present logic and control metadata necessary for providing compliance assistance services. Regulation 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 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, and which provisions do not need to be investigated. Each type of logic or control processing metadata can be associated with any regulation provision in the document. In the regulation framework described in this thesis, these 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, a federal used oil regulation

(40 CFR 279) has been manually tagged with regulation logic metadata, with user-interface logic metadata, and with control processing metadata.

The XML structure described in the previous chapter is a critical component for the development of a web-based regulation assistance system. The RAS makes use of the regulations annotated with meta-data to provide compliance assistance services.

This chapter investigates the development of a regulation assistance system. The design and development of the regulation compliance-assistance system is based on first order predicate calculus. The required extensions to the XML regulation standard are also examined in detail in this chapter. First, propositional and predicate logic are briefly introduced and logic sentences are expressed as a form of metadata. Second, additional types of metadata that are added to the XML regulations to enable a logic-based compliance-assistance system are discussed. Third, the algorithms used for compliance checking are examined. Fourth, the use of the RAS system is illustrated. Finally, some of the related research work in this area is reviewed.

4.2 Logic

This section introduces the types of metadata used to annotate XML regulations. Symbolic logic is a representational formalism used to describe concepts, ideas and knowledge. The formal representation of knowledge can be used to reason about the information and to draw new conclusions or look for contradictions. Use of formal symbolic logic can also be used to communicate information between systems [35]. Propositional logic and predicate logic are two symbolic logic languages, each of which will be briefly introduced below. For a more in-depth treatment of this subject please refer to [112].

4.2.1 Propositional Logic

Propositional logic sentences consist of abstract statements and connectives between them. These sentences can take on the truth-values TRUE or FALSE. The statements in a propositional sentence are called propositions, which may be abstract statements or one of the truth symbols *true* or *false*. The propositions that are abstract statements can have either truth-value TRUE or FALSE. The truth symbol *true* always has the truth value TRUE, while the truth symbol *false* always has the truth value FALSE. The connectives between propositions can be "and" ($\wedge$), "or" ($\vee$), "not" ($\neg$), "implies" ($\Rightarrow$), or "equivalent" ($\equiv$).

An example of a propositional logic sentence is the following:

$$\text{OilHasBeenUsed} \wedge \text{OilIsContaminatedFromUse} \Rightarrow \text{OilMeetsUsedOilDefinition}$$

This propositional logic sentence states that if oil has been used, and it has been contaminated by the usage, then it meets the definition of used oil. The propositional sentence above illustrates how complex sentences can be built from basic propositions and logical connectives. Propositional sentences of an arbitrary length can be written by chaining together propositions with the set of connectives.

One can determine the truth-value of a propositional sentence by examining the truth-value of individual propositions and applying evaluation rules associated with the connectives. The rules for evaluating propositional sentences are as follows:

- The sentence *true* has the truth value TRUE.
- The sentence *false* has the truth value FALSE.
- Sentences composed of a single proposition have the truth-value assigned to that proposition.

- The sentence $\neg$A has the truth value TRUE if A is FALSE, and the truth value FALSE if A is TRUE.

- The conjunction of two sentences (A $\wedge$ B) is TRUE if both A and B are TRUE, and FALSE otherwise.

- The disjunction of two sentences (A $\vee$ B) is TRUE if either A or B is TRUE, and FALSE if both A and B are FALSE.

- The implication (A $\Rightarrow$ B) is FALSE if A is TRUE and B is FALSE, and is TRUE otherwise. This is different from a common non-technical understanding of implications. If the antecedent is false, the statement is always true. Therefore a statement such as, “dogs can fly $\Rightarrow$ cats can fly”, is considered TRUE.

- The equivalence (A $\equiv$ B) is TRUE if the truth value of A is the same as the truth-value of B, and is FALSE otherwise.

Inference rules can be used to reason about knowledge formalized in propositional logic. What can be represented in propositional logic is somewhat limited, which limits the usefulness of this system of logic for representing regulation rules. Propositional logic cannot represent the concept of an object or relationships between objects. Without the ability to express objects, it is not possible to express the qualities of an object. Propositional logic also cannot represent the idea of quantification. For example, it is not possible to express in propositional logic the tautology that, “all water is not polluted, or there exists some water that is polluted.”

Since propositional logic is rather limited in its expressive power, the more powerful formalism of predicate logic is used to annotate environmental regulations in the work described in this thesis.

4.2.2 Predicate Logic

Predicate logic is similar to propositional logic, but allows quantification and the usage of objects. For example, Figure 4.3 shows the logic tautology “all water is not polluted, or there exists some water that is polluted”, expressed in predicate logic.

The syntax of predicate logic is similar to that of propositional logic. Predicate logic sentences are composed of connectives, truth symbols (*true* or *false*), constants, variables, predicate symbols and function symbols. Constants and variables denote objects. Predicates define relationships between objects. Functions define functions on the objects. Predicates and functions have defined arities that are the number of arguments associated with their use. These arguments are called terms and are separated by commas. Terms may be constants, variables or function expressions. Similar to propositional logic, the connectives between elements in a predicate logic sentence can be “and” ($\wedge$), “or” ($\vee$), “not” ($\neg$), “implies” ($\Rightarrow$), or “equivalent” ($\equiv$). The rules for truth determination are the same as those listed for propositional calculus.

Quantifiers are used to quantify predicate logic variables as universally or existentially quantified. An example of a universally quantified ($\forall$) variable is in the first half of the disjunction from Figure 4.3, which can be read as, “for all x, if x represents water then x is not polluted”. An example of an existentially quantified ($\exists$) variable is the second half of the disjunction from the water pollution example in Figure 4.3, which can be read as, “there exists an x, where x represents water and x is polluted”. An example of representing knowledge in predicate logic is shown in Figure 4.4, which states that if something is an industrial boiler, utility boiler or oil-fired space heater then it can also be known more generally as simply a boiler. Figure 4.4 also states some information about an object “unit1”, which is described as an IndustrialBoiler. Using these components of predicate logic the logic sentences in Figure 4.3 and Figure 4.4 can be constructed.

$\forall$x.(water(x) $\Rightarrow$ $\neg$polluted(x))
$\vee$
$\exists$x.(water(x) $\wedge$ polluted(x))

Figure 4.3 Example of predicate logic tautology

IndustrialBoiler(unit1)

$\forall$x. ((industrialBoiler(x) $\vee$ utilityBoiler(x) $\vee$ oilFiredSpaceHeater(x)) $\Rightarrow$ boiler(x))

Figure 4.4 Predicate logic examples

There has been a great deal of research on inference procedures for predicate logic. Using these inference procedures it is possible to reason about knowledge represented in predicate logic. For example, using the logic sentence “industrialBoiler(unit1)” and the implication rule above, it is possible to infer, “boiler(unit1)”.

Many inference procedures have been developed that maintain completeness and soundness. Soundness of a proof procedure means that it does not generate incorrect results. Given a set of premises, a sound inference rule generates new sentences that logically follow from the premises. Completeness for a proof procedure means that if there exists a proof for the logic problem posed, then the inference procedure is capable of finding that proof.

There are many proof procedures that maintain both soundness and completeness for proofs by contradiction. When doing a proof by contradiction, the goal of the proof is negated. The theorem prover knows the proof is complete when it derives a contradiction between the starting premises and the negated conclusion. For example, to prove that “boiler1” is in fact a boiler, we would include the two logic sentences from Figure 4.4 and the negated conclusion “$\neg$boiler(unit1)” in an input file for the theorem prover. The theorem prover can complete a proof by contradiction by deriving the sentence “boiler(unit1)” which contradicts the input sentence “¬boiler(unit1)”.

4.2.3 Metadata for Logic and Control Processing

In order to facilitate the development of compliance assistance systems, XML elements are included to provide processing information for systems interpreting the regulation document. We have developed control elements to specify what regulation provisions need to be processed, and logic elements to represent compliance information.

4.2.3.1 Control Processing Elements

The control element is a wrapper element that may contain one or more instructions within it in the form of one or more of three sub-elements: goto, switchTo, or end. These control elements allow regulation designers to specify what regulation provisions may or may not need to be investigated. While not FOPC in nature, control elements provide processing logic and therefore may be used within the logic XML elements that will be discussed in Section 4.2.3.2. These control elements are currently added manually, but it seems possible to take advantage of the automatically-generated reference elements to partially-automate the process of adding control elements.

The goto control element specifies a regulation provision that the system should process next; returning to the current provision once the specified provision has completed its check. This is similar to a standard programming subroutine call. The goto element is useful when it is necessary to check an additional regulation provision without abandoning the current line of processing. For example, frequently a regulation provision will refer readers to another regulation provision that should be read before continuing. The goto element instructs a system to temporarily go to the specified regulation provision, but to return to the current provision eventually.

Similarly, the switchTo element specifies a regulation provision to process next, but processing should not return to the current provision once the specified provision has completed its check. This is useful when a regulation provision specifies some conditions under which a different regulation provision will apply. The switchTo element instructs the system that the check against the current regulation provision is complete, and that processing should continue starting with the regulation provision specified by the switchTo element.

Figure 4.5 illustrates the goto and switchTo elements. This example instructs the system to process section 279.65, and once processing for that section is complete to switch processing to section 279.73. Note that in order to direct processing control to move to a reference within the regulation the control attribute “target” is used. For example, target = “40.cfr.279.65”, refers the compliance processing system to section 279.65 in 40 CFR.

The end element signals that the check of the specified provision is complete. This is useful when the regulation specifies that under certain conditions the check against the current provision need not go any further. Since regulation checks may be done at any level of the regulation document, it is important to specify a target reference for the end element. For example, if during a check of the regulation section 40 CFR 279.12 an end element is encountered that specifies that 40 CFR 279.12(a) is complete, it is important to realize that the check against the higher level section 40 CFR 279.12 is not finished and should continue. On the other hand, if subsection 40 CFR 279.12(a) is being checked and an end element is encountered that specifies 40 CFR 279 is complete, processing of the current provision can stop because the subsection is contained within Part 279. Figure 4.6 illustrates the end element. This example instructs the system that the compliance checking for provision 40 CFR 279.12 is complete.

```
<control>
  <goto target = “40.cfr.279.65” />
  <switchTo target = “40.cfr.279.73” />
</control>
```

Figure 4.5 Illustration of the goto and switchTo elements

```
<control>
  <end target = “40.cfr.279.12” />
</control>
```

Figure 4.6 Illustration of the end element

4.2.3.2 Adding Logic to XML Regulations

Logic can be added to the XML-based regulation document to facilitate manipulation and interpretation of the document. Internal contradictions within the regulation can be checked for, contradictions between regulation documents can be identified, and compliance checking systems can be built to verify that a user is in compliance with the regulation. This section discusses the tagging of environmental regulations with logic, and will be followed by a discussion of the algorithm used by a prototype system for checking compliance with the regulations.

Our goal is to target the narrow area of regulation compliance assistance tools. The broader issue of legal reasoning is beyond the scope of this work. We use FOPC to model regulations in this research work because it offers a flexible, standardized, and computable representation. The choice of FOPC also introduces a great deal of flexibility for the choice of a reasoning system, since there are many reasoners available for working with FOPC. FOPC does not have the expressive power to deal with issues of open texture, deontic modality, or subjunctive conditionals. These are active areas of research [45, 85], and work in these areas could be incorporated into future work extending the research work presented in this thesis.

The current system, using FOPC, cannot precisely model the regulations. FOPC does, however, allow us to model the regulation rules in a simplified form that is sufficient for constructing a system to guide users through regulations and identify potential conflicts with the regulation rules. Even though we use a simplified representation of the regulation rules, the logic metadata and the other metadata may be useful for a variety of systems [53]. While the exceptions in the regulation rules can introduce an element of non-monotonicity, the closed domain of the regulation scope may make this a tractable problem in FOPC. Open texture issues, or managing ambiguous concepts, are treated by leaving resolution of the ambiguity up to the user. The question and answer system that will be described next asks the user to resolve issues that are ambiguous.

The approach of tagging XML structured regulations with FOPC introduces an open platform consisting of structured text and embedded logic that is an improvement over unstructured text variants. Logic elements can be added to the XML structure within the regElement XML elements. The logic elements are denoted by “logic” tags, and may contain either logicSentence or logicOption elements. These elements are described in detail below. All of the logic added to regulations in the current implementation is done manually.

Logic sentences representing the ideas laid out by a regulation are added to an XML regulation document in logicSentence elements that may be placed within “logic” elements. These logicSentence elements may be used to tag regulation provisions throughout the document with their logical meaning. The flexible placement of the logicSentence element enables the tagging of any provision within the document with a meaning. For example, tagging the root regulation element with a logicSentence element specifies that the logic sentence should be applied to the entire document. The logicSentence elements are generally used to define the rules and concepts expressed in a regulation.

Figure 4.7 illustrates a logicSentence element. The logicSentence element describes a rule that used oil may not be used as a dust suppressant. The rule states that for all objects “o”, if “o” is used oil then “o” cannot be a dust suppressant. The use of “ForwardImplies” instead of the more common logic syntax “->” is necessitated by the XML standard, and is described in greater detail in Section 4.2.3.3.

Another logic element was introduced to handle the issue of user input. The logicOption element can be used to build a structured question and answer system that constructs logic sentences based on the user’s input. Without the “logicOption” elements, interacting with the system would require the user to work with FOPC sentences.

The logicOption element contains one question element and one or more answer elements. The question specifies the text that can be used to prompt a user for input. The answer element contains a possible answer to the question and the logic that should be associated with that answer. Since answers are tied to logic statements, the user can interact with the system in plain English, but the answers are mapped to logic statements so that they can be used for compliance checking. The logicOption element allows logic statements to be specified for compliance checking without requiring the user to construct FOPC sentences on their own.

Figure 4.8 illustrates the usage of a logicOption element that assists with gathering user input. This particular element maps the user’s response to a question about the use of used oil to logic statements that reflect the user’s answer. For example, a compliance assistance system might ask the question “is the used oil used as a dust suppressant?” and provide the option of answering “yes" or “no”. If a user selects the “yes” answer, the system would know to match the response to the logic sentence “(usedOil(oil1) AND dustSuppressant(oil1)).”.

```
<logicSentence>
  all _o  (usedOil(_o) ForwardImplies -(dustSuppressant(_o))).
</logicSentence>
```

Figure 4.7 Illustration of the logicSentence element

```
<logicOption>
  <question>
    Is the used oil used as a dust suppressant?
  </question>
  <logicOpt answer = "yes">
    <logicAns>
      (usedOil(oil1) AND dustSuppressant(oil1)).
    </logicAns>
  </logicOpt>
  <logicOpt answer = "no">
    <logicAns>
      (usedOil(oil1) AND (-(dustSuppressant(oil1)))).
    </logicAns>
  </logicOpt>
</logicOption>
```

Figure 4.8 Illustration of a logicOption element

4.2.3.3 Standard Logic Syntax and XML Standards

One drawback to the use of XML for storing logic representations of regulations is that there are syntactic limitations that must be met to comply with the XML standard. For example, XML elements are defined by the XML standard to start with "<", as in "<regText>". This conflicts with the standard logic syntax used for reverse implications, "<-", and equivalences, "<->". A simple substitution of text provides the solution for this problem, where the illegal XML character sequences are replaced with legal ones.

The substitutions currently being used to represent FOPC in an XML compliant syntax are shown in Table 4.1. Note that the substitutions for "->" and "|" are not necessitated by XML standards, but are done so that the XML logic uses a consistent representation formalism. The substitutions for "<-", "<->", and "&" are required by XML standards. The substitutions must be reversed by the logic processing systems that read the XML regulation so that the standard syntax is used when providing the data to a logic reasoner. The XML compliant substitutions also become reserved words in the logic representation language. Since the words in the right column of Table 4.1 will be substituted with the logic symbols in the left column, words in the right-hand side of the table are reserved words that cannot be used for logic predicates or function names.

4.2.4 Nested logicOption Elements

To increase the flexibility in specifying the control flow for compliance checking, the logicOption elements are allowed to be nested within the logicOpt portion of other logicOption elements. Since logicOption elements are used to build the user-interaction portion for compliance checking systems, nesting these elements allows more finely tuned questions to be asked about a provision. The nested elements also provide a clean way to control the compliance checking processing based on user responses. Only the logicOption elements may be nested. Nesting of logicSentence and control elements is not allowed. It is legal to place control elements within the logicOpt elements, but logicOption elements cannot be placed within control elements. These restrictions maintain the integrity of the logic and control elements, while allowing significant flexibility within the logicOption elements for more accurate control of the processing system. Figure 4.9 shows nested logicOption elements. The nested option_logic elements in Figure 4.9 demonstrates how the elements can be nested to improve the questions a user interacts with. Improved questions can help model the user's situation more precisely in FOPC.

Table 4.1 Substitutions for XML compliant logic sentences

Standard logic syntax	**XML compliant substitution**
->	ForwardImplies
<-	ReverseImplies
<->	EquivalentTo
&	AND
|	OR

```
<logicOption>
  <question>
    Is the offspecification used oil burned in an industrial boiler?
  </question>
  <logicOpt answer = "yes">
    <logicOption>
      <question>
        Is the industrial boiler located on the site of a facility engaged in a manufacturing
        process where substances are transformed into new products?
      </question>
      <logicOpt answer = "yes">
        <logicAns>
        (usedOil(oil1) & burnedIn(oil1, method1) & industrialBoiler(method1) &
        burnedAt(oil1, facility1) & engagedIn(facility1, process1) & manufacturingProcess(process1)).
        </logicAns>
      </logicOpt>
      <logicOpt answer = "no">
        <logicAns>
        (usedOil(oil1) & burnedIn(oil1, method1) & industrialBoiler(method1) &
        burnedAt(oil1, facility1) & engagedIn(facility1, process1) & -manufacturingProcess(process1)).
        </logicAns>
      </logicOpt>
    </logicOption>
  </logicOpt>
  <logicOpt answer = "no">
    <logicAns>
    (usedOil(oil1) & burnedIn(oil1, method1) & -industrialBoiler(method1)).
    </logicAns>
  </logicOpt>
</logicOption>
```

Figure 4.9 Nested logicOption elements

4.3 Logic-Based Compliance System

A major research focus for this project has been the development of a web-based Regulation Assistance System (RAS) tool that is built upon the XML regulation framework. This section explains in detail how the XML regulation framework can be used to support compliance assistance services. First the structure of the prototype regulation assistance system developed in this research work is explained. Second, the process used to check for compliance is discussed.

4.3.1 System Structure

The compliance-checking functionality of the RAS system is implemented through a web interface built on top of a compliance-checking component (RCCsession), shown in Figure 4.10. 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, both of which have been discussed earlier in Section 4.2.3. 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 are passed back to the web interface. A user response is mapped to corresponding FOPC user-interface logic sentence for that response according to the associated logicOption element's contents. 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 developed at Argonne National Laboratory [110]. Figure 4.10 shows the organization of the compliance assistance system, which includes a web interface (RASweb), the RCCsession component, and the Otter automated-deduction program.

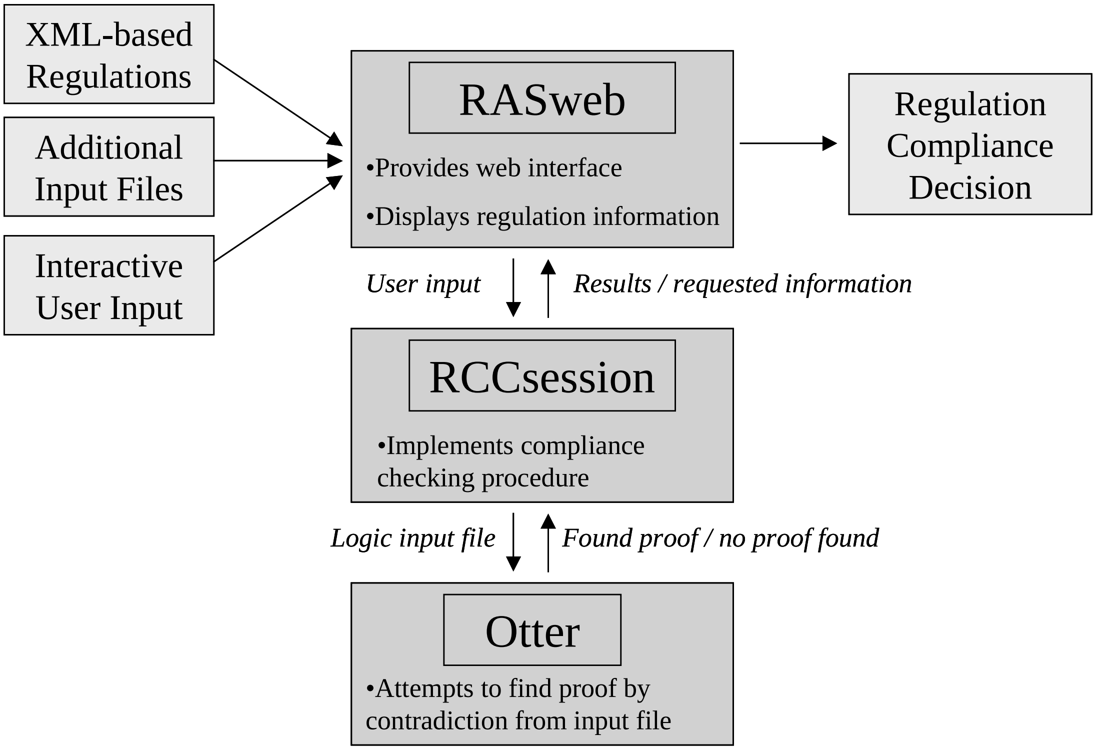

XML-based Regulations
Additional Input Files
Interactive User Input
RASweb
•Provides web interface
•Displays regulation information
Regulation Compliance Decision
User input
Results / requested information
RCCsession
•Implements compliance checking procedure
Logic input file
Found proof / no proof found
Otter
•Attempts to find proof by contradiction from input file

Figure 4.10 Diagram of the Regulation Assistance System's structure

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.

This section has provided a brief overview of the compliance-checking system’s structure. The next section describes the compliance-checking procedure in detail.

4.3.2 Compliance-Checking Process

The RAS compliance-assistance processing algorithm proceeds in three stages. First, the XML regulation is verified. Second, interactive processing is done as the RAS system moves dynamically through the regulation. Third, results of the analysis are compiled and presented to the user. Each of the stages is discussed in detail below. These three stages are illustrated in Figure 4.11, Figure 4.12, and Figure 4.17, respectively.

4.3.2.1 XML Regulation Verification

The RAS system performs two layers of verification checks on an XML regulation before it is used to assist with compliance checking procedures. The first step in verifying an XML regulation is to perform structural verification of the XML. This is done by verifying the regulation against the regulation DTD, which is shown in Appendix A, that defines the valid structure for the XML regulations.

The second step in verifying the XML regulations is to verify that all the logic rules contained in logicSentence elements are consistent. The system extracts all the logicSentence elements from the target regulation and builds an appropriate input file for the theorem prover, Otter. If the theorem prover does not find a contradiction in the logic sentences within a given time period, the logic sentences defining the regulation rules are assumed to be consistent.[42] This check attempts to ensure that the set of logic rules embedded in the XML regulation do not contain a contradiction. The initial check for contradictions in regulation rules does not guarantee that there are no contradictions in the rules, since Otter is not guaranteed to find a contradiction if one exists. In practice, however, this initial logic check has been fairly robust.

[42] This time limit has been set to 30 seconds after a series of experiments has shown that most proof attempts can be completed in only a few seconds by either finding a proof or exhausting reasoning possibilities without finding a proof.

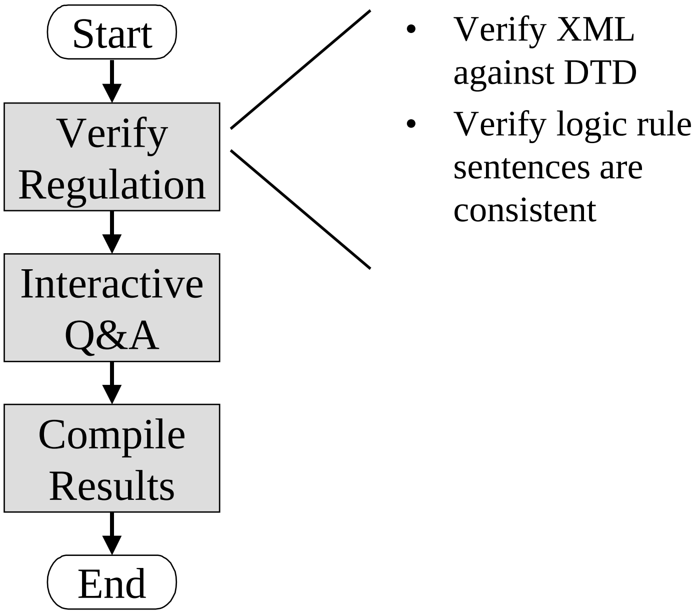


Figure 4.11 Overview of verifying the XML regulation

The initial check for problems with the logic rules is important, since if the rules contain a contradiction the logic system will not be able to assist with compliance checking against the regulation. The RAS system identifies potential conflicts with regulation rules by identifying logical contradictions between user input and regulation rules. Therefore, if the regulation rules themselves contain a contradiction the algorithm used by the RAS system will not work.

4.3.2.2 Gather and Process Logic Sentences

Processing of the regulation document is done by a depth-first tree traversal of the XML structure starting at a provision selected for compliance checking. This initially selected provision may be anywhere in the XML regulation tree. A provisions-to-process (PTP) stack maintains a list of regulation provisions that need to be investigated, and an already-processed-provisions (APP) list keeps track of provisions for which processing is

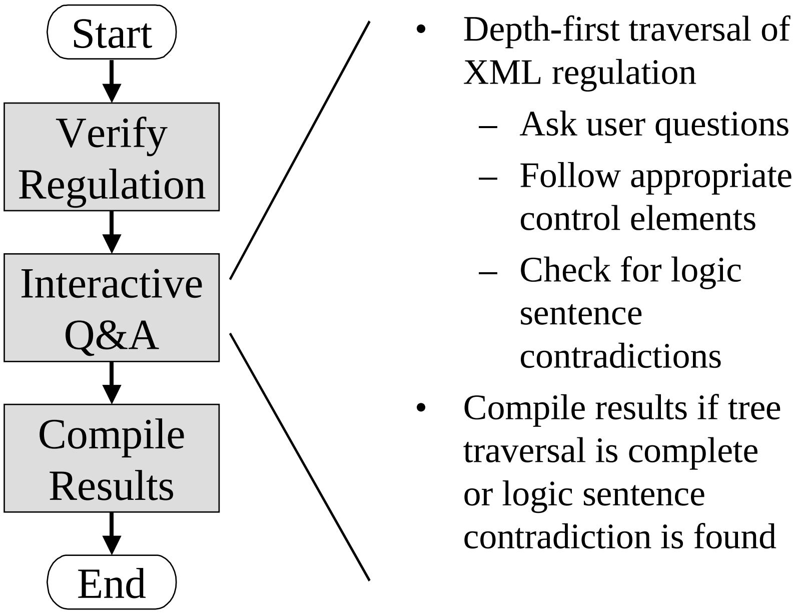

Start
Verify Regulation
Interactive Q&A
Compile Results
End
• Depth-first traversal of XML regulation
– Ask user questions
– Follow appropriate control elements
– Check for logic sentence contradictions
• Compile results if tree traversal is complete or logic sentence contradiction is found

Figure 4.12 Overview of the interactive question and answer compliance processing

complete. Any XML "control" elements encountered while traversing the regulation tree-structure redirect the flow of processing. The effect of the three control elements on the PTP stack and APP list will be discussed next.

The effect of the goto control element on the PTP stack and the APP list is shown in Figure 4.13. The initial PTP stack is shown on the left and the resulting stack after taking the control elements into account is shown on the right side of the figure. As the first example shows, in the simplest case the goto element adds the provision specified to the PTP stack. As the second example shows, only a single call to a particular regulation provision may be in the PTP stack at a time. Additional attempts to add the same provision to the stack are ignored so as to prevent infinite loops. The third example illustrates the idea that provisions in the APP list cannot be added to the PTP stack, since they have already been processed. The fourth example demonstrates that even if the system is processing a sub-provision of the top PTP provision, the goto element operates as would be expected in that the specified provision is added to the PTP stack.

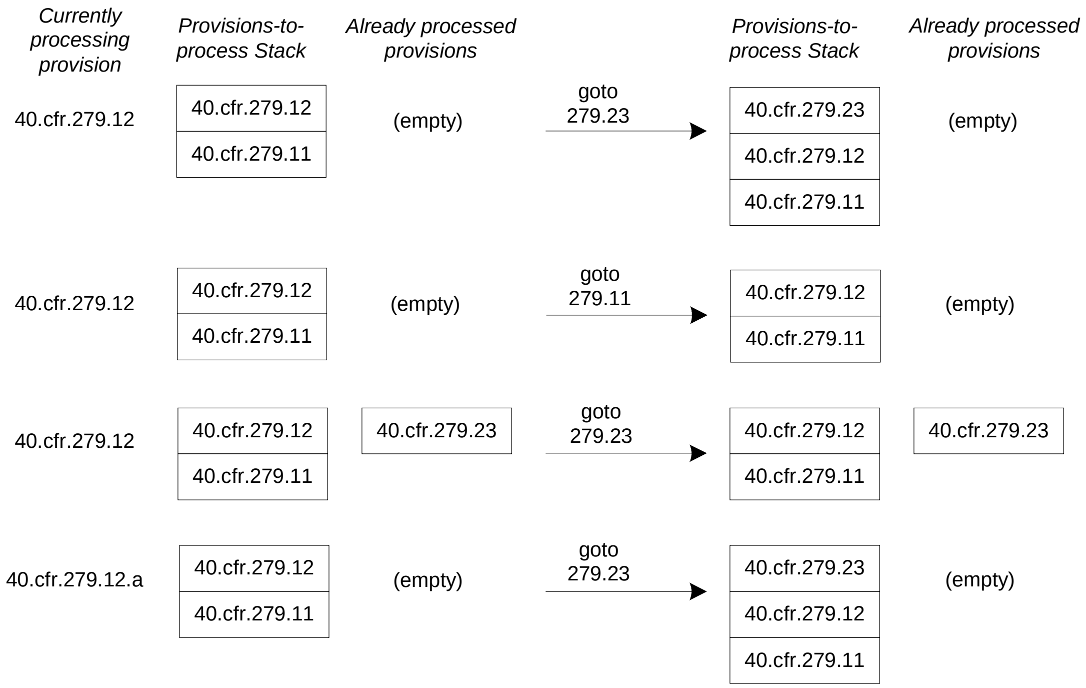


Figure 4.13 The goto element

The effect of the end control element on the PTP stack and the APP list is shown in Figure 4.14. Control elements of end type result in the targeted provision being removed from the PTP stack if the provision exists in the stack, and being added to the APP list. The first example in Figure 4.14 shows the element's basic effect. The end element has no effect if the specified provision is already in the APP list, as shown in the second example. The third example illustrates how sub-provisions of provisions in the PTP stack can be added to the APP list. If provisions in the PTP stack are sub-provisions of a provision targeted by an end element, the sub-provisions will be removed from the PTP stack.

The effect of the switchTo control element on the PTP stack and the APP list is shown in Figure 4.15. The first example in Figure 4.15 shows how in the basic case the top

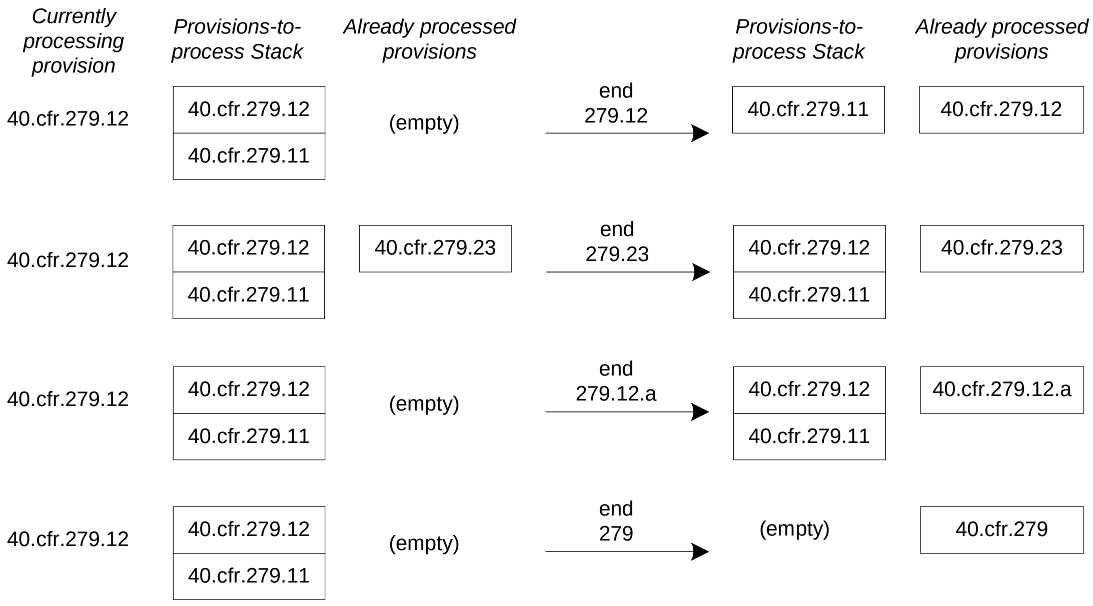


Figure 4.14 The end element

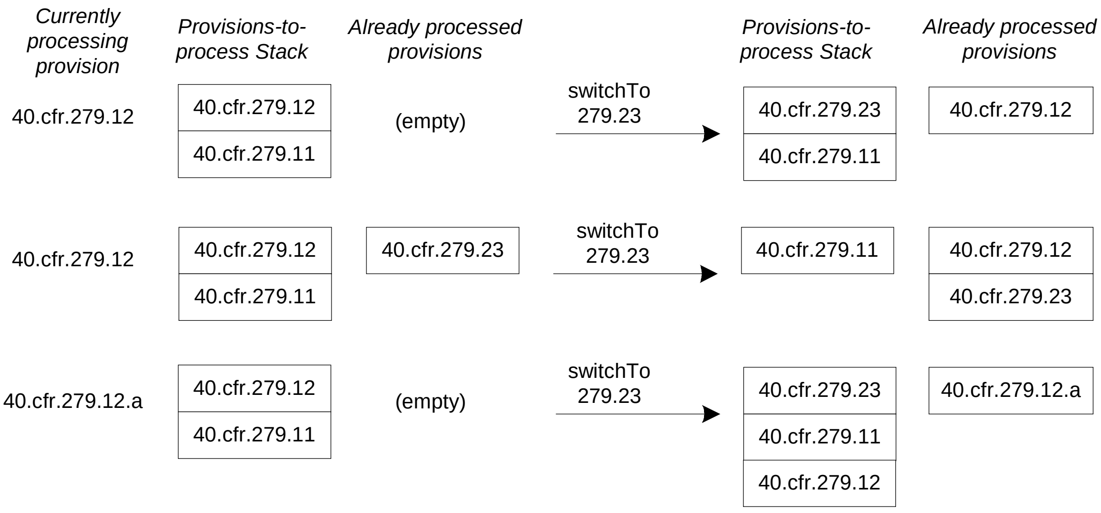


Figure 4.15 The switchTo element

provision is removed from the PTP stack and added to the APP list, and the provision specified by the switchTo element is added to the PTP stack. The second example demonstrates how in cases where the switchTo element refers to previously processed provisions the referenced provision is not added to the PTP stack. The third example shows that if the system is processing a sub-provision of the top PTP provision, the switchTo element adds the provision currently being processed to the APP list and the provision specified by the switchTo element to the PTP stack. The switchTo element is provided for convenience, since it has the same effect as a goto element combined with an implied end element for the current provision.

As logicOption XML elements are encountered during processing, these elements are used to prompt the user for answers to the questions they contain. The logicOption elements provide a mapping from user responses to logic sentences, which can then be verified against logic rules in the regulation. After each question is answered, the logic associated with the selected answer is recorded and any control elements associated with the answer are noted. Otter is then used to check for a contradiction between the logic associated with the user's answers, called response logic, and the rules specified by the regulation.

Figure 4.16 shows a flowchart diagramming the procedure for identifying contradictions between user response logic sentences and regulation rule logic sentences. First, an input file is prepared for the theorem prover, Otter, with all the regulation rules encountered during processing, all the logic sentences selected by users in response to questions, and the logic sentences stating that the provisions to check for compliance with are satisfied. Second, Otter is run using the input file. Third, RAS reads the output file from Otter to see if the theorem prover was able to find a contradiction in the input logic sentences. If no proof was found, the logic sentences are assumed to be consistent. If a proof was found, the proof steps are read to find the input logic sentences that contributed to the

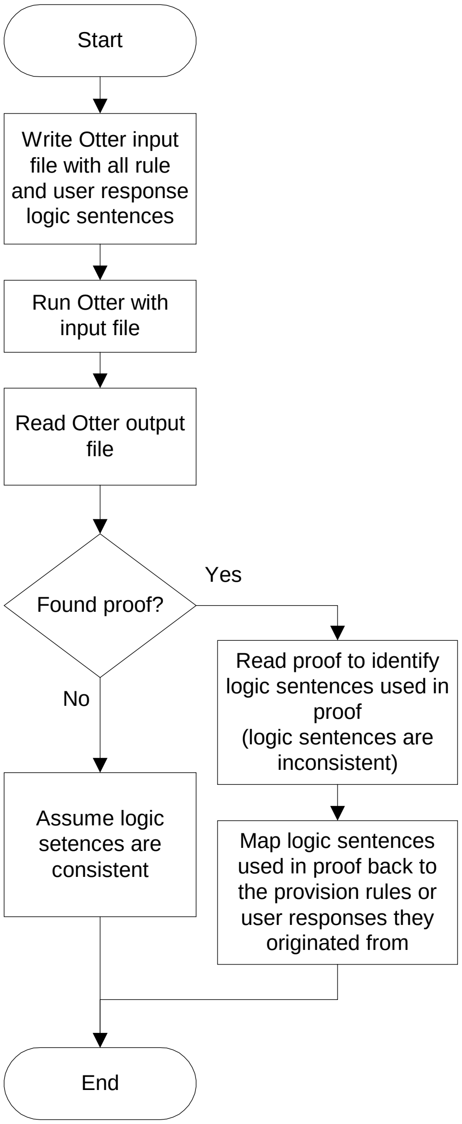

Start
Write Otter input file with all rule and user response logic sentences
Run Otter with input file
Read Otter output file
Found proof?
Yes
No
Read proof to identify logic sentences used in proof (logic sentences are inconsistent)
Map logic sentences used in proof back to the provision rules or user responses they originated from
Assume logic setences are consistent
End

Figure 4.16 Processing FOPC with Otter

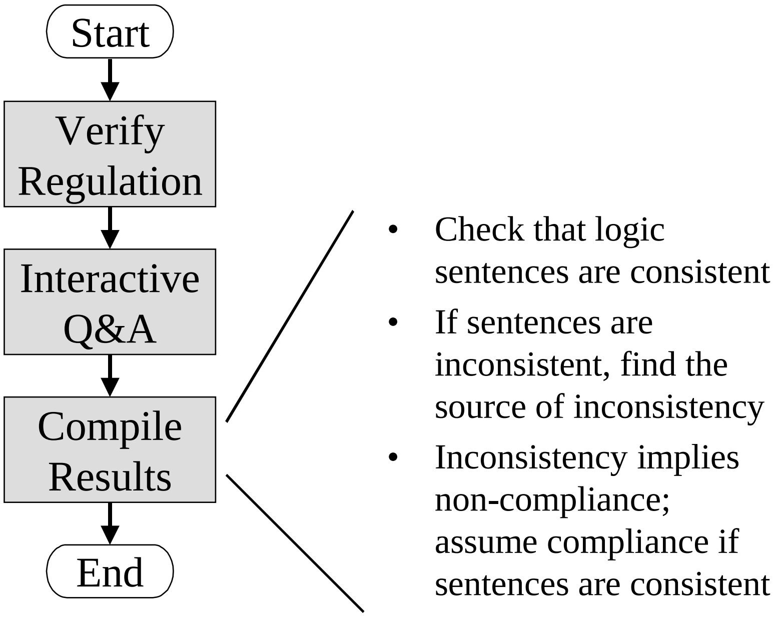
contradiction. These input logic sentences are then mapped back to the provision rules or the user responses from which they originated. This allows the system to identify what is contributing to the logical contradiction, i.e., non-compliance with the regulation.

It is important to provide the extra logic sentences stating that the provisions selected to check for compliance are satisfied. This is because there are two ways in which the procedure may find a compliance problem. First, there may be a contradiction between a user's response to questions and regulation rules that must be satisfied in order to be in compliance with the regulation. Second, a provision that is used to initiate the compliance check may not be satisfied. This provision may not be a regulation rule that always needs to be satisfied, so it may not by itself trigger a contradiction with the regulation rules. Initiating a compliance checking session with a particular regulation provision is asking whether one is in compliance with that particular provision, not necessarily the entire regulation. Not all provisions in a regulation must be satisfied to be in compliance with the regulation. Oftentimes only a subset of a complete regulation needs to be investigated, and only a subset of this subset needs to be satisfied to be in compliance with the regulation. For example, a regulation provision X may state that if section Y is not satisfied, then section Z must be satisfied. Therefore, failure to satisfy section Y does not necessarily imply that one is not in compliance with the regulation because section Z may be satisfied. If section Y were used to initiate a compliance checking session however, then compliance would require that section Y be satisfied.

Answers to the questions contained in the logicOption elements are recorded in a file. This file enables the system to automatically process questions that have been answered in the past. This log file of answers also forms a detailed audit trail that can be provided to the user.

4.3.2.3 Compilation of Results

The questioning procedure terminates when either a logical contradiction is found or the PTP stack is empty. When the questioning procedure ends due to an empty PTP stack, the system returns a result stating that it appears the user is in compliance with the regulation. Note that this is different from proving that the user is in compliance with the regulation. The procedure attempts to show noncompliance with the regulation. Failure to show noncompliance is assumed to mean that the user is in compliance with the regulation.



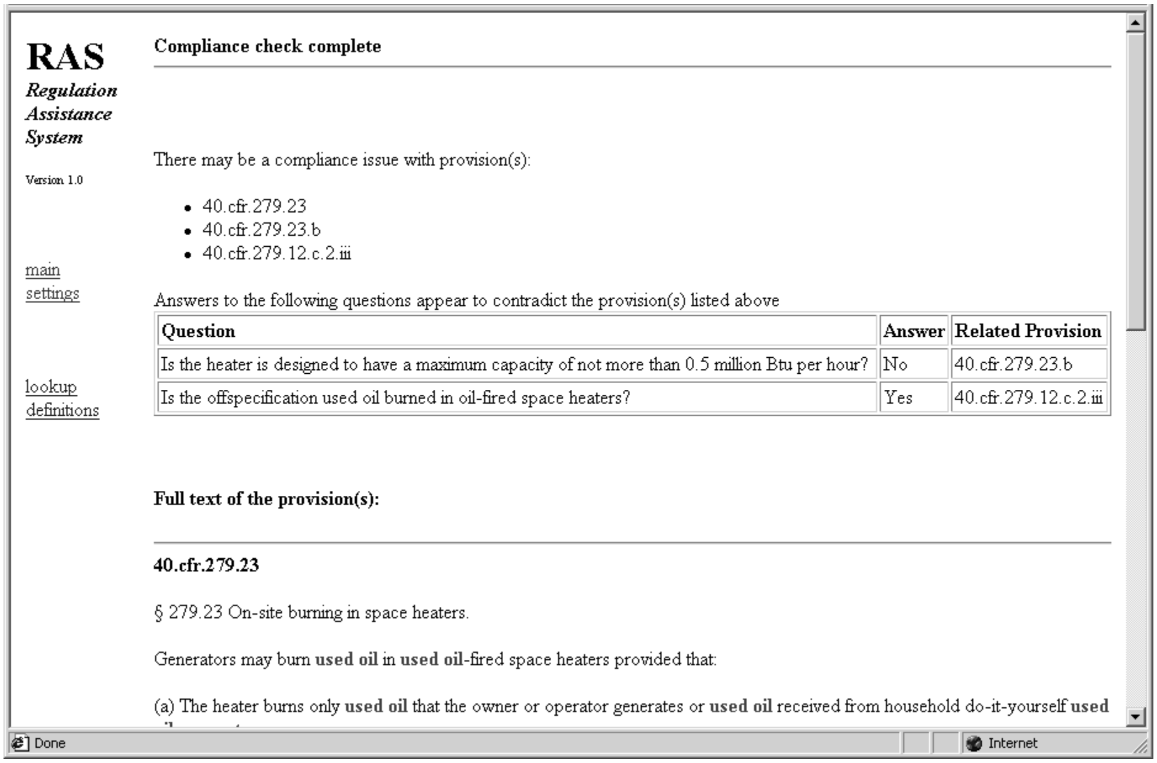
- Check that logic sentences are consistent
- If sentences are inconsistent, find the source of inconsistency
- Inconsistency implies non-compliance; assume compliance if sentences are consistent

Figure 4.17 Overview of compiling results of a compliance check

When the questioning procedure ends due to a logical contradiction being found, the system returns a result stating that there is a compliance problem and a detailed report is returned to the user to help identify the problem. All the questions, answers and relevant provisions that contributed to the logical contradiction are then displayed for the user. An example screen shot of this process is shown in Figure 4.18.

Annotating XML regulations with logic elements and processing them in the manner described above has a performance advantage over simply building a large Knowledge Base (KB) of logic sentences. The primary advantage of the approach described is that the number of logic sentences that need to be handled by the reasoning subsystem (i.e., Otter) is reduced. Doing logic proofs is computationally intense, and significantly reducing the number of extraneous logic sentences greatly reduces the processing time for proofs and increases the complexity of problems that can be handled. If a RAS compliance-checking session were to trace over and dynamically load the logic from every provision in a particular regulation, the performance of the RAS system would not be better than that of a system that simply used a complete KB of all the logic sentences in a regulation for reasoning. The expectation of the RAS system is that in general not all provisions in a regulation need to be processed, so the corresponding number of logic sentences needed for the compliance-checking logic proofs will be smaller than in the case of a system that simply uses a complete KB for all logic checks.

This section has described the logic and control elements added to the XML regulations, along with the algorithm that uses these elements for compliance-assistance purposes. A flowchart of the overall processing algorithm is shown in Figure 4.19.



Compliance check complete

There may be a compliance issue with provision(s):

- 40.cfr.279.23
- 40.cfr.279.23.b
- 40.cfr.279.12.c.2.iii

Answers to the following questions appear to contradict the provision(s) listed above

Question	Answer	Related Provision
Is the heater is designed to have a maximum capacity of not more than 0.5 million Btu per hour?	No	40.cfr.279.23.b
Is the offspecification used oil burned in oil-fired space heaters?	Yes	40.cfr.279.12.c.2.iii

Full text of the provision(s):

40.cfr.279.23

§ 279.23 On-site burning in space heaters.

Generators may burn **used oil** in **used oil**-fired space heaters provided that:

(a) The heater burns only **used oil** that the owner or operator generates or **used oil** received from household do-it-yourself **used**

Figure 4.18 Compliance summary with questions contributing to non-compliance shown

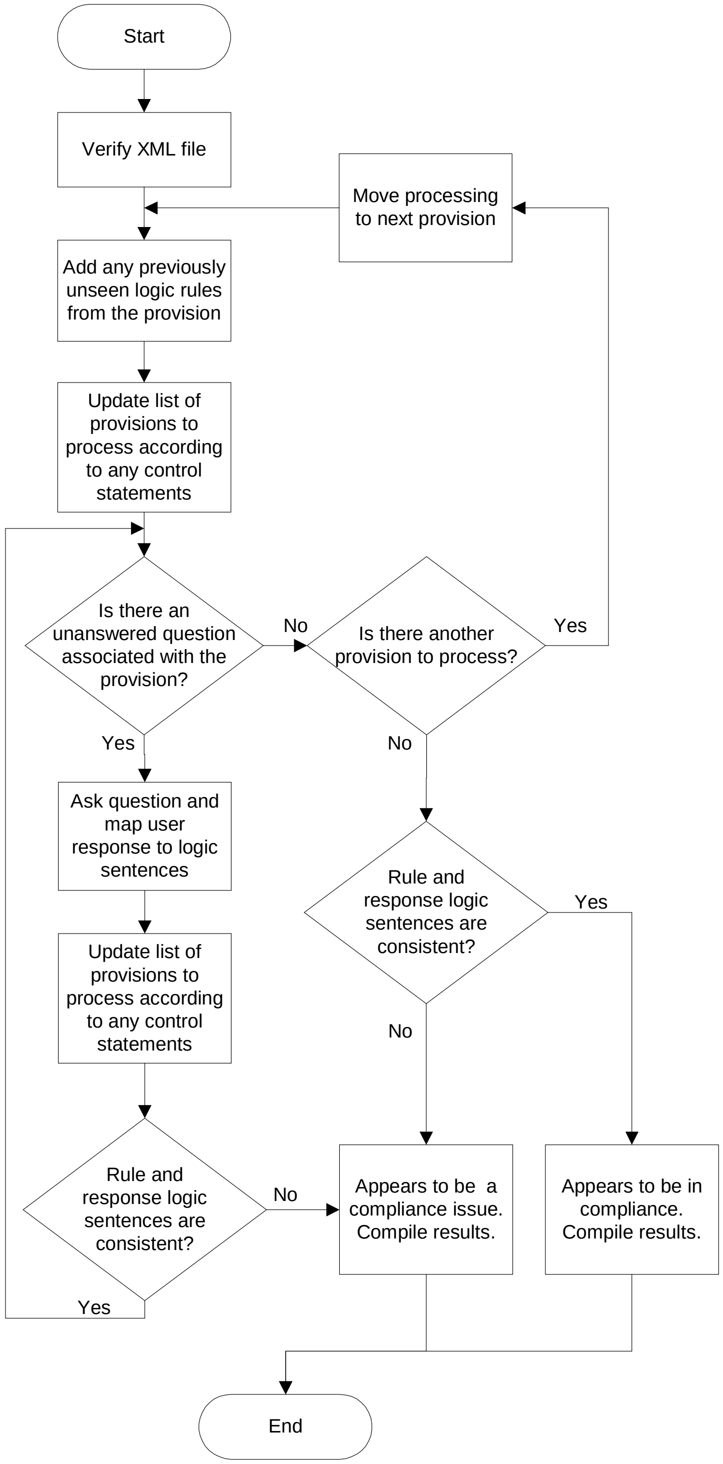

Start
Verify XML file
Move processing to next provision
Add any previously unseen logic rules from the provision
Update list of provisions to process according to any control statements
Is there an unanswered question associated with the provision?
No
Is there another provision to process?
Yes
Yes
No
Ask question and map user response to logic sentences
Rule and response logic sentences are consistent?
Yes
Update list of provisions to process according to any control statements
No
No
Rule and response logic sentences are consistent?
No
Appears to be a compliance issue. Compile results.
Appears to be in compliance. Compile results.
Yes
End

Figure 4.19 Determining compliance with a regulation

4.3.2.4 Logic-Based Control Statements

One weakness of the control elements introduced in Section 4.2.3 is that they only specify immediate and unconditional changes in the processing control. The goto, switchTo and end elements all define actions that should be executed immediately, without regard for any information other than what provisions are in the PTP stack and APP list. There are logical constructions in regulations that require greater flexibility than the control elements alone can provide. For example, Figure 4.20 shows a regulation provision where the applicability of a regulation subpart depends partly on information that may not be currently available. In this example the section 40 CFR 279.23 may not have been encountered yet, so it may not be possible to determine if subpart G should be added to the PTP stack. The control elements introduced in Section 4.2.3 cannot represent this situation because there is a gap between the FOPC logic used to check for compliance issues and the control processing elements that specify what provisions to examine.

We introduce a logic sentence construction to bridge this gap and enable conditional control processing statements. This logic sentence construction allows one to specify that under certain logical conditions a regulation provision should be checked or should be avoided. The logic construction used is of the form “X implies provision Y applies” and “X implies provision Y does not apply”. Figure 4.21 illustrates how the provision in Figure 4.20 can be represented in FOPC. Using this representation, if a company is a used oil generator that also burns the used oil it generates, subpart G will apply if 40 CFR 279.23 is not satisfied. The complete logic representation for 40 CFR 279.20(b)(3) should include an element that directs the system to check for compliance with 40 CFR 279.23 if someone is a used oil generator that also burns used oil.

40.cfr.279.20.b.3 states:
Generators who burn off-specification used oil for energy recovery, except under the on-site space heater provisions of §279.23, must also comply with subpart G of this part.

Figure 4.20 A provision from 40 CFR 279

```
<logicSentence>
   all _client _oil ((generator(_client) & usedOil(_oil) &
   burnsForEnergy(_client, _oil) & -satisfied(40_cfr_279_23)) ->
   provApplies(40_cfr_279_G)).
</logicSentence>
```

Figure 4.21 Logic representation for conditional control statement

A process similar to that used to determine compliance using Otter is used to process these logic-based control statements. For each logic-based control statement, a logic input file is created with all the relevant logic rule and answer sentences, along with a logic sentence negating the target of the logic-based control statement. If Otter finds a contradiction, the logic-based control statement is executed. If the statement is of the "provApplies" variety, the targeted provision is added to the PTP stack. If the statement is of the "provDoesNotApply" variety, the targeted provision is added to the APP list and removed from the PTP stack if it is located there. The targeted provision will not necessarily be in the PTP stack, since the provision may not have been previously added to the stack. Figure 4.22 shows a flowchart of how this process is structured. This procedure is carried out whenever new logic sentences are added to the KB. In the flowchart for the compliance process (Figure 4.19), this occurs any time a process symbol labeled "Update list of provisions to process according to any control statements" is encountered. This approach to conditional control statements maintains the modular design of using separate compliance checking and theorem proving components.

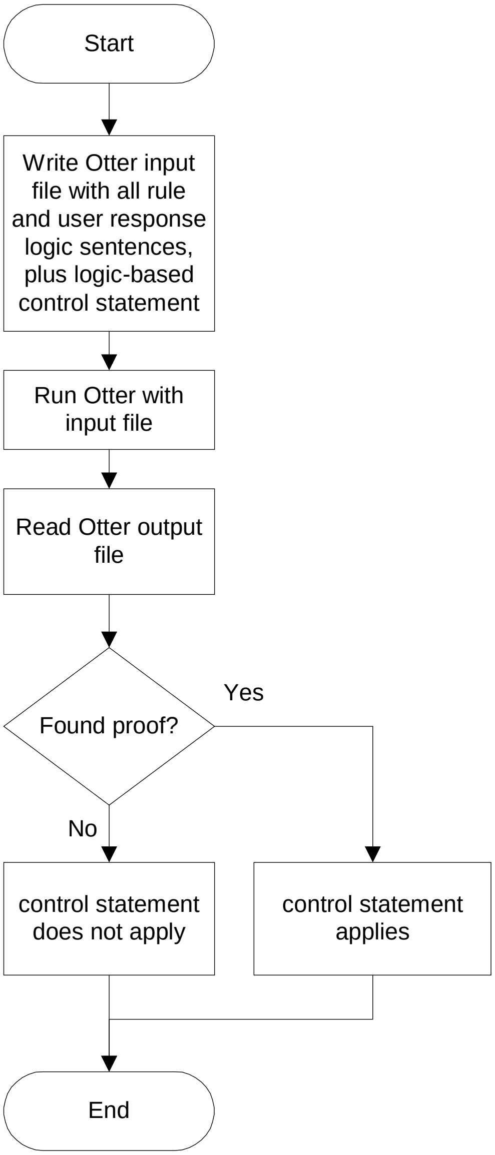

Start
Write Otter input file with all rule and user response logic sentences, plus logic-based control statement
Run Otter with input file
Read Otter output file
Found proof?
Yes
No
control statement does not apply
control statement applies
End

Figure 4.22 Processing logic-based control statements with Otter

4.4 Web-Based System

The web-based Regulation Assistance System (RAS) system allows clean viewing and usage of the information contained in XML regulations, as well as a flexible web interface for compliance checking. The regulation assistance system also provides a demonstration platform to illustrate some of the tools achievable with the metadata in the XML regulation framework. For example, the user can view the regulation text along with questions about the regulation while running the compliance checking procedures.

The RAS system is written as a Java servlet. Servlets are java-based programs that run on servers, similar in usage to CGI programs [22]. The RAS system is run as a web application using Tomcat, a java-based web-server being produced by the Apache Jakarta Project [18].

4.4.1 Overview of RAS Regulation Viewing Features

The web-interface system is able to take advantage of the word and acronym definitions specified in the XML regulations. This is important since the large number of domain-specific terms and acronyms that appear in regulations can make regulation text difficult for novices to understand. The web-based RAS system is able to incorporate explicit definitions of terms and acronyms into its user interface by highlighting words with definitions, and providing pop-up definition or acronym explanations when a user moves the mouse over the highlighted terms. An example of this feature is shown in Figure 4.23.

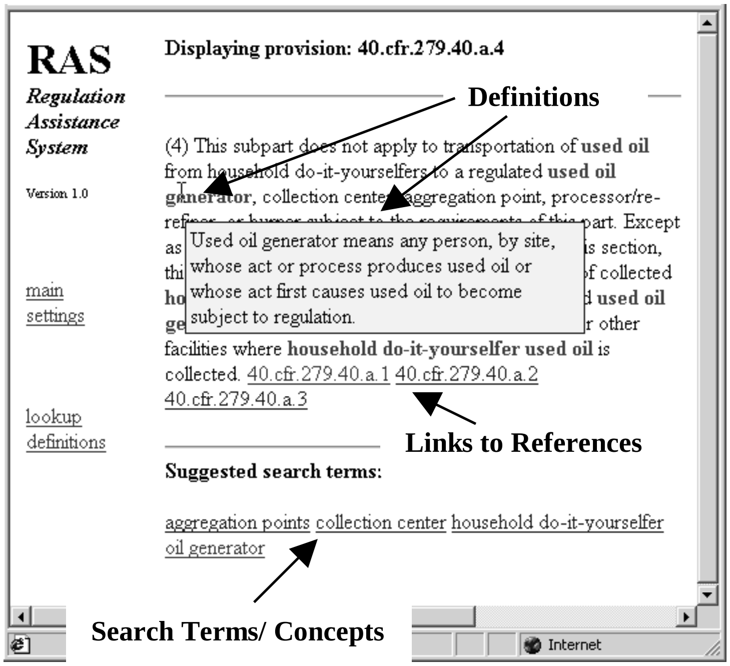
Regulation provisions tend to contain a large number of casual English references to other provisions. These references are cumbersome to look up manually. Moreover, they reduce the readability of the regulation text itself. The RAS system addresses this issue by making use of the references provided by the XML regulation to automatically link to the referenced regulation provisions with hyperlinks so that reading the regulation is less cumbersome. Examples of these reference links are shown in Figure 4.23 as the underlined links following each regulation provision in which a reference occurs.

An important issue for assisting users in determining compliance with regulations is making available appropriate information that assists in determining the meaning of regulation provisions. A variety of documents such as guidance documents, letters of interpretations, and administrative decision provide valuable information that can clarify ambiguous portions of regulations. The RAS system assists the user in locating these documents by using the “concept” elements in the XML regulation to link regulation provisions to the document repository described in Chapter 2. By identifying documents in the document repository that share “concepts” in common with a particular regulation provision, supplementary information that is relevant to that provision can be identified. An example of this feature is shown in Figure 4.23 and Figure 4.24. Figure 4.23 shows the concepts from the XML regulation, which function as predefined search terms, linking to the document repository. Figure 4.24 shows how concept searches can lead to the document repository, and from there users can locate relevant supplementary documents.

4.4.2 Example Usage

The user interface consists of a left-side panel that provides the overall navigation control for the program. This side panel has links to the main control menu for working with the XML regulations, and a “settings” screen where the user can configure RAS.

The main control menu provides a brief overview of the RAS system, along with links to start the system in various modes. From the main control menu the user can initiate a compliance checking session by entering references for the regulation provisions the user



Displaying provision: 40.cfr.279.40.a.4

Definitions

(4) This subpart does not apply to transportation of **used oil** from household do-it-yourselfers to a regulated **used oil generator**, collection center, aggregation point, processor/re-refiner, or burner subject to the requirements of this part. Except as is section, thi of collected **ho d used oil ge** r other facilities where **household do-it-yourselfer used oil** is collected. 40.cfr.279.40.a.1 40.cfr.279.40.a.2 40.cfr.279.40.a.3

Used oil generator means any person, by site, whose act or process produces used oil or whose act first causes used oil to become subject to regulation.

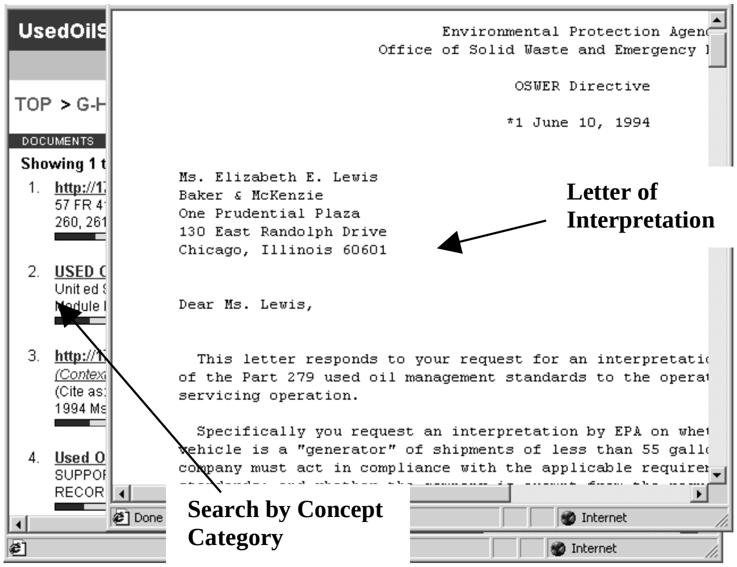
Links to References

Suggested search terms:

aggregation points collection center household do-it-yourselfer oil generator

Search Terms/ Concepts

Figure 4.23 Accessing the document repository through linked concepts



```
                              Environmental Protection Agenc
                           Office of Solid Waste and Emergency I

                                         OSWER Directive

                                        *1 June 10, 1994


Ms. Elizabeth E. Lewis
Baker & McKenzie
One Prudential Plaza
130 East Randolph Drive
Chicago, Illinois 60601


Dear Ms. Lewis,


  This letter responds to your request for an interpretatic
of the Part 279 used oil management standards to the operat
servicing operation.

  Specifically you request an interpretation by EPA on whet
vehicle is a "generator" of shipments of less than 55 gallc
company must act in compliance with the applicable requirem
```

Letter of Interpretation

Search by Concept Category

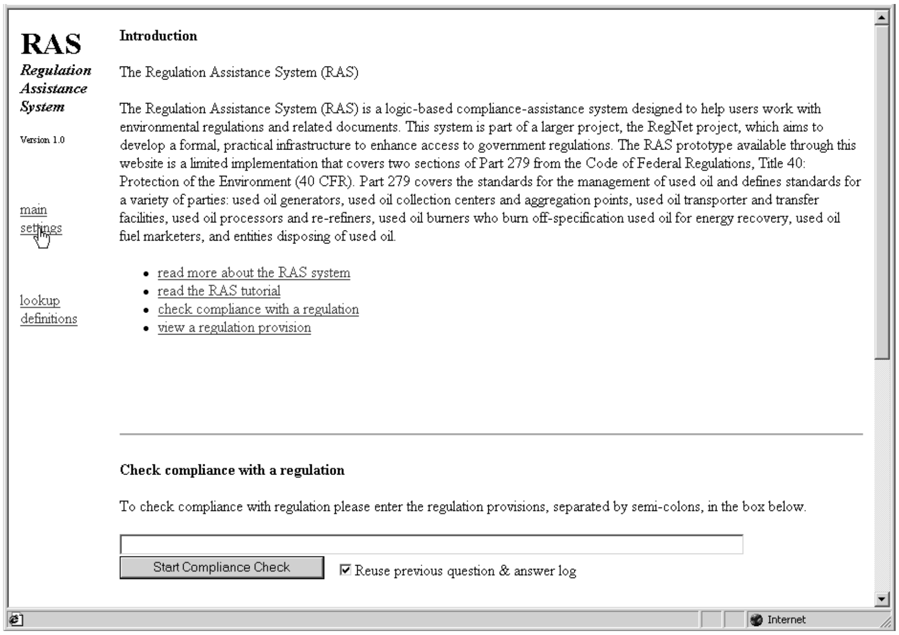
Figure 4.24 Identifying relevant documents though concepts linked from the RAS

would like to check compliance against. The user can also enter a regulation provision to view the text and meta-data for that regulation. This main menu is shown as a screen shot in Figure 4.25.

The RAS settings web page provides the user with a number of viewing configuration options. During the usage of the RAS system, various portions of the XML regulation can be displayed for the user. For example, during compliance checks it is probably useful to see the text of regulation provisions that the compliance questions are in reference to. This enables the user to read the provisions for more information, or simply have a starting point from which to understanding the compliance checking process. The simultaneous viewing of the compliance checking questions and the associated regulation text is possible in the RAS system. Additional check boxes in the settings web page enable the display of other information from the XML regulations, such as the regulation logic, concepts, regulation references, and legal interpretations of the provision.

To initiate a compliance checking session, the user starts from the main menu and selects the compliance checking option. The regulation provision or provisions to be checked are entered in the starting text box by reference, and the compliance check is initiated. Figure 4.26 shows a compliance checking screen shot of the system. The text for the current section of the regulation is shown during the compliance checking process, with the exact provision that the system is asking a question about shown in bold.

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. Just as when viewing regulations, the system automatically inserts hyperlinks to any referenced regulation provisions when performing compliance checks. If selected, these referenced provisions will be displayed for viewing in a new window. The system also displays 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



Introduction

The Regulation Assistance System (RAS)

The Regulation Assistance System (RAS) is a logic-based compliance-assistance system designed to help users work with environmental regulations and related documents. This system is part of a larger project, the RegNet project, which aims to develop a formal, practical infrastructure to enhance access to government regulations. The RAS prototype available through this website is a limited implementation that covers two sections of Part 279 from the Code of Federal Regulations, Title 40: Protection of the Environment (40 CFR). Part 279 covers the standards for the management of used oil and defines standards for a variety of parties: used oil generators, used oil collection centers and aggregation points, used oil transporter and transfer facilities, used oil processors and re-refiners, used oil burners who burn off-specification used oil for energy recovery, used oil fuel marketers, and entities disposing of used oil.

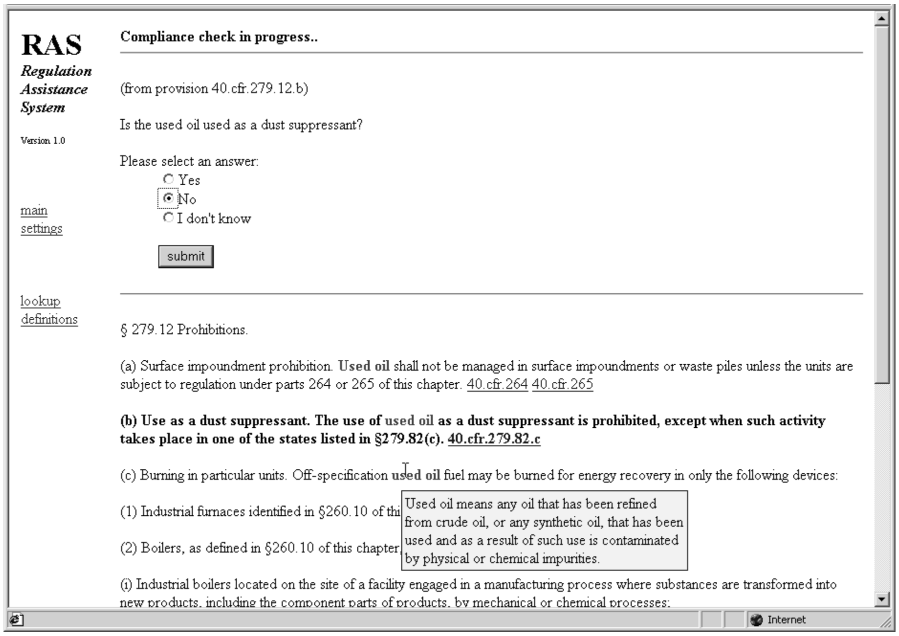
- read more about the RAS system
- read the RAS tutorial
- check compliance with a regulation
- view a regulation provision

Check compliance with a regulation

To check compliance with regulation please enter the regulation provisions, separated by semi-colons, in the box below.

Start Compliance Check ☑ Reuse previous question & answer log

Figure 4.25 Regulation Assistance System main menu



Compliance check in progress..

(from provision 40.cfr.279.12.b)

Is the used oil used as a dust suppressant?

Please select an answer:
○ Yes
● No
○ I don't know

§ 279.12 Prohibitions.

(a) Surface impoundment prohibition. **Used oil** shall not be managed in surface impoundments or waste piles unless the units are subject to regulation under parts 264 or 265 of this chapter. 40.cfr.264 40.cfr.265

(b) Use as a dust suppressant. The use of used oil as a dust suppressant is prohibited, except when such activity takes place in one of the states listed in §279.82(c). 40.cfr.279.82.c

(c) Burning in particular units. Off-specification **used oil** fuel may be burned for energy recovery in only the following devices:

(1) Industrial furnaces identified in §260.10 of thi

(2) Boilers, as defined in §260.10 of this chapter,

Used oil means any oil that has been refined from crude oil, or any synthetic oil, that has been used and as a result of such use is contaminated by physical or chemical impurities.

(i) Industrial boilers located on the site of a facility engaged in a manufacturing process where substances are transformed into new products, including the component parts of products, by mechanical or chemical processes;

Figure 4.26 Regulation Assistance System example compliance check in progress

provision can be displayed and linked, enabling instant access to repository documents related to the provision with which one is checking for compliance. 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. The regulation text may assist the user in better understanding the questions and learning to navigate the regulation. Displaying the references, definitions, and concept links to related documents provides the user with additional information on the regulation provision. Providing all this additional information that is directly relevant to the regulation provision at hand leverages the power of the web to make checking compliance with regulations easier to understand.

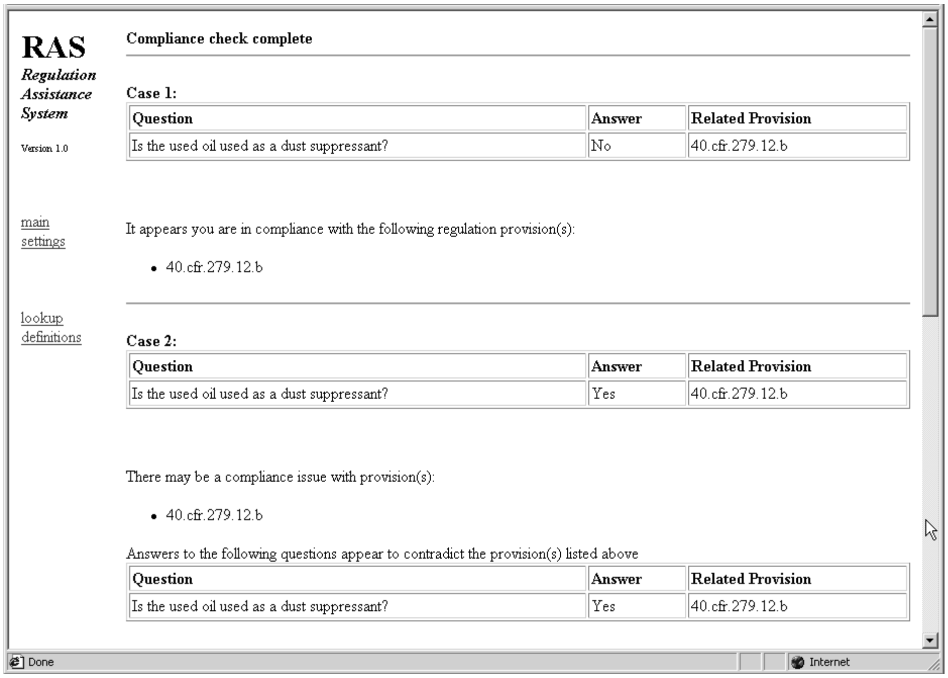
4.4.3 Exploring Possible Compliance Cases

Users of a regulation assistance system do not always know the answer to every question that the system asks them. In addition, users sometimes want to explore different possible answers to investigate how the regulation works. The system developed in this research work allows users to fork the compliance checking procedure along the different answer choices so that they can explore these answer alternatives while the system tracks the different requirements.

In addition to the answers contained in the XML regulation’s logicOption elements, the system adds one more answer for the user to choose from. This additional answer is, “I don’t know the answer”, which provides substancial flexibility to the system. When this option is selected, three options are offered in response. First, the user can choose to explore each of the possible answers for compliance-checking purposes. Exploring each answer means that the system will proceed with the compliance check by checking for compliance based on each of the available answers. If the user answers additional questions with, “I don’t know the answer”, the compliance check process expands to check all the additional answers for the additional questions as well. Once the user has answered all the questions for all the compliance checking cases, the results are presented in the format of, “If X is the case, then you are in compliance. If Y is the case you are not in compliance”. A screenshot of this feature of the RAS system is shown in Figure 4.27.

The second option available to the user when he or she selects an, “I don’t know”, answer is to investigate additional information that may assist with clarifying the question as presented. There are many types of additional information that could be useful, and the RAS system provides those that are available to the user to investigate. For example, the user can be referred to the text of the regulation, which may provide additional background information on the question. The user may be referred to the definitions at the start of the regulation to clarify terminology. There may be legal interpretations or supplementary documents relevant to the provision that RAS can help a user find as well. All these additional information sources can assist the user in determining the answer to the question that they originally were unable to answer.

The third option available to users who select an “I don’t know” answer is to simply stop the compliance checking procedure and resume it at a later time. Users may quit the current compliance check and download a log file containing the questions and answers that they have already answered. After the user determines the information needed to answer the question for which they selected the “I don’t know” option, the user may proceed with the compliance check by uploading the log file and restarting the compliance check. The ability to pause the compliance checking procedure so the user can gather further information is most useful when the user understands the question, but needs to get more information about a business process in order to answer the question.



Compliance check complete

Case 1:

Question	Answer	Related Provision
Is the used oil used as a dust suppressant?	No	40.cfr.279.12.b

It appears you are in compliance with the following regulation provision(s):

- 40.cfr.279.12.b

Case 2:

Question	Answer	Related Provision
Is the used oil used as a dust suppressant?	Yes	40.cfr.279.12.b

There may be a compliance issue with provision(s):

- 40.cfr.279.12.b

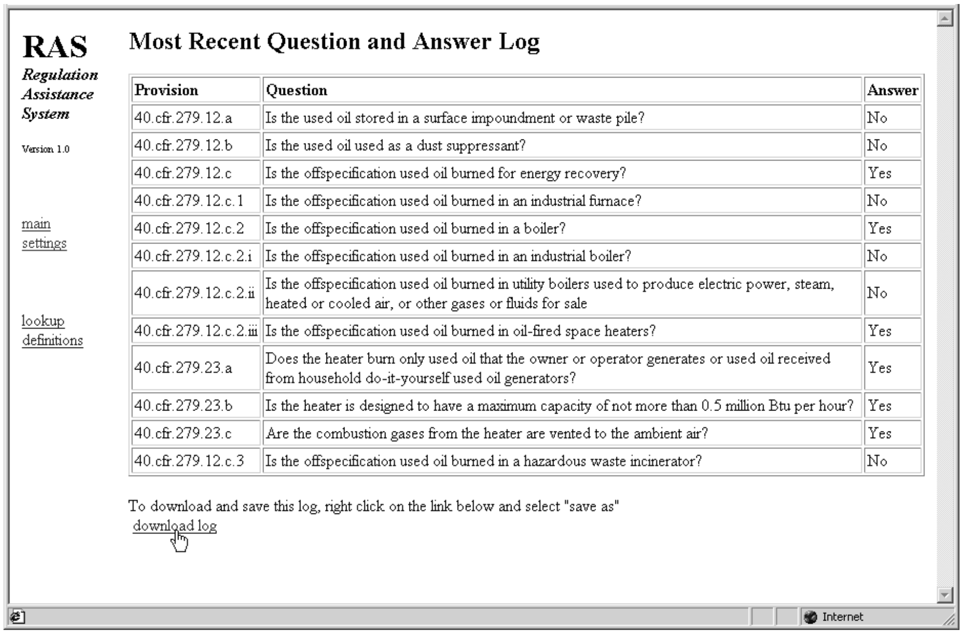
Answers to the following questions appear to contradict the provision(s) listed above

Question	Answer	Related Provision
Is the used oil used as a dust suppressant?	Yes	40.cfr.279.12.b

Figure 4.27 Example of checking multiple answers during compliance checking

4.4.4 Tracking Compliance with an Audit Trail

Following a completed compliance check, a user may view and download a log of the compliance checking session. This is shown in Figure 4.28. This is a feature companies we met with felt would be valuable for record keeping or when revisiting the regulations at a later date. Log files are not only useful for record keeping audit trails; they may be uploaded to the system at a later date. Uploading log files allows users to check for compliance against regulations that have been modified since the previous compliance check. Log files may also be modified to reflect changing operations or allow checking of different scenarios. Modifications to the log file are made by simply removing the answers that a user does not wish to keep. This is shown in Figure 4.29.

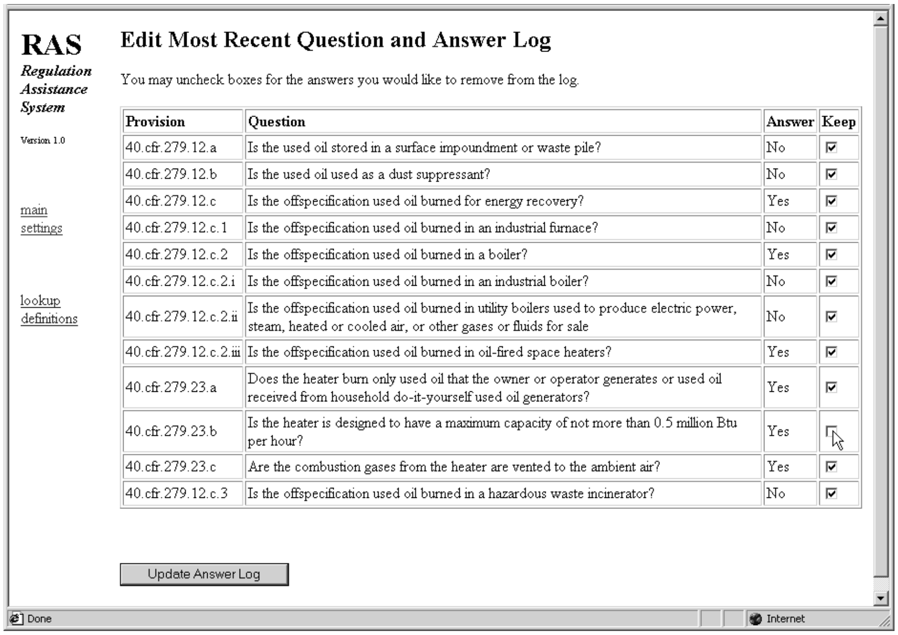


Most Recent Question and Answer Log

Provision	Question	Answer
40.cfr.279.12.a	Is the used oil stored in a surface impoundment or waste pile?	No
40.cfr.279.12.b	Is the used oil used as a dust suppressant?	No
40.cfr.279.12.c	Is the offspecification used oil burned for energy recovery?	Yes
40.cfr.279.12.c.1	Is the offspecification used oil burned in an industrial furnace?	No
40.cfr.279.12.c.2	Is the offspecification used oil burned in a boiler?	Yes
40.cfr.279.12.c.2.i	Is the offspecification used oil burned in an industrial boiler?	No
40.cfr.279.12.c.2.ii	Is the offspecification used oil burned in utility boilers used to produce electric power, steam, heated or cooled air, or other gases or fluids for sale	No
40.cfr.279.12.c.2.iii	Is the offspecification used oil burned in oil-fired space heaters?	Yes
40.cfr.279.23.a	Does the heater burn only used oil that the owner or operator generates or used oil received from household do-it-yourself used oil generators?	Yes
40.cfr.279.23.b	Is the heater is designed to have a maximum capacity of not more than 0.5 million Btu per hour?	Yes
40.cfr.279.23.c	Are the combustion gases from the heater are vented to the ambient air?	Yes
40.cfr.279.12.c.3	Is the offspecification used oil burned in a hazardous waste incinerator?	No

To download and save this log, right click on the link below and select "save as"
download log

Figure 4.28 Viewing log of compliance check



Edit Most Recent Question and Answer Log

You may uncheck boxes for the answers you would like to remove from the log.

Provision	Question	Answer	Keep
40.cfr.279.12.a	Is the used oil stored in a surface impoundment or waste pile?	No	☑
40.cfr.279.12.b	Is the used oil used as a dust suppressant?	No	☑
40.cfr.279.12.c	Is the offspecification used oil burned for energy recovery?	Yes	☑
40.cfr.279.12.c.1	Is the offspecification used oil burned in an industrial furnace?	No	☑
40.cfr.279.12.c.2	Is the offspecification used oil burned in a boiler?	Yes	☑
40.cfr.279.12.c.2.i	Is the offspecification used oil burned in an industrial boiler?	No	☑
40.cfr.279.12.c.2.ii	Is the offspecification used oil burned in utility boilers used to produce electric power, steam, heated or cooled air, or other gases or fluids for sale	No	☑
40.cfr.279.12.c.2.iii	Is the offspecification used oil burned in oil-fired space heaters?	Yes	☑
40.cfr.279.23.a	Does the heater burn only used oil that the owner or operator generates or used oil received from household do-it-yourself used oil generators?	Yes	☑
40.cfr.279.23.b	Is the heater is designed to have a maximum capacity of not more than 0.5 million Btu per hour?	Yes	☐
40.cfr.279.23.c	Are the combustion gases from the heater are vented to the ambient air?	Yes	☑
40.cfr.279.12.c.3	Is the offspecification used oil burned in a hazardous waste incinerator?	No	☑

Figure 4.29 Editing a compliance checking log

4.5 Related Research

Representation of regulations and laws has been an active research area for decades. There has been a great deal of work done on building expert systems for law [87, 104, 111]. T. Bench-Capon provided a review on the applications of knowledge-based systems for legal applications, particularly the research and development efforts related to the Alvey DHSS Demonstrator project in the U.K. [10]. The reference includes a large number of citations that appeared before 1990 that are related to logic and rule based approaches and their application in legal systems. Much of the earlier work in IT and law focused on building systems to optimize decisions with respect to laws, particularly tax law [59]. Some of the recent work has focused on investigations into case-based reasoning and information retrieval [19, 97]. Methodologies for tailoring legal documents to users' needs have also been studied [82]. While legal knowledge representation and reasoning has been an active research topic [73-75], an integrated approach covering the management of regulations, efficient access and retrieval of documents and tools for compliance checking is missing.

In the past thirty years great strides have been achieved in advancing theorem-proving technology [110]. Research for new formalisms and specialized logics [40, 88] continue to improve reasoning speed and non-monotonic reasoning capabilities. The work presented in this thesis builds on these trends in both legal informatics and theorem proving technology.

Two research projects in particular are closely related to the work presented in this thesis. Royles wrote his thesis on the intelligent presentation and tailoring of online legal information [83]. His goal was to take a new look at how expert systems could be interacted with over the Internet to provide advice and document customization services [82]. A prototype implementation was built to provide private consultations with users to help them identify relevant benefits they might be able to collect from the government.

Some of the main features of this prototype are document customization to reduce the amount of unnecessary information provided to users, the use of simple knowledge representation via HTML meta-tags, and a proxy-server to do the intelligent document customization and protect the user’s privacy from the primary information source. The problem areas addressed by Royles’ research work are complementary to those the work this thesis addresses. Royles’ work provides important guidance for how many of the privacy questions that might arise from work in this thesis could be addressed, and provides a model for how a tiered implementation of the compliance system might be implemented.

Jie Wang [106] discussed in his thesis the development of an integrated and distributed information management infrastructure to support hazardous waste compliance, research work that was a precursor to the work presented in this thesis and laid much of the groundwork for it. This thesis develops an information broker model as a solution to the problem. The thesis investigates the information organization of regulations, a distributive framework for the compliance process, and issues of information interoperability for the compliance process. These are important issues in the design of any regulation compliance assistance system.

4.6 Summary

This chapter has discussed the development of a regulation assistance system and related work. The development of a first order predicate calculus based regulation compliance-assistance system and the required extensions to the XML regulation standard has been examined in detail in this chapter. We have described in detail the prototype effort for the regulation assistance system, and the regulation assistance system is illustrated in the domain of used oil management. First, an overview of how the regulation assistance system works, and the motivation for it, was provided as a point of reference. Second, propositional and predicate logic were briefly introduced. Third, the types of metadata added to XML regulations to enable a logic-based compliance-assistance system were discussed. Fourth, the algorithms used for compliance checking were examined. Fifth, use of the RAS system was illustrated. Finally, related research work in this area was reviewed.

Chapter 5

Broader Compliance Perspective

5.1 The Overall Compliance Process

The work presented in the previous chapters has primarily focused on the issue of how one can determine the requirements for a specific regulation provision. Given a particular regulation provision, the work in earlier chapters helps one determine if one is in compliance with the regulation, and what needs to be done if one is not in compliance. This chapter presents how that work fits in the overall compliance process.

The compliance process from the perspective of the regulated community can be broken-down into three general steps, which are illustrated in Figure 5.1. In the first step, one must determine which sets of regulations one must comply with. As mentioned earlier, this can be a very challenging task. Searching for relevant regulations can be done by checking with trade groups, regulators, or searching the Internet. In unfortunate cases, relevant regulations can still be identified by environmental inspectors when they audit the company. This initial step of identifying regulations to comply with is critical, since

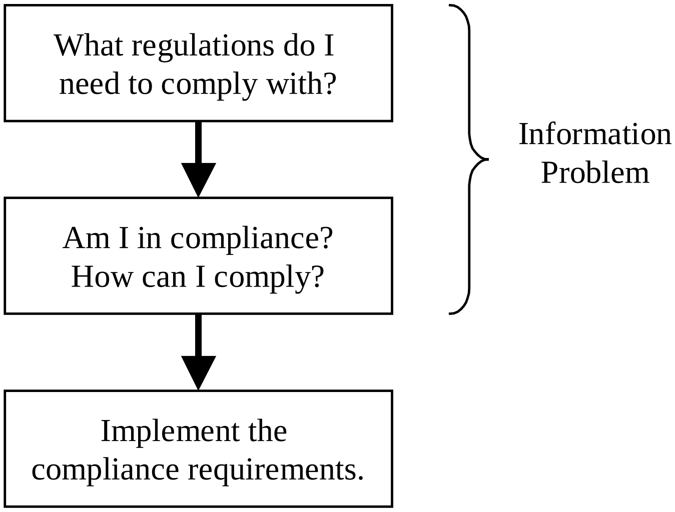

What regulations do I need to comply with?
Am I in compliance? How can I comply?
Implement the compliance requirements.
Information Problem

Figure 5.1 Three general steps for the compliance process

if regulations are overlooked the company is unlikely to be in compliance with them. Identifying regulations that need to be complied with is clearly an information problem, so information technology can be applied to address this problem.

In the second step of compliance process, one must determine if one is in compliance with the regulations, or what needs to be done in order to comply with them. This has been discussed at length in this thesis, and can be a very complicated task. Determining if one is in compliance with a regulation can require extensive research into the regulation requirements, often by investigating supplementary information associated with the regulation.

The third step of the compliance process is to implement the requirements of the relevant regulations in one’s business processes. Implementing the required processes for compliance may be expensive and time-consuming, but the cost associated with complying with the regulations is generally difficult to address with information technology. From an economic perspective, environmental regulation is often the government’s approach to forcing polluters to internalize their external costs by somehow taking into account the cost of particular waste discharges [69]. External costs are costs that are incurred by the environment or society, but are not adequately reflected in market prices. An example of an external cost might be the damage done to the environment when a company disposes of its waste into a river. The government might force a company to internalize this cost by taxing waste disposed into the river or requiring a different method of disposal altogether. If wastes must be disposed in a different, more expensive manner to prevent them from contaminating the environment, there may not be an information technology solution that significantly reduces the cost of this disposal. Information technology may be applicable in some instances to reduce the cost of complying with a regulation. For example, this might be possible by reducing the cost of reporting requirements by streamlining the paperwork and data transfer. In general though, the cost of complying with the requirements of a regulation is associated with taxes or changes in business processes. Unlike the previous two steps in the compliance process, this third step of complying with a regulation generally is not primarily an information problem.

This thesis presents work that has sought to address the information problem component of the compliance process. The RAS system primarily addresses the second of these two compliance steps, i.e., helping to determine if one is in compliance with a regulation, by guiding users through regulations. The RAS system is designed, however, such that it can be used as a component in a larger system that would also assist a user in identifying regulations that need to be investigated, thus addressing the broader information problem of the first two steps in the compliance process. Towards that end, the RAS system is designed such that it can initiate compliance checks at any point within a regulation, and a compliance check can be started by connecting to the RAS system with a target regulation encoded in a web browser's URL. With this design, our work contributes towards reducing the overall information problem associated with complying with environmental regulations.

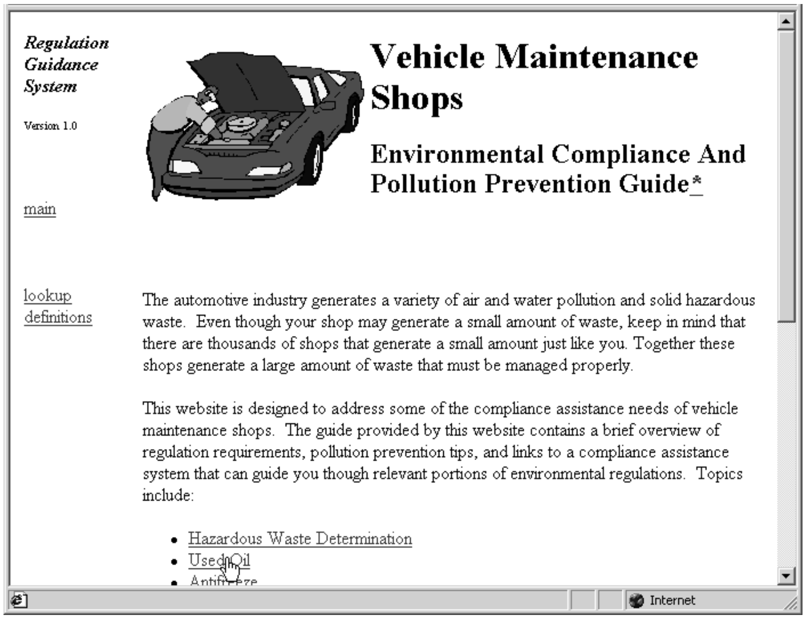
5.2 Example Internet-Enabled Guidance System

To demonstrate how straightforward it can be to build a compliance guide for a specific application utilizing the RAS system and the document repository as a component system, a sample online guide has been built for vehicle maintenance shops. The vehicle maintenance shop online guide is adapted from a paper-based guide developed by the New York State Department of Environmental Conservation Pollution Prevention Unit [67]. Our adaptation is for demonstration purposes only since the original guide provides state regulation references while our online guide links users to federal regulations analogous to the state requirements. Linking to federal regulations is sufficient for developing a research demonstration system, since implementing the state's version of the regulations in a fully annotated XML form can be done using the same methodology that is used for the federal regulations and discussed earlier. In addition, for the case of used oil regulations the New York state regulations are similar to the federal regulations, so linking to federal regulations adequately illustrates the functionality possible with the system.

The online guide for vehicle maintenance shops explains what regulations apply to typical work done in that industry. The guide explains in plain language why vehicle maintenance shops are regulated, and how it is important for the health of the environment that vehicle maintenance shops follow the regulations. The guide then lists a number of common materials and activities used by vehicle maintenance shops in the course of business. Each of these materials or activities has a web page dedicated to explain in plain language regulatory requirements governing the material or activity. The original paper-based guide explains general requirements and then references applicable regulations for more detail. This creates a problem, because when readers are referred to the regulation, they are back to the original dilemma that the guide is attempting to address; the problem of dealing with all the issues associated with finding, working with, and interpreting regulations. Our online adaptation provides a solution to the reference problem in the form of an additional feature that links references to the RAS system. These links enable users to click on referenced regulations, which will connect to the RAS system, to check for compliance by guiding the user through the regulation itself.

Figure 5.2 through Figure 5.4 illustrate the link between the vehicle maintenance shop online guide and the regulation assistance system. Figure 5.2 shows the starting web page for the vehicle maintenance shop online guide, from which users may access information on specific materials or processes, like used oil. Selecting the used oil link brings the user to the web page illustrated in Figure 5.3, which shows the regulatory requirements for used oil. Note the reference in Figure 5.3 to a regulatory provision, 40 CFR 279.23, which is used as a link to the regulation assistance system. Figure 5.4 shows the RAS system, as accessed from the used oil web page of the vehicle maintenance shop online guide. From the RAS system users can check for compliance with the referenced used oil regulation provision or connect to the document repository to look for related supplementary documents. The linking to the RAS system allows users to check for compliance with specific regulatory requirements referenced from the vehicle maintenance shop guide.

Figure 5.5 illustrates how the regulation assistance system and many different online regulation guides can work together. There can be many online guides with functionality similar to that of the vehicle maintenance guide. An individual (the client) attempting to comply with regulatory requirements could identify a relevant online guide addressing the appropriate industry focus for his or her situation. The online guide can then refer the client to the relevant parts of applicable regulations by using hyperlinks to a regulation assistance system. This design allows many different online guides to all refer back to a single regulation assistance system.



Regulation Guidance System

Version 1.0

main

lookup definitions

Vehicle Maintenance Shops

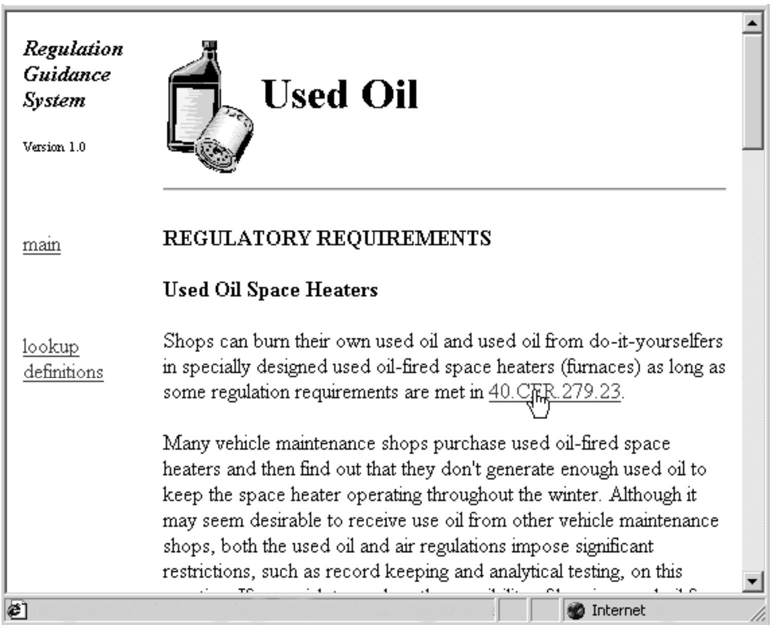
Environmental Compliance And Pollution Prevention Guide*

The automotive industry generates a variety of air and water pollution and solid hazardous waste. Even though your shop may generate a small amount of waste, keep in mind that there are thousands of shops that generate a small amount just like you. Together these shops generate a large amount of waste that must be managed properly.

This website is designed to address some of the compliance assistance needs of vehicle maintenance shops. The guide provided by this website contains a brief overview of regulation requirements, pollution prevention tips, and links to a compliance assistance system that can guide you though relevant portions of environmental regulations. Topics include:

- Hazardous Waste Determination
- Used Oil
- Antifreeze

Figure 5.2 Vehicle maintenance shop compliance guide introduction.



Regulation Guidance System

Version 1.0

main

lookup definitions

Used Oil

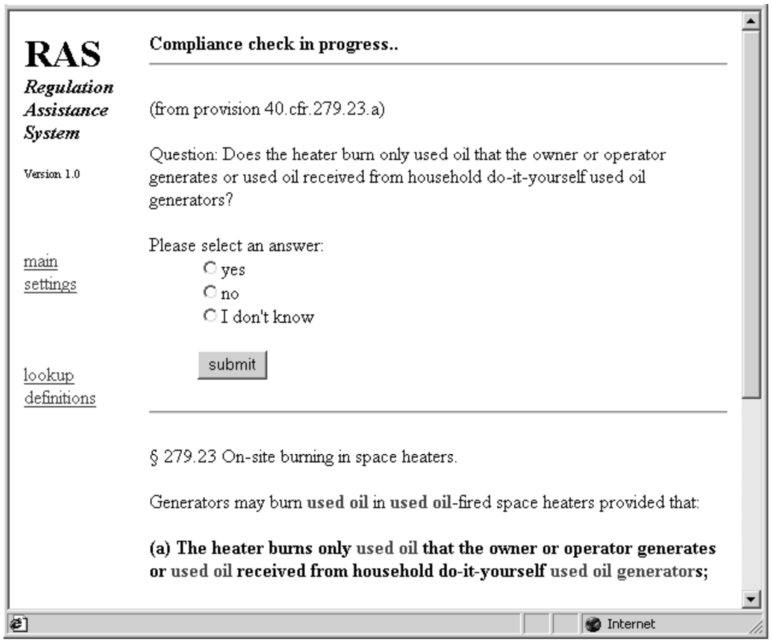
REGULATORY REQUIREMENTS

Used Oil Space Heaters

Shops can burn their own used oil and used oil from do-it-yourselfers in specially designed used oil-fired space heaters (furnaces) as long as some regulation requirements are met in 40.CFR.279.23.

Many vehicle maintenance shops purchase used oil-fired space heaters and then find out that they don't generate enough used oil to keep the space heater operating throughout the winter. Although it may seem desirable to receive use oil from other vehicle maintenance shops, both the used oil and air regulations impose significant restrictions, such as record keeping and analytical testing, on this

Figure 5.3 Vehicle maintenance shop compliance guide for used oil.



Compliance check in progress..

(from provision 40.cfr.279.23.a)

Question: Does the heater burn only used oil that the owner or operator generates or used oil received from household do-it-yourself used oil generators?

Please select an answer:

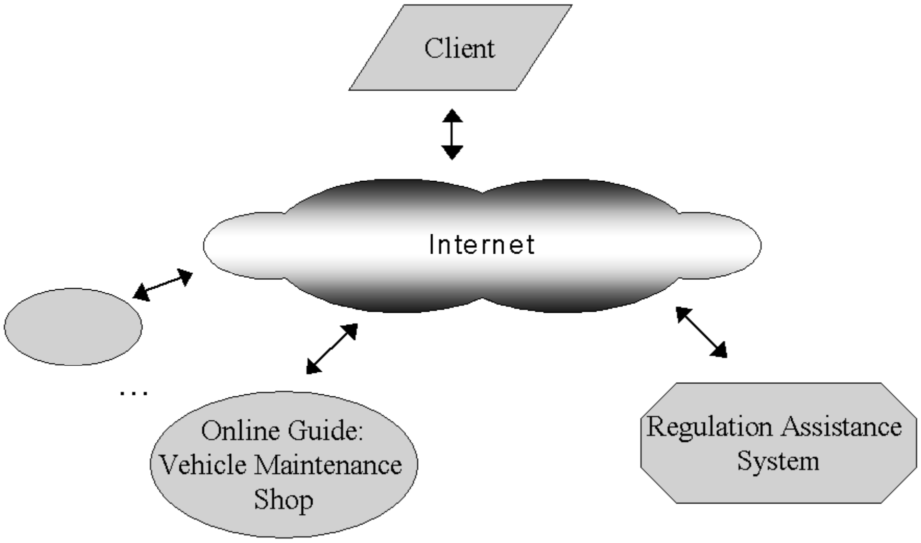
○ yes
○ no
○ I don't know

§ 279.23 On-site burning in space heaters.

Generators may burn **used oil** in **used oil**-fired space heaters provided that:

(a) The heater burns only used oil that the owner or operator generates or used oil received from household do-it-yourself used oil generators;

Figure 5.4 Vehicle maintenance shop compliance guide linked into RAS.



Client
Internet
...
Online Guide: Vehicle Maintenance Shop
Regulation Assistance System

Figure 5.5 Illustration of how online guides can build on a RAS

5.3 Summary

This chapter has presented how the regulation assistance system can be used as a component in a larger and distributed regulation management framework. Online regulation guides, such as the vehicle maintenance shop example, located anywhere around the globe on the Internet, can build upon the compliance-checking capabilities of the RAS system simply by passing target regulations in the web browser's URL as a query string. A variety of compliance guides provided by regulators, industry trade groups, or commercial third party assistance providers could build upon this paradigm by developing online plain language compliance guides linked to regulation assistance systems.

Chapter 6

Summary and Discussion

6.1 Summary and Contributions

This thesis addresses some important problems in the domain of environmental regulation compliance, and has developed an Internet-enabled software framework that can provide regulation compliance assistance services. There are four main contributions made in this thesis:

- First, a document repository containing regulations and supplemental documents has been designed to help categorize these regulatory documents and make them more accessible.

- Second, an XML framework has been proposed and implemented to structure regulation documents and to enable their annotation with a variety of metadata.

- Third, a regulation assistance system has been built that can guide users through regulation requirements to help them determine if they are in compliance and also identify relevant supplementary documents.

- Fourth, the entire system can be used as a component in online industry-specific compliance guides. This addresses the issue of identifying appropriate regulations with which to comply, and illustrates the modularity of our solution.

The development of the regulation software framework is anticipated to accomplish three technical goals:

- Developing a formal structure for representing and annotating regulations to foster future research on, or commercial development of, software systems that address the environmental regulatory information problem.

- Developing a software framework that allows interested parties to more easily identify relevant regulatory information. There are four parties that are interested in making regulations easier to work with. First, businesses, in particular small businesses, that are seeking to comply with regulations often need assistance. Second, policymakers that are developing new regulations need to identify relevant existing regulatory material. Third, regulators that are enforcing regulations require comprehensive access to existing regulatory material. Fourth, the general public and various advocacy groups may wish to critique the current regulatory structure and may desire help learning about it.

- Providing a regulation assistance system to foster higher rates of compliance by helping companies understand the requirements of regulations.

Each of these three goals has been addressed by the work presented in this thesis. First, all of the work presented in this thesis addresses the goal of developing a formal structure for representing and annotating regulations. Second, throughout this research project we gathered feedback from relevant parties to ensure our framework would assist in easily identifying appropriate regulatory information. Third, the development of a regulation assistance system presented in Chapter 4 should facilitate better understanding of regulation requirements.

The work presented in this thesis focuses on an important, but narrow, component of the overall regulation compliance problem. There are a variety of issues, ranging from areas as diverse as regulatory rulemaking to computer security, that need to be investigated from the perspectives of software-based regulation compliance assistance tools. This chapter discusses how the research work described in this thesis fits into the wider context of the legal issues, regulatory issues, and security and privacy issues. Each of these issues will be examined in this chapter, and relevant future work will be identified throughout.

6.2 Future Research

This thesis opens up a wide field of potential research and applications, as well as technologies developed in order to support them. While the work in this thesis has addressed many aspects of the problem of regulation management and compliance, it could only do so to a limited depth, and in a limited application area. Other application areas are bound to present new challenges. While the technology can be shared, the breadth and knowledge needed for the many domains that are affected by regulations cannot be managed in a single setting. We expect to find many distinct efforts in the future, but hope that they will profit from each other's work and be able to interoperate. This section describes some of the most significant avenues deemed valuable for future research in regulatory information management and compliance assistance.

6.2.1 Identifying Regulations for Compliance Checking

In Chapter 5 we introduced an online guide for vehicle maintenance shops that could help them identify regulations that applied to their business. The strategy described in Chapter 5 for addressing the broader compliance problem provides an approach for manually building systems that can assist users in identifying relevant regulations. It is possible for various interested parties to create a range of online guides as web services, each targeting a different audience, so that most businesses will have access to a guide specific to their area of work.

It is useful to investigate the research issue further with the goal of automating to some degree the process of identifying regulations with which one needs to comply. Manually creating and updating a large number of online guides, while an improvement over paper-based guides, is still a difficult task. Investigating approaches to automate the first step of the compliance process, in which one must find the relevant regulations with which to comply, is an important area for future research.

6.2.2 Extending the XML and Logic Framework

The XML structure described in this thesis was tested by implementing the federal used oil regulations, 40 CFR 279, using the XML structure and fully annotating it with metadata. This implementation example has demonstrated the capability of the XML regulation structure to support regulation viewing and compliance assistance services.

There are a number of research issues to be addressed in order to increase the robustness and functionality of the regulation framework. First, software tools should be developed to assist with the logic-based annotation of regulations. The logic and control elements for the 40 CFR 279 implementation were all developed manually. Tools that facilitate the annotation of the regulations with logic could take advantage of the reference metadata or the regulation text itself to improve the speed and quality of developing logic representations. These tools should also assist with maintaining the logic annotations of the regulations. For example, they should strive to identify logical contradictions that may accidentally be added to the regulations and they should help organize sets of logic predicates that have been created.

Second, an approach to mediating logic rules between regulations should be developed. The 40 CFR 279 implementation example was fairly self-contained, but many regulations contain extensive references to other regulations. The current system does not mediate logic between different regulations; all rules are grouped together for compliance checking regardless of where the regulation rules originated. This may create a problem when developing logic annotations for more intertwined regulations, because in order to operate correctly all the regulations must be developed and tested together. If logic representations for two regulations are developed independently, the regulations may have contradictions or other problems [39]. For example, different logic terminology or overlapping namespaces (i.e., the regulations may use the same logic term for two different things) could cause contradictions. A solution to this problem needs to be investigated. Mediation technologies, which provide a systematic approach to handling semantic differences between information sources, may be useful for addressing this problem [109]. Creation of an all-encompassing development environment could also be an alternative.

Third, further research on extending the XML regulation structure should be done. These extensions could allow for multiple logic representations, in different formalisms, within the same regulation document. More advanced annotation structures for the regulations might also allow experts in different fields to provide industry-specific annotations with a variety of legal interpretations.

6.2.3 Legal Issues

There are a number of ways in which the research work described in this thesis will eventually have an impact on the regulatory and rulemaking processes. The regulatory process could be affected through an information feedback loop. The rulemaking process could be affected through the use of logic representations in regulations. In addition we encountered other issues related to the law in this research project, such as the legality of regulation advice systems and the occurrence of ambiguity and contradictions in regulations.

These issues have received a light treatment in this research project, but the results of our preliminary work are sketched out in the following sections. This section also identifies some areas for future research.

6.2.3.1 Legality of Regulatory Guidance Systems

The practice of law by non-lawyers in the United States is illegal. For that reason, the legality of software systems that provide services in the legal arena can be somewhat contentious [8, 98]. For example, in 1998 Parson Technology Inc. was sued in Texas for manufacturing “Quicken Family Lawyer”, a self-help legal software package. A federal District Court in Texas ruled against Parsons, finding that Parsons violated Texas's unauthorized practice of law statute [99]. In response, the Texas Legislature quickly passed a law exempting self-help software with proper disclaimers from the state's unauthorized practice of law statute. The updated Texas law makes an exception for products that, “clearly and conspicuously state that the products are not a substitute for the advice of an attorney” [2].

Despite the Texas case, however, “the larger war over unauthorized practice of law remains anything but settled” [51]. The issue of software that borders on the practice of law is still an issue being discussed in the legal literature and continues to be an area of active research [34, 51]. In light of the growing use of expert systems by the federal government to explain regulations, it appears that the general trend is one of allowing self-help legal software, as long as it contains a legal disclaimer about its use.

6.2.3.2 Precisely Modeling Regulations with Logic

This section will examine problems in attempting to more precisely model regulations than was done in this research work, which modeled regulations in an approximate form. The treatment of ambiguity and contradictions are key issues when using logic to represent the meaning of regulations, particularly when attempting to precisely represent the meaning of a regulation. This is because both ambiguity and contradictions are difficult to represent in logical form.

The issues of ambiguity and contradictions arise in regulation texts intentionally and unintentionally. There are a number of reasons for why ambiguity and contradictions come into existence in regulations.

Sometimes the ambiguity is intended to be built into a regulation. Ambiguity may result from an inability of drafters developing the regulations to compromise on more detailed provisions in a regulation. Therefore the details are not specified. Ambiguity may also occur in regulations because it provides room for discretion on evolving issues. In addition, ambiguity may be accidental or the result of changing technologies.

Contradictions in laws and regulations arise for a number of reasons. Contradictions may result because some regulation provisions are written for some stakeholders, while other provisions are written for other stakeholders. Contradictions may also result from the structuring of laws, or regulations being issued by different governing bodies.

Watershed management in the U.S. Southwest provides an example of how laws originating from different sources can end up conflicting. Kara Gillon, in a law review article on watersheds, describes the situation in the Southwest [37]:

> “The United States has developed separate laws for clean water, clean air, endangered species, irrigated agriculture, and land use management, for implementation at the federal level. … Each of the western states has developed similar, yet diverse, laws governing the allocation and use of water rights,

> administration of groundwater resources, and wildlife management. … Separate agencies administer these laws at federal, state, and tribal levels. … Each entity has different missions, authorities, and modes of operation.
>
> What we are left with is a patchwork of statutes that recognize jurisdictions of state, federal and tribal agencies regarding countless issues affecting a watershed. Where these authorities overlap, it is often difficult for governmental entities to cooperate and share power among themselves as well as the regulated community."

This watershed situation illustrates a situation in which contradictions may occur, intentionally or unintentionally. Contradictions may occur when different agencies are promulgating regulations and these regulations overlap. Contradictions in the regulations may also happen within a single agency's regulations by accident due to carelessness, when one person is combining regulations from different areas of the agency into one document, or when revisions are made over time that interfere with previous rules.[43]

As mentioned in Chapter 4, the research work presented in this thesis relies on first order predicate calculus (FOPC) for modeling regulations, and this form of logic was sufficient for approximately modeling regulations in this research. There is an extensive body of research on other types of logic [44, 45, 85]. The work on advanced forms of logic could be used to investigate modeling regulations more precisely. One reason FOPC was selected for use in the research work presented in this thesis is that it is a standard form of logic that has well-developed and efficient tools for reasoning with it. This may not be the case for some of the alternative forms of logic that may be more expressive, and is an issue that would have to be overcome in order to use them for compliance checking.

[43] In the course of this research work we found an error in the federal used oil regulations, 40 CFR 279, which were first promulgated in 1992. A subsection, 40 CFR 279.64 (e), is currently incorrectly titled "Secondary containment for existing aboveground tanks" instead of "Secondary containment for new aboveground tanks." While this is a simple subsection titling problem rather than a more serious regulation contradiction, it illustrates the difficulty of managing massive regulations and attempting to keep them flawless.

There would be a distinct advantage to using a form of logic that allowed greater precision in modeling regulations than FOPC. Greater precision in logically modeling regulations would result in enhanced capabilities for providing regulation compliance assistance services. In addition, more precise logic modeling of regulations would also enhance the identification of contradictions and ambiguity in regulations.

There are many areas of future research possible in precisely modeling regulations. An area of future research that would advance the work presented in this thesis would be to investigate the application of advanced logics to modeling and computing regulation rules within the XML framework for the purpose of compliance checking.

6.2.3.3 Rulemaking with Logic Representation

What if, when regulation rules are being drafted, the rulemaking agency had to include a logic-based specification of the regulation, in addition to the regulation texts? What effect might this have on the rulemaking process? Answering these questions is an important area for future research, and a preliminary sketch of the area is provided in this section.

An issue that lends support to the idea of developing regulations that contain a logical component from the start is that ambiguities and contradictions must be resolved eventually. Resolving these issues in advance, during the regulation-writing period, makes their resolution more public. The regulation-writing process is designed to be formal and open. The process is designed such that interested parties can keep informed about what is going on, and have an opportunity to influence the process. When ambiguities are “resolved” after a regulation is written, the process is generally not nearly as open. For example, an agency may use its discretion with regard to ambiguous provisions when deciding whom to prosecute and whom not to prosecute for violations. It is generally not possible for interested parties to keep informed about what is occurring. They may never know about potential prosecutions that were not pursued for unstated policy reasons. These issues all point to the idea that regulations should be complete, i.e., containing minimal ambiguity and contradictions, when they are promulgated.

Including a logic component in regulation writing could remove the ability of agencies to "paper-over" issues. A logic component might force the agency to identify more potential contradictions in regulations and resolve more ambiguities, since these contradictions and ambiguities are difficult to model in logical form. Since this would require very careful examination of the regulation rules, this process would result in very carefully structured regulations. However, such an approach would also reduce the discretion available to people enforcing the regulations.

There are also benefits in terms of knowledge sharing and regulatory quality that could result from modeling regulations. Researchers in the Netherlands, working with the Dutch Tax and Customs Administration to model new legislation, have found that there are significant side benefits to modeling legislation for knowledge-based systems. They note that, "these knowledge-based systems proved to be useful by themselves, but perhaps even more important were the side effects caused in creating them" [31]. The researchers cite three major positive side effects of building knowledge-based systems. First, knowledge about the legislation was made explicit through the representation of this knowledge in the document. The explicit representation of knowledge about legislation is important because the knowledge would otherwise remain implicit, and a later need for this knowledge would require identifying an expert on the relevant part of the legislation. Second, explicitly defining knowledge from all the appropriate experts for a piece of legislation makes it easier to determine the validity of an expert's knowledge about the legislation. Third, explicitly defining the knowledge of various experts about legislation allows wider dissemination of this information, thus allowing an organization to more effectively use their likely sparse set of experts and improve the quality of its enforcement actions.

A key feature of modeling regulations as they are developed is that it provides a central repository of expert knowledge. Because expert knowledge, that would have gone undocumented, now is encoded in the regulation, there will be greater sharing of this information within the regulatory community. Modeling regulations in a form of logic at the time they are developed could help ensure regulations are clear and free of logical problems [32]. This is significant because, from a policy standpoint, undocumented ambiguity in regulations is problematic in that it can lead to different enforcement policies and unequal rulings.

One drawback to developing regulations that include a logical specification is that it would add a significant burden to the regulators' task of developing regulations. This is important to note, since it would further increase the workload for already overburdened agencies. In addition to the issue of adding work to develop a regulation, if regulations were not ambiguous, perhaps they would not be passed in the first place. Forcing a resolution of all issues during the development of the regulation might be so overwhelming that the regulation would never be completed.

The issues described above are fertile ground for future research on the effects of logic modeling on regulation drafting. Developing a better understanding of how logic specifications could positively or negatively affect the development of regulations could impact the adoption of logic representations in the regulatory community. This is an important area for future research with important implications for the way regulations are drafted.

6.2.3.4 Regulatory Implications

The research work presented in this thesis does not directly change the regulatory process, but it could affect it in the future. In addition to potentially laying a foundation for using logic representations in the rulemaking process, this research work could have an effect on the regulatory process if it includes a feedback loop that provides regulators with more information.

There are many forms of feedback information that a regulation assistance system could gather for regulators. The following are forms of feedback that our regulatory contacts indicated would be of value to the regulatory community:

- A regulation assistance system could help regulators identify provisions that users feel are ambiguous or confusing. Provisions of this nature could be identified by the frequency with which users select the “I don't know” option during compliance checks.

- A compliance assistance system could provide statistics on what regulatory provisions people are viewing or working with most often. Information of this nature would be straightforward to collect from a regulation assistance system and could provide statistics such as what provisions are the most or least commonly used.

- A compliance assistance system could provide information about how companies are complying with regulations. This includes information about what types of companies are using particular provisions. This could help identify regulation provisions being used differently than expected. Perhaps a minor exception has taken on great importance for many companies. With this information, one could also better predict impact of changes in the regulation. This information might also be used by trade groups to advise members on best practices, or regulators to optimize the structure of regulatory regimes.

- A regulation assistance system could provide more direct feedback from the regulated community. For instance, people could file feedback during a compliance check, with reference to a particular provision. This could assist not only in identifying confusing provisions, but would also quickly bring new industry processes to the attention of regulators.

Online information gathering tools would also be useful for identifying how effective a regulation assistance system is in achieving the goal of improving the rate of compliance and understanding of regulations as a result of using the system. The EPA has a requirement to measure how effective any compliance assistance efforts, such as training, on-site visits, or paper guides, are in terms of changes made by the user as a result of the compliance assistance effort. The approach used by the EPA could be applied to the evaluation of a software regulation assistance system tool as well.

The EPA typically evaluates the results of compliance assistance efforts along three dimensions, applied to regulated entities receiving direct assistance.[44] The first metric is the percentage of entities that report an increased understanding of environmental requirements as a result of the assistance. The second metric is the percentage of entities that report the improvement of environmental management practices as a result of assistance. The third metric is the percentage of entities that report a reduction in pollution as a result of assistance, along with how much pollution was reduced.

Identifying the best way to provide the above types of information, in addition to any other important information that can be gleaned from a regulation assistance system, are important areas for future research. Assessing the impact of both regulatory feedback and system evaluation information on the regulatory process also could be an important area for future research. Gathering this type of information, however, raises some other important questions in the area of privacy and security.

[44] Per email correspondence with an EPA regulatory advisor.

6.2.4 Privacy and Security Issues

The ability of a regulation system to gather the information described in the previous section in a way not previously possible raises a number of privacy and security concerns. In fact, even if a regulation-assistance service provider performs no explicit information gathering, the exchange of sensitive information in the compliance process makes security and privacy a major concern. While privacy and security concerns are not a research focus in this project, it is useful to briefly address them here because they are of such significance for future research.

An important research question is how a regulation assistance system might maintain security and gather aggregate compliance information from the regulated community. Even if regulators have access to information about how companies are complying with regulations, it is important to protect an individual company's information from being inappropriately accessed by regulators or competitors.

There is a large literature on performing secure transactions over the Internet. Various encryption techniques are generally used to provide for secure communication over the Internet [94]. Mediation technologies for creating secure management of information, which could be used to manage access to compliance information, have been recently developed [55, 108]. A great deal of this work could be applied to designing a secure regulation assistance system.

In his thesis on tailoring online legal documents to individuals, Christopher Royles explores a software system design involving a proxy server and a local program to improve the security and privacy of his document customization system [83]. Under Royles' model no significant user information needs to be sent across the Internet, where it might be intercepted, or stored on a remote machine.

The regulation assistance system described in this thesis is currently implemented as a web application that is run on a remote server. The system could also be implemented to run on a client's local machine, as a proxy server, or using some combination of these design configurations. For example, one could implement the regulation assistance system to run on a proxy server maintained by a trade association or some other trusted source. Encryption could be used to protect information exchange between the company performing a compliance check and the trusted proxy server. Another alternative would be to implement a regulation assistance system that runs entirely on a client's machine, only requesting updated regulations to be downloaded from the trusted source. Determining which of these many possible designs would be best for providing a compliance assistance service could be a focus area for future research work.

Research on how to protect privacy and maintain security while providing compliance assistance services is an important area for future research in order for a regulation assistance system to become a practical tool. In addition, the policy question of how much information regulators should have access to with regard to how companies are complying with regulations could prove to be an interesting area for research.

6.2.5 Implementation Issues

This section examines the issue of how this research work might be moved beyond a research prototype and implemented for production use. Our brief exploration of the topic here is in terms of who might implement a compliance assistance system and what legal standing such systems might have.

If the EPA implemented a logic-based regulation assistance system they would most likely treat it as a guidance document. That way the EPA would not be required to abide by the compliance/noncompliance rulings of the system, since the EPA is not legally bound by its own guidance documents. Although the regulated community would probably be more enthusiastic about the technology if the EPA treated logic-based systems like a regulation equivalent, since the system would then be legally binding and thus offer more legal protection, the regulated community would probably still be interested if the logic system was considered a guidance document.

A disclaimer designed to protect against litigation in the event the system provides incorrect advice, something all of the current expert systems described in the introduction possessed, might also accompany any implementation by the EPA. For example, the following disclaimer accompanies all the expert systems that are part of OSHA's elaws suite of systems:[45]

> “STATUTORY AND REGULATORY DISCLAIMER
>
> The Department of Labor is providing this information as a public service. The regulations and related materials are maintained on this system to enhance public access to information on Department of Labor programs. This is a service that is continually under development. While we try to keep the information timely and accurate, there will often be a delay between official publication of the materials and their appearance in or modification of this system. Therefore, we make no express or implied guarantees. The Federal Register and the Code of Federal Regulations remain the official source for regulatory information published by the Department. We will make every effort to correct errors brought to our attention. Further, the advice about rights and obligations provided by this interactive compliance assistance guide depends on the accuracy and completeness of your responses to the questions asked.”

Another possibility is that someone other than the EPA, someone with fewer constraints, could implement such a system. For example, a trade group, commercial service provider, or a company’s internal IT group might build a regulation assistance system. A trade association or other third party building such a system might also add a liability disclaimer. While the regulated community would probably prefer that the EPA implements the system and thereby provides some legal cover, they would probably still

[45] Located at the web address http://www.dol.gov/elaws/, accessed on July 19, 2003.

find the system useful.[46] Although the regulatory assistance system provided by a third party would not have any legal standing, it might demonstrate good faith efforts to comply with the regulations if any regulatory action were taken against companies using it. In addition, the record-keeping benefits of such a system would still have significant value for companies. Another issue to consider is that if the system were implemented outside the EPA, it would be easier for the ambiguities in regulations to continue to exist. This is important if the EPA did not want to remove the ambiguities, since it would result in less pressure on EPA to remove them from the regulations.

Research work on the best way to create a practical service for providing regulation assistance services could be an important area for future research, since this might increase the rate at which such systems are adopted for use.

6.2.6 Summary of Future Directions

This research has attained the goal of building an initial document repository, XML structure for regulations, and a regulation assistance system. There are a variety of important technical research issues that remain to be addressed, however, particularly with respect to the regulation assistance system. These research issues are important for enhancing the capabilities of the system and resolving important issues to make the system more practical outside of the research domain. Some of the more salient research questions for future research are:

- How can automated tools be built to help entities find information on state and federal laws, as well as identify sources of assistance with compliance questions and problems?

[46] In presenting our research work for industry contacts, several companies expressed interest in a commercial version of the system. This provides anecdotal evidence that even if a regulation assistance system does not have legal standing or support from the EPA, companies might still be interested in taking advantage of its functionality.

- How can the XML regulation structure presented in this thesis be extended, particularly to allow other logic formalisms and more advanced annotation with legal interpretations?
- Can more advanced forms of logic be incorporated to more precisely represent the regulation?
- What are the implications of regulators modeling new regulations in logic at the time they are promulgated?
- How will multiple, domain specific systems, interoperate?
- What type of information can be gathered from users of a regulation assistance system, such that the information can improve the regulatory process and the users of the system are not concerned about privacy issues? How might this information affect the regulatory process?
- How can security and privacy be provided when using a compliance assistance system?
- What would be the best model with which to implement a compliance assistance system outside the research domain? Should this be done by regulators, trade groups, commercial service providers, or internally by companies themselves. What are the implications of each of these models?

6.3 Conclusions

Some of the positive outcomes of this work are that it may improve the rate of compliance, reduce the cost and time for both companies and regulators working on compliance, educate individuals involved in the compliance process, and facilitate greater tracking of compliance. Each of these effects is discussed in this section.

First, the work presented in this thesis can improve the rate of compliance. As noted in the introduction, there is evidence that companies would like to be in compliance with environmental regulations, and their failure to do so is often inadvertent. By making it easier to understand the requirements for compliance, thus helping companies that want to comply with the regulations be better equipped to do so, the work presented in this thesis could improve the rate of compliance.

Second, the work presented in this thesis can reduce the cost and time for both companies and regulators working with regulations. Determining how to comply with regulations using the current paper-based approach to compliance checking is time-consuming, risky, and expensive for companies. Regulators also spend a significant amount of time and money helping companies figure out compliance requirements and auditing companies to ensure that they comply with the rules. A software system that reduces the time and cost of these activities may result in a significant cost savings for all parties involved.

Third, the work presented in this thesis can be helpful for educating companies, regulators, or other parties involved with the regulation process. A software system that educates users on compliance requirements, and also fosters convenient research of supporting documents, may improve the overall knowledge of regulation requirements in industry. The approach to working with regulations by guiding users through them may also be useful for training regulators. During interviews with regulators, there was anecdotal evidence that the system would be useful for training, with some regulators commenting that it would be helpful for training new inspectors.

Fourth, the work presented in this thesis can be useful for tracking compliance. Using systems like RAS, users would retain a complete audit trail of how their compliance processes fit with the regulation. This audit trail could facilitate communication of the company's approach to complying with the regulations to different employees or auditors.

This would be particularly useful over a longer period of time, since it could be used as a form of institutional memory. It would also make it easier to note when changes to a regulation might affect the company, and would facilitate quickly rechecking for compliance when regulations are updated.

In conclusion, this thesis presents work that demonstrates how software tools can facilitate compliance with environmental regulations. This interdisciplinary research work has also identified a number of important non-technical areas for future research. It is hoped that the work presented in this thesis will provide the theoretical groundwork for practical implementations of environmental compliance-assistance tools.

Appendix A: XML Regulation DTD

```
<!ELEMENT regulation (logic?, regElement+)>
<!ATTLIST regulation
                id CDATA #REQUIRED
                name CDATA #REQUIRED
                type CDATA #REQUIRED
                versionDate CDATA #REQUIRED
                source CDATA #REQUIRED
>
<!ELEMENT regElement (regElement* | regText? | concept* | reference* | defs? | legalInterpretation? | logic?)*>
<!ATTLIST regElement
                id CDATA #REQUIRED
                name CDATA #REQUIRED
>
<!ELEMENT regText (#PCDATA | paragraph | table | pre | img)*>
<!ELEMENT paragraph (#PCDATA | paragraph | table | pre | img)*>
<!ELEMENT table (td)*>
<!ELEMENT td (tr)*>
<!ELEMENT tr (#PCDATA | paragraph | table | pre | img)*>
<!ELEMENT pre (#PCDATA)>
<!ELEMENT img EMPTY>
<!ATTLIST img
                source CDATA #REQUIRED
>
<!ELEMENT concept EMPTY>
<!ATTLIST concept
                name CDATA #REQUIRED
                times CDATA #IMPLIED
>
<!ELEMENT reference EMPTY>
<!ATTLIST reference
                id CDATA #REQUIRED
                times CDATA #IMPLIED
>
<!ELEMENT legalInterpretation (#PCDATA | paragraph | table | pre | img)*>
<!-- Logic elements -->
<!ELEMENT logic (logicSentence | logicOption | control)*>
```

```
<!ATTLIST logic
                comment CDATA #IMPLIED
>
<!ELEMENT logicSentence (#PCDATA)>
<!ATTLIST logicSentence
                comment CDATA #IMPLIED
>
<!ELEMENT logicOption (question, logicOpt+)>
<!ATTLIST logicOption
                comment CDATA #IMPLIED
>
<!ELEMENT question (#PCDATA)>
<!ATTLIST question
                comment CDATA #IMPLIED
>
<!ELEMENT logicOpt (logicAns | control | logicOption)*>
<!ATTLIST logicOpt
                answer CDATA #REQUIRED
                comment CDATA #IMPLIED
>
<!ELEMENT logicAns (#PCDATA)>
<!ATTLIST logicAns
                comment CDATA #IMPLIED
>
<!ELEMENT control (goto | switchTo | end)*>
<!ATTLIST control
                comment CDATA #IMPLIED
>
<!ELEMENT goto EMPTY>
<!ATTLIST goto
                target CDATA #REQUIRED
>
<!ELEMENT switchTo EMPTY>
<!ATTLIST switchTo
                target CDATA #REQUIRED
>
<!ELEMENT end EMPTY>
<!ATTLIST end
                target CDATA #REQUIRED
>
<!ELEMENT defs (definition+)>
<!ELEMENT definition (term, definedAs)>
<!ELEMENT term (#PCDATA)>
<!ELEMENT definedAs (#PCDATA)>

<!ENTITY % iso-tech PUBLIC "ISO 8879:1986//ENTITIES General Technical//EN//XML"
                "http://www.oasis-open.org/docbook/xmlcharent/0.3/iso-tech.ent">
%iso-tech;
<!ENTITY % iso-num PUBLIC "ISO 8879:1986//ENTITIES Numeric and Special Graphic//EN//XML"
                "http://www.oasis-open.org/docbook/xmlcharent/0.3/iso-num.ent">
%iso-num;
<!ENTITY % iso-grk3 PUBLIC "ISO 8879:1986//ENTITIES Greek Symbols//EN//XML"
                "http://www.oasis-open.org/docbook/xmlcharent/0.3/iso-grk3.ent">
```

```
%iso-grk3;
<!ENTITY % iso-amsr PUBLIC "ISO 8879:1986//ENTITIES Added Math Symbols:
Relations//EN//XML"
                    "http://www.oasis-open.org/docbook/xmlcharent/0.3/iso-amsr.ent">
%iso-amsr;
<!ENTITY % iso-grk1 PUBLIC "ISO 8879:1986//ENTITIES Greek Letters//EN//XML"
                    "http://www.oasis-open.org/docbook/xmlcharent/0.3/iso-grk1.ent">
%iso-grk1;
<!ENTITY % iso-amsb PUBLIC "ISO 8879:1986//ENTITIES Added Math Symbols: Binary
Operators//EN//XML"
                    "http://www.oasis-open.org/docbook/xmlcharent/0.3/iso-amsb.ent">
%iso-amsb;
<!ENTITY min "'">
<!ENTITY sec "&quot;">
<!ENTITY inch "in.">
<!ENTITY Tau "&Tgr;">
```

Appendix B: Reference Parser Grammar and Lexicon

Parser Grammar:

```
REF --> LEV0'
REF --> ASSUME_LEV0 LEV1r'

REF --> LEV2' BackRefKey LEV0

REF --> ASSUME_LEV0 LEV2' BackRefKey LEV1r'
REF --> ASSUME_LEV0 SEC LEV3'
REF --> ASSUME_LEV0 SecSymb LEV3'
REF --> ASSUME_LEV0 SecSymb SecSymb LEV3'

REF --> ASSUME_LEV0 PARA LEV4' BackRefKey LEV3
REF --> ASSUME_LEV0 ASSUME_LEV1 LEV2'
REF --> ASSUME_LEV0 LEV2' BackRefKey LEV1a

LEV0' --> LEV0
LEV0' --> LEV0 CONN' LEV0'

LEV0 --> INT CFR LEV1a'
LEV0 --> INT CFR LEV3'

LEV1a' --> LEV1a
LEV1a' --> LEV1a CONN' LEV1a'
LEV1a' --> LEV1a INTERP LEV1a'

LEV1r' --> LEV1r
LEV1r' --> LEV1r CONN' LEV1a'
LEV1r' --> LEV1r INTERP LEV1a'

LEV1a' --> LEV1s
LEV1a' --> LEV1p
LEV1r' --> LEV1p
```

```
LEV1a --> LEV1s
LEV1a --> LEV1p
LEV1r --> LEV1p

LEV1s --> INT
LEV1s --> LEV1_SELFREF
LEV1p --> PART INT CONL2

LEV1_SELFREF --> txt(this) txt(part) ASSUME_LEV1

CONL2 --> txt(,) LEV2'
CONL2 --> e

LEV2' --> LEV2_SELFREF
LEV2' --> SUBPART UL'

LEV2_SELFREF --> txt(this) txt(subpart) ASSUME_LEV2
LEV2_SELFREF --> txt(this) txt(Subpart) ASSUME_LEV2

UL' --> UL
UL' --> UL CONN' UL'
UL' --> UL INTERP UL'

LEV3' --> LEV3
LEV3' --> LEV3 CONN' LEV3'
LEV3' --> LEV3 INTERP LEV3

LEV3 --> LEV3_SELFREF
LEV3 --> DEC CONL4
LEV3 --> PART DEC CONL4

LEV3_SELFREF --> txt(this) txt(section) ASSUME_LEV3

CONL4 --> e
CONL4 --> LEV4'

LEV4' --> LEV4
LEV4' --> LEV4 CONN' LEV4'
LEV4' --> LEV4 INTERP LEV4'

LEV4 --> BRAC_LL CONL5

CONL5 --> e
CONL5 --> LEV5'

LEV5' --> LEV5
LEV5' --> LEV5 CONN' LEV5'
LEV5' --> LEV5 INTERP LEV5'

LEV5 --> BRAC_INT CONL6

CONL6 --> e
CONL6 --> LEV6'
```

```
CONN' --> CONN
CONN' --> SEP CONN

LEV6' --> LEV6
LEV6' --> LEV6 CONN' LEV6'
LEV6' --> LEV6 INTERP LEV6'

LEV6 --> BRAC_ROM CONL7

CONL7 --> e
CONL7 --> LEV7'

LEV7' --> LEV7

LEV7 --> BRAC_UL CONL8
LEV6' --> LEV6 CONN' LEV6'
LEV6' --> LEV6 INTERP LEV6'

CONL8 --> e
```

Parser Lexicon:

```
CONN --> and
CONN --> or
CONN --> ,

SEP --> ,
SEP --> ;

INTERP --> through
INTERP --> between
INTERP --> to

PART --> part
PART --> parts
PART --> Part
PART --> Parts

SUBPART --> subpart
SUBPART --> subparts
SUBPART --> Subpart
SUBPART --> Subparts


SEC --> section
SEC --> sections
SEC --> Section
SEC --> Sections
```

```
PARA --> paragraph
PARA --> paragraphs
PARA --> Paragraph
PARA --> Paragraphs

BackRefKey --> of
BackRefKey --> in

CFR --> CFR
CFR --> cfr
```

Parse Tree Interpreter Grammar:

```
REF --> LEV0 LEV1 LEV2
REF --> LEV0 LEV3 LEV4 LEV5 LEV6 LEV7
```

Parse Tree Interpreter Lexicon:

```
PTERM --> INT
PTERM --> CFR
PTERM --> DEC
PTERM --> UL
PTERM --> BRAC_INT
PTERM --> BRAC_LL
PTERM --> BRAC_UL
PTERM --> BRAC_ROM

NPTERM --> PARA
NPTERM --> PART
NPTERM --> SUBPART
NPTERM --> SEC
NPTERM --> SecSymb
NPTERM --> txt
NPTERM --> e

SKIPNEXT --> BackRefKey

REFBREAK --> CONN
REFBREAK --> SEP
REFBREAK --> CONN'

INTERPOLATE --> INTERP
```

Bibliography

[1] Administrative Procedures Act (APA), 1946, in 5 U.S.C. §§ 551-59, 701-06, 1305, 3105, 3344, 5372, 7521.

[2] An Act Relating to the Definition Of The Practice Of Law, HB 1507, 76th Legislature, Texas, 1999, (available on the web at: http://www.capitol.state.tx.us/tlo/76R/billtext/HB01507F.HTM).

[3] Anderson, F., Chirba-Martin, M. A., Elliott, E. D., Farina, C., Gellhorn, E., Graham, J. D., Gray, C. B., Holmstead, J., Levin, R. M., Noah, L., Rhyne, K., and Wiener, J. B., "Regulatory Improvement Legislation: Risk Assessment, Cost-Benefit Analysis, and Judicial Review," *Duke Environmental Law and Policy Forum*, pp. 89-138, Volume 11, Number 1, Fall 2000.

[4] Appalachian Power Co. v. Environmental Protection Agency, 208 F.3d 1015, (D. C. Circuit, 2000).

[5] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern Information Retrieval*, Addison-Wesley, 1999.

[6] Bailey, K. D., “Typologies and Taxonomies: An Introduction to Classification Techniques,” *Sage University Papers*, Series on Quantitative Applications in the Social Sciences, 07-102, 1994.

[7] Barker, V. E., O'Connor, D. E., Bachant, J., and Soloway, E., “Expert Systems for Configuration at Digital: XCON and Beyond,” *Communications of the ACM*, Volume 32, Issue 3, pp. 298-318, March 1989.

[8] Baum, M. L. and Lopez, A. S., “Law on the Internet: Unauthorized Practice or Public Access?,” *Proceedings of IASTED International Conference on Law and Technology*, San Francisco, CA, IASTED/ACTA Press, pp. 68-72, 2000.

[9] Beazer East, Inc. v. U.S. EPA, 963 F.2d 603, 3rd Cir. 1992.

[10] Bench-Capon, T. J. M., *Knowledge Based Systems and Legal Applications*, The APIC Series 36, Academic Press, 1991.

[11] Bergmark, D., “Automatic Extraction of Reference Linking Information from Online Documents,” *Technical Report CSTR 2000-1821*, Department of Computer Science, Cornell University, 2000.

[12] Bergmark, D., Phempoonpanich, P., and Zhao, S., “Scraping the ACM Digital Library,” *ACM SIGIR Forum*, Volume 35, Issue 2, pp. 1-7, 2001.

[13] Bertino, E. and Ferrari, E., “Secure and Selective Dissemination of XML Documents,” *ACM Transactions on Information and System Security*, Volume 5, Issue 3, pp. 290-331, 2002.

[14] Boer, A., Hoekstra, R., and Winkels, R., “METALex: Legislation in XML,” *Proceedings of Jurix 2002: Fifteenth Annual International Conference on Legal Knowledge and Information Systems*, London, UK, IOS Press, pp. 1-10, 2002.

[15] Boer, A., Hoekstra, R., Winkels, R., Engers, T. v., and Willaert, F., “Proposal for a Dutch Legal XML Standard,” *Proceedings of EGOV2002: First International Conference of Electronic Government (DEXA 2002)*, Berlin, Germany, Springer Verlag, pp. 142-149, 2002.

[16] Botkin, A., “Wizards, Advisors and Websites, Oh My! Interactive Electronic Tools for Compliance Assistance,” presented at the *National Compliance Assistance Providers Forum*, co-sponsored by U.S. Environmental Protection Agency and Texas Commission on Environmental Quality, San Antonio, December 2002.

[17] Brin, S. and Page, L., “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” *Proceedings of The Seventh International Conference on World Wide Web*, Brisbane, Australia, Elsevier Science, pp. 107-117, 1998.

[18] Brittain, J. and Darwin, I. F., *Tomcat: The Definitive Guide,* O'Reilly & Associates, 2003.

[19] Brüninghaus, S. and Ashley, K. D., “Finding Factors: Learning to Classify Case Opinions Under Abstract Fact Categories,” *Proceedings of Sixth International Conference on Artificial Intelligence and Law*, Melbourne, Australia, ACM Press, pp. 123-131, 1997.

[20] Bush Administration, “E-Government Strategy, Implementing the President's Management Agenda for E-Government,” 2003, (available on the web at: http://www.whitehouse.gov/omb/egov/2003egov_strat.pdf).

[21] *Business Compliance One Stop Workshop*, Small Business Administration, Queenstown, MD, July 24-26th, 2002.

[22] Callaway, D. R., *Inside Servlets: Server-Side Programming for the Java Platform*, 2nd Edition Ed., Addison-Wesley, 2001.

[23] Cannataro, M., Guzzo, A., and Pugliese, A., “Knowledge Management and XML: Derivation of Synthetic Views Over Semi-Structured Data,” *Proceedings of First European Workshop on XML and Knowledge Management Best Papers*, ACM Press, pp. 33-36, 2002.

[24] Carnell, J., Linwood, J., and Zawadzki, M., *Professional Struts Applications: Building Web Sites with Struts, Object Relational Bridge, Lucene, and Velocity*, Wrox Press Inc, 2003.

[25] Chakrabarti, S., “Data Mining for Hypertext: A Tutorial Survey,” *ACM SIGKDD*, Volume 1, Issue 2, pp. 1-11, 2000.

[26] Chalmer, P., Butner, S., and Geyer, G., “500 Channels and Nothing On: What's Missing on the Web?,” presented at the *National Compliance Assistance Providers Forum*, co-sponsored by U.S. Environmental Protection Agency and Texas Commission on Environmental Quality, San Antonio, December 2002.

[27] Chen, H. and Dumais, S., “Bringing Order to the Web: Automatically Categorizing Search Results,” *Proceedings of CHI 2000 Conference on Human Factors in Computing Systems*, SIGCHI, ACM Press, pp. 145-152, 2000.

[28] City of Chicago v. Environmental Defense Fund, 511 U.S. 328, U.S. Supreme Court, 1994.

[29] *Code of Federal Regulations*, Title 40, Part 261, Identification and Listing of Hazardous Waste, section 4, subsection b, paragraph 1., 40 CFR 261.4(b)(1), 2002.

[30] Dawson, E. L. and Davies, L. L., “Book Review: Environmental Law And Policy: Nature, Law, And Society. By Zygmunt J.B. Plater, Robert H. Abrams, William Goldfarb, And Robert L. Graham,” *Stanford Environmental Law Journal*, Volume 19, Number 2, pp. 469-478, May 2000.

[31] Engers, T. M. v., Gerrits, R., Boekenoogen, M., Glassée, E., and Kordelaar, P., “POWER: Using UML/OCL for Modeling Legislation - An Application Report,” *Proceedings of The 8th International Conference on Artificial Intelligence and Law*, St. Louis, Missouri, ACM Press, pp. 157-167, May 2001.

[32] Engers, T. M. v. and Boekenoogen, M., “Improving Legal Quality - An Application Report,” *Proceedings of 9th International Conference on Artificial Intelligence and Law*, Edinburgh, Scotland, UK, ACM Press, pp. 284-292, June 2003.

[33] Fountain, J. E., “Information Institutions and Governance: Advancing a Basic Social Science Research Program for Digital Government,” Report of the May 2002 National Workshop, *Developing a Basic Research Program for Digital Government*, National Center for Digital Government, John F. Kennedy School of Government, Harvard University, 2002.

[34] Fountaine, C. L., “When is a Computer a Lawyer?: Interactive Legal Software, Unauthorized Practice of Law, and the First Amendment,” *University of Cincinnati Law Review*, Volume 71, pp. 147- 179, 2002.

[35] Genesereth, M., *Knowledge Interchange Format*, Technical Report Logic-92-1, Computer Science Department, Stanford University, 1992.

[36] Giles, C. L., Bollacker, K. D., and Lawrence, S., “CiteSeer: An Automatic Citation Indexing System,*” Proceedings of the Third ACM conference on Digital Libraries*, Pittsburgh, Pennsylvania, United States, ACM Press, pp. 89-98, 1998.

[37] Gillon, K., “Watershed Down?: The Ups and Downs of Watershed Management in the Southwest,” *University of Denver Water Law Review*, Volume 5, pp. 395-425, Spring 2002.

[38] Graham, M., Kennedy, J. B., and Hand, C., "A Comparison of Set-Based and Graph-Based Visualisations of Overlapping Classification Hierarchies," *Proceedings of Working Conference on Advanced Visual Interfaces*, SIGCHI, ACM Press, pp. 41-50, 2000.

[39] Gregoire, E., "About the Fusion of Legal Knowledge with Exceptions," *Proceedings of 9th International Conference on Artificial Intelligence and Law*, Edinburgh, Scotland, ACM Press, pp. 91-92, June 2003.

[40] Greiner, R., Darken, C., and Santoso, N. I., "Efficient Reasoning," *ACM Computing Surveys*, Volume 33, Issue 1, pp. 1-30, 2001.

[41] Heenan, C., *Manual and Technology-Based Approaches to Using Classification for the Facilitation of Access to Unstructured Text*, (unpublished manuscript), Engineering Informatics Group, Department of Civil and Environmental Engineering, Stanford University, 2002, (available on the web at: http://eil.stanford.edu/publications/charles_heenan/ClassificationPaper_S.pdf).

[42] Heffron, F. A. and McFeeley, N., *The Administrative Regulatory Process*, Longman, 1983.

[43] Hsu, C.-N. and Dung, M.-T., "Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web," *Information Systems*, Volume 23, Issue 8, pp. 521-538, 1998.

[44] Johnston, B. and Governatori, G., "Induction of Defeasible Logic Theories in the Legal Domain," *Proceedings of 9th International Conference on Artificial Intelligence and Law*, Edinburgh, Scotland, UK, ACM Press, pp. 204-213, June 2003.

[45] Jones, A. J. I. and Sergot, M., "On the Characterisation of Law and Computer Systems: the Normative Systems Perspective," in *Deontic Logic in Computer*

Science: Normative System Specification, J.-J.C. Meyer and R.J. Wieringa, Editors. John Wiley Publ. Co., pp. 275-307, 1993.

[46] Jurafsky, D. and Martin, J. H., *Speech and Language Processing*, Prentice Hall, Inc., New Jersey, 2000.

[47] Krovetz, R. and Croft, W. B., “Lexical Ambiguity and Information Retrieval,” *ACM Transactions on Information Systems*, Volume 10, Issue 2, pp. 115-141, 1992.

[48] Kules, B., Shneiderman, B., and Plaisant, C., “Data Exploration with Paired Hierarchical Visualizations: Initial Designs of PairTrees,” *Proceedings of National Conference on Digital Government Research*, Boston, MA, pp. 255-260, May 2003.

[49] Kushmerick, N., Weld, D. S., and Doorenbos, R., “Wrapper Induction for Information Extraction,” *Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI)*, Nagoya, Japan, Morgan Kaufmann, pp. 729-735, 1997.

[50] Lam, W. and Lai, K.-Y., “A Meta-Learning Approach for Text Categorization,” *Proceedings of 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* ACM Press, pp. 303-309, 2001.

[51] Lanctot, C. J., “Scriveners in Cyberspace: Online Document Preparation and the Unauthorized Practice of Law,” *Hofstra Law Review*, Volume 30, pp. 811-854, 2002.

[52] Lau, G., Law, K. H., and Wiederhold, G., “Similarity Analysis on Government Regulations,” (to appear) *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, Washington DC, 2003.

[53] Lauritsen, M., “Knowing Documents,” *Proceedings of the Fourth International Conference on Artificial Intelligence and Law*, Amsterdam, The Netherlands, ACM Press, pp. 184-191, 1993.

[54] Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., and Lederberg, J., *Application of Artificial Intelligence for Chemistry: The DENDRAL Project*, McGraw-Hill, New York, 1980.

[55] Liu, D., Law, K. H., and Wiederhold, G., “CHAOS: An Active Security Mediation System,” in B. Wangler and L. Bergman, Eds.: *Advanced Information Systems Engineering (CAISE 12)*, Springer LNCS vol.1789, Stockholm, Sweden, pp.232-246, June 2000.

[56] Manning, C. D. and Schutz, H., *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts, 1999.

[57] Marchetti, A., Megale, F., Seta, E., and Vitali, F., “Using XML as a Means to Access Legislative Documents: Italian and Foreign Experiences,” *ACM SIGAPP Applied Computing Review*, Volume 10, Issue 1, pp. 54-62, 2002.

[58] Mauldin, M. L., “Retrieval Performance in Ferret, a Conceptual Information Retrieval System,” *Proceedings of 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Chicago, IL, ACM Press, pp. 347-355, 1991.

[59] McCarty, T., “Reflections on Taxman: An Experiment in Artificial Intelligence and Legal Reasoning,” *Harvard Law Review*, Volume 90, pp. 837-893, 1977.

[60] McCune, W. W., “Otter 3.0 Reference Manual and Guide,” ANL-94/6, Mathematics and Computer Science Division, Argonne National Laboratory, 1994.

[61] Miller, N. E., Wong, P. C., Brewster, M., and Foote, H., “TOPIC ISLANDS – A Wavelet-Based Text Visualization System,” *Proceedings of Conference on*

Visualization, Research Triangle Park, North Carolina, IEEE Computer Society Press, pp. 189-96, 1998.

[62] Muslea, I., “Extraction Patterns for Information Extraction Tasks: a Survey,” *Proceedings of AAAI '99: Workshop on Machine Learning for Information Extraction*, Orlando, Florida, AAAI Press, pp. 1-6, 1999.

[63] Muslea, I., Minton, S., and Knoblock, C., “A Hierarchical Approach to Wrapper Induction,” *Proceedings of Proceedings of the Third Annual Conference on Autonomous Agents*, Seattle, Washington, ACM Press, pp. 190-197, 1999.

[64] *National Compliance Assistance Providers Forum*, co-sponsored by the U.S. Environmental Protection Agency and Texas Commission on Environmental Quality, San Antonio, Texas, December 2002.

[65] Navarro, A., White, C., and Burman, L., *Mastering XML*, SYBEX Inc., 2000.

[66] Needle, J., “The Automatic Linking of Legal Citations,” *The Journal of Information, Law and Technology (JILT)*, Issue 3, 2000, (available on the web at: http://elj.warwick.ac.uk/jilt/00-3/needle.html).

[67] New York State Department of Environmental Conservation Pollution Prevention Unit, *Environmental Compliance And Pollution Prevention Guide for Vehicle Maintenance Shops*, April 2002.

[68] Nolon, J. R., “In Praise of Parochialism: The Advent of Local Environmental Law,” *Harvard Environmental Law Review*, Volume 26, Number 2, pp. 365-416, 2002.

[69] Ortolano, L., *Environmental Regulation and Impact Assessment*, Wiley, New York, 1997.

[70] Paliwala, A., Cartwright, A., and Terrett, A., “User Needs in Electronic Law Reporting: A Research Study of the Law Reports,” *The Journal of Information, Law and Technology (JILT)*, Issue 2, 1997, (available on the web at: http://elj.warwick.ac.uk/jilt/leginfo/97_2pal/).

[71] Pirolli, P., Card, S., and Van Der Wege, M., “The Effect of Information Scent on Searching Information: Visualizations of Large Tree Structures,” *Proceedings of Working Conference on Advanced Visual Interfaces*, Palermo, Italy, ACM Press, pp. 161-172, 2000.

[72] Porter, M. F., “An Algorithm for Suffix Stripping,” *Program*, Volume 14, Issue 3, pp. 130-137, 1980.

[73] *Proceedings of the 7th International Conference on Artificial Intelligence and Law*, Oslo, Norway, ACM Press, June 1999.

[74] *Proceedings of the 8th International Conference on Artificial Intelligence and Law*, St. Louis, U.S., ACM Press, May 2001.

[75] *Proceedings of the 9th International Conference on Artificial Intelligence and Law*, Edinburgh, Scotland, ACM Press, June 2003.

[76] *Proceedings of the National Conference on Digital Government Research*, Los Angeles, CA, Digital Government Research Center, May 2001.

[77] *Proceedings of the National Conference on Digital Government Research*, Los Angeles, CA, Digital Government Research Center, May 2002.

[78] *Proceedings of the National Conference on Digital Government Research*, Boston, MA, Digital Government Research Center, May 2003.

[79] Rechtschaffen, C. “Competing Visions: EPA and the States Battle for the Future of Environmental Enforcement,” *Environmental Law Reporter*, 30 Envtl. L. Rep. 10803, 2000.

[80] *Regulatory Flexibility Act (RFA)*, 5 U.S.C. §§ 601 et seq, 1980.

[81] Romine, M., “Politics, the Environment, and Regulatory Reform at the Environmental Protection Agency,” *Environmental Lawyer,* Volume 6, Number 1, pp. 1-97, 1999.

[82] Royles, C. A. and Bench-Capon, T. J. M., “Dynamic Tailoring of Law Related Documents to User Needs,” *Proceedings of 9th International Workshop on Database and Expert System Applications*, Vienna, Austria, IEEE, pp. 609-613, 1998.

[83] Royles, C. A., *Intelligent Presentation and Tailoring of Online Legal Information*, Ph.D. Thesis in Department of Computer Science, University of Liverpool, Liverpool, U.K., 2000.

[84] Ruckelshaus, W. D., “Environmental Regulation: The Early Days at EPA,” *EPA Journal*, March 1988.

[85] Sanders, K. E., “Representing and Reasoning About Open-Textured Predicates,” *Proceedings of Proceedings of the Third International Conference on Artificial Intelligence and Law*, ACM Press, Oxford, England, pp. 137-144, 1991.

[86] Savola, T., Westenbroek, A., and Heck, J., *Special Edition Using HTML*, Rolan Elgey, 1995.

[87] Sergot, M. J., Sadri, F., Kowalski, R. A., Kriwaczek, F., Hammond, P., and Cory, H. T., “The British Nationality Act as a Logic Program,” *Communications of the ACM*, Volume 29, Issue 5, pp. 370-386, 1986.

[88] Shanahan, M., *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*, MIT Press, Cambridge, MA, 1997.

[89] Shneiderman, B., Feldman, D., Rose, A., and Grau, X. F., "Visualizing Digital Library Search Results with Categorical and Hierarchical Axes," *Proceedings of Fifth ACM Conference on Digital Libraries*, ACM Press, pp. 57-66, 2000.

[90] Shortliffe, E. H., *MYCIN: A Rule-Based Computer Program for Advising Physicians Regarding Antimicrobial Therapy Selection*, Ph.D. Thesis, Graduate Division Special Programs, Medical Information Sciences, Stanford University, 1974.

[91] Skrzycki, C., "The Regulators; Compliance Education Goes Self-Service", The Washington Post, May 23rd, 2000.

[92] *Small Business Regulatory Enforcement Fairness Act (SBREFA)*, Pub Law No. 104-121, March 29 1996.

[93] Spence, D. B., "Paradox Lost: Logic, Morality, and the Foundations of Environmental Law in the 21st Century," *Columbia Journal of Environmental Law*, Volume 20, Issue 1, pp. 145-182, 1995.

[94] Stallings, W., *Cryptography and Network Security: Principles and Practice*, Prentice Hall, 1999.

[95] Stewart, R. B., "A New Generation of Environmental Regulation?," *Capital University Law Review*, Volume 29, pp.21-95, 2001.

[96] Stoll, R. G., "Court Forces EPA to Comply with Due Process Standards," *Andrews Hazardous Waste Litigation Reporter*, 21 No. 4 Andrews Hazardous Waste Litig. Rep. 10, January 5th, 2001.

[97] Stranieri, A. and Zeleznikow, J., "The Evaluation of Legal Knowledge Based Systems," *Proceedings of Seventh International Conference on Artificial Intelligence and Law*, Oslo, Norway, ACM Press, pp. 18-24, 1999.

[98] Ticer, M. A., "Self-Helpless: Will Anybody be Harmed when Legal Software Stands in for Professional Counsel? You Bet," *The Recorder*, San Francisco, March 10, 1999.

[99] Unauthorized Practice of Law Committee v. Parsons Technology, Inc., No. Civ.A. 3:97CV-2859H, United States District Court, N.D. Texas, Dallas Division, Jan. 22, 1999.

[100] U.S. v. Plaza Health Laboratories, 3 F.3d 643, 2nd Cir. 1993.

[101] Usdin, T. and Graham, T., "XML: Not a Silver Bullet, but a Great Pipe Wrench," *StandardView*, Volume 6, Issue 3, pp. 125-132, 1998.

[102] Vincoli, J. W., *Basic Guide to Environmental Compliance*, Van Nostrand Reinhold, New York, 1993.

[103] Voorhees, E. M., "Using WordNet to Disambiguate Word Senses for Text Retrieval," *Proceedings of 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, Pennsylvania, ACM Press, pp. 171-180, 1993.

[104] Wahlgren, P., *Automation of Legal Reasoning*, Kluwer Law and Taxation Publishers, 1992.

[105] Wall, L., Christiansen, T., and Schwartz, R. L., *Programming Perl*, O'Reilly & Associates, Inc., Sebastopol, CA, 1996.

[106] Wang, J., *Distributed Information Organization and Management for Hazardous Waste Regulation Compliance Checking*, Ph.D. Thesis, Department of Civil and Environmental Engineering, Stanford University, 2003.

[107] Widdison, R., "New Perspectives in Legal Information Retrieval,*"* *International Journal of Law and Information Technology*, Vol. 10, Issue 1, pp. 41-70, 2002.

[108] Wiederhold, G., Bilello, M., Sarathy, V., and Qian, X., "A Security Mediator for Health Care Information," *Proceedings of AMIA Conference*, Washington D.C., pp. 120-124, 1996.

[109] Wiederhold, G. and Genesereth, M. R., "The Conceptual Basis for Mediation Services," *IEEE Expert, Intelligent Systems and their Applications*, Volume 12, Issue 5, pp. 38-47, 1997.

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

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

[112] Zohar, M. and Waldinger, R., *The Deductive Foundations of Computer Programming*, Addison-Wesley Publishing Company, Inc., 1993.