For: Garrett
Printed on: Fri, May 3, 1996 10:48:40

Document: bledpper
Last saved on: Fri, May 3, 1996 07:48:30

A WWW–BASED REGULATION BROKER

Garrett, Jr. J. H., Fenves S. J. and Stasiak D. M.

ABSTRACT: There is an ever growing body of regulations covering many aspects of design, operation and disposal of engineered systems, artifacts and products. Constructed facilities are governed by an exceptionally large number of applicable regulations. We have developed a WWW–based broker for providing access and support for using these regulations. In this paper, we describe the underlying representations of the documents and classification systems, the search engine used, and the basic operations conducted by the broker. We also discuss further work we intend to conduct.

KEYWORDS: Regulations, World Wide Web, Full–Text Search, Classifier–based Search

1. INTRODUCTION

There is an ever growing body of regulations covering many aspects of design, operation and disposal of engineered systems, artifacts and products. Constructed facilities are governed by an exceptionally large number of applicable regulations. As globalization increases, engineering organizations must also become aware of and comply with an increasing number of national and international regulations.

To deal with this issue, many large corporations have personnel dedicated to being aware of all regulations that affect their company. These groups may write internal corporate documents which essentially outline or summarize these regulations for those in the corporation who are responsible for ensuring the compliance of that corporation's products with all applicable regulations. However, while this procedure reduces their burden of being aware of the ever changing state of national and international regulations, middle management and designers must now be aware of the still large body of internal corporate documents describing relevant, interpreted regulatory requirements. Simply reading through each and every document may still take a manager or designer a significant amount of time. Hence, even with this internal system for maintaining awareness and compliance within a corporation, it is still unrealistic to expect that a regulation user can easily locate and be aware of every internal document that applies to his or her project. In smaller companies, personnel responsible for ensuring compliance must directly access and search every applicable regulation.

2. THE BROKER CONCEPT

To assist companies, both large and small, in tracking, accessing and using regulations, we propose the concept of a regulation broker. The basic concept of the broker is to provide access to electronic forms of regulations over a distributed wide–area network. The broker serves as the link between regulatory agencies and regulation users. Regulatory agencies provide electronic forms of their regulations in a format compatible with the broker and register the location of these regulations with the broker. Thus, agencies maintain ownership of the document and can update their regulations as needed, but the regulations are centrally indexed, publicized3, searched for and accessed. The broker allows users to enter information about their activities and receive back from the broker the regulations that apply. Regulation users are then able to use the brokers' search facilities to search the regulation down to the relevant, specific provisions.

The broker could be an important computer–based support tool for firms dealing with all aspects of engineered systems, acting as a virtual centralized location where these companies could go to search and access regulations applicable to their activities. Because the broker is operating over a distributed wide–area network, national and international regulatory agencies would be able to access the broker to

register regulations; as regulations are changed, they can be immediately posted. Hence, companies would be able to better track evolving regulations. If most regulatory agencies register their regulations with the broker, companies could easily generate lists of regulations, and specific provisions within those regulations, that are applicable to their business. In summary, the broker appears to be a natural, and necessary, part of any system aimed at helping companies become and remain compliant with applicable regulations.

3. THE PROTOTYPE BROKER—OVERVIEW

We have implemented a prototype broker that is accessible over the world wide web (WWW). The broker is intended to serve as a central access point for regulations; it does not maintain copies of registered regulations. The broker main database only contains the titles, Internet addresses, and keywords of each registered regulation. This eliminates the problem of maintaining large amounts of information on a single server and adds the advantage that the most recent revision of a regulation is always retrieved. In order for a document to be registered and usable within the broker, it needs to contain markup that identifies the various parts of the regulation. Section 4.1. discusses the Broker Form of a document in more detail.

The broker supports several types of searches: full–text searches; classifier–based searches; and browsing. Full–text searching is useful if the words provided in the user query appear in the text, but it does not aid in finding implied concepts. To aid in the latter, the broker also incorporates a classification system. The author of a registered regulation may define several multi–level facets of classification when composing the broker form of a document. Each facet of classification is a tree of mutually exclusive classifiers. Sub–documents are then classified with at most one classifier from each facet of classification. The representation of facets of classification is discussed in Section 4.1.2.

The broker permits a user to browse a document to a certain point, and then issue a query consisting of full–text phrases and additionally specified classifiers. The broker then searches only those provisions below that point in the document and retrieves those provisions matching both the full–text phrase and the specified classifiers. To perform this search, we use Inquery, a probabilistic search engine developed at the University of Massachusets (Callan, 1992). Inquery supports a combination of full text searching with searching over specific fields of the document not visible to the document user (e.g., fields in which classifiers are attached to specific provisions). We describe Inquery in more detail in Section 4.2. We describe how the broker interacts with Inquery in Section 4.4.

4. COMPONENTS OF THE BROKER

The components of the prototype broker, shown in Figure 1, are: (1) the main broker database in which addresses and associated classifiers for registered regulations, in broker form, are stored; (2) the Register Script, a program (CGI script) invoked by the broker to register a regulation in broker form provided by a regulatory agency; (3) the Hypertext Script, a program (CGI script) invoked by the broker to respond to a user's request to traverse the hypertext in a regulation in broker form; (4) the Query Script, a program (CGI script) invoked by the broker to respond to a user's request to search for regulations or provisions in regulations; (5) Inquery, the search engine used to respond to user queries; and (6) the HTML forms used to interact with a broker user via a WWW browser. Each of these components is described in more detail in the following subsections.

4.1. Broker Form of Regulations

A regulation in broker form is a collection of Inquery documents. An Inquery document is a delimited chunk of text to which additional information is attached (see Fig. 3). A regulation can be represented

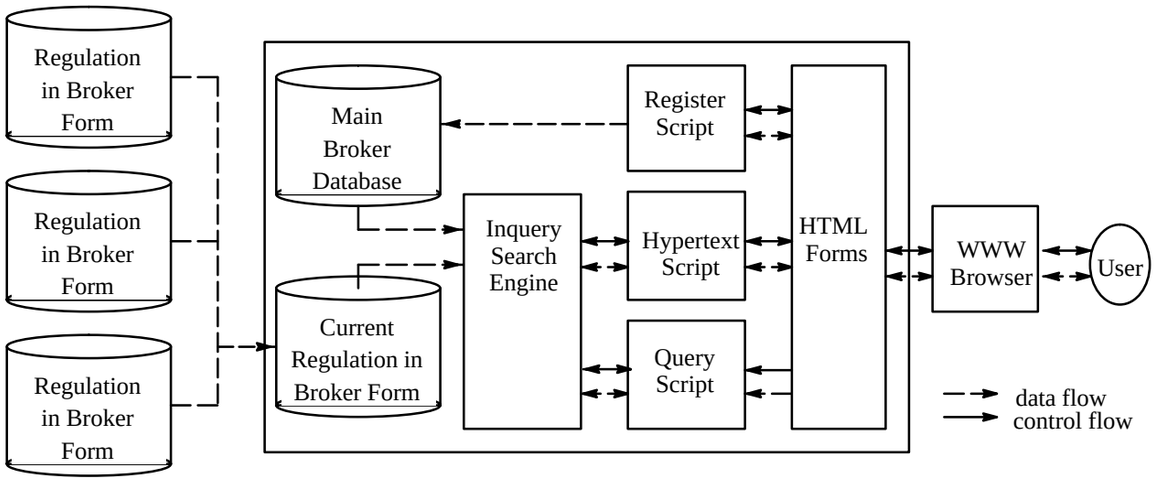


FIG. 1: Components of the Prototype Broker

by one Inquery document or by many. For regulations in broker form, we use many Inquery documents, all stored in the same file, to represent the various pieces of text (sections, subsections, provisions, subparts of provisions, etc.) that comprise a regulation. All Inquery documents are defined using a text markup language (see Fig. 2 for an example of this markup language), which is defined using the Standard Generalized Markup Language (SGML). The Inquery documents are then collected into an ASCII file to create the broker form of the regulation. Broker database files are ASCII text files, so any text editor can be used to create them.

```
<DOC>
<DOCNO>1.2.2</DOCNO>
<TITLE>Phase I – Application Requirements</TITLE>
</DOC>
```

FIG. 2: Example Regulation Document

A regulation in broker form is actually a collection of instances of two types of Inquery documents: (1) regulation documents; and (2) classifier documents. Regulation documents are used to represent the hierarchical structure of the regulation based on the table of contents. Classifier documents are used to represent the various facets of classification by which each of these regulations documents might be classified. The basic structure of each type of Inquery document is the same; additions are made to the basic structure depending on the type being defined (see Fig. 3).

4.1.1. Regulation Documents

As illustrated in Fig. 3, there are three subtypes of regulation documents: (1) internal nodes; (2) provisions; and (3) sub–parts of a provision.

Internal node documents are used to represent the hierarchical structure of the actual regulation document. In other words, they represent the sections and subsections of a regulation. As such, they only have a document number, title, document type set to “internal node”, and associated classifiers. They are assigned document numbers that define how the section fits into the overall structure of the document (see Fig. 4a). This organizational structure is used extensively when the user is browsing the regulation.

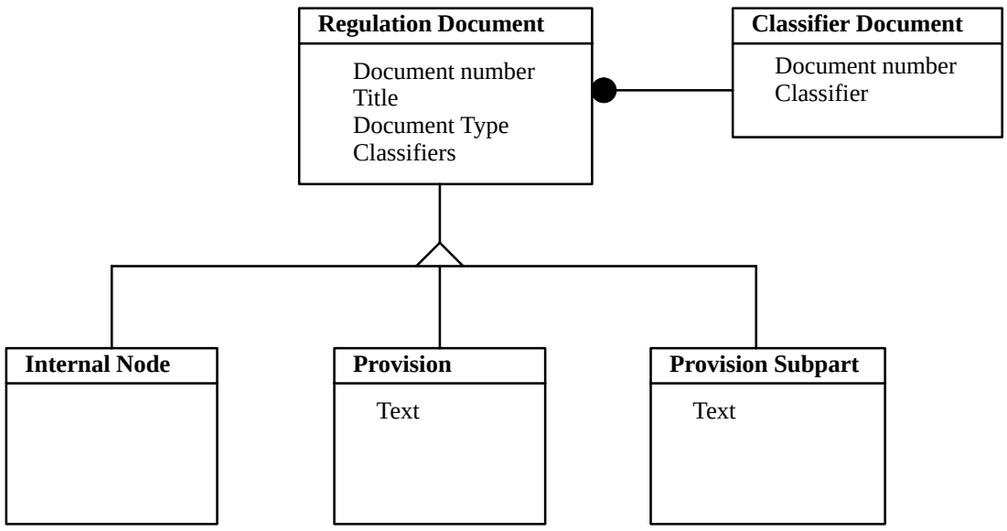


FIG. 3: Classes and Subclasses of Inquery Documents used in Broker Form of Regulation

Provision documents are the first level regulation documents where the text of the regulation appears. Hence, each provision document possesses a document number, a title, the text associated with that provision, a document type set to "provision", and any associated classifiers from the various classification facets.

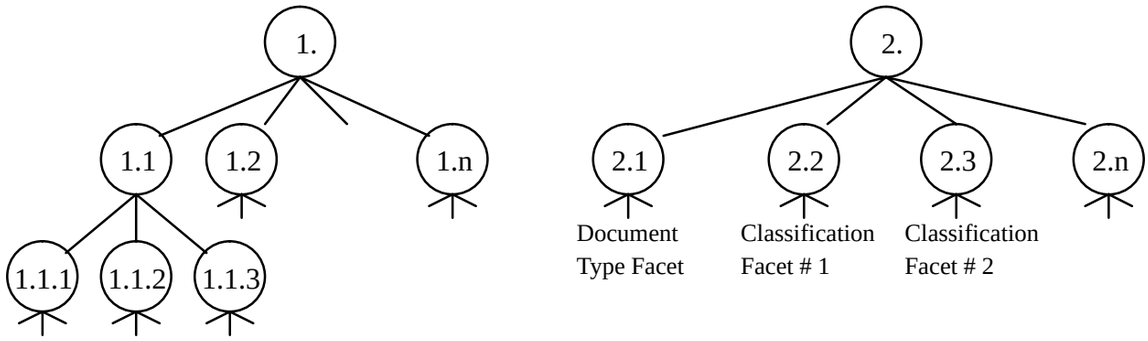
Sub–part documents are further subdivisions of provision documents. Use of sub–part documents is at the discretion of the developer of the model of the regulation if he or she feels that a further breakdown of the document would allow for more specific classification or clarity. The sub–part documents for a single provision may or may not contain all of the text of the provision. A sub–part document may contain non–contiguous parts of the original provision, if the developer feels that these parts may be grouped together and assigned a more specific classifier.

4.1.2. Classifier Documents

Each of the three subtypes of regulation documents may be assigned classifiers from the various facets of classification relevant to the regulation. The classifiers in these classification facets, like the regulation document structure, are also defined in a hierarchically numbered collection of classifier documents (see Fig. 4b). A classifier document defines one classifier of one classification facet. The document number assigned to a classifier places it within the hierarchy of the classification facet. A regulation document is then able to refer to one or more of these classifiers by referring to their document numbers (see Fig. 4c). A regulation document does not have to have a classifier for each facet of classification, if all are not relevant. By representing the classification system along with the regulation text, different regulations can use different classification systems within the broker.

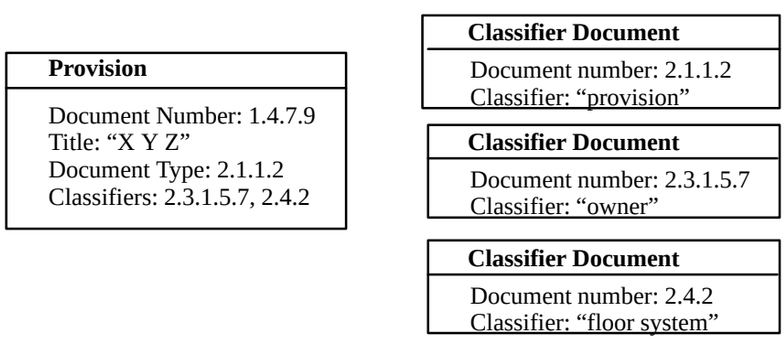
4.2. INQUERY

Inquery is a probabilistic Information Retrieval (IR) system, developed at the University of Massachusetts (Callan, 1992). Based on a form of probabilistic retrieval called the inference network, Inquery retrieves from a collection of documents those documents that are relevant to a keyword query. The inference network models both the collection of documents in an Inquery database and the queries used in retrieving relevant documents. The Inquery system includes software to build an Inquery



a.) Basic Regulatory Document Organization

b.) Classification Facets



c.) Classification of Regulation Document Instances

FIG. 4: Organization and Classification of Regulation and Classification Documents

database from a collection of documents (SGMLbuild) and software to search that database (Inquery search engine).

For a given query, the inference network is used by the Inquery search engine to retrieve documents from an Inquery database in order of decreasing belief value. The belief value of a retrieved document is a value between 0 and 1 that gauges how well a document matches a given query. A retrieved document with a high belief value matches a query more closely than one with a low value.

The inference network essentially employs a bayesian inference network to represent both the document collection (document network) and a user–defined query (query network). A bayesian inference network is a directed acyclic graph in which nodes represent variables and arcs represent dependencies that are weighted by probabilities (Charniak, 1991). The nodes are either true or false and the arcs are assigned values between 0 and 1.

The document network (i.e., the Inquery database), shown in Figure 5, consists of document nodes ($d_1..d_n$) as well as representation nodes ($r_1..r_n$). Document nodes represent the individual documents in the document collection and representation nodes represent concepts found in the document (i.e., words, phrases, dates, numbers). The arcs between these two node types are conditional probabilities, $P(r_i/d_j)$, that the representation, r_i, will be present in the text of a document, d_j.

The query network, also shown in Figure 5, consists of a query node (q), which is the user–defined query text, and concept nodes ($c_1..c_n$). The concept nodes are generated from the text of the query node and represent the individual concepts present in the query. These concept nodes are either true or false

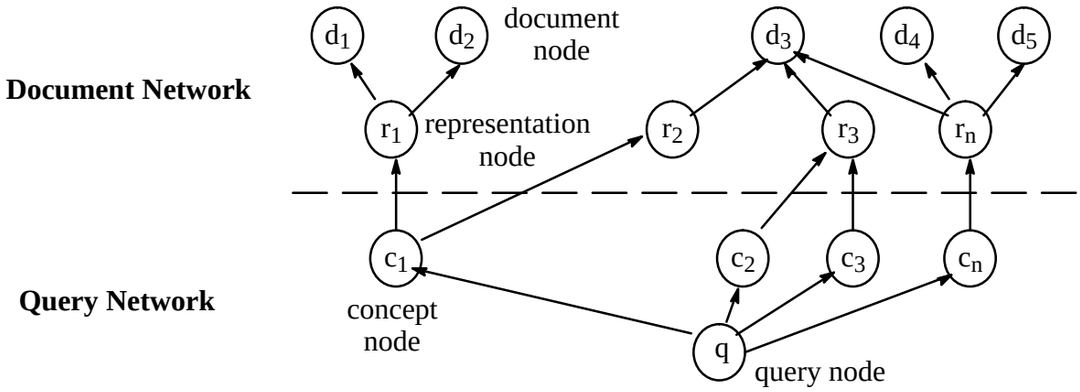


FIG. 5: Inquery's Inference Network Model (Callan, 1992)

and are recombined to form a belief value for the query node using formulae appropriate for the query operators (e.g. AND, OR, SUM).

The query network is dynamically attached to the document network by arcs. These arcs depict that representation nodes in the document network are evidence for concept nodes in the query network. Hence, to calculate the belief values of the query node, the Inquery search engine reasons with the bayesian inference network that links the document nodes ($d_1..d_n$) to the query node (q). Those documents that lead to a belief value for the query node that is greater than a certain threshold are returned as a result of the query.

4.3. Broker Operations Supporting Regulatory Agencies

Regulatory agencies are able to access, login to and use the broker to register a new regulation in broker form with the broker's main database. Once a regulation is in broker form, the regulatory agency must register the correct information with the broker. First, the regulatory agency must login using an id and password issued to the regulatory agency. The regulatory agency then indicates that it intends to register a regulation with the broker. The broker then supplies a broker registration form which prompts the regulatory agency for the title, high level classifiers, and network address of the regulation file. Once the information is submitted, the three arguments are passed to the Register script. The Register script simply adds an entry in the broker's main database containing the classifiers and the network address of the broker database file for the regulation. However, when this information is added, the main database is not yet searchable by the Inquery search engine. To be searchable by the Inquery search engine, the broker's main database must be updated using the SGMLbuild program. After this process is complete, a new regulation is successfully registered and is fully searchable.

4.4. Broker Operations Supporting Users

Regulation users are able to access, login to, and use the broker to: (1) retrieve a hypertext list of applicable regulations from the broker's main database; (2) view a retrieved regulation through its hierarchical levels via hypertext traversal; and (3) search a retrieved regulation for relevant provisions using keywords, predefined classifiers, or both. In these latter three operations, the broker is using Inquery to support its search over the set of documents representing a regulation. Each of these operations is described in more detail in the next three subsections. Due to space limitations, we can only illustrate the described operations with a few screen images of the broker in operation. For more

complete examples of broker usage, please refer to (Krofchik, 1995), (Krofchik, et al., 1995) and (Stasiak, et al., 1996)

4.4.1. Retrieving Applicable Regulations

To retrieve a hypertext list of applicable regulations, the prototype broker first displays a broker search form and prompts the regulation user to enter high level classifiers that describe the type of regulations for which to search (see Fig. 6). Upon submitting one or more classifiers, the arguments are passed to the Hypertext script. The Hypertext script submits an Inquery query containing the regulation user's high level classifiers to the search engine to use in searching over the broker's main database. The main database resident within the broker is searched and a list of applicable regulations is located by the Inquery search engine. The Hypertext script receives the returned list from the search engine and prints on the regulation user Web browser screen a hypertext list of regulations that were found (see Fig. 7). The hypertext list is generated "on the fly" and no physical HTML documents are generated by the system. Each hyperlink to a regulation that appears on the regulation user screen has an associated network address to the broker form file that will be retrieved when the regulation user chooses that hyperlink.

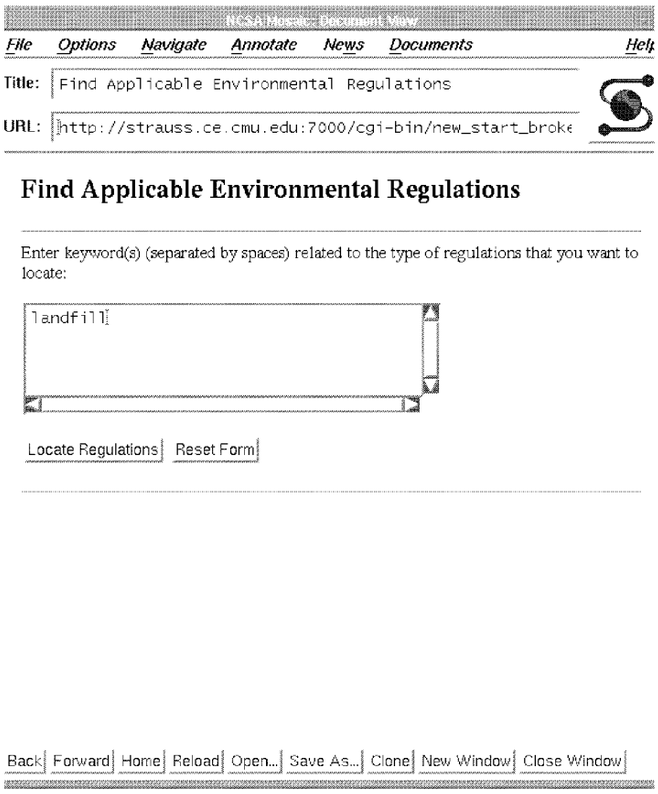


FIG. 6: Broker Search Form

4.4.2. Hypertext Traversal of a Regulation

The process of downloading a chosen regulation to the broker computer and then conducting a hypertext traversal of the regulation is as follows. After a regulation user selects one of the hyperlinks to a regulation, its associated network address is sent directly to the Hypertext script. The Hypertext

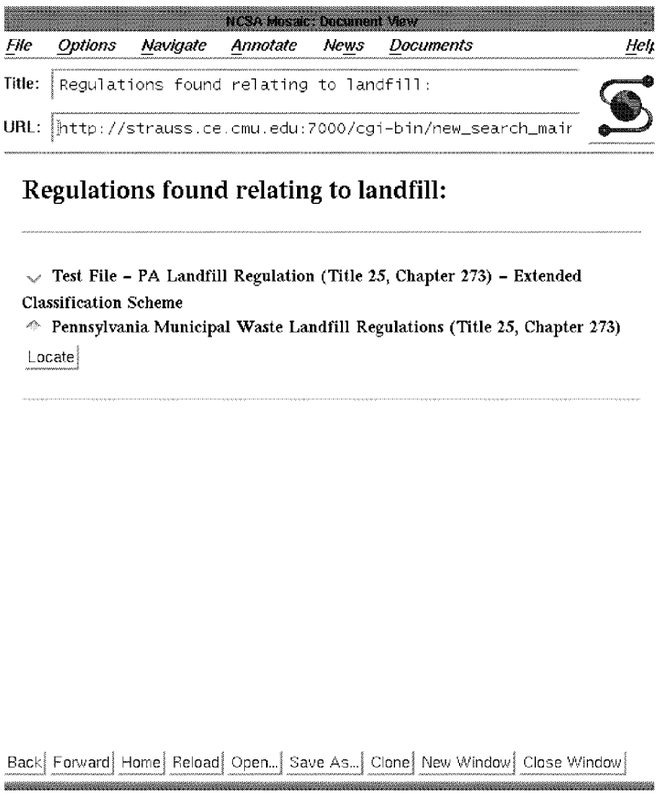


FIG. 7: List of Regulations Returned

script first creates a temporary directory for the regulation user on the broker computer. Then, the regulation file in broker form is transferred from the regulatory agency computer to the temporary user directory on the broker computer. The regulation file is then converted into Inquery form, searchable by the Inquery search engine, using the SGMLbuild program provided as part of the Inquery software. Finally, the first page of the regulation, containing a hypertext list of the first hierarchical level of the regulation, is displayed on the regulation user screen (see Fig. 8). Each of these hyperlinks possesses an associated Inquery query that can be processed by the Hypertext script to retrieve the next hierarchical level of the regulation.

Once this process of downloading the regulation to the broker computer and creating its Inquery form is complete, full hypertext traversal of the regulation can be performed. When the regulation user chooses to delve deeper into the regulation by choosing a hyperlink from the first page of the regulation, the associated Inquery query is sent to the Hypertext script. The Hypertext script then submits the query to the Inquery search engine. The Inquery form of the regulation, in the temporary user directory, is searched and the appropriate text from the next hierarchical level of the regulation is located by the search engine. The text is sent to the Hypertext script and another hypertext list is displayed "on the fly" on the regulation user screen. Each hyperlink in the new list again possesses an associated Inquery query that can be processed by the Hypertext script. This process is repeated until the regulation user can view the text of the provisions within the regulation, which is the lowest level in the hierarchy of a regulation (see Fig. 9). Using this hypertext capability, the user can traverse all levels of the regulation.

4.4.3. Searching a Regulation for Applicable Provisions

In addition to simply browsing a located regulation via hypertext traversal, a regulation user can conduct a search over the regulation for applicable provisions. Each hypertext list (hierarchical level)

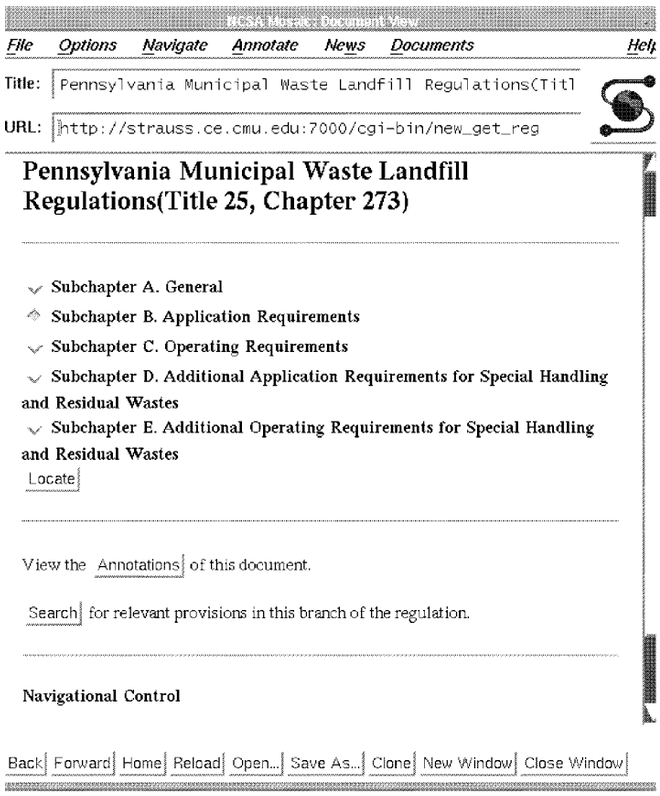


FIG. 8: First Level of Regulation

of a regulation that is displayed during hypertext traversal also presents an option for a regulation user to perform a search for provisions that appear below the current hierarchical level in the regulation (see Fig. 10).

When a regulation user chooses the option to perform a search for provisions, a broker query form is generated "on the fly," into which the user must specify certain arguments with which to search a regulation. The broker query form has a text field for the regulation user to enter text for full text searching of a regulation. Also, the form displays the set of classification facets by which provisions in the regulation have been classified. This list of classifiers is retrieved from the classification documents defined in the regulation file of the regulation (in the temporary directory) and printed "on the fly" in the broker query form presented to the regulation user.

To search for relevant provisions, the regulation user enters any text that is desired, chooses any of the supplied classifiers in the form, or both. These arguments are sent to the Query script, which assembles the arguments into an Inquery query and supplies it to the search engine. The search engine searches and retrieves the document numbers and titles of the provisions satisfying the query. The text is sent to the Query script that, in turn, displays a hypertext list of located provisions on the regulation user's screen (see Fig. 11). At this point the regulation user can choose any of the hyperlinks to view the text of any of the located provisions.

5. FURTHER WORK

Our plans are to expand the present broker prototype to support a realistic field trial, where a representative subset of environmental regulations would be made available to a number of

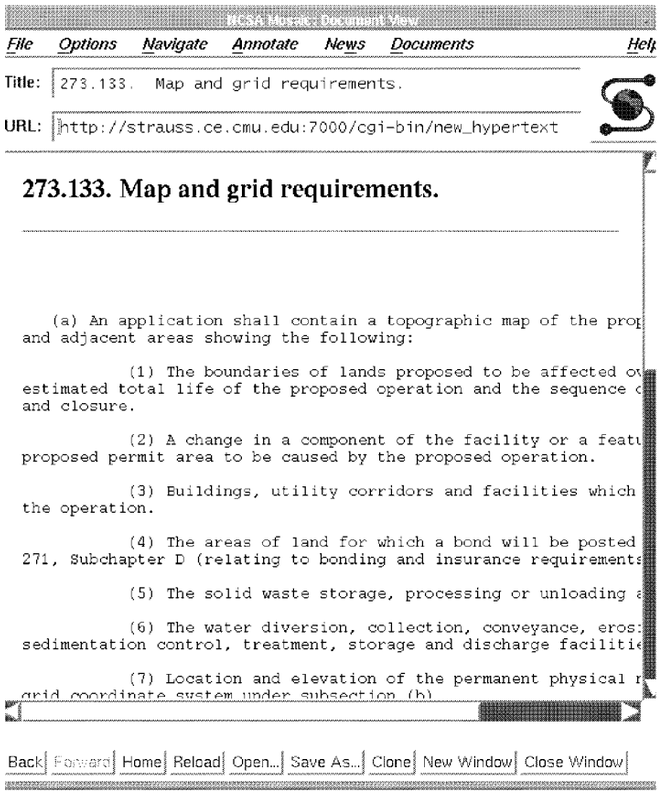


FIG. 9: Provision in Regulation

organizations who would use the broker and provide critiques of capabilities and suggestions for further functionalities. To reach this demonstration level, further work is needed at four levels.

First, the present authoring process to convert the plain text file of a regulation into a searchable database of many small sub–documents needs to be largely, if not fully, automated. Ideally, the author should have to do no more than highlight a portion of the text to be treated as a unit (typically, a provision in the document or a subdivision of a provision); all the markup, indexing and linking to other portions should be done automatically by the system.

Second, the present manual assignment of classifiers needs to be computer–augmented. This activity has two aspects. Classifier facets dealing with the subject of provisions could be readily assisted by a thesaurus, which could be used to match alternate terms, and even provide language translation, for terms appearing in the text. Other potential classification facets, such as the intent of provisions or the parties responsible for approving compliance, are frequently only implicitly stated. It is not yet clear what kinds of aids may be applicable to assist authors in assigning such classifiers, or whether it would be advantageous to depart from the strict tree hierarchy of mutually exclusive and collectively exhaustive classifier trees.

A third, related, issue is that in the present broker the top–level keywords pertaining to regulations as a whole (see Fig. 6) are also manually assigned. The keywords available at this level should be automatically generated as abstractions of the detailed classifiers referring to individual provisions.

Finally, the system needs to be augmented and made more interactive by supporting annotations that parallel the regulation provisions. Several classes of annotation may be incorporated: (1) explanatory

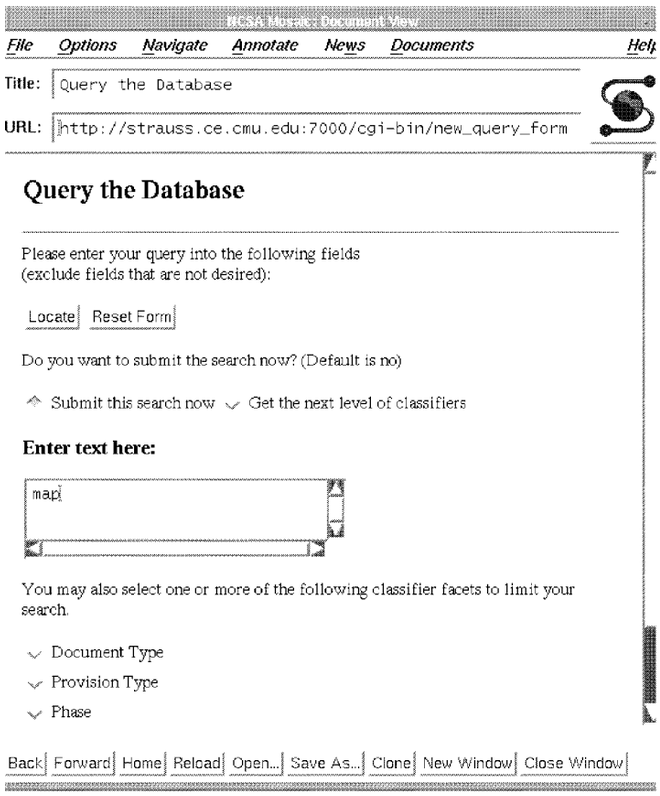


FIG. 10: Launching Search from Lower Level of Regulation

material, references, guidelines, etc., provided by regulatory and enforcement agencies to supplement the regulations proper; (2) corporate annotations whereby large corporations provide guidelines to individual users within a corporation; and (3) user comments, requests for clarification, etc. that may be collected by the broker and supplied to the regulatory agencies. All such annotations can be readily organized as documents with a hierarchical structure paralleling those shown in Fig. 4.

6. SUMMARY

The proposed broker addresses a number of needs. Regulatory agencies retain ownership of the regulations; the broker may eventually be used to impose access and security control and a charging policy for access, if appropriate. Regulations users can access a wide variety of regulations registered with the broker, pinpoint their queries through a combination of full–text and classifier–based searches, and receive only the relevant provisions of the applicable regulations, suitable for immediate use in compliance evaluation. Finally, with the provision of annotations, the broker can facilitate an active two–way interaction between regulatory agencies and users.

7. ACKNOWLEDGEMENTS

Support for this work from IBM under the grant entitled “Product Design for the Environment” and the Green Design Consortium is gratefully acknowledged.

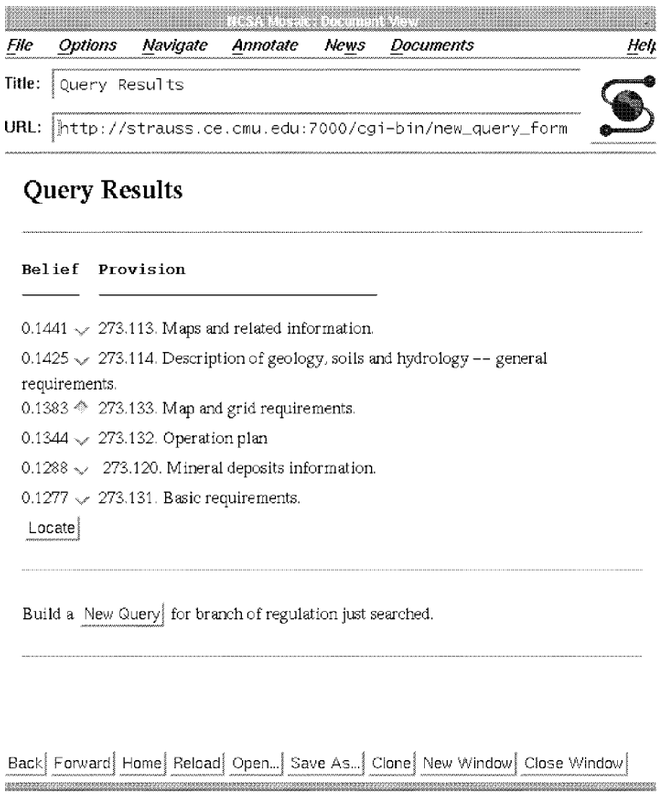


FIG. 11: Results of Search

8. REFERENCES

Callan, J. P. (1992). The Inquery Retrieval System, Proceedings: Third National Conference on Database and Expert Systems Applications, Valencia, Spain, pp. 78–83.

Charniak, E. (1991). Bayesian Networks Without Tears, AI Magazine, pp. 50–63, Winter.

Krofchik, M. A. (1995). A Broker for Delivering Environmental Regulations over the World Wide Web, unpublished MS. Thesis, Department of Civil and Environmental Engineering, Carnegie Mellon University.

Krofchik, M. A., J. H. Garrett, Jr., and S. J. Fenves (1995). A Broker for Delivering and Accessing Environmental Regulations, Proceedings: Second ASCE Congress on Computing in Civil Engineering, pp 36–43.

Stasiak, D. M., J. H. Garrett, Jr., and S. J. Fenves (1996). A Broker for Tracking, Delivering and Using Regulations over the World Wide Web, Proceedings: 1996 IEEE International Symposium on Electronics and the Environment, Dallas, TX, April.