Knowledge-Based Validation, Aggregation and Visualization of Meta-data: Analyzing a Web-Based Information System

Heiner Stuckenschmidt[1] and Frank van Harmelen[2,3]

[1]Center for Computing Technologies, University of Bremen
[2]AIdministrator BV, Amerfoort,
[3] AI Department, Vrije Universiteit Amsterdam

Abstract. As meta-data become of ever more importance to the Web, we will need to start managing such meta-data. We argue that there is a strong need for *meta-data validation and aggregation*. We introduce the WebMaster Approach for verifying semi-structured information and show how it can be used to validate, aggregate and visualize the Meta-Data of an existing Information System. We conclude that the possibility to verify and aggregate meta-data is an added value with respect to content based access to information.

1 Motivation: Meta-Data on the Web

The information society demands large-scale availability of data and information. With the advent of the World Wide Web huge amounts of information is available in principle, however size and the inherent heterogeneity of the Web makes it difficult to find and access useful information. A suitable information source must be located which contains the data needed for a given task. Once the information source has been found, access to the data therein has to be provided. A common approach to this problem is to provide so-called meta-data, i.e. data about the actual information. This data may cover very different aspects of information: technical data about storage facilities and access methods co-exist with content descriptions and information about intended uses, suitability and data quality. Concerning the problem of finding and accessing information, the role of meta-data is two-fold: On the side of information providers it serves as a means of organizing, maintaining and cataloguing data, on the side of the information users meta-data helps to find, access and interpret information. Recently, standards have been proposed that cover different aspects of meta-data, especially the syntax for coding, the model structure and content of a meta-data model. Some of these standards are:

- Syntactic Standards: HTML, XML, RDF (see http://www.w3c.org)
- Structural Standards: RDF schemas (see http://www.w3.org/TR/rdf-schema/), Topic Maps (see http://topicmaps.org/)

– Content Standards: Dublin Core (see http://dublincore.org/)

These standards mentioned provide good guidance to design and encode meta-data for information resources on the world-wide web. However, there are still some severe problems that are addressed neither by structural nor by content standards. These problems are concerned with the relation between information and meta-data about it. Some of the most important are:

Completeness: In order to provide full access to an information source, it has to be ensured that all the information is annotated with the corresponding meta-data. Otherwise, important or useful parts of an information source may be missed by meta-data driven search methods or cannot be indexed correctly.

Consistency: Meta-data about the contents of available information is only useful if it correctly describes this contents. In fact, meta-data that is not consistent with the actual information is an even bigger problem than missing meta-data, because mechanisms relying on meta-data will produce wrong results without warnings.

Accessibility: In order to be useful, meta-data has to be accessible not only to the information provider but especially for users that want to access it. Therefore, an important question is how a comprehensive description of an information source can be provided and accessed by potential users.

In this paper we describe a system for the validation of semi-structured information that can be used to check the completeness and consistency of meta-data with respect to the information it describes. We apply this approach to an existing information system and show how it can be used to generate and visualize an aggregated meta-data model for a large part of the information system in such a way that accessibility is improved.

2 BUISY: A Web-Based Environmental Information System

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 Technologies 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

Fig. 1. The Meta-Data Driven Document Search Facility

internal use as well as for the publication of approved information on the internet.

Meta-data plays an important role in the BUISY system. They control the access to individual web pages. Each page in the BUISY system holds a set of meta-data annotations that reflects its content and state [7]. The current version of BUISY supports a set of meta tags annotating information about the data-object's type, author, dates of creation- and expiration, and relevant keywords and topic area of the page. The "Status" meta-tag indicates whether the data-object is part of the Internet or (restricted) Intranet section of BUISY.

```
<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"/>
```

At the moment, this meta-data is used to provide an intelligent search facility for publications of the administration concerned with environmental protection. The user selects a document type and a topic area. Based on the input, a list of available publications is generated (see figure 1).

3 The WebMaster Approach

We have developed an approach to solve the problems of completeness, consistency and accessibility of meta-data that were identified above. This

is done on the basis of rules that must hold for the information found in the Web site, both the actual information and the meta-data (and possibly their relationship) [6]. This means that besides providing Web site contents and meta-data, an information provider also defines classification rules (also called: integrity constraints) that should hold on this information. An inference engine then applies these integrity constraints to identify the places in the Web site which violate these constraints. This approach has been implemented in the WebMaster system, developed by the Dutch company AIdministrator (www.aidministrator.nl). In this section, we will describe the different steps of our approach.

Step 1. Constructing a Web-site ontology The first step in our approach to content-based verification and visualisation of web-pages is to define an ontology of the contents of the web-site. Such an ontology identifies classes of objects on our web-site, and defines subclass relationships between these classes. For example, pages can be about water. These can again be subdivided into new subclasses: *Gewässer* (watercourses), *Weser* (A river in Bremen) *Grundwasser* (Groundwater) *Abwasser* (wastewater) and *Anlagen* (Technical Installations). Further, we included some classes corresponding to types of documents that might appear in the system. We chose *Berichte* (Reports) and *Verordnungen* (legislations). This leads to a hierarchy of pages that is based on page-contents, such as the example shown in Figure 2.

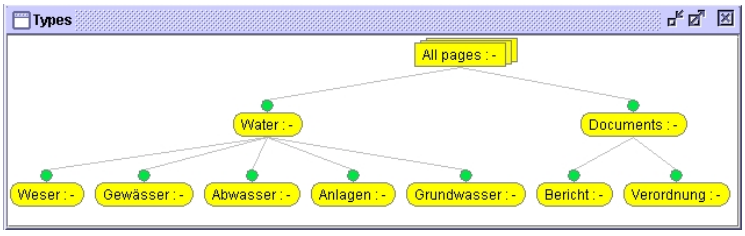

Fig. 2. An Example Classification Tree

A subtle point to emphasize is that the objects in this ontology are *objects in the web-site*, and not objects in the real-world that are described by the web-site. For example, the elements in the class "river-drainage" are not (denotations of) different river-drainage systems in the environment of Bremen, but they are *web-pages* (in this case: web-pages talking about river-drainage systems). As a result, any properties we can validate for these objects are properties of the *pages on the web-site*, as desired for our validation purposes.

Step 2. Defining the classification criteria for the ontology The first step only defines the classes of our ontology, but does not tell us which instances belong to which class. In the second step, the user defines rules that determine which Web pages will be members of which class. In this section, we will briefly illustrate these rules by means of three examples.

Figure 3 specifies that a rule is about "water" if the keyword "wasser" appears in the meta-information of the web-page. The rule succeeds if the following code appears in the web-page:

```
<meta name="Keywords" content="wasser">
<meta name="Typ" content="Bericht>
```

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 9 for example, the class 'missing-attributes' will contain all pages that do NOT contain the attribute NAME in the <META>-tag.

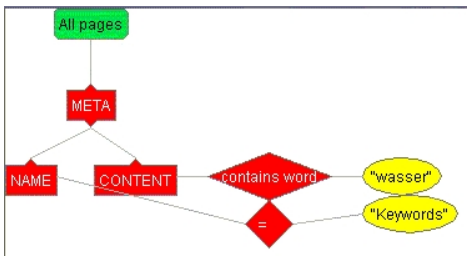

Fig. 3. Example of a Classification Rule Using Meta-Data

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). The rule format (adopted from [5]) 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.

4 Applying WebMaster to the BUISY System

The ability of the WebMaster System to classify web-pages according to the meta-data contained in every page enables us to use the system to perform the tasks we claimed to be necessary for meta-data management on the internet, i.e. the validation, aggregation and visualization of the meta-data annotations in the BUISY system. At that time the BUISY system contained approximately 1500 pages which are not maintained centrally, but the different topic areas of the systems had been supplemented by different persons after the initial development phase that ended in 1998. Due to this fact, we expected to be faced with incomplete and inconsistent meta-data annotations in the different parts of the system. We performed some validation and some aggregation experiments on this meta-data that are reported in the next sections.

4.1 Validating Meta-Data

Checking Meta-Attributes and Values After we 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. 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.

Check for Missing Keywords 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. We chose the keyword according to the classes described in step 1. We used the keywords to compare the keyword annotations with the contents of the page using a full text search on the whole page.

The validation revealed that most pages containing a keyword in the text did not have this keyword in the meta data annotation. Using WebMaster we were able to identify these pages and present them to the systems administrator who has to decide if the keyword has to be added.

4.2 Aggregating Meta-Data

The validation of meta-data discussed in the previous section is all done on the <META>-tags which are distributed across the 1500 pages of the BUISY system. At construction time, such a distributed organization of the meta-data is rather attractive: each page can be maintained separately, containing its own meta-data. Page-authors can directly update the meta-data annotations when updating a page, and no access to a central meta-data repository is needed. However, when we want to use the meta-data to create content-based navigation maps (as in the next section), or as the basis for a meta-data-based search engine, such a distributed organization of the meta-data is no longer attractive. We would then much rather have fast access to a central meta-data repository instead of having to make remote access to 1500 separate pages when looking for certain meta-data.

Using the validation process described in section 3 we analyzed the Web-site with respect to membership of pages to different topic areas. The result of this step is a classification of pages into a number of classes, based on the application of the classification rules to the <META>-tags in the pages. This yields a populated class-hierarchy of pages. Such a populated class hierarchy can be stored in a combined RDF and RDF Schema format [2]. The following statements are taken from the RDF Schema encoding of the Webmaster type hierarchy. The first three show how of the types "water", "Gewässer" and "Weser" and their subtype relationship are encoded in standard RDF Schema.

```
<rdfs:Class rdf:ID="water"/>

<rdfs:Class rdf:ID="Gewässer">
   <rdfs:subClassOf rdf:resource="#water"/>
</rdfs:Class>

<rdfs:Class rdf:ID="Weser">
   <rdfs:subClassOf rdf:resource="#water"/>
</rdfs:Class>

...
```

The following is an example of an RDF encoding of instance information: the URL mentioned in the "about" attribute is declared to be a member of the class "water" (and consequently of all its supertypes, by virtue of the RDF Schema semantics.

```
<rdf:Description
    about="http://www.umwelt.bremen.de/buisy/scripts/buisy.asp?
        doc=Badegewaesserguete+Bremen">
   <rdf:type resource="#Gewässer"/>
</rdf:Description> ...
```

These automatically generated annotations constitute an aggregated description of a web site that can be used to get an overview of its content. The annotations are machine-readable, but they are hard to use by a human WebMaster. This is the reason why we not only generate am aggregated meta-data model, but also provide a condensed visualization on the basis of the aggregated model. We will discuss this visualization, that is intended for human use in the next section.

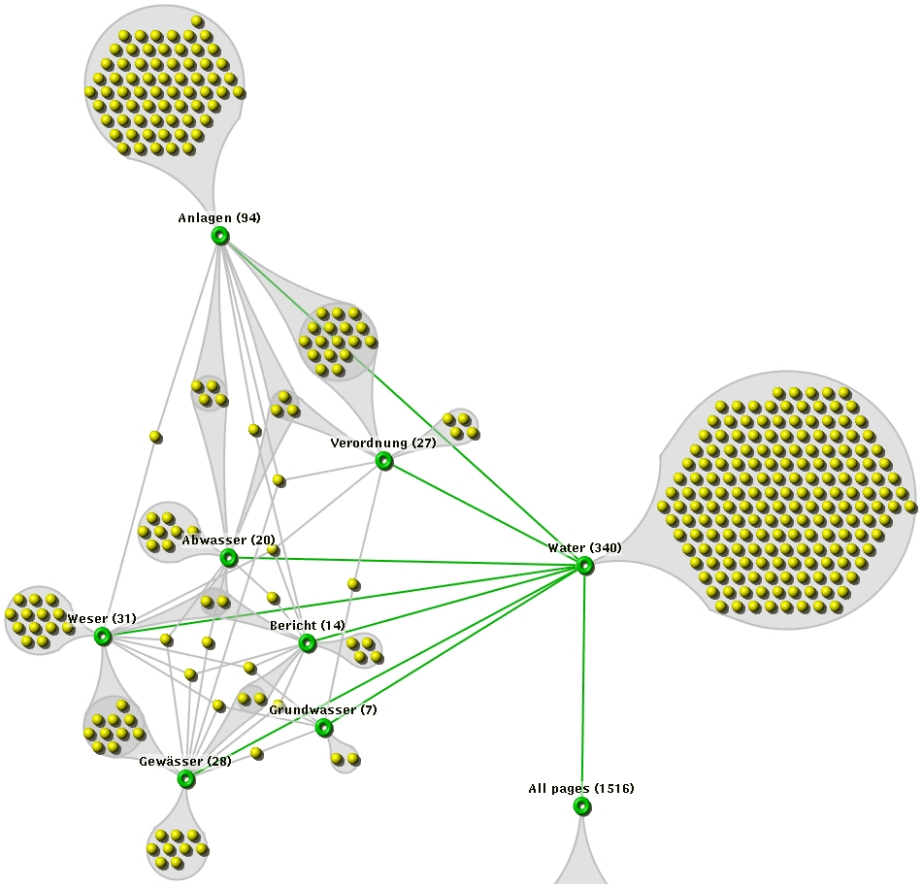

Fig. 4. Cluster Map of the Water Subsystem

4.3 Meta-Data Visualization

The WebMaster system supports the automatic generation of so-called cluster maps about a web-site. A cluster map visualizes an instantiated hierarchy of pages by grouping pages from the same class into a cluster. These clusters may overlap if pages belong to more than one class.

The map generated from the classes described above (figure 4) shows 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 contents, 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 membership in these classes are displayed. We can clearly identify pages about technical facilities and waste water as well as pages containing information about legislation concerning one or both of these topics.

Automatically constructed figures such as figure 7 are compact enough to display many hundreds of pages in a single small image (the map contains 340 pages). This should be compared with output from traditional search engines, where a set of more than 300 pages is typically presented as 15 pages with 25 URLS each. The format of figure 4 is much more useable in practice.

5 Discussion

We argued that meta-data plays an important role in information management on the Internet and mentioned existing problems. We identified the need for validation, aggregation and visualization of meta-data and presented a knowledge-based approach to these tasks. We introduced the WebMaster System which implements the approach and presented some results in applying it to a web-based environmental information system. Two mayor implications are drawn from our experiments.

Meta-Data Validation is Necessary and Possible While meta-data standards cover many questions of what kinds of meta-data to use and how to represent it, guaranteeing completeness, consistency and accessability of meta-data is still a problem in web-age information management. There is a need for methods to check information sources for the existence of meta-data and to relate it to the actual content of the information source.

Our experiments also showed that the WebMaster system implements a promising approach to meta-data management. It enables us to perform completeness, consistency and plausibility checks on meta-data. We can locate pages with missing meta-data, compare information contents and meta-data and produce hints towards missing keywords. The graphical interface of the systems supports the inspection of large information systems, the aggregation of meta-data annotations, and the visualization of the aggregated model.

Aggregated Meta-Data is an Added Value The ability to create an aggregated content model of an information system and store it in RDF format provides opportunities that go far beyond an inspection of existing meta-data. The aggregated model can serve as intelligent interface for an information source. This interface can be used by next generation search engines to get a quick overview of the contents in order to decide whether a detailed search is promising or not. Appropriate inference services like the one reported in [3] can be used to answer queries about the contents of an information source on the basis of the meta-data model.This ability could even be enhanced by using more expressive languages for the representation of aggregated meta-data. A promising language for this purpose is described in [4]. The idea of using aggregated meta-data for answering queries about information is of great interest with respect to the notion of a semantic web as aspired by Tim Berners Lee [1] and has to be further investigated in the future.

References

1. Tim Berners-Lee and Mark Fischetti. *Weaving the Web : The Original Design and Ultimate Destiny of the World Wide Web by its Inventor.* Harper, San Francisco, Oktober 1999.
2. Pierre-Antoine Champin. Rdf tutorial. Available at http://www710.univ-lyon1.fr/ champin/rdf-tutorial/, June 2000.
3. Stefan Decker, Dan Brickley, Janne Saarela, and Jürgen Angele. A query and inference service for rdf in proceedings of ql'98 - the query languages workshop. 1998.
4. D. Fensel, I. Horrocks, F. Van Harmelen, S. Decker, M. Erdmann, and M. Klein. Oil in a nutshell. In *12th International Conference on Knowledge Engineering and Knowledge Management EKAW 2000*, Juan-les-Pins, France, 2000.
5. M.-C. Rousset. Verifying the world wide web: a position statement. In F. van Harmelen and J. van Thienen, editors, *Pro-ceedings of the Fourth European Symposium on the Validation and Verification of Knowledge Based Systems (EUROVAV97)*, 1997.
6. Frank van Harmelen and Jos van der Meer. Webmaster: Knowledge-based verification of web-pages. In M. Ali and I. Imam, editors, *Proceedings of the Twelfth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, (IEA/AEI99)*, LNAI. Springer Verlang, 1999.
7. Thomas Vögele, Heiner Stuckenschmidt, and Ubbo Visser. Buisy - using brokered data objects for environmental information systems. In Klaus Tochtermann and Wolf-Fritz Riekert, editors, *Hypermedia im Umweltschutz*, Marburg, 2000. Metropolis Verlag.