Landes-Umweltportale

Vernetzung von Informationen in den Umweltportalen von Baden-Württemberg, Sachsen-Anhalt und Thüringen unter Einsatz einer kommerziellen Suchmaschine

T. Schlachter; W. Geiger; R. Weidemann; G. Zilly
Forschungszentrum Karlsruhe GmbH
Institut für Angewandte Informatik
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen

R. Ebel; M. Tauber
Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg
Griesbachstr. 1
76185 Karlsruhe

K. Zetzmann; A. Sawade; R. Mayer-Föll
Umweltministerium Baden-Württemberg
Kernerplatz 9
70182 Stuttgart

V. Bachmann; B. Köther
Ministerium für Landwirtschaft und Umwelt des Landes Sachsen-Anhalt
Olvenstedter Straße 4
39108 Magdeburg

W. Rott; D. Keil
Thüringer Ministerium für Landwirtschaft, Naturschutz und Umwelt
Beethovenstr. 3
99096 Erfurt

1. Warum Landes-Umweltportale?

Die EU-Richtlinie über den Zugang der Öffentlichkeit zu Umweltinformationen sowie die darauf aufbauenden Umweltinformationsgesetze des Bundes und der Länder verpflichten die Behörden, den Bürgern Umweltinformationen zugänglich zu machen und diese Informationen aktiv zu verbreiten. Die relevanten Umweltinformationen liegen aber in sehr verschiedenen Formen (Fachdokumente, Mess- und andere Sachdaten, Geoinformationen) und, soweit überhaupt öffentlich zugänglich, verteilt über zahlreiche Internetangebote vor.

Um den Bürgern einen singulären Einstiegspunkt und damit den vom Umweltinformationsgesetz (UIG) geforderten „leichten Zugang" zu den behördlichen Umweltinformationen bereitzustellen, aber auch als behörden-interne Arbeitserleichterung, sollen Landes-Umweltportale die behördlichen Umweltinformationen eines Bundeslandes mit übergreifenden Such- und Navigationsfunktionen möglichst umfassend erschließen. Sie ergänzen damit PortalU®[1], das Umweltportal auf Bundesebene, z.B. dadurch, dass auch kommunale Internetangebote aufgenommen werden und damit die Kommunen eine Plattform für die Erfüllung ihrer gesetzlichen Pflichten aus dem UIG erhalten.

2. Die bisherigen Landes-Umweltportale

Die bisherige, erste Generation der Landes-Umweltportale von Baden-Württemberg und Sachsen-Anhalt, die von 2003 bis 2007 noch ohne Thüringen entwickelt wurde, konzentriert sich im Wesentlichen auf textuelle Umweltinformationen. Hauptkomponenten der Systeme sind die Metadaten-Verwaltung auf Basis des Content Management Systems (CMS) WebGenesis®[2], die Suchmaschine ht://Dig[3] und das eigentliche Portal. Die über die web-basierte CMS-Autorenoberfläche gepflegten Metadaten parametrisieren sowohl die Volltextsuche als auch die Benutzeroberfläche des Portals. Die vom Volltext-Crawler indizierten Seiten werden zusätzlich über die Semantic Network Services /1/ verschlagwortet. Eine Schlagwortsuche über den UBA-Thesaurus (Umweltthesaurus des Umweltbundesamtes) und die Unterstützung der Volltextsuche durch kontext-sensitiv vorgeschlagene, zu den Suchbegriffen ähnliche Begriffe sind derzeit in den Portalen noch nicht freigeschaltet. Eine detailliertere Beschreibung der ersten Generation der Landes-Umweltportale wurde bereits an anderer Stelle veröffentlicht /2/.

Außer Baden-Württemberg und Sachsen-Anhalt bietet bisher noch kein anderes Bundesland ein Landes-Umweltportal im umfassenden Sinne an, es werden jedoch einzelne Teilaspekte abgedeckt. Niedersachsen integriert in das Webangebot des Umweltministeriums eine auf niedersächsische Informationsanbieter vorkonfektionierte Suchanfrage an das PortalU®. Schleswig-Holstein betreibt mit dem InfoNet-Umwelt[4] in Ergänzung zum Landesportal ein Informationssystem, in das Firmen und Privatpersonen umweltrelevante Informationen aus dem Land einstellen können. Darüber hinaus führt eine Reihe weiterer Länder Metadatenka-

[1] http://www.portalu.de
[2] http://www.iitb.fraunhofer.de/servlet/is/2223/
[3] http://htdig.sourceforge.net/
[4] http://www.umwelt.schleswig-holstein.de/

taloge über vorhandene Informationsangebote. Abgesehen von der niedersächsischen Lösung fehlt hier jedoch die unmittelbare Erschließung der Originalquellen.

Gewonnene Erfahrungen und Anforderungen an die neue Generation

Geänderte Randbedingungen, technischer Fortschritt und die Erfahrungen aus dem bisherigen Betrieb der Landes-Umweltportale führten zum Entschluss, die nächste Generation der Umweltportale von Grund auf neu zu konzipieren. Da die bisher eingesetzte Open-Source-Volltextsuchmaschine ht://Dig inzwischen nicht mehr weiterentwickelt wird und durch den Umfang der zu indizierenden Inhalte die Grenze ihrer Leistungsfähigkeit erreicht hat, ist die Ablösung dieser Suchmaschine unabdingbar und vordringlich /3/. Weiterhin soll zukünftig die Einbindung strukturierter Informationen, insbesondere von Datenbankinhalten, verbessert und ausgebaut werden. Die anstehende Ablösung des Umweltdatenkatalogs durch eine entsprechende Erweiterung der InGrid®-Software /4/, /5/ macht ebenso wie die angestrebte, möglichst redundanzfreie, bessere Anbindung der Umweltinformationen an PortalU® ein Redesign notwendig.

Eine zentrale Bedeutung bei der Neukonzeption kommt der Volltextsuche zu. Wie zahlreiche Untersuchungen belegen /6/ und auch durch eine eigene Nutzungsanalyse untermauert wurde /7/, prägen die Internet-Suchmaschinen und hier vor allem Google das Suchverhalten der meisten Nutzer. Benutzer erwarten, dass sie mit minimalem Aufwand (Eingabe eines oder mehrerer Begriffe in einem Suchfeld) zu einer passenden Ergebnisliste kommen. Die Wahrscheinlichkeit, dass ein Suchergebnis die Aufmerksamkeit des Nutzers erringt, ist für die ersten Listenelemente am höchsten und nimmt danach schnell ab. Dass einzelne Ergebnisse aus Sicht des Nutzers völlig unpassend sind, wird als lästig aber „normal“ empfunden. Dagegen werden komplexere Zugangswege, die eine längere Navigation oder eine spezielle Parametrisierung der Suchanfrage erfordern, vergleichsweise wenig, wohl hauptsächlich von Intensivnutzern oder Experten verwendet, auch wenn auf diesem Weg die Ergebnismenge besser eingeschränkt werden kann. Dem Gelegenheitsnutzer ist der Aufwand, sich mit einer unbekannten oder ungewohnten Oberfläche vertraut zu machen, oft zu hoch.

Für das Redesign der Landes-Umweltportale ergeben sich daraus wichtige Konsequenzen: Da die breite Öffentlichkeit eine Hauptzielgruppe der Landes-Umweltportale bildet, ist davon auszugehen, dass die Volltextsuche auch weiterhin den primären Zugangsweg darstellen wird. Deshalb sollten über diesen Weg möglichst alle Inhalte, auch wenn diese nicht unmittelbar in Textform vorliegen (z.B. Kartendarstellungen), erreicht werden können. Die adäquate Sortierung der Suchergebnisse und damit die Bewertungsfunktion für die Relevanz der einzelnen Suchergebnisse hat hohe Bedeutung für die Nutzerakzeptanz. Suchmaschinen verwenden in der Regel mehr oder weniger ausgefeilte statistische und strukturelle Kriterien zur Relevanzbewertung. Eine Verbesserung der Suchergebnisse könnte erreicht werden, wenn die Suchmaschine auch die Semantik von Anfragen und Inhalten berücksichtigen würde, sei es über eigene Funktionalitäten oder über geeignete Schnittstellen. Dabei könnte z.B. die Indizierung von Inhalten oder der Suchvorgang auf der Basis externer Wissensstrukturen wie Thesauri oder Ontologien semantisch angereichert werden.

Auch wenn die Volltextsuche eine derart zentrale Stellung einnimmt, heißt dies nicht, dass auf andere Zugangswege gänzlich verzichtet werden kann. So sollen auch weiterhin thema-

tische und anwenderspezifische Zugänge, Metadaten-Suche oder Newsfeeds bei gezielten Fragestellungen speziell darauf zugeschnittene Unterstützung liefern.

3. Architektur-Alternativen

Basierend auf den geschilderten Randbedingungen wurden drei Architektur-Alternativen für die zweite Generation der Landes-Umweltportale von Baden-Württemberg, Sachsen-Anhalt und Thüringen entworfen und bewertet:

(1) Austausch von ht://Dig durch eine funktional gleichwertige Suchmaschine

Bei dieser „kleinen“ Lösung wird ht://Dig durch ein aktuelleres Produkt ausgetauscht. Die Funktionalität der bisherigen Umweltportale kann erhalten werden, es sind aber auch keine wesentlichen Mehrwerte zu erwarten. Konkret wurden die Open-Source-Produktkombination Lucene/Nutch sowie das kommerzielle Produkt dtSearch als Ersatz für ht://Dig näher untersucht.

(2) Einsatz einer kommerziellen Suchmaschine

Kommerzielle Suchmaschinen, wie die im Projekt betrachteten Google Search Appliance (GSA) und Oracle Secure Enterprise Search (OSES), haben praktische Vorteile in Bezug auf Performance, Skalierbarkeit, Stabilität, Unterstützung etc. Funktional gehören sie einer anderen, leistungsfähigeren Produktkategorie als die unter (1) aufgeführten Produkte an. So ergeben sich Überschneidungen mit Funktionen der Umweltportale, die bisher außerhalb der Suchmaschine im CMS abgedeckt wurden. Eine Neustrukturierung der Umweltportale ist damit nicht zu vermeiden. Hauptnachteil der Werkzeuge sind fehlende Schnittstellen zur Einbettung in ein Gesamtsystem, insbesondere was die Konfiguration und die Eingabe notwendiger Metadaten betrifft. Die GSA ist als Stand-Alone-Werkzeug konzipiert.

(3) Einsatz von InGrid®

Als dritte Alternative wurde der Einsatz der InGrid®-Software, die als Basis für PortalU® /4/ entwickelt wurde und noch wird, konzeptionell im Detail untersucht. Der größte Vorteil hier ist die relativ problemlose Integration mit PortalU® durch die gleichzeitige Anbindung der InGrid®-Schnittstellenkomponenten an PortalU® und die Landes-Umweltportale. Dies erleichtert die redundanzfreie Erfassung erforderlicher Metadaten und verhindert die Mehrfach-Indizierung von Informationsquellen. Allerdings sind einige Anforderungen an die Landes-Umweltportale (z.B. Themensuche) nur durch die Erweiterung der InGrid®-Software realisierbar. Eine praktische Evaluierung der Software von August bis Oktober 2007 ergab außerdem, dass InGrid® noch keine für einen breiten Einsatz durch Dritte notwendige Produktreife erreicht hatte.

Auswahl einer Alternative

Die Abwägung der Vor- und Nachteile der drei Alternativen führte zu dem Schluss, dass mit dem Einsatz von InGrid® die im Gesamtzusammenhang mit Blick auf Schnittstellen, Wiederverwendbarkeit und Dienstearchitektur (SOA) eleganteste Lösung zu erreichen wäre. Das Risiko durch die fehlende Produktreife, der zu erwartende Aufwand und die Dringlichkeit ei-

ner Ablösung der Suchmaschine ht://Dig gaben jedoch den Ausschlag, diese Lösung auf einen späteren Zeitpunkt zu verschieben und kurzfristig eine Lösung mit der Google Search Appliance (GSA) anzugehen. Für die GSA sprachen insbesondere deren Produktreife, die durch die Nutzer der funktional in großen Teilen gleichen Google-Internetsuchmaschine täglich evaluiert wird, die Nutzerakzeptanz, die Funktionalität und die in Referenzinstallationen bestätigten kurzen Entwicklungszeiten. Zusätzlich wird der hohe Bekanntheitsgrad von Google für die Öffentlichkeitsarbeit der Umweltverwaltungen auf Entscheiderebene als positiver Werbeeffekt gesehen.

4. Konzept für ein GSA-basiertes Umweltportal

Bei der GSA[5] handelt es sich um eine kombinierte Hardware-/Softwarelösung mit einem einfachen Lizenzmodell, das allein auf der Anzahl maximal indizierbarer Seiten beruht. Die GSA realisiert eine Volltextsuche, die über eine eigene web-basierte Administrationsoberfläche parametrisiert wird. Relevante Inhalte werden mit einfachen URL-Mustern oder regulären Ausdrücken durch die Angabe von Start-URLs und Positiv-/Negativ-Listen beschrieben, vom GSA-Crawler ermittelt und anschließend indiziert. Zusätzlich kann die GSA Inhalte von Datenbanken indizieren und Eingaben („Feeds“) zur automatisierten Konfiguration heranziehen.

Anfang 2008 wurde mit der Evaluierung der GSA begonnen. Das vorhandene Grobkonzept für die Landes-Umweltportale wird derzeit verfeinert. Wegen der als gegeben zu betrachtenden Funktionalität der GSA und der eingeschränkten Mächtigkeit ihrer Schnittstellen erfolgt diese Verfeinerung teilweise bottom-up, was bedeutet, dass eventuell einige im Grobkonzept vorgesehene Funktionen mit vertretbarem Aufwand nicht realisierbar sein werden und dafür u.U. andere, bisher nicht vorgesehene Funktionen durch die GSA ohne wesentlichen Zusatzaufwand bereitgestellt werden können.

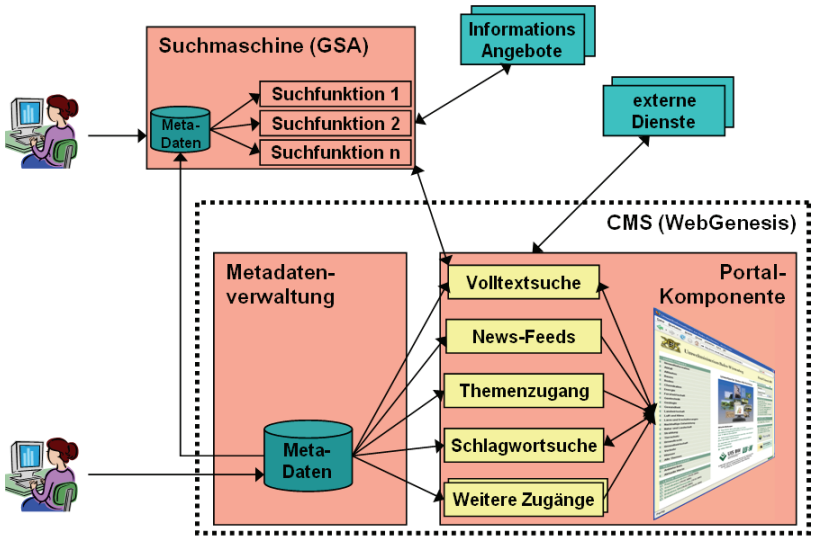

Abbildung 1: Struktur der Landes-Umweltportale mit Google Search Appliance.

[5] http://www.google.de/enterprise/gsa/features.html

Um die GSA möglichst schnell in den Produktivbetrieb zu bringen und andererseits mittelfristig deren volle Funktionalität nutzen zu können, ohne große Verzögerungen durch die dafür notwendige Einarbeitungs- und Entwicklungszeit in Kauf nehmen zu müssen, wurde ein schrittweises Vorgehen gewählt. Das Gesamtkonzept geht davon aus, dass die Landes-Umweltportale zumindest in der ersten Stufe weiterhin aus den drei Komponenten Volltextsuchmaschine, Metadaten-Verwaltung und Portal-Komponente bestehen (vgl. Abbildung 1).

Die GSA ersetzt anfangs lediglich die bisherige Volltextsuchmaschine und nutzt dabei nur einen Teil ihrer vorhandenen Funktionalität. Anschließend werden sukzessive die weiteren Funktionen der GSA evaluiert und in das Konzept eingepasst. Soweit sich dabei Überschneidungen mit bereits im CMS bzw. in der Umwelt-Portal-Software vorhandenen Funktionen ergeben, werden diese aufgelöst. Das Ziel ist, nach vollständiger Evaluierung der GSA eine Entscheidungsgrundlage zu haben, ob deren Funktionalität ausreicht, um die Landes-Umweltportale künftig in geänderter Form, z.B. durch Integration in vorhandene Portale der Länder, betreiben zu können, oder ob das bisherige, dreikomponentige Konzept (ggf. mit Änderungen) weiterverfolgt werden muss, um die vorhandenen Anforderungen zu erfüllen.

4.1 Erschließung von Datenbankinhalten

Eine große Menge vorliegender Umweltdaten konnte bisher mit der verwendeten Volltextsuchmaschine ht://Dig nicht erschlossen werden, da der in solchen Systemen übliche Mechanismus eines „Crawlers“, der ausgehend von einer Startseite deren Inhalt indiziert und rekursiv den vorhanden Links zu weiteren Inhaltsseiten folgt, nicht greift.

Für die folgenden Betrachtungen werden Datenbankinhalte in drei Kategorien eingeteilt:

- Inhalte, die ausschließlich über eine Datenbankschnittstelle (z.B. per SQL-Abfrage) verfügbar sind. Solche Daten stehen meist übergeordneten Anwendungen zur Verfügung und sind nicht über eine Web-Schnittstelle erreichbar.
- Inhalte, die als Basis von Webanwendungen dienen, welche die Daten ausschließlich über einen Abfragemechanismus (z.B. Formularauswahl) zur Verfügung stellen.
- Inhalte, die als Basis von Webanwendungen dienen, welche über Links in entsprechenden Menüstrukturen alle Daten zugänglich machen.

Für die Indizierung durch eine Internet- bzw. Web-Volltextsuchmaschine sind nur die ersten beiden Kategorien problematisch. Die verlinkten Webanwendungen der dritten Kategorie können durch den Crawler der Suchmaschine indiziert werden.

4.1.1 Indizierung von Daten über eine Datenbankschnittstelle

Für den ersten Punkt bietet es sich an, in der Datenbank eine Sicht auf die Daten zu generieren, welche jeweils alle wesentlichen Informationen zu einem Sachobjekt enthält. Hierzu müssen in der Regel Daten aus mehreren Tabellen zusammengeführt werden.

Die Volltextsuchmaschine wird per Datenbankabfrage mit dieser Sicht verbunden und bekommt eine Anzahl von „Zeilen“ der Sicht geliefert, welche nun innerhalb der Suchmaschine zwischengespeichert und indiziert werden können. Wenn bei einer Suchanfrage an die Volltextsuchmaschine ein Treffer innerhalb einer solchen Zeile auftritt, muss eine Sicht auf diese

Trefferzeile generiert werden, denn der Inhalt steht lediglich als Ergebnis der Datenbankabfrage zur Verfügung. Eine solche „Stellvertreterseite“ kann beispielsweise durch die Verwendung eines XSLT-Stylesheets dargestellt werden (siehe Abbildung 2).

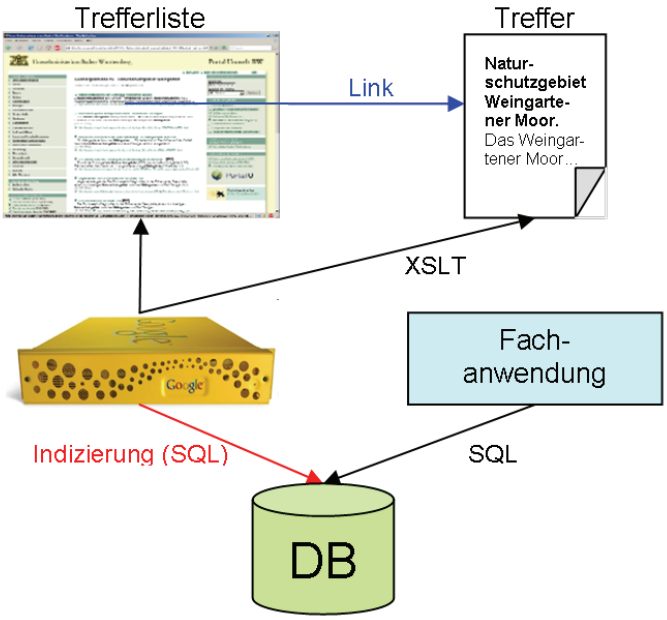

Abbildung 2: Direkte Indizierung von Datenbankinhalten und Darstellung der Inhalte über Stellvertreterseiten, welche per XSLT durch die Suchmaschine erzeugt werden.

4.1.2 Indizierung von Web-Inhalten hinter einer Formularabfrage

Im Gegensatz zum vorigen Fall geht es hier um Daten, welche bereits in einer Webanwendung zur Verfügung stehen, die aber, weil die Daten nur über eine Formularauswahl angesprungen werden und nicht anderweitig verlinkt sind, für Volltext-Crawler nicht erreichbar sind.

Zur Lösung dieses Problems gibt es mehrere Wege, von denen drei in den folgenden Abschnitten beschrieben werden.

Generieren von Jump-Pages oder Sitemaps

Eine für alle Volltextsuchmaschinen – egal ob Google, Yahoo & Co. oder eigene – praktikable Lösung ist das Generieren von künstlichen Jump-Pages, welche Links zu allen verfügbaren Inhalten enthalten. Solche müssen separat, z.B. durch eine Datenbankabfrage, generiert werden und haben den Vorteil, dass man durch sie eine Kontrolle über alle zu indizierenden Inhalte hat. Statt dem Erzeugen von Jump-Pages ist auch die Generierung von Sitemap-Dateien möglich, welche jedoch nicht von allen Suchmaschinen verstanden werden /8/.

Erzeugen von Links zu Inhaltsseiten per Datenbank-Feed

Eine weitere Möglichkeit zur Erschließung solcher Inhalte ist es, den Suchmaschinen per Datenbank-Feed die vollständigen URLs aller zu indizierenden Seiten mitzuteilen. Hierzu erzeugt eine Datenbank-Sicht ggf. neben weiteren Inhalten auch die URLs der zu indizieren-

den Seiten und teilt diese der Volltextsuchmaschine per Feed mit. Bei diesen Feeds handelt es sich im Prinzip ebenfalls um Sitemaps, die jedoch aktiv von der Datenbank an die Suchmaschine übertragen werden. Vorteil dieser Methode ist es, dass neben dem Inhalt der Seiten, auf welche die URLs verweisen, auch die Inhalte der Datenbankabfrage in den Index aufgenommen werden können. Hierdurch ist es möglich auch Inhalte zu indizieren, welche auf der Seite gar nicht dargestellt werden. Ein wesentlicher Nachteil dieser Methode ist allerdings, dass sie nur für solche Suchmaschinen funktioniert, die diese Feeds verstehen und auch tatsächlich damit gefüttert werden. Insbesondere funktioniert dieser Weg bei allen Internet-Suchmaschinen nicht.

Erzeugen von Links aus Primärschlüsseln

Eine mit der vorigen Methode verwandte Möglichkeit zur Anbindung an eine Volltextsuchmaschine ist es, der Suchmaschine eine feste Basis-URL sowie die relevanten Schlüssel per Datenbankabfrage zur Verfügung zu stellen, welche die Suchmaschine dann zu vollständigen URLs verbinden kann. Zum Beispiel wird bei der Abfrage eines Umweltdatenkatalogs (UDK) der Primärschlüssel als Parameter PK der festen Basis-URL übergeben:

```
http://www.udk-domain.de/wwwudk/UDKDetailServlet?Type=Data&PK={docid}
```

Ein Vorteil dieser Methode ist es, dass die Datenbank nichts über die URLs bzw. Adressierung der aus ihren Inhalten erzeugte(n) Webanwendung(en) wissen muss.

Auch bei dieser Variante besteht die Möglichkeit, zusätzlich zum Inhalt der Seiten jeweils auch passende Datenbankinhalte indizieren zu lassen (siehe Abbildung 3). Leider hat aber auch diese Methode den Nachteil, nur für explizit so angeschlossene Suchmaschinen zu funktionieren.

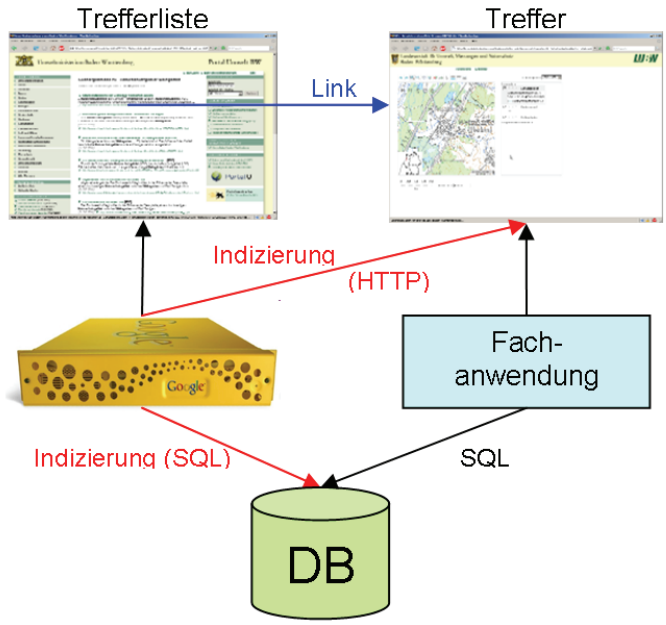

Abbildung 3: Indizierung von Webanwendungen über Datenbankfelder und Seiteninhalt.

4.2 Untermengen und Themensuche

In der Volltextsuche kann ein Mehrwert erreicht werden, wenn sich Suchanfragen auf definierte Untermengen der indizierten Inhalte beschränken lassen. Dies kann genutzt werden, um Abfragen auf einzelne Web-Sites oder Informationsanbieter einzuschränken oder auch Untermengen nach fachlichen Kriterien zu bilden.

PortalU®/InGrid® führt für anbieterspezifische Abfragen den Informationsanbieter als Metadatum bei allen indizierten Inhalten mit. Bei der GSA hingegen lassen sich Teilindexe über reguläre Ausdrücke (URL-Muster) aus dem Gesamtindex ausschneiden. Diese sogenannten Collections können sich überlappen und bei einer Suche über Mengenoperationen verknüpft werden. So lassen sich z.B. mehrere Angebote desselben Anbieters bündeln.

Über solche Collections lassen sich nicht nur die Suchanfragen der drei Landes-Umweltportale länderspezifisch beantworten; sie werden auch benutzt, um anbieter- und themenspezifische Anfragen zu ermöglichen. So ist z.B. die einzelne Abfrage der Umweltdatenkataloge der beteiligten Länder über solche Collections möglich.

4.3 Semantische Erweiterung der Suchanfragen

Alle Suchanfragen werden durch das in der GSA enthaltene deutsche Wörterbuch automatisch um Flexionsformen erweitert. Darüber hinaus konnte auch der UBA-Thesaurus in die GSA integriert werden. Hierdurch werden auch verwandte Suchbegriffe aus dem Bereich Umwelt gefunden bzw. die Suche nach umgangssprachlichen Begriffen um die entsprechenden Fachbegriffe ergänzt. Eine Suche nach „Müll“ enthält beispielsweise auch Treffer zu „Abfall“ und „Schrott“. Diese direkte Integration des Umweltthesaurus ersetzt große Teile der bisher verwendeten Anreicherung durch „ähnliche Suchbegriffe“ bzw. die Verschlagwortung aller Dokumente durch die Semantic Network Services (SNS) /1/.

Neben der Erweiterung der Suchanfragen durch Wörterbücher und Thesauri ist mit der GSA auch die manuelle Pflege von besonders relevanten Ergebnissen für bestimmte Suchanfragen (Key-Matches) möglich. So kann zum Beispiel bei der Eingabe eines Begriffes wie „Feinstaub“ auf eine passende Übersichts-Webseite hingewiesen werden (Abbildung 4).

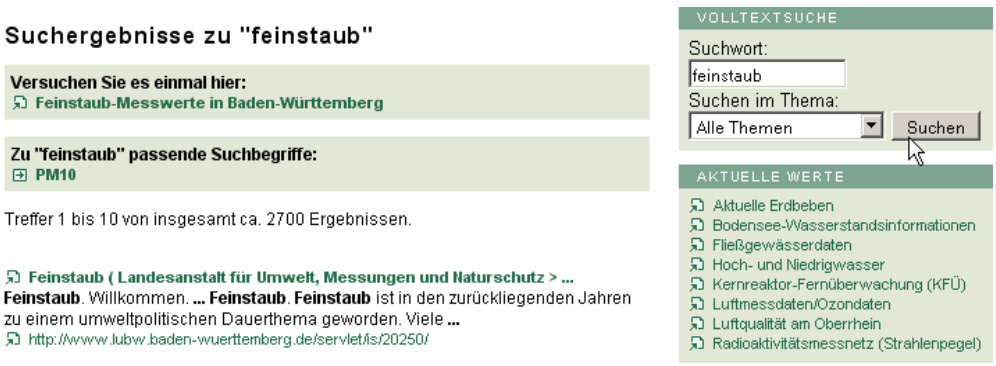

Abbildung 4: Automatisierte Erweiterung der Suchanfrage und Anzeige von Key-Matches (Ausschnitt aus der Ergebnisliste einer Volltextsuche im Portal Umwelt-BW).

Zusätzlich lassen sich in der Ergebnisliste Hinweise auf verwandte Suchbegriffe einblenden. Dieser Mechanismus kann zum Beispiel verwendet werden, um Benutzer beim Aufkommen aktueller Umweltfragen wie „Feinstaub“ oder „Gammelfleisch“ auf die entsprechenden Fachtermini hinzuweisen, noch ehe diese eine Aufnahme in den Umweltthesaurus gefunden haben.

Eine dritte, sehr flexible Möglichkeit zur Erweiterung der Ergebnisliste sind sogenannte One-Boxes (Abbildung 5). Ausgelöst durch bestimmte Suchbegriffe können neben der eigentlichen Ergebnisliste weitere Suchergebnisse eingeblendet werden. Diese können entweder aus einer parallelen Suche in bestimmten Collections kommen oder durch die Online-Abfrage eines weiteren Informationssystems gewonnen werden. Zum Beispiel könnten künftig auf diese Weise bei der Suche nach einem Ortsnamen die aktuellen Immissionsdaten einer Station aus dem Luftmessnetz eingeblendet werden.

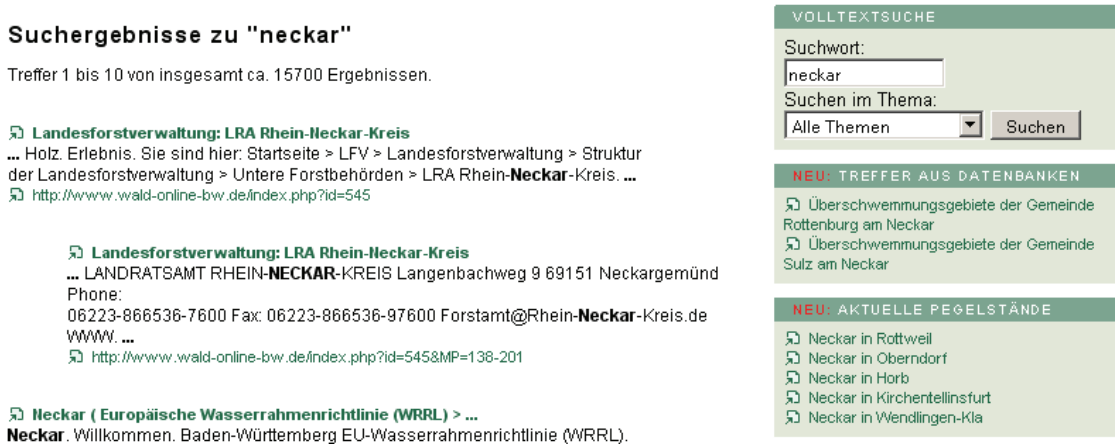

Abbildung 5: OneBoxes liefern neben der Trefferliste passende Hinweise auf Datensätze aus Datenbanken und aktuelle Pegelstände.

4.4 Einbindung in die bestehenden Umweltportale

Das Kernproblem bei der Einbettung einer Komponente in ein Gesamtsystem ist die Frage nach geeigneten Schnittstellen (API). Zur Einbettung einer Volltextsuchmaschine werden insbesondere Schnittstellen benötigt, um Suchanfragen und Ergebnislisten austauschen zu können. Weiterhin sollte eine genügend mächtige Schnittstelle zur Verwaltung der Suchmaschine zur Verfügung stehen, um beispielsweise die Start-URLs pflegen zu können.

Die Einbindung der GSA-Suche, hier insbesondere in die Portalkomponente, erweist sich als unproblematisch. Eine Suche wird entweder direkt über ein Formular oder über eine parametrisierte URL angestoßen. Die verschiedenen Parameter und die zugehörigen Optionen sind gut dokumentiert. Das Suchergebnis kann in einem Google-XML-Format oder über ein XSLT-Stylesheet bereits in ein gewünschtes Zielformat transformiert zurückgeliefert werden. Dies könnte auch das OpenSearch-Format[6] sein, das als standardisierte Form der Ausgabe von Suchergebnissen immer weitere Verbreitung findet.

[6] http://www.opensearch.org/

Hauptproblem der Architektur der neuen Landes-Umweltportale ist die Pflege der in den verschiedenen Komponenten benötigten Metadaten. Metadaten werden zum einen in der Portal-eigenen CMS-gestützten Metadaten-Komponente und zum anderen in der Administrationskomponente der GSA benötigt (s. Abb. 1). Über die Portal-Metadaten-Komponente werden die Navigationsfunktionen der Portalkomponente gesteuert und mit Inhalten versorgt, z.B. mit Angaben zu Anbietern und themenbezogenen Einstiegspunkten. Ein Teil der Metadaten wird sowohl innerhalb der GSA für die Parametrisierung des Crawlers als auch außerhalb für die Navigationsfunktionen der Portal-Komponente benötigt. Die Administration der GSA ist derzeit jedoch nur über eine Web-Oberfläche möglich; eine API, die man zum Abgleich der Metadaten nutzen könnte, gibt es bisher nicht. Um die betreffenden Metadaten trotzdem nicht doppelt erfassen zu müssen, ist vorgesehen, die web-basierte Admin-Oberfläche der GSA über ein Skript aufzurufen und die entsprechenden Metadaten automatisiert in das GSA-Formular zu übertragen.

5. Erfahrungen beim Betrieb

Seit Januar 2008 wird die GSA evaluiert. Nach den bisherigen Erkenntnissen ist das Laufzeitverhalten sehr gut. Die Administration ist teilweise etwas umständlich, da beispielsweise gewisse Informationen (z.B. URL-Muster) an mehreren Stellen jeweils erneut eingegeben werden müssen. Insbesondere die mangelnde Mandantenfähigkeit erweist sich als hinderlich, da es auf direktem Weg nicht möglich ist, dass die beteiligten Länder die sie betreffenden Teile isoliert sehen und pflegen können. Hier wird versucht, das Problem mit einigen speziell zu entwickelnden Skripts zu entschärfen. Die in die Administrationskomponente integrierten Analysewerkzeuge erweisen sich als sehr hilfreich, da bereits in kurzer Zeit eine Vielzahl bisher unentdeckter Fehler in den indizierten Informationsangeboten gefunden werden konnten (tote Links, mehrfach indizierte Seiten). Als Nebeneffekt der Fehlerbeseitigung für die GSA erfolgt gleichzeitig eine Optimierung der Angebote in Bezug auf Internet-Suchmaschinen.

Insgesamt betrachtet, konnten sehr schnell erste Indizierungen und Suchen durchgeführt werden. Innerhalb von drei Wochen war der erste vollständige, optimierte Volltextindex der drei Umweltportale im Umfang von ca. 250.000 Webseiten für Tests verfügbar. Inzwischen wurden sukzessive neue Informationsquellen angeschlossen, darunter auch Datenbanken, wie z.B. ausgewählte Inhalte (Naturschutzgebiete) von „Umwelt-Datenbanken und -Karten online“[7]. Die GSA-Suche wurde ohne größere Probleme als Ersatz für ht://Dig in die Landes-Umweltportale von Baden-Württemberg und Sachsen-Anhalt sowie ein neues, prototypisches Portal für Thüringen integriert. In Baden-Württemberg wird die GSA nicht nur die Volltextsuche im Landes-Umweltportal ersetzen, sondern die Volltextsuche im Umweltinformationssystem des Landes generell vereinheitlichen. Verschiedene Suchräume werden dabei durch Collections implementiert. Als erstes wurde die Suche in der Web-Site der LUBW bzw. des Umweltministeriums Ende April 2008 zur öffentlichen Nutzung testweise freigeschaltet. Erste Nutzerreaktionen sind sehr positiv und bestätigen die erwartete hohe Akzeptanz der neuen Suche.

[7] http://brsweb.lubw.baden-wuerttemberg.de/

6. Ausblick

Die Realisierung der neuen Landes-Umweltportale folgt einem anspruchsvollen Zeitplan. Die Funktionalität der bisherigen Volltextsuche konnte innerhalb weniger Monate bis zur Produktionsreife gebracht und sogar erweitert werden (vgl. Abbildung 6). Die Ablösung der bestehenden Umweltportale in Baden-Württemberg und Sachsen-Anhalt ist bereits erfolgt, die Freigabe des ersten Umweltportals für Thüringen soll Mitte des Jahres erfolgen.

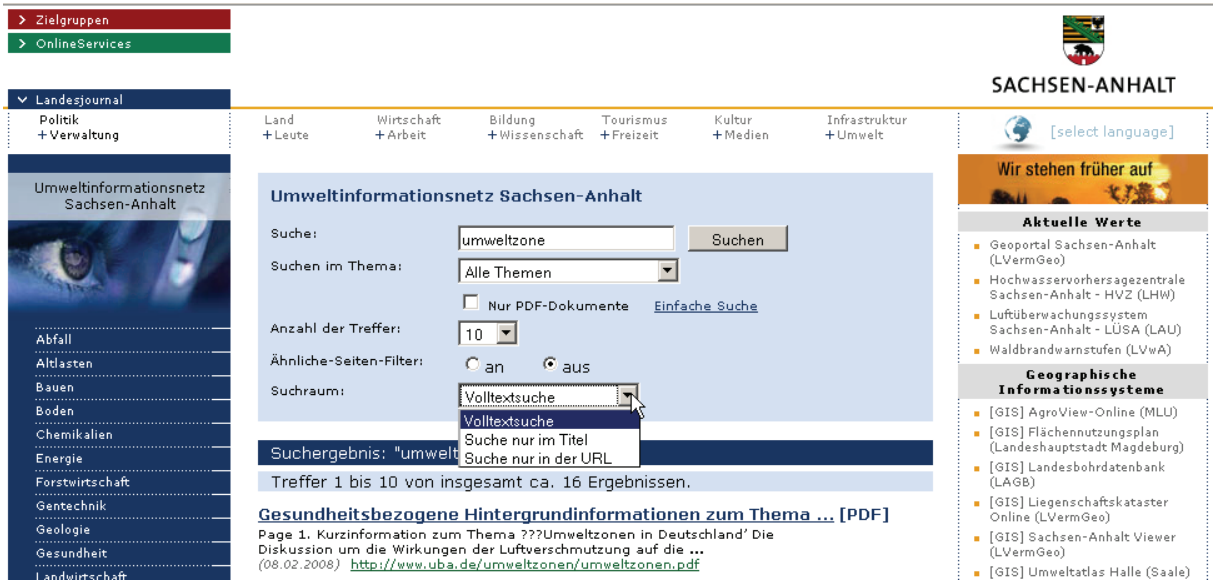

Abbildung 6: Erweiterte Suchmöglichkeiten im Umweltinformationsnetz Sachsen-Anhalt.

Schwerpunkt in der zweiten Jahreshälfte 2008 wird die Evaluierung der weiteren Funktionalität der GSA sein, insbesondere für eine stärkere semantische Unterstützung von Suche und Navigation z.B. über eine durch Suchworte getriggerte Einbindung spezieller Informationsangebote. Bei der Umsetzung entsprechender Portalfunktionen erfolgt ein Abgleich bzw. eine Zusammenführung mit den vorhandenen, CMS-basierten Zugangsfunktionen. Bis Ende des Jahres werden zwei wesentliche Projektergebnisse erwartet: Auf der einen Seite sollte festgestellt werden können, inwieweit eine Suchmaschine wie die GSA zur Realisierung von Landes-Umweltportalen ausreicht bzw. welche Zusatzfunktionen erforderlich sind. Auf der anderen Seite sollte die schrittweise Einführung der GSA dazu führen, dass die Nutzung der Landes-Umweltportale wegen der für viele Nutzer attraktiven Verwendung der Google-Suche steigt, ohne Einschränkungen in der bisherigen Funktionalität hinnehmen zu müssen. Insgesamt verspricht die Umsetzung des neuen Konzepts neben einer höheren Planungssicherheit durch Verwendung einer dauerhaft gepflegten Basis-Software auch eine wirtschaftliche Erstellung der Landes-Umweltportale durch Einsatz einer leistungsfähigen kommerziellen Suchmaschine.

7. Literatur

/1/ Rüther, M., Bandholtz, T., Menger, M. (2006): SNS Environmental Vocabulary – from Terms to Ontology. Conf. Semantics 2006, From Visions to Applications - Semantics: The New Paradigm Shift in IT, Wien, 28.-30. November 2006.

/2/ Schlachter, T. et al. (2007): Accessing administrative environmental information. In Tatnall, A.; Hrsg: Encyclopedia of Portal Technologies and Applications, Vol.1, S.20-25, Hershey, Pa.: Information Science Reference.

/3/ Schlachter, T. et al. (2007): UINBW und UINST – Ausbau der Umweltinformationsnetze von Baden-Württemberg und Sachsen-Anhalt; technische Weiterentwicklung. In: Mayer-Föll, R., Keitel, A., Geiger, W.; Hrsg.: F+E-Vorhaben KEWA Kooperative Entwicklung wirtschaftlicher Anwendungen für Umwelt, Verkehr und benachbarte Bereiche in neuen Verwaltungsstrukturen Phase II 2006/2007, Forschungszentrum Karlsruhe, Wissenschaftliche Berichte, FZKA 7350, S. 7-20.

/4/ Klenke, M., Kruse, F., Lehmann, H., Riegel, T., Vögele, T. (2006): InGrid® 1.0 - The Nuts and Bolts of PortalU®. In Tochtermann, K.; Scharl, A. (Hrsg.): Managing Environmental Knowledge, Shaker-Verlag, Aachen.

/5/ Vögele, T., Klenke, M., Kruse. F. (2007): Metadata Creation and Management of Distributed Data Catalogs with PortalU and InGrid 1.1, EnviroInfo 2007, 21st Int. Conf. on Informatics for Environmental Protection, 12.-14. September 2007, Warschau, Polen.

/6/ Zhang, J., Dimitroff, A. (2005): The impact of webpage content characteristics on webpage visibility in search engine results (Part I), Information Processing & Management, Volume 41, Issue 3, Cross-Language Information Retrieval, S. 665-690, Mai 2005.

/7/ Weidemann, R., Ebel, R., Mayer-Föll, R.; Hrsg. (2005): Fachdokumentenmanagement im Umweltinformationssystem Baden-Württemberg, Forschungszentrum Karlsruhe, Wissenschaftliche Berichte, FZKA-7200, http://www.lubw.baden-wuerttemberg.de/servlet/is/29855/

/8/ http://www.sitemaps.org/de/, besucht am 13.03.2008