

A Hypertext-based Information Retrieval Network for Environmental Protection Regulations

Wolf-Fritz Riekert and Lilijan E. Kadric¹

Abstract / Zusammenfassung

Since the beginning of 1996, the FAW Research Institute for Applied Knowledge Processing in Ulm, Germany, commissioned by Daimler-Benz AG, has been working on the development of an easy-to-use hypertext system which supplies environmental protection regulations. This hypertext system, referred to as H.I.R.N. (Hypertext Information Retrieval Network), allows for the retrieval of such regulations and straightforward display and usage of the texts that have been found. Furthermore, the system is enriched by an authoring component for entering new texts into the system.

Ein hypertextbasiertes Informationsrecherche-Netzwerk für Umweltschutzvorschriften: *Seit Anfang 1996 arbeitet das Forschungsinstitut für anwendungsorientierte Wissensverarbeitung (FAW) Ulm im Auftrag der Daimler-Benz AG an der Entwicklung eines komfortablen Hypertextsystems zur Bereitstellung von Umweltschutzvorschriften. Das aus dieser Entwicklung hervorgegangene System H.I.R.N. (Hypertext-Informationen-Recherche-Netzwerk) ermöglicht die Recherche nach Umweltvorschriften sowie die unmittelbare Anzeige und Nutzung der gefundenen Texte. Darüber hinaus verfügt das System über eine Autorenkomponente zur Eingabe neuer Texte.*

1 Introduction

The traditional way of documenting and disseminating environmental protection regulations, i.e., through loose leaf collections, is increasingly viewed as outdated and uneconomical by modern industrial companies. The compilation of loose leaf collections is expensive, the process of updating them by manual insertion and removal is time-consuming for the user, and the retrieval and usage of information is impeded by the medium paper.

¹ Dr. W.-F. Riekert and L.E. Kadric, FAW – Research Institute for Applied Knowledge Processing, PO Box 2060, D-89010 Ulm, Germany, phone: (+49) 731 501 500, fax: (+49) 731 501 999, email: riekert | kadric @faw.uni-ulm.de

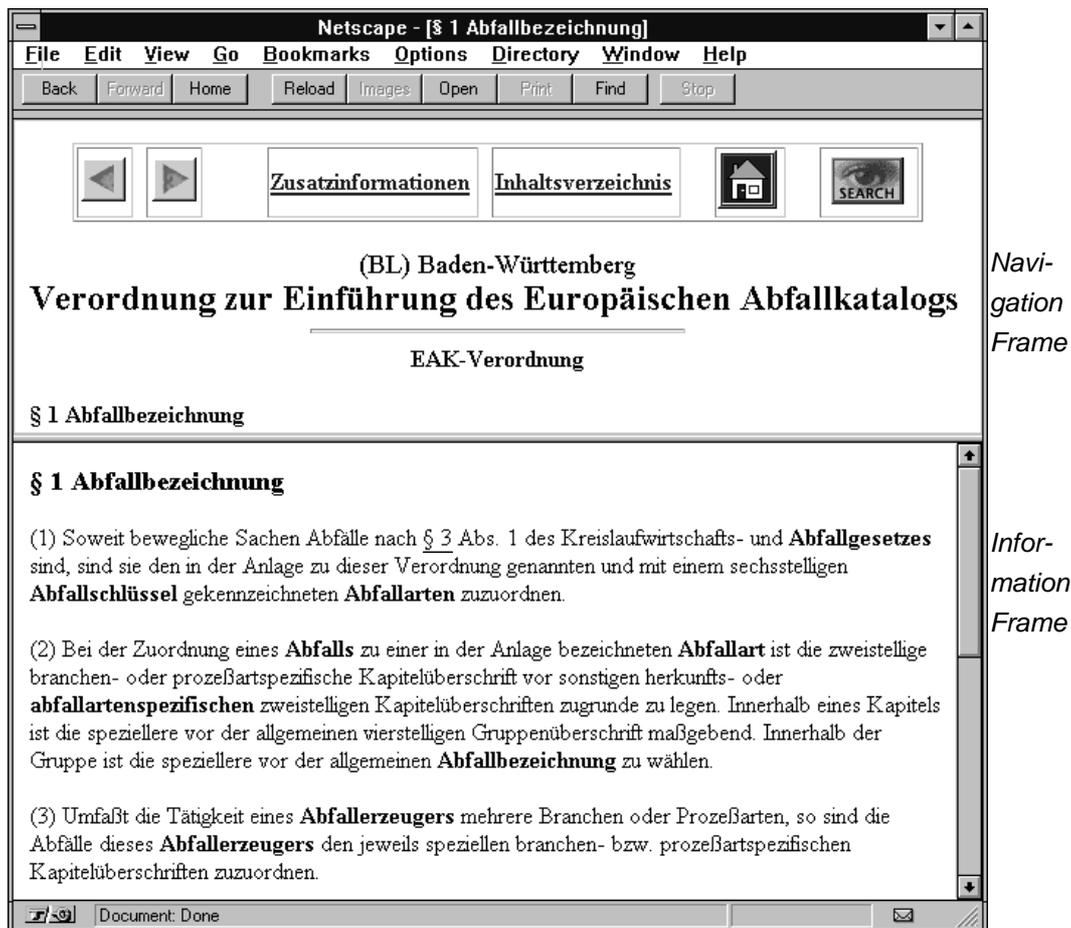


Figure 1: The user interface of the H.I.R.N. system. The upper part of the hypertext document is the *Navigation Frame* containing the title of the regulation and the name of the current section as well as function buttons to reach other pages such as the previous and next sections, the annotation page for the section, a list of contents, the H.I.R.N. homepage, and the search facility. The lower part of the document referred to as *Information Frame* contains a scrollable display of the current section. Cross references to other sections (here: “§3” of another regulation) are presented as hyperlinks that can be followed by a mouse click. The page shown is the result of a search for the keyword “Abfall” (i.e. waste), therefore all words containing this text string are displayed in bold typeface.

This fact was the essential trigger for Daimler-Benz to cooperate with the FAW in developing H.I.R.N.², a hypertext-based information retrieval network for environmental protection regulations. The usage of commercially available legal information systems was not intended because these systems are, on the one hand,

² H.I.R.N. is a trademark of Daimler-Benz AG. An earlier version of the system bore the name HyperLex which has now become obsolete (Riekert 1996).

not adapted to the information requirements of a car manufacturer and, on the other hand, do not allow for a subsequent adaptation to the users' requirements.

H.I.R.N., however, is enriched by an open authoring component. With this authoring component it is possible to include exactly those regulations into the system which are required by the users. A rapid updating cycle offers immediate access to any changes concerning the laws and regulations for environmental protection. The cumbersome work of updating information, e.g., maintaining loose leaf collections, is thus avoided. Instead, the stock of documents in the server systems is updated by exchanging storage media or by transferring data via networks.

The retrieval techniques offered allow for a rapid look-up of the relevant regulations. The texts of these regulations need no longer to be searched for in unwieldy loose leaf collections or procured from a library. Instead, they can immediately be retrieved through a computer monitor and directly transferred into a text processing system, e.g., to compile studies or formulate expert opinions. In addition, users may add individual comments to the regulations stored in the system.

The system has been designed to manage original texts rather than mere references to the regulations. At the final stage, environmental protection regulations at international, national, and regional levels as well as those of specified communities will be integrated in the system. In addition, the system manages internal environmental protection regulations of the Daimler-Benz company.

An essential goal of the development of the H.I.R.N. system was the creation of a scaleable solution which can be used in the Internet, in corporate networks, in local networks, as well as in stand-alone installations. A further goal was to offer the functions of H.I.R.N. on all major client platforms. Costly and inflexible proprietary solutions were to be avoided.

For these reasons, a system concept was chosen which is based on the World-Wide Web technology. The Web allows for the realization of client-/server solutions according to the hypertext-paradigm. It also allows for the integration of data, multimedia information, and functions through a standardized interface. The Web is a non-proprietary standard which is offered on all modern computer platforms. The software components required for using the Web are inexpensive or may even be used free of cost as public domain software.

A relational database system is used for managing the regulations. The link with this database is realized mainly in a supplier-independent manner by using the ODBC³ interface.

Relatively inexpensive PC systems running Windows NT can be used as server systems. However, the H.I.R.N. server software can easily be adapted to other system platforms since it is programmed in PERL, a script language which is independent of the underlying operating system.

³ Open Database Connectivity

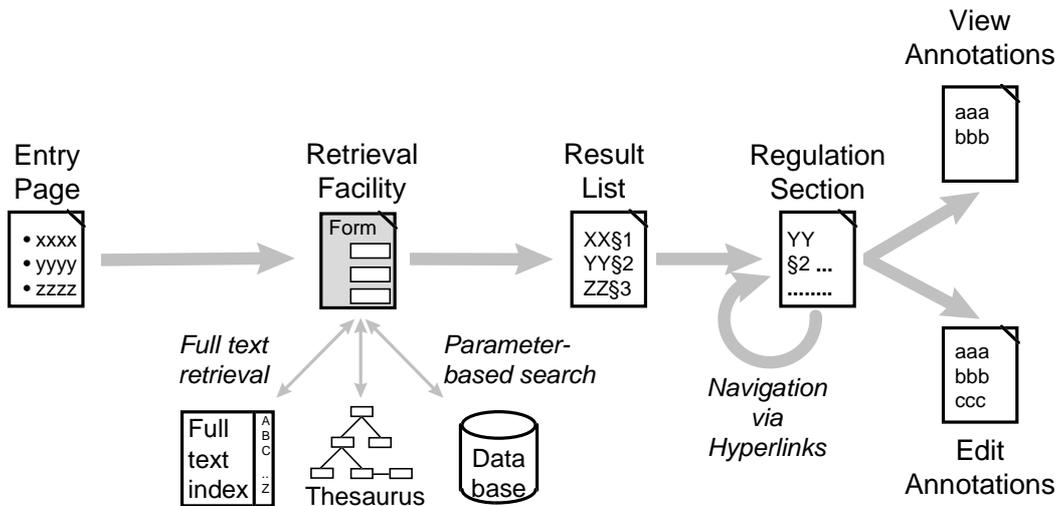


Figure 2: H.I.R.N. from the user's point of view

2 H.I.R.N. from the User's Point of View

Figure 2 shows the functions of H.I.R.N. from the user's point of view. The user interface follows the hypertext paradigm. Using the system is, for the user, comparable to navigating in hypertext. The various retrieval functions offered by the system can be reached through an entry page.

The *fulltext retrieval* facility allows searching for and retrieving regulation sections containing specific words or character sequences. These fulltext retrieval criteria can also be combined by logical operators such as “and”, “or”, and “not”.

Currently, a *thesaurus component* is being integrated into the system. The thesaurus consists of a semantic network of keywords and their relations (such as associations with synonyms as well as broader, narrower, and related keywords). A thesaurus browser lets the user navigate through the semantic network and select the keywords which are most suited for the retrieval of the information in question. In addition, the semantic relations between keywords can automatically be evaluated by the system, thus allowing a search for a keyword which does not occur literally in the regulations but is implicitly associated with the text in question because it contains a semantically related keyword, e.g., a synonym.

The search can be limited to specific regulations by specifying individual *document parameters*, such as the type of the regulation, the name of the regulation or the name of the legislative body that issued the regulation.

The result of using these retrieval facilities is a list of relevant section titles which are summed up on a result page. By clicking onto such a title, the user can reach the respective text section, e.g., a paragraph which is shown on a separate hypertext page. Now it is possible to browse through this paragraph or to jump to neighboring sections and a table of contents of the regulation. In addition, cross

references to other paragraphs are presented in the form of hyperlinks that can be followed by a mouse click. Public and private annotations can be added to each section via a separate hypertext page. Users can view the annotations already added – as long as they possess the appropriate access rights.

Figure 1 shows the user interface of the H.I.R.N. system, which is presented with the help of a commercial Web browser that supports the *Frame* feature, i.e., the presentation of hypertext pages in multiple rectangular display areas referred to as frames. The upper frame contains buttons for the activation of generic functions and an exact identification of the displayed document section. The content of the current section is shown in the lower frame where it can be scrolled up and down.

3 H.I.R.N. from the Author's Point of View

Figure 3 shows the functions of H.I.R.N. from the author's point of view. Here, the term "author" stands for the person who edits and enters a new regulation text into the system.

The most favored representation for the texts to be entered into the H.I.R.N. system is a Microsoft Word for Windows file. Prior conversion is necessary if a regulation cannot be delivered in this form. In case the original document exists in printed form only, it must first be scanned and converted into a textual representation with the help of an Optical Character Recognition (OCR) software.

The Microsoft Word text processor is the interface for the subsequent steps of structuring the environmental protection regulations. The information collected in

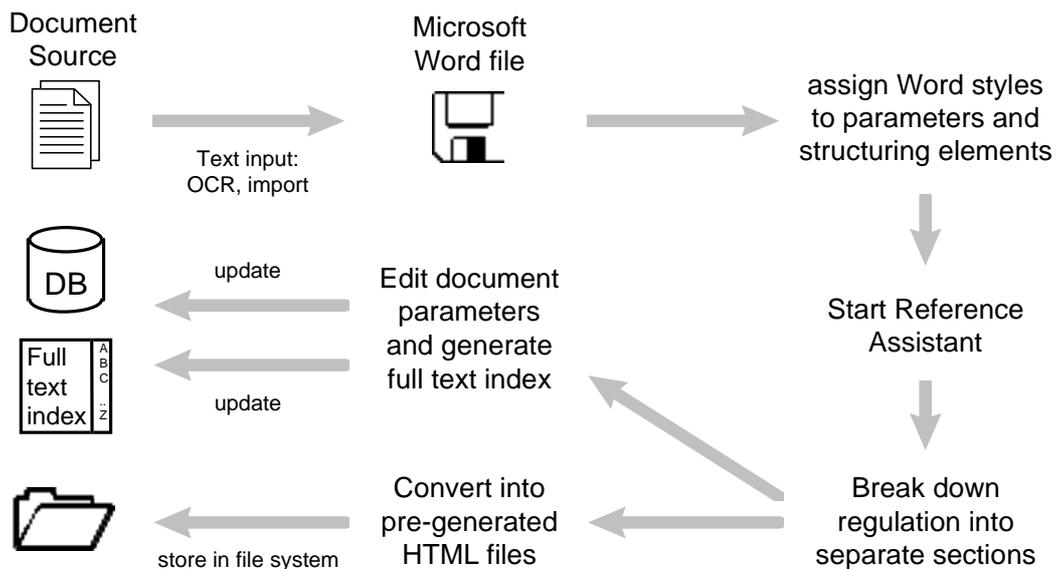


Figure 3: H.I.R.N. from the author's point of view



Figure 4: The H.I.R.N. Reference Assistant

these steps is used for both generating the hypertext structure of the document and deriving various document parameters to be stored in the H.I.R.N. database.

By assigning specific Word styles to the text components of a regulation, the author identifies parameters and structuring elements of the document such as the regulation title, the revision date, the document type and the legislative body issuing the regulation, as well as the hierarchy of the parts and paragraphs of the regulation. In addition, we developed a Word extension referred to as *Reference Assistant* that supports the author in detecting and defining cross references to other regulation sections in a semi-automatic way (figure 4). For this purpose, the Reference Assistant searches for characteristic patterns such as paragraph (“§”) signs and other syntactic constructs. For every potential cross reference, the system recommends the most likely destination in the same or another regulation. The author can accept, cancel or override these suggestions. The cross references found will later be displayed as hyperlinks to the end user of the system (figure 1).

Corresponding to the structure of the document, the text of the regulation is then broken down into separate sections. The individual sections of the environmental protection regulation are converted into partially pre-generated HTML⁴ files and stored in the file system. Together with the information in the database

⁴ The *Hypertext Markup Language* HTML is used to define hypertext pages in the World-Wide Web.

H.I.R.N. - Dokumentbeschreibung

 **Mercedes-Benz**

Verteilungskategorie: Lokal Global

Autor: Dr. K. Heldt

Kurztitel: Kreislaufwirtschafts- und Abfallgesetz

Abkürzung: [KrW-/AbfG]

Titel 1: Gesetz zur Vermeidung, Verwertung und Beseitigung von Abfällen

Titel 2: Gesetz zur Förderung der Kreislaufwirtschaft und Sicherung der umweltverträglichen Beseitigung von Abfällen

Letzte Änderung: 11. 11. 1996 Stand: vom 27. September 1994

Gültigkeitsstatus: in Kraft

Dokumenttyp: Gesetz

Urheber: Bundesrepublik Deutschland (BRD)

Rechtsgebiet: Abfall

Figure 5: Entering document descriptions into the system. The fields in this form contain document parameters such as title information, the revision date, the document type, the legislative body, etc. The content of most of these fields is derived automatically from the Word file representing the regulation in question.

(see next paragraph), these files are for use later by the retrieval component for generating a visual presentation of the regulation (as shown in figure 1).

Finally, the database and the full text index have to be updated. For this purpose, H.I.R.N. generates a form containing the document parameters which have already been identified by the author in the Word document source (figure 5). After having been completed and confirmed by the author, the parameters are stored in the database together with additional system-internal information about the individual document sections and their hierarchical structure. Additional descriptors for the individual document sections are generated by the *SWISH* full text indexing tool (Hughes 1995).

Regulations can deliberately be exchanged between H.I.R.N. systems. An export facility allows to select the regulations to be transferred and generates a self-contained installation file for each selected regulation. Installation files can easily be sent to the desired destination via email or diskette. At the destination site, they can be unpacked and installed by an import facility. The implemented installation concept is very modular and even allows regulations originating from multiple authoring sites to coexist at a single installation site.

4 Summary and Further Outlook

It has turned out that the approach used for H.I.R.N. is very general and goes well beyond the original requirements of an information system for environmental protection regulations to be used by Daimler-Benz. Particularly, further types of documents, such as standards or patent information can be added to the system. We also expect a growing user community in the near future, since a huge part of the environmental protection regulations integrated in the system is not only relevant to Daimler-Benz but also to subcontractors and other car manufacturers. Since Daimler-Benz is an international company, there are also plans to extend the H.I.R.N. system by multilingual components. This would include a multilingual user interface and a multilingual thesaurus, thus allowing a keyword-based document search in multiple languages. Such a broadening of the H.I.R.N. user group would facilitate the amortization of the expenditure for incorporating and updating the regulation texts. Here, H.I.R.N. could develop into an automated service center, offering information and services on the growing electronic market-place for multimedia contents.

References

- Hughes, K. (1995): SWISH – Simple Web Indexing System for Humans. Enterprise Information Technologies (EIT), Menlo Park, California.
URL: <http://www.eit.com/software/swish/swish.html>
- Riekert, W.F. / Kadric, L.E. (1996): HyperLex: Ein hypertextbasiertes Dokumentations- und Informationssystem für Umweltschutzvorschriften. In: Hilty, L.M. / Rautenstrauch, C. / Schoop, E. / Schraml, T. (eds.): Prozeßorientierte Dokumentation in Betrieblichen Umweltinformationssystemen. Metropolis-Verlag, Marburg.

Acknowledgement

The authors are grateful to Dr. D. Greiner and Dr. K. Heldt of Daimler-Benz AG for their dedicated support of the work presented here and for their valuable suggestions during the composition of this paper.