Knowledge-Based Meta-Data Validation:
Analyzing a Web-Based Information System

Frank van Harmelen

AI Department, Free University of Amsterdam

frankh@cs.vu.nl

Arjohn Kampman

Aidministrator, Amersfoort

akam@aidministrator.nl

Heiner Stuckenschmidt, Thomas Vögele

Intelligent Systems Group, Center for Computing Technologies, University of Bremen

{heiner, vogele}@tzi.de

Abstract

Web-based information systems play an important role in today’s practice of data-analysis and reporting in the field of environmental protection. While these systems solved some technical problems concerning the integration and visualization of information they show some problems even harder to handle. These problems are concerned with content and organization of the information within the information system and arise as a result of de-centralized authoring and provision of information. We present an AI approach for analyzing and structuring web-based information systems in order to check validity and consistency of information and to provide a content driven navigation structure. We show the usefulness of the approach by applying it to an existing environment information system that we analyze and structure according to different thematic categories. We also discuss further potentials of knowledge-based approaches to the problem of housekeeping in web-based information systems.

Introduction

Detailed, machine-readable information about the state of the environment and the degree of pollution caused by human activities have become a crucial factor in today's environmental protection. This holds for administration as well as for commercial organizations. The reasons for this are two-fold:

Environmental Legislation: The legislation in many European countries now demands for detailed reports on the execution of commercial activities on the environment. Further, the administration is obliged to provide information about the state of the nature to all citizen of a country. In order to fulfill this duty, these organizations have to collect, aggregate and visualize large amounts of chemical, biological and geographic data.

Preventative Protection: Efficient protection of the environment has to start before serious damages that often take years to recover from have taken place. The main problem of a precautionary protection lies in the perception of progressing damage that is often not recognized until it is too late. Therefore, preventative protection demands for constant monitoring and assessment of air, water and soil with respect to critical impacts.

For both purposes, information technology has proven useful in terms of automated acquisition of data using sensor networks, storage of the data in specialized databases and quick access to stored data through the built-in access mechanisms of the databases used. A problem that remained unsolved for a long time is an integrated access to different information sources. In practice, the task of environmental monitoring is most often split up among several organizational units each performing their own measurement programs whose results are stored in stand-alone databases, thus only shifting the problem of providing an overall picture of the situation. The advent of web-based information systems came with an attractive solution to the problem of providing integrated access to environmental information according to the duties and needs of modern environmental protection. Many information systems were set up either on the Internet in order to provide access to environmental information for everybody or in intra-nets to support monitoring, assessment and exchange of information within an organization. One of the most recent developments in Germany is BUISY, an environmental information system for the city of Bremen that has been developed by the Center for Computing Technology of the University of Bremen in cooperation with the public authorities. The development of the system was aimed at providing unified access to the information existing in the different organizational units for internal use as well as for the publication of approved information on the internet. While the use of web-based information systems solves the integration problem on a technical level, new problems arise that are even harder to tackle. These problems mainly concern content and organization of the information in the system. Some problems that were recognized after the BUISY system had been successfully installed are the following:

1. How can we guarantee that the information from the different sources is consistent and up-to-date?

2. How can de-centrally authored information be inserted at the 'right position' in the system?

3. How can we provide adequate access to the information to different user groups (especially internet vs. intranet users)?

A basis for a solution to these problems had already been laid in the development phase of the system by assigning a basic set of meta-information to each page in the system. In this paper, we present an approach for intelligent analysis and structuring of web-based information systems, based on document- structure, content and meta-data. The approach is based on the classification of web pages according to rules that can be specified by the web-administrator using the WebMaster system that has been developed by the Dutch company Aidministrator (www.aidministrator.nl). We discuss the general approach and show how the above-mentioned problems can be addressed in the WebMaster framework. The paper is organized as follows: We brief review the meta-data in the BUISY system and its current use. We describe how the same meta-data can be used to support the analysis and structuring of the system using WebMaster in two steps: We describe the WebMaster system and present the results of some experiments we carried out in applying the system to the BUISY (www.umwelt.bremen.de) website. We summarize with some lessons learned from the experiments and point towards further potential applications of AI techniques to the problem of housekeeping for web-based information systems.

Meta-Data in the BUISY System

Meta-data play an important role in the BUISY system. They control the access to individual web pages or "data-objects", as well as the location of individual objects in the BUISY information space. The BUISY information space is a multi-dimensional space, where "dimensions" are defined by the various classification principles that can be used to organize the data-objects (Visser et al., 2000). Typical dimensions are for example the „geographic“ dimension, describing the data-object's relevance with respect to a location in space, or the „organizational“ dimension, describing it's relevance with respect to the structure of an organization. Each data object in BUISY holds a set of meta-data that reflects the dimensions of the BUISY information space.

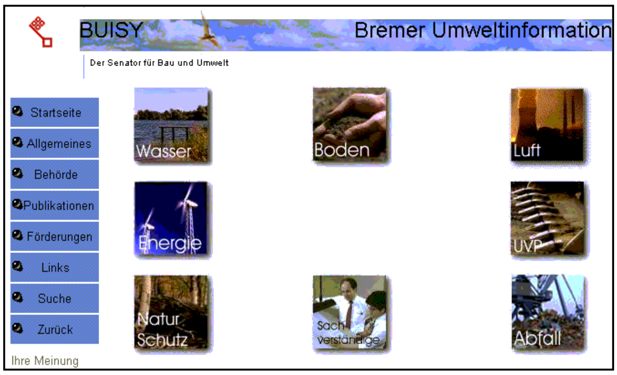


Figure 1: The BUISY main page with links to its subsystems

Conceptually, the basic BUISY architecture is similar to that of a distributed, object-oriented database, where each "information unit " is implemented as a stand-alone data-object. Such data-objects may consist of raw data (i.e. text, images etc.), methods working on these data (i.e. JavaScript code, ASP-code, ActiveX elements, etc.), and meta-data describing the object. Because the BUISY data-objects were implemented as HTML-pages, standard HTML META-tags in the document header could be used to annotate the meta-data. However, the system could easily be extended to support other formats as well: In XML, for example, meta-data are described in the respective DTD and implemented in the XML-document as tags and attributes (W3C, 1999). Even for most MS-Office formats, it is possible to attach meta-data to data-files.

The meta-data are parsed and evaluated by a data-broker which works like a search engine that automatically (and frequently) scans all data-objects, collects the meta-data and stores them in an index table. Through this dynamic meta-data index table, the data-broker controls the selection, formatting and display of all data-objects in a given data space.

Technically, the system was implemented on a Windows NT platform. This was necessary to make the system consistent with the existing hard- and software infrastructure at the Senate of Bremen, and to be able to leverage existing IT know-how in the organization. BUISY therefore uses standard Microsoft tools such as MS Internet Information Server (MS IIS), MS IndexServer, and ASP. The broker architecture was implemented as a combination of the MS IndexServer and custom-built ASP-code.

Meta-Data Annotations

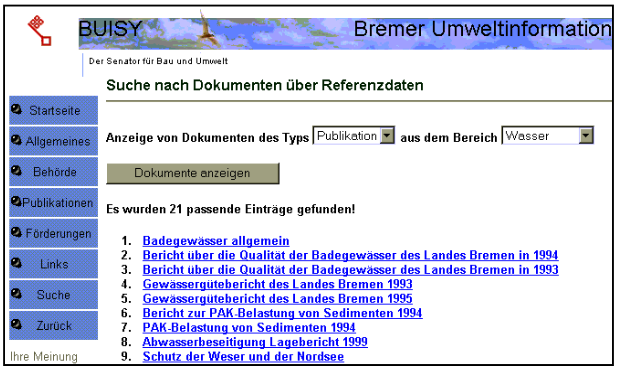
The current version of BUISY supports four dimensions that describe each data-object's location within a geographic space (meta tags "Ort" and "RechtsHoch"), a thematic space (meta tab "Bereich"), an organizational space (meta tag "Organization"), and a legal framework (meta tag "Rechtsbezug"). A set of additional meta tags annotates information about the data-object's type, author, creation- and expiration dates, and relevant keywords. The "Status" meta-tag indicates whether the data-object is part of the Internet or (restricted) Intranet section of BUISY.

A typical BUISY metadata will therefore look like the following:

```
<meta name="Status"
content="Freigegeben">
<meta name="Typ" content="Publikation">
<meta name="Author" content="TJV">
<meta name="Date" content="10-04-1999">
<meta name="Expires" content="31-12-
2010">
<meta name="Keywords" content="Wasser,
Gewässergüte, Algen">
<meta name="Bereich" content="Wasser">
<meta name="Rechtshoch"
content="3487020,5885240">
<meta name="Ort" content="Finndorf">
<meta name="Rechtsbezug"
content="VaWS">
<meta name="Organisation"
content="Abt.6">
```

Using Meta-Data for Document Search

The combination of meta-data with a broker architecture opens the door to sophisticated and intelligent ways of information retrieval, for example through the application of knowledge-based methods, such as ontology type meta-data combined with logic reasoning (Fensel et al., 1998). The term ontology was originally used in philosophy to describe a theory of "being and existence". In the area of artificial intelligence it was adopted to describe knowledge models that provide definitions of vocabulary used to describe a certain domain (Gruber, 1992). In combination with a logic inference machine, ontologies can be used to transform implicit knowledge hidden in the data-objects into explicit knowledge available to the user of an information site.



Suche nach Dokumenten über Referenzdaten

Anzeige von Dokumenten des Typs Publikation aus dem Bereich Wasser

Es wurden 21 passende Einträge gefunden!

1. Badegewässer allgemein
2. Bericht über die Qualität der Badegewässer des Landes Bremen in 1994
3. Bericht über die Qualität der Badegewässer des Landes Bremen in 1993
4. Gewässergütebericht des Landes Bremen 1993
5. Gewässergütebericht des Landes Bremen 1995
6. Bericht zur PAK-Belastung von Sedimenten 1994
7. PAK-Belastung von Sedimenten 1994
8. Abwasserbeseitigung Lagebericht 1999
9. Schutz der Weser und der Nordsee

Figure 2: Meta-Data driven document search

Such an "intelligent" information site would for example "know", through rules and definitions provided in a formal ontology, that agricultural production is frequently linked to nitrate contamination of ground water. A search for the keyword "nitrate pollution" in a specific area would then not only produce reports on nitrate measurements in ground water of that area (i.e. documents that contain the specified keyword), but also background information about the recent increase of agricultural activities nearby areas (i.e. related information that doesn't contain the keyword or synonyms of the keyword). With the help of the geographic meta-data and spatial reasoning, all relevant data-objects within a region of interest could be selected.

Content-driven classification of Web pages

As mentioned in the introductory section, both maintenance and navigation of the contents of web sites are important open problems, in particular for large web sites that contain material originating from different sources.

Navigation: Easy accessibility of huge amounts of information (on Intranets or Internet sites) becomes ever more important. High demands are made upon disclosure of such information: not only must the information be always up-to-date and available, but it must also be organised in a meaningful way and be easily searchable. A related problem is that different persons have different information needs and therefore demand different kinds of classification and navigation structures. Existing search- and navigation-tools do not satisfy these high demands. Search-tools are usually keyword based, resulting in low precision and recall. The navigation structure is limited to hand-made menu's and index pages that must be used by everyone. One additional disadvantage is that this way of structuring quickly ages and therefore requires a lot of maintenance. Visual aids are rarely exploited.

Maintenance: Maintaining the content of Web sites is an open and urgent problem on the current World Wide Web. Anybody who has used the WWW has experienced the amounts of outdated, missing, and inconsistent information on many Web sites, even on those sites that are of crucial importance to individuals, companies or organisations. Websites are large, frequently updated, and constructed by multiple authors. All this makes it impossible to do manual maintenance on contents of Websites. Although many current tools deal with problems such as broken links and missing images, very few solutions exist for maintaining the contents of Web sites.

WebMaster is a software tool developed by the Dutch company AIdministrator that aims at solving these problems. This is done in a four-step process:

Step 1. Constructing a classification of Web-pages

The builder or administrator of a site defines a class-hierarchy of the different types of pages that appears on the web site. For example, pages can be about water, soil, air, energy, etc. Each of these types of pages can again be subdivided into new subclasses: water-pages can be about wastewater, drinking water, river-drainage, etc. This leads to a hierarchy of pages that is based on page-*contents*, such as the example shown in Figure 5.

Step 2. Defining the classification criteria.

For each of the classes defined in step 1, the user defines a rule that determines which Web pages will be members of that class. These rules can exploit the entire structure and contents of the specific pages. For example, the rule in Figure 3 specifies that a rule is about water if the keyword "wasser" appears in the meta-information of the web-page.

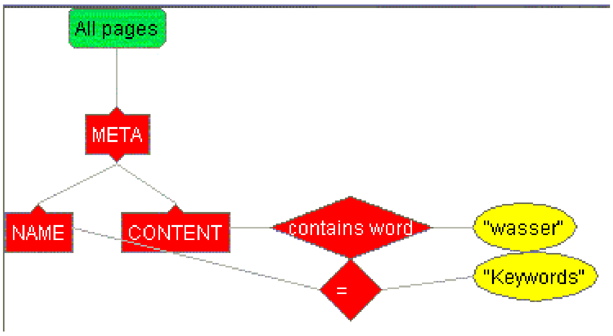


Figure 3: Rule for pages with the keyword "Wasser"

But the classification rules need not be limited to meta-data. Figure 4 shows a rule which states that a page is about water if it contains a link to a page whose URL conatins the string "/buisy/wasser".

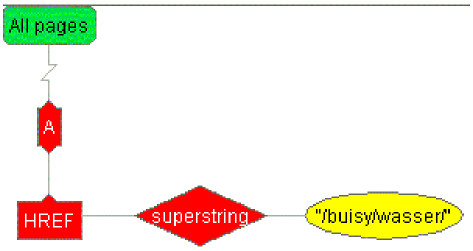


Figure 4: Rule for pages stored in the directory "/buisy/wasser/"

In the typical case, a page belongs to a class if the rule that is defined for that class succeeds for the page. However, it is also possible to define classes by negation: a page belongs to a class when the corresponding rule fails on that page. This is indicated by a rectangle in the class-hierarchy (instead of a rounded box). In Figure 5, the class META-tags will contain all pages that do not contain a <META>-tag.

The rule-format of WebMaster allows general conditions to be imposed on the contents and structure of a Web-page, visualised to emphasise the required nesting of text, tags and attributes that must appear in a page before it is classified as belonging to a certain type. We refer to (van Harmelen & van der Meer 1999) for a more detailed discussion of the rule format and the corresponding visualisation.

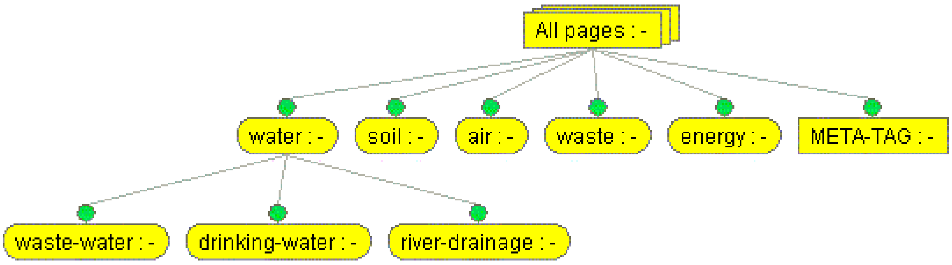


Figure 5: A Classification Tree from the water domain

Step 3. Classifying individual pages

Whereas the human user of the WebMaster system performs the previous steps, the next step is automatic. The definition of the hierarchy in step 1 and the rules in step 2 allows the WebMaster inference engine to automatically classify each page in the class hierarchy. Notice that classes may overlap (a single page may belong to multiple classes).

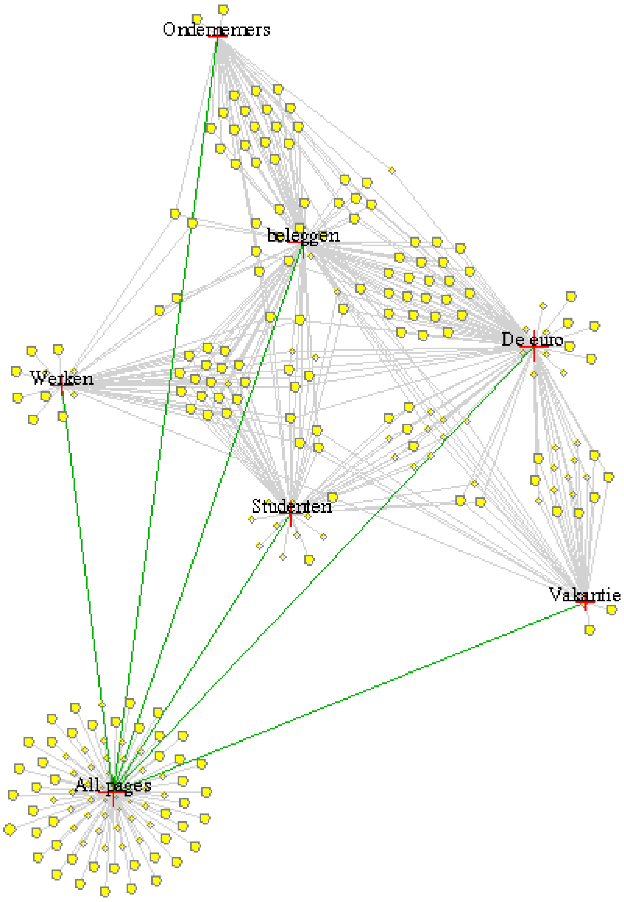


Figure 6: Example of a content map

The rule format has been defined in such a way as to provide sufficient expressive power while still making it possible to perform such classification inference on large numbers of pages (many thousands in human acceptable response time).

After these three steps, we have a class hierarchy that is populated with all the pages of a given site. This finally puts us in a position to tackle the problems of verification and navigation:

Step 4a. Detecting errors in web pages

A very effective strategy to detect errors in the contents of Web pages is to use step 2 to define integrity constraints on web pages (i.e. rules expressing properties that must hold for all pages or a subset of pages), and then use the rule-negation mechanism to find all pages that violate such a constraint. A simple example is the class from Figure 5 defined by all pages that do not contain a <META>-tag. Other examples will follow in the next section.

Step 4b. Visualising site-contents

Besides detecting the errors, the populated class hierarchy of pages can also be used as the basis for navigation, namely through so called semantic site-maps. Images such as Figure 6 are automatically constructed from a populated type hierarchy: clusters of nodes in the figure correspond to types from the hierarchy, and can be used to navigate the site based on contents. Furthermore, pages that belong to more than one class are located in between the classes to which they belong. As a result, classes that have many overlapping pages (and which are therefore apparently semantically close) also appear visually close in the site-map.

Automatically constructed figures such as Figure 6 are compact enough to display many hundreds of pages in a single small image (Figure 6 maps around 250 pages).

Validation of Meta-Data

In section two we described the use of meta-data in the BUISY system. This approach heavily depends on the completeness and consistency of the meta-data itself, as well as its relation with the actual content of the page. In this section, we describe how the WebMaster system can be used to check the meta-data in the present system. We start with the selection of pages that actually contain environmental information and proceed by checking the existence of the META attributes and values categorizing pages according to subsystems. We also check the keyword annotations in the water-subsystem by combined meta-data and full-text search.

Pre-Selection of Relevant Pages

Before we can start to analyze the meta-annotations of the content pages in the BUISY system, we have to sort out those pages that are not meant to host information. These pages are, for example, pages defining frame-sets or customized navigation bars. We can easily sort out these pages by defining a class with a rule that checks all pages for the html-tag <FRAMESET>. We found that more than half of the approximately one thousand five hundred pages of the BUISY website, namely 870, fall into this category and are therefore irrelevant for our analysis. In order to exclude them from the further analysis, we defined a corresponding constraint class containing all pages that do not define frame-sets.

In a similar way, it is possible to identify and exclude the pages containing the navigation bars for the different sub-systems. The identifying property of these pages is that they all call the same JAVA script (menu.js). Using WebMaster this can also be checked by defining a rule that claims the existence of <SCRIPT SRC =“menu.js”>.

Checking Meta-Attributes and Values

After we have extracted the pages that are actually supposed to contain information, we can start to check the completeness of the annotated meta-information. In our analysis, we focused on the meta-information that assigns a page to a certain topic area. In the BUISY system this information is stored in the meta-attribute named “Bereich”. So the first task is to check whether all pages that passed the pre-selection contain the meta-attribute “Bereich”. The result of this test was rather negative. We found that about one hundred of the six hundred fifty content pages do not contain the “Bereich” attribute. Another three pages did contain the attribute but without a value. The reason for this is manifold. A first reason could be that we missed some pages in the pre-selection. Another reason could be found in the integration of the different heterogeneous subsystems. It is very likely that not all pages that have been included into the BUISY system are annotated yet. However, using the WebMaster system, we are able to find these pages and to decide whether meta-data has to be added or not.

After dropping the pages without usable meta-information, we got a core set of content pages annotated with appropriate meta-information about the topic area. In a next step, we classified these remaining pages according to their content area using the value of the META-tag. The result of the classification shown in Figure 7 reveals that the topic area “water” contains by far the most pages. This is not astonishing as the BUISY System was built on top of the water quality information system EISA that therefore constitutes the most elaborated part of the system. We can also see that the topic areas “soil”, “air, “waste” and “nature” also contain a reasonable amount of pages, whereas the topic areas impact assessment with nine and energy with zero pages are very likely to be the place to look for missing meta-information that disabled the WebMaster to find the corresponding pages.

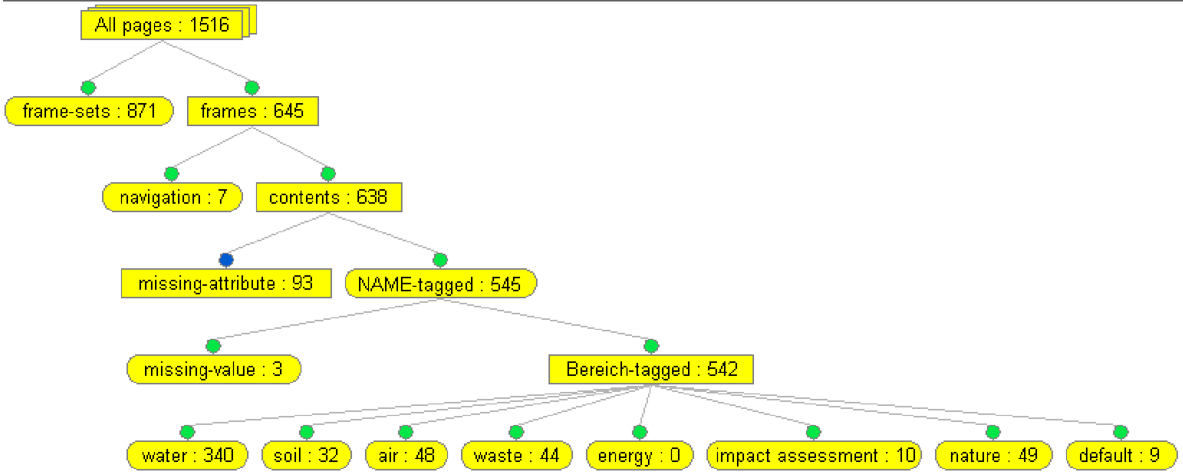


Figure 7: Classification tree used for the validation of meta-data

Check for Missing Keywords

The keyword annotations designated to be part of every content page are an important tool for finding relevant information within one of the topic areas of the BUISY system. Therefore, the validation of the keyword annotations that actually exists in the system is the next step of our analysis. In order to judge the quality of the present annotations we defined some keywords covering important aspects of the information found in the system. These keywords included the names of rivers, different categories of water (e.g. groundwater, watercourses or waste water). We used these keywords to compare the keyword annotations with the content of the page using a full text search on the whole page. Figure 8 shows a corresponding class definition rule.

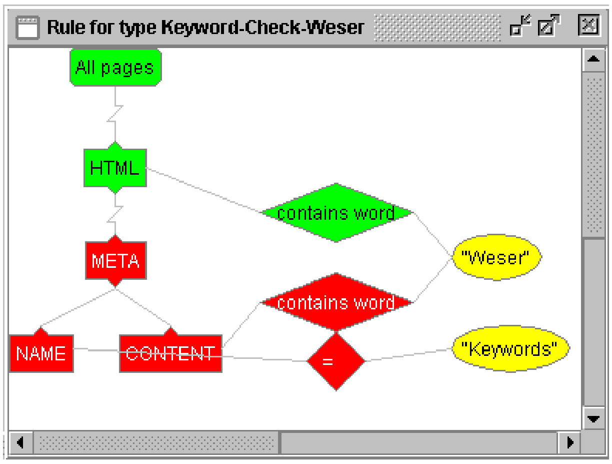


Figure 8: Classification rule for keyword validation

The rule states that if the web page contains the word “Weser” (the main river in Bremen) then there has to be a meta tag where the value of the NAME attribute equals “Keywords” and the value of the CONTENT attribute contains the word “Weser”.

The results of the keyword validation were even worse than the ones reported before. It turned out that there is a complete lack of a common keyword vocabulary or standard. Most pages just contain, if any, the nouns that appear in their title as keywords. As a consequence, the keywords do not provide more information than the page titles. Consequently, no keyword-based search is possible at the moment. The search mechanisms provided by the BUISY System completely rely on full text search in the page content.

Generation of a Content Map

After the analysis of the meta-information, we tried to generate a content map of the BUISY system. However, it turned out that the classes defined in the first step are not suitable for this purpose. Up to now we mainly classified the pages according to the topic area they belong to. This property is defined by a single meta-attribute and therefore produces disjoint classes. When producing a content map, however, the connections between different classes are of special interest. A normal procedure would be to classify the pages according to their keyword annotations and use the content map in order to visualize the position of the pages relative to these keywords. As described in the last section, this was also not possible due to the absence of suitable keyword annotations. Therefore, we decided to classify the content pages using the result of the full text search performed in order to validate the keyword annotations. In the following we briefly describe the classes defined and discuss the resulting content-maps shown in Figure 9 and 10.

Defining Classes

We defined some keywords for the topic area ‘water’ that are expected to appear often in the content of the pages belonging to that area. We defined a class for every keyword using a rule claiming that the word appears somewhere in the text body of the page. We chose the following keywords that describe different areas of interest.

- Gewässer: Watercourses
- Weser: A river in Bremen
- Grundwasser: Groundwater
- Abwasser: Wastewater
- Anlagen: Technical Installations

Further, we included some words corresponding to types of documents that might appear in the system. We chose the following.

- Berichte: Reports
- Verordnungen: Legislations

Other types of information appearing frequently could easily extend this list. Good examples are online-databases requests and forms used for the interaction of the user with the system. In the present state of the system, this does not make much sense, but in the course of further developments other kinds of documents will become more important and have to be included in a content map.

Results of the Map Generation

The maps generated from the classes described above (Figure 9 and 10) show some interesting features. The first thing that attracts attention is the fact that again most of the pages could not be classified into one of the keyword classes. The better part of the approximately one thousand pages analyzed do not even contain information about the topic area water. This can be explained by the fact that a content map always contains all pages of a web site. However, there are also many pages that contain relevant content, but do not belong to one of the keyword classes (page cluster at the right hand side of the page). The interesting part of the content map is its left side where the pages from the different keyword classes and their memberships in these classes are displayed. Even a quick look at the map reveals a significant difference between the bottom of the left side where the classes ‘Bericht’, ‘Grundwasser’, ‘Gewässer’ and ‘Weser’ are located and the top of the left side with the classes ‘Abwasser’, ‘Verordnungen’ and ‘Anlagen’. While the former is sparsely populated and does not show much regularities, despite an overlap of the classes ‘Weser’ and ‘Gewässer’ the latter part reveals clear structures: We can clearly identify pages about technical facilities and waste water as well as pages containing information about legislations on one or both of these topics. If, on the other hand, we try to find a page on watercourses the map does not provide much help. This result underlines the need for an improvement of suitable keyword annotations that could be used to search for pages effectively.

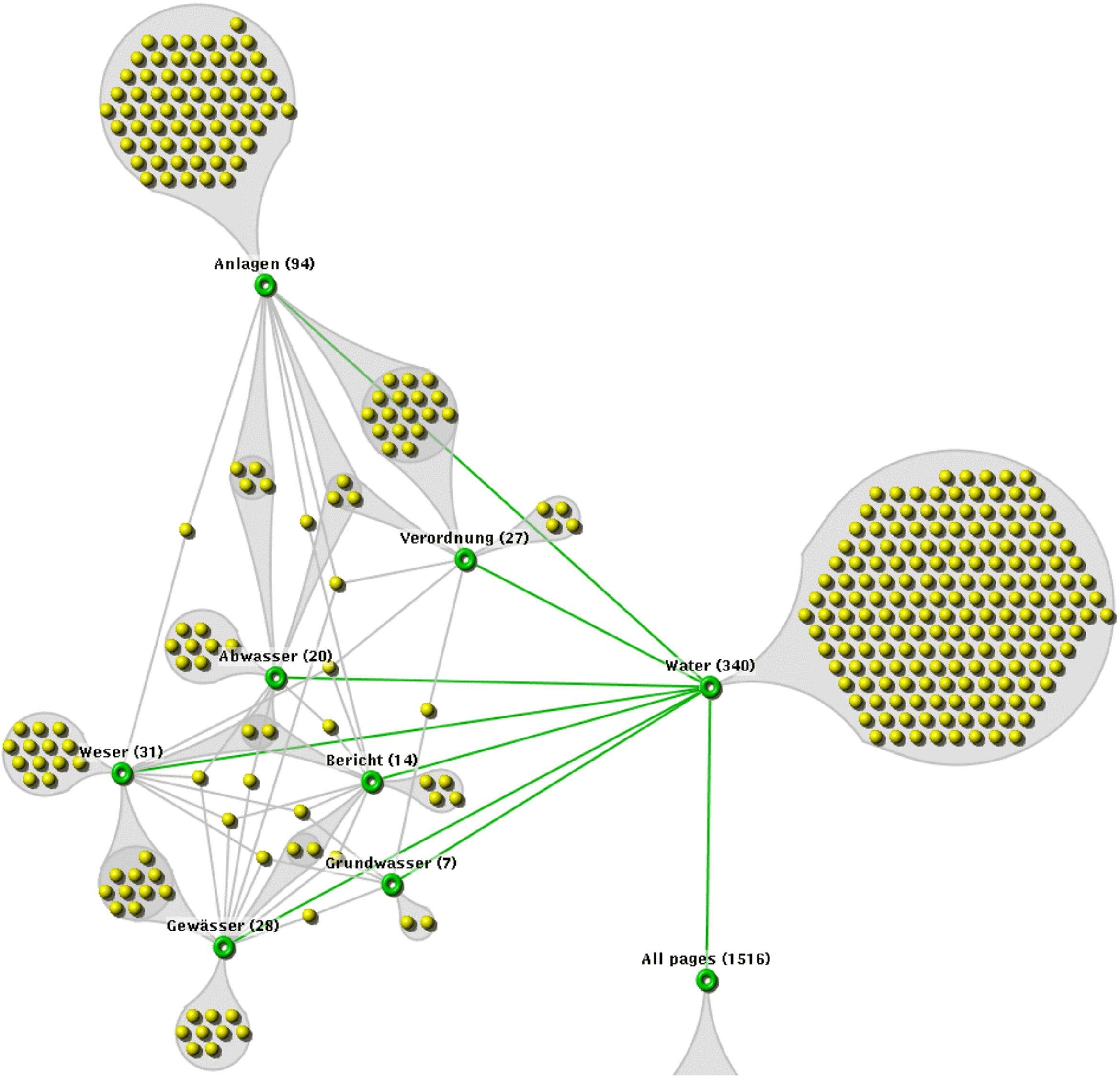


Figure 9: Cluster Map of the Water Subsystem

Discussion

In this paper we presented the meta-data concept of the environmental information system BUISY and used WebMaster, a knowledge-based analysis tools for weakly structured information in order to evaluate the actual implementation of the concept. We found several shortcomings in the current system. The inconsistent use of keyword annotations was identified to be one of most striking problems. Much of the aspired functionality of the system (Stuckenschmidt & Ranze 1999) relies on a consistent use of these annotations. The Implications of this result is two-fold. On the one hand we got some evidence that the classification approach used in WebMaster could be used as a way to implement content-driven search and context dependent navigation structures. On the other hand, it became clear that much work has to be done on the BUISY system in order to lay a foundation for such an implementation. One of the most urgent topics is to define a standard keyword vocabulary to be used as meta-information. Another topic is the automatic annotation of existing and new pages using this vocabulary. This could again be done using the WebMaster classification approach. A content-driven access to the information in the system could be built on top of the annotated pages.

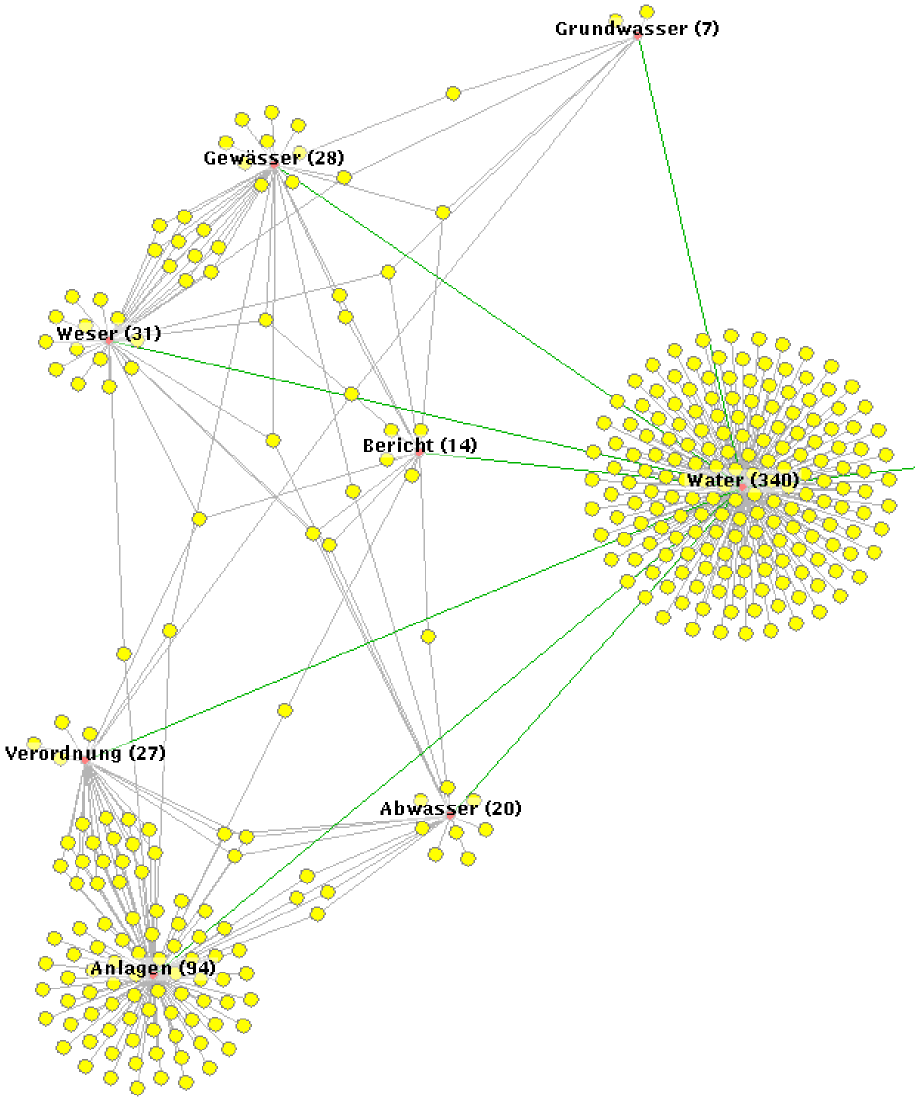


Figure 10: Site-Map of the Water SubSystem

References

Dieter Fensel,, Stefan Decker, Michael Erdmann., and Rudi Studer. (1998): Ontobroker - The Very High Idea, FLAIRS-98 11th International Conference, Sanibal Island, USA

Thomas Gruber. (1992) Ontolingua: A Mechanism to Support Portable Ontologies, technical Report KSL-91-66 Computer Science Department, Stanford University.

Frank van Harmelen and Jos van der Meer (1999) *WebMaster: Knowledge-based Verification of Web-pages.* Proceedings of the Twelfth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, (IEA/AEI'99), M. Ali and I. Imam (eds.), Springer Verlang, LNAI, 1999.

Heiner Stuckenschmidt, K. Christoph Ranze (1999). *Intelligenter Zugang zu Umweltinformationen durch Ontologie-basiertes Information-Retrieval.* In Christian Dade und Bernhard Schulz (Hrsg.): Management von Umweltinformationen in vernetzten Umgebungen Metropolis Verlag, Marburg, 1999

Ubbo Visser, Heiner Stuckenschmidt, Gerhard Schuster & Thomas Vögele (2000): Ontologies for Geographic Information Integration, Computers & Geosciences, submitted.

Thomas Vögele, Heiner Stuckenschmidt, Ubbo Visser (2000) *BUISY - Using Brokered Data Objects for Environmental Information Systems.* Klaus Tochtermann, Wolf-Fritz Riekert (Hrsg.): Hypermedia im Umweltschutz Metropolis Verlag, Marburg 2000.

W3C (1999): Extensible Markup Language (XML) 1.0; W3C Recommendation 10. February 1998 (http://www.w3.org/TR/REC-xml).