

Texte

**01**  
**09**

ISSN  
1862-4804

## Umweltinformationssysteme

Suchmaschinen und  
Wissensmanagement -  
Methoden und Instrumente

Umwelt  
Bundes  
Amt 

Für Mensch und Umwelt





# Umweltinformationssysteme

## Suchmaschinen und Wissensmanagement - Methoden und Instrumente

Workshop des Arbeitskreises  
„Umweltdatenbanken/Umweltinformationssysteme“ der Fachgruppe  
„Informatik im Umweltschutz“, veranstaltet im Umweltbundesamt  
Dessau-Roßlau am 05. und 06. Juni 2008

Diese Publikation ist ausschließlich als Download unter  
<http://www.umweltbundesamt.de>  
verfügbar.

Die in den Beiträgen geäußerten Ansichten  
und Meinungen müssen nicht mit denen des  
Herausgebers übereinstimmen.

**Herausgeber:** Umweltbundesamt  
Postfach 14 06  
06813 Dessau-Roßlau  
Tel.: 0340/2103-0  
Telefax: 0340/2103 2285  
Internet: <http://www.umweltbundesamt.de>

**Redaktion:** Fachgebiet IV 2.1  
Gerlinde Knetsch  
Karin Jessen  
Ines Beckmann

Dessau-Roßlau, Januar 2009

## Vorwort

Der Arbeitskreis „*Umweltdatenbanken/ Umweltinformationssysteme*“ der Fachgruppe *"Informatik im Umweltschutz"* veranstaltete am 05. und 06. Juni 2008 in der Bauhausstadt Dessau-Roßlau im Umweltbundesamt den jährlichen Workshop, der einleitend auf das 20-jährige Bestehen Bezug nahm. Anlass ein kurzes Resümee zur Historie und Entwicklung des Arbeitskreises zu ziehen.

1988 von einem kleinen Kreis interessierter Experten gegründet, die sich mit Konzepten und methodischem Design von Umweltdatenbanken befassten, entwickelte sich diese „Interessensgemeinschaft“ zu einem Netzwerk von Spezialisten der Informationstechnologie und Fachanwendern aus Behörden, Instituten und Forschungseinrichtungen des deutschsprachigen Raumes. Die Angliederung des Arbeitskreises an die Gesellschaft für Informatik prägte die inhaltliche Ausrichtung und die Zielstellungen der Arbeitskreistreffen. Während Ende der 80er und Anfang der 90er Jahre vorwiegend Konzepte, Methoden und Techniken zum Aufbau von Datenbanken im Fokus der Veranstaltung standen, Fachanwendungen vorgestellt und diskutiert wurden, erweiterten sich die Themen ab Mitte der 90er Jahre auf die Standardisierung von Fachverfahren, deren Vernetzung untereinander und die Bedeutung von Metadaten für das Auffinden und Bewerten von Umweltdaten. Neue Technologien, insbesondere die Möglichkeiten der Nutzung des Internet beeinflussten die weiteren Entwicklungen von Datenbanken und Umweltinformationssystemen mit Beginn des 3. Jahrtausends. Der Arbeitskreis bot ein Podium für Entwickler und Anwender von Fachapplikationen, die zunehmend auf Webtechnologien aufbauten, die eine Verbreitung und Vernetzung von Umweltdaten und Informationen für verschiedene Nutzergruppen ermöglichten.

Die Neufassung der Europäischen Richtlinie 2003/4/EG des Europäischen Parlaments und des Rates vom 28.01.2003 über den Zugang der Öffentlichkeit zu Umweltinformationen und die damit verbundene Novellierung der deutschen Umweltinformationsgesetze schuf den rechtlichen Rahmen für den freien Zugang zu Umweltinformationen bei informationspflichtigen Stellen. Diese rechtlichen Rahmenbedingungen prägten die technologischen Weiterentwicklungen; der Aufbau von Portalen und Umweltinformationssystemen in nationalen und internationalen Netzwerken, die Anwendung von internationalen Standards für Metadaten rückten in den Fokus der Themen des Arbeitskreises. Mit dem zunehmenden Trend des

Angebots an Umweltdaten im Internet steigen die Anforderungen an ein zielgerichtetes Auffinden dieser Informationen.

Die Schwerpunktthemen Suchmaschinen und Wissensmanagement, die Entwicklung von offenen Systemarchitekturen mit dienste-orientiertem Charakter und semantische Dienste standen auf der Tagesordnung des von ca. 45 TeilnehmerInnen besuchten Workshops im Umweltbundesamt in Dessau.

Durch die ausgewogene Mischung an Beiträgen zu Lösungen des Erschließens von Inhalten über Internetportale von Bundes- und Länderbehörden konnten verschiedene methodische Ansätze vorgestellt und diskutiert werden. Für den Bereich des Krisen- und Risikomanagements sind derartige strukturierte Suchen mit einer hohen Erfolgsquote von besonderer Bedeutung, um schnell und zielgerichtet Maßnahmen einzuleiten. Ein Beispiel bildet hierbei das EU-Projekt ORCHESTRA - Open Architecture and Spatial Data Infrastructure for Risk Management. Die Nutzung von Klassifizierungsverfahren und der Einsatz von Ontologien für die Informationsstrukturierung bieten Möglichkeiten des zielgerichteten Auffindens von Umweltinformationen. Semantische Netzwerke wie der SNS – Semantic Network Service – unterstützen den Gebrauch einer gemeinsamen Fachterminologie für deutsche Umweltinformationen und weisen damit den Weg einer dienste-orientierten Aufbereitung von Inhalten mit Umweltbezug.

Die Beiträge des Workshops werden erstmals mit einem englischen Abstract veröffentlicht. Das EU - Projekt *ICT ENSURE (Information and Communication Technologies - Environmental Sustainability Research - [www.ict-ensure.eu/](http://www.ict-ensure.eu/))* befasst sich unter anderem mit dem Aufbau eines europäischen Internetportals zu Informations- und Kommunikationstechnologien. Die in einer Literaturdatenbank zu verwaltenden Abstracts zu themenverwandten Veröffentlichungen stellen eine Säule des Portals dar. Einen Beitrag für die internetgestützte Datenbank leistet dieser Workshopbericht und eine „intelligente“ Suchmaschine wird diese Abstracts einem interessierten Nutzerkreis zu Tage fördern.



Gerlinde Knetsch

Umweltbundesamt  
Fachgebiet Informationssysteme Chemikaliensicherheit  
Dezember 2008

# Inhaltsverzeichnis

<b>Erschließen von Datenbank-Inhalten durch die Volltextsuche in Landes-Umweltportalen</b>	<b>1</b>
<i>Thorsten Schlachter, Renate Ebel, Rainer Weidemann</i>	
<b>Zur Methode automatisierter Kategorisierungsverfahren der Internetsuchmaschine des Bayern-Portals</b>	<b>9</b>
<i>Erich Weihs</i>	
<b>KnowledgeNavigator: Ontologiebasierte Datenintegration in der Umweltforschung</b>	<b>21</b>
<i>Matthias Faerber, Oliver Archne, Florent Jochaud, Stefan Jablonski</i>	
<b>Ein Umweltziel- und Kennzahlensystem für die Umweltverwaltung auf der Basis von Topic Maps</b>	<b>31</b>
<i>Hans-Knud Arndt, Stephan Jacob, Henner Graubitz</i>	
<b>Integrierte Umweltdatenbanken: Geodatenhaltung, Architekturalternativen, Inhaltliche und technische Aspekte der Modellierung</b>	<b>43</b>
<i>Gergely Lukács, Wassilios Kazakos, Dominik Kumer</i>	
<b>Semantische Dienste in ORCHESTRA – Stand und Ausblick</b>	<b>49</b>
<i>Ulrich Bügel, Dr. Désirée Hilbring, Thomas Usländer</i>	
<b>Nutzung von Ontologien zur Informationsstrukturierung im Themenpark Umwelt</b>	<b>59</b>
<i>Claudia Greceanu, Clemens Döpmeier, Renate Ebel</i>	
<b>5 Jahre Semantic Network Service (SNS) Aktueller Stand und Ausblick</b>	<b>69</b>
<i>Maria Rüter, Thomas Bandholtz</i>	
<b>Grundwasser-Online – Internetbasiertes UIS für Deutschland?</b>	<b>89</b>
<i>Thomas Gutzke</i>	
<b>Vom Umweltdatenkatalog zum InGrid Data Catalog - Aktuelle Anpassungen des UDK-Datenmodells und der Software InGrid im Hinblick auf INSPIRE</b>	<b>93</b>
<i>Martin Klenke, Fred Kruse, Stefanie Uhrich, Christiane Giffei, Sybille Peters</i>	

<b>Entwicklung einer Maßnahmendatenbank für die Aufstellung von Bewirtschaftungsplänen und Maßnahmenprogrammen zur Umsetzung der Wasserrahmenrichtlinie</b>	<b>107</b>
<i>Friedhelm Hosenfeld, Dirk Behrens, Marcus Lempert, Andreas Rinker, Michael Trepel, Angelika Steingräber</i>	
<b>ETOX: Informationssystem Ökotoxikologie und Umweltqualitätsziele - Ein Werkzeug zur Entwicklung von Umweltqualitätsanforderungen</b>	<b>119</b>
<i>Dieter Schudoma</i>	

# Erschließen von Datenbank-Inhalten durch die Volltextsuche in Landes-Umweltportalen

Thorsten Schlachter<sup>1</sup>, Renate Ebel<sup>2</sup>, Rainer Weidemann<sup>1</sup>

[Thorsten.Schlachter@iai.fzk.de](mailto:Thorsten.Schlachter@iai.fzk.de)

[Renate.Ebel@lubw.bwl.de](mailto:Renate.Ebel@lubw.bwl.de)

[Rainer.Weidemann@iai.fzk.de](mailto:Rainer.Weidemann@iai.fzk.de)

## Abstract

Comprehensive provision of environmental information to the citizen is the main goal of the environmental portals of the states of Baden-Wuerttemberg, Saxony-Anhalt and Thuringia. The second generation of these environmental portals goes beyond presenting data from classical web sites and also opens up additional information from databases and database driven information systems. This is realized by embedding a commercial search engine into the system.

## Zusammenfassung

Die übergreifende Bereitstellung von Umweltinformationen für den Bürger ist das Ziel der Umweltportale der Länder Baden-Württemberg, Sachsen-Anhalt und Thüringen. Die zweite Generation dieser Umweltportale geht über die Erschließung von Daten von klassischen Websites hinaus und erschließt zusätzlich Informationen aus Datenbanken und Informationssystemen, die auf Datenbanken aufsetzen. Hierzu wird eine kommerzielle Suchmaschine eingesetzt.

## 1 Einleitung

Die Umweltinformationsgesetze der EU, des Bundes und der Bundesländer 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, typischerweise verteilt über zahlreiche Internetangebote vor.

Um den Bürgern, aber auch Behörden selbst, einen zentralen Einstiegspunkt und damit den vom UIG geforderten „leichten Zugang“ zu den behördlichen Umweltinformationen bereitzustellen, wurden seit 2003 für die Länder Baden-Württemberg und Sachsen-Anhalt Landes-Umweltportale entwickelt. Sie ergänzen damit PortalU® [PortalU 2008], das Umweltportal auf Bundesebene. Einen

---

<sup>1</sup> Forschungszentrum Karlsruhe, Institut für Angewandte Informatik

<sup>2</sup> Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg (LUBW)

wesentlichen Zugang stellt dabei die Volltextsuche dar, welche textuelle Daten in Webseiten und den dort verlinkten Dokumenten erschließt. Mit der Umstellung auf eine neue, leistungsfähigere Volltextsuchmaschine im Jahr 2008 ist auch der Zugang zu Informationen möglich, die nur in Datenbanken vorliegen und deshalb per Volltext-Crawler nicht erreichbar sind. Der folgende Beitrag gibt einen Überblick über die Möglichkeiten zur Erschließung solcher, für Internet-Suchmaschinen nicht sichtbarer, Inhalte.

## **2 Erste Generation der Landes-Umweltportale**

Die 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 auf Websites. Hauptkomponenten der Systeme sind die Metadaten-Verwaltung auf Basis des Content Management Systems (CMS) WebGenesis, die Volltext-Suchmaschine ht://Dig und das eigentliche Portal. Die vom Volltext-Crawler indizierten Seiten werden zusätzlich über die Semantic Network Services (SNS) [SNS 2008] verschlagwortet. Anfragen werden automatisch um Synonyme und ähnliche Begriffe aus den SNS bzw. dem Umwelt-Thesaurus UmThes erweitert. Eine detailliertere Beschreibung der ersten Generation der Landes-Umweltportale wurde bereits an anderer Stelle veröffentlicht [Schlachter 2007a+b].

## **3 Erschließung von Datenbankinhalten per Volltextsuche**

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 vorhandenen Links zu weiteren Inhaltsseiten folgt, versagt.

Wir möchten für unsere Betrachtungen Datenbankinhalte in drei Kategorien einteilen:

- 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 die ersten beiden Kategorien problematisch. Die dritte Kategorie, verlinkte Webanwendungen, können über den Crawler der Suchmaschine indiziert werden.

### **3.1 Indizierung von Daten über eine Datenbankschnittstelle**

Zur Lösung des ersten Problems bietet sich der folgende Weg an: In der Datenbank wird eine Sicht auf die Daten generiert, welche alle wesentlichen Informationen zu

einem Sachobjekt enthält. Hierzu müssen in der Regel Daten aus mehreren Tabellen zu einer Sicht 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.

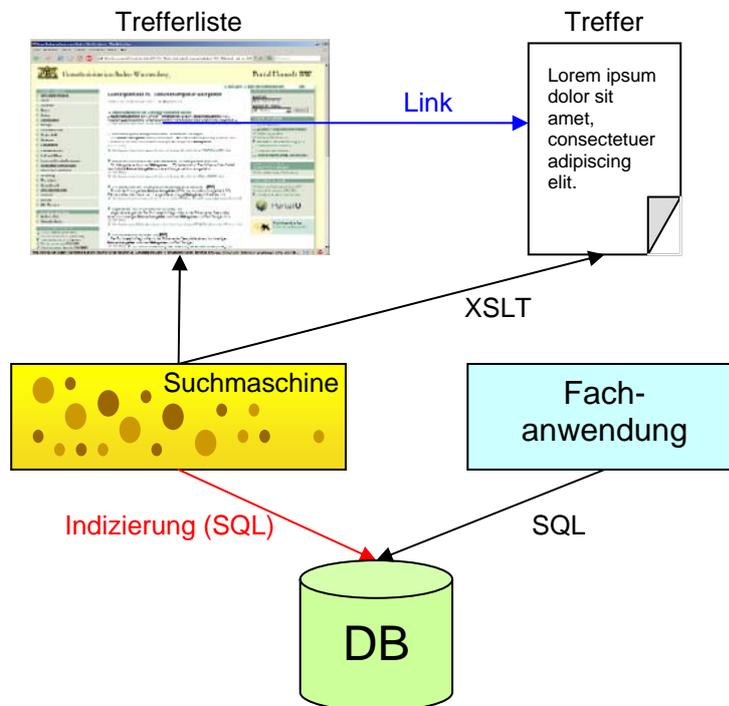


Abbildung 1: Direkte Indizierung von Datenbankinhalten und Darstellung der Inhalte über Stellvertreterseiten, welche per XSLT durch die Suchmaschine erzeugt werden.

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. Die Suchmaschine simuliert gewissermaßen eine Web-Ansicht auf diese Trefferzeile („Stellvertreterseite“). Dies kann beispielsweise durch die Verwendung einer Formatvorlage oder eines Stylesheets geschehen.

### 3.2 Indizierung von Web-Inhalten hinter einer Formularabfrage

Im Gegensatz zum gerade geschilderten Fall geht es hier um Daten, welche bereits in einer Webanwendung zur Verfügung stehen, die aber, weil die Daten nur über eine Formularauswahl angesprochen 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.

#### 3.2.1 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 Jump-Pages müssen separat, z. B.

durch eine Datenbankabfrage, generiert werden und haben den großen 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 [Sitemaps 2008].

### **3.2.2 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 neben weiteren Inhalten auch die URLs der zu indizierenden 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.

### **3.2.3 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; hierum kümmert sich die (mit der Basis-URL versorgte) Suchmaschine.

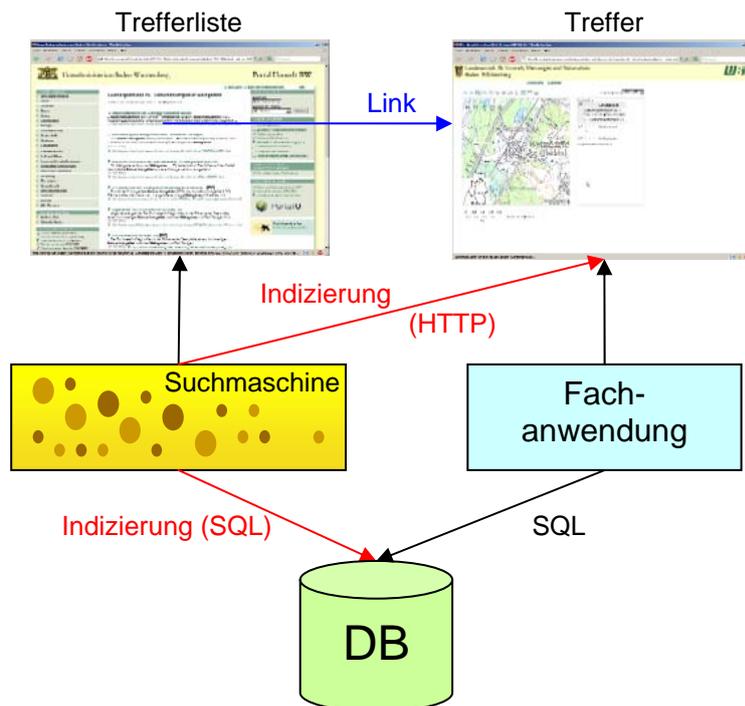


Abbildung 2: Gleichzeitige Indizierung von Webanwendungen über Datenbankfelder und Seiteninhalt

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

#### 4 Neue Generation der Landes - Umweltportale

Mittels der beschriebenen Wege ist es – eine geeignete Volltextsuchmaschine vorausgesetzt – möglich, auch Datenbankinhalte (z.B. Sachdaten) und datenbankgestützte Webanwendungen zu indizieren.

Die zweite Generation der Umweltportale der Länder Baden-Württemberg, Sachsen-Anhalt und Thüringen nutzt seit 2008 eine solche leistungsfähige Suchmaschine. Nachdem Tests der InGrid®-Software von PortalU®, welche neben einem Volltext-Crawler auch einen Datenbank-Connector bietet, zur Verwendung in diesen Portalen leider nicht befriedigend verlaufen waren, wurden mehrere Alternativen geprüft. Letztlich wurde das Produkt „Google Search Appliance“ (GSA) [Google 2008] beschafft, welches neben der Volltextindizierung per Crawler auch Methoden zum Zugriff auf Datenbankinhalte bietet.

Trefferlisten können dabei direkt durch die GSA mittels XSLT erzeugt oder über ein XML-Format in eigene Anwendungen integriert werden. Dabei ist es möglich, Suchanfragen durch die Verwendung von fertigen Wörterbüchern und manuell gepflegten Listen zu erweitern. Verschiedene Filter bieten die Möglichkeit die Suche z.B. auf bestimmte Dokumenttypen, Quellen, Meta-Angaben einzugrenzen.

Auch ist es möglich, parallel zur eigentlichen Suchanfrage weitere Datenquellen zur Laufzeit anzufragen und die Ergebnisse zusätzlich zur Ergebnisliste zu präsentieren.

Im Kontext der Suche nach einem Gewässernamen können so zum Beispiel die aktuellen Pegelstände eingeblendet werden, während die Suche nach einem Ortsnamen die aktuellen Immissionsdaten liefert. Diese Methode ist in vielen Fällen sinnvoller als große Mengen von Meßdaten per Volltextsuche zu indizieren. Hierdurch wird nicht nur die Indizierung großer gleichartiger Datenbestände vermieden, welche die Trefferlisten überschwemmen würden, sondern man kann auf diesem Weg auch sicher stellen, dass tatsächlich jeweils die aktuellsten Daten angezeigt werden.

## **5 Erste Erfahrungen mit der neuen Suchmaschine**

Ende Januar 2008 wurde mit dem Testbetrieb der Google Search Appliance begonnen. Innerhalb weniger Tage konnten die Inhalte der Umweltportale Baden-Württemberg und Sachsen-Anhalt (zusammen rund 300.000 Dokumente) indiziert und Sammlungen zum landes- bzw. themenscharfen Zugriff auf diese Inhalte angelegt werden. Darüber hinaus wurden weitere Sammlungen für einzelne Behörden angelegt. Nach der Implementierung einer Schnittstelle für das verwendete CMS konnte die bisherige Volltextsuche im Umweltministerium Baden-Württemberg und der Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg (LUBW) Mitte April 2008 durch die neue Suchmaschine abgelöst werden. Seit Anfang Mai 2008 ist auch die Suche im Portal Umwelt-BW auf die GSA-Suche umgestellt. Die Indizierung (durchschnittlich werden täglich über 100.000 Dokumente (re-)indiziert) sowie die Beantwortung der Suchanfragen laufen performant und bisher problemlos.

Zur Indizierung von Umweltdaten des Statistischen Landesamtes Baden-Württemberg wurde eine Jump-Page gemäß Abschnitt 3.2.1 erzeugt. Über diese Jump-Page-Generierung, welche durch eine Datenbankabfrage innerhalb einer Active Server Page (ASP) realisiert wurde, erfolgt gleichzeitig eine Eingrenzung auf Daten des gesamten Landes; die Daten zu weiteren rund 1.100 Regionaleinheiten sowie Daten ohne Umweltrelevanz werden dagegen nicht indiziert.

Auch komplexe Webanwendungen wie das auf Basis von disy-Cadenza [Disy 2008] realisierte System „Umwelt-Datenbanken und -Karten online“ der Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg können nun wie in Abschnitt 3.2.2 beschrieben indiziert werden. Da in dieser Software sehr viele verschiedene Fachanwendungen stecken, die jeweils über eigene Abfragen (Selektoren) verfügen, wurde auf Datenbankebene eine Sicht generiert, mit deren Hilfe es möglich ist, die verschiedenen Selektoren inklusive aller notwendigen Daten mit nur einer einzigen Datenbankverbindung in die GSA zu integrieren.

Hier werden durch die zusätzliche Indizierung der Datenbankinhalte auch Kartendarstellungen zu Naturschutzgebieten gefunden, obwohl deren Namen nicht als textueller Inhalt in den HTML-Seiten enthalten sind.

Die landesscharfe Anbindung der Umweltdatenkataloge von Baden-Württemberg, Sachsen-Anhalt und Thüringen ist nach der in Abschnitt 3.2.3 präsentierten Methode gelungen. Dabei werden die URLs durch eine Datenbankabfrage erzeugt und die entsprechenden Inhalte anschließend auf der Website [www.portalu.de](http://www.portalu.de) indiziert.

Alle Suchanfragen werden durch das in der GSA enthaltene deutsche Wörterbuch automatisch um alle Flexionsformen erweitert. Darüber hinaus konnte auch der

Umweltthesaurus des Umweltbundesamtes 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“.

## 6 Fazit

Der Anschluss von Umwelt-Datenbanken und von datenbankbasierten Webanwendungen an die Umweltportale macht für Bürger und Fachleute noch mehr Umweltinformationen verfügbar.

Für den Nutzer bieten die Umweltportale an zentraler Stelle eine breite Fülle von Umweltinformationen, welche sich durch den eingegrenzten Suchraum, die hohe Qualität behördlicher Informationen und durch exklusiv verfügbare Inhalte von den Internet-Suchmaschinen abheben.

## 7 Literaturverzeichnis

[Disy 2008] [http://www.disy.net/cadenza\\_web.html](http://www.disy.net/cadenza_web.html), besucht am 13.03.2008

[Google 2008] <http://www.google.com/enterprise/gsa/>, besucht am 13.03.2008.

[PortalU 2008] <http://www.portalu.de/>, besucht am 13.03.2008.

[Schlachter 2007a] Schlachter, T., Geiger, W., Weidemann, R., Ebel, R., Tauber, M., Mayer-Föll, R., Sawade, A., Bachmann, V., Köther, B.: Accessing administrative environmental information. In Tatnall, A. [Ed.]: Encyclopedia of Portal Technologies and Applications, Vol.1, S. 20-25, Information Science Reference ISBN 978-1-59140-989-2, Hershey, Pa., 2007.

[Schlachter 2007b] Schlachter, T., Grieß, C., Geiger, W., Weidemann, R., Zilly, G., Ebel, R., Tauber, M., Bachmann, V., Köther, B., Sawade, A., Mayer-Föll, R.: UINBW und UINST – Ausbau der Umweltinformationsnetze von Baden-Württemberg und Sachsen-Anhalt, technische Weiterentwicklung. In Mayer-Föll, R., Keitel, A., Geiger, W. (Hrsg.): Kooperative Entwicklung wirtschaftlicher Anwendungen für Umwelt und Verkehr in neuen Verwaltungsstrukturen, Phase II 2006/07, Forschungszentrum Karlsruhe, Wissenschaftliche Berichte, FZKA 7350, S. 7 - 20, 2007.

[Sitemaps 2008] <http://www.sitemaps.org/de/>, besucht am 13.03.2008.

[SNS 2008] [http://www.semantic-network.de/doc\\_intro.html?lang=de](http://www.semantic-network.de/doc_intro.html?lang=de), besucht am 13.03.2008.



# Zur Methode automatisierter Kategorisierungsverfahren der Internetsuchmaschine des Bayern-Portals

Erich Weihs<sup>3</sup>  
[erich.weihs@web.de](mailto:erich.weihs@web.de)

## Summary

For the Bavaria portal, [www.bayern.de](http://www.bayern.de), a "find machine" is used which covers the data space of public administration and closely related institutions, such as chambers of commerce, universities, employment offices, etc. What makes the find machine special is that it divides up the results lists according to life event. The life events are based on an agreement of a working group of the ministers of the interior at the federal and state levels. The life events are based on the gender mainstreaming approach, which categorizes by requirements and not in relation to a specific organization or a specific field. Overall, 650 life events were used for classification, divided up into citizen life events and corporate life events. The search engine of the Bavaria portal uses new technological methods to develop and process the public web-offerings (government, authorities, municipalities, universities etc.) according to the life events of citizens.

The find machine of the Bavaria portal can beside semantic classification categorize automatically using the mathematic-statistic support vector method (SVM). The advantage of the SVM method is enhanced stability of the results when the data space changes over time. Since the approach evaluates the distributions of the word space, it does not depend on the construct of a semantic definition, but instead only on the probability of its appearing. Another method which is increasingly being employed is to use a meta tag related to life event and place which enables a direct classification.

The search space of the Bavarian search engine is limited by official offerings, municipalities, institutions of higher learning, employment offices, chambers of commerce, etc. In certain cases, additional institutions such as churches are included in order to take account of the regional supply of kindergartens, for instance. On the whole, approx. 6000 URLs are tracked, with approx. 2½ million indexed sites. The data material is extremely varied. Along with html pages, it also reads and completely indexes pdf and doc formats, among others. This means that such things as municipal bulletins and other reports consisting of several hundred pages are collected and completely indexed. While reports can sometimes be thematically circumscribed section by section, this is often not the case with bulletins such as community magazines, etc. Here it is frequently a question of a chronological or editorial mixture of news, reports and statements, whose content cannot be assigned in detail to the life event classifications.

---

<sup>3</sup> Erich Weihs, Insterburger Strasse 7, 81929 Munich, Germany, [erich.weihs@web.de](mailto:erich.weihs@web.de)

# 1 Einführung

Das Internet ist innerhalb weniger Jahre zu einer unverzichtbaren Infrastruktur unserer Gesellschaft geworden. Den wichtigsten Dienst des Internets stellen die Suchmaschinen dar, die als Universalschnittstelle mit der digitalen Welt agieren und dem Nutzer die gewünschten Informationen in Form einer Linkliste liefern. Wir befinden uns "definitiv auf dem Weg in die Google-Gesellschaft".

Wenn wir uns auch am Weg von der Suchmaschine zur „Findemaschine“ im praktischen Einsatz noch am Anfang befinden, so wird hier ein Weg gezeigt, wie mit der Suchmaschine des bayerischen Portals ([www.bayern.de](http://www.bayern.de)) das öffentliche Angebot (Regierung, Behörden, Gemeinden etc.) strukturiert nach einem erweiterten Lebenslagenprinzip erschlossen werden kann. Nicht das Ordnungskriterium der Medien oder Organisationen steht im Vordergrund sondern das Ordnungskriterium des Bedarfs der Öffentlichkeit: Dem Bürger sollen zu einer bestimmten Lebenslage (Umzug, Hochzeit, Geburt, Umweltfragen, Gesundheit etc.) alle relevanten Informationen der Informationsanbieter bereitgestellt werden, wie die in der Situation erforderlichen Verwaltungsleistungen, Kontaktdaten zuständiger Behörden, nötige Formulare oder weiterer Informationen der behördennahen Anbieter, der Behörden und Gemeinden. Dadurch wird die Serviceorientierung und Dienstleistungsqualität erheblich gesteigert. Insgesamt werden ca. 650 hierarchisch gegliederte Kategorien ausgewertet.

Die Kategorisierung der etwa 2 ½ Mio. Seiten des öffentlichen Angebots Bayerns erfolgt nach Methoden, die in der Ausarbeitung an Hand konkreter Ergebnisse bewertet und beschrieben werden.

Ein inhaltlicher Schwerpunkt wird auf Umwelt bezogene Aspekte der Lebenslagen gelegt.

Der Lebenslagenansatz eignet sich einerseits als Analyseinstrument für die zielgruppenorientierte Erfassung von Politikfeldern und damit als Grundlage für die Politikgestaltung nach Gender Mainstreaming. Andererseits ist der Lebenslagenansatz auch geeignet, das Angebot einer Recherche von Suchmaschinen an Daten und Information nach den Kategorien der Lebenslagen in Ergebnislisten aufgeschlüsselt anzubieten.

Die in einer Bund/Länder Arbeitsgruppe <sup>4,5</sup> erarbeiteten Lebenslagenkategorien wurden in die Bereiche „Lebenslagenmodell für Bürgerinnen und Bürger“ und „Lebenslagen für Unternehmen“ aufgeteilt. Die Definition lässt Raum für Erweiterungen. Das Gliederungsprinzip wurde für die bayerische Suchmaschine (SUMA) übernommen. Um die Gemeinden nach dem bayerischen Modell ausreichend einzubeziehen, muss das Lebenslagenprinzip z.B. durch zusätzliche Kategorien wie „Leben in meiner Gemeinde“, „Fremdenverkehr“, etc. erweitert werden. Ebenso wurde die Oberkategorie Umwelt eingeführt. Insgesamt existieren etwa 50 Kategorien, die nach dem UIG relevant sind.

Das Lebenslagenprinzip deckt in seiner Gliederung eine Vielzahl von „Umweltlagen“ ab, wie beispielhaft:

---

<sup>4</sup> UAG Lagen (2006) "Gemeinsamen Vorschlag der UAG Lagen für DOL (Juli 2008)" der DOL-Unterarbeitsgruppe "Lagen" vom 12.07.2006-1

<sup>5</sup> Arbeitskreises VI der Innenministerkonferenz (2002)

<p><b>Bauen</b> Denkmalschutz <b>Gesellschaft und Politik</b> Agenda21 <b>Gefahrenabwehr und Sicherheit</b> Unwetterwarnungen Zivil- und Katastrophenschutz <b>Gesundheit und Vorsorge</b> Ernährung Forschung und Medizin Verbraucherschutz <b>Tierhaltung und Jagd</b> Artenschutz Ein- und Ausfuhr von Tieren Fischen Hundehaltung Jagd Tierärzte/Veterinärwesen Tierhaltung und Tierzucht Tierschutz Tierseuchenbekämpfung Tierzucht/Tierhaltung Veterinärüberwachung</p>	<p><b>Wohnen</b> Ver- und Entsorgung <b>Meine Umwelt</b> Allgemeines Landschaftspflege Besonderheiten aus Natur und Landschaft Energie Garten / Obstprodukte aus der Region Grundwasser Karten, Bilder Klima Lebensmittel / Lebensmittelüberwachung Messwerte aus Luft, Boden und Wasser Müll, Abfall, Wertstoffe Nachhaltigkeit Naturdenkmäler Schutzgebiete Seen / Gewässer Straßenbau / Straßenunterhalt / Verkehrswesen Umweltgesetze und Verordnungen</p>	<p>Umweltinformationen im Landkreis <b>Tourismus in meiner Gemeinde</b> Ausflugsziele Erholung <b>Land- und Forstwirtschaft</b> <b>Informationsdienste</b> <b>Formulare</b> <b>Forschung</b></p>
---	--	--

Tabelle 1: Umweltlebenslagen, Ausschnitt.

Dargestellt sind nur Umweltlebenslagen mit ihren hervorgehobenen Oberkategorien. Viele Umwelt bezogene Lebenslagen finden sich unter anderen Oberbegriffen, wie z.B. unter "Wohnen" die Kategorie „Ver- und Entsorgung“, vgl. Tabelle 1: Umweltlebenslagen, Ausschnitt.

## 2 Der Suchraum

### 2.1 Die Abgrenzung des Suchraums

Der Suchraum der bayerischen Suchmaschine (SUMA) ist durch das Behördenangebot, Gemeinden, Hochschulen, Arbeitsämter, Handelskammern usw. und weiteren Behörden nahen Institutionen begrenzt. Bei der Auswahl der Angebote wurde darauf geachtet, dass nur die gewünschten Informationen durch geeignete URLs indexiert werden (z.B. sollen aus dem Hochschulbereich Studienunterlagen etc. nicht indexiert werden – die Zahl der indexierten Seiten würde sich verdoppeln – und entsprechen auch nicht dem Lebenslagenangebot). Neben dem Behörden- und Kommunenangebot werden auch weitere Angebote von Institutionen wie Handelskammern, Arbeitsämtern, und Religionsgemeinschaften usw. einbezogen, um das regionale Angebot z.B. an Kindergärten oder caritativen Leistungen zu berücksichtigen. Insgesamt werden ca. derzeit 5000 URLs erfasst, mit derzeit ca. 2 1/2 Mio. indexierter Seiten.

## 2.2 Die Daten des Suchraums

Das indexierte Datenmaterial ist äußerst unterschiedlich. Neben html Seiten werden u.a. pdf und doc-Formate gelesen und vollständig indexiert. Darunter sind unter anderem auch Berichtsbände mit mehreren hundert Seiten oder Mitteilungsblätter der Gemeinden mit unterschiedlichsten Inhalten erfasst und vollständig indexiert. Während Berichtsbände sich manchmal abschnittsweise thematisch abgrenzen lassen, ist das bei Mitteilungsblättern wie Gemeindespiegel, usw. oft nicht der Fall. Hier haben wir es häufig mit einer chronologischen oder redaktionell gestalten Mischung von Nachrichten, Berichten und Statements zu tun, die sich einer inhaltlich detaillierten Einordnung in die Lebenslagen-Klassifikation entziehen und bei der Klassifizierung zu berücksichtigen sind.

Die Vollständigkeit des Suchraums soll auf Gemeindeebene durch eine einheitliche Erfassungsmethode der URLs sichergestellt werden. Erfahrungsgemäß wird hier für die QS immer eine Nacharbeit erforderlich sein, da verschiedene Angebote "outsourced" werden, deren URLs oft Provider abhängig sind und der Gemeinde unter Umständen nicht vor Augen sind. Vergleichbares gilt für in den Angeboten verwandten URLs. Die weiteren „behördennahen“ URLs müssen explizit gepflegt werden. Für die quantitative Kontrolle der Suchräume (sind diese befüllt) stehen der Administration verschiedene weiter unten behandelte Hilfsmittel zur Verfügung.

## 2.3 Unterteilung der Suchräume

Aus technischer Sicht ist eine Unterteilung der Suchräume möglich und erfolgt

- zur Prioritätssteuerung bei der Indexierung
- zur Unterscheidung der Suchräume für unterschiedlicher Datenquellen (z.B. Internet, Datenbanken)
- für unterschiedliche Indexierungsverfahren
- Tests

## 3 Das Klassifikationsverfahren

Für die Aufteilung einer recherchierten Ergebnisliste nach Lebenslagen (oder einer anderen Klassifikation) bestehen prinzipiell 2 Möglichkeiten:

- die Klassifikation erfolgt jeweils ad hoc nach der jeweiligen Recherche durch Aufbereitung der Ergebnisliste oder
- die Klassifikation erfolgt über die indexierte Grundgesamtheit aller durch die Suchmaschine erfassten Links jeweils nach der Indexierung, bevor die Daten für Recherchen freigegeben werden.

Die ad hoc Lösung wird hier nicht weiter verfolgt, da die Klassifizierungsergebnisse von der Recherche abhängen – und bei kurzen Listen statistisch gesehen nicht signifikant sind.

Grundsätzlich sind folgende Vorgehensweisen möglich:

- Kennzeichnung der Internetseiten durch den Informationsanbieter z.B. durch Metatags,

- Definition von (semantischen) Wortprofilen, als Grundlage der Zuordnung
- Mathematisch statistische Verfahren.
- Explizite Definition von URL's, denen z.B. bei der Erfassung durch die Suchmaschine bereits eine Lebenslage zugewiesen wird
- manueller Eingriff auf von der Suchmaschine erfassten Seiten, z.B. bei Werbung, Verstoß gegen die guten Sitten oder im Ranking.

### 3.1 Metatags

Die für den Betreiber der Suchmaschine fachlich sicherste Klassifikation ist die, die vom Datenanbieter selbst vorgenommen wird. Besonders für neu zu erstellende Seiten oder Dokumentationen eines Internetangebotes lässt sich eine Klassifikation nach Lebenslage, Raum und Zeit vornehmen:

```
<meta name="ByLebenslagen"
content="ByL1_MeineUmwelt_Energie" />
```

Für den Raumbezug nach Ortsnamen (PLZ, Koordinaten) lassen sich mit folgendem Syntax erfassen:

```
<meta name="ByRaumbezug" content="ByRb_GN[Amberg|Bogen]"/>
```

Für die Suchmaschine stehen 3 Kategorisierungsmethoden zur Verfügung, die bei der Qualitätssicherung unterschiedlich zu behandeln sind:

- Kategorisierung via Metatags
- Semantische Kategorisierung
- Mathematische Methode

Die Kategorisierung soll eine automatische Zuordnung der gefundenen Dokumente zu den Lebenslagen ermöglichen. Das erfordert eine sorgfältige Definition der Kategorien nach den unterschiedlichen unten beschriebenen Verfahren, die in der Regel miteinander kombiniert werden. Besondere Schwierigkeiten ergeben sich bei Unter-Kategorien die mit "Allgemein," bezeichnet sind. Allgemein bedeutet nicht eine Zusammenfassung der anderen Unterkategorien der Lebenslage sondern den Nachweis allgemeiner Informationen. Das stößt oft auf definitorische Schwierigkeiten einer kategorialen Eingrenzung. In einigen Fällen wurde die Kategorie "Allgemein" daher nicht definiert. Ebenso ist zu prüfen, ob ausreichende Fallzahlen für die Kategorie vorhanden sind/ sein können. Gleichmaßen deuten zu hohe Fallzahlen (über 1500) eventuell auf eine unscharfe Definition hin.

Die Kategorisierung via Metatags nimmt eine Sonderstellung ein, da Zuordnung durch den Informationsanbieter in seinen Web-Seiten erfolgt.

Die Anwendung der mathematischen Kategorisierungsmethode verallgemeinert die Kategorisierungsergebnissen der Anwendung der semantischen Methode in einer zweiten aufbauenden Stufe.

#### 3.1.1 Qualitätssicherung (QS) bei Metatags

Mit der Anwendung der für die Suchmaschine eigens definierten Metatags kann eine automatische Klassifizierung der Angebotsseiten nach Lebenslagen (fachliche Sicht) und Ort (Raumbezug) erreicht werden. Demzufolge liegt die Verantwortung der

richtigen Zuordnung beim Dienste Anbieter <sup>6</sup>. Die QS hat zumindest stichprobenartig darauf zu achten, dass die fachlichen Definitionen durch die Anbieter eingehalten werden, z.B. durch Kontrolle der einfach auffindbaren Metatags im Quelltext einer Seite.

### 3.1.2 Semantischen Klassifizierung

Die Definition der Lebenslagen-Kategorien erfolgt bei der semantischen Klassifizierung durch Bedingungen aus verschiedenen Begriffen (einzelne Wörter oder Zeichenketten wie "Lebenslagen Prinzip"), die m.E. mittels boolescher Operatoren (und, oder, nicht, near by, Stoppwort), in Reihe und durch Klammersetzung verknüpft werden - vgl. Abbildung 1.

"regionales Obst" oder "Bauernmärkte und Obst" oder "Bauernmarkt und Obst" oder "regionale Obstsorten" oder (Obstanbau und Region) oder (Obstanbau und regional) oder "regionales Gemüse" oder "Bauernmärkte und Gemüse" oder "Bauernmarkt und Gemüse" oder "regionale Gemüsesorten" oder (Gemüseanbau und Region) oder (Gemüseanbau und regional) oder (regionaltypische near[99] Lebensmittel) oder (Obstprodukte near[55] Garten) oder (Gartenbau near[55] Obst) oder (Region near[55] obst) oder Bundesgartenschau oder (Weinbau near[55] Garten) oder (Gartenbau near[55] Landespflege) oder (Streuobstwiese oder Streuobstwiesen) **Nicht Buergemeister nicht Buergemeisterin nicht Amtsblatt nicht Anzeigenpreise nicht impressum nicht Stadtbranchenbuch nicht Schulanzeiger nicht Sportartikel**

Abbildung 1: Semantische Definition der Kategorie „Garten / Obstprodukte aus der Region“

In Abbildung 2: Optimierung der Near-Parameter für Lebenslagen im Kontext mit Abwasser und Gebühren“ wurde der near Parameter Gebühren near[n] Abwasser mit  $n = 1, 2, \dots, 99$  variiert und ein Optimum bei  $N \approx 20$  gefunden. Eine weitere Überprüfung der Ergebnisse in den Lebenslagen "Meine Gemeinde", "Verordnungen" zeigt, dass hier ab etwa 20 Wörtern Abstand des erwähnten near-Kommandos die Wahrscheinlichkeit, aussagefähige Treffer zu erhalten, verwässert wird. Das Optimum wurde durch den Quotient aus den Treffern aus "Meine Umwelt" und "Meine Gemeinde" gebildet, der die Empfehlung nochmals verdeutlicht ( $\Delta \text{Treffer} / \Delta n$ ).

Aus Abbildung 2 kann geschlossen und m.E. auch aus anderen Ergebnissen gefolgert werden, dass für die Definition die fachlichen Kategorien Near-Befehle mit n-Parametern zwischen 15 und 50 anzuwenden sind.

<sup>6</sup> UAG Lagen (2006) "Gemeinsamen Vorschlag der UAG Lagen für DOL (Juli 2008)" der DOL-Unterarbeitsgruppe "Lagen" vom 12.07.2006-1

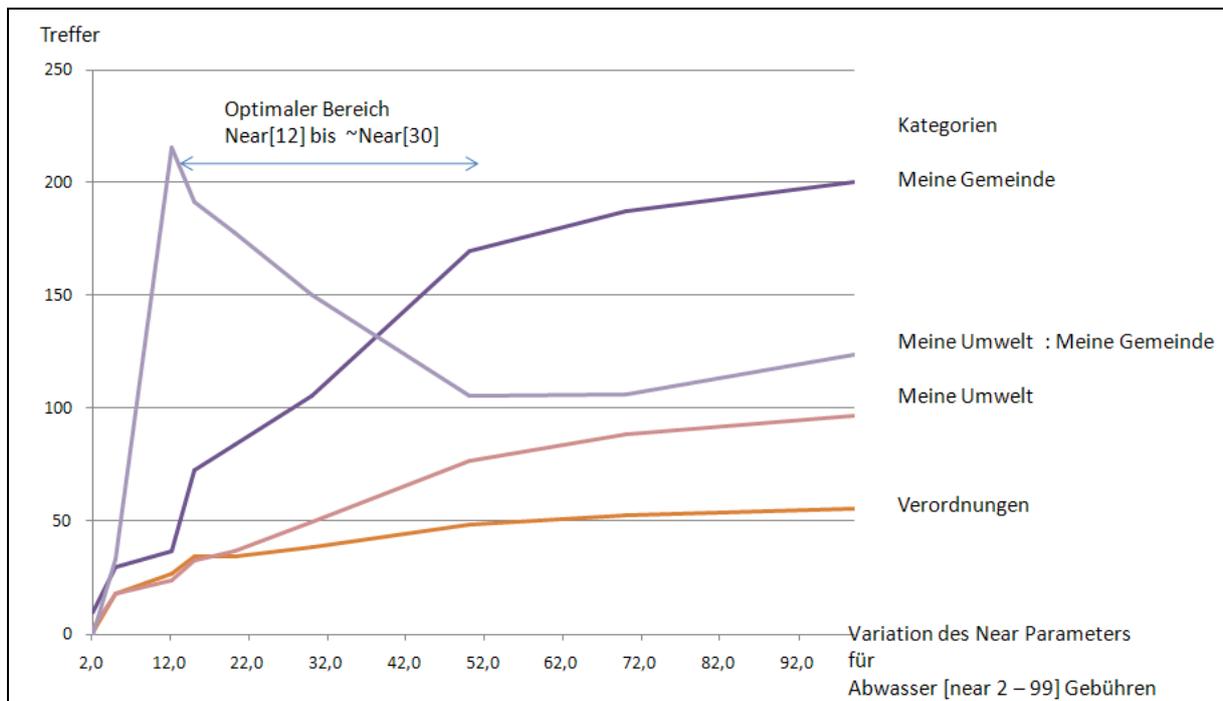


Abbildung 2: Optimierung der Near-Parameter für Lebenslagen im Kontext mit Abwasser und Gebühren

### 3.1.3 Automatische Klassifikation

Die automatische Klassifizierung soll den Aufwand für die Klassifizierung reduzieren und diese von der engen Bindung an booleschen Bedingungen der semantischen Definitionen lösen.

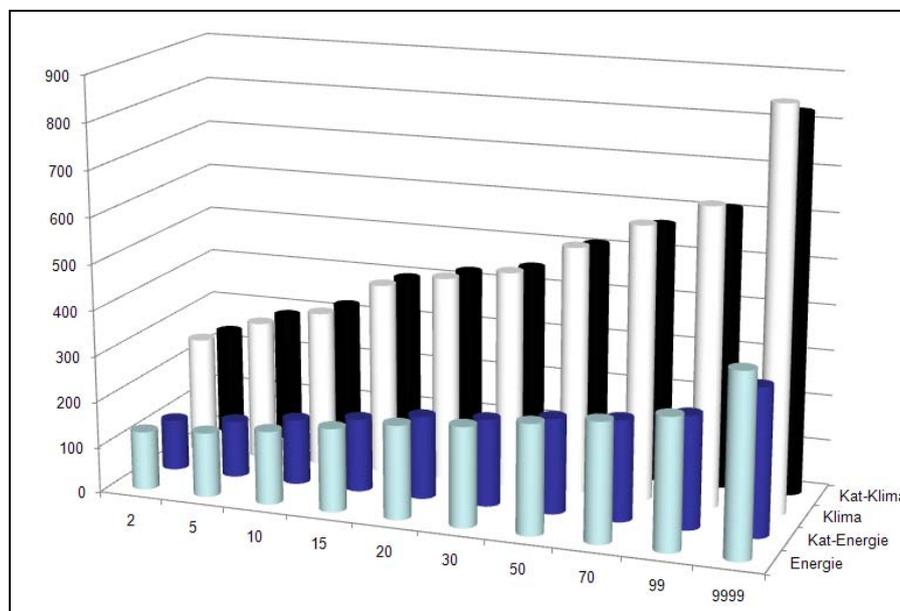


Abbildung 3: Vergleich der Treffer-Verteilungen zwischen semantischer Definition und Klassifikation nach der mathematischen Methode „Supported Vector Machine“ (SVM) als Plausibilitätstest.

Die SVM-Methode basiert auf folgendem Modell: Vorgegeben werden je Klasse zwischen 50 bis 200 Web-Seiten, die das Thema der Lebenslage umreißen. Die SVM Methode als multivariates mathematisch-statistisches Verfahren „charakterisiert“ diese „Lern“-Seiten in den zu bildenden Variablenraum durch die Cluster (Kategorien) abgrenzenden Hyperebene und sucht dann auf Grund dieser Charakterisierung (SV) die Seiten des Datenraums nach einem Wahrscheinlichkeitsmaß zuzuordnen.

Die Praxis hat gezeigt, dass es zweckmäßig ist, diese Lernseiten nicht manuell aus dem Webangebot zu selektieren sondern dafür eine Stichprobe aus den Daten der semantischen Klassifikation zu ziehen.

Da die booleschen Bedingungen nur für die Generierung der Lernseiten eine Rolle spielten, ist die Methode – im Gegensatz zur semantischen Definition - nicht mehr an die strengen booleschen Bedingungen gebunden, die bei nachfolgenden Crawling-Läufen ja so nicht erfüllt sein müssen, da sich das Web-Angebot laufend ändert. Trotzdem bleiben die nach der SV Methode erzielten „Lebenslagen-Cluster“ am Thema, da es weniger auf den einzelnen Begriff als auf die Verteilung aller Begriffe ankommt.

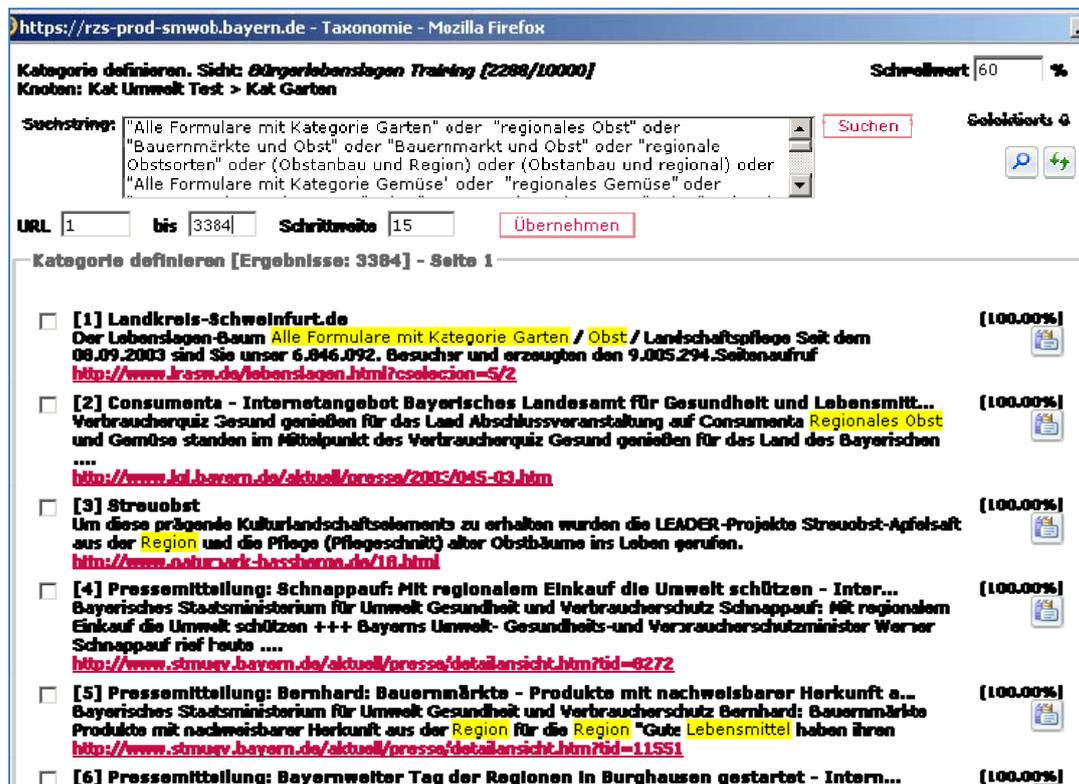


Abbildung 4: Bildung der Lernseiten als Stichprobe aus den Treffern der semantischen Definition

Die Qualität der SVM Methode ist gemäß Abbildung 3 stichprobenartig in weiteren Klassen zu überprüfen. Die QS hat folgendermaßen vorzugehen: Die Klassifizierung erfolgt getrennt nach beiden Methoden. Anschließend werden, ähnlich wie bei der Überprüfung der semantischen Methode, die Ergebnisse durch Vergleich der Verteilungen bewertet. Die Verteilungen entstehen durch Begriffspaare, die durch near-Kommandos variiert werden. In Abbildung sind exemplarisch die Verteilungen die Treffer der Kategorien „Energie,“ und „Klima“ verglichen. Es ist unmittelbar ersichtlich, dass die Ergebnisse in ihrer Verteilung sehr gut übereinstimmen.

### 3.1.4 Weitere Methoden

Die Praxis zeigt, dass es zweckmäßig oder erforderlich ist, für bestimmte Kategorien URLs explizit aufzunehmen z.B. bekannte eindeutige URLs eines Informationsdienstes oder diese auszuschließen.

## 3.2 Raumbezug

Neben der fachlichen Zuordnung nach Lebenslagen erfolgt eine ortsbezogene Zuordnung des Webangebots über Gemeindegchlüssel, Orts-(teil)namen, Koordinaten (Orts-(teil)namen und Koordinaten derzeit nicht unterstützt).

Die Zuordnung erfolgt entweder über

- Metatags oder mittels
- Zuordnungstabelle zu den URLs der Provider.

Damit wird auch eine Recherche ermöglicht, die den Ortsbezug berücksichtigt.

## 4 Schnittstellen

Die Suchmaschine ist durch SOA-Schnittstellen über das Web erreichbar. Bedient wird die öffentlich zugängliche Weboberfläche. Darüber hinaus besteht die Möglichkeit des Zugriffs auf andere Portale mit eventuell eingeschränktem Datenraum über eine eigene SOA-Schnittstelle. Dadurch ist es möglich, dass einzelne Anbieter, zum Beispiel Gemeinden, die Suchmaschine in ihre eigene Weboberfläche einbinden können. Ebenso denkbar ist die Einbindung in Dienste Dritter wie PortalU. Neben der SOA-Schnittstelle bestehen noch Schnittstellen zu Webdiensten wie dem Geodienst der Vermessungsverwaltung und zu einem Synonymverzeichnis.

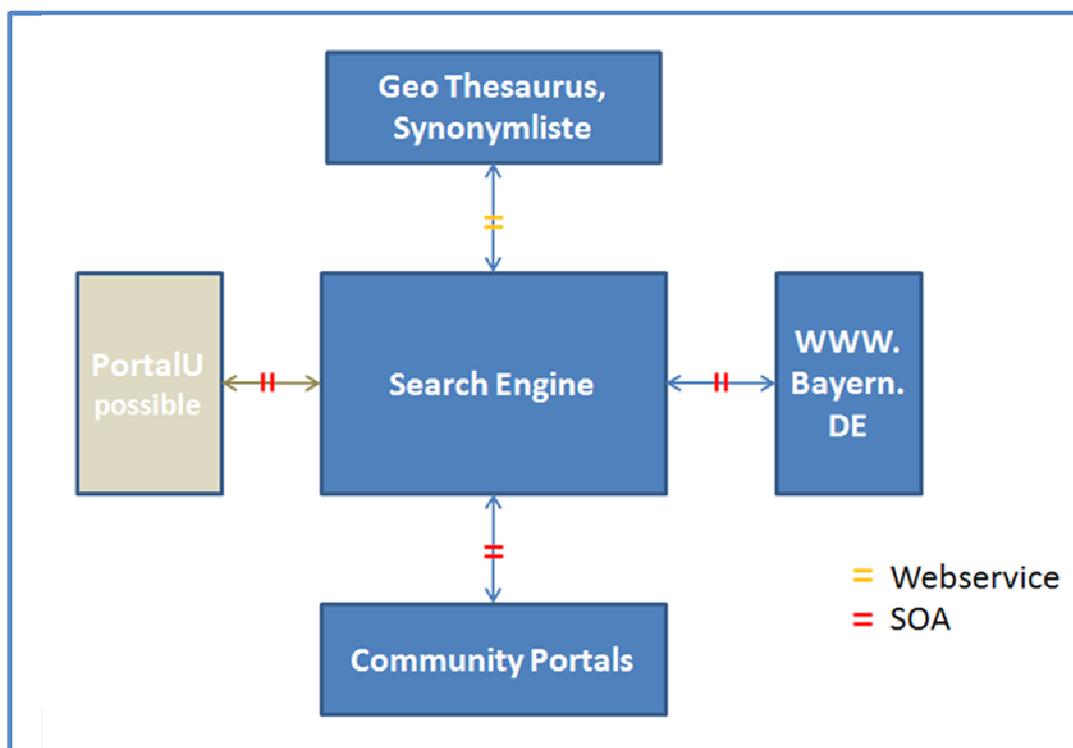


Abbildung 5: Die Schnittstellen der Suchmaschine

## 5 Ergebnisse

Die Web-Oberfläche der Suche bietet die Auswahl der Themen (Bürgerlebenslagen und Geschäftslbenslagen (1) und die Auswahlmöglichkeit der Georeferenzierung (in der Abbildung „Weinberg“, (2)). Dargestellt ist in der Abbildung 6 die Ergebnisliste und der Lebenslagenbaum (3), Durch einen Klick auf die gewünschte Kategorie, hier Geburt, wird die zugehörige Ergebnisliste ausgegeben.

Einen Sonderfall stellt die Behandlung von homonymen Suchwörtern dar. In der Regel erfolgt eine Zuordnung in der jeweils „richtigen“ Kategorie.

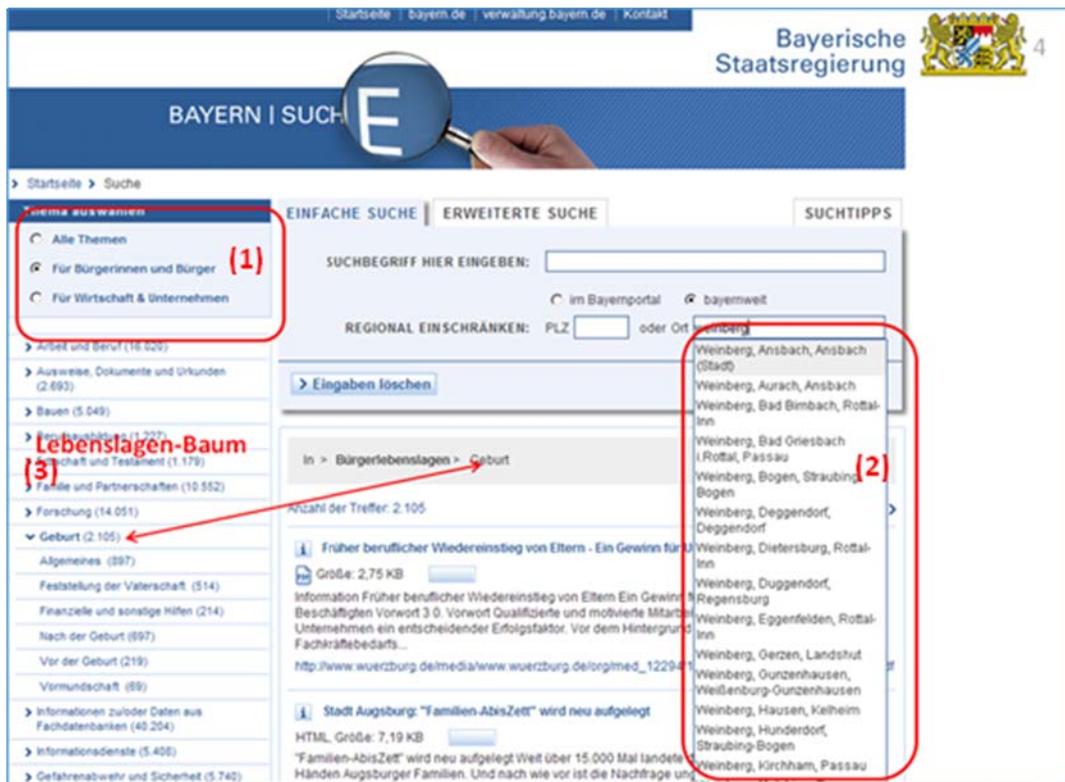


Abbildung 6: Die Web-Oberfläche der einfachen Suche

In Abbildung 7 ist das Ergebnis einer Suche nach "Umzug" dargestellt. In der Lebenslage "Kultur in meiner Gemeinde" sind Faschingsaufzüge, Prozessionen usw. aufgelistet (1). In den Kategorien "Mobilität" (2) und "Tierhaltung und Jagd" (3) werden z.B. Fragen beantwortet, die mit Hundehaltung und häuslichem Umzug zusammenhängen. Entsprechend verhält es sich bei den anderen Lebenslagen.

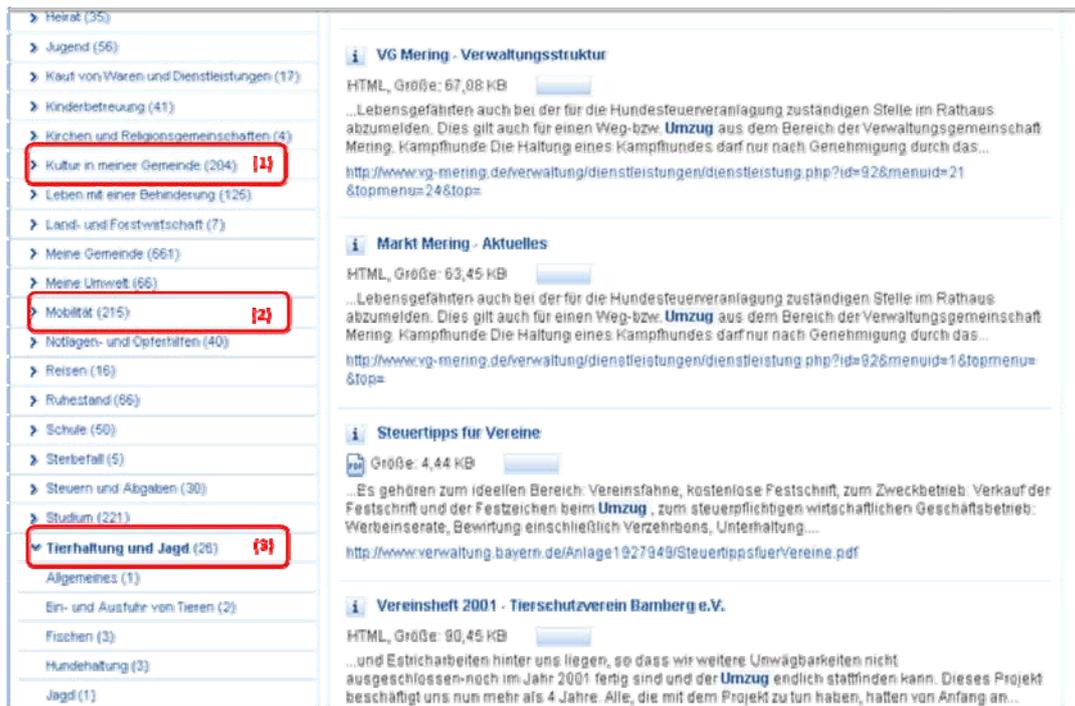


Abbildung 7: Die Behandlung von Homonymen

## 6 Zusammenfassung

Die Findemaschine des Bayernportals ist durch folgende Features charakterisiert:

- Die Abgrenzung des Contents öffentlicher Datenanbieter wie Gemeinden, Staatsregierung und Ämter und andere Einrichtungen, die der öffentlichen Verwaltung nahe stehen, wie Arbeitsämter, Handelskammern, Universitäten oder Fachhochschulen etc. Derzeit werden etwa 2 ½ Millionen über das Web zugängliche Seiten erfasst.
- Verwendung des Lebenslagen Prinzips zur Kategorisierung der Daten. Das Lebenslagenprinzip ist auf einzelne Lebenssituationen wie Geburt, Arbeit, Gesundheit usw. bezogen und nicht auf Organisationsformen von Behörden oder auf Fachkriterien.
- Verwendung verschiedener Kategorisierungsmethoden zur Definition der über 600 Kategorien, die in Lebenslagen der Bürger und Unternehmungen geteilt sind.
- Die Verwendung der mathematisch statistischen Supported Vector Methode (SVM) erlaubt eine statistisch abgesicherte Klassifizierung.
- Die Verwendung von Metatags erlaubt eine Klassifizierung nach Lebenslage und Raumbezug unmittelbar durch den Datenanbieter.
- Die Suchmaschine ist über SOA-Schnittstellen mit dem Internet verbunden.

## 7 Literaturnachweis

UAG Lagen (2006) "Gemeinsamen Vorschlag der UAG Lagen für DOL (Juli 2008)"  
der DOL-Unterarbeitsgruppe "Lagen" vom 12.07.2006-1

Arbeitskreises VI der Innenministerkonferenz (2002) Aktivitäten zur Staats- und  
Verwaltungsmodernisierung in den Ländern und beim Bund, Unterausschuss  
"Allgemeine Verwaltungsorganisation" des Arbeitskreises VI der  
Innenministerkonferenz, [http://www.inneres.sachsen-  
anhalt.de/min/r14/download/brosch\\_verwref.pdf](http://www.inneres.sachsen-anhalt.de/min/r14/download/brosch_verwref.pdf), S 6 ff.

Bleek, W.-G. (2001): Situations in Life to Support the Use and Modelling of Municipal  
Information Systems. In: Remenyi, D.; Bannister, F. (Hrsg.): Proceedings of the  
ECEG. Dublin: MCIL, S. 49–60.

Bleek, W.-G. :, (2002) Lebenslagen als Unterstützung bei der Benutzung und  
Modellierung von städtischen Portal-Webseiten, In: Herczeg, M.; Prinz, W.;  
Oberquelle, H. (Hrsg.): Mensch und Computer 2002. Stuttgart u.a.: B.G.  
Teubner, S.75-84., 2002

Jasmin Fischer: (2007) Support Vector Machines (SVM) Seminar "Statistische  
Lerntheorie und ihre Anwendungen",

[http://www.mathematik.uni-ulm.de/stochastik/lehre/ss07/seminar\\_sl/ausarbeitung\\_fischer.pdf](http://www.mathematik.uni-ulm.de/stochastik/lehre/ss07/seminar_sl/ausarbeitung_fischer.pdf)

# KnowledgeNavigator: Ontologiebasierte Datenintegration in der Umweltforschung

Matthias Faerber, Florent Jochaud, Stefan Jablonski<sup>7</sup>, Oliver Archner<sup>8</sup>  
 [{Matthias.Faerber, Florent.Jochaud, Stefan.Jablonski}@uni-bayreuth.de](mailto:{Matthias.Faerber, Florent.Jochaud, Stefan.Jablonski}@uni-bayreuth.de)  
[oliver.archner@bayceer.uni-bayreuth.de](mailto:oliver.archner@bayceer.uni-bayreuth.de)

## Zusammenfassung

Heterogenität in Daten, die aus dem Bereich der Umweltforschung kommen ist ein wichtiges und komplexes Problem. Daten werden in aller Regel aus einer Vielzahl an Sensoren und externen Systemen extrahiert. Obwohl sich diese Systeme auf die gleiche Domäne beziehen, ist das Datenschema jedoch häufig unterschiedlich. Die Integration dieser Daten sowohl in syntaktischer als auch semantischer Hinsicht ist eine wichtige Herausforderung. Dieser Beitrag stellt ein schlankes und flexibles Wissensmanagement System zur Lösung dieses Problems vor. Für jedes Forschungsgebiet im Bereich der Geoökologie können eigene Datenmodule entworfen werden, die anschließend in einer gemeinsamen OWL-basierten Datenbasis integriert werden. Der Beitrag beschreibt sowohl die Architektur als auch den Ansatz zur Integration der Daten mit Hilfe von Importschnittstellen. Über eine gemeinsame Nutzerschnittstelle können Anfragen einfach an das System gestellt werden und die ursprünglich heterogenen Daten abgefragt werden.

## Abstract

Heterogeneity in ecological data is an important and complex problem. Data usually comes from different sources and is expressed using different schemas (which can change over time) even if they are related to the same concept and describe the same object of interest. Integrating this data is then an interesting challenge as one must take into account the schema heterogeneity as well as the semantic heterogeneity. This paper introduces the knowledge management system “KnowledgeNavigator” which tries to solve this problem by providing a mechanism to integrate heterogeneous ecological data. The system is able to cope with syntactic and semantic mismatch between data sets. For each research area within ecology single modules are built and then integrated using an OWL based data store. In this paper, we will first introduce the architecture of the system and its components. We will then explain the external OWL-based ontologies we used and extended and demonstrate how we can model our data to fit this model. We will also present the different mechanisms we used to incorporate the different dataset in our knowledge base. Finally, we present the user interface which offers an integrated view of the originally heterogeneous data.

---

<sup>7</sup> Angewandte Informatik, Universität Bayreuth

<sup>8</sup> BayCEER, Universität Bayreuth

# 1 Einleitung

Wissenschaftliche Umweltforschung ist charakterisiert durch die Notwendigkeit umfangreiche, heterogene Daten automatisiert zu erheben und zu verarbeiten. Die typische experimentelle Feldforschung nutzt dabei sowohl strukturierte Datenbanken als auch unstrukturierte Dateien als Datenquelle.

Dank moderner Übertragungstechniken sind Sensornetzwerke zur Aufnahme von Routineparametern wie Bodentemperatur oder Niederschlag heute Stand der Technik. Die Sensoren sind in der Regel fest über mehrere Jahre an einem bestimmten Standort installiert und liefern Messwerte mit einer zeitlichen Auflösung von mehreren Hertz bis zu einer Stunde. Zur Speicherung der Messwerte kommen proprietäre Datenformate zum Einsatz, die über Konvertier- und Importprozesse in eine Datenbank übertragen werden.

Für speziellere fachliche Fragestellungen sind die Informationen aus dem Routinemessprogramm oftmals unzureichend. Wissenschaftler führen deshalb zusätzliche experimentelle Arbeiten im Feld durch. Dies kann z. B. die Messung des Brusthöhendurchmessers (BHD) oder die Aufnahme von Bodenprofilen in Kombination mit einer chemischen Analyse von Horizonten sein. Die Beobachtungen finden meist in Form von mehreren Messkampagnen verteilt über die Projektlaufzeit statt. Nach jeder Kampagne muss der Forscher Datenquellen integrieren um den Ziel der Arbeit näher zu kommen. So erfordert z. B. die Beantwortung der Frage "Wie wirken sich bodenchemischen Unterschiede als Einflussgröße auf das Wachstumsverhalten von Bäumen aus?" die wiederholte räumliche und zeitliche Integration von Datenbank- und Dateiquellen über mehrere Jahre, ehe signifikante Veränderungen nachgewiesen werden können. Derzeit realisiert der Forscher dies durch aufwendige manuelle Integration der Daten innerhalb seiner Analysesoftware. Ein standardisiertes übergreifendes Datenschema kommt dabei aus folgenden Gründen nur sehr selten zum Einsatz:

- ein übergreifendes Datenmodell ist nicht vorhanden oder nur schwer zu erweitern
- der Fokus liegt auf einem sehr speziellen Teilausschnitt
- mangelndes technisches Knowhow

Insgesamt führen diese Umstände zu einer sehr großen Anzahl an unstrukturierten dateibasierten Datenquellen mit hoher syntaktischer und semantischer Heterogenität. Deren Nachteile wie z. B. die mangelnde Reproduzierbarkeit von Ergebnissen verhindern den Aufbau einer gemeinsamen Wissensbasis. Unser Ziel ist es daher ein System zu entwickeln, das es den Forschern ermöglicht Daten kontextsensitiv zu erheben und in eine semantische Wissensdatenbank zu integrieren. Mit diesem Ansatz sollen verschiedene Teilziele erreicht werden:

- Forscher sollen weiterhin in der Lage sein, ihre Daten fokussiert auf ein bestimmtes Auswertungsziel zu erheben und zu speichern. Nur so kann ein effektives und effizientes Arbeiten gewährleistet werden.
- Nach der Übertragung der Daten in die Wissensdatenbank sollen die Daten über einen langen Zeitraum zur Verfügung gestellt werden.
- Dank der semantischen Annotation werden Daten so beschrieben, dass sie zu einem späteren Zeitpunkt einfach ausgewertet werden können.

Um dieses Ziel zu erreichen wurde das Wissensmanagement System KnowledgeNavigator entwickelt. Den Kern des Systems bildet dabei der Wissensspeicher, in dem Beobachtungsdaten abgelegt werden. In Kombination mit einer Ontologie für Umweltbeobachtungen ermöglicht es die semantische Suche und Integration von Umweltinformationen.

In dem weiteren Teil dieses Artikels versuchen wir das System detailliert zu beschreiben. Der folgende Abschnitt 2.1 erläutert den grundlegenden Aufbau des Wissensmanagement Systems und seine Architektur. Abschnitt 2.2 versucht einen groben Überblick über die eingesetzte Ontologie zu vermitteln, da diese fundamental für das Verständnis der im Abschnitt 2.3 beschriebenen Import Schnittstellen ist. Im Anschluss daran stellen wir im Abschnitt 2.4 die Benutzerschnittstelle vor und beschließen unsere Ausführungen mit einem Ausblick zu geplanten Erweiterungen.

## 2 Systemarchitektur

### 2.1 Überblick

Die Architektur des Systems baut auf Techniken des Semantic Web [Berners-Lee, 2001] auf. Zur Implementierung des zentralen Wissensspeichers wurde ein RDF Triple Store OpenRDF Sesame<sup>9</sup> gewählt. Dieser ist in der Lage OWL-basierte Ontologien zu speichern und zu verarbeiten. Ontologien bieten im Vergleich zu konventionellen Datenbankkonzepten verschiedene Vorteile, die in diesem System genutzt werden sollen.

- Flexibilität bei der Erweiterung um neue Konzepte
- Nutzung eines vordefinierten Standardmodells
- Einfacher Datenaustausch durch integrierte semantische Beschreibung

Die Architektur des Wissensnavigators besteht aus mehreren aufeinander aufbauenden Schichten. Diese sind in Abbildung 11 schematisch abgebildet:

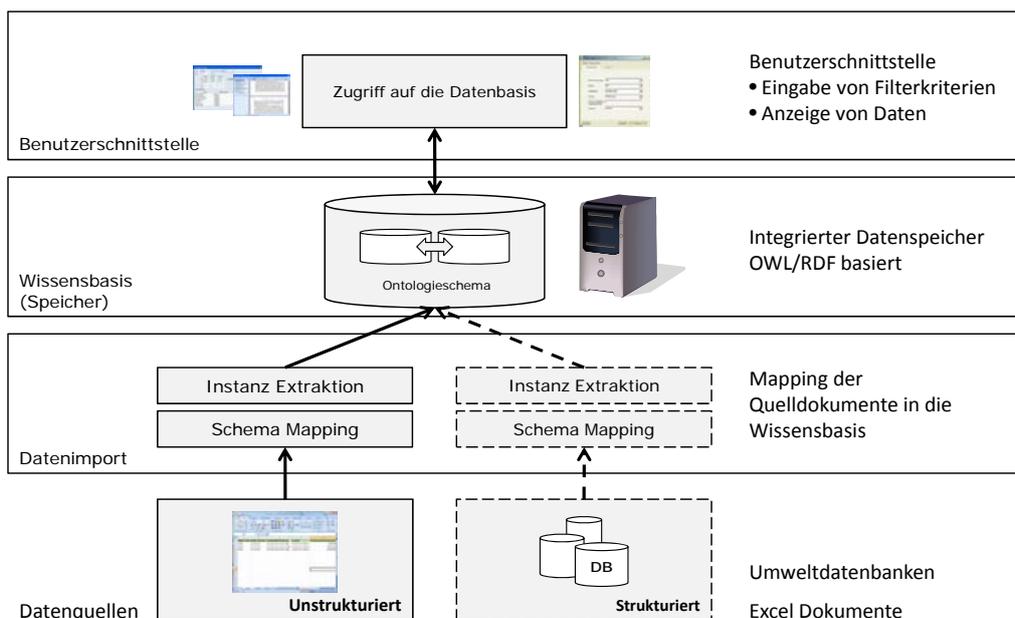


Abbildung 1: Architektur Knowledge Navigator

<sup>9</sup> OpenRDF Sesame: <http://openrdf.org>

Die Datenquellen bezeichnen die zu importierenden Umweltdaten und Dateien. An dieser Stelle unterscheiden wir zwischen strukturierten und unstrukturierten Datenquellen:

Strukturierte Datenquellen bieten eine Beschreibung der eigenen Daten an. Hier existiert zusätzlich zu den eigentlichen Daten ein Meta-Modell in dem die Struktur der gespeicherten Daten beschrieben ist. Dieses Meta-Modell ist in einem für IT-Systeme lesbaren Format gespeichert und kann automatisiert ausgewertet werden. Ein typisches Beispiel hierfür sind Datenbanken. Im ER-Schema ist gespeichert, welche Entitäten in einer Datenbank vorhanden sind und wie diese in Verbindung stehen.

Unstrukturierte Datenquellen bieten im Gegensatz zu den strukturierten Datenquellen keine Beschreibung der eigenen Datenstrukturen. Zwar müssen auch hier die Daten in einheitlicher Struktur abgelegt sein (z.B. einheitliche Spalten und Wertebereiche etwa in Excel oder CSV). Allerdings gibt es keine maschinenlesbare Beschreibung dieser Strukturen. Eine solche Beschreibung existiert häufig nur in Form von (Code-)Dokumentation und kann nicht automatisiert ausgewertet werden. Die in diesem Beitrag beschriebenen Datenquellen, also Excel Dateien in denen Umwelt-Messreihen beschrieben sind gehören in diese Klasse der unstrukturierten Datenquellen. Hier muss in einer externen Datei die Struktur der Excel-Dokumente beschrieben werden. Der Ansatz den wir gewählt haben wird in Abschnitt 2.3 (Datenimport und Integrationsschicht) dargestellt.

Die unterste Schicht des Systems bildet die Datenimportschicht. Sie ist dafür verantwortlich Daten aus den Datenquellen in die Wissensbasis zu importieren. Diese Schicht hat im Wesentlichen zwei Aufgaben. Sie muss einerseits das Mapping der beiden Datenschemas gewährleisten; andererseits aber auch die Instanzen (Tupel) aus der Quelldatei extrahieren und korrekt in das Ziel (Wissensbasis) integrieren. Analog zu den Datenquellen ist auch hier zwischen einem Mechanismus für strukturierte und unstrukturierte Datenquellen zu unterscheiden.

Bei strukturierten Datenquellen können die dort verfügbaren Meta-Modelle genutzt werden um das Schema-Mapping vorzubereiten. Hierzu wird aus dem vorhandenen Schema ein Ontologieschema erstellt, das anschließend noch in das Schema der Wissensbasis eingebunden werden muss. Obwohl Ansätze zum automatisierten Mapping der Modelle existieren (vgl. Modelmanagement [Bernstein, 2007]) ist hier häufig noch manuelle Nacharbeit zu leisten.

Bei unstrukturierten Datenquellen können keine vorhandenen Informationen genutzt werden und das Mapping des Schemas der Datenquellen an das der Wissensbasis muss komplett manuell erstellt werden. Dazu muss zunächst die Struktur der Datenquellen (in unserem Fall der Excel-Dokumente) beschrieben werden. Diese Beschreibung muss anschließend mit dem Schema des Ziels abgeglichen werden, so dass die Daten korrekt importiert werden können.

Mit Wissensspeicher bezeichnen wir die Schicht, in der die Daten gespeichert werden. Die Grundlage bildet ein ontologiebasierter Triplestore. Hier werden die Daten aus den verschiedenen Datenquellen in einem einheitlichen Datenschema gespeichert. Basis dieses Datenschemas ist die „Observation-Ontology“. Diese wurde mit dem Ziel entwickelt Messwerte aus wissenschaftlichen Umweltbeobachtungen in einem integrierten Datenschema speichern zu können. Im folgenden Abschnitt gehen wir näher auf dieses Datenschema ein und beschreiben, wie dieses genutzt wird um die Daten aus den Messungen darin zu speichern.

Die oberste Schicht des KnowledgeNavigators ist Schnittstelle für die Nutzer. Eine webbasierte Oberfläche bereitet die Daten zur Anzeige für den Benutzer auf. Mittels vordefinierter oder freier Abfragen können diese auf alle in der Wissensbasis vorhandenen Daten einfach zugreifen und diese Auswerten.

## 2.2 Wissensbasis (Observation Ontologie)

Das globale Schema unserer Wissensbasis basiert auf der aus dem SEEK<sup>10</sup> Project hervorgegangene Extensible Observation Ontologie (OBOE) [Madin, 2007]. Im Gegensatz zu reinen Standards zur Daten-Schemata wie z. B. ABCD<sup>11</sup> oder EML<sup>12</sup> [Arndt, 2000] ist sie in der Lage auch die Semantik der Daten zu erfassen. Sie bedient sich eines klaren schematischen Aufbaus, bestehend aus den Klassen Observation, Entity, Measurement, Characteristic und MeasurementStandard.

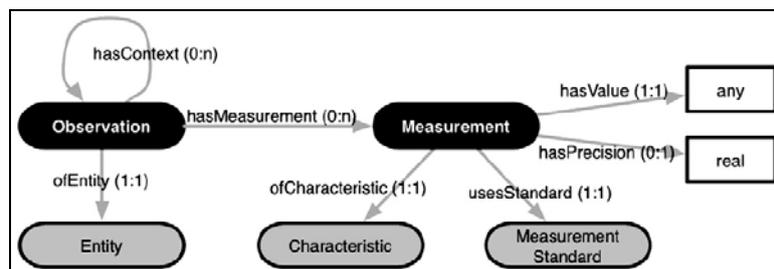


Abbildung 2: Hauptklassen und Relationen der OBOE Ontologie [Madin, 2007]

In OBOE ist eine Umweltbeobachtung (engl.: Observation) genau einem unter Beobachtung stehenden Objekt (engl.: Entity) zugeordnet. Als Objekte kommen hier sowohl materielle als auch nicht materielle Objekte in Frage. Die Beobachtung kann zudem im Kontext zu anderen Beobachtungen stehen, beispielsweise bei einem zeitlichen oder räumlichen Zusammenhang. Um Daten einer Beobachtung zu erhalten wird eine Messung (engl.: Measurement) einer Eigenschaft (engl. Characteristic) des Objekts mit einem bestimmten Standard (engl. MeasurementStandard) durchgeführt. Liefert die Messung Interval- oder Ratiozahlen verbirgt sich hinter dem Standard die Einheit der Messung (z. B. Meter oder Gramm). Eine mögliche Übersetzung des engl. Begriffs „Characteristic“ wäre die im deutschen Sprachgebrauch übliche Bezeichnung „Parameter“ oder „Messgröße“. Der eigentliche Messwert wird zusammen mit der Genauigkeit als primitiver Datentyp der Messung zugeordnet. In Abbildung 2 ist der beschriebene Ausschnitt aus der Struktur der OBOE in einem Diagramm dargestellt.

Abbildung 3 zeigt die Anwendung der Ontologie am Beispiel der Messung des Brusthöhendurchmessers eines Baumes. Der Beobachtung eines Baumobjekts (:Stem) wird eine Messung (:Measurement) zugeordnet. Der Doppelpunkt vor dem Namen zeigt an, dass es sich hierbei um eine Instanz der jeweiligen Klasse handelt. Die Messung des Parameters (:BreastHeightDiameter) wurde in der Längeneinheit Meter aufgenommen und ergab einen Messwert von 0.52 bei einer Genauigkeit von 0.01 Metern.

<sup>10</sup> Science Environment for Ecological Knowledge, SEEK, <http://seek.ecoinformatics.org/>

<sup>11</sup> Access to Biological Collections Data , ABCD, <http://www.bgbm.org/TDWG/CODATA/Schema/>

<sup>12</sup> Environmental Markup Language, EML, <http://www.xml-eml.org/>

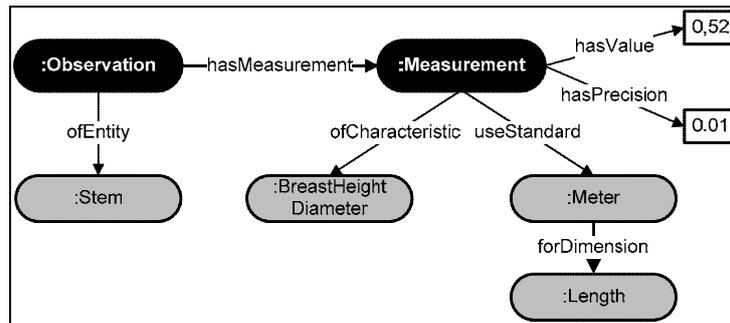


Abbildung 3: Beschreibung einer Messung des Brusthöhendurchmessers eines Baumes in der OBOE Ontologie.

Eine positive Eigenschaft von OBOE ist die leichte Erweiterbarkeit der Ontologie. Basierend auf einem sehr umfangreichen Gerüst aus bereits vorhandenen Characteristics und MeasurementStandards können eigene Klassen die bestehende Ontologie erweitern. Das gilt auch für die Klasse Entity die im ursprünglichen OBOE Standard lediglich als Erweiterungspunkt vorgesehen ist. In der untenstehenden Abbildung 4 zeigen wir dies exemplarisch anhand der Umsetzung eines Ausschnitts der Bodenkundlichen Kartieranleitung für Deutschland [Sponagel, 2005]. Entitäten werden dabei über eine transitive „is-A“ Beziehungen verbunden. Zur Verfeinerung der Hierarchie lassen sich sämtliche in OWL-DL zur Verfügung stehenden sprachlichen Mittel einsetzen. So lässt sich über eine neue ObjektProperty „composedOf“ in Kombination mit der Existential Restriction (some) ausdrücken, dass sich Boden aus biotischen und abiotischen Objekten zusammensetzt und ein Bodenhorizont natürlich auch Boden enthalten muss.

*Soil* ≡ (composedOf some BioticEntity) and (composedOf some AbioticEntity)

*SoilHorizon* ≡ composedOf some Soil

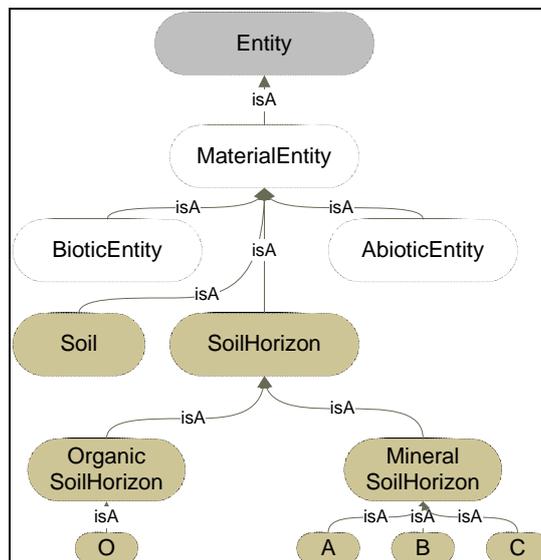


Abbildung 4: Erweiterung der Klasse Entity für eine Klassifikation von Bodenhorizonten in Anlehnung an die Bodenkundliche Kartieranleitung für Deutschland [Sponagel, 2005].

## 2.3 Datenimport und Integrationsschicht

Die Integrationsschicht ist wie schon beschrieben dafür zuständig die Daten aus den Datenquellen in die Wissensbasis abzubilden. Ziel bei der Entwicklung des Systems war es einen möglichst generischen Ansatz zu Datenintegration zu finden. Generisch bedeutet in diesem Zusammenhang, dass die Lösung durch einfache Parametrisierung an verschiedene Datenschemata angepasst werden kann und aufwändige Mehrfachentwicklungen von Integrationsmodulen entfallen können.

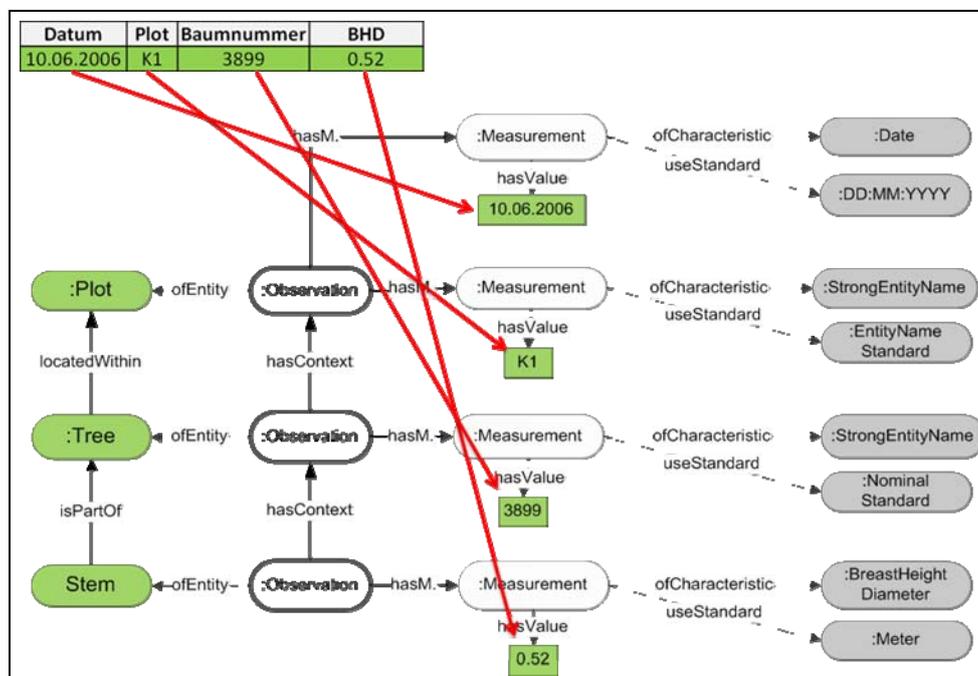


Abbildung 5: Abbildung der Tupel aus der Excel-Datei in die Ontologie

In Abbildung 5 ist die grundsätzliche Idee zur Integration der Daten aus den Excel-Dokumenten in die Datenstruktur der Wissensbasis dargestellt. In der Abbildung 5 ist oben die Excel-Tabelle mit den Attributen Datum, Plot, Baumnummer und BHD (Breast Height Diameter) eingezeichnet. Der Zeileneintrag in der Tabelle beschreibt die Messung, die am 10.06.2006 im Plot K1 am Baum 2889 durchgeführt wurde und einen Umfang von 0,52 ergab. Die verschiedenen Werte dieses Eintrags müssen dann den entsprechenden Stellen in der Ontologie zugeordnet werden. Diese Zuordnung ist in der Darstellung durch Pfeile zwischen den Einträgen in der Tabelle und den Knoten der Ontologie abgebildet.

Wie schon zu Beginn dieses Abschnitts beschrieben, war das Ziel eine Softwarearchitektur zu finden, die sich durch Parametrisierung so anpassen lässt, dass das in Abbildung 6 beschriebene Mapping zwischen Datendatei und Ontologie einfach realisiert werden kann. Mit dem Framework RDF123 [Han, 2007] steht eine Möglichkeit zur Verfügung, wie dieses Mapping zwischen Datei und Ontologie deklarativ umgesetzt werden kann. In Abbildung 6 ist dieses Mapping dargestellt. Auf der linken Seite sind die Entitäten Stamm (local:Stem), Baum (Ex:Baum) und Plot (Ex:Plot) dargestellt. Diesen werden direkt die Attribute aus der Datei zugeordnet. Diese Zuordnung wird durch die an Knoten und Label angegebenen String „Ex:Baum+\$3“ realisiert. Der Teil „\$3“ bezeichnet dabei das dritte Attribut der Tabelle. Für die anderen Klassen der Ontologie wird die Zuordnung analog vorgenommen.

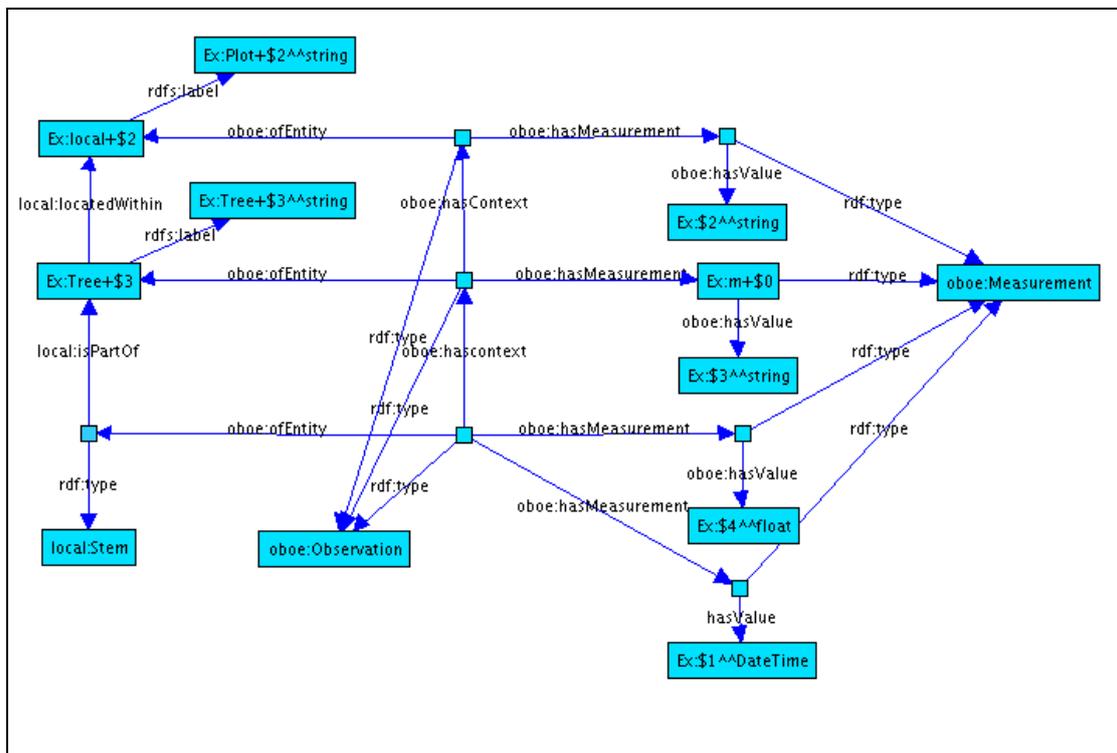


Abbildung 6: Mapping-Datei

## 2.4 Zugriffsschicht

Die Benutzerschnittstelle des KnowledgeNavigators wurde als Webseite realisiert. Auf diese Weise können den Mitarbeitern und Forschern einfach die Inhalte der Wissensbasis zugänglich gemacht werden. In Abbildung 7 ist das Ergebnis einer Suchabfrage an die Wissensbasis dargestellt nachdem der Nutzer einen Baum aus der Datenbasis ausgewählt hat. Zu einem Baum werden die verschiedenen Zeitpunkte, zu denen Messungen erstellt wurden, dargestellt.

Knowledge Navigator

Tree: 3899

Date	BHD	
10.06.2006	0,52	open Soil
15.01.2007	0,54	open Soil

Plot: K1

Layer	Elevation	Ca <sup>2+</sup> [mmol /kg]	K <sup>+</sup> [mmol /kg]	Mg <sup>2+</sup> [mmol /kg]
Oh	5	53,87	2,09	3,11
Ah	-11	41,90	2,35	2,40
Bh	-15	52,57	0,87	1,45
Bs	-31	24,15	0,63	0,69
Bv	-59	4,41	0,69	0,17
Cv		2,98	1,21	0,23

other Trees in the Plot  
Treenummer: 3067

Lehrstuhl für Angewandte Informatik IV -- Florent Jochaud / Matthias Faerber

Abbildung 7: Benutzerschnittstelle des Knowledge Navigators

Die Inhalte der Webseite werden über Anfragen an die Wissensbasis erzeugt. Dazu wird auf die Sprache SparQL [W3C, 2008] zurückgegriffen. Diese ermöglicht es ähnlich zu SQL im Bereich der Relationalen Datenbanken Inhalte aus RDF-basierten Wissensspeichern abzufragen. Der KnowledgeNavigator stellt eine Reihe von vorkonfigurierten Abfragen an die Wissensbasis bereit. Somit ist es für die einzelnen Benutzer nicht nötig die Anfragesprache zu erlernen. Die verschiedenen Anfragemasken (wie die in Abbildung 7 dargestellte) lassen sich von einem Administrator ohne großen Aufwand einfach in das System integrieren. Für spezielle Anfragen und für Nutzer mit Kenntnissen in SparQL wird ein generisches Interface zur Verfügung gestellt, in dem sich Anfragen absetzen lassen. Die Ergebnisse werden dann in einer Tabelle dargestellt.

### 3 Fazit und Ausblick

In diesem Artikel haben wir gezeigt, wie die von uns entwickelte Software KnowledgeNavigator zum Aufbau einer Wissensbasis im Bereich der Umweltbeobachtung eingesetzt werden kann. Auf Basis der Framework Ontologie OBOE in Kombination mit dem Tool RDF123 ist es uns gelungen dateibasierte Quellen in semantischer als auch in syntaktischer Hinsicht zu integrieren. Aus den Erfahrungen unserer Arbeit beim Aufbau von Umweltdatenbanken sind es gerade diese Informationen, die in der Vergangenheit oftmals nur sehr schwer erschlossen werden konnten.

In Zukunft möchten wir in unser System auch Datenbanken als Datenquellen aufnehmen. Wir arbeiten deshalb gerade daran die am BayCEER vorhandene Umweltbeobachtungsdatenbank mit in die Wissensbasis zu integrieren und im RDF Format zu publizieren. Von diesem Schritt versprechen wir uns eine Aufwertung der Attraktivität der Wissensbasis für unsere Nutzer. Als Ergänzung des Systems planen wir zusätzlich zu der hier beschriebenen Web-Oberfläche eine Zugriffsschicht für externe Programme wie Statistikprogramme (z. B. „R<sup>13</sup>“) oder Geographische Informationssysteme (z. B. ArcGIS<sup>14</sup>). Auf diese Weise sollen den Nutzern neben den in der Web-Oberfläche fest vorgegebenen Abfragen eine Möglichkeit gegeben werden, freie Anfragen zu stellen und diese in die Auswertungstools einzubinden.

### 4 Literaturverzeichnis

- Madin, Joshua; et al.: An ontology for describing and synthesizing ecological observation data, *Int. J. Ecol. Informatics* 2 (2007), pp. 279–296.
- Arndt, Hans-Knud; Christ, Mario: XML als Metasprache zum Entwurf einer standardisierten Beschreibung von Umweltdaten, in Arndt, H. (ed.), *Environmental markup language (EML)*, Metropolis-Verl., 2000 .
- Han, Lushan; et. al: *RDF123: a mechanism to transform spreadsheets to RDF*, University of Maryland, Baltimore County, 2007.
- Sponagel, Herbert (ed.): *Bodenkundliche Kartieranleitung mit ... 103 Tabellen und 31 Listen*, Schweizerbart, 2005.
- Bernstein, Philip A.; Melnik, Sergey: *Model management 2.0: manipulating richer mappings*. Proceedings of the 2007 ACM SIGMOD international conference on Management of data. Beijing, China, 2007.

---

<sup>13</sup> The R Project for Statistical Computing: <http://www.r-project.org/index.html>

<sup>14</sup> ArcGIS, <http://www.esri.com/software/arcgis/index.html>

W3C: SPARQL Query Language for RDF, Seaborne, Andy; Prud'hommeaux, Eric, Editors, W3C Recommendation, 15 January 2008, <http://www.w3.org/TR/rdf-sparql-query/>  
Berners-Lee, Tim; Hendler, Jim; Lassila, Ora: The Semantic Web, Scientific American, 2001.

# Ein Umweltziel- und Kennzahlensystem für die Umweltverwaltung auf der Basis von Topic Maps

Hans-Knud Arndt, Stephan Jacob und Henner Graubitz<sup>15</sup>  
[hans-knud.arndt@iti.cs.uni-magdeburg.de](mailto:hans-knud.arndt@iti.cs.uni-magdeburg.de)

## Abstract

Given strategies, e.g. for environmental policies or environmental goals are one of the most successful tools to realize sustainable environmental protection. For organizations it is possible to coordinate internal and external relationships resulting in reactions onto these influences; for public environmental departments it is a perfect verification which environmental precautions should be followed and which costs for planning, doing and checking will be accrued.

But beside these facts in the real world there still exist non-clear environmental oriented strategies or miss totally. With this paper we represent an idea of how to use the fundamentals of the management instrument "Balanced Scorecards" and one of the semantic network standards ISO/IEC 13250 for Topic Maps and derive strategies and goals from only modeled general conditions.

## 1 Steuerung von Organisationen mit Strategien und Zielen

Strategien und Ziele haben eine hohe Bedeutung im Managementprozess einer Organisation. Eine Organisation besteht im Inneren aus Beziehungen von sozialen und technischen Komponenten, worin die Bezeichnung sozio-technisches System begründet ist. Darüber hinaus ist eine Organisation gekennzeichnet durch eine Vielzahl von Beziehungen mit seiner äußeren Umwelt, die in folgende Gruppen eingeteilt werden (Steinmann/Schreyögg 2005):

- makro-ökonomische Umwelt,
- technologische Umwelt,
- politisch-rechtliche Umwelt,
- sozio-kulturelle Umwelt,
- natürliche Umwelt.

Der Erfolg der Organisation hängt stark davon ab, in welchem Maße die internen und externen Beziehungen koordiniert und beherrscht werden. Einen Beitrag zur Koordination und Beherrschung von Managementprozessen leisten die sogenannten Managementsysteme. Managementsysteme haben sich vor allen in Problembereichen von Organisationen etabliert wie z.B. Qualität, Umweltschutz, Arbeitsschutz, Risiken etc. und können als ein Prozess der kontinuierlichen Verbesserung aufgefasst werden, welcher aus folgenden Systemelementen und Schritten besteht:

---

<sup>15</sup> Otto-von-Guericke-Universität Magdeburg, AG Wirtschaftsinformatik – Managementinformationssysteme –, Institut für Technische und Betriebliche Informationssysteme, Universitätsplatz 2, D-39106 Magdeburg, Germany, [http://www.iti.cs.uni-magdeburg.de/iti\\_mis](http://www.iti.cs.uni-magdeburg.de/iti_mis)

- Politik (Vision): Die Festlegung der Politik der Organisation hat durch die oberste Leitung zu geschehen. Aus der Politik (Vision) können umsetzungsbezogene Strategien als Grundlage der Planung abgeleitet werden.
- Planung: Die Planung umfasst Verfahren zur Ermittlung der organisationsspezifischen Aspekte, Gesetze und anderer Forderungen, die Festlegung von Zielsetzungen und Einzelzielen der Organisation im Einklang mit der Politik sowie die Erstellung eines oder mehrerer Managementprogramme zur Verwirklichung dieser Ziele.
- Implementierung und Durchführung: Die Implementierung und Durchführung setzt sich aus der Festlegung von Organisationsstruktur und Verantwortung, der Einführung von Verfahren zur Schulung, Bewusstseins- und Kompetenzbildung, der Aufrechterhaltung der Kommunikation, der Dokumentation des Managementsystems, Verfahren zur Lenkung der Dokumente und des Ablaufs sowie Anweisungen zur Notfallvorsorge und Maßnahmenplanung zusammen.
- Kontroll- und Korrekturmaßnahmen: Das Systemelement „Kontroll- und Korrekturmaßnahmen“ dient der Überwachung und Messung umweltbezogener Tätigkeiten, der Maßnahmenplanung zur Behandlung von Abweichungen, zur Einleitung von Korrekturen und der Vorsorge, der Einführung von Verfahren zur Dokumentation sowie der Durchführung eines Auditing des Managementsystems.
- Bewertung durch die oberste Leitung: Die Bewertung durch die oberste Leitung der Organisation stellt in bestimmten Abständen die Angemessenheit und Wirksamkeit des Managementsystems sicher.

Die Ablauffolge eines Managementsystems entspricht dem herkömmlichen Kreislauf der betriebswirtschaftlichen Planung bzw. des Controlling.

In vielen Umweltverwaltungen bzw. Organisationen jeglicher Art existieren oftmals unklare Strategievorgaben bzw. fehlen gänzlich, z.B. aufgrund von:

- Mangelnder Bereitschaft der obersten Leitungsebene sich bezüglich des Umweltschutzes konkret festzulegen (auch aus politischen bzw. taktischen Gründen);
- Unzureichendem Wissen der obersten Leitung bezüglich des Umweltschutzes;
- Unkenntnis, etc.

Auch ohne diese Vorgaben kommt gerade den mittleren und unteren Führungsebenen die Aufgabe zu, Umweltschutz erfolgreich zu betreiben. Durch das Fehlen einer konkreten Strategie existiert kein geeignetes Mittel zur Bewertung verschiedener Handlungen. So ist auch die Wahl zwischen verschiedenen Handlungsalternativen nur schwer begründbar. Die Strategie gibt eine Ausrichtung der Organisation vor. Geplante bzw. umgesetzte Handlungen können mit Hilfe dieser Ausrichtung beurteilt werden und ihr Beitrag zum Erfolg ist ableitbar.

Ein verbreitetes Managementinstrument (auch im Umweltschutz) zur Umsetzung von Strategien und Zielen mit Hilfe von Kennzahlen stellt die sogenannte Balanced Scorecard dar.

## 2 Die Balanced Scorecard als Managementinstrument

Mit der Balanced Scorecard (BSC) stellen Kaplan und Norton ein Managementinstrument zur Verfügung, welches die Gestaltung des Planungs-, Steuerungs- und Kontrollprozesses in einer Organisation unterstützt (Kaplan/Norton 1997). Hervorgegangen aus dem ZVEI Leistungsindikatorensystem bildet die BSC neben finanziellen Steuerungsgrößen, welche vergangenheitsorientiert sind (Wagner 2002) auch Messgrößen ab, welche zukünftig angestrebte Leistungen steuerbar machen.

Die Hauptbestandteile der BSC bilden die insgesamt vier Perspektiven (Kaplan/Norton 1997 und Wagner 2002):

- **Finanzielle Perspektive:** Finanzielle Kenngrößen dienen der Darstellung der vergangenen finanziellen Leistung einer Organisation, wie z.B. dem Return on Investment. Schwieriger gestaltet sich diese Perspektive für nicht-gewinnorientierte Organisationen wie z.B. eine öffentliche Verwaltung. Doch auch sie können monetäre Kennzahlen zur Ausrichtung der Strategie nutzen.
- **Kundenperspektive:** Sie beinhaltet alle Indikatoren, die die Beziehungen zwischen einer Organisation und deren Kunden beschreiben. Kunden- und Marktsegmente werden identifiziert. Die Tätigkeiten der Organisation werden durch Indikatoren, wie z.B. Kundenzufriedenheit, Kundentreue oder den Gewinn- und Marktanteil in einem Segment beurteilt.
- **Interne Geschäftsprozessperspektive:** Sie gibt einen Überblick über alle kritischen Prozesse, in denen Verbesserungsschwerpunkte gesetzt werden können. Einerseits ermöglicht diese Perspektive den Unternehmen, Kundenanforderungen zu erfüllen und somit die Kundentreue zu erhöhen, andererseits können Erwartungen von Anteilseignern bezüglich der finanziellen Gewinne erhöht werden. Ein besonderes Augenmerk gilt den Prozessen, die den größten Einfluss auf die Kundenzufriedenheit haben.
- **Innovationsperspektive:** Diese vierte Perspektive misst Kennzahlen, die die Lernkurve eines Unternehmens und somit zum Wachstum und zur Verbesserung der Organisation beiträgt. Mit zunehmendem Wettbewerb ist es für eine Organisation wichtig, seine Kernkompetenzen kontinuierlich weiterzuentwickeln.

Grundsätzlich ist das Konzept der Balanced Scorecard nicht nur in am Markt tätigen Organisationen einsetzbar, sondern auch in öffentlichen Verwaltungen (vgl. Krems 2007). Dabei ist der ursprüngliche Aufbau der Scorecard, welcher die finanziellen Ziele als das Ende der Ursache-Wirkungskette zwischen den verschiedenen Zielen der Perspektiven sieht, an eine Verwaltung anzupassen. Für eine Verwaltung sind nicht die finanziellen Erfolge wichtig. Diese Organisationen sehen ihr Ziel in der Schaffung von Werten, welche für die Gesellschaft von Nutzen sind und somit die Wohlfahrt erhöhen (Wellisch 2000). Zwar müssen Verwaltungen aufgrund knapper Ressourcen Wirtschaftlichkeitskriterien einhalten, aber ihr eigentliches Ziel ist es, ihren Leistungsauftrag zu erfüllen (König/Rehling 2002).

Die Balanced Scorecard (BSC) ist aber nur bedingt in der Lage, in diesem Szenario (bei fehlender Strategievorgabe) geeignete Handlungen zu identifizieren. Ungeachtet fehlender Strategievorgaben besteht weiterhin die Notwendigkeit, verschiedene Aktionen zu planen und Alternativen zu beurteilen. Aus diesem Grund ist das Ziel dieser Arbeit, ein Konzept auf Grundlage des Topic Maps Standards (ISO/IEC 2002)

anzuzeigen, welches auch ohne explizite Strategievorgabe und langfristige Zielformulierung Wirkungszusammenhänge bezüglich des Umweltschutzhandelns darstellt.

### 3 Der Topic Map Standard

Das Konzept der Topic Map basiert auf der Idee der Davenport Group (Pepper 2000), semantische Netze grafisch als Modell darzustellen (Widhalm/Mück 2002). Ein Modell beschreibt ein abstraktes, immaterielles Abbild realer Strukturen bzw. realen Verhaltens für die Zwecke des Subjektes (dem Nutzer des Modells). Es bildet somit ein adäquates, vereinfachendes und idealisierendes Abbild der Realität (Rosemann 1996).

Topic Maps wurden primär zur Verarbeitung von Wissen in Indexen, Glossaren und ähnlichen Verzeichnissen entwickelt (Smolink 2005) und im Jahre 1999 von der International Standard Organization (ISO) als Standard 13250:2000 verabschiedet. Im Jahr 2001 wurde der Standard um eine Version erweitert, welche auf der eXtensible Markup Language (XML) basiert. Topic Maps sind demnach in der Lage, die mitunter komplexen Strukturen einer Wissensbasis abzubilden und somit eine Verarbeitung dieser zu erleichtern. Im Gegensatz zu klassischen semantischen Netzwerken, bei denen die Bezeichnungen der Kanten und das damit verbundene Wissen im Vordergrund stehen, konzentriert sich das Konzept der Topic Maps auf die Navigation durch das Netzwerk (Freese 2000).

Der ISO Standard spezifiziert folgende Elemente einer Topic Map (vgl. Widhalm/Mück 2002):

- **Topics:** Topics sind die zentralen Elemente einer Topic Map. Diese bilden alle Subjekte der Realwelt ab (vgl. Smolink 2005). Demnach repräsentieren sie eine Entität im modellierten Kontext (vgl. Widhalm/Mück 2002). Ein Topic ist grundsätzlich ein Stellvertreter für ein Subjekt (vgl. XTM 2001). Das bedeutet, dass das eigentliche Subjekt nicht in die Topic Map kopiert wird. In einer konsistenten Topic Map wird jedes Subjekt durch genau ein Topic repräsentiert. Existieren in einer Topic Map mehrere Topics, die das gleiche Subjekt referenzieren, können diese zu einem Topic verschmolzen werden (XTM 2001). Weiterhin besitzt jedes Topic Eigenschaften, zu denen die Namen der Topics (Topic Names), relevante Informationsobjekte der Topics (Topic Occurrences) und die Rollen der Topics in Beziehungen mit anderen Topics einnimmt (Associations).
- **Topic Names:** Jedem Topic wird ein bestimmter Name zugeordnet, der als Grundform im Topic als `baseName` beschrieben wird (XTM 2001). Weiterhin erlaubt die XML-Version des Topic Map Standards beliebig viele Arten von Namensvarianten, wohingegen im Original nur ein „display name“ (zur Darstellung der Topics) und ein „sort name“ (um Topics zu sortieren) angegeben werden kann.
- **Topic Occurrences:** Sie beschreiben die Verbindungen der einzelnen Topics und deren dazugehörige Informationsobjekte und werden ebenfalls als Topics in der Topic Map modelliert. Dazu gehören nicht zwangsläufig nur Informationsobjekte, die in einer vorliegenden Topic Map repräsentiert werden. Mittels eines Uniform Resource Identifier (URI) können auch externe Quellen (Bücher, Artikel, Bilder, Video- oder Audioeinspielungen) im Topic mit eingebunden werden.

- **Subject Descriptor:** Es ist möglich, dass zwei unterschiedliche Topic Maps dasselbe Topic mit unterschiedlichen Informationen enthalten. Zur eindeutigen Identifikation existiert in der ISO 13250:2000 der Subject Descriptor. Er veranlasst eine Topic Map dazu, zwei gleiche Topics mit unterschiedlichen Informationen zu einem Topic zusammenwachsen zu lassen.
- **Associations:** Sie können als multidirektionale Relationen einzelner Topics beschrieben werden, wobei die Relationen symmetrisch, transitiv oder reflexiv sein können (Widhalm/Mück 2002). Auch Associations werden in Topics beschrieben und können Klassenhierarchien und einen Typ beinhalten, durch den die Möglichkeit ausgeschöpft werden kann, Associations zu klassifizieren.
- **Scopes:** Topics mit einem bestimmten Namen können unterschiedliche Bedeutung haben. Zur Vermeidung dieser Homonymie-Probleme wurde das Konzept der Scopes eingeführt. Scopes beschreiben, in welchem Kontext die Eigenschaften eines Topics gültig sind (Pepper/Gronmo 2002). Existiert in einem Topic kein Scope so drückt das Topic eine Allgemeingültigkeit aus. Weiterhin bieten sie auch die Möglichkeit, die Navigation innerhalb des Topic Map Dokumentes zu erleichtern (Pepper 2002).

Der im Folgenden vorgestellte Ansatz erfasst zunächst in einer Topic Map die Rahmenbedingungen sowie die Auswirkungen von Handlungen. Dieses erlaubt wiederum, Aussagen über den Beitrag der Aktionen zum Erfolg von Organisationen zu treffen. In diesem Zusammenhang wird die Wahl von Handlungsalternativen unterstützt.

## **4 Ein Ansatz zur Steuerung von Umweltschutzaufgaben bei Strategie- und Zieldefekten**

### **4.1 Das Konzept der Topic Map-basierten Balanced Scorecard**

Der vorliegende Ansatz hat zum Ziel, auch ohne konkrete Strategievorgaben mit Hilfe eines semantischen Netzes die Wirkungszusammenhänge von Zielen und Kennzahlen untereinander darzustellen. Dies erleichtert das Abschätzen von Auswirkungen der Änderungen der Kennzahlen oder Ziele auf andere Größen. Dadurch wird es im begrenzten Maße möglich, eine Handlungsalternative zu bewerten. Allerdings kann das System die Beurteilung der Aktionen im Sinne einer Strategie nicht vornehmen.

Ziele und Kennzahlen in einem System stehen untereinander in Beziehung und beeinflussen sich somit. Durch die Änderung eines dieser Elemente durch eine Aktion treten Änderungen an anderen Elementen auf. Dadurch sind die Auswirkungen messbar. Zur Bewertung verschiedener Alternativen können die Auswirkungen dieser Aktionen durch das semantische Netz ermittelt werden. Zur endgültigen Bewertung einer Alternative bedarf es aber einer Vorstellung, welche Änderungen erwünscht sind. Ist diese Vorstellung nicht gegeben, können nur implizite Faktoren zur Festlegung der Präferenz der Alternativen herangezogen werden.

In einer Topic Map können die Kennzahlen als ein Topic dargestellt werden, der mit dem Namen einer Kennzahl beschriftet wird und somit auch die Bedeutung der Kennzahl wiedergibt.

Dabei sind eng mit den Kennzahlen die Ziele verknüpft. Die Kennzahlen beschreiben die aktuelle Ausprägung eines Ziels. Somit repräsentieren Kennzahlen oftmals

(zumindest implizit) ein Ziel. Es ist somit sinnvoll, auch Ziele als Topic darzustellen. Weiterhin ist eine Verknüpfung der Ziele mit den entsprechenden Kennzahlen wünschenswert.

Neben der Zuordnung von Zielwerten zu Zielen ist eine Verknüpfung der Ziele mit Kategorie-Topics sinnvoll. Somit besteht die Möglichkeit der Ablesung einer Hierarchieebene aus der Topic Map. So können sie in z.B. finanzielle Ziele, Kundenziele oder andere Gruppen gegliedert oder Kategorien z.B. für qualitative oder quantitative Ziele vergeben werden. Quantitative Kennzahlen können mit als Topics dargestellten mathematischen Operatoren verbunden werden.

Zur eindeutigen Identifikation der Verknüpfung zwischen einem Ziel und den entsprechenden Kennzahlen sollte ein bestimmter Assoziationstyp verwendet werden. So können sich Ziele in die Ausprägungen Identität, Neutralität, Komplementarität, Zielkonkurrenz und Zielantinomie einteilen. Die Verwendung des Ausprägungstyps als Assoziationstyp ist daher naheliegend.

Gleichzeitig besteht auch die Möglichkeit, Relationen zwischen Kennzahlen auszudrücken. Die Beziehungen zwischen diesen Maßzahlen werden ähnlich wie im Scorecard-Maps-Ansatz (Arndt et. al 2007) modelliert. Durch die Verknüpfung von Kennzahlen mit mathematischen Operatoren besteht die Möglichkeit, neue Kennzahlen zu errechnen. Dazu wird eine Association in der Topic Map abgebildet, welche die Eingabekennzahlen, den mathematischen Operator und die Ausgabekennzahl enthält. Die Darstellung des Operators erfolgt dabei analog als ein Topic. In der eigentlichen Beziehung wird eine Instanz dieses Operators eingebunden.

Neben der allgemeinen Darstellung von verschiedenen Beziehungstypen ist auch die Angabe einer Beziehungsstärke wünschenswert. Mit deren Hilfe ist es beispielsweise möglich, die Beziehung zwischen zwei Zielen nicht nur als konfliktär zu klassifizieren, vielmehr kann die Stärke der Hemmung des einen Ziels durch die der anderen ausgedrückt werden. Im Rahmen der Darstellung von Kennzahlen besteht die Möglichkeit, eine Relation zwischen diesen Messgrößen allgemein als positiv oder negativ zu klassifizieren. Hierbei ist zusätzlich die Modellierung der Stärke der Beeinflussung einer Kennzahl umsetzbar.

Dabei erfolgt zunächst die Deklaration von Topics, welche die Art der Beziehungsstärke ausdrücken. Erfolgt die Angabe der Stärke zwischen zwei Kennzahlen in Prozent so wird ein Topic, welches den Prozent-Typ beschreibt, modelliert. Die Verknüpfung zwischen den beiden Kennzahlen wird anschließend um ein Topic, das eine Instanz des Prozent-Typs ist und als baseName den Prozentwert enthält, erweitert.

In einem Zielsystem besteht auch die Möglichkeit, dass Ziele verschiedener Aggregationsebenen koexistieren (z.B. Kosten und Personalkosten). Diese Aggregationsebenen drücken einen Detaillierungsgrad der inhaltlichen Bedeutung des Ziels bzw. der Kennzahl aus. Mit Hilfe der Aggregationsebenen ist es möglich, die Komplexität der Darstellung des Systems zu reduzieren. Werden nur die Elemente mit einer hohen Aggregationsebene angezeigt, verringert sich die Anzahl der darzustellenden Elemente im System, da die untergeordneten Teilmengen dieser Elemente nicht visualisiert werden.

Die Umsetzung der Aggregationsebenen eines Ziels oder einer Kennzahl erfolgt über den Topic Map-Mechanismus der Scopes. Jedem Topic wird ein Scope mit der Aggregationsebene zugeordnet. Die Modellierung dieser Ebene erfolgt wiederum als Topic in der XTM. Mit Hilfe der verschiedenen Scopes besteht die Möglichkeit

unterschiedlichen Nutzerinteressen innerhalb des semantischen Systems gerecht zu werden. Je nach Betrachtungsweise des Nutzers kann dieser den Detaillierungsgrad der Anzeige verändern oder auch von einer bestimmten Nutzersicht komplett ausgeschlossen werden.

## **4.2 Eine Topic Map-basierte Balanced Scorecard am Beispiel der Umweltverwaltung**

Auch wenn keine expliziten umweltschutzbezogenen Strategie- oder Zielvorgaben in einer Umweltverwaltung vorliegen, existieren Rahmenbedingungen, welche die (möglichen) Handlungsalternativen einschränken. Diese müssen bei der Modellierung eines Ziel- bzw. Kennzahlensystems beachtet werden. Die Rahmenbedingungen sind unter anderem Naturgesetze, welche nicht beeinflusst werden können. Darüber hinaus wirken auch rechtliche Faktoren wie Gesetze auf die Umweltverwaltung ein. Diese Gesetze werden oftmals organisationsintern durch Normen oder ähnliche Regelungen erweitert. All diese Statuten schränken den Handlungsspielraum der Umweltverwaltung und somit auch die Menge der potenziellen Ziele ein. Aus diesem Grund sind die Erfassung und Modellierung der relevanten Rahmenbedingungen ein wichtiger Bestandteil zur Abbildung eines semantischen Netzes, welcher die Wirkungszusammenhänge innerhalb der Organisation darstellt.

Bei Beachtung der gültigen Rahmenbedingungen sind die möglichen Handlungsalternativen stark eingeschränkt werden. Die denkbaren Alternativen werden durch gesetzliche Bestimmungen eingegrenzt. Die legalen Möglichkeiten können entweder durch monetäre, gesellschaftliche Faktoren oder durch beide verringert werden. Auch besteht die Möglichkeit der Umsetzung von Aktionen, welche nicht finanziellen oder gesellschaftlichen Einflüssen unterliegen. Diese Maßnahmen werden z.B. durch den Gesetzgeber vorgeschrieben und sind unabhängig von weiteren Faktoren im definierten Umfang durchzuführen.

Nach der Gemeindeordnung z.B. des Landes Sachsen-Anhalt ist die Gemeinde der Träger der gesamten öffentlichen Aufgaben. Diese Aufgaben können zum einen Pflichtaufgaben und zum anderen freiwillige Tätigkeiten sein. Als dritte Form können auch Aufgaben identifiziert werden, welche vom Land oder dem Bund an die Gemeinde übertragen wurden. Aus dieser Unterteilung der Aufgaben ist ersichtlich, dass nur ein geringer Spielraum bei der Ausgestaltung der angebotenen Leistungen besteht. So müssen die Pflichtaufgaben und die übertragenen Aufgaben in qualitativ geeigneter Weise erbracht werden. Einzig im Bereich der freiwilligen Leistungen besteht die Möglichkeit der Variation oder auch der Kürzung von Leistungen. Die verschiedenen Aufgaben sind durch Gesetze oder Satzungen konkretisiert. Diese sind auch für die Verwaltung bei ihren Handlungen bindend. Als Beispiel können hier Gesetze über die Betreuung von Kindern, zum Brandschutz oder auch die Gewerbeordnung genannt werden.

Die zweite Form von Bedingungen, an welche die Gemeinde gebunden ist, wird durch die gesellschaftliche Struktur beschrieben. Diese ist teilweise eng mit den gesetzlichen Bestimmungen verbunden. Herrscht z.B. eine hohe Anzahl an Erwerbslosen in einer Gemeinde, sollte diese auch größere Bestrebungen in Ansiedlung von Gewerbe betreiben.

Als dritte Form der Restriktionen sind die finanziellen Mittel zu nennen. Im Haushalt der Gemeinde werden die Budgetausgaben beschränkt. Diese Beschränkungen müssen im Rahmen der einzelnen Aktivitäten beachtet werden. Zwar besteht die Möglichkeit der Umverteilung von Finanzmitteln, allerdings sind die Pflichtaufgaben

der Gemeinde so fixiert, dass auch der Bedarf an monetären Ressourcen feststeht. Die Umverteilung ist demnach nur im Rahmen der freiwilligen Aufgaben möglich.

Neben diesen drei konkreten Restriktionen beeinflussen auch politische Entscheidungen und Programme die Zielfindung der Gemeinde. Diese können evtl. geplante (freiwillige) Handlungen unterbinden und auch selbst Aktivitäten initiieren.

Zur Ableitung eines Kennzahlensystems, welches die Wirkungszusammenhänge in einer Gemeinde darstellt und somit Schlüsse auf mögliche Ziele zulässt, werden im ersten Schritt die beschriebenen Bedingungen erfasst. Anschließend können den verschiedenen Aufgaben Kennzahlen zugeordnet werden. Diese Kennzahlen beachten immer auch die Restriktionen.

Die folgende Abbildung 1 visualisiert die verschiedenen Rahmenbedingungen, welche bei der Ableitung von Steuergrößen beachtet werden müssen. Diese werden schematisch als Topic Map dargestellt. Hierbei ist zu bemerken, dass die Beziehungen innerhalb der Ebenen und zwischen den Ebenen aus Gründen der Übersichtlichkeit nicht beschriftet sind. Weiterhin ist festzuhalten, dass die Abbildung 1 als ein Topic Map-Dokument dargestellt werden kann. Zur besseren Veranschaulichung wurde dies aber in Ebenen unterteilt. Die Beziehungen innerhalb dieser Ebenen werden mit Hilfe von durchgezogenen Linien dargestellt. Gestrichelte Linien beschreiben ebenenübergreifende Verbindungen.

Abbildung 1 beschreibt auf oberster Ebene eine Topic Map, welche die Gesetzesgrundlage der Gemeindeordnung widerspiegelt. Das Topic der Gemeindeordnung untergliedert sich in die Topics der Pflichtaufgaben, freiwilligen Aufgaben und übertragenen Aufgaben. Die Beziehungen innerhalb dieser Ebene beschreiben, dass die Gemeindeordnung der Gemeinde verschiedene Aufgaben überträgt.

Auf der nächst tieferen Ebene sind weitere Gesetze und Satzungen modelliert. Die Pflichtaufgaben und übertragenen Aufgaben aus der höheren Ebene werden auf dieser Ebene durch Gesetze konkretisiert. Dabei erfolgt die Beschreibung, wie und in welchem Umfang die Leistungen zu erbringen sind.

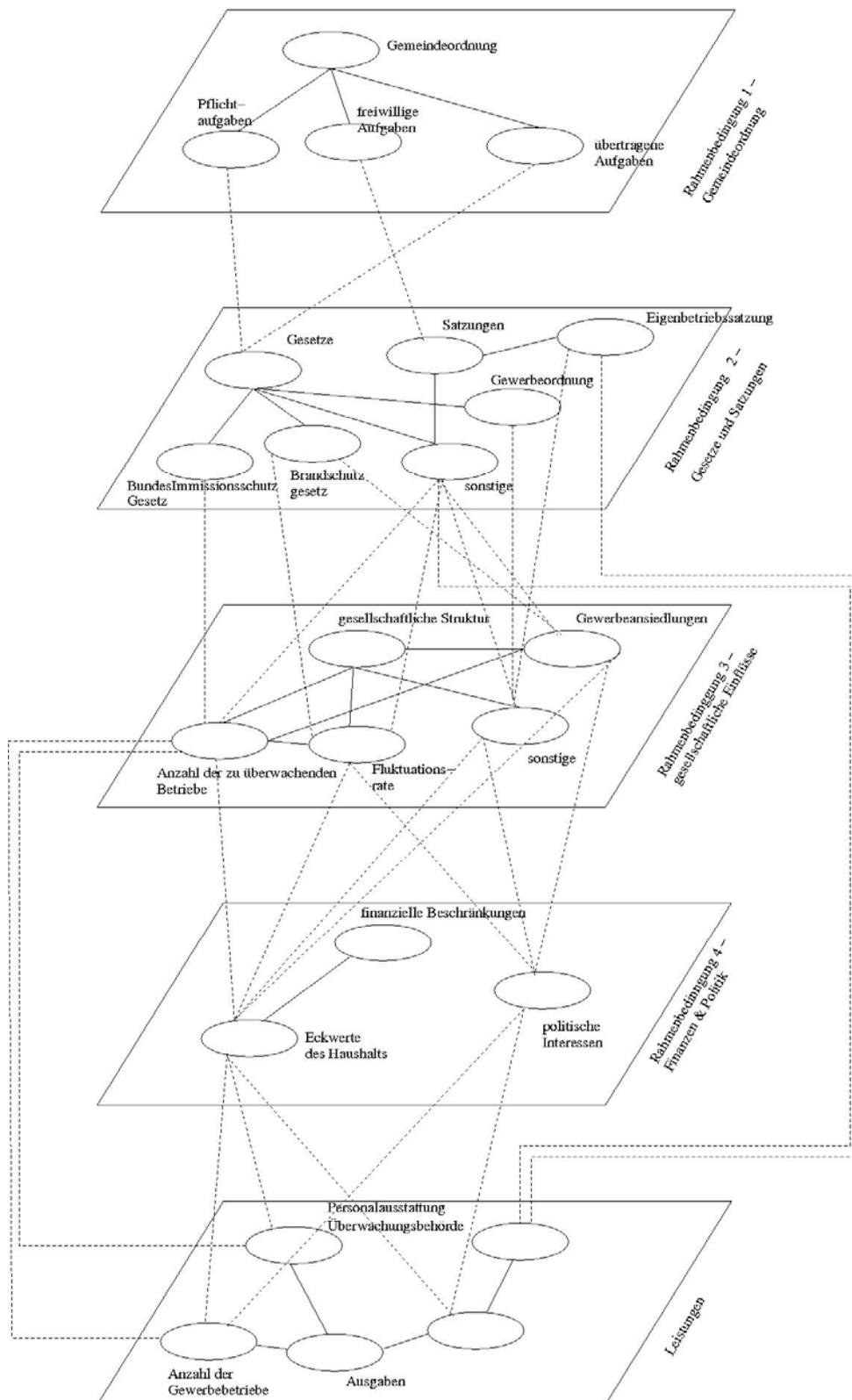


Abbildung 1: Topic Map der verschiedenen Ebenen der Rahmenbedingungen für umweltbezogene Leistungen einer Gemeinde

Die Beziehungen zwischen der ersten und zweiten Ebene beschreiben eine Zuordnung der Rechtsvorschriften zu den zuvor ermittelten Aufgaben und begrenzen somit diese. Als Beispiel für relevante Gesetze zeigt die Abbildung 1 das BundesImmissionsschutzgesetz (BImSchG) und das Brandschutzgesetz.

Die Beziehungen innerhalb der Ebene beschreiben eine Kategorisierung der einzelnen Rechtsvorschriften. Die darauf folgende Ebene beschreibt den Einfluss der gesellschaftlichen Struktur auf die Bildung von Handlungsalternativen. Hierbei sind exemplarisch einige Kenngrößen dieser Struktur abgebildet. Diese sind die Anzahl der Gewerbebetriebe, welche bereits in der Gemeinde angesiedelt sind sowie die Fluktuations- und Ansiedlungsrate. Die Größen dieser Ebene können oftmals durch statistische Erhebungen ermittelt werden. Die Beziehungen auf dieser Ebene beschreiben zum einen die Zuordnung der Faktoren zum Topic der gesellschaftlichen Struktur. Zum anderen können aber auch Einflüsse zwischen der Ansiedlungs- und Abwanderungsrate und der Anzahl der angesiedelten Betriebe identifiziert werden. Die Anzahl wird durch die Ansiedlungen oder Abwanderung von Gewerbebetrieben beeinflusst. Die Elemente der dritten Ebene beschreiben Faktoren, welche bei der Einhaltung der gesetzlichen Bestimmungen aus gesellschaftlicher Sicht relevant sind. Daher erfolgt die Verknüpfung der entsprechenden Elemente der zweiten Ebene mit den Faktoren der dritten.

Auf der vierten Ebene erfolgt die Modellierung der finanziellen Beschränkungen und der politischen Interessen als Rahmenbedingungen. Innerhalb der Ebene ist nur die Konkretisierung der finanziellen Beschränkungen in die Eckwerte des Haushalts (vergleichbar mit Budgetbeschränkungen) zu nennen. Die Gesetze und die gesellschaftliche Struktur erfordern Handlungen, deren Umsetzung durch monetäre Größen beeinflusst wird. Auch politische Interessen werden oftmals durch gesellschaftliche Strukturen geprägt.

So kann eine Veränderung in der Gesellschaft die Parteien zur Initiierung einer Aktion veranlassen. Aus diesem Grund erfolgt die Verknüpfung der dritten und vierten Ebene, wobei die Beziehungen als Einfluss der dritten Ebene auf die Elemente der vierten Ebenen interpretierbar sind.

Die letzte Ebene bildet schließlich konkrete Kennzahlen ab. Die Abbildung 1 zeigt eine exemplarische Auswahl. Zwischen diesen Kennzahlen können Wirkungszusammenhänge bestehen, welche durch die ebeneninternen Beziehungen angedeutet werden. Diese Kennzahlen bzw. die Ausprägungen dieser werden zum einen durch die monetären und politischen Faktoren der vierten Ebenen beeinflusst. Zum anderen erfolgt auch eine Einflussnahme durch höhere Ebenen auf diese Kennzahlen. So wirkt die Anzahl der Gewerbebetriebe aus der gesellschaftlichen Ebene auf die Anzahl der Überwachungen nach dem BImSchG und damit auf die Personalausstattung der Überwachungsbehörden ein.

Die Kennzahlen auf der letzten Ebene können einen Beitrag zur Bildung von Handlungsalternativen leisten. In diesem Rahmen werden die Bestimmungen eingehalten, bzw. diese Restriktionen können Handlungen erzwingen. Somit wird die Menge der möglichen Ziele bereits eingeschränkt.

Die Komplexität und Vielfältigkeit der Einflüsse der Verpflichtungen wird durch die Abbildung 1 deutlich. Bereits bei der exemplarischen Auswahl von wenigen Faktoren lassen sich viele Relationen bestimmen. Dies erschwert das Verarbeiten der gesamten Einschränkungen.

## 5 Ausblick

Bereits die Modellierung eines kleinen Systems mit einer geringen Anzahl von Elementen kann durch die hohe Anzahl von Abhängigkeiten zu einem schwer überschaubaren Modell heranwachsen. Werden zu viele Elemente aus einem Modell gewählt steigt die Komplexität sehr schnell an, werden dagegen zu wenig Elemente modelliert sind die Ursache-Wirkung-Zusammenhänge nicht klar erkennbar.

Ein komplexes Ursache-Wirkung-Gefüge kann IT-gestützt durch den Einsatz von Scopes dargestellt werden, ohne Informationen zu verlieren. Gleichzeitig ist durch den Einsatz von Operatoren die Ableitung von Zielen und Strategien möglich.

## 6 Literaturverzeichnis

- [Arndt et. al 2007] Arndt, H.-K.; Graubitz, H.; Köppen, V. (2007): Topic Maps for Representing Balanced Scorecards In: Conference of Computational Economics and Financial and Industrial Systems (CEFIS 2007), IFAC Symposium, October 9 - 11, Istanbul, Turkey
- [Freese 2000] Freese, E. (2000): Using Topic Maps for the representation, management & discovery of knowledge. (26.05.2008)  
<http://www.infoloom.com/gcaconfs/WEB/paris2000/S22-01.HTM>
- [ISO 13250:2000] ISO (1999): ISO/IEC 13250 Topic Maps. 1st Edition.
- [ISO 13250:2003] ISO (2002): ISO/IEC 13250 Topic Maps. 2nd Edition.
- [Kaplan/Norton 1997] Kaplan, R. S.; Norton, D. P. (1997): Balanced Scorecard – Strategien erfolgreich umsetzen– Aus dem Amerikanischen von Horváth, P.; Kuhn-Würfel, B.; Vogelhuber, C. Schäffer-Poeschel Verlag. Stuttgart.
- [König/Rehling 2002] König, S.; Rehling, M. (2002): Zur Übertragbarkeit der Balanced Scorecard auf ein zukunftsgerichtetes Personalmanagement in der öffentlichen Verwaltung.<http://www.uni-oldenburg.de/orgpers/DiskussionsPapier-01-02.pdf> (26.05.2008)
- [Krems 2007] Krems, B. (2007): Balanced Scorecard (BSC).  
<http://www.olev.de/b/bsc.pdf> (26.05.2008)
- [Pepper 2000] Pepper, S. (2002): The TAO of Topic Maps - Finding the Way in the Age of Infoglut.<http://www.ontopia.net/topicmaps/materials/tao.html> (26.05.2008)
- [Pepper/Gronmo 2002] Pepper, S.; Gronmo, G. O. (2002): Towards a General Theory of Scope.<http://www.ontopia.net/topicmaps/materials/scope.htm> (26.05.2008)
- [Rosemann 1996] Rosemann, M. (1996): Komplexitätsmanagement in Prozessmodellen – Methodenspezifische Gestaltungsempfehlungen für die Informationsmodellierung. Gabler Verlag. Wiesbaden.
- [Steinmann/Schreyögg 2005] Steinmann, H.; Schreyögg, G. (2005): Management – Grundlagen der Unternehmensführung – Konzepte – Funktionen – Fallstudien. 6. Auflage. Gabler Verlag. Wiesbaden.
- [Wagner 2002] Wagner, J. (2002): Balanced Scorecard als Managementsystem. Eul Verlag. Köln

- [Wellisch 2000] Wellisch D. (2000): Rechtfertigung der Staatstätigkeit. Vahlen Verlag. München.
- [Widhalm/Mück 2002] Widhalm, R.; Mück, T. (2002): Topic Maps – Semantische Suche im Internet. Springer Verlag. Berlin, Heidelberg
- [XTM 2001] XTM (2001): XML Topic Maps (XTM) 1.0. <http://www.topicmaps.org/xtm/>  
(26.05.2008)

# **Integrierte Umweltdatenbanken: Geodatenhaltung, Architekturalternativen, Inhaltliche und technische Aspekte der Modellierung**

Gergely Lukács, Wassilios Kazakos, Dominik Kumer, disy GmbH,

[lukacs@disy.net](mailto:lukacs@disy.net)

## **Einleitung**

Die Betrachtung komplexer Sachverhalte ist typisch für viele Aufgaben im Umweltschutz. Um diese Aufgaben zu unterstützen, werden integrierte Umweltinformationssysteme benötigt. Dieser Beitrag gibt eine Übersicht über diverse Ansätze, die es erlauben, integrierte oder später erweiterbare und integrierbare Umweltdatenbanken aufzubauen.

Neben dem Basisansatz für integrierte Umweltdatenbanken – einem Data Warehouse (DW) mit Geodatenhaltung in der Datenbank – werden DW-Architekturalternativen vorgestellt: DWs mit integrierter Fachanwendung, DWs mit externer Geodatenhaltung und DWs mit Fachanwendung für Massendaten (Messdaten).

Außerdem werden einige inhaltliche und technische Themen der Datenmodellierung aufgezeigt, die für integrierte Umweltdatenbanken von zentraler Bedeutung sind. Im Einzelnen sind dies Einheiten, chemische Stoffe, biologische Taxonomie, (aggregierte) Messwerte, Konventionen für Schemaobjekte, Objektartenkatalog und Versionsverwaltung.

## **Abstract**

For many tasks regarding environmental protection the observation of rather complex situations is characteristic. As support for these tasks integrated environmental information systems are necessary. This paper gives an overview over some approaches which allow the development of either integrated environmental data bases or ones that can be extended and integrated at a later date.

Firstly, the basic approach for integrated environmental databases is presented – the data warehouse (DW) with integrated management of geographical data – as well as alternatives: DWs with integrated application for data collection, with external management of geographical data, and with specific applications for mass data such as measured values.

Secondly, some aspects of data modelling with respect to content and techniques are pointed out; in detail these are units, chemical substances, biological taxonomy, (aggregated) measured values, conventions for database schemata, object type catalogues and version management.

## 1 Einführung

Unsere Natur ist ein komplexer Organismus, bei dem die Menschheit bis heute nicht alle Zusammenhänge von Ursache und Wirkung verstanden hat. Die Umweltforschung deckt immer wieder neue Kausalitäten im Umweltgeschehen naturwissenschaftlich auf, welche im Rahmen des operativen Umweltschutzes Berücksichtigung finden sollten.

Die ganzheitliche Sicht auf die Natur erhält auch in der nationalen und EU-weiten Gesetzgebung immer mehr Gewicht und verlangt deshalb von den beurteilenden und Maßnahmen steuernden Organen eine angemessene Berücksichtigung. Für die Umsetzung der Wasserrahmenrichtlinie, um ein Beispiel zu nennen, müssen Wassermenge und Qualität von Oberflächengewässern und Grundwasser, gemeinsam mit dem Niederschlag, der Wassernutzung und anderen Aspekten betrachtet werden. Zwei weitere EU-Richtlinien aus den letzten Jahren schreiben eine ganzheitliche Betrachtung für den Lärm und für die Umweltverschmutzung aus großen industriellen Quellen vor.

Umweltinformationssysteme, die Fachleuten zur Erledigung der operativen Aufgaben zur Verfügung gestellt werden oder die Öffentlichkeit informieren, müssen für diese ganzheitliche Betrachtung bzw. für die ständige Erweiterung der fachlichen Themen gerüstet sein und den betreffenden Nutzergruppen adäquate Benutzerschnittstellen zur Verfügung stellen. Auch wenn zunächst nur ein isoliertes Informationssystem aufgebaut werden sollte, kann die Berücksichtigung einiger Überlegungen eine spätere Erweiterung oder Integration deutlich erleichtern.

In diesem Beitrag werden die Erfahrungen aus zahlreichen Projekten in unterschiedlichen Bundesländern und bei einigen Bundesbehörden zusammengefasst.

## 2 Geodatenhaltung

Die Verwaltung geographischer Daten ist ein zentrales Thema im Bereich der Umweltinformationssysteme. Da das Thema sowohl in der Industrie als auch in der Literatur sehr häufig und oft sogar ausschließlich behandelt wird, siehe beispielsweise [DatenbankSpektrum, 2007], möchten wir an dieser Stelle, ohne auf das Thema näher einzugehen, nur einen einzelnen Punkt herausgreifen.

Marktführende (objekt-)relationale Datenbanken sind auch für Geodatenhaltung gerüstet. Die Haltung von Geodaten – gemeinsam mit den Sachdaten – direkt in den Datenbanken hat den Vorteil, dass auf SQL-Ebene auch Konstrukte, wie Funktionen oder Prädikate, für die Verwaltung von Geodaten zur Verfügung stehen. Damit ist es möglich, Sach- und Geodaten gemeinsam zu verwalten. Die Haltung von Geodaten in proprietären, binären Formaten, auch wenn sie in einer Datenbank erfolgt, hat diesen Vorteil nicht.

## 3 Architekturalternativen

Eine Auswertedatenbank, also ein Data Warehouse, ist die erste Wahl, wenn vorhandene, semantisch heterogene Datenbanken gemeinsam abgefragt und ausgewertet werden müssen [Lukács et al., 2003]. Durch die Trennung der Ebenen der Datenerfassung (Fachdatenbanken mit Fachanwendungen) und der integrierten Abfrage/Auswertung (Data-Warehouse-Datenbank und Anwendung) kann die Komplexität sehr stark reduziert werden. Die Daten werden dann automatisch und

periodisch durch ETL-Prozesse (ETL steht für Extract, Transfer, Load) aus den Datenbanken der Fachanwendungen in das Data Warehouse überführt. Für das Data Warehouse werden geeignete Schemakonstrukte und Werkzeuge benötigt, um die Vorteile voll auszuschöpfen.

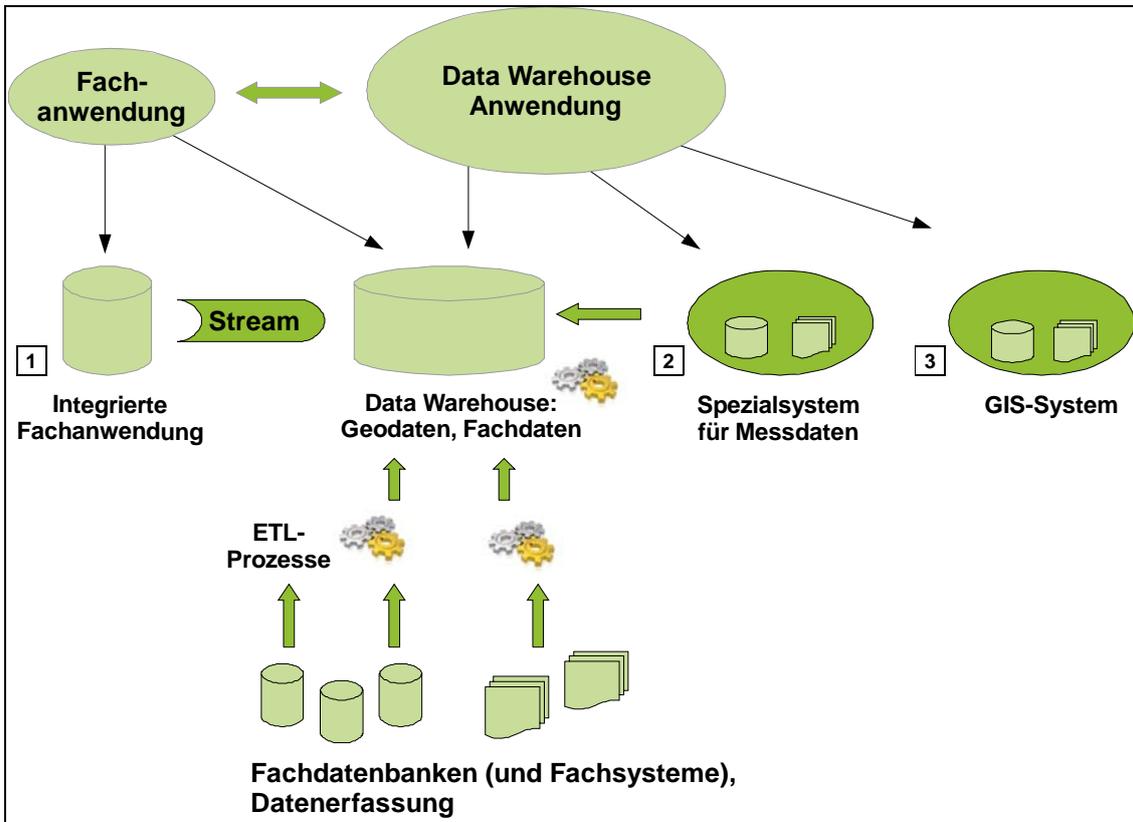


Abbildung 1: Architekturalternativen: Data Warehouse; (1) mit integrierter Fachanwendung, (2) mit Spezialsystem für Massendaten, (3) mit externen GIS-Diensten

Wenn auch die Entwicklung der Fachdatenbanken bestimmten Regeln folgt, kann der Aufbau des Data Warehouse deutlich vereinfacht werden. Es ist sogar möglich, eine Fachanwendung mit dem Data Warehouse zu integrieren [Nagypál et al., 2008]. Bestimmte Schemateile werden aus der Datenbank der Fachanwendung benutzt. Die nur im Data Warehouse benutzten übergreifenden Schemateile werden asynchron, mittels so genannter „Datenbank-Streams“ gepflegt. Die Vorteile der integrierten Fachanwendung sind der reduzierte Gesamtaufwand für die Entwicklung und Pflege, die Eliminierung der Zeitverzögerung durch die periodisch laufenden ETL-Prozesse und – falls nicht nur auf der Datenbankebene, sondern auch auf der Anwendungsebene eine Integration erfolgt – die einheitlichere Benutzeroberfläche.

Geodaten können ebenfalls in der Datenbank verwaltet werden. Dies hat den Vorteil, dass die Daten integriert, unter gemeinsamer Betrachtung der räumlichen, zeitlichen und fachlichen Kriterien abgefragt und ausgewertet werden können. Wenn die Geodaten in spezialisierten GIS-Systemen vorgehalten werden, die unter anderem mehr Spezialfunktionalität für Geodaten anbieten, ist die Berücksichtigung der zeitlichen und fachlichen Dimensionen eingeschränkter. Die Integration der Geodaten erfolgt in diesen Fällen meist über die Anwendungsebene, d.h., die Anwendung für das Data Warehouse greift über standardisierte Schnittstellen auf GIS-Dienste zu.

Datenquellen mit sehr großen Datenmengen, z.B. mit Pegelwerten, die häufig mehrmals pro Stunde automatisch gemessen werden, können ebenfalls auf der Anwendungsebene integriert werden. Wenn diese Daten mit anderen, im Data Warehouse vorliegenden Daten gemeinsam analysiert werden müssen, ist oft die Übernahme der aggregierten Daten die Lösung. Durch die Aggregation wird die Datenmenge reduziert, die aggregierten Daten reichen aber normalerweise für die gemeinsame Analyse aus.

Föderierte Datenbanken mit identischem Schema oder identischer Schnittstelle werden vor allem für Metadatenbanken benutzt, wenn jeder Teilnehmer auch die physikalische Hoheit über seine Daten behalten möchte. Diese Architektur hat wenig Flexibilität und ist vor allem für kleinere oder standardisierte Schemata geeignet.

#### 4 Modellierung inhaltlicher Aspekte

Die Entwicklung von Datenbankschemata ist nur zu einem kleineren Teil ein technisches, zu einem deutlich größeren Teil ein inhaltliches Thema [Silverston, 2001]. Beispielsweise kann die Normalisierung recht leicht gelernt werden. Für einen Anwendungsbereich ein passendes Schema zu entwickeln, erfordert aber Hintergrundwissen im Anwendungsbereich und auch Erfahrung, um die impliziten bzw. zuerst vom Anwender nicht genannten Anforderungen mit in die Abstimmung und Entwicklung einbeziehen zu können. Im Folgenden sind einige inhaltliche Schwerpunkte bei Umweltdatenbanken beschrieben:

Einheiten sind im Umweltbereich sehr vielfältig und von zentralem Interesse. Die Modellierung der Einheiten muss u.A. die Umrechnung von Werten in verwandte Einheiten oder die Prüfung von Formeln unterstützen.

Die Modellierung von Messwerten ist weitaus komplizierter als die von einfachen Zahlen. Neben dem eigentlichen numerischen Wert und seiner Einheit sind z.B. die Umstände der Messung (z.B. Messmethode), die Genauigkeit oder die Kennung (z.B. unter Nachweisgrenze) von zentraler Bedeutung. Aggregierte Messwerte werden oft mit speziellen Aggregationsmethoden oder für spezielle Zeiträume berechnet. Die Schemamodellierung von Messwerten und aggregierten Messwerten muss diesen Anforderungen Rechnung tragen (Abbildung 2).

MESSWERTAGGR	
<u>QUELLSYSTEM_NR</u>	NUMBER(4)
<u>OBJEKTART_NR</u>	NUMBER(8)
<u>OBJEKT_NR</u>	NUMBER(14)
<u>AGGRART_NR</u>	NUMBER(10)
<u>INTERVALLART_NR</u>	NUMBER(10)
<u>DATUMAB</u>	DATE
<u>DATUMBIS</u>	DATE
<u>PARAMETER_NR</u>	NUMBER(10)
<u>VORBEHANDLUNG_NR</u>	NUMBER(10)
<u>EINHEIT_NR</u>	NUMBER(10)
<u>VORZEICHEN</u>	CHAR(1 BYTE)
<u>WERTAGGR</u>	VARCHAR2(120 BYTE)
<u>WERTAGGRZAHL</u>	NUMBER
<u>ANZAHLWERTE</u>	NUMBER(10)
<u>DATUM</u>	DATE
<u>BERECHNUNGSMATHEMATISCH</u>	NUMBER(1)
<u>BERECHNUNGQUELLSYSTEM</u>	NUMBER(1)

Abbildung 2: Modellierung aggregierter Messwerte (wasserbezogene Daten)

Chemische Stoffe sind ein weiteres zentrales Thema. Hier sind z.B. die Verwaltung von Synonymen (oft werden für den gleichen Stoff in unterschiedlichen Kontexten unterschiedliche Namen benutzt) und die Beschreibung von Stoffgruppen relevant.

Auch die biologische Taxonomie kommt in unterschiedlichen Kontexten vor. Die Basis für die Taxonomie entspricht abstrakt einer Hierarchie. Jedoch machen Aspekte wie Alter, Geschlecht und Synonyme die Modellierung komplizierter. Für die praktische Handhabung kann auch eine Vereinfachung der Hierarchie hilfreich sein.

## **5 Technische Fragen der Modellierung**

Konventionen sind für die Entwicklung von Datenbankschemata generell hilfreich. Für integrierte und in weiteren Schritten auszubauende Datenbanken sind sie aus Sicht der Autoren sogar Pflicht, um die Komplexität zu reduzieren.

Schemakonventionen decken eine Reihe von Einzelaspekten ab; hierzu zählen unter anderem die Namensgebung für Tabellen und Attribute, die Bildung von Attributgruppen und die Reihenfolge der Attribute.

Die Entwicklung und Benutzung eines Objektartenkatalogs ist ein weiterer Eckpunkt integrierter Datenbanken. Das Objektartenkatalog beschreibt die vorkommenden Objektarten, wie z.B. Oberflächengewässer, Messpunkte oder Wasserfassungen. Dies kann z.B. in einer flachen oder in einer hierarchisch strukturierten Form erfolgen.

Der möglichst einheitliche Aufbau der Primär- (und damit der Fremd-)schlüssel ist ebenfalls empfehlenswert und erleichtert die Integration bzw. den Ausbau der Datenbanken erheblich.

Die Versionsverwaltung der Daten ist im Umweltbereich aus zwei Gründen interessant. Zum einen ist es wegen der Langlebigkeit bzw. wegen des komplexen Messverfahrens der Umweltdaten oft gewünscht, auf die historischen Zustände zuzugreifen und zeitlich langjährige Auswertungen durchführen bzw. Datenänderungen protokollieren zu können. Zum anderen ist, z.B. für die Planung von Maßnahmen, oft notwendig, unterschiedliche Zukunftsszenarien parallel zu beschreiben und deren Konsequenzen miteinander zu vergleichen. Auch wenn die Versionsverwaltung von Datenbankdaten deutlich komplizierter ist als die anderer Daten bzw. nur Teilfunktionalitäten gängiger Versionsverwaltungssysteme wie cvs oder subversion abdeckt, sind passende Lösungen für viele Anforderungen machbar und wünschenswert.

Vor allem im Data-Warehouse-Kontext enthalten Teile der Datenbank Daten, die im Voraus, zum Zeitpunkt der Einlagerung oder der Aktualisierung, berechnet wurden. Die Aktualisierung dieser Auswerteschemateile sowie die Prüfung der Aktualität ihrer Daten ist ebenfalls eine wichtige Erleichterung bei der Verwaltung der Daten.

## **6 Recherche und Auswertung**

Der Fokus unseres Beitrags ist die Ebene der (objekt-)relationalen Datenbank. Alle bisherigen Punkte haben sich auf diese Ebene bezogen.

Es ist allerdings wichtig im Auge zu behalten, dass die Recherche bzw. Auswertung einer integrierter Umweltdatenbank auch eine geeignete Anwendung erfordert. Eine feste „Ausprogrammierung“ der Anwendung für bestimmte Datenbankschemata, Abfrage- und Darstellungsmöglichkeiten ist bei der Komplexität integrierter Umweltdatenbanken nicht mehr möglich und verhindert die einfache Integration

neuer Teilschamata. Eine mögliche Alternative für eine geeignete, „generische“ Anwendung ist die Plattform disy Cadenza [Hofmann et al., 2007].

## 7 Zusammenfassung

In diesem Beitrag haben wir die wichtigsten Aspekte kurz angerissen, die beim Aufbau einer integrierten Umweltdatenbank – oder einer Datenbank, die später, wie im Umweltbereich typisch, ausgebaut wird – zu berücksichtigen sind. Diese decken die Haltung der Geodaten, die Datenbankarchitektur sowie inhaltliche und technische Modellierungsfragen ab.

Auch wenn es keine einzig richtige Lösung gibt, die alle Aspekte in ganzer Tiefe abdeckt, ist die Behandlung dieser Themen und die kontinuierliche Weiterentwicklung der dahinter stehenden Lösungen nach der Einschätzung der Autoren zentral. Grundlegend steht hinter allen Überlegungen die Bereitstellung von Systemumgebungen, mit denen es den Organisationen möglich ist, in Zeiten knapper Finanzmittel gut recherchier- und auswertbare integrierte Umweltinformationssysteme für das operative Geschäft bereit zu halten.

## 8 Literaturverzeichnis

[DatenbankSpektrum, 2007] Datenbank Spektrum, Sonderheft Geo- und Umweltdatenbanken, 2007.

[Lukács et al., 2003] Lukács, Gergely et al.: Data Data Warehouse for Water Management – Geographic Data and ETL processes. In A. Gnauck an R. Heinrich (editors): Proc of the 17th International Conference Informatics for Environmental Protection, Metropolos Verlag, Marburg, 2003.

[Nagypál et al., 2008] Nagypál, Gabor et al.: Die Landesweite Datenbank für die Wasserwirtschaft und das Wasserbuch in Niedersachsen: Ein Data Warehouse mit integrierter Fachanwendung auf der Basis von disy Cadenza. Wird erscheinen im Konferenzband für die EnviroInfo 2008.

[Silverston, 2001] Silverston, Len: The Data Model Resource Book. Wiley, New York, 2001.

[Hofmann et al., 2007] Hofmann, Claus et al.: disy Cadenza/GISterm - Plattform für Berichte, Auswertungen und Geographische Informationssysteme bei Bund und Ländern, KEWA-Bericht Umweltministerium Baden-Württemberg, 2007, <http://www.uis.baden-wuerttemberg.de/servlet/is/28429/>

# Semantische Dienste in ORCHESTRA – Stand und Ausblick

Ulrich Bügel, Fraunhofer IITB [buegel@iitb.fraunhofer.de](mailto:buegel@iitb.fraunhofer.de)

Dr. Désirée Hilbring, Fraunhofer IITB [hilbring@iitb.fraunhofer.de](mailto:hilbring@iitb.fraunhofer.de)

Thomas Usländer, Fraunhofer IITB [uslaender@iitb.fraunhofer.de](mailto:uslaender@iitb.fraunhofer.de)

## 1 Einführung und Motivation

Das ORCHESTRA-Projekt [1] und die im Rahmen des Projektes entwickelte Dienst-Architektur wurden in früheren Workshops des Arbeitskreises vorgestellt [2]. Durch Anpassung des ISO Referenzmodells für „Open Distributed Processing (RM-ODP)“ an die Anforderungen einer „Service-Oriented Architecture (SOA)“ wurde ein „Reference Model for the ORCHESTRA Architecture (RM-OA)“ spezifiziert und im Kontext der Arbeitsgruppen „Risk and Crisis Management“ und „Architecture“ des Open Geospatial Consortium (OGC) publiziert. Das Referenz-Modell hat mittlerweile den Status eines OGC Best Practice Papers erreicht [3].

Bis heute konzentrierte man sich bei Standardisierungsarbeiten im Bereich „Distributed Processing“ hauptsächlich auf die Definition der Transfersyntax (Signatur, Schnittstellen) und der Struktur der transferierten Daten. Diese beiden Aspekte stellen die physikalische Konnektivität und den Datenaustausch sicher und sind die Basis zur Nutzung einer verteilten Architektur. Sie berücksichtigen jedoch nicht die inhaltlichen Aspekte des Datenaustauschs: ähnlicher Inhalt auf verschiedenen Systemen kann sich auf einen völlig unterschiedlichen Kontext beziehen. Konsequenterweise können die Systeme daher nicht auf semantischer Ebene zusammenarbeiten.

Dieser Beitrag beschreibt die Arbeiten von ORCHESTRA zur Einbindung semantischer Technologien in serviceorientierte Architekturen mit Systemen zur raumbezogenen Informationsverarbeitung unter Beachtung existierender ISO- und OGC-Standards. Sie umfassen die Definition, Implementierung und Pilot-Anwendung „semantischer Dienste“. Mit diesem Grundgerüst ist jedoch das Nutzungspotential semantischer Technologien bei weitem noch nicht ausgeschöpft. Geplante Weiterentwicklungen, die sich derzeit in der Konzeptions- bzw. Realisierungsphase befinden, werden am Ende des Beitrages skizziert.

## 2 Der ORCHESTRA Ansatz zur Interoperabilität

Interoperabilität (nach ISO 19119) bedeutet „die Fähigkeit, die es funktionalen Einheiten ermöglicht zu kommunizieren, Programme auszuführen oder Daten auszutauschen, wobei der Benutzer möglichst wenig oder kein Wissen über die charakteristischen Eigenschaften dieser Einheiten benötigt“.

Die wesentliche Forderung ist hier eindeutig die Entlastung des Benutzers auch von den inhaltlichen Aspekten des Informationsaustausches. Dies betrifft auch die Überwindung sprachlicher und fachterminologischer Hindernisse. Die semantische Erweiterung der ORCHESTRA-Architektur verfolgt den Ansatz, dass der Zugriff auf Daten, Information und Dienste nicht direkt, sondern über eine Vermittlungsschicht

erfolgt, die sich auf formale, semantische Beschreibungen stützt. Dies bedeutet konkret, daß alle Zugriffsobjekte formal beschrieben werden müssen in einer Form, die dem Inhalt eine eindeutige Interpretation zuordnet. Der formale Aspekt stellt die Möglichkeit zur maschinellen Verarbeitung von Inhalten sicher.

Ontologien sind das Mittel der Wahl, mit der die Bedeutung von Inhalten dargestellt werden kann. Daten und Diensten wird ontologiebasierte Meta-Information zugeordnet, so dass diese auf semantischer Ebene miteinander in Beziehung gesetzt werden können.

Heutige Geo-Informationssysteme stützen sich weitgehend auf bewährte OGC-Standards. Der Einbezug semantischer Aspekte ist in diesen Standards noch nicht berücksichtigt. Es gibt jedoch einige Initiativen wie z.B. das „Geospatial Semantic Web (GSW)“ [4], die zu beachten sind. Die Arbeiten in ORCHESTRA zur semantischen Interoperabilität stützen sich weitgehend auf Standards, die vom World Wide Web Konsortium und der European Semantic Systems Initiative (ESSI) speziell im Bereich semantischer Dienste vorangetrieben werden. Sie wurden vor allem von folgenden Anforderungen bestimmt:

- Die Definition von Anwendungs-Schemata bleibt kompatibel zu der von der OGC empfohlenen Vorgehensweise, stellt aber zusätzlich ein Regelwerk zur Beschreibung von ontologiebasierter Meta-Information zur Verfügung.
- Dienste-orientierte Architekturen sind durch das Einbringen semantischer Dienste erweiterbar. Dieser Aspekt wird in den folgenden Kapiteln weiter ausgeführt.

### **3 Semantische Basisdienste**

Die ORCHESTRA-Architektur unterscheidet zwischen thematischen Diensten – z.B. Dienste zur Ausführung geostatistischer Berechnungen oder Simulationsdienste – und Architekturdiensten, die die generische Infrastruktur bereitstellen und die thematischen Dienste unterstützen – z.B. Katalogdienste, Kartendienste oder Dienste für die Nutzung von Sensoren. Das Konzept zur Nutzung semantischer Technologien umfasst die Definition von „semantischen Diensten“ als Architekturdienste. Diese erweitern die ORCHESTRA-Dienstarchitektur additiv und führen somit nicht zu einem vollständigen Re-Design der Architektur, was oft in anderen Ansätzen beobachtet werden kann. Semantische Dienste dienen zur Verarbeitung von Informationen, die in Sprachen zur ontologiebasierten Wissenrepräsentation (z.B. RDF(S), OWL) abgefasst sind. Für die thematischen Dienste ist diese Verarbeitung transparent, d.h. diese werden von der expliziten Handhabung semantischer Informationen entlastet.

Semantische Dienste stellen die Interoperabilität auf semantischer Ebene sicher. Sie wurden in ORCHESTRA so definiert, dass eine Komposition von Diensten in semantischen Szenarien möglichst vereinfacht wird.

#### **3.1 Annotationsdienst**

Eine Basisaufgabe bei der Realisierung des Semantic Web ist die semantische Annotation von textuellen Inhalten (z.B. Web-Seiten, Dokumenten) aber auch aktiven Bausteinen (z.B. Dienste, Datenbanken, Anwendungen) mit formaler Meta-Information. Diese wird mit Hilfe von Ontologien in einen Kontext eingebettet und kann durch Anwendungen weiterverarbeitet werden. Während heutige Annotationswerkzeuge meistens benutzergeführt oder semi-automatisch arbeiten,

generiert der ORCHESTRA-Annotationsdienst automatisch Meta-Information und setzt diese in Bezug zu den Elementen einer Ontologie (z.B. Konzepte, Instanzen, Relationen).

In seiner ersten Implementierung ist der ORCHESTRA-Annotationsdienst auf die Annotation unstrukturierter und semi-strukturierter Quellen (z.B. Text in natürlicher Sprache, Dokumente oder Webseiten) ausgelegt. Er basiert auf der Verwendung eines Werkzeuges zur automatischen Informations-Extraktion (IE) [5], das in der Lage ist, im Text auftretende „named entities“ (z.B. Straßen, Flüsse, Personen) zu identifizieren. Dazu werden Methoden der Verarbeitung natürlicher Sprache genutzt. Die Zuordnung von „named entities“ zu ontologischen Elementen basiert auf der Verwendung einer Wissensbasis (siehe Kapitel 3.3), die voreingestelltes Wissen (z.B. „der Rhein ist ein Fluss“) enthält.

Geplante Implementierungen sehen die Erweiterung des Annotationsdienstes dahingehend vor, dass nicht nur vorab bekannte Entitäten entdeckt werden, sondern auch neues Wissen entdeckt und in die Wissensbasis eingetragen werden kann. Dort kann es dann mit Hilfe einer Abfragesprache abgerufen werden.

Mit Hilfe einer einfachen Testanwendung können Texte am Bildschirm automatisch annotiert werden, d.h. die entdeckten Entitäten werden farblich markiert und den Elementen einer Ontologie zugeordnet.

### **3.2 Ontologie-Zugriffsdienst**

Der „Ontology Access Service“ unterstützt den Lesezugriff auf Ontologien und erlaubt den Import und Export formal spezifizierter Ontologien in eine Ontologie-Basis für den programmierten Zugriff. Er gestattet den Zugriff auf ganze Ontologien sowie auf einzelne Ontologie-Elemente und stellt Funktionen für das Ontologie-Management zur Verfügung. Wenn andere Architekturdienste oder thematische Dienste Ontologie-Management Funktionen benötigen, können diese mit der Schnittstelle des Ontologie-Zugriffsdienstes erweitert werden.

### **3.3 Wissensbasis-Schnittstelle**

Der Ontologie-Zugriffsdienst – und jeder beliebige andere Dienst – kann mit einer Schnittstelle zur Ablage und Abfrage gespeicherten Wissens ausgestattet werden. Auf diese Weise können beispielsweise RDF Datenbanken eingebunden und mit SPARQL-Queries abgefragt werden. Grundsätzlich sind jedoch alle semantischen Dienste unabhängig von konkreten Sprachen auf abstrakter Ebene definiert; die Nutzung konkreter Sprachen ist implementierungsabhängig. Optional kann der Wissenszugriff mit Reasoning-Funktionalität erweitert werden.

## **4 Semantischer Katalog**

Im Unterschied zu Suchmaschinen wie Google oder Yahoo erlauben Kataloge nach der OGC-Spezifikation auch den Zugriff auf Meta-Information über raumbezogene Ressourcen, z.B. topografische Daten oder Anwendungen zur geo-statistischen Informationsverarbeitung. OGC-Kataloge werden häufig auch von Nicht-Experten genutzt, die die geeigneten Suchbegriffe für die zumeist komplexe Katalogschnittstelle nicht exakt kennen. Die Resultate solcher Suchanfragen sind oft unbefriedigend.

Der in ORCHESTRA entwickelte semantische Katalog (SemCat) erlaubt daher die Konstruktion von Suchanfragen durch Selektion, Analyse und Nutzung geeigneter Ontologien in Kooperation mit dem Ontologie-Zugriffsdienst. Der semantische Katalog stellt Funktionen für die semantische Erforschung von Kataloginhalten konventioneller Kataloge zur Verfügung.

Der semantische Katalog setzt sich aus drei Architekturebenen zusammen:

- Die erste Architekturebene umfasst konventionelle Kataloge (z.B. OGC-konforme-, ORCHESTRA-Kataloge oder Bildkataloge) und andere Metainformationsquellen wie Google oder Yahoo. ORCHESTRA Kataloge können in diese Architekturebene direkt integriert werden, während OGC-konforme Kataloge mit Hilfe von Adaptern eingebunden werden können. Es wurden im ORCHESTRA Projekt zwei Adapter entwickelt. Der Erste realisiert das Einbinden von OGC-konformen Katalogen mit dem ISO19115 Profil, der Zweite realisiert das Einbinden von Katalogen mit dem ebRIM (ISO/TS 15000-3) Profil. Sowohl die OGC-konformen Kataloge als auch die ORCHESTRA Kataloge nutzen als Abfragesprache die „OGC Filter Encoding“ Spezifikation.
- Die zweite Architekturebene realisiert eine kaskadierte Katalogarchitektur, in der jede Komponente mit einer ORCHESTRA Katalogschnittstelle ausgestattet ist. In der kaskadierten Architektur werden Anfragen für die Bearbeitung in Subkataloge weitergeleitet und die Antworten in einer gemeinsamen Ergebnismenge gesammelt.
- Die dritte Architekturebene realisiert die semantischen Erweiterungen des Katalogs. Sie verfügt wie die anderen Ebenen über eine ORCHESTRA Katalogschnittstelle, damit auf alle Katalogkomponenten über die konventionelle Katalogschnittstelle zugegriffen werden kann. Nur Klienten, die die semantischen Erweiterungen nutzen wollen, müssen ihre Funktionen für die Nutzung der semantischen Schnittstelle des Katalogs erweitern.

Die semantische Schnittstelle des Katalogs nutzt Ontologien für eine automatische oder interaktive Erweiterung von Kataloganfragen. Dafür wird eine kontextspezifische Ontologie benötigt, die wichtige Konzepte und Beziehungen im thematischen Interessensgebiet der Nutzer abbildet. Das kann bedeuten, dass die Ontologie für die Bedürfnisse einer bestimmten semantischen Kataloginstanz oder auf einen bestimmten Nutzer abgestimmt wird.

Der semantische Katalog begegnet diesen Anforderungen flexibel und realisiert auf der dritten Architekturebene einige Operationen des Ontologie-Zugriffsdienstes, damit die Ontologien, die für die Erweiterung der Kataloganfragen verwendet werden, beliebig ausgetauscht werden können.

Ein Ziel der Architektur des semantischen Katalogs ist es, das Einbinden von konventionellen Katalogen ohne Änderungen an diesen Katalogen zu ermöglichen. Trotzdem müssen die Konzepte der fachspezifischen Ontologien mit den Abfrageschlüsselwörtern der konventionellen Kataloge in Verbindung gebracht werden. Das wird durch die Integration der Abfrageschlüsselwörter in die Beschreibung der Konzepte in der Ontologie erreicht.

Für die Erweiterung der Kataloganfrage realisiert die semantische Schnittstelle des Katalogs eine neue Operation: *improveQuery*. Diese Operation erhält eine normale Kataloganfrage. In der Ontologie wird nach Konzepten gesucht, die den Suchbegriffen der Anfrage ähneln. Ausgehend von den identifizierten Konzepten

werden über die Ontologie verwandte Begriffe (Eltern, Kinder oder über Beziehungen verknüpfte Begriffe) identifiziert und dem Client als Optionen für die interaktive Erweiterung der Kataloganfrage angeboten.

Eine automatische Variante der Anfragenerweiterung ist in der *search* Operation des semantischen Katalogs realisiert. Die ursprüngliche Kataloganfrage wird mit Hilfe der Operation *improveQuery* erweitert und anschließend automatisch an die Subkataloge weitergeleitet. Die in den Katalogen gefundenen Ergebnisse werden sortiert nach Suchbegriffen in einem Semantischen Report zurückgeliefert.

Der semantische Katalog wurde im „ORCHESTRA Federated Pilot“ (förderierter Pilot der Pilotanwendungen, URL: <http://semcat.iitb.fraunhofer.de>) getestet (Abbildung 1).

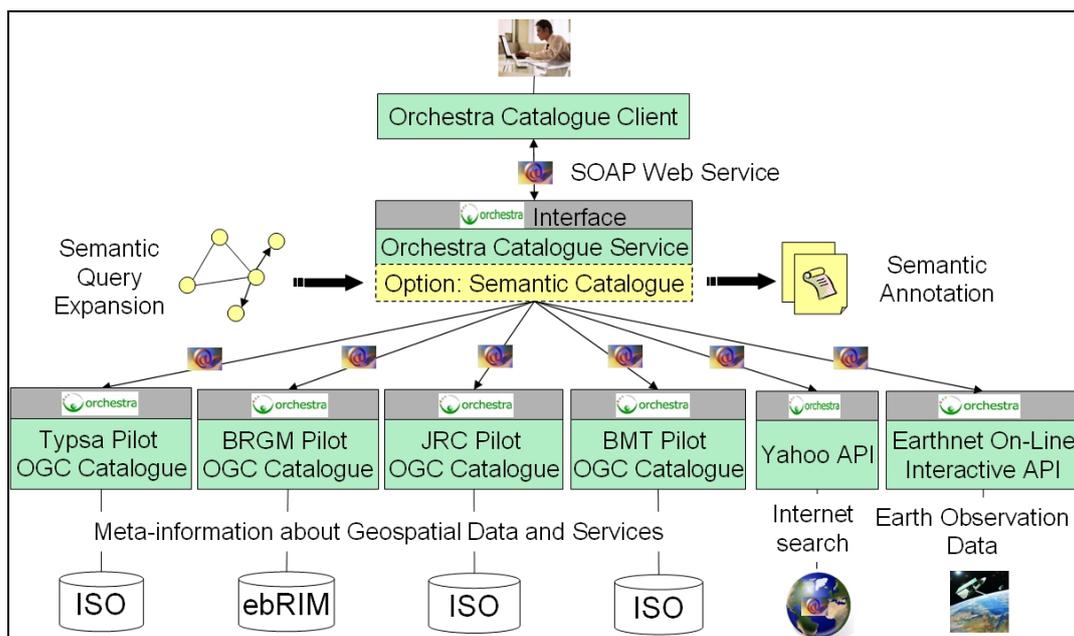


Abbildung 1: Der Semantische Katalog im ORCHESTRA Federated Pilot

Das kaskadierte Szenario umfasst die vier Kataloge der ORCHESTRA Pilotanwendungen, einen Zugriff auf eine Internet-Suchmaschine (Yahoo) und einen Zugriff auf einen Bildkatalog mit Satellitenbildern verschiedener Erdbeobachtungssatelliten. Dementsprechend enthält die Ontologie Konzepte, die sich aus der Beschreibung des ORCHESTRA Projektes ergeben und Konzepte aus dem Fernerkundungsbereich. Ein Nutzer, der sich über ORCHESTRA informieren möchte oder jemand, der auf der Suche nach Satellitenbildern ist, ohne die in der Branche üblichen Bezeichnungen zu kennen, kann seine Suchbegriffe mit Hilfe der Ontologie solange verfeinern, bis er gewünschte Ergebnisse findet.

Zusätzlich zur Suchfunktion bietet der Klient des Semantischen Katalogs noch weitere ontologiebasierte Funktionen an. Es ist möglich, textbasierte Ergebnisse mit Hilfe des Annotationsdienstes zu annotieren (Abbildung 2). Die Annotationen werden mit Hilfe eines Ontologie-Browsers konkreten Textstellen zugeordnet und dienen dem Aufrufer als umfassende Verständnishilfe. Über den Ontologie-Browser können verwandte Konzepte identifiziert werden. Der Nutzer hat zusätzlich die Möglichkeit Konzepte, die mit Hilfe des Annotationsdienstes identifiziert wurden, in den Katalogklienten zu übertragen und von neuem mit der Suche nach Informationen zu beginnen.

**Chapter 1: Introduction**

**Ontology(OAS-Name):** MaritimeAuthority  
**URL:** http://semcat.iitb.fhg.de/servlet/is/1998/?command=downloadContent&filename=MaritimeAuthori

Tributyltin (TBT) and triphenyltin (TPT) are among the most toxic compounds deliberately released into the marine environment by man (Fent, 1996). The purpose of the application of TBT, and to a lesser degree also TPT, is to prevent ship's hulls from fouling by organisms such as barnacles, bivalves, and algae. Anti-fouling paint has an intended beneficial effect for ships' sailing and manoeuvring properties by keeping the hull smooth and hence lowering the drag coefficient of the ship and associated fuel consumption. However, TBT does not remain at the ship's surface, it leaches further into the sea, adheres to particles and biota and settles into the sediment, where it can last for decades (Chau, Maguire, Brown, Fang, & Batchelor, 1997; Maguire, 2000; Maguire, Chau, & Thompson, 1997). Recently TBT has also been found to be transported by marine organisms (Tessier, & Donard, 2000) allowing a rapid transport over large distances. This is the case for the transport of organotin compounds in the liver of male sperm whales (Physeter macrocephalus) stranded on the Dutch and Danish coasts, although these deep-sea animals had apparently not been feeding after they accidentally entered the shallow North Sea (Ariese, Hattum, Hopman, Boon, & ten Hallers-Tjabbes, 1998). Input of TBT from antifouling paints in the open sea is expected to be larger than 25 m in EU countries. A major reduction of TBT input is expected since the implementation of the Convention on Harmful Anti-fouling Systems for Ships (IMO, 2001) has entered into force.

In the marine environment, non-target organisms have frequently been found to be affected by toxic effects of organotin biocides (Fent, 1996). The most sensitive organisms are gastropods (snails) and oysters; these animals can already be affected by tributyltin (TBT) concentrations between 1 and 10 ng Sn l<sup>-1</sup> sea water (Matthiessen & Gibbs, 1998). In snails, female organisms develop a penis and a vas deferens or spermatheca (Spoooner, Gibbs, Bryan, & Burt, 1987). The phenomenon of penis and vas deferens formation in females is called imposex. In some species such as the well-investigated dogwhelk (Nucella lapillus), females become sterile because of the development of the female genital pore by the developing sperm duct (Bryan, Gibbs, Burt, & Hummerstone, 1987; Gibbs & Bryan, 1988). Recently, the biochemical mechanism elucidated in the mud snail *Hydrobia ulvae* (Oberdörster & McClellan-Grigg, 1998) triggering imposex is the neuropeptide APGWamide, which is involved in the control of sexual differentiation in molluscs. The male steroid hormone testosterone and its derivative 11-ketotestosterone are also involved in the control of sexual differentiation. Laboratory experiments with *Buccinum undatum* showed that juveniles were more sensitive than adults (Mensink, Everaarts, Kragt, ten Hallers-Tjabbes, & Boon, 1996; Monecniak et al., 2002). In the juveniles, TBT concentrations 57 ng Sn l<sup>-1</sup> induced imposex, whereas in adults it did not. In contrast to *N. lapillus*, the developing vas deferens did not show a tendency to overgrow the female genital pore.

**Auswahl für Catalogue Client Service:**

Lookup Reset

Abbildung 2: Semantische Annotation textbasierter Ergebnisse einer Katalogsuche

Bei der Implementierung des semantischen Katalogs im föderierten ORCHESTRA Pilot stellte sich heraus, dass die Nutzung semantischer Technologien in offenen, grenz- und domänenübergreifenden IT-Umgebungen zwar eine ambitionierte Aufgabenstellung ist, deren Ergebnisse jedoch für die beteiligten Institutionen von beträchtlichem Nutzen sein können.

## 5 Weiterführende Arbeiten

Die ORCHESTRA-Architektur und deren semantische Erweiterungen werden in Projekten auf verschiedener Ebene angewendet und weiterentwickelt. Beispielsweise erfolgt im EU-Projekt Sensor Anywhere (SANY) [9] die Erweiterung der Architekturdienste und Informationsmodelle um die funktionale Domäne der Sensorik. Der semantische Katalog wird in SANY für die Suche nach Umweltsensoren, Modellen, Messwerten und sensorbezogenen Diensten eingesetzt. Dazu wird eine Ontologie definiert, die Umweltparameter-, Sensor- und Dienstetypen miteinander in Beziehung setzt. Stellvertretend für funktionale Weiterentwicklungen des semantischen Katalogs werden nachfolgend zwei Themen beschrieben, die eine

verbesserte Nutzung der Ergebnis-Annotation über den Charakter einer Lesehilfe hinaus zum Ziel haben.

## 5.1 Ontologiebasiertes Ranking

Durch Einbezug eines Algorithmus für ontologiebasiertes Ranking [6] wird die Möglichkeit zur Reduktion und Ordnung der Trefferliste in Bezug auf die selektierte Ontologie ermöglicht. Diese Algorithmen werden derzeit im Semantischen Katalog getestet. Zurzeit wird ein simpler Algorithmus verwendet, der zuerst mit Hilfe der Operation *improveQuery* des semantischen Katalogs verwandte Konzepte des originalen Suchbegriffs identifiziert. Die Suchergebnisse des Katalogs werden dann nach den identifizierten Konzepten durchsucht und Treffer werden entsprechend gewichtet. Die Ergebnisliste der Suchanfrage wird anschließend anhand der Gewichtungen sortiert.

## 5.2 Automatische Klassifikation

Zur Erfüllung von Aufgaben im Umwelt und Naturschutz wird der Zugriff zu (Geo-) Fachdaten anderer Verwaltungen wie Geologie, Forsten, Landwirtschaft, Bauleitplanung, Raumordnung, Regional- und Landesplanung, Statistik, Straßenbau, Verkehr usw. benötigt. Diese Fachdaten liegen in strukturierter Form in Datenbanken und Informationssystemen vor. In integrierten Umweltinformationssystemen und Geodatenverbänden wie WIBAS (Baden-Württemberg) oder ALKIS/ATKIS werden für die Fachbereiche alle Objektarten, die Realwelthänomene repräsentieren, in sog. Objektartenkatalogen (bei ISO/OGC auch Feature Type Catalogues genannt) geführt und damit der Datenaustausch erleichtert.

Ergänzend zur strukturierten Information wird in den Fachinformationssystemen auch ein dynamischer Bestand an unstrukturierter Information in Form von Texten, Dokumenten etc. verwaltet, z.B. Merkblätter, Durchführungsverordnungen, Umweltberichte oder Hintergrundinformation zu den Daten. Diese Dokumente haben i.d.R. Bezug zu Einträgen in den Objektartenkatalogen. Um diesen Bezug explizit zu machen, wäre es vorteilhaft, die Dokumente objektartspezifisch nach dem Anwendungsschema der Kataloge zu klassifizieren. Ziel weiterführender Arbeiten im ORCHESTRA-Kontext ist daher die Realisierung einer automatischen (maschinellen) Klassifikation, um so eine intelligente, automatische Ablage neu hinzukommender Dokumente zu schaffen.

Für die automatische Klassifikation unstrukturierter Information gibt es zwei Ansätze:

## 6 Nutzung des ORCHESTRA Annotationsdienstes

Die Objektartenkataloge können in verschiedener Form vorliegen, z.B. als einfaches Dokument oder als Datenbank-Schema oder –Anwendung. Das Anwendungsschema kann auch in Form von Konzepten einer formalen Ontologie spezifiziert werden. Damit ist die Voraussetzung für die Nutzung des ORCHESTRA–Annotationsdienstes geschaffen: die Dokumente können semantisch annotiert werden. Eine anschließende Auswertung der Annotationen – z.B. durch Gewichtungen wie beim ontologiebasierten Ranking (Kap 5.1) – kann das am besten passende Ontologie-Konzept ermitteln und somit eine Kategorisierung vornehmen (Abbildung 3).

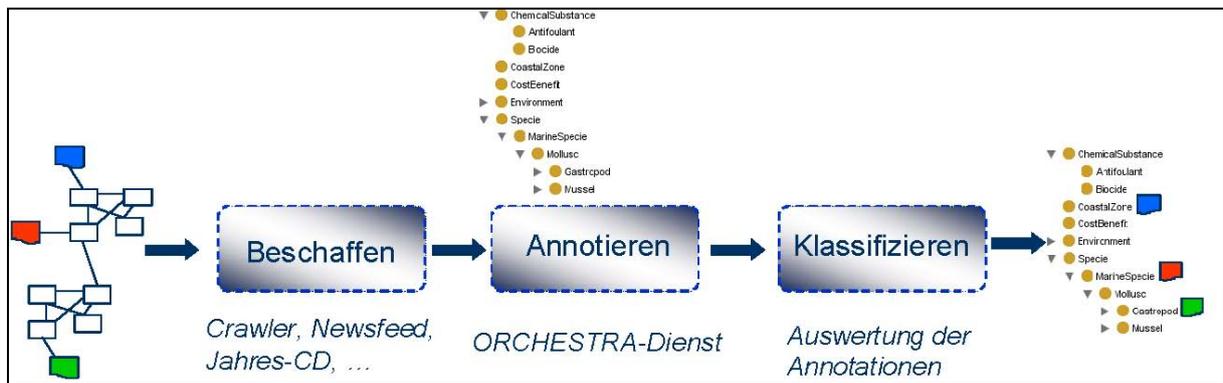


Abbildung 3: Ontologiebasierte automatische Klassifikation

Die Beschaffung der Dokumente aus der Fachdomäne kann – je nach Anforderung – durch Hintergrundprozesse (z.B. Crawler, Newsfeeds) mit anschließender Indizierung erfolgen, so dass neu hinzukommende Dokumente automatisch in die Ablage gestellt werden. Alternativ können auch Medien wie z.B. Jahres-CDs eingelesen werden.

## 7 Klassifikation auf Basis von Trainingsdaten

Kann das Anwendungsschema nicht in eine Ontologie umgesetzt werden oder ist überhaupt kein Schema vorhanden, besteht die Alternative, einen Teil des bisherigen Dokumentenbestandes durch Fachexperten manuell vorzuklassifizieren. Idealerweise wird der Bestand schon manuell in einer solchen Ablagestruktur geführt.

Sind diese Voraussetzungen gegeben, kann diese Ablage als Trainingsdatensatz für die Generierung eines Modells verwendet werden, welches neu hinzukommende Dokumente dann automatisch klassifizieren kann (Abbildung 4).

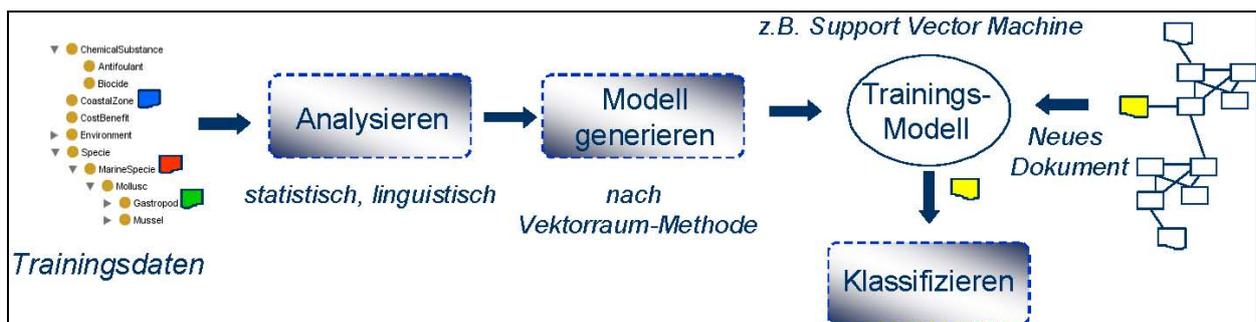


Abbildung 4: Statistische automatische Klassifikation nach dem Vektorraummodell

Die Dokumente müssen dazu statistisch aufbereitet werden, d.h. aus den Dokumenten werden Terme extrahiert, die das Dokument in Bezug auf eine durchzuführende Klassifikation beschreiben. Diese Termliste kann durch linguistische Analyse – dazu gehört z.B. die Rückführung der Terme auf deren Stammform – weiter reduziert werden. Mit Hilfe einer Gewichtung der im Dokument vorkommenden Terme wird jedes Dokument in Form eines Vektors nach dem Vektorraummodell beschrieben. Mit Hilfe einer speziellen Lern-Software wird aus den so repräsentierten Dokumenten ein Modell erzeugt, das in der Lage ist, neu hinzukommende Dokumente automatisch zu klassifizieren.

Hat man den Trainingsdatensatz erstellt, können die Schritte zur Analyse und Modellgenerierung weitgehend maschinell durchgeführt werden, geeignete Werkzeuge stehen hierfür zur Verfügung (z.B. „Rapid Miner“ [7]).

## 8 Schlussbemerkung

Die Entwicklung raumbezogener Informationssysteme ist bis heute maßgeblich durch die Nutzung von OGC-Standards vorangetrieben. Diese reflektieren die Anforderungen der Nutzer und haben zur Realisierung verteilter Systeme mit Interoperabilität auf syntaktischer Ebene geführt. Es gibt zwar auch Arbeiten über semantische Themen in den OGC Standardisierungsgremien, die Hauptantriebskraft zur semantischen Interoperabilität kommt aber heute noch hauptsächlich von der Semantic Web Initiative und der Standardisierung im Bereich der W3C Gemeinde. Technologien aus dem Bereich des Semantic Web sind heute noch weitgehend akademische Entwicklungen, von Seite der Endanwender im Bereich der Umwelt-Informationssysteme wird darauf noch recht wenig Einfluss genommen.

Im Rahmen des ORCHESTRA-Projektes wurde angestrebt, einen Beitrag zur Zusammenführung dieser heute noch getrennten Welten zu leisten und Initiativen mit praktischem, anwendergetriebenen Hintergrund zu fördern. Dies hat sich als anspruchsvolle Aufgabe herausgestellt; die Entwicklung und Implementierung der ORCHESTRA-Architektur, die in der Lage ist, semantische Dienste unter Beibehaltung der Kompatibilität zu OGC-Standards zu integrieren, hat aber zu einem „proof-of-concept“ beigetragen. Positive Rückmeldungen in Anwender-Workshops lassen den Schluss zu, dass die Integration semantischer Dienste zu einem signifikanten Mehrwert führt und den Weg zu grenz- und domänenübergreifender Interoperabilität öffnet.

## 9 Literatur

- [1] Integriertes Projekt ORCHESTRA (Open Architecture and Spatial Data Infrastructure for Risk Management), IST FP6-511679, Laufzeit 9/2004-2/2008, <http://www.eu-orchestra.org>
- [2] Bügel, U.; Usländer, T.; Denzer, R.: Europäisches Risikomanagement – Architekturansatz und Pilotanwendungen des ORCHESTRA-Projektes. AK Umweltdatenbanken der GI-Fachgruppe 4.6.1, Höxter, 15.5.2006.
- [3] Usländer, T. (ed.), 2007. Reference Model for the ORCHESTRA Architecture Version 2 (Rev. 2.1). OGC Best Practices Document 07-097. [http://portal.opengeospatial.org/files/?artifact\\_id=23286](http://portal.opengeospatial.org/files/?artifact_id=23286).
- [4] Lieberman, J., Pehle, T., Dean, M. 2007: Semantic Evolution of Geospatial Web Services. <[http://www.w3.org/2005/04/FSWS/Submissions/48/GSWS\\_Paper.html](http://www.w3.org/2005/04/FSWS/Submissions/48/GSWS_Paper.html)>, 2007/03/01.
- [5] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, C. Ursu, M. Dimitrov, M. Dowman, N. Aswani, and I. Roberts. Developing Language Processing Components with GATE Version 4 (a User Guide), 2007.
- [6] David Vallet, Miriam Fernández, and Pablo Castells: An Ontology-Based Information Retrieval Model. European Semantic Web Conference (ESCW) 2005: Heraklion, Crete, Greece
- [7] [http://rapid-i.com/component/option,com\\_frontpage/Itemid,1/lang,de/](http://rapid-i.com/component/option,com_frontpage/Itemid,1/lang,de/)

- [8] Kopp, Alina: Design and Implementation of an Automatic Semantic Annotation Service. Diplomarbeit. Universität des Saarlandes / Fraunhofer IITB, Karlsruhe, 2007
- [9] Integriertes Projekt SANY (Sensors Anywhere), IST FP6-033564, Laufzeit 9/2006-8/2009, <http://www.sany-ip.eu>, Deliverable D2.3.2 Specification of the Sensor Service Architecture V0

# Nutzung von Ontologien zur Informationsstrukturierung im Themenpark Umwelt

Claudia Greceanu, Clemens Döpmeier<sup>16</sup>

[claudia.greceanu@iai.fzk.de](mailto:claudia.greceanu@iai.fzk.de); [clemens.duepmeier@iai.fzk.de](mailto:clemens.duepmeier@iai.fzk.de)

Renate Ebel<sup>17</sup>

[renate.ebel@lubw.bwl.de](mailto:renate.ebel@lubw.bwl.de)

## Abstract

The Theme Park Environment is a web-based environmental information system for the general public developed within the Environmental Information System Baden-Württemberg (UIS). It serves as a platform for the presentation of environmental information with close reference to Baden-Württemberg for the interested public.

This paper presents how authors can easily combine information objects to composed structured presentations of topics in the Theme Park by using an ontology model and the underlying web framework WebGenesis®. The ontology defines a semantically rich structure to the otherwise largely unstructured multimedial information parts (web pages) of the Theme Park. Object properties (i.e. relations) between information objects within the ontology are not only used for dynamically generating the main navigation structures in the Theme Park Environment but also allow easy reuse of information parts in different information contexts.

## Zusammenfassung

Der „Themenpark Umwelt“ ist ein web-basiertes Umweltinformationssystem für die Öffentlichkeit und wird im Rahmen des Umweltinformationssystems Baden-Württemberg (UIS) entwickelt. Er dient als Plattform für die Präsentation von Umweltinformationen mit engem geographischen Bezug zu Baden-Württemberg für die interessierte Öffentlichkeit.

Der vorliegende Beitrag beschreibt, wie Autoren die Informationsobjekte im Themenpark Umwelt auf einfache Art und Weise zu zusammengesetzten strukturierten Präsentationen von Themen durch Nutzung eines Ontologiemodells und des zugrunde liegenden Web-Frameworks WebGenesis® verknüpfen können. Die Ontologie definiert eine semantikkreiche Struktur zu den ansonsten größtenteils unstrukturierten multimedialen Informationsbestandteilen (Webseiten) des Themenparks. Objekt-Eigenschaften (d. h. Beziehungen) zwischen Informationsobjekten innerhalb der Ontologie werden nicht nur für die dynamische Generierung der Hauptnavigation im Themenpark Umwelt genutzt, sondern erlauben auch die

---

<sup>16</sup> Forschungszentrum Karlsruhe, Institut für Angewandte Informatik

<sup>17</sup> Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg

einfache Wiederverwendung von Informationsteilen in unterschiedlichen Informationskontexten.

## 1 Einleitung

Der „Themenpark Umwelt“ ist ein web-basiertes Umweltinformationssystem für die Öffentlichkeit und wurde im Auftrag des Umweltministeriums Baden-Württemberg vom Institut für Angewandte Informatik des Forschungszentrums Karlsruhe und der Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg (LUBW) gemeinsam mit anderen Projektpartnern im Rahmen des Umweltinformationssystems Baden-Württemberg (UIS) entwickelt.

Der Themenpark Umwelt dient als Plattform für die Präsentation von Informationen für die interessierte Öffentlichkeit mit den Schwerpunkten Boden, Geologie und Natur und vermittelt Umweltinformationen mit engem geographischen Bezug zu Baden-Württemberg in multimedialer und leicht verständlicher Weise im Unterschied zu Umweltfachinformationssystemen.

### Motivation für Ontologien

Bei der Entwicklung eines ersten Prototyps des Themenparks wurde – am Beispiel des Bodens – im Rahmen einer Diplomarbeit gezeigt, dass Umweltinformationssysteme für die breite Öffentlichkeit anderen Anforderungen genügen müssen als dies für fachbasierte Umweltinformationssysteme gilt [Ruchter, 2002], [Ruchter/Düpmeier, 2002].

Um die interessierte Öffentlichkeit als Zielgruppe zu erreichen und ihr Umweltsachverhalte zu vermitteln, ist der regionale Bezug der Inhalte und auch der inhaltliche Querbezug zu anderen den Bürgern interessierenden Themen (aktuelle Umweltpolitik, Kulturhistorie etc.) von großer Bedeutung. Ein direkter örtlicher Bezug sowie eine ansprechende Darstellung unter Verwendung von Multimedia-Elementen wie Bildern und Videos erscheint am besten geeignet, einerseits das Umwelt- bzw. Bodenbewusstsein zu fördern und andererseits die Neugier auf die Umwelt „direkt vor der Haustür“ zu wecken. Der Betrachter wird im Themenpark animiert, nach draußen zu gehen und die im Themenpark Umwelt vorgestellten Umwelterlebnisorte selbst zu erleben. Die vermittelten Umweltthemen sind dabei in Anlehnung an den Umweltplan Baden-Württembergs mit den politisch bedeutsamen Umweltthemen in Baden-Württemberg verknüpft.

Um die Verknüpfbarkeit der Themenparkinhalte mit anderen Informationen – geobasierten, anderen thematischen Angeboten im Internet, Sachdaten aus Sachdaten-systemen des gesamten Umweltinformationssystems Baden-Württembergs (UIS) – zu gewährleisten, wurde bei der Entwicklung des Themenparkkonzeptes nach einer Möglichkeit gesucht, kleinere Informationsbestandteile (Multimediaelemente, textuelle Beschreibungsseiten) und ihre Verknüpfung miteinander in einer semantikhreichen Art und Weise zu beschreiben.

Das Web-Framework und Content-Management-System WebGenesis® des Fraunhofer-Instituts für Informations- und Datenverarbeitung IITB verfügt über eine solche Beschreibungsmöglichkeit in Form eines integrierten Ontologie-Datenmodells. Dieses Ontologie-Datenmodell wurde für die Implementierung des Themenparks

genutzt. Im Folgenden sollen das hierbei verwendete Ontologiemodell und seine Vor- und Nachteile genauer erläutert werden.

## 2 Das Ontologiemodell des Themenparks Umwelt

### 2.1 Ontologien in WebGenesis

Das Ontologiekonzept von WebGenesis ist an die Struktur der OWL (Web Ontology Language<sup>18</sup>) angelehnt und ermöglicht Erstellern von Webanwendungen, unstrukturierten Informationen wie multimedialen Texten semantische Strukturinformationen auf Basis eines Ontologiemodells zuzuordnen [WebGenesis, 2007]. Eine Ontologie beschreibt eine Vererbungshierarchie von Begrifflichkeiten („Konzepten“), denen einfache Informationsattribute (DatatypeProperties, Datentyp-Eigenschaften) und Beziehungen zwischen Konzepten (ObjectProperties, Objekt-Eigenschaften) zugeordnet werden können.

Ontologien sowie deren Konzepte und Objekt-Eigenschaften können ohne Programmierung in der Autorenumgebung eines WebGenesis-Systems angelegt werden. Die Ontologien werden dabei normalerweise nicht von Autoren erstellt, sondern von Systembetreiberseite vordefiniert, während die konkreten Individuen und die Beziehungen zwischen Individuen implizit durch die Autoren des Systems beim Einfüllen von Informationen angelegt werden. Hierzu werden Webinhalten des WebGenesis-Systems von Autoren Konzepte der Ontologie zugeordnet, die zugehörigen Metadaten (DatatypeProperties) aufgefüllt und hierbei neue Individuen erzeugt. Weiter definieren Webautoren Beziehungen zwischen Webinhalten (Informationsobjekten) und damit implizit Relationen zwischen Konzepten der Ontologie bzw. deren Individuen.

### 2.2 Nutzung des Ontologiemodells von WebGenesis im Themenpark Umwelt

Das Ontologiekonzept des „Themenparks Umwelt“ besteht aus zwei ganz unterschiedlichen Konzeptarten: generische Konzepte zur Informationsstrukturierung wie Beschreibungsteil, Beschreibungsseite und Multimediaobjekt sowie fachspezifische Konzepte.

Modelliert wurden fachspezifische Konzepte wie Moor, Geotop, Schutzgebiet, Bodenlandschaft etc. Diese können je nach Konzeptklasse spezifische Metadaten enthalten, die Individuen dieser Konzepte u. a. mit Sachdaten aus dem Sachdatensystem des UIS Baden-Württembergs verknüpfen:

- FFC<sup>19</sup>, OAC<sup>20</sup> und OID<sup>21</sup>: genormte Schlüsselattribute im Rahmen des UIS Baden-Württemberg für Umweltobjekte wie Moore, Geotope und Schutzgebiete im Zugang „Unsere Umwelt“ (durch diese Angaben lassen sich die raumbezogenen Umweltobjekte eindeutig definieren)
- Längen- und Breitengrad für Erlebnisorte und für Medienelemente mit Ortsbezug (Bilder und Videos)
- Autor- und Copyrightangaben für Medienelemente

---

<sup>18</sup> <http://www.w3.org/2004/OWL/>

<sup>19</sup> Fachführungscode

<sup>20</sup> Objektartencode

<sup>21</sup> Objektidentifikator

Die fachspezifischen Konzepte sind nicht nur über ihre Hierarchie, sondern auch über Objekt-Eigenschaften (im Geokontext z. B. Relationen wie „besteht aus“, „enthält“, „wird untergliedert in“) miteinander und mit generischen Konzepten der Themenpark-Ontologie verknüpfbar.

Beispielhaft für den Zugang „Unsere Umwelt“ des Themenparks lassen sich in einem Diagramm – angelehnt an die UML<sup>22</sup>-Notation eines Klassendiagramms – die fachspezifischen Konzepte und deren Abhängigkeiten zeigen:

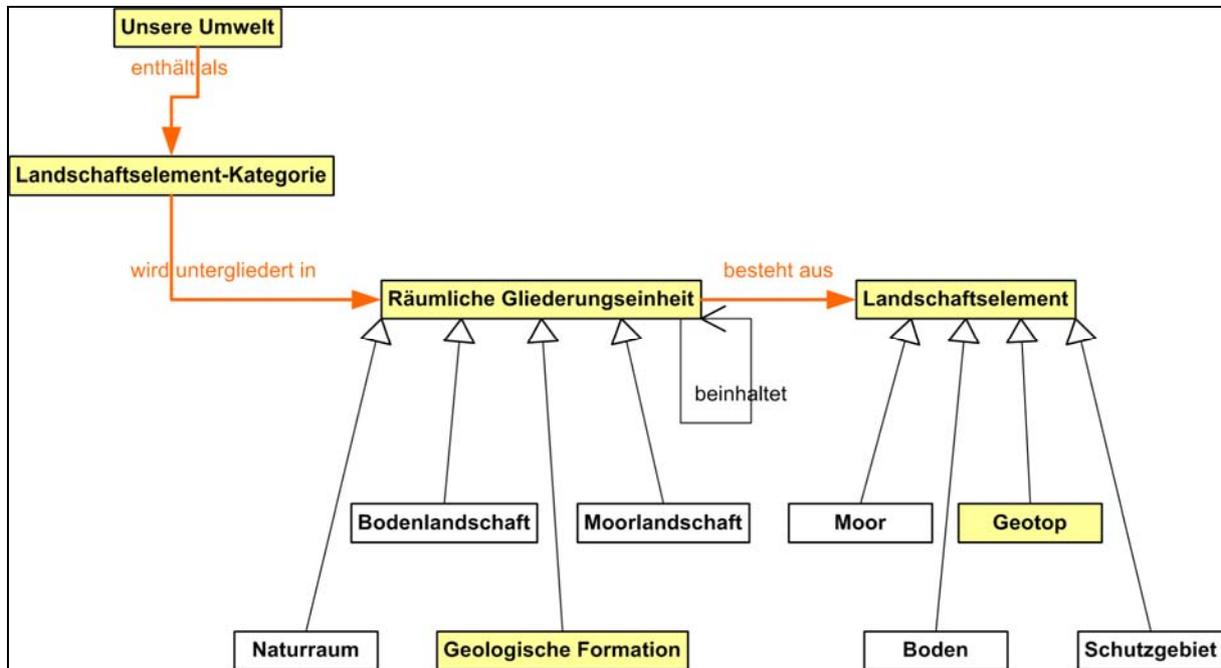


Abbildung 1: Ontologie und fachspezifische Konzepte des Zugangs „Unsere Umwelt“

Darüber hinaus gibt es wie eingangs erwähnt generische Konzepte für wiederverwendbare Seitenbestandteile wie Medienelement (Bild, Video), oder größere Beschreibungseinheiten wie Beschreibungsteil, Beschreibungseite, Artikel etc., d. h. Elemente, mit denen sich Dokumentstrukturen gestalten lassen. Abbildung 2 zeigt die Zusammenhänge und Abhängigkeiten der übergreifenden Ontologie und der generischen Konzepte:

<sup>22</sup> Unified Modeling Language

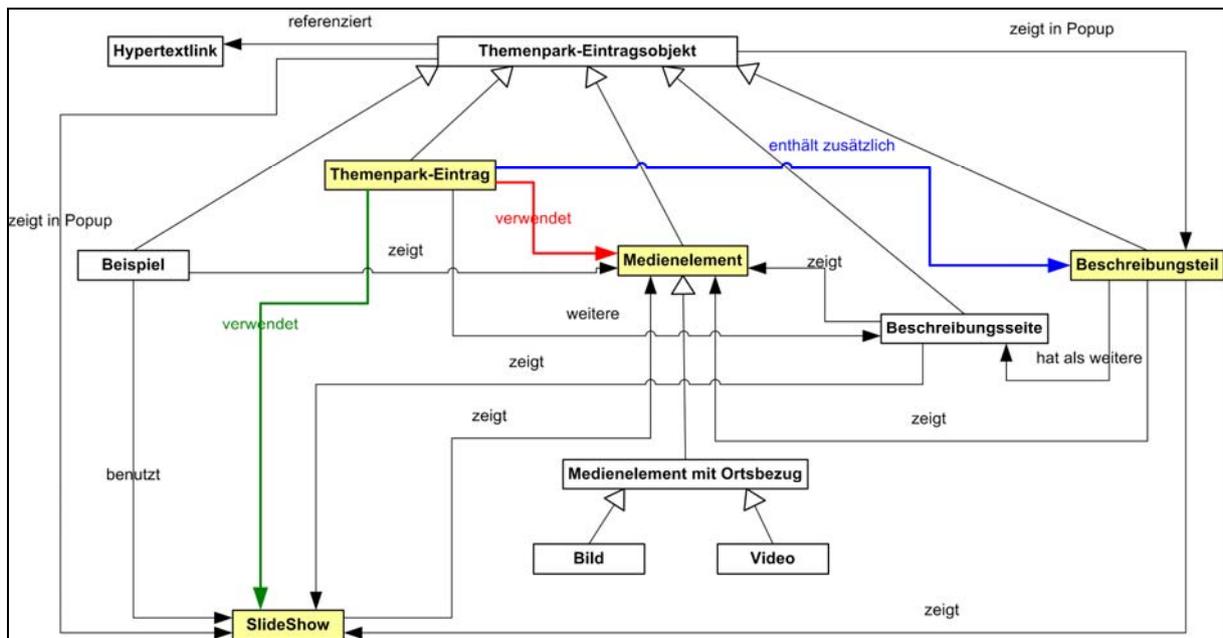


Abbildung 2: Übergreifende Ontologie und generische Konzepte als Diagramm

Mittels der vordefinierten Objekt-Eigenschaften können die Autoren sowohl fachspezifische Konzepte – bzw. die Instanzen solcher Konzepte – über Relationen untereinander verknüpfen (Bsp. Schutzgebiet gehört zu Räumliche Gliederungseinheit) als auch fachspezifische Konzepte mit generischen Konzepten (Bsp. Themenpark-Eintrag verwendet Medienelement) verbinden.

Eine solche Inhaltsorganisation könnte beispielsweise für ein Moor folgendermaßen aussehen: Ein Moor besteht aus mehreren Beschreibungsteilen (die wiederum aus mehreren Beschreibungsseiten bestehen können). Beschreibungsteile sowie –seiten können dabei Text und verschiedene Medienelemente (Bilder, Videos) enthalten. Beschreibungsteilen und -seiten kann dabei eine inhaltliche Klassifikation (z. B. „Entstehung“, „Nutzung“, „Kulturgeschichte“, „Pflanzen- und Tierwelt“ etc.) zugeordnet werden, die multimediale textuelle Informationen dadurch semantisch einordnet.

Die Zuordnung der Beschreibungsteile und -seiten zu Fachobjekten erfolgt dabei über die Objekt-Eigenschaften „Themenpark-Eintrag enthält zusätzlich Beschreibungsteil“ und „Beschreibungsteil hat als weitere Beschreibungsseite“, die Zuordnung der Medienelemente zu Beschreibungsteil und –seite über die Objekt-Eigenschaften „Beschreibungsseite zeigt Medienelement“ bzw. „Beschreibungsteil zeigt Medienelement“.

Die Ontologie wird im Themenpark Umwelt nun u. a. dazu genutzt, um Navigationsstrukturen dynamisch anzulegen. Dabei sind die logischen Organisationsstrukturen der Autoren (Inhaltsablage) strikt von den Navigationsstrukturen und inhaltlichen Beziehungsstrukturen auf der Präsentationsebene getrennt. Die inhaltliche und semantische Navigation zwischen Inhaltsobjekten im Themenpark wird auf Präsentationsebene ausschließlich über die Beziehungstypen der Ontologie aufgebaut. Hierbei werden Objekte über Beziehungen wie „Themenbereich enthält Thema“ verknüpft; solche Relationen erlauben es dann die üblichen hierarchisch- baumartigen Navigationsstrukturen aufzubauen, wie folgendes

Beispiel zeigt: Zugang „Unsere Umwelt“ → Schutzgebiete → Naturraum „Mittleres Oberrhein-Tiefland“ → Naturschutzgebiet „Taubergießen“<sup>23</sup>.

Beziehungen wie „Beispiel gehört zu Thema“ erlauben es, ein Objekt vom Typ Beispiel (z. B. die Beschreibung eines Weinanbaugebietes) mit beliebigen anderen Objekten vom Typ Thema (z. B. Umweltthema „Boden liefert Nahrungsmittel“) zu verknüpfen, wie in Abbildung 3 exemplarisch dargestellt wird:

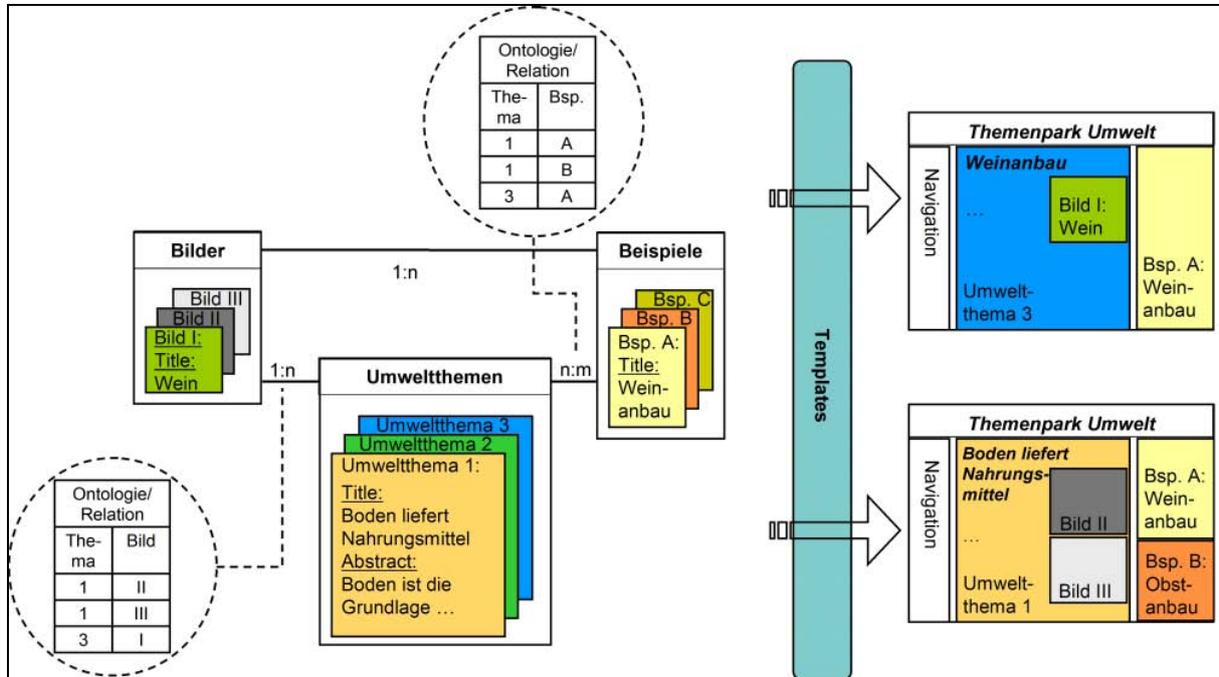


Abbildung 3: Beziehungen zwischen Objekten (verschiedener Autoren)

Solche Beziehungen können dazu genutzt werden, um Informationsobjekte in unterschiedlichen inhaltlichen Kontexten mehrfach zu nutzen. Damit wird im Themenpark Umwelt das wichtige Ziel der Wiederverwendbarkeit von Informationsobjekten erreicht. Auf Autorenebene nutzen die Autoren dabei die standardmäßig hierarchisch über Verzeichnisse und WebGenesis-Inhaltsobjekte definierte Organisationsstruktur zur Ablage ihrer Inhalte. Außerdem ermöglicht die Themenpark-Ontologie den Autoren, auch Informationselemente, die von anderen Autoren angelegt wurden, schnell aufzufinden und für eigene Inhalte zu verknüpfen<sup>24</sup>. Durch dieses Konzept bleiben die Autoren der verschiedenen Institutionen in der Autorenumgebung durch unterschiedliche Verzeichnisse voneinander getrennt und sie können auch bezüglich des Zugriffsschutzes einfach verwaltet werden.

Die durch die Autoren erstellten Relationen bilden ein Netzwerk zwischen den Informationsobjekten. Die Templates zur Navigationsstrukturierung in den verschiedenen Zugängen (Unsere Umwelt, Umweltthemen, Umwelt beobachten, Umwelt erleben) können die Relationen auswerten und somit die Navigationsstruktur zwischen den Informationsobjekten dynamisch darstellen und Informationsseiten aus verschiedenen Informationsobjekten zusammenbauen. Somit beruht die Navigation in der Präsentationsebene auf der Ontologie bzw. den konkreten Relationen, aber nicht auf der physikalischen Lage der Informationsobjekte innerhalb der Autorenumgebung. Dadurch ist die Navigation auf der Präsentationsseite auch

<sup>23</sup> <http://www.themenpark-umwelt.baden-wuerttemberg.de/servlet/is/9765/?path=4422;6350;9760>

<sup>24</sup> sofern der Autor ausreichende Rechte vergeben hat, damit andere Benutzer darauf zugreifen dürfen

unabhängig von einer Verschiebung der Inhaltsobjekte innerhalb der Autoren-Ablagestruktur.

Am Beispiel des Geotops „Eisinger Loch“<sup>25</sup> innerhalb der Geologischen Formation des Muschelkalks soll der Zusammenbau von Informationsbausteinen zu einer Seite samt Navigation erläutert werden: Abbildung 4 zeigt als Diagramm – angelehnt an die UML-Notation eines Objektdiagramms – die Verknüpfung der Instanzen der verwendeten Konzepte an:

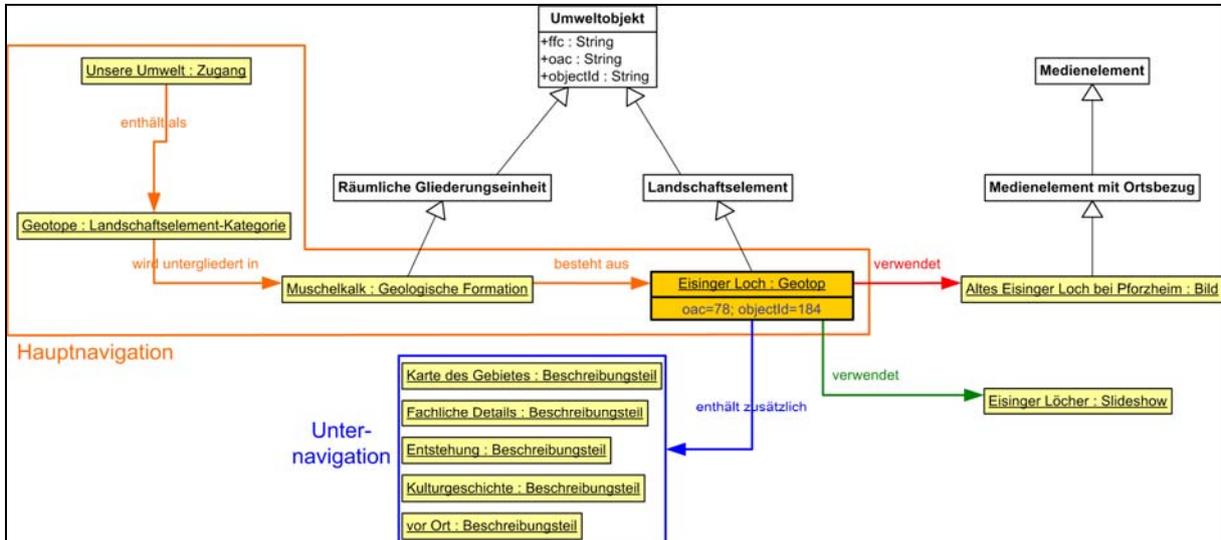


Abbildung 4: Diagramm für Geotop „Eisinger Loch“

Abbildung 5 zeigt die generierte Präsentationssicht mit der Hauptnavigation im Kopfbereich und in der linken Spalte sowie der Unternavigation in der rechten Spalte:

<sup>25</sup> <http://www.themenpark-umwelt.baden-wuerttemberg.de/servlet/is/6335/?path=4422:6277:6293>

The screenshot shows the 'Themenpark Umwelt' website interface. At the top, there are logos for the Baden-Württemberg state government, the 'Umweltinformationssystem Themenpark Umwelt', and other partners like 'igkb', 'LGRB', and 'LUBW'. The navigation menu includes 'Eingangseite', 'Unsere Umwelt', 'Umweltthemen', 'Umwelt beobachten', 'Umwelt erleben', and 'Kontakt'. The main content area is titled 'Eisinger Loch' and is part of the 'Geotope' > 'Muschelkalk' > 'Eisinger Loch' hierarchy. It features a map of Baden-Württemberg with a red pin indicating the location. Below the map is a search bar and a list of geotopes to choose from, including 'Bodenlandschaften', 'Geotope', 'Moore', 'Schutzgebiete', and 'Bodensee'. The 'Geotope' section is expanded to show 'Muschelkalk' and 'Eisinger Loch'. The main text describes the 'Altes und Neues Eisinger Loch' as karst formations and collapse dolines in the Muschelkalk. It mentions that these are located about 10 km north of Pforzheim and have been designated as a natural monument since 1985. A photograph shows the entrance to the 'Altes Eisinger Loch' in a forest. The page also includes a sidebar with 'Eisinger Loch' and 'Altes Eisinger Loch - Eingangsbereich' and a 'Weitere Informationen' section with links to 'Karte des Gebietes', 'Fachliche Details', 'Entstehung', 'Kulturgeschichte', and 'vor Ort'.

Abbildung 5: Geotop „Eisinger Loch“ innerhalb der Geologischen Formation des Muschelkalks

Für die Umweltobjekte mit Raumbezug im Zugang „Unsere Umwelt“ ist der Kartenservice der LUBW integriert [Düpmeyer, 2006].

Durch die für dieses Geotop vom Autor ausgefüllten spezifischen Metadaten FFC, OAC und OID (s. Kap. 2.2) werden in der Präsentationsebene in der rechten Spalte automatisch die zwei Links „Karte des Gebietes“ sowie „Fachliche Details“ generiert, die dem Nutzer über eine serviceorientierte Schnittstelle den Zugriff auf die Fachkarte des Gebietes und auf die Sachdaten gewähren.

### 3 Fazit und Ausblick

Der vorliegende Beitrag zeigt, wie durch Nutzung des Ontologiemodells von WebGenesis im Themenpark Umwelt flexible Navigationsstrukturen auf Präsentationsebene unabhängig von der Lage der durch die Autoren organisierten Inhaltsobjekte in der Ablagestruktur realisiert werden können. Das Ontologiekonzept ist weiter für die Wiederverwendbarkeit bestimmter Informationsobjekte durch die Verwendung von für Autoren vorgegebene Beziehungstypen sowie für die Verknüpfung von unstrukturierten mit strukturierten (Fach-)Daten nützlich.

Als eher nachteilig zu bewerten ist, dass die Verknüpfung zwischen Informationsobjekte keinen Baum bildet, d. h. durch die Wiederverwendbarkeit ergeben sich mehrdeutige Navigationsstrukturen, was die Programmierung komplizierter gestaltet.

Dennoch hat sich die Nutzung des Ontologiekonzepts im Themenpark Umwelt so bewährt, dass ein solches Konzept auch in dem Fachdokumentenmanagement im UIS BW (FADO<sup>26</sup>) angewendet wird.

Noch können im Themenpark die Potenziale von Ontologien nicht völlig ausgeschöpft werden, da der Ontologie-Dienst von WebGenesis kein volles OWL-Modell unter Einschluss eines Reasoners<sup>27</sup> implementiert. Daher lassen sich nur bedingt Schlussfolgerungen aus bestehenden Strukturen ziehen, mit denen die Ontologie selbstständig neue Beziehungen generieren kann. Die Themenpark-Ontologie ist daher im Wesentlichen durch die Autoren manuell zu pflegen. Hier steckt aber durchaus noch einiges Potenzial im algorithmischen Bereich. So können z. B. räumliche Bezüge zwischen Themenparkobjekten automatisch durch GIS-Berechnungen (z. B. zur räumlichen Nähe) abgeleitet und über eine Beziehung „befindet sich in der Nähe von“ in die Themenpark-Ontologie eingepflegt werden.

Des Weiteren könnte man durch semantische Attributierung von Informationen im gesamten UIS Baden-Württemberg basierend auf einem UIS-weiten Ontologiekonzept automatisch Beziehungen von Fachobjekten im Themenpark mit Informationseinheiten anderer Systeme z. B. über einen semantischen Crawler aufbauen.

Der Themenpark Umwelt ist unter folgender Adresse erreichbar:

<http://www.themenpark-umwelt.baden-wuerttemberg.de/>

#### 4 Literaturverzeichnis

[Ruchter, 2002] Ruchter, M.: Web-basierte Umweltinformationssysteme für die Öffentlichkeit: Konzepte, Potentiale und Realisierbarkeit. Forschungszentrum Karlsruhe, Wissenschaftliche Berichte, FZKA-6760, 2002, <http://bibliothek.fzk.de/zb/berichte/FZKA6760.pdf>

[Ruchter/Düpmeier, 2002] Ruchter, M.; Düpmeier, C.; Geiger, W.: Web-basierte Bodeninformationssysteme für die Öffentlichkeit: Konzepte, Potentiale und Realisierung am Beispiel des Themenpark Boden. GI Workshop Umweltdatenbanken, Ilmenau, 13.-14. Juni 2002

[Düpmeier, 2006] Düpmeier, C.; Geiger, W.; Greceanu, C. et al.: Themenpark Umwelt. Integration von anderen Informationssystemen in das Internetportal Themenpark Umwelt. In: Mayer-Föll, R., Keitel, A., Geiger, W. (Hrsg.): F+E-Vorhaben KEWA. Kooperative Entwicklung wirtschaftlicher Anwendungen für Umwelt und Verkehr in neuen Verwaltungsstrukturen. Phase I 2005/06. Forschungszentrum Karlsruhe, Wissenschaftliche Berichte, FZKA-7250, S.105-10, 2006, <http://bibliothek.fzk.de/zb/berichte/FZKA7250.pdf>

[Düpmeier, 2007] Düpmeier, C.; Geiger, W.; Greceanu, C. et al.: Themenpark Umwelt - Fortentwicklung des Themenparks Umwelt, Erprobung von Web 2.0 Technologien. 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 2006/07. Forschungszentrum Karlsruhe, Wissenschaftliche Berichte, FZKA-7350, S. 45-52, 2007, <http://bibliothek.fzk.de/zb/berichte/FZKA7350.pdf>

[WebGenesis, 2007] WebGenesis-Hilfe <http://www.iitb.fraunhofer.de/servlet/is/83/>

---

<sup>26</sup> Fachdokumente Online, Nachfolgesystem für die XfaWeb-Systemfamilie

<sup>27</sup> [http://www.ag-nbi.de/lehre/07/S\\_MWT/Material/05\\_ReasonersFrameworks.pdf](http://www.ag-nbi.de/lehre/07/S_MWT/Material/05_ReasonersFrameworks.pdf)



# 5 Jahre Semantic Network Service (SNS) Aktueller Stand und Ausblick

Maria Rüther, Umweltbundesamt, [maria.ruether@uba.de](mailto:maria.ruether@uba.de)

Thomas Bandholtz, innoQ Deutschland GmbH, [thomas.bandholtz@innq.com](mailto:thomas.bandholtz@innq.com)

## Abstract/ Einleitung

Semantic Network Service (SNS) provides shared terminology access and indexing services for environmental and geospatial information systems since five years (<http://www.semantic-network.de>). Having started closely related to the German Environmental Information Network (GEIN), SNS has emerged as a trusted reference for many systems, such as the German Environmental Information Portal (PortalU), the German Spatial Data Infrastructure (GeoPortal.Bund) on the federal level, and others on the state level. From applying the librarian approach of subject indexing to the Web, the focus has moved towards a more intuitive full text search assistance, and is currently shifting to semantic data integration.

Seit fünf Jahren bietet der Semantische Netzwerk Service (SNS) Umwelt- und raumbezogenen Informationssystemen Zugang zu einer gemeinsamen Fachterminologie und zu semantischen Diensten (<http://www.semantic-network.de>). Aus einer anfänglich engen Kopplung mit dem Umweltinformationsnetz Deutschland (German Environmental Information Network, GEIN) hat sich SNS zu einer bewährten Referenz für zahlreiche Systeme entwickelt, darunter das Umweltportal Deutschland (PortalU) und das Portal der Geodaten-Infrastruktur (GeoPortal.Bund) auf nationaler Ebene, sowie weitere auf Länderebene. Der Schwerpunkt von SNS hat sich dabei von einer dokumentarischen Erschließung (Verschlagwortung) zu einer mehr intuitiven Unterstützung der Volltextsuche verlagert, und bewegt sich derzeit weiter in Richtung einer semantischen Datenintegration.

## 1 Einführung

Seit 2003, als SNS in Produktion ging, haben das Semantic Web und Ontologiebasierte Ansätze auch in der Umweltinformatik zunehmend Beachtung gefunden. Zahlreiche Vokabulare sind entstanden und weiterentwickelt worden, und dabei wurde mit neuen formalen Sprachen und verwandten Technologien experimentiert. Nationale und europäische Strategien wie Infrastructure for Spatial Information in Europe (INSPIRE), Shared Environmental Information System (SEIS) oder Single Information Space for the Environment in Europe (SISE) verlangen nach einer semantischen Grundlage, und eine Vielzahl von Konferenzen und Forschungsprojekten hat sich diesen Themen gewidmet und zu ersten Lösungen beigetragen. Diese Entwicklung gibt uns Anlass zu einer Zwischenbilanz nach 5 Jahren des erfolgreichen Betriebs und zu einem Blick in die Zukunft.

## 2 Rückblick (1998-2003)

Die Ursprünge von SNS gehen zurück auf das Umweltinformationsnetz Deutschland (German Environmental Information Network GEIN, 1998-2006, heute abgelöst durch PortalU). GEIN hatte die thesaurus-basierte Erschließung der Dokumentation auf Umweltinformation im Web angewendet. Die Ziele von SNS bestanden darin, das ursprüngliche Vokabular und die semantischen Verfahren von GEIN weiter zu entwickeln und die Ergebnisse in Form von Web Services verfügbar zu machen, um so ein gemeinsames Vokabular der deutschen Umweltinformation als Bezugsrahmen der Integration zu ermöglichen.

### 2.1 Die Ursprünge: Thesaurus-basierte Verschlagwortung in GEIN 2000

Das Forschungsprojekt GEIN 2000<sup>28</sup> wurde 1998 begonnen, gefördert vom Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit / Umweltbundesamt. Das Ergebnis war das erste gemeinsame öffentliche Umweltportal der deutschen Umweltbehörden im Internet und als solches ein Beitrag zur Weltausstellung EXPO 2000. GEIN blieb bis 2006 der zentrale öffentliche Zugang zur Umweltinformation, und wurde dann durch das heutige PortalU abgelöst.

GEIN 2000 hat zu seiner Zeit eine ganze Reihe von Innovationen in der Umweltinformatik eingeführt, so z.B. den übergreifenden Einsatz von Internet-Suchmaschinen auf den Web-Seiten aller Informationsanbieter oder die Anbindung von Datenbanken mit XML-Protokollen über HTTP. Dieser Beitrag fokussiert auf die dritte Innovation: die systematische Anwendung eines gemeinsamen Fachvokabulars.

In den späten 1990ern bestand dieses Vokabular zunächst aus dem UMTHES®<sup>29</sup> Thesaurus des Umweltbundesamts, der für die inhaltliche Erschließung von Fachliteratur und der Forschungsdatenbank des Amtes entwickelt worden war. Für GEIN 2000 sollte jedoch ein „dreidimensionaler“ Index über „Thema, Raum und Zeit“ aufgebaut werden. Um alle drei Dimensionen mit einem Vokabular zu hinterlegen, übernahm GEIN den UMTHES® für „Thema“ und entwickelte die beiden fehlenden Komponenten selbst:

- Der „Geo-Thesaurus“ entstand in Zusammenarbeit mit dem Geographischen Informationssystem Umwelt (GISU)<sup>30</sup> und enthielt neben den administrativen Gebietseinheiten (Gemeinde, Kreise usw.) auch verschiedene Kategorien von Schutzgebieten, Berge, Flüsse, Seen, Meeresteile, Wassereinzugsgebiete und Landschaften sowie die Lagebeziehungen all dieser Gebiete untereinander.
- Die Umweltchronologie verzeichnete wichtige Ereignisse in der Entwicklung der Umwelt und verknüpfte die gebräuchlichen Bezeichnungen und Beschreibungen dieser Ereignisse mit dem exakten Zeitpunkt („Was geschah seit Tschernobyl?“).

GEIN 2000 verwendete dieses dreidimensionale Vokabular beim Aufbau des Index. In regelmäßigen Abständen wurde zunächst mit „Crawler“ und „Indexer“ ein aktueller Volltextindex erstellt. Auf dieser Grundlage erzeugte ein vollautomatischer, Thesaurus-basierter Algorithmus einen „Qualifizierten Index“, in welchem alle

<sup>28</sup> UFOPLAN No. 29811603/01, UBA/BMU; Projektdokumente und Abschlussbericht sind verfügbar unter <http://www.kst.portalu.de/archiv/gein/>

<sup>29</sup> <http://www.umweltbundesamt.de/uba-info/dokufabib/thes.htm>

<sup>30</sup> <http://gisu.uba.de>

Webseiten zusätzlich mit dem drei-dimensionalen Vokabular vernetzt wurden – ein einfacher Vorfahr dessen, was heute “text mining” (Heyer, 2006) bzw. “geoparsing” und “geotagging” (Scharl, 2007) genannt wird. Diesem „Qualifizierten Index“ entsprach ein ebenso Thesaurus-basierter „Recherche Assistent“, der den Anwender von seinen umgangssprachlichen Suchbegriffen zum Fachvokabular und von dort zu der zugehörigen Information führte. Dieser Assistent wird weiter unten näher beschrieben, mit Bezug auf die spätere Version von 2003, die bereits von SNS unterstützt wurde.

## 2.2 Das SNS Forschungsprojekt (2001-2003)

Die Idee des SNS Forschungsprojekts<sup>31</sup> bestand darin, den Umgang mit den Vokabularen von GEIN zu erweitern und in einen eigenständigen Web Service auszulagern, der nicht mehr allein einem einzigen Informationssystem zur Verfügung stand. Vielmehr sollte der Umweltinformation in Deutschland generell ein gemeinsames Vokabular und eine Verschlagwortungsmaschine als Integrationswerkzeug zur Verfügung gestellt werden (Angrick, 2002).

Als wir 2001 nach einem geeigneten Standard für die Strukturierung von Thesaurus-übergreifenden Vokabularen suchten, entsprachen Topic Maps (ISO 13250)<sup>32</sup> den Anforderungen am besten. Eine Topic Map besteht allgemein aus Topics, die mit Assoziationen untereinander vernetzt sind. Topics und Assoziationen können typisiert werden, und so brauchten wir nur Topic-Typen wie „Deskriptor“ (Thesaurus), „Ort“ und „Ereignis“ zu definieren, sowie Assoziationstypen wie „Oberbegriff“, „Räumliche Überschneidung“ oder „Verwandtes Ereignis“. Auf diese Weise konnten wir die drei unterschiedlichen strukturierten Vokabulare in eine gemeinsame formale Struktur einbetten, die normalisiert und erweiterbar war. Auf der Grundlage dieser Topic Map implementierten wir die grundlegenden semantischen Methoden und machten sie über Web Services<sup>33</sup> im Internet verfügbar. 2003 ging SNS in Produktion, indem es zunächst die früheren eingebauten Methoden von GEIN ablöste.

Um dieselbe Zeit wurde GEIN überarbeitet und erhielt dabei auch eine neue Oberfläche und den geänderten Namenszug gein®. Abbildung 1 zeigt eine Anwendung des überarbeiteten und nun von SNS unterstützten Recherche-Assistenten. Der Anwender hat in diesem Fall den vollständigen Text einer (englischen<sup>34</sup>) Pressemeldung als Suchbedingung eingegeben (im Bild links oben zu sehen unter „Input:“). gein® hat diesen Text unverändert an den SNS-Indexierungsdienst autoClassify übergeben und zeigt nun die Ergebnisse unter den Rubriken „Topic“ (Thema) und „Area“ (Raum) an. Jeder Begriff funktioniert als Hyperlink, von hier aus kann zu einer weiteren Webseite verzweigt werden, die alle Eigenschaften und Assoziationen des Begriffs anzeigt, und von wo auch weiter durch das Vokabular navigiert werden kann. Zusätzlich ist jeder Begriff mit einer Checkbox versehen, mit der er für die anschließende Suche im „Qualifizierten Index“ ausgewählt werden kann.

---

<sup>31</sup> UFOPLAN 20111612, gefördert von UBA/BMU. Der Abschlussbericht ist verfügbar auf [http://www.semantic-network.de/doc\\_intro.html](http://www.semantic-network.de/doc_intro.html)

<sup>32</sup> <http://isotopicmaps.org/>

<sup>33</sup> <http://www.w3.org/2002/ws/>

<sup>34</sup> UMTHESES® und Umweltchronologie sind durchgängig zweisprachig in Deutsch und Englisch.

Abbildung 1: *gein*® Recherche-Assistent mit SNS Unterstützung (ab 2003)

### 3 SNS Services und ihre Anwender heute

SNS bietet heute sieben Service Operationen, darunter drei grundlegende semantische Methoden, die schon 2003 verfügbar waren, und vier komplexere Methoden, die auf den grundlegenden aufbauen und in den folgenden Jahren entwickelt wurden. Eine vollständige technische Dokumentation ist zum Download verfügbar (SNSWS, 2008).

Der Nutzerkreis dieser Operationen hat sich auf derzeit sechs ausgedehnt, drei weitere befinden sich in Vorbereitung

#### 3.1 Grundlegende Service Operationen (seit 2003)

Drei grundlegende Service Operationen sind seit Beginn der Produktion (Rüther, 2004) verfügbar und werden in optimierter Form auch heute angeboten.

- findTopics
- getPSI
- autoClassify

### 3.1.1 findTopics: Suche nach Fachbegriffen im Vokabular

Suche nach Begriffen anhand von Wörtern. Diese können Namen von Begriffen sein, aber auch in Beschreibungen oder anderen Attributen vorkommen. Die Namen der Begriffe müssen in SNS nicht zwingend eindeutig sein ("Homographen"), und sie können sich aus mehreren Wörtern zusammensetzen. Innerhalb des Thesaurus können auch Flexionsformen der Namen erkannt werden. Dieser Dienst kann durch zahlreiche Optionen genauer eingegrenzt werden. Im Ergebnis der Suche ist auch die Topic ID enthalten, die für die Verwendung von getPSI benötigt wird.

Service Operationen funktionieren über XML Protokolle und sind zur Nutzung durch Maschinen gedacht. Abbildung 2 zeigt einen SNS-Dialog, der eine interaktive Nutzung von findTopics erlaubt. Man sieht hier die subtile Wirkung der Parameter. Der Suchbegriff „uig“ findet mit der hier gewählten Einstellung den Thesaurusbegriff „Bodenraugigkeit“, da diese Buchstabenfolge im Namen (und damit in den Wortformen) enthalten ist. Zugleich wird auch das Ereignis „Bürger erhalten besseren Zugang zu Umweltinformationen“ gefunden, worin die Zeichenfolge „uig“ nicht enthalten ist. Dies liegt daran, dass für Ereignisse ein freierer Umgang mit den Wortformen gepflegt wird. Da Ereignisse in der Regel komplexere Bezeichner („Namen“) tragen, sind gewollt zutreffende Suchbegriffe hier auch als Wortform eingetragen. Will man Treffer wie „Bodenraugigkeit“ ausschließen, sollte man anstatt „Wortteil“ besser „exakte Übereinstimmung“ wählen, dies hängt aber vom jeweiligen Suchbegriff und den Absichten des Anwenders ab.

The screenshot shows a web interface for the findTopics service. It is divided into two main sections: 'Eigenschaften' (Properties) and 'Suchparameter' (Search Parameters).

**Eigenschaften:**

- Suchbegriff: A text input field containing 'uig'.
- Topic-Typ: A dropdown menu currently set to 'Topic'.

**Suchparameter:**

Suche nach	Suche in
<input type="radio"/> exakter Übereinstimmung	<input checked="" type="radio"/> Wortformen
<input type="radio"/> Wortanfang	<input type="radio"/> phonetische Übereinstimmung
<input checked="" type="radio"/> Wortteil	<input type="radio"/> Namen
	<input type="radio"/> allen Datenfeldern

Suchergebnisse	
Topics	
<a href="#">Bodenraugigkeit</a>	Thesauruseintrag > Nichtdeskriptor
<a href="#">Bürger erhalten besseren Zugang zu Umweltinformationen</a>	Ereignis > Rechtliches
<a href="#">Messgenauigkeit</a>	Thesauruseintrag > Deskriptor
<a href="#">Umweltinformationsgesetz</a>	Ereignis > Rechtliches

Datensätze 1 bis 4 von 4. | [Ende](#)

Abbildung 2: SNS Dialog zu findTopics und Ergebnisliste

Die Option „phonetische Übereinstimmung“ sollte bei Unsicherheiten in der genauen Schreibweise helfen, befindet sich aber noch immer im experimentellen Zustand von 2003, da von den SNS-Nutzern kein Bedarf an einer Verbesserung geäußert wurde. Die Optionsliste unter „Topic-Typ“ erlaubt auch die Eingrenzung der Ergebnisse auf einzelne TopicTypen.

### 3.1.2 getPSI: Alles über einen Begriff

Der Dienst getPSI bietet Zugriff auf einen Begriff über seine ID. Das Kürzel "PSI" verweist auf das Topic Map Konzept der "Published Subjects Indicators", nach dem jeder Begriff eines kontrollierten Wortguts durch eine eindeutige URL identifiziert und damit "veröffentlicht" wird. Dieser Dienst gibt alle Eigenschaften (Namen, Beschreibung, Quelle usw.) sowie seine unmittelbaren Assoziationen zu anderen Begriffen zurück. Abbildung 3 zeigt eine solche Darstellung in der GUI von SNS. Leider sind längst nicht alle Topics so reichhaltig attribuiert wie die von SNS selbst gepflegten Ereignisse.

## Bürger erhalten besseren Zugang zu Umweltinformationen

**Typ**

Ereignis > Rechtliches

**Weitere Namen**

**Thema**

Ein neues Umweltinformationsgesetz tritt in Kraft. Für die Bürgerinnen und Bürger wird der Zugang zu Umweltinformationen deutlich verbessert. So werden künftig alle Stellen der öffentlichen Verwaltung des Bundes sowie bestimmte private Stellen zur Herausgabe von Umweltinformationen verpflichtet. Die Auskunftspflichten der Landesverwaltung werden künftig in landesrechtlichen Vorschriften geregelt, bis dahin gilt für Behörden der Länder und Gemeinden die Umweltinformationsrichtlinie unmittelbar. Die Bundesverwaltung wird verpflichtet, umfassender als bisher Umweltinformationen aktiv zu verbreiten. Dabei soll zunehmend das Internet als modernes und schnelles Medium genutzt werden. Mit dem Umweltinformationsgesetz wird die neu gefasste Umweltinformationsrichtlinie der EU (2003/4/EG) umgesetzt. Zugleich werden die Verpflichtungen aus der Aarhus-Konvention erfüllt.

**Definitionen**

[Umweltinformationsgesetz \(UIG\)](#)

**Verwandte Informationen**

[BMU Pressemitteilung](#)  
[EU: Freier Zugang zur Information](#)

**Ereignis**

14.02.2005

**Assoziationen**

<a href="#">Ereignis-Beschreibung</a>	<a href="#">Aarhus-Konvention</a> (Deskriptor)
<a href="#">Ereignis-Beschreibung</a>	<a href="#">Europäische Union</a> (Deskriptor)
<a href="#">Ereignis-Beschreibung</a>	<a href="#">Internet</a> (Deskriptor)
<a href="#">Ereignis-Beschreibung</a>	<a href="#">Landesrecht</a> (Deskriptor)
<a href="#">Ereignis-Beschreibung</a>	<a href="#">Staatsbürger</a> (Deskriptor)

Abbildung 3: SNS Dialog zu getPSI

### 3.1.3 autoClassify: Automatische Verschlagwortung

Dieser Dienst ermittelt mit Hilfe einer automatischen Textanalyse geeignete Schlagwörter für Textdokumente. Den Text übergibt man entweder direkt beim Aufruf oder aber in Form einer URL, die dann vom Dienst ausgewertet wird. Ein Beispiel einer Ergebnisdarstellung ist Abbildung 1 von 2003. Heute werden auch Ereignisse zurückgegeben.

### 3.2 Neuere komplexere Service Operationen (2004-2007)

Seit 2004 lag der Schwerpunkt auf einer fortgeschrittenen Unterstützung der End-Anwender im Unterschied zu der eher am Bibliotheks- und Dokumentationswesen orientierten „Verschlagwortung“. Als Ergebnisse liegen bis heute vor:

- findEvents
- anniversary
- getSimilarTerms
- getHierarchy

#### 3.2.1 findEvents: Nach Ereignissen suchen

Dies ist eine Spezialisierung von findTopics für Ereignisse der Umwelt-Chronologie. Hier kann zusätzlich mit Zeitangaben gesucht werden.

**Auswahlkriterien**

**Ereignistypen**

<input checked="" type="checkbox"/> Alle Ereignistypen	<input checked="" type="checkbox"/> Historisches	<input checked="" type="checkbox"/> Rechtliches
<input checked="" type="checkbox"/> Maßnahme	<input checked="" type="checkbox"/> Konferenz	<input checked="" type="checkbox"/> Natur des Jahres
<input checked="" type="checkbox"/> Unfall-Störfall-Katastrophe	<input checked="" type="checkbox"/> Jahrestag	<input checked="" type="checkbox"/> Internationales Jahr
<input checked="" type="checkbox"/> Publikation		
<input checked="" type="checkbox"/> Beobachtung		

**Zeitbezug**

Von:  Bis:  (TT.MM.JJJJ)

Am:  (TT.MM.JJJJ)

**Suchbegriff**

1 - 6 von 6 Einträgen in **Umweltchronik** (Seite 1 von 1) | 1

**Greenpeace Aktion gegen Atommüllverklappung**  [Recherche](#)

Ereignistyp: Historisches  
Datum: 1978

Großbritannien und andere europäische Länder kippen, wie Greenpeace zufällig entdeckt, radioaktive Abfälle in den Nordatlantik. Als der britische Frachter Gem mit 2 000 t Atommüll

Abbildung 4: PortalU nutzt findEvents

### 3.2.2 anniversary: Was geschah heute vor X Jahren?

Anstelle von "heute" kann auch ein beliebiges anderes Datum eingegeben werden. Die Antwort bevorzugt taggenaue Ereignisse. Sind solche nicht vorhanden, werden "runde" Jahrestage (25, 50, 100) ermittelt.

Jury Umweltzeichen
<b>Typ</b>
Ereignis > Konferenz
<b>Weitere Namen</b>
<b>Thema</b>
Die Jury Umweltzeichen hat ihre erste Sitzung. Das Umweltzeichen wird für den Konsumenten ein Wegweiser zum umweltgerechteren Einkauf. Die Vergabe des Umweltzeichens ("Blauer Engel") erfolgt unter enger Abstimmung mit den Verbraucherverbänden, den Umweltorganisationen und der Wirtschaft. Mittlerweile hat das Label viele Nachahmer. Um Klarheit zu schaffen, veröffentlicht das Umweltbundesamt 1998 einen Wegweiser für 40 Ökosiegel.
<b>Definitionen</b>
<b>Verwandte Informationen</b>
<b>Ereignis</b>
05.06.1978
<b>Assoziationen</b>
<a href="#">verwandtes Ereignis</a> <a href="#">30 Jahre Blauer Engel, 30 Jahre Innovation</a> (Ereignis)

Abbildung 5: Heute vor 30 Jahren (SNS Darstellung)

PortalU zeigt unten rechts auf seiner Homepage ein laufend aktualisiertes Ergebnis.

**UMWELTCHRONIK**

**Vor 16 Jahren: Richtlinie des Rates zur Erhaltung der natürlichen Lebensräume sowie der wildlebenden Tiere und Pflanzen (21.05.1992)**

[» Details](#)

[» Weitere Umweltereignisse](#)

Abbildung 6: PortalU nutzt *anniversary*

### 3.2.3 getSimilarTerms: Nach ähnlichen Begriffen suchen

getSimilarTerms ermittelt "ähnliche" Begriffe aus dem Umwelt-Thesaurus zu einem Ausgangswort. Die Ähnlichkeit bezieht sich nicht auf die Schreibweise, sondern auf die Bedeutung. Typische Anwendungsfälle sind etwa Vorschläge für die Volltextsuche oder auch die Wortwahl beim Verfassen von Texten.



Abbildung 7: PortalU bietet Ergebnisse von *getSimilarTerms* an

Ergebnisse wie die von *getSimilarTerms* könnten auch durch eine Sequenz von *findTopics* und mehreren *getPSI* ermittelt werden. Die Zusammenfassung in einen Dienst erfolgte vor allem aus Performanzgründen, aber auch um die Ergebnismengen aufgrund einiger Erfahrungsregeln zu optimieren. Die genaue Spezifikation wurde zwischen 2005 und 2008 im Rahmen des VKoopUIS<sup>35</sup> Projekts 9 „WebService Thesaurusdienste“ in einer Reihe von Projektsitzungen unter dem Titel „Ähnliche Begriffe“ gemeinsam erarbeitet. Bayern hatte hierbei die Federführung, UBA und LUBW waren Partner.

In der aktuellen Version (seit 2007) dieser Operation werden die Ergebnisse nicht mehr zur Aufrufzeit ermittelt, sondern aus einer Ergebnisdatenbank. Die Ergebnisdatenbank wird jeweils asynchron nach Änderung des Umwelt-Thesaurus durch eine einmalige Batchprozedur erzeugt.

### 3.2.4 *getHierarchy*: Rückgabe einer Begriffshierarchie

Ausgehend von einem Begriff des Umwelt-Thesaurus wird eine Begriffshierarchie (Ober-/Unterbegriffe) zurückgegeben. Diese kann nach Richtung und Tiefe begrenzt werden. Aufgrund der Struktur des Wortguts ist dieser Dienst heute nur für deutschsprachige Deskriptoren verfügbar. Eine Erweiterung für die Verwaltungshierarchien (administrative Gebietseinheiten) ist grundsätzlich denkbar.

<sup>35</sup> Vereinbarung über die Kooperation bei Konzeptionen und Entwicklungen von Software für Umweltinformationssysteme

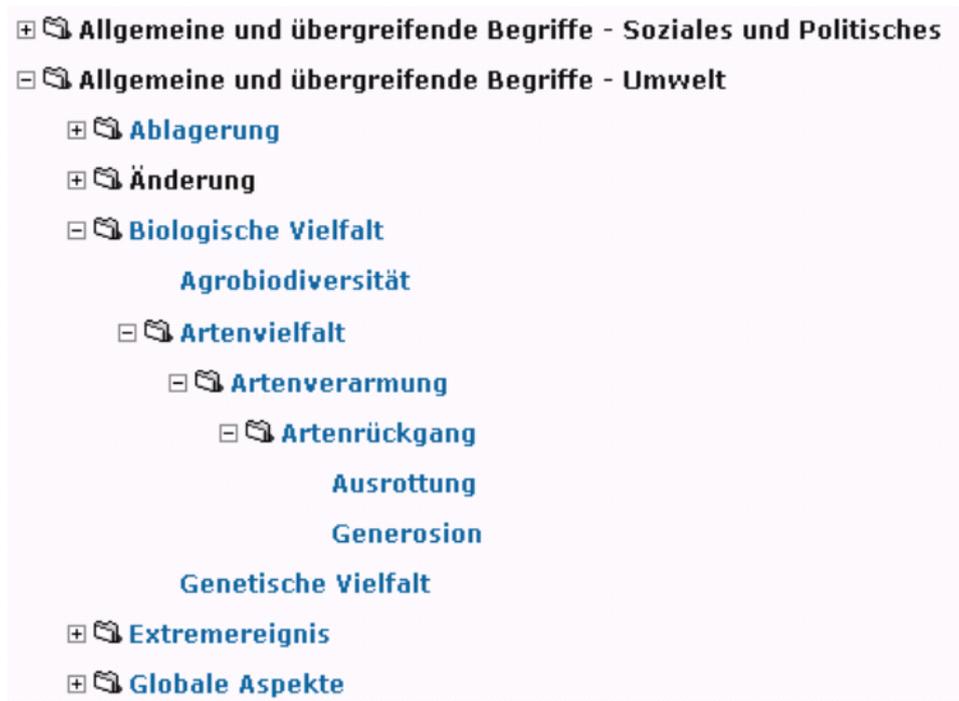


Abbildung 8: *getHierarchy* in PortalU

Ähnlich wie bei *getSimilarTerms* könnte dasselbe Ergebnis auch durch eine Sequenz von einzelnen *getPSI* aufgebaut werden. Auch hier wurde die komplexere Operation aus Performanzgründen zusammengefasst.

### 3.3 Anwender der SNS Service Operationen

Heute dient SNS neun Anwendungen der Umweltinformation und der Geodateninfrastruktur, wobei sich drei noch in der Erprobungsphase befinden:

- PortalU verwendet seit 2006 alle angebotenen Services und wird ab Mitte 2008 auch die „keywords“ von INSPIRE Katalogdaten mit SNS pflegen.
- Das Umweltportal Rheinland-Pfalz nutzt *autoClassify* zur Generierung von Metadaten.
- Der bayerische Umweltobjektkatalog (UOK) integriert mit Hilfe von SNS den UMTHES® mit weiteren, meist spezielleren Fachthesauri.
- Die Umweltportale Baden-Württemberg, Sachsen-Anhalt und Thüringen planen den SNS Einsatz, insbesondere *getSimilarTerms*. Der Einsatz der Google Search Appliance ab 2008 wirft hier allerdings das Problem auf, dass sich Web Services nicht ohne weiteres in die Google Suchmaske einbinden lassen. SNS entwickelt daher derzeit eine zusätzliche Schnittstelle, die Google's „OneBox“ Spezifikation entspricht (vgl. Kap. 4.4.3).

Geografische Namen werden mit *findTopics* verfügbar in:

- Nationales Geo-Portal Portal Deutschland „GeoPortal Bund“
- Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV)
- Geo-Portal des Landes Thüringen.

## 4 Vom Forschungsprojekt zum ausgereiften Produktivsystem

Neben der Erweiterung der Service Operationen wurden verschiedene weniger sichtbare Eigenschaften des SNS weiterentwickelt, um einen angemessenen Reifegrad als produktives und verlässliches System zu erreichen.

### 4.1 Performanz und Skalierbarkeit

Von Anfang an musste sich SNS mit einer verbreiteten Skepsis hinsichtlich der Antwortzeiten und der zuverlässigen Erreichbarkeit eines gemeinsam genutzten Web Service auseinandersetzen. Obwohl dieses Thema absolut nichts mit den speziell semantischen Schwerpunkten der Funktionalität zu tun hat, handelt es sich dennoch um eine überlebenskritische Herausforderung.

Auf der einen Seite entstehen Spitzen von über 300.000 Aufrufen von autoClassify durch PortalU bei der maschinellen Aktualisierung des Dokumentenindex. Gleichzeitig wird erwartet, dass SNS die interaktiven Nutzungen durch menschliche Anwender umgehend beantwortet, so dass ein zügiger Fluss der Interaktionssequenzen nicht behindert wird. Dies hat zunächst wiederholt zu Problemen geführt. SNS musste in dieser Hinsicht zu einem Reifegrad weiterentwickelt werden, wie er von einem bloßen Forschungsergebnis üblicherweise nicht erwartet wird. Die Performanz wurde durch Serveroptimierung mehrfach verbessert, und derzeit wird auf eine Multi-Server-Architektur mit Lastverteilung umgestellt, um den steigenden Anforderungen nachhaltig zu genügen.

Eine grundsätzlich unbegrenzte Skalierbarkeit für jede sinnvolle Anwendung der Dienste ist eine strategische Anforderung, ohne deren Erfüllung auch ein Dienst von höchster inhaltlicher Qualität seine Glaubwürdigkeit einbüßt.

### 4.2 Versionierung und Erweiterung des Vokabulars

SNS hat wiederholbare und erweiterbare Import- und Integrationsprozeduren implementiert, die auf recht heterogene Originalquellen des Vokabulars eingehen.

Der Umwelt-Thesaurus wird aus einem Datenbank-Dump des UMTHESES® Originalbestands geladen, dessen Datenmodell den Standard ISO 2788 (1986) mit einigen nicht-trivialen Erweiterungen ergänzt.

Die geographischen Bezüge („Geo-Thesaurus“) sind ursprünglich aus ESRI Shapefiles importiert worden, werden aber seit 2007 aus dem Web Feature Service „Geographische Namen Deutschlands“ (WFS GN-DE) des Bundesamts für Kartographie und Geodäsie (BKG)<sup>36</sup> aktualisiert. Hier tritt SNS einmal selbst als Nutzer eines externen Web Service auf. Hinzu kommen Verzeichnisse der Schutzgebiete, die das Bundesamt für Naturschutz (BfN)<sup>37</sup> bisher separat bereitstellt.

Die Schwierigkeit der Importe liegt hier weniger in der Umsetzung der jeweiligen Protokolle und Datenmodelle als im Abgleich der Versionsstände aus unterschiedlichen Quellen. Dabei müssen Fragen entschieden werden wie: Was ist unverändert geblieben? Was ist ähnlich, aber nicht genau gleich? Was wurde mit Sicherheit geändert? Was ist offensichtlich ungültig geworden?

Um diesen Fragen gerecht zu werden, hat SNS eine transparente Versionspolitik implementiert: Bevor eine neue Version in Produktion geht, werden alle registrierten

---

<sup>36</sup> <http://www.geodatenzentrum.de>

<sup>37</sup> <http://bfm.de/>

Nutzer über den wirksamen Zeitpunkt und den Umfang der Änderungen informiert. Dabei bleiben die Versionen stets abwärts kompatibel. Um eine wirklich dauerhafte Begriffsreferenz zu bieten, werden SNS Topics „nie“ gelöscht, selbst wenn sie aus den Originalquellen entfernt worden sind. Dies geschieht z.B. durch administrative Gebietsreformen, bei denen zuvor wohlbekannte Gemeinden oder Landkreise plötzlich ihre Existenz verlieren. Dadurch verschwinden ihre Namen aber nicht auch gleichzeitig aus den nach wie vor verfügbaren Informationsbeständen. SNS führt sie deshalb unverändert weiter und markiert sie lediglich als „verfallen“ (expired).

### 4.3 Semantische Struktur und Attributierung

Gegenüber dem Anfangsmodell der Topic Map ist die semantische Struktur ständig weiterentwickelt worden. Aktuelle Neuerungen von 2007 sind:

- Die Attributierung der Topics wurde angereichert, insbesondere durch das „Verfallsdatum“ (expired), das Mitführen von verschiedenen speziellen Schlüsselssystemen der Quellen (nativeKey), durch die Umweltklassifikation und die Wortart im Umwelt-Thesaurus.
- SNS führte in Übereinstimmung mit UMTHESES® eine asymmetrische Mehrsprachigkeit (deutsch/englisch) im Thesaurus ein. Dies folgt den Empfehlungen der International Federation of Library Associations and Institutions (IFLA), „Guidelines for Multilingual Thesauri“.
- Um die Qualität von autoClassify zu verbessern, wird der Umgang mit zusammengesetzten Begriffen überarbeitet. Das erste Ergebnis ist bisher: wir haben die bisherige, unzureichende Sonderbehandlung nahezu vollständig (außer bei gekennzeichneten Eigennamen) aufgegeben – manchmal ist weniger einfach mehr.

### 4.4 Interoperabilität der Schnittstellen

Im SNS Forschungsprojekt wurde internationalen Standards eine hohe Bedeutung eingeräumt. Dies hat letztlich einen praktischen Grund: es ist anzunehmen, dass sich die angestrebte Vielzahl der Nutzer darauf insgesamt leichter einstellen kann als auf jede denkbare proprietäre Schnittstelle. Dieses Prinzip hat sich in der Praxis bewährt, muss aber auch laufend mit den Realitäten abgeglichen werden.

#### 4.4.1 Web Services und XML Topic Maps

2003 ging SNS mit einer Schnittstelle in Betrieb, die den Standards XML Topic Maps<sup>38</sup> („XTM“, ISO 19350) und Web Services<sup>39</sup> (W3C) folgt. Web Services regeln das Protokoll zwischen Anwender und SNS bei der Nutzung der Services, XTM behandelt die Formatierung des SNS Vokabulars bei der Rückgabe der Ergebnisse.

Web Services haben sich seitdem immer stärker durchgesetzt und werden aufgrund der guten Werkzeugunterstützung auf den verschiedensten Plattformen von den meisten Technikern bevorzugt eingesetzt, insbesondere im Zusammenhang mit serviceorientierten Architekturen. Bei neueren Web Anwendungen („Web 2.0“) verbreiten sich daneben einfachere Protokolle wie Atom-Feeds oder Microformats. XML Topic Maps konnten in der Praxis problemlos eingesetzt werden, haben in der globalen Praxis semantischer Anwendungen jedoch eher an Bedeutung verloren.

SNS trägt beiden Trends durch Weiterentwicklung Rechnung.

---

<sup>38</sup> <http://www.w3.org/ws/>

<sup>39</sup> <http://isotopicmaps.org/>

#### 4.4.2 Topic Maps und Semantic Web

2004 hat die internationale ECOInformatics Initiative unter dem Namen Ecoterm<sup>40</sup> eine „Thesaurus und Terminologie“ Aktivität aufgenommen, um „ein gewisses Niveau von Interoperabilität zwischen diesen Terminologien“ herbeizuführen. SNS hat bei dieser Initiative von Anfang an mitgewirkt und dort auch den Ansatz über XTM und Web Services eingebracht. Ecoterm hat sich jedoch vor allem mit den formalen Sprachen des Semantic Web beschäftigt, insbesondere mit Simple Knowledge Organisation System (SKOS)<sup>41</sup> und der komplexeren Web Ontology Language (OWL)<sup>42</sup>. SNS hat sich darauf eingestellt. Da SKOS aus Sicht von SNS zu stark auf Thesauri und Klassifikationen beschränkt ist, haben wir zunächst eine OWL-Lite Repräsentation des SNS Vokabulars (Bandholtz, 2006) als allgemeines Austauschformat entworfen.

Die Vereinbarung gemeinsamer Protokolle und Schnittstellen zwischen den Anbietern von Terminologie ist noch nicht abgeschlossen. Sie stellt jedoch eine Voraussetzung für die effiziente Nutzung verschiedener terminologischer Quellen innerhalb eines Metadatenbestands dar.

#### 4.4.3 Google OneBox

Einige an SNS interessierte Betreiber von Umweltportalen setzen heute als Suchmaschine die Google Search Appliance<sup>43</sup> ein (vgl. Kap. 3.3). Aus dieser Anwendung heraus lassen sich Web Services nicht einbinden. Wenn die semantischen Dienste von SNS hier trotzdem angeboten werden sollen, muss SNS sich auf die von Google unterstützten Schnittstellen einstellen.

Google bietet hier vor allem unter dem Namen „OneBox“ einen Mechanismus, mit dem externe Quellen in eine Suche einbezogen werden. Das wohl bekannteste Beispiel ist die Einbindung von Google Maps in die Standard Trefferliste, wenn das Suchwort ein Ortsname ist. Es gibt darüber hinaus zahlreiche weitere „eingebaute“ Beispiele<sup>44</sup>, zum Teil nur auf Englisch.



Abbildung 9: *getSimilarTerms* als Google OneBox (Entwurf)

SNS untersucht derzeit näher, in welcher Weise die eigenen semantischen Dienste sinnvoll in die Google Trefferliste eingebunden werden können. Ein erstes Beispiel ist die Operation „getSimilarTerms“ (vgl. Abb. 9), die in Kürze in dieser Form verfügbar sein wird. Die zurückgegebenen Begriffe werden dabei als Hyperlinks ausgeführt, mit denen jeweils die Google Suche mit diesem Begriff ausgelöst wird.

<sup>40</sup> <http://www.eea.europa.eu/cooperations/eco-informatics/area-of-activities-information-technologies>

<sup>41</sup> <http://www.w3.org/2004/02/skos/>

<sup>42</sup> <http://www.w3.org/2004/OWL/>

<sup>43</sup> <http://code.google.com/apis/searchappliance/>

<sup>44</sup> <http://www.google.com/help/features.html>

## 5 Die Herausforderung: Semantische Datenintegration

Nachdem das Interesse an semantischer Unterstützung bei Suchmaschinen und Informationsportalen in den letzten Jahren deutlich zugenommen hat, erwarten wir künftig eine stärkere Beachtung von semantischen Verfahren bei der Datenintegration. Hierbei sehen wir für SNS zwei Schwerpunkte:

- Die Umsetzung der INSPIRE “Metadata Implementing Rules” durch das Bereitstellen sinnvoller „keywords“, um damit die Auffindbarkeit von und den Zugriff auf Datenbestände zu vereinfachen;
- Die technische Umsetzung des Shared Environmental Information System (SEIS) durch die Einführung einer semantischen Schicht im Single Information Space for the Environment (SISE).

### 5.1 INSPIRE Metadaten und Vokabulare

Mit der Integration des Umweltdatenkatalogs (UDK) in PortalU geht auch eine Anwendung der SNS Operationen findTopics, getHierarchy and autoClassify für die Vergabe von UMTHES® Deskriptoren in den UDK Metadaten einher.

Diese Metadaten werden künftig auch für INSPIRE bereitgestellt. Die INSPIRE Metadata Implementing Rules (INSPIRE, 2008) fordern aber für jeden Datensatz die Verwendung von „wenigstens einem“ Keyword des General Multilingual Environmental Thesaurus (GEMET<sup>45</sup>) und für alle Keywords die Angabe des jeweilig verwendeten Vokabulars („citation of the originating Controlled Vocabulary“).

Um dies auch für die UDK Beiträge durchzusetzen, muss SNS zumindest Verweise von UMTHES® Deskriptoren auf GEMET Deskriptoren einarbeiten. Dies ist zunächst recht einfach, denn UMTHES® war eine der Quellen für GEMET und enthält daher solche Verweise auf GEMET Begriffe in Deutsch und Englisch, die lediglich bisher von SNS nicht berücksichtigt worden sind.

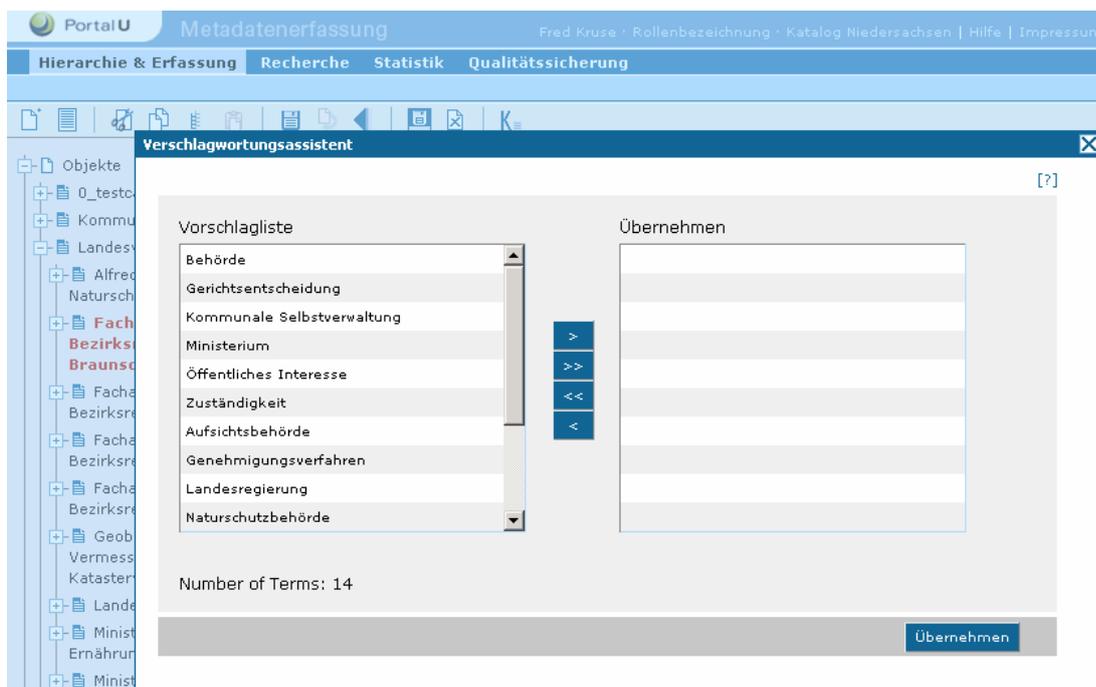


Abbildung 10: SNS bei der Verschlagwortung des UDK in PortalU

<sup>45</sup> <http://www.eionet.europa.eu/gemet>

Allerdings enthält UMTHESES® nicht die spezifische Struktur des GEMET, also Supergroups, Groups, Themes und deren Hierarchie und begriffliche Zuordnung. Dies ist auch nicht notwendig, denn INSPIRE wird GEMET für die eigene Nutzung in aktualisierter Form verfügbar machen. Daher muss SNS nicht den gesamten GEMET integrieren, sondern kann sich durch externe Verweise auf ihn beziehen. Da GEMET im Kontext von Ecoterm bereits seit 2004 im Semantic Web Format SKOS vorliegt, können etablierte Verfahren wie Ontology Matching (Euzenat, 2007) benutzt werden, um einen nahtlosen Übergang zwischen GEMET und den feinkörnigeren und reicheren Vokabularen herzustellen. Die Ontology Alignment Evaluation Initiative (OAEI)<sup>46</sup> hat 2007 GEMET auf diese Weise mit den Thesauri AGROVOC<sup>47</sup> und NAL<sup>48</sup> abgeglichen.

Hinsichtlich der Ausführung von Begriffsverweisen in den INSPIRE Metadaten bestehen allerdings noch einige offene Punkte, die sich bei der beabsichtigten Nutzung als hinderlich herausstellen könnten.

```

<gmd:descriptiveKeywords>
  <gmd:MD_Keywords>
    <!-- Part B 3.1 Keyword Value -->
    <gmd:keyword>
      <gco:CharacterString>Hydrography</gco:CharacterString>
    </gmd:keyword>
    <!-- Part B 3.2 Originating Controlled Vocabulary -->
    <gmd:thesaurusName>
      <gmd:CI_Citation>
        <gmd:title>
          <gco:CharacterString> GEMET Thesaurus version 1.0 </gco:CharacterString>
        </gmd:title>
      </gmd:CI_Citation>
    </gmd:thesaurusName>
  </gmd:MD_Keywords>
</gmd:descriptiveKeywords>

```

Abbildung 11: Kodierung von Keywords in INSPIRE. Quelle: (INSPIRE, 2008)

Das Beispiel in Abbildung 11 benennt sowohl das Keyword selbst als auch dessen Ursprungsvokabular lediglich durch Namen, und nicht durch eindeutige Identifier oder URL. Eine solche Kodierung führt zu mehreren Restriktionen:

- Es können lediglich Vokabulare mit eindeutiger Namenskonvention verwendet werden. Dies trifft für die meisten traditionellen Thesauri zu, aber nicht für neuere Vokabulare, die mehrdeutige Namensschreibweisen („Homographen“) zulassen und Eindeutigkeit allein über eine URL herstellen.
- Wenn Anwendungen diese Metadaten maschinell auswerten wollen, können sie nicht direkt auf die Quellvokabulare zugreifen, um die Definitionen, den semantischen Kontext und externe Querverweise auszuwerten.
- Offenbar sind solche Einschränkungen von INSPIRE bisher nicht berücksichtigt worden, während sie aus Sicht des Semantic Web sofort ins Auge springen. Katalogdaten in dieser Form sind abhängig von spezialisierten Anwendungen, die implizites Wissen implementieren.

<sup>46</sup> <http://www.few.vu.nl/~wrvhage/oeai2007/environment.html>

<sup>47</sup> [http://www.fao.org/aims/ag\\_intro.htm](http://www.fao.org/aims/ag_intro.htm)

<sup>48</sup> <http://agclass.nal.usda.gov/agt/agt.shtml>

## 5.2 SEIS und SISE

Das Konzept eines Shared Environmental Information System (SEIS) wurde bereits auf der EnviroInfo 2007 öffentlich vorgestellt und ist im Februar 2008 von der Europäischen Kommission verabschiedet worden. SEIS zielt auf "streamlining and simplification" des bestehenden europäischen Berichtswesens und darf nicht als zusätzliche Berichtsverpflichtung missverstanden werden. Aus technischer Sicht gibt es zwei zentrale Leitlinien:

- "information should be managed as close as possible to its source", and
- "information is provided once and shared with others for many purposes".

Die vorausgegangenen Kapitel sollten deutlich gemacht haben, das SNS schon seit 2003 nach diesen Grundsätzen vorgeht, auch wenn es selbst keine Umweltinformation anbietet, sondern lediglich das Vokabular und semantische Methoden, die das Auffinden, das Verständnis und die Integration solcher Information ermöglichen. Um dies auch für SEIS verfügbar zu machen, sollte SNS entsprechend erweitert und als Baustein der künftigen SEIS Infrastruktur genutzt werden.

Während die SEIS Dokumente überwiegend administrative Themen behandeln, widmet sich eine damit verbundene Forschungsinitiative "Towards a Single Information Space for the Environment in Europe" (SISE, 2008) der erforderlichen technischen Infrastruktur. Auf deren ersten Beratungsworkshop gab es Beiträge von PortalU (Uhrich, 2008) und SNS (Bandholtz, 2008), die zu der Priorisierung der künftigen Forschungsschwerpunkte im Ergebnisbericht beitrugen. Dort werden „Semantics, thesauri, ontology services and standardisation“ als Voraussetzung für eine „flexible Verkettung verteilter Umweltdienste“ ("Flexible chaining of distributed environmental services") aufgeführt. Es geht also hier nicht mehr allein um (menschenslesbare) Information, sondern um die dynamische Komposition von Daten und Methoden quer durch die gesamte Europäische Umweltinformatik.

SNS und seine bis auf GEIN zurückreichenden Erfahrungen könnten hier einen Ausgangspunkt für den Aufbau einer verteilten, semantisch vernetzten Infrastruktur bieten.

## 5.3 Ein möglicher Lösungsweg: Linking Open Data

Betrachtet man das Thema aus Sicht des Semantic Web, und möchte man nicht zuerst die Ergebnisse weiterer langjähriger Forschung abwarten, so bietet das erst Mitte 2007 begonnene Community Projekt „Linking Open Data“<sup>49</sup> reichhaltiges Anschauungsmaterial.

Das Semantic Web wird hier als „decentralised world-wide information space for sharing machine-readable data with a minimum of integration costs“ (W3C, 2008) gesehen. Dies ist eine Formulierung, die ebenso gut aus einem Entwurf für SISE stammen könnte. Abbildung 12 zeigt die im März beteiligten Datenbanken. Zum letzten „Zähltermin“ (Oktober 2007) waren insgesamt gute 2 Milliarden Fakten („RDF triples“) mit 3 Millionen Querbeziehungen versehen und öffentlich verfügbar.

---

<sup>49</sup> <http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

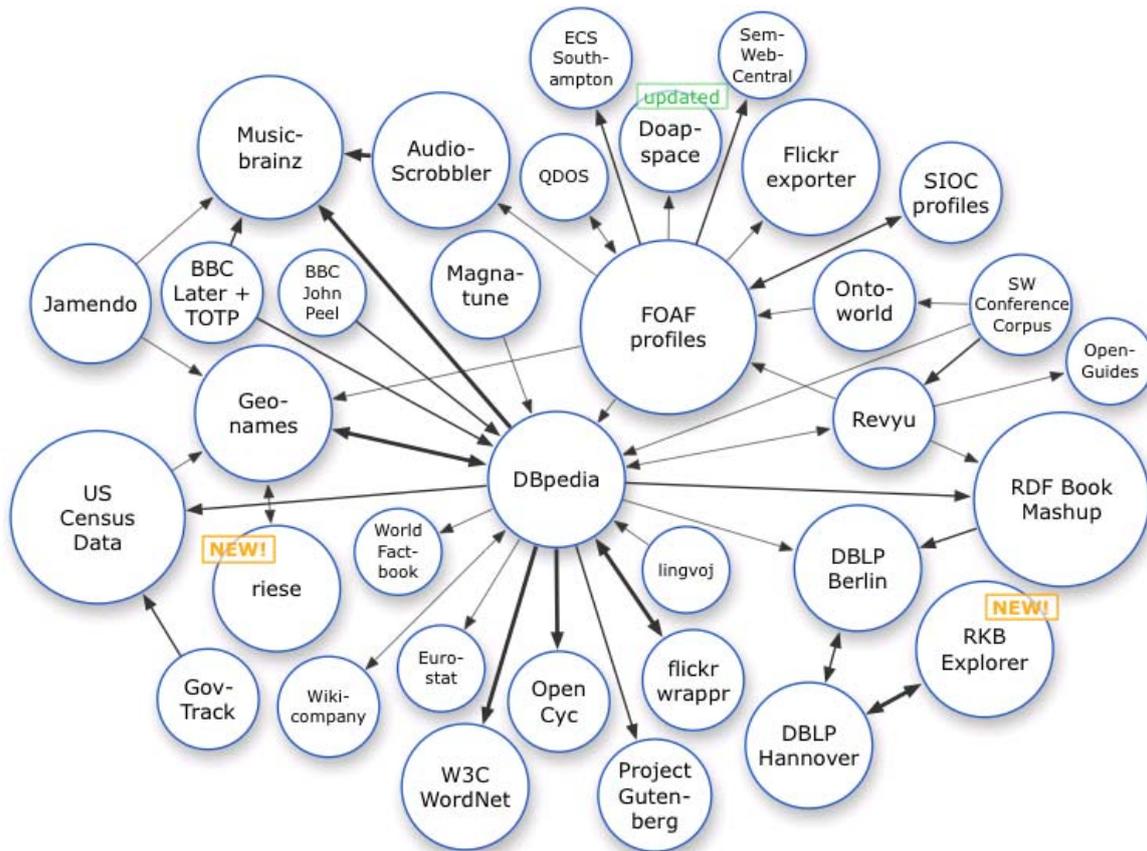


Abbildung 12: Linking Open Data im März 2008

Linking Open Data basiert auf dem Resource Description Framework (RDF) und einer Representational State Transfer (REST) Architektur. REST ist darüber hinaus die Grundlage des allgegenwärtigen Internet. Somit besteht die Infrastruktur für Linking Open Data vollständig aus frei zugänglichen Standards und technologischen Bausteinen.

Wie beim „Surfen“ im Internet, kann in diesem Netzwerk frei über Hyperlinks navigiert werden. Es gibt darauf zugeschnittene Suchmaschinen und Browser, und man kann sich die bekannten Lesezeichen auf jeden einzelnen Datensatz anlegen. Im Unterschied zum bekannten Internet haben hier alle Informationen eine primär maschinenlesbare Struktur (RDF). Natürlich sind Werkzeuge verfügbar, die dies in eine für Menschaugen eher geeignete Form umsetzen, aber die Maschinenlesbarkeit von RDF bietet gerade die entscheidende Grundlage für jene Werkzeuge, mit denen jeweils relevante Daten identifiziert, flexibel verkettet und in allen erdenklichen Formen aufbereitet werden können.

Dabei spielen semantische Vokabulare wie „WordNet“<sup>50</sup> oder „Geonames“<sup>51</sup> die Rolle von Bindegliedern und Referenzpunkten, um Datenbestände wie die „US Census Data“, „Eurostat“ oder Literaturdatenbanken untereinander in Beziehung zu setzen.

#### 5.4 Semantische Vernetzung von Umwelt-Dateninseln

Linking Open Data ist zweifellos weder auf Umweltinformation spezialisiert, noch werden die mit SISE angesprochenen Verarbeitungsmethoden implementiert.

<sup>50</sup> <http://wordnet.princeton.edu/>

<sup>51</sup> <http://www.geonames.org/>

Dennoch finden sich hier entscheidende Anregungen, wie große Mengen sehr unterschiedlicher Daten nahe bei ihren Quellen lediglich einmal publiziert werden und dennoch in ihrer Gesamtheit einen „gemeinsamen Informationsraum“ (shared information space) ausmachen, in dem es sich mit frei verfügbaren Mitteln unbegrenzt navigieren lässt.

Wenn sich z.B. Eurostat in dieser Weise publizieren und integrieren lässt, so dürfte es kaum einen Grund geben, warum man mit beliebigen Umweltmessnetzen nicht ebenso verfahren kann. Ein „RDF Book Mashup“ bietet das Modell für die Umweltforschungsdatenbank, und „Geonames“ zeigt den Weg einer Integration mit INSPIRE.

So könnte z.B. ein an der Umwelt interessierter Hobbyangler bei der Wochenendplanung einen Fluss entdecken, in dem die Brasse heimisch ist. Er hätte dann die Wahl, weitere Fischarten näher zu betrachten, oder aber zu entdecken, dass seit 1993 die Anreicherung von „p,p'-DDD“ in der Muskulatur dieser Tiere beobachtet wird. Wenn er nicht weiß, was „p,p'-DDD“ ist, genügt ein Klick, und er befindet sich an der betreffenden Stelle eines toxikologischen Glossars. Will er es genauer wissen, kann er sich ebenso einfach die Messmethode „GC-MS“ erläutern lassen oder die Referenzen des ausführenden Messinstituts einsehen.

Ähnlich wie hier der „interessierte Laie“ in die Umweltinformation gewissermaßen hineinstolpert, könnten anspruchsvollere Anwendungen ganze Zeitreihen verarbeiten, die Vergleichbarkeit von Messmethoden beurteilen, Korrelationen analysieren usw., befreit von komplizierten Zugangs- und Transformierungsprozeduren, sogar ohne den Umweg über Metadatenkataloge, da die Daten selbst unmittelbar miteinander vernetzt wären.

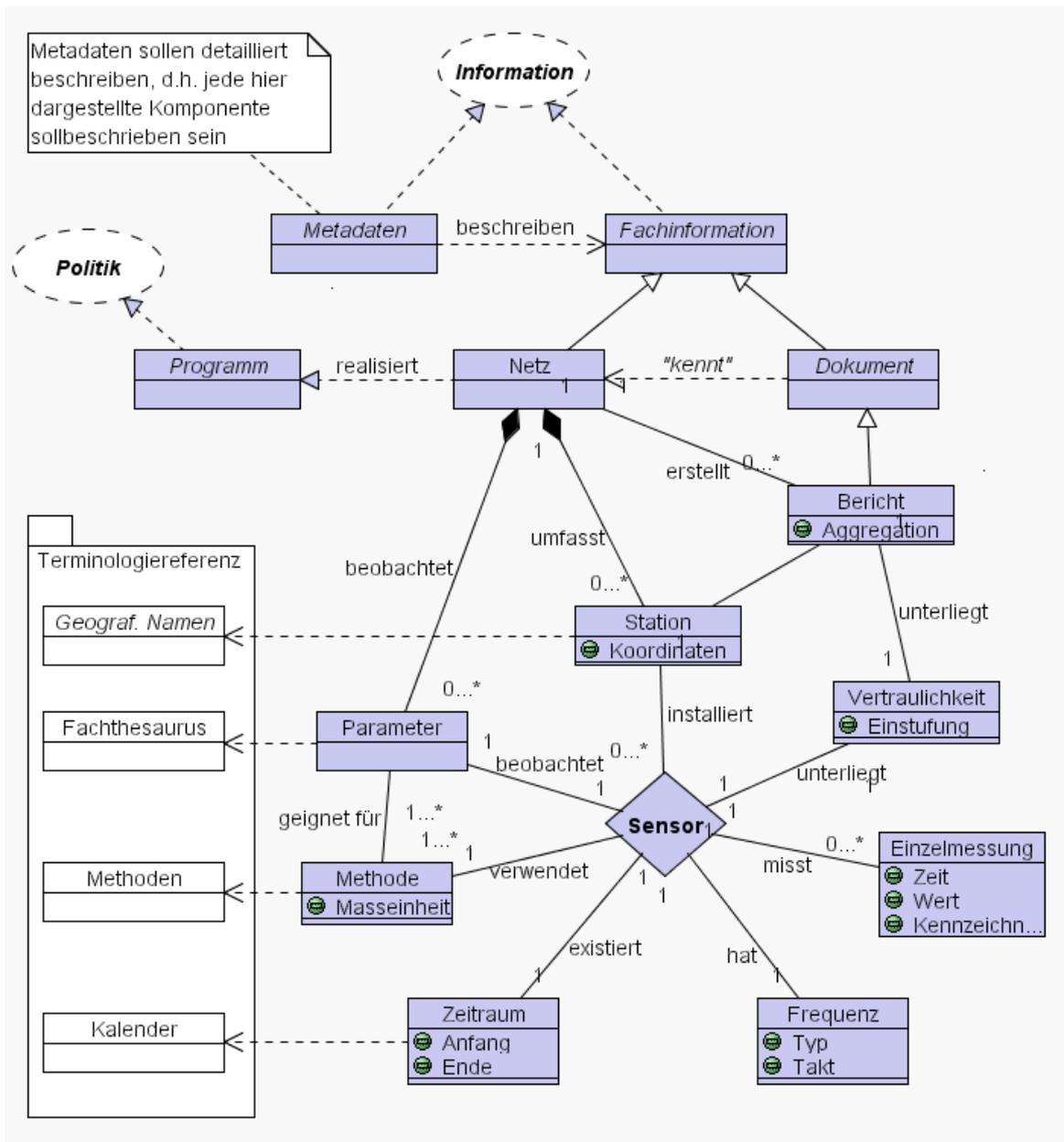


Abbildung 13: Semantisches Datenmodell für Beobachtungsnetze (Bandholtz, 2004)

Eine ähnliche Perspektive wurde bereits 2004 in der Machbarkeitsstudie für eine Integrationsschicht Umweltbeobachtung (Bandholtz, 2004) entworfen, ist aber bisher nicht weiter verfolgt worden. Ein erweitertes SNS war hier als übergreifende „Terminologiereferenz“ für die Datenbanken vorgesehen, wodurch das Auftauchen gleicher Begriffe in unterschiedlichen Quellen den Ansatzpunkt für eine Vernetzung der Dateninseln bieten würde.

## 6 Zusammenfassung und Schlussfolgerung

Im Verlauf der letzten 10 Jahre hat die Kombination von Services und Semantik ihren Machbarkeitsnachweis geliefert und entwickelt sich heute zur „großen Hoffnung“ für die künftige Umweltinformation. Entstanden aus der dokumentarischen Erschließung, entwickeln kontrollierte Vokabulare eine strategische Rolle für das Auffinden und die Integration relevanter Daten. Ausgereifte semantische Dienste wie SNS sind Bausteine für eine solche Zukunft und lassen sich leicht für die konkreteren Anforderungen spezialisieren.

## 7 Literaturverzeichnis

- Angrick, M. et al. (2002). Semantic Network Services (SNS). 16th International Conference "Informatics for Environmental Protection" (EnvirolInfo). Vienna 2002.
- Bandholtz, T. (2004). Machbarkeitsstudie Integrationsschicht Umweltbeobachtung. Abschlussbericht. Umweltbundesamt 2004.
- Bandholtz, T. et al. (2006). SNS Environmental Vocabulary – from Terms to Ontology. In: Proceedings of the SEMANTICS 2006, Vienna 2006.
- Bandholtz, T. (2008). Subject oriented data provision. SISE Workshop 2008. [http://cordis.europa.eu/fp7/ict/sustainable-growth/workshops/ws-20080215\\_en.html](http://cordis.europa.eu/fp7/ict/sustainable-growth/workshops/ws-20080215_en.html)
- Euzenat, J, Shvaiko, P. (2007). Ontology Matching. Berlin 2007. <http://www.ontologymatching.org/>
- Heyer, G. et al. (2006). Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse. Leipzig 2006.
- INSPIRE (2008). Metadata Implementing Rules based on ISO 19115 and ISO 19119 (2008-04-25). <http://inspire.jrc.it/reports.cfm>
- Rüther, M. (2004). Sharing Environmental Vocabulary. 18th International Conference "Informatics for Environmental Protection", Geneva 2004.
- Scharl, A., Tochtermann, K. (eds.). The Geospatial Web. London 2007.
- SEIS (2008). Commission Communication towards a Shared Environmental Information System (SEIS). Brussels 2008. <http://ec.europa.eu/environment/seis/index.htm>
- SISE (2008). Towards a Single Information Space for the Environment in Europe. FP7 & ICT PSP exploratory workshop. Final Report. Brussels 2008. [ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/sustainable-growth/sise-workshop-report-08\\_en.pdf](ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/sustainable-growth/sise-workshop-report-08_en.pdf)
- SNSWS (2008). SNS Web Services 2.1, Version 1.02 vom 28. März 2008. <http://www.semantic-network.de/sns-ws.pdf>
- Uhrich, S. (2008). The Software of the German Environmental Information Portal PortalU® as Basis for a European Environmental Information Portal? SISE Workshop 2008. [http://cordis.europa.eu/fp7/ict/sustainable-growth/workshops/ws-20080215\\_en.html](http://cordis.europa.eu/fp7/ict/sustainable-growth/workshops/ws-20080215_en.html)
- W3C 2008 Cool URIs for the Semantic Web. W3C Interest Group Note 31 March 2008. <http://www.w3.org/TR/cooluris/>

# Grundwasser-Online – Internetbasiertes UIS für Deutschland?

Thomas Gutzke<sup>52</sup>,  
[info@envi-systems.com](mailto:info@envi-systems.com)

## Abstract

The management of groundwater resources for the supply of German cities and suburban areas has become a matter of public interest during the last years. Negative headlines dealt with cracks in buildings as well as damaged woodlands and inundated agriculture areas as an effect of varying groundwater levels.

At present the required data is logged and stored by each legal authority of each federal state and the local water supply companies since an holistic management of groundwater information does not exist.

This paper describes an approach to combine all information witch are related to water supply and groundwater monitoring (e.g., water level, water quality, rainfall, conveyance-masses, drilling profiles). In the south of hessian this approach is already realised – since 2003. Within the internetportal [www.grundwasser-online.de](http://www.grundwasser-online.de) all water supply companies, legal authorities and engineering societies share there data and provide selected data for the public.

This paper doesn't give an answer to the question how this portal can be extended, but it is collecting ideas and points out possibilities.

## 1 Einleitung

Das Internetportal Grundwasser-Online befindet sich seit 2003 im Produktivbetrieb. In Südhessen erfassen und verwalten alle an der Überwachung und Steuerung beteiligten Akteure ihre gesamten Datenbestände über dieses Internetportal. Durch die bereitgestellten Möglichkeiten, die Datenbestände gezielt untereinander auszutauschen, ist eine unternehmensübergreifende Auswertung innerhalb des Projektgebiets auf Basis hochaktueller Daten möglich.

Dieser Beitrag stellt kurz die technischen Aspekten des Softwaresystems vor und beschreibt die derzeitige Leistungsfähigkeit. Darauf aufbauend wird aufgezeigt, wie das vorhandene System deutschlandweit (und bei Bedarf darüber hinaus) erweitert und bereitgestellt werden kann.

## 2 Grundwasser-Online

Grundwasser-Online ([www.grundwasser-online.de](http://www.grundwasser-online.de)) ist eine bisher deutschlandweit einzigartige Plattform für die unternehmens- und gebietsübergreifende Überwachung und Steuerung der Grundwassersituation. Eingesetzt wird dieses Internetportal im Hessischen Ried, einem der bedeutendsten Grundwasserreservoirs Deutschlands

---

<sup>52</sup> Gerrit Seewald envi-systems GmbH – Ingenieurbüro für Umweltinformatik

mit aktiver Bewirtschaftung (Förderung und Infiltration). Die führenden Wasserversorgungsunternehmen der Region sowie Ingenieurbüros und Behörden tauschen über diese Plattform ausgewählte Daten automatisiert untereinander aus. Seit Oktober 2007 steht das Portal auch für neue interessierte Unternehmen und Behörden offen.

### 3 Nutzungsmöglichkeiten

Neben einem öffentlichen Portal, in dem ausgewählte Ganglinien tagesaktuell auf Basis eines WebGIS zur Verfügung gestellt werden, stehen umfassende Module zur Datenerfassung, -verwaltung und -auswertung in Form von Listen, Diagrammen sowie flächenhaften Auswertungen den Fachanwendern zur Nutzung über das Internet zur Verfügung. Der Zugriff auf die vorhandenen Datenbestände wird mittels feingranularer Rechte für jede einzelne Benutzergruppe bis unterhalb der Objektebene abgebildet und sichergestellt. Diese Rechteverwaltung ist Voraussetzung für den freiwilligen Datenaustausch zwischen den beteiligten Unternehmen, den Verwaltungsbehörden, den Ingenieurbüros und schließlich auch der interessierten Öffentlichkeit.

Die Basis für Grundwasser-Online bildet eine zentrale Datenbank. Jedes Unternehmen besitzt zudem die Möglichkeit, über lokale und mobile Softwaremodule die hausinternen Erfassungs- und Auswertungsprozesse gezielt zu unterstützen. Diese Module arbeiten auf Basis von lokalen (bzw. mobilen) Datenbanken, die sich mittels triggerbasierter, optimistisch-assynchroner Replikationsverfahren über die zentrale Serverdatenbank abgleichen, so dass letztlich zwischen allen (herstellerunabhängigen) Datenbanken innerhalb des Netzwerks ein Datenaustausch erreicht wird (Abbildung 1):

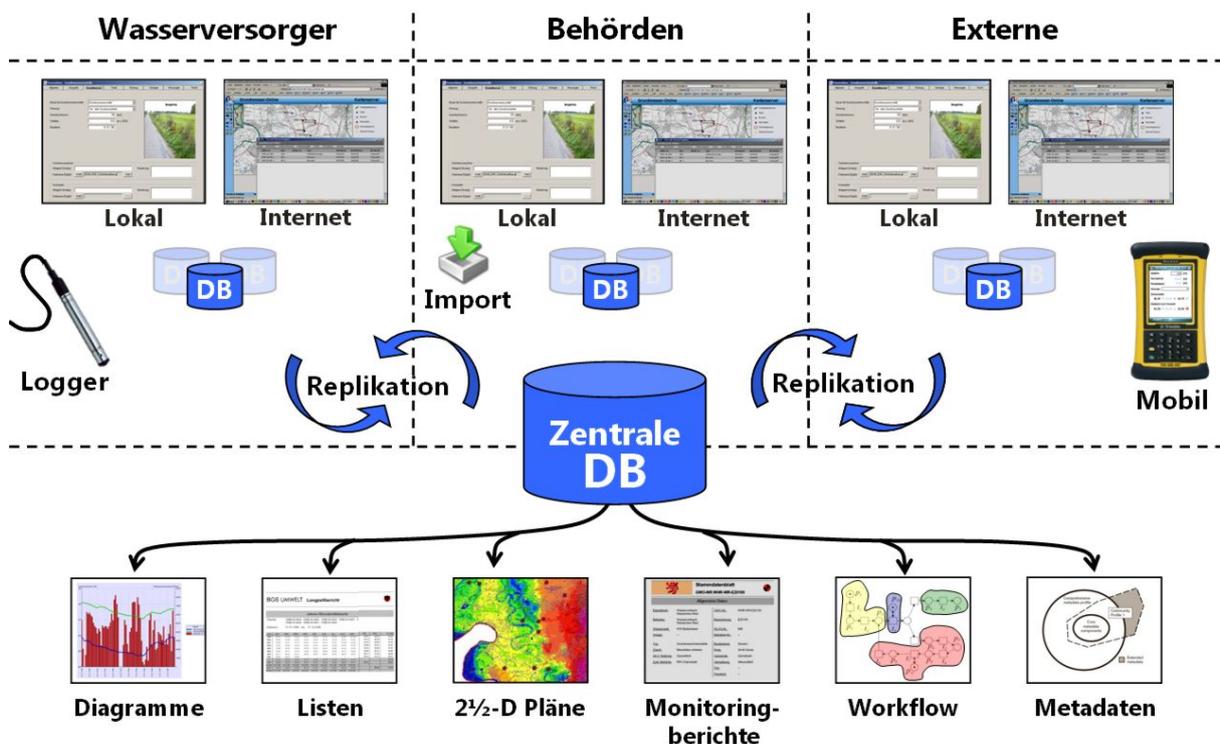


Abbildung 1: Systemskizze mit Erfassungs- und Auswertungsmöglichkeiten

## 4 Perspektive

Für eine räumliche Ausweitung des Projektraums sind bereits jetzt aus technischer Sicht alle Voraussetzungen vorhanden. Entscheidend für eine erfolgreiche Etablierung des Systems werden die Aspekte Inbetriebnahme, Abwicklungsorganisation sowie die Motivation der verschiedenen Gruppen zur aktiven Teilnahme sein. Die beteiligten Gruppen und ihre Aufgaben bzw. Interessen lassen sich wie folgt abbilden (Abbildung 2).

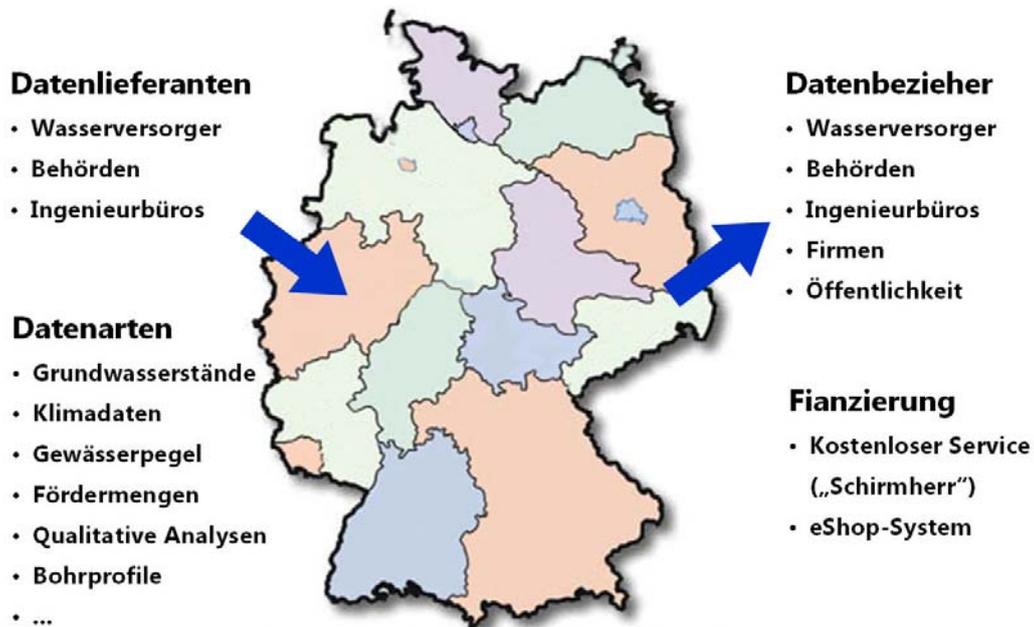


Abbildung 2: Deutschlandweite Bereitstellung von Grundwasserinformationen

## 5 Motivation

Für die einzelnen Gruppen (Datenlieferanten, Datenbezieher, Schirmherr, Portalbetreiber) ergeben sich unterschiedliche Gründe für eine Beteiligung an Grundwasser-Online.

### Datenlieferant:

- Reduzierung der Auskunftstätigkeiten an Dritte (Behörden, benachbarte WVU, Ingenieurbüros, Öffentlichkeit).
- Imagepflege durch offenen Umgang mit unternehmenseigenen Umweltdaten.
- Nutzung eines preiswerten (evtl. kostenfreien) Services.
- Kein zusätzlicher Aufwand der Datenaufbereitung und -bereitstellung durch (teil-) automatisierte Datenübertragung.
- Möglichkeiten der zusätzliche Einnahmen durch Nutzung des bereitgestellten eShop-Systems (Ausbaustufe II).

### Datenbezieher:

- Zentrale Anlaufstelle für den Erhalt von Grundwasser- und Umweltdaten.
- Angefragte Daten werden in einheitlichen Auswertungsarten (Listen, Diagramme) und -formaten bereitgestellt.
- Der Basis-Service ist kostenlos.

- Nach dem Aufbau eines eShop-Systems bleiben die Kosten transparent bzw. kalkulierbar.

#### **Schirmherr:**

- Die hohe Präsenz und Außendarstellung trägt zur Imagepflege und Profilbildung bei.
- Grundsätzlich ist von einer Steigerung der Datenqualität auszugehen, da durch die Datenbereitstellung im Internet unternehmensinterne Prozesse ggf. optimiert werden.
- Der Zugriff auf den gesamten, unternehmensübergreifenden Datenbestand vereinfacht sich signifikant.

#### **Portalbetreiber:**

- Zusätzlich lassen sich lokale und mobile Erfassungs- und Auswertungskomponenten vertreiben, um die Datenbereitstellung noch komfortabler zu gestalten
- Steigerung der Bekanntheitsgrades und Imagepflege.

Ziel sollte es sein, das System nach einer für die Datenlieferanten und Datenbezieher kostenlosen Startphase über einen eShop zu finanzieren, so dass es sich langfristig selbst trägt.

# Vom Umweltdatenkatalog zum InGrid® Data Catalog - Aktuelle Anpassungen des UDK-Datenmodells und der Software InGrid® im Hinblick auf INSPIRE

Martin Klenke, Fred Kruse, Stefanie Uhrich, Christiane Giffei, Sybille Peters<sup>53</sup>,  
[martin.klenke@portal.u.de](mailto:martin.klenke@portal.u.de)

## Abstract

Currently, a web-based application for the collection and maintenance of environmental metadata is under development within the scope of the project InGrid® 2.0. This new application will supersede the Windows-UDK desktop application, actually being widely distributed within the German environmental authorities. Along with the front-end development several data model adjustments are necessary in order to improve the performance of external interfaces as well as to calibrate to new demands with respect to content. Regarding the latter, of most importance are requirements emerging within the scope of the INSPIRE1 directive. For the metadata community, of particular significance are the INSPIRE Implementing Rules for Metadata which are supposed to be published officially in the second half of 2008. In the first part of our paper we describe modifications of the ISO 19115/ ISO 19119 conformal UDK metadata model, initiated by the INSPIRE Draft Implementing Rules for Metadata (Version 4).

While the INSPIRE implementing rules are getting more concrete, discussions arise within the German environmental and geospatial data and metadata communities on how to establish an efficient national metadata infrastructure to serve INSPIRE. At present, different approaches are discussed in the context of the GDI-DE2 (Spatial Data Infrastructure Germany) pilot project Geodatenkatalog-DE. One aim of the pilot project is to establish the prototype of a central German metadata-node for INSPIRE. In the second part of our article we give a brief overview on the state of discussion in the pilot project as well as on some desired InGrid®-functionalities, probably being relevant for the INSPIRE metadata flow in the near future.

## Zusammenfassung

Im Zuge der Entwicklung und Bereitstellung einer web-basierten Metadatenerfassungs- und Pflegekomponente im Rahmen des Projekts InGrid® 2.0 [Klenke et al., 2007] als Ersatz für den in der deutschen Umweltverwaltung verbreiteten Windows-UDK (Umweltdatenkatalog) werden verschiedene Anpassungen am ISO 19115- und ISO 19119-konformen UDK-Datenmodell erforderlich, teils um die Abfrageperformanz externer Schnittstelle zu erhöhen, teils um neuen Funktionalitäten und inhaltlichen Anforderungen Rechnung zu tragen. Von besonderer Relevanz sind inhaltliche Erweiterungen, die derzeit auf Grundlage der

---

<sup>53</sup> Koordinierungsstelle PortalU im Niedersächsischen Ministerium für Umwelt und Klimaschutz

INSPIRE<sup>54</sup> Draft Implementing Rules for Metadata (Version 4) durchgeführt werden, um sicher zu stellen, dass über PortalU®<sup>55</sup>/InGrid® die Inhalte der angeschlossenen Umweltdatenkataloge INSPIRE-konform abgegeben werden können. Im ersten Teil unseres Artikels beschreiben wir die diesbezüglich durchgeführten Anpassungen am UDK-Datenmodell.

Mit der zunehmenden Konkretisierung der verschiedenen INSPIRE-Durchführungsbestimmungen stellt sich im nationalen Umfeld die Frage, mit welchem Architekturkonzept GDI-DE<sup>56</sup> (Geodateninfrastruktur Deutschland) den Anforderungen der INSPIRE-Richtlinie begegnet. Verschiedene Ansätze werden aktuell im Rahmen des GDI-DE-Modellprojekts Geodatenkatalog-DE diskutiert. Im Modellprojekt soll u.a. prototypisch ein zentraler, deutscher INSPIRE-Metadatenknoten entwickelt werden. Im zweiten Teil unseres Artikels geben wir einen kurzen Überblick über den aktuellen Stand der Diskussion im Modellprojekt sowie über geplante Erweiterungen der PortalU®-Software InGrid®, die im Rahmen von INSPIRE relevant erscheinen.

## 1 Einführung: INSPIRE und PortalU®

Am 15. Mai 2007 trat die Richtlinie 2007/2/EG zur Schaffung einer Geodateninfrastruktur in der Europäischen Gemeinschaft (INSPIRE, Infrastructure for Spatial Information in Europe) in Kraft. Ziel der Richtlinie, die innerhalb von zwei Jahren von den Mitgliedstaaten der EU jeweils in nationales Recht umgesetzt werden muss, ist es, die grenzübergreifende Nutzung von Daten in Europa zu erleichtern. INSPIRE fordert webbasierte Online-Dienste für die Suche, die Visualisierung und den Download der Daten. Die technischen Einzelheiten zur Umsetzung der Richtlinie werden über Durchführungsbestimmungen geregelt, die zurzeit erarbeitet werden.

Als erste Durchführungsbestimmung soll diejenige zum Metadaten-Modell von INSPIRE veröffentlicht werden. Der Regelungsausschuß „INSPIRE Committee“ hat am 14. Mai 2008 den Durchführungsbestimmungen zu Metadaten ohne Gegenstimmen und Enthaltungen zugestimmt. Das Votum des Regelungsausschusses wird von der EU-Kommission nun dem Europäischen Parlament (EP) zugeleitet. Das EP hat eine Rückäußerungsfrist von 4 Wochen. Im Anschluss werden die Durchführungsbestimmungen dem „EU-Kabinett“ (EU-Kommissare) zur Zustimmung vorgelegt. Danach können die Durchführungsbestimmungen veröffentlicht werden und treten in Kraft. Die EU-Kommission rechnet damit, dass die Metadaten-Durchführungsbestimmungen zum Ende des Sommers veröffentlicht werden. Sie sind dann unmittelbar geltendes Recht in den Mitgliedsstaaten.

Die inhaltliche Ausarbeitung, an der die Koordinierungsstelle PortalU als Mitglied des INSPIRE Drafting Teams Metadata direkt beteiligt war, ist nach der Zustimmung des Regelungsausschusses in einem Stadium, dass für die Veröffentlichung keine Änderungen gegenüber dem derzeitigen Entwurf (Draft Implementing Rules for Metadata, Version 4) mehr erwarten läßt. Für die PortalU®-Software InGrid®, deren Metadatenkatalog (UDK) bislang ISO 19115 / ISO 19119 konform gestaltet ist, wurden die für INSPIRE erforderlichen Datenmodellerweiterungen daher aus dem Draft / Version 4 abgeleitet.

---

<sup>54</sup> <http://www.ec-gis.org/inspire/>

<sup>55</sup> <http://www.portalu.de/>

<sup>56</sup> <http://www.gdi-de.org/>

## 2 Erforderliche Anpassungen am UDK-Datenmodell

In den folgenden Unterkapiteln werden die notwendigen Anpassungen und Erweiterungen am Metadatenmodell von InGrid® (UDK-Datenmodell) feldweise beschrieben. Dort, wo es sich um echte Erweiterungen handelt, ist zu erwarten, dass die beschriebenen Anpassungen auch für andere, nicht auf InGrid®-Software basierende, ISO-19115/19119-konforme Kataloge Relevanz besitzen.

Der Umstieg auf das neue Datenmodell erfordert für die bestehenden Umweltdatenkataloge eine Altdatenübernahme. Wo möglich, soll die Befüllung der neuen Felder initial automatisiert erfolgen. Die Überführungsregeln werden, soweit vorhanden, ebenfalls dargestellt. Die überführten Umweltdatenkataloge werden im Folgenden auch als InGrid®-Datenkataloge (IDCs) bezeichnet.

### 2.1 Konformität (Draft: 2.2.7 Conformity)

Die Feldgruppe Konformität ist eine neue Pflichtfeldgruppe, in der beschrieben wird, in wie weit und zu welcher Spezifikation die in dem Metadatenobjekt beschriebenen Daten konform sind. Die Feldgruppe besteht aus einer 1-zu-N-Zuordnung von Feldpaaren.

Bei der Spezifikation der Konformität soll angegeben werden, zu welcher Durchführungsbestimmung der INSPIRE-Richtlinie die beschriebenen Daten konform sind. Hierbei soll die Möglichkeit bestehen mehr als einen Eintrag vorzunehmen. Es können grundsätzlich also auch Nicht-INSPIRE-Spezifikationen aufgenommen werden. Pro Spezifikationen sollen der Name der Spezifikation und ein entsprechendes Referenzdatum (Datum der Publikation, Datum der letzten Überarbeitung oder Erstellungsdatum) aufgeführt werden. Das Feld Grad der Konformität wird über eine Auswahlliste mit folgenden Einträgen zu füllen sein:

- konform: Die Datenquelle ist vollständig konform zur zitierten Spezifikation
- nicht konform: Die Datenquelle ist nicht konform zur zitierten Spezifikation
- nicht evaluiert: Die Konformität der Datenquelle wurde noch nicht evaluiert

Standardvorgabe bei der Altdatenübernahme für Grad der Konformität ist nicht evaluiert, Standardvorgabe für Spezifikation der Konformität ist die INSPIRE-Richtlinie. Zukünftig, wenn die verschiedenen Durchführungsbestimmungen (insbesondere Data Specifications) verabschiedet sind, kann das Freitextfeld Spezifikation der Konformität ggf. durch eine Auswahlliste ersetzt bzw. unterstützt werden.

### 2.2 Eindeutiger Identifikator der Datenquelle (Draft: 2.2.1.5 Unique resource identifier)

Das Feld Identifikator der Datenquelle ist ein neues Pflichtfeld, das für jeden in den Objekten beschriebenen Datensatz einen eindeutigen Identifikator enthalten soll. Das Feld ist nur für die UDK-Objektklasse Geoinformation/Karte vorgeschrieben.

Bisher ist unklar, wie dieser Identifikator aussehen wird, dieser Punkt soll über die Durchführungsbestimmungen zu den Data Specifications geklärt werden.

Der Fachbezug der UDK-Objektklasse Geoinformation/Karte wird um ein einzeliges Textfeld erweitert, dessen Befüllung verpflichtend ist.

Im Rahmen der Altdatenübernahme wird das Feld Identifikator der Datenquelle für jedes Objekt der Objektklasse Geoinformation/Karte automatisch mit der

Kombination aus dem Katalognamen gefolgt von einem Doppelpunkt und dem Identifikator des Metadaten-Objektes (automatisch für jedes UDK-Objekt generierte UUID) gefüllt.

### **2.3 Zugangs- und Nutzungsbeschränkungen (Draft: 2.2.8. Constraint related to access and use)**

Die INSPIRE-Anforderungen werden mit einer zweiseitigen Tabelle erfüllt, die sich aus den Feldern Zugangsbeschränkungen und Nutzungsbedingungen zusammensetzt. Mindestens ein Eintrag ist verpflichtend, wobei beide Felder ausgefüllt sein müssen.

#### **2.3.1 Zugangsbeschränkungen (Draft: 2.2.8.2 Limitations on public access)**

Das Feld Zugangsbeschränkungen bezieht sich auf die Auflistung der in Artikel 13 der INSPIRE-Richtlinie erlaubten Beschränkungen. Pro Zeile der Tabelle Zugangsbeschränkungen und Nutzungsbedingungen ist ein Eintrag verpflichtend.

Das Feld soll über die folgende, fest vorgegebene, Auswahlliste befüllt werden:

- Keine
- Aufgrund der Vertraulichkeit der Verfahren von Behörden
- Aufgrund internationaler Beziehungen, der öffentlichen Sicherheit oder der Landesverteidigung
- Aufgrund laufender Gerichtsverfahren
- Aufgrund der Vertraulichkeit von Geschäfts- oder Betriebsinformationen
- Aufgrund der Rechte des geistigen Eigentums
- Aufgrund der Vertraulichkeit personenbezogener Daten
- Aufgrund des Schutzes einer Person
- Aufgrund des Schutzes von Umweltbereichen

Die Standard Vorgabe für den ersten Eintrag ist *keine*.

#### **2.3.2 Nutzungsbedingungen (Draft: 2.2.8.1 Conditions applying to access and use)**

Das Feld Nutzungsbedingungen soll die Bedingungen zur Nutzung des beschriebenen Datensatzes bzw. des Services enthalten. Es sollen beispielsweise die Kosten für die Nutzung der Daten angegeben werden. Pro Zeile der Tabelle Zugangsbeschränkungen und Nutzungsbedingungen ist ein Eintrag verpflichtend.

Bei frei nutzbaren Daten bzw. Services soll laut Durchführungsbestimmungen Keine Einschränkungen eingetragen werden. Sind die Bedingungen unbekannt, ist Unbekannte Nutzungsbedingungen einzutragen. Die UDK-Felder Nutzungsanmerkungen und Kosten in der UDK-Rubrik Verfügbarkeit entfallen.

#### **2.3.3 Altdatenübernahme**

Bei der Altdatenübernahme wird geprüft, ob eines der existierenden, derzeit optionalen UDK-Felder Nutzungsanmerkungen bzw. Kosten in der UDK-Rubrik Verfügbarkeit einen Eintrag enthält. Es werden folgende Fälle unterschieden:

- Kein Eintrag: Das Feld *Zugangsbeschränkungen* erhält einen Eintrag *Keine*, das Feld *Nutzungsbedingungen* erhält einen Eintrag *Keine Einschränkungen*
- Existierender Eintrag: Das Feld *Zugangsbeschränkungen* erhält einen Eintrag *Aufgrund der Rechte des geistigen Eigentums*. Das Feld *Nutzungsbedingungen* wird mit dem Inhalt der jetzigen beiden Felder gefüllt. Sind in beiden Feldern Einträge enthalten, werden sie durch eine Absatzmarke getrennt zusammengeführt.

## 2.4 Fachliche Grundlage (Draft: 2.2.6.1 Lineage)

In den Durchführungsbestimmungen zu INSPIRE wird das Feld Fachliche Grundlage zum Pflichtfeld für Geodaten (Draft: spatial datasets) erklärt, also für diejenigen Objekte, die über die UDK-Objektklasse "Geoinformation/Karte" beschrieben werden. Es soll u.a. Angaben zur Entstehungsgeschichte sowie zur Qualität des Datensatzes enthalten.

Das Feld Fachliche Grundlage ist derzeit im UDK-Fachbezug Geoinformation/Karte ein optionales Feld. Es wird zukünftig als Pflichtfeld gekennzeichnet und zu den weiteren Pflichtfeldern an den Anfang des Fachbezugs verschoben. Beim Speichern wird, wie bei allen Pflichtfeldern, die Befüllung überprüft. Ist das Feld in einem existierenden Katalog nicht ausgefüllt, wird es im Rahmen der Altdatenübernahme mit dem Eintrag Unbekannt befüllt.

## 2.5 Service-Klassifikation (Draft: 2.2.2.2 Spatial data service type)

Das aktuelle UDK-Feld Servicetyp im Fachbezug der UDK-Objektklasse Dienst / Anwendung / Informationssystem wird in das Feld Service-Klassifikation überführt. Es wird wie bisher ein über eine Auswahlliste zu füllendes Pflichtfeld ausschließlich für die Objektklasse Dienst / Anwendung / Informationssystem. Die bisherige Auswahlliste wird durch die folgende geschlossene Auswahlliste (Draft: B.3 Spatial data service type) ersetzt:

- Suchdienste (CSW) [Discovery Service]
- Visualisierungsdienste (WMS) [View Service]
- Zugriffsdienste (WFS) [Download Service]
- Transformationsdienste (WCTS) [Transformation Service]
- Verkettete Geodatendienste [Invoke Spatial Data Service]
- Andere Dienste [Other Service]

Im Rahmen der Altdatenübernahme wird der existierende Service-Eintrag anhand von Tabelle 1 umgesetzt.

Bisheriger Eintrag	Neue ID
CSW	1
WMS	2
WFS	3
Alle anderen Einträge (einschl. leeres Feld)	6

Tabelle 1: Service-Klassifikation

## 2.6 Typ des Dienstes (Draft: 2.2.3 Keyword, Annex B.4)

Dieses Feld wird mit INSPIRE für spatial data services verpflichtend. Umgesetzt wird es als einspaltige Tabelle im Fachbezug der UDK-Objektklasse Dienst / Anwendung / Informationssystem. Die von INSPIRE fest vorgegebene Auswahlliste (Draft: Annex B.4) wird für InGrid®-Datenkataloge um die ID=901: Non Geographic Service erweitert.

Bei der Altdatenübernahme wird gemäß Tabelle 2 die erste Zeile der neuen Tabelle mit einem Wert vorbelegt, der sich aus dem existierenden UDK-Feld Service-Klassifikation (siehe 2.5) ergibt.

<i>ID Service-Klassifikation</i>	<i>ID Typ des Dienstes</i>
1	207
2	202
3	201
6	901

Tabelle 2: Typ des Dienstes

## 2.7 Erstellungsmaßstab (Draft: 2.2.6.2 Spatial resolution)

Die bereits im Fachbezug Geoinformation/Karte vorhandene Feldgruppe Erstellungsmaßstab wird zusätzlich im Fachbezug der Objektklasse Dienst / Anwendung / Informationssystem eingeführt. Einträge in diese Feldgruppe sind optional.

## 2.8 Email-Adresse (Draft: 2.2.10.1 Metadata point of contact)

Für alle Adressen gilt, dass die Email-Adresse, abweichend vom aktuellen UDK-Datenmodell, ein Pflichtfeld wird. Im InGrid®-Datenkatalog wird die Angabe mindestens einer Email-Adresse daher verpflichtend. Die zugehörige Feldgruppe wird zu den anderen Pflichtfeldern an den Anfang der Erfassungsmaske gerückt und optisch als Pflichtfeld gekennzeichnet. Beim Speichern einer Adresse wird die Angabe mindestens einer Email-Adresse abgeprüft.

Die Angabe der Postadresse, die derzeit im UDK verpflichtend ist, wird durchgängig zu einem optionalen Feld, d.h. es verbleibt lediglich die Email-Adresse als Pflichtfeld in der Erfassungsmaske.

### 2.8.1 Altdatenübernahme

Ist für einen Adresseintrag keine Email-Adresse angegeben, wird in der UDK-Adress-Hierarchie schrittweise aufwärts bis zur Institution nach einer Email-Adresse gesucht. Die erste bei diesem Vorgang gefundene Email-Adresse wird dem Adresseintrag zugeordnet.

Ist der gesamten Institution keine Email-Adresse zugeordnet, wird die Email-Adresse des Katalogadministrators allen Adressen der Institution zugeordnet. Hierzu wird die Eingabe der Email-Adresse des Katalogadministrators beim Start der Altdatenübernahme vorgesehen. Im Normalfall sollen die Kataloge vor der Altdatenübernahme dahingehend überarbeitet werden, dass zumindest jeder Institution eine Email-Adresse zugeordnet ist.

## 2.9 Verschlagwortung (Draft: 2.4 Keyword)

Im UDK ist bislang die Vergabe von drei Thesaurusschlagworten pro Metadatenobjekt obligatorisch. Genutzt wird das Wortgut des UMTHEs<sup>57</sup>, das zurzeit noch im Windows-UDK-Erfassungswerkzeug integriert ist und ab InGrid® 2.0 über den Semantic-Network-Service<sup>58</sup> des Umweltbundesamts angeschlossen wird.

Für INSPIRE wird die Vergabe eines GEMET<sup>59</sup>-Schlagwortes (General Multilingual Environmental Thesaurus) verpflichtend. Da der Wortgutumfang des GEMET erheblich geringer ist als der des UMTHEs und darüber hinaus der deutsche GEMET-Anteil komplett im UMTHEs abgebildet ist, ergeben sich für InGrid® 2.0 nur marginale Änderungen. Auch zukünftig werden bei der UDK/IDC Metadatenerfassung drei UMTHEs-Schlagworte verpflichtend sein. Über einen Prüfmechanismus wird abgeprüft, ob sich unter den drei gewählten UMTHEs-Deskriptoren ein GEMET-Schlagwort befindet. Ist dies nicht der Fall, werden die Bearbeiter vor dem endgültigen Verspeichern des Metadatenobjekts bzw. vor Übersendung an die Qualitätssicherung zur Eingabe eines GEMET-Schlagwortes verpflichtet. Dabei werden sie von verschiedenen Thesaurus-Assistenten unterstützt: Dem Thesaurus-Navigator, der ein blättern in der Thesaurushierarchie (Abbildung 1) sowie eine Suche im UMTHEs-Wortgut (Abbildung 2) ermöglicht sowie einem Verschlagwortungsassistenten, der Inhalte von Textfeldern an den SNS übergibt und vom Service Schlagwort-Vorschläge zurück erhält.

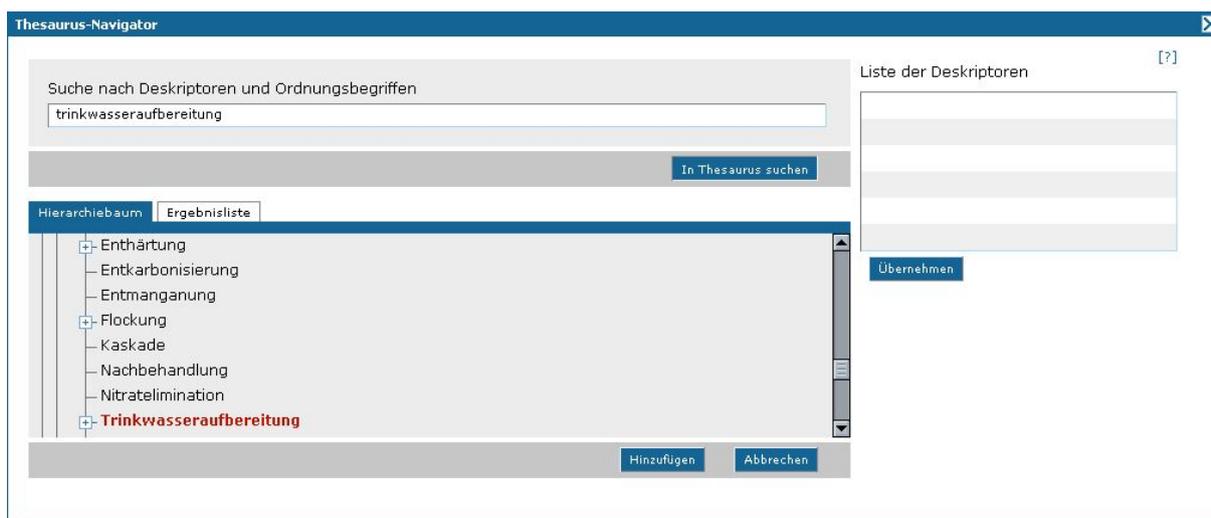


Abbildung 1: InGrid® 2.0 Thesaurus-Navigator / Hierarchiebaum

Für die Altdatenübernahme ist für jedes Objekt eine automatische Überprüfung der existierenden Schlagworte vorgesehen. Wird auf diesem Wege zunächst kein GEMET-Schlagwort gefunden, wird in der UMTHEs-Hierarchie von den eingetragenen Schlagworten aus so weit „nach oben“ gewandert, bis ein entsprechender GEMET-Deskriptor gefunden wird. Dieser wird als zusätzliches Schlagwort zum bestehenden Metadatenobjekt hinzu gefügt.

<sup>57</sup> <http://www.umweltbundesamt.de/uba-info/dokufabib/thes.htm>

<sup>58</sup> <http://www.semantic-network.de>

<sup>59</sup> <http://www.eionet.europa.eu/gemet>

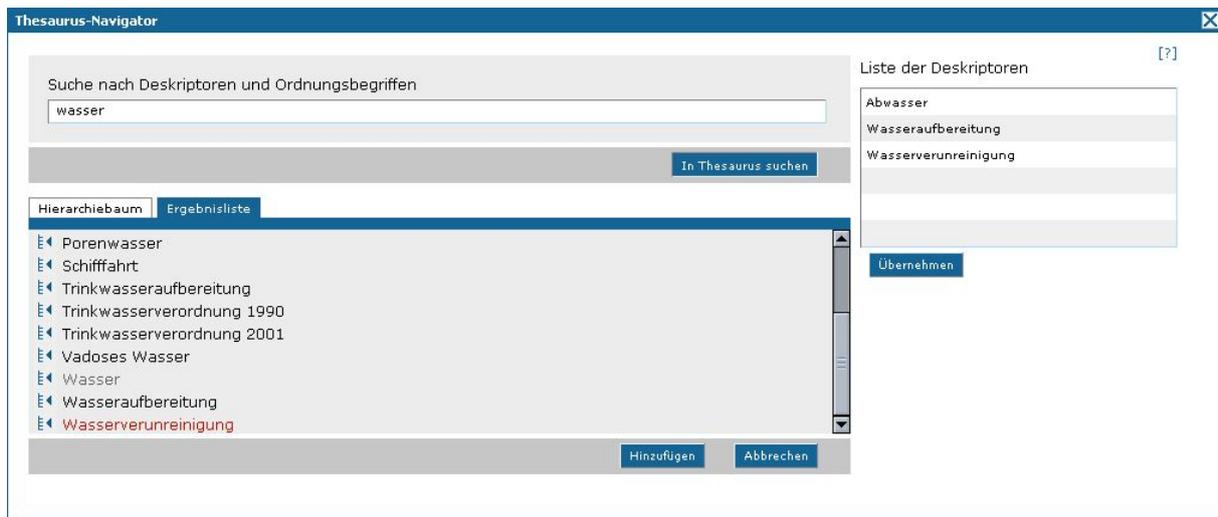


Abbildung 2: InGrid® 2.0 Thesaurus-Navigator / Suche im Thesaurus

### 3 GDI-DE/ INSPIRE Metadaten Infrastruktur

Im Folgenden werden einige Überlegungen zur Frage, wie eine deutsche Metadateninfrastruktur für INSPIRE realisiert werden könnte, vorgestellt. Die Darstellung bewegt sich dabei entlang des Diskussionsstands im GDI-DE Modellprojekt Geodatenkatalog-DE, in dem die Koordinierungsstelle PortalU aktiv mitarbeitet. Aus den GDI-DE- und INSPIRE-Anforderungen werden, als Diskussionsgrundlage, verschiedene Weiterentwicklungsoptionen für die Software InGrid® abgeleitet.

#### 3.1 GDI-DE Modellprojekt Geodatenkatalog-DE

Ziel des Modellprojekts ist der Aufbau einer Topologie existierender Katalogdienste und deren Bündelung im Sinne einer zentralen Abfragemöglichkeit der verteilten Dienste. Hierzu ist es erforderlich, typische Metadateninformationssysteme auf höheren Verwaltungsebenen (z.B. Bundes-MIS, Landes-MIS, länderübergreifende Fach-MIS) im Projekt zu verknüpfen. Konkret sollen eine Auswahl an fachübergreifenden Geodatenkatalogen der Länder, der Geodatenkatalog des Bundes sowie die für INSPIRE absehbar relevanten und länderübergreifenden Fachkataloge PortalU® und AdV-MIS<sup>60</sup> zusammengeführt werden.

Nutzeranforderungen werden im Projekt u.a. über die Beteiligung des JRC<sup>61</sup> und der GIW-Kommission<sup>62</sup> berücksichtigt.

Die technische Umsetzung wird den ersten Prototyp eines zentralen, deutschen INSPIRE-Metadatenknotens bilden. Die angestrebte Lösung soll einen performanten Zugriff und eine hohe Verfügbarkeit sowie Redundanzfreiheit gewährleisten.

#### 3.2 Anforderungen seitens GDI-DE und INSPIRE

Im GDI-DE-Architekturkonzept<sup>63</sup> ist als Spezifikation für Katalogdienstschnittstellen das OpenGIS® Catalogue Services Specification 2.0.2 - ISO Metadata Application Profile (1.0.0)<sup>64</sup> festgeschrieben. Dieses Applikationsprofil wird, unter

<sup>60</sup> <http://www.geodatenzentrum.de/isoinfo/>

<sup>61</sup> <http://ec.europa.eu/dgs/irc/>

<sup>62</sup> <http://www.geobusiness.org/Geobusiness/Navigation/giwk.html>

<sup>63</sup> [http://www.gdi-de.org/de/download/GDI\\_ArchitekturKonzept\\_V1.pdf](http://www.gdi-de.org/de/download/GDI_ArchitekturKonzept_V1.pdf)

<sup>64</sup> [http://portal.opengeospatial.org/files/?artifact\\_id=21460](http://portal.opengeospatial.org/files/?artifact_id=21460)

Berücksichtigung des aktuellen Diskussionsstands im INSPIRE Drafting Team Network Services, auch der fachliche Katalogdienst-Standard für INSPIRE werden.

Die technisch/funktionalen Anforderungen an INSPIRE-konforme Katalogdienste werden im Drafting Team Network Services diskutiert und in den Draft Implementing Rules for Discovery and View Services dokumentiert. Wenngleich die zu erfüllenden Performanzwerte noch nicht endgültig fest stehen, zeichnet sich ab, dass Harvesting- und Indexierungsvorgänge erforderlich sein werden um einen INSPIRE-konformen deutschen Metadatanknoten bereit zu stellen. So werden in den Draft Implementing Rules for Discovery and View Services (IR1) Version 2.0 u.a. folgende Anforderungen gestellt:

- Maximale Zeit zum Senden der Rückmeldung nach einer Suchanfrage (in 90% der Fälle, bei 250 gesendeten Metadatensätze): 3 Sekunden
- Minimum möglicher gleichzeitiger Anfragen an den Service: 100

Die Suchkriterien für INSPIRE-Suchdienste werden in Artikel 11(2) der INSPIRE-Richtlinie festgelegt. Danach sind zumindest Kombinationen von folgenden Suchkriterien zu gewährleisten:

- Schlüsselwörter
- Klassifizierung von Geodaten und Geodatendiensten
- Qualität und Gültigkeit der Geodatensätze
- Grad der Übereinstimmung mit den in Artikel 7(1) vorgesehenen Durchführungsbestimmungen
- geografischer Standort
- Bedingungen für den Zugang zu und die Nutzung von Geodatensätzen und -diensten
- Behörden, die für die Erzeugung, Verwaltung, Erhaltung und Verbreitung von Geodatensätzen und -diensten zuständig sind

Für existierende Kataloge stellen sich damit Probleme analog zu den bereits weiter oben im Kontext der Metadatendurchführungsbestimmungen diskutierten. Zum Beispiel wird das Feld Konformität als neues Pflichtfeld eingeführt, um das in Punkt (4) definierte Suchkriterium zu erfüllen. Im Rahmen von Suchdiensten kommt hinzu, dass sichergestellt werden muss, dass dieses Feld durchsuchbar ist.

Ähnliches gilt auch für die Nutzungsvereinbarungen (6), die z.B. im InGrid®-Metadatenmodell derzeit als optionales Feld enthalten sind. Zum Einen wird dieses Feld zum Pflichtfeld erklärt, zum Anderen muss es auch über die Schnittstelle durchsuchbar gemacht werden.

Bezüglich der Funktionen der INSPIRE-Suchdienste fordern die Draft Implementing Rules for Discovery and View Services gegenüber CSW AP ISO 1.0 erweiterte Funktionalitäten, indem die CSW-Operationen Manager.Harvest und Manager.Transaction von optional auf verpflichtend umgesetzt werden.

Unter einem Harvesting-Mechanismus wird die Fähigkeit eines Suchdienstes verstanden, einen anderen Suchdienst abfragen zu können, wobei die Daten des abgefragten Suchdienstes in dem des Anfragenden gespeichert werden. Mit der Software-Version InGrid® 2.0 wird PortalU® diese Fähigkeit besitzen.

Unter einem Transaktionsmechanismus wird die Funktionalität eines Suchdienstes verstanden, über die CSW- (bzw. die CSWT-) Schnittstelle in den Datenbestand hineinschreiben zu können und zwar mit den Funktionen Einfügen, Ändern und Löschen. Diese Funktionalitäten, die von außen auf einem Katalog durchgeführt werden können, wurden für PortalU® als zu risikoreich und damit als explizit nicht gewünscht erachtet.

Aus Sicht der Koordinierungsstelle PortalU können beide Funktionalitäten für einzelne Suchdienste interessant sein. Eine verpflichtende Vorgabe für alle INSPIRE-Suchdienste erscheint aber sowohl aus Kostengründen als auch im Licht der fachlichen Anforderungen nicht sinnvoll und kann insbesondere für den Transaktionsdienst mit Risiken für die Datensicherheit verbunden sein. Die Koordinierungsstelle hat daher in der Kommentierungsphase zu den Draft Implementing Rules for Discovery and View Services gefordert, die Funktionalitäten Harvesting und Transaktion auf optional zurückzustufen.

### **3.3 Topologie der Katalogdienste**

Grundsätzlich kann die Strukturierung von Diensten einem fachbezogenen Aufbau folgen, vergleichbar dem Ansatz von Adv MIS und PortalU®, oder auch streng administrativ, basierend auf zentralen Kommunen- und Landeskatalogen bis hin zum Bundesknoten. Eine weitere Möglichkeit besteht im Verzicht auf eine mehrstufige Topologie, d.h. unterhalb des zentralen Knotens werden alle Kataloge unabhängig von ihrer Herkunft direkt angeschlossen. Faktisch existieren derzeit alle genannten Topologiearten parallel und in Mischformen und sollen im Modellprojekt berücksichtigt werden. In einem ersten Arbeitsschritt müssen die vorhandenen Topologien charakterisiert und in eine Registry überführt werden. Der Registry-Gedanke wird derzeit in einem weiteren GDI-DE-Modellprojekt konkretisiert.

Insbesondere die hohen Performanzanforderungen der Draft Implementing Rules for Discovery and View Services (IR1) Version 2.0 (s.o.) stecken den Rahmen für geeignete Implementierungstechnologien ab. Ohne Harvesting- und Indexierungsmechanismen wird eine INSPIRE-konforme Katalogimplementierung nicht möglich sein. Die Vorteile solcher Verfahren liegen in der Erhöhung der Antwortgeschwindigkeit sowie in der Möglichkeit gemischte, datenquellen-übergreifende und nach Relevanz im Hinblick auf den Suchbegriff sortierte Ergebnislisten zu erstellen. Allerdings entstehen durch das Harvesting in zentrale Repositorien auch Probleme wie Dublettenpotential, redundante Datenhaltung, Daten-Aktualität, rechtliche Fragestellungen und das Erfordernis, große Datenmengen regelmäßig über das Netz zu transportieren. Diese sollen im Modellprojekt adressiert und praktisch analysiert werden.

### **3.4 Erweiterungen der Software InGrid®**

Um den steigenden Anforderungen an Brokersysteme wie PortalU® gerecht zu werden, sollen insbesondere die InGrid®-CSW-Komponenten erweiterte Funktionalitäten erhalten. Die Anpassungen befinden sich aktuell im Abstimmungsstadium und sollen im Folgenden vorab zur Diskussion gestellt werden.

#### **3.4.1 Harvesting und Indexierung**

In einem ersten Schritt ist ein Mechanismus vorzusehen, der über Harvesting-Abfragen in der Lage ist, den gesamten Dateninhalt einzelner Kataloge zu ernten

und anschließend zu indexieren. Als Indexierer ist Apache Lucene<sup>65</sup> zu verwenden, um die Kompatibilität des Rankings mit anderen Schnittstellenadaptern (iPlugs) von InGrid® zu gewährleisten. Aus Gründen der Skalierbarkeit, Lastverteilung und Flexibilität erhält jede angeschlossene Datenquelle einen eigenen Index, dessen Harvestingrhythmus individuell konfigurierbar ist. Der Index wird am jeweiligen iPlug vorgehalten und konfiguriert. Die Zusammenführung der unterschiedlichen Indizes erfolgt zum Suchanfragezeitpunkt am InGrid®-iBus [Klenke et al., 2006].

Das Harvesting soll über die Abfrage GetRecords (ElementSetName=brief) mit einer Wildcard-Query und anschließendem Herunterladen der Metadaten (GetRecordById, ElementSetName=full) realisiert werden. Die Inhalte werden am jeweiligen iPlug in Form einer XML-Datei temporär gespeichert.

Der Indexprozess wird nach dem Harvesting automatisch angestoßen. Er indiziert die CSW-XML-Datei(en). Die Indizierung erfolgt nach einer Indexfeld-zu-XPath-Abbildungsregel. Indexfelder werden dabei einem bestimmten XPath zugeordnet. Damit eine einfache Integration in die InGrid® 1.x Systemlandschaft möglich ist, werden als Feldnamen die Bezeichnungen verwendet, die im iPlug auf Basis des InGrid®-Datenmodells vorgegeben sind. Es wird ein Konfigurationsmechanismus implementiert, der das Austauschen der Mapping-Klassen vorsieht. Dadurch wird es möglich, unaufwendig verschiedene Mappings im iPlug zu unterstützen, z.B. AP ISO 0.9.3. und AP ISO 1.0.

Es ist sicherzustellen, dass trotz der Abbildung der Indexinhalte auf die iPlug-Feldnamen eine Abfrage der originären Inhalte der geharvesteten XML-Daten möglich ist (ElementSetName=original).

Standardmäßig erfolgt die Beantwortung von Anfragen performant aus dem Index, alternativ ist ein Mechanismus vorzusehen, der es ermöglicht, für die Detaillierungsergebnisdarstellung (GetRecordById) auch auf die Originaldatenquelle zurück zu greifen.

Es ist, insbesondere bei mehrstufigen Abfragekaskaden, möglich, dass aus unterschiedlichen CSW-Quellen die selben Datensätze geliefert werden. Ein Auftreten von Duplikaten in der resultierenden Antwort im Portal bzw. an CSW- und OpenSearch<sup>66</sup>-Schnittstelle ist daher zu berücksichtigen. Jeder CSW-Datensatz hat eine eindeutige ID, die zur Filterung von Dubletten verwendet werden kann. Sollten über die ID-Dubletten gefunden werden, ist zu prüfen, welcher Datensatz gemäß Zeitstempel der aktuellere ist. Dieser ist weiter zu geben. Der Harvest-Prozess sowie der iBus sind um einen entsprechenden Mechanismus zu erweitern.

### **3.4.2 Erweiterung des CSW-Profiles um ein PortalU®-Resultset**

Das AP ISO 1.0 wird um ein schmales PortalU®-Resultset erweitert. Ziel ist es, Informationssysteme zu unterstützen, die nur ein eingeschränktes, nicht CSW/ISO-konformes Datenmodell besitzen, jedoch über eine eigene Web-Darstellung der Datensätze verfügen. Das Resultset ist optional, das heißt, ISO- oder INSPIRE-konforme CSW müssen es nicht unterstützen. Tabelle 3 zeigt Inhalte des Resultsets.

---

<sup>65</sup> <http://lucene.apache.org/>

<sup>66</sup> <http://www.opensearch.org/>

CSW-Name	Pflicht (M) / Optional (O)	Beschreibung
Title	M	Titel des Datensatzes
Abstract	M	Beschreibung des Datensatzes
OnlineResource	M	URL zur Darstellung des Datensatzes im Original-System
Subject	O	Schlagworte
GeographicDescriptionCode	O	Name des geographischen Objekts
BoundingBox	O	geografische Bounding Box (WestBoundLongitude, NorthBoundLatitude, EastBoundLongitude, SouthBoundLatitude)
TemporalExtent	O	Zeitbezug der beschriebenen Daten (TempExtent_begin, TempExtent_end)

Tabelle 3: PortalU<sup>®</sup>-Resultset

### 3.4.3 Server für PortalU<sup>®</sup>-Resultset

Es wird ein einfacher Server erstellt, der es einem Anwender erlaubt, seine Datenbank auf das PortalU<sup>®</sup>-Resultset abzubilden.

Hierzu wird eine Administrationsoberfläche entwickelt, die dem Administrator die Struktur der Datenbank grafisch anzeigt und über die er das Mapping auf die Felder des PortalU<sup>®</sup>-Resultsets festlegen kann.

Eine Anfrage wird dann über die gemappten Felder der Datenbank geführt und das PortalU<sup>®</sup>-Resultsets zurückgegeben.

Es müssen nur Anfragen beantwortet werden, die auch von InGrid<sup>®</sup> unterstützt werden. Die Unterstützung weitergehender Anfragen, die ggf. auch durch das CSW-Profil definiert werden, ist nicht gefordert.

## 4 Zusammenfassung und Ausblick

Wenngleich die INSPIRE-Durchführungsbestimmungen aktuell an Kontur gewinnen bestehen noch große Unsicherheiten hinsichtlich deren praktischer Umsetzung. Grundsätzlich erscheinen die relevanten existierenden deutschen Katalogsysteme fachlich, durch ihre Unterstützung der ISO-Standards 19115 und 19119, zukunftsfähig aufgestellt. Auch auf technischer Ebene werden momentan bei vielen Betreibern, nicht zuletzt auf Grundlage des aktuellen GDI-DE-Architekturkonzepts, die Weichen in Richtung einer AP ISO 1.0-konformen CSW-Schnittstelle gestellt.

Die Ausgestaltung der Topologie deutscher Katalogdienste im Hinblick auf die ab 2010 verpflichtende Bereitstellung INSPIRE-relevanter Informationen ist demgegenüber noch völlig offen und birgt verschiedenste Schwierigkeiten, die nicht primär konzeptionell/theoretisch sondern insbesondere anhand konkreter technischer Umsetzungen beleuchtet und gelöst werden können. Insofern kommt dem GDI-DE-Modellprojekt Geodatenkatalog-DE eine maßgebliche Bedeutung für die Ausgestaltung der deutschen INSPIRE-Architektur zu.

## 5 Literaturverzeichnis

[Klenke et al., 2007] Klenke, Martin; Vögele, Thomas; Kruse, Fred; Lehmann, Hanno: PortalU® & InGrid® - Werkzeuge zur Erstellung, Recherche und Verteilung von Metadaten. AGIT 2007 - Symposium und Fachmesse für Angewandte Geoinformatik an der Universität Salzburg, 2007. <http://kurzlink.de/kle-2007-2>

[Klenke et al., 2006] Klenke, Martin; Kruse, Fred; Lehmann, Hanno; Riegel, Thomas; Vögele, Thomas: InGrid® 1.0 – The Nuts and Bolts of PortalU®. EnviroInfo 2006, Graz, 2006. <http://kurzlink.de/kle-2006-1>



# Entwicklung einer Maßnahmendatenbank für die Aufstellung von Bewirtschaftungsplänen und Maßnahmenprogrammen zur Umsetzung der Wasserrahmenrichtlinie

Friedhelm Hosenfeld<sup>67</sup>, Dirk Behrens<sup>68</sup>, Marcus Lempert<sup>69</sup>, Andreas Rinker<sup>1</sup>,  
Michael Trepel<sup>70</sup> und Angelika Steingraber<sup>4</sup>

## Abstract

The development of a central measure database (MDB) as a web application was initiated by state authorities of Schleswig-Holstein in order to manage all measures planned in water bodies in Schleswig-Holstein. This measure database supports the generation of river management plans and programmes of measures required by the European Water Framework Directive (WFD).

The first stage version of the measure database application was designed and implemented as a web application accessible by the responsible authorities and organisations via Internet. Access management with well-defined rights and roles ensures data security. Existing data about measures planned at river water bodies are interactively imported and completed by the project members.

The application combines all required information and relations to meet the various evaluation and reporting demands of regional, national, and European reporting tasks.

In a second stage the measure database will be extended to a universal instrument for planning, documenting, and financial purposes to support all aspects of measure management covering ideas and funding of measures as well as inspection and monitoring.

## Zusammenfassung

Von den schleswig-holsteinischen Landesbehörden wurde die Entwicklung einer zentralen Maßnahmendatenbank (MDB) als Web-Applikation initiiert, um die in den Wasserkörpern des Landes geplanten Maßnahmen zu verwalten. Diese Maßnahmendatenbank unterstützt die von der EG-Wasserrahmenrichtlinie (WRRL) geforderte Erstellung von Bewirtschaftungsplänen und Maßnahmenprogrammen, in denen die Maßnahmen zur Verbesserung der Gewässerqualität und zur

---

<sup>67</sup> DigSyLand - Institut für Digitale System Analyse & Landschaftsdiagnose, Zum Dorfteich 6, D-24975 Husby, EMail: {hosenfeld|rinker}@digsyland.de, Internet: <http://www.digsyland.de/>

<sup>68</sup> MLUR – Ministerium für Landwirtschaft, Umwelt und ländliche Räume des Landes Schleswig-Holstein, Mercatorstraße 3, D24106 Kiel, EMail: [Dirk.Behrens@mlur.landsh.de](mailto:Dirk.Behrens@mlur.landsh.de), Internet: <http://www.mlur.schleswig-holstein.de/>

<sup>69</sup> KISTERS AG, Arnold-Dehnen Straße 49, D-47138 Duisburg, Germany, EMail: [Marcus.Lempert@kisters.de](mailto:Marcus.Lempert@kisters.de), Internet: <http://www.kisters.de/>

<sup>70</sup> LANU – Landesamt für Natur und Umwelt des Landes Schleswig-Holstein, Hamburger Chaussee 25, D-24220 Flintbek, EMail: {mtrepel|asteingr}@lanu.landsh.de, Internet: <http://www.lanu-sh.de/>

Verminderung von Defiziten in Hinblick auf definierte Ziele zusammengestellt werden.

Die erste Ausbaustufe der Maßnahmendatenbank wurde als Web-Applikation entworfen und umgesetzt, die von allen verantwortlichen Behörden und Organisationen über das Internet genutzt werden kann. Eine Zugriffsverwaltung mit definierten Rechten und Rollen garantiert die Datensicherheit. Bereits existierende Daten über an den Wasserkörpern geplante Maßnahmen werden interaktiv importiert und von den Projektmitgliedern vervollständigt.

Die Maßnahmendatenbank kombiniert alle notwendigen Informationen und Beziehungen, um die verschiedenen Auswertungs- und Berichtsanforderungen der regionalen, nationalen und EU-bezogenen Berichtspflichten zu erfüllen.

In einer geplanten zweiten Ausbaustufe wird die Maßnahmendatenbank-Anwendung zu einem universellen Instrument erweitert, das Informationen und Prozesse zur Planung, Dokumentation und Finanzierung verwaltet, um alle Aspekte des Maßnahmenmanagements von der Maßnahmenidee über die Kosten bis zu Überprüfung und Monitoring zu unterstützen.

## **1 Einführung**

Mit dem Inkrafttreten der EG-Wasserrahmenrichtlinie (WRRL, Directive 2000/60/EC, 2000) entstanden neue Anforderungen an die Gewässerüberwachung und Bewirtschaftung (Kallis & Butler 2001, Kaika 2003). Speziell die Berichterstattung für die Europäische Union (EU) verlangt die Transparenz der Zustandsbewertung und der Maßnahmenplanung, um bestehende Defizite an Gewässern zu vermindern (DWA 2008, Spitzer 2008). In den Bewirtschaftungsplänen werden die Zielsetzungen bezüglich der Wasserqualität aufgelistet. Im Maßnahmenprogramm werden die zu treffenden Vorkehrungen zur Erreichung dieser Ziele zusammengestellt. Zur Erstellung der geforderten Bewirtschaftungspläne und Maßnahmenprogramme müssen alle in den Flussgebietseinheiten geplanten Maßnahmen zentral zusammengestellt und ausgewertet werden.

Für das Land Schleswig-Holstein beauftragten das Ministerium für Landwirtschaft, Umwelt und ländliche Räume (MLUR) und das Landesamt für Natur und Umwelt (LANU) die Unternehmen KISTERS AG und DigSyLand mit der Entwicklung einer zentralen Datenbankapplikation zur Verwaltung der WRRL-Maßnahmen.

Folgende Aufgaben sollen durch die Maßnahmendatenbank erfüllt werden:

- Zusammenstellung aller existierenden Informationen über die Wasserkörper (als kleinster relevanter Raumeinheit) und die an den Wasserkörpern geplanten Maßnahmen.
- Verknüpfung der Maßnahmen mit lokalen, regionalen und überregionalen Zielen sowie mit geschätzten Kosten und Auswirkungen.
- Auswertungen auf unterschiedlichen räumlichen Ebenen, wie z.B. Wasserkörper, Planungsräume und das gesamte Land Schleswig-Holstein zur Unterstützung verschiedener Steuerungs-, Dokumentations- und Berichtsanforderungen.

## **2 Voraussetzungen und Rahmenbedingungen**

Für die Umsetzung der ersten Ausbaustufe stand nur ein kurzer Zeitraum zur Verfügung. Nach dem Projektstart im Juni 2007 war die Dateneingabe in einer ersten Produktivversion für August 2007 geplant. Langfristig soll die Maßnahmendatenbank weitere umfassende Aufgaben des WRRL-Maßnahmenmanagements in Schleswig-Holstein wahrnehmen. Daher wurde eine zweistufige Vorgehensweise beschlossen:

- Eine erste Ausbaustufe sollte in einem kurzen Zeitraum entwickelt werden mit dem Ziel, 2007 in den Produktivbetrieb gehen zu können, um alle Berichtsanforderungen im Frühjahr 2008 erfüllen zu können.
- Die zweite Ausbaustufe ist als Modul der KISTERS-Softwareumgebung K3-Umwelt geplant. Für die Konzeption und Umsetzung dieser zweiten Ausbaustufe steht mehr Zeit zur Verfügung. Dafür soll diese Version die umfassenden Anforderungen als universelle Lösung zur Maßnahmenverwaltung für die WRRL in Schleswig-Holstein erfüllen.

Weitere wichtige Anforderungen bestanden in der Integration aller relevanten, bereits existierenden Daten über die geplanten Maßnahmen und den Zustand sowie andere Eigenschaften der Wasserkörper.

Bereits vor der Entwicklung der Maßnahmendatenbank waren die an den Fließgewässern konzipierten Maßnahmen in einem ausführlichen Diskussionsprozess durch etwa 30 Arbeitsgruppen zusammengestellt und unter anderem in GIS-Karten abgebildet worden. Diese dezentral verwalteten Informationen sollten in die Maßnahmendatenbank übernommen werden. Zusätzliche Informationen wurden in unterschiedlichen Datenquellen gehalten und sollten ebenfalls mit den Wasserkörpern verknüpft und in die MDB integriert werden.

Gewünscht wurde eine einfach zu bedienende Applikation mit einer schlichten Navigation, um Aufwand und Zeit bis zum produktiven Arbeiten mit der Anwendung zu minimieren. Die Datenbank-Anwendung sollte sicherstellen, dass alle Nutzenden nur die Daten lesen bzw. bearbeiten können, die in den eigenen Zuständigkeitsbereich fallen, und dass keine unberechtigten Zugriffe auf die verwalteten Daten möglich sind.

## **3 Konzeption**

### **3.1 Softwareumgebung**

Aufgrund der beschränkten Entwicklungszeit wurde entschieden, die erste Ausbaustufe als Web-Applikation zu entwickeln, so dass auf Module ähnlicher Anwendungen aufgesetzt werden konnte, die DigSyLand bereits für das Umweltressort des Landes umgesetzt hatte (Rammert & Hosenfeld 2003). Die dadurch ermöglichte breite Verfügbarkeit war ein weiterer Grund zum Einsatz von Web-Technologien.

Zunächst waren nur Landesbehörden als Nutzer der Maßnahmendatenbank vorgesehen, die innerhalb des Landesnetzes Zugriff auf das Intranet des Landesamtes LANU besitzen. Jedoch sollte prinzipiell auch externen Organisationen ein Zugang zu der Applikation gewährt werden, damit diese die Umweltbehörden bedarfsgerecht bei der Maßnahmenplanung unterstützen. So sah das

Entwicklungskonzept vor, die Applikation zunächst im LANU-Intranet einzusetzen und bei Bedarf später auf einen im Internet zugänglichen Server zu migrieren.

### **3.2 Sicherheitsmanagement**

Das Entwicklungskonzept enthält zahlreiche Aspekte zur Gewährleistung der Sicherheit der Datenbankinhalte. Die Rolle der Standard-Projektmitglieder erlaubt nur das Bearbeiten der Wasserkörper aus dem eigenen Verantwortungsbereich. Daten anderer Wasserkörper können gelesen aber nicht verändert werden. Die im LANU angesiedelte Administration kann unter anderem neue Benutzerkennungen einrichten sowie Rechte und Rollen zuordnen.

Während der Bearbeitung wird der aktuelle Wasserkörper für Schreibzugriffe gesperrt, kann aber von anderen Benutzern gelesen werden.

### **3.3 Grundfunktionalität**

Die erste Hauptaufgabe der MDB bestand in dem Transfer der Wasserkörper-Daten, die bereits in vorangegangenen Planungs- und Diskussionsphasen zu den beabsichtigten Maßnahmen erzeugt worden waren.

Im Fall der über 600 Fließgewässer-Wasserkörper wurden alle Maßnahmen lokal mit Hilfe des Geographischen Informationssystems (GIS) ESRI ArcView geplant. Für jedes der über 30 Bearbeitungsgebiete wurden die Maßnahmen dezentral in ArcView-Shape-Dateien verwaltet, so dass eine Upload-Möglichkeit dieser Informationen in die zentrale Maßnahmendatenbank realisiert werden musste.

Das Konzept sah das Hochladen der GIS-Dateien mit einer Überführung der Maßnahmen-bezogenen Sachdaten in die Datenbank vor. Die GIS-Dateien werden auf dem Server gespeichert und über die Web-Applikation unter Beachtung der Datenzugriffsrechte zum Download verfügbar gemacht.

Zusätzlich können weitere Dateien, die Karten oder Dokumentationen enthalten, hochgeladen und in der Anwendung verfügbar gemacht werden.

Aufgrund des engen Zeitrahmens konnte keine GIS-Funktionalität in die Anwendung aufgenommen werden. Alle Informationen werden in dem relationalen Datenbankmanagementsystem verwaltet, ohne die räumlichen Bezüge abzubilden. In der zweiten Ausbaustufe der Maßnahmendatenbank werden GIS-Verknüpfungen berücksichtigt, um GIS-basierte Auswertungen zu realisieren.

### **3.4 Datenmodell**

Das Datenmodell gliedert sich in drei Bereiche:

- Systemmanagement und Steuerungsinformationen,
- Wasserkörper-bezogene Informationen und
- Maßnahmen-bezogene Informationen.

Soweit wie möglich werden die Daten zu den verschiedenen Gewässerkategorien Fließgewässer, Seen, Küstengewässer und Grundwasser in einheitlichen Strukturen verwaltet. Der Bezug zum Wasserkörper als kleinster Raumeinheit erlaubt datenbankbasierte Aggregationen und Auswertungen auf verschiedenen räumlichen Ebenen.

Gemeinsam vom Ministerium MLUR und dem Landesamt LANU wurde ein standardisierter, hierarchisch organisierter Maßnahmenkatalog entwickelt, der

einerseits umfassende Auswertungen und Analysen ermöglicht, andererseits aber auch die unterschiedlichen Anforderungen seitens nationaler Gremien und der EU flexibel unterstützt. Alle in der MDB verwalteten Maßnahmen werden einer Maßnahmengruppe des Katalogs zugeordnet. Die Maßnahmengruppen sind mit technischen und fachlichen Eigenschaften sowie Bezügen zu weiteren Klassifikationen versehen, wie etwa

- Gewässerkategorien,
- räumliche Dimensionen (Punkt, Linie, Fläche),
- signifikante Belastungen,
- Referenzen für Berichtsschablonen der Maßnahmenprogramme und
- Klassifikationen für WISE (Water Information System for Europe).

Durch die Zuweisung bestimmter Werte für Belastungen und Wirkungen (z.B. Nährstoff-Reduzierung) zu den im Maßnahmenkatalog aufgeführten Maßnahmengruppen können wiederum die Auswirkungen der geplanten Maßnahmen für jeden Wasserkörper andere Raumeinheiten abgeschätzt werden.

Zur Anpassung von Berichten an zusätzliche Klassifikationsschemata müssen diese im allgemeinen nur auf den Maßnahmenkatalog abgebildet werden, um die entsprechenden Berichtsvorgaben für alle Maßnahmen in den Wasserkörpern zu erfüllen. Mit dem Maßnahmenkatalog wird eine Standardisierung der Maßnahmeninformationen erreicht, die ein wichtiges Fundament der Maßnahmendatenbank darstellt.

Das Europäische Wasserinformationssystem WISE wird zur Zeit aufgebaut, um die in der WRRL geforderten Berichte und Daten der Mitgliedsstaaten und den jeweils aktuellen Stand der Umsetzung zu verwalten und zugänglich zu machen.

## **4 Implementierung**

Die MDB-Anwendung wurde mit PHP 4 auf der Basis des Oracle 9 Datenbankmanagementsystems realisiert.

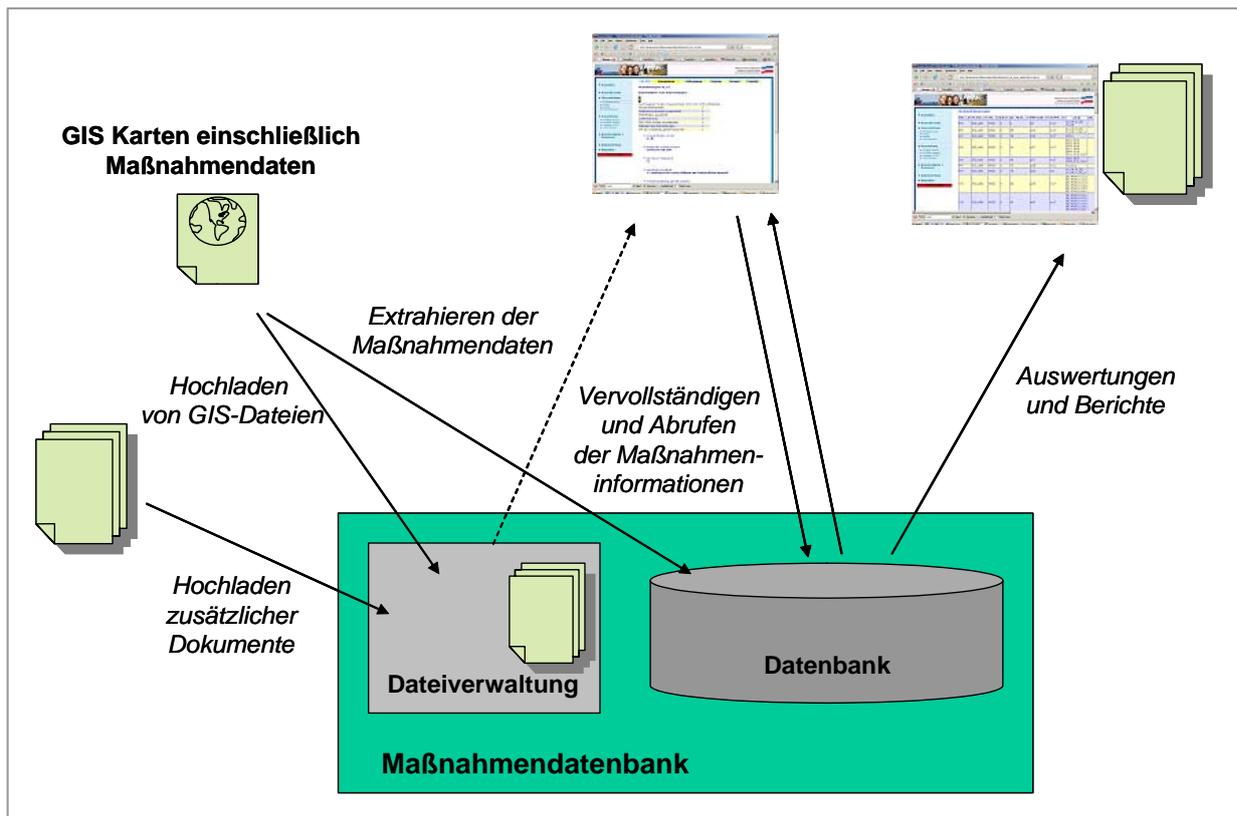


Abbildung 1: Datenfluss und Organisation der Maßnahmendatenbank.

#### 4.1 Hochladen von Daten

Die zunächst vordringlich zu erfüllende Aufgabe der MDB bestand in dem Hochladen, Analysieren und dem Datentransfer der ArcView-Shapedateien, die die Maßnahmen der Fließgewässer enthalten (s. Abb. 1). Während des Hochlade-Prozesses unterstützt die Anwendung die Nutzenden durch inhaltliche (z.B. Schlüssellistenvergleiche) und strukturelle (z.B. Formate, Dateiname und -typ, ) Analysen, um die korrekte Herstellung der Bezüge zu den Wasserkörpern zu erleichtern.

Mithilfe eines Versionierungsmechanismus wird sichergestellt, dass unterschiedliche Versionen von Geodaten bewahrt werden. In der Datenbank kann jederzeit nachvollzogen werden, welche Daten für welchen Wasserkörper gültig sind. Auf dem Server werden die GIS-Dateien in einer hierarchischen Struktur abgelegt, die die Gliederung in Bearbeitungsgebiete und deren Wasserkörper widerspiegelt. Je nach Art der Geometrie (Punkt, Linie, Fläche) von Maßnahmen werden diese in unterschiedlichen GIS-Dateien verwaltet. Eine GIS-Datei kann wiederum Maßnahmendaten zu verschiedenen Wasserkörpern enthalten. Der Versionierungsmechanismus muss berücksichtigen, dass zu einzelnen Wasserkörpern korrigierte Dateien nachträglich hochgeladen werden können. Die Datenbank verwaltet für jeden Wasserkörper die Angaben darüber, in welchen Dateien die aktuellen Maßnahmenkarten vorliegen.

Aus der Anwendung können gemäß den entsprechenden Zugriffsrechten alle hochgeladenen Dateien abgerufen werden.

Neben den GIS-Dateien werden weitere Dokumente wie etwa Kartendarstellungen im PDF oder PNG-Format von der Web-Anwendung verwaltet und zugänglich gemacht.

## 4.2 Zusätzliche Wasserkörperinformationen

Zur Unterstützung der Bewertung und Bearbeitung der Maßnahmen bietet die Maßnahmendatenbank zusätzliche Informationen über die Wasserkörper an. Dies umfasst unter anderem Daten über die Umweltziele, ökologische Qualitätskomponenten, Natura 2000-Gebiete und Badegewässer.

Nr.	WK-Kürzel	Name des Wasserkörpers	Typ gem. Codelist	Einst.	Zielst.	BGNr	Bearb.-Geb.	Vollst.	Anz. MN	MN zugeordnet	Pflichtf.	Zieler.
1	tr_01	Zulauf Tressee	16	2: erheblich verändert	2: gutes ökologisches Potenzial	6	Treene	🚦	15	X	X	o
2	tr_02	Bondenau OL	16	2: erheblich verändert	2: gutes ökologisches Potenzial	6	Treene	🚦	16	X	o	o
3	tr_03	Bondenau	19	2: erheblich verändert	2: gutes ökologisches Potenzial	6	Treene	🚦	5	X	X	o
4	tr_04	Mühlenstrom	16	2: erheblich verändert	2: gutes ökologisches Potenzial	6	Treene	🚦	8	X	o	o
5	tr_06	Kielstau/Bondenau	16	1: natürlich	1: guter ökologischer Zustand	6	Treene	🚦	43	X	X	X
6	tr_07	Zulauf Sankelmarker See	14	2: erheblich verändert	2: gutes ökologisches Potenzial	6	Treene	🚦	4	X	X	o
7	tr_08_a	Treene OL	14	2: erheblich verändert	2: gutes ökologisches Potenzial	6	Treene	🚦	6	o	o	o
8	tr_08_b	Treene	14	1: natürlich	1: guter ökologischer Zustand	6	Treene	🚦	11	X	o	X
9	tr_08_c	Bek	14	2: erheblich verändert	2: gutes ökologisches Potenzial	6	Treene	🚦	29	X	X	X
10	tr_08_d	Büschau	14	2: erheblich verändert	2: gutes ökologisches Potenzial	6	Treene	🚦	13	X	o	o
11	tr_08_e	Jernsbek	14	2: erheblich verändert	2: gutes ökologisches Potenzial	6	Treene	🚦	29	o	o	o

Abbildung 2: Bearbeitungsstatus für Fließgewässer (Auszug).

## 4.3 Workflow

Um die benötigten Maßnahmeninformationen zur Erstellung der Maßnahmenprogramme und zur Erfüllung der Berichtspflichten zu erfassen, müssen für jeden Fließgewässer-Wasserkörper folgende Arbeitsschritte durchgeführt werden (s. Abb. 1):

- Hochladen der GIS-Dateien zum Transfer der Maßnahmendaten in die Datenbank.
- Hochladen weiterer Dokumente zu den Maßnahmen wie Karten als PDF/PNG und automatisiert erstellte PDF-Dateien (sogenannte digitale Beurteilungsbögen).
- Vervollständigung der Maßnahmeninformationen um weitere Angaben:
  - Kostenschätzung,
  - Zuordnung zu dem Maßnahmenkatalog (wird automatisiert unterstützt) und
  - Einschätzung der Zielerreichung.

Wenn alle erforderlichen Angaben zu einem Wasserkörper vorhanden sind, kann dieser als „vollständig bearbeitet“ markiert werden. Diese Kennzeichnung wird von der MDB überprüft und ist nur möglich, wenn alle Pflichtinformationen vorhanden sind. Im anderen Fall wird auf fehlende Angaben hingewiesen.

Für eine übersichtliche Visualisierung des aktuellen Bearbeitungsstatus wird jeder Wasserkörper durch ein Ampel-Symbol gekennzeichnet: Grün steht für vollständig bearbeitete Wasserkörper, während gelb und rot anzeigen, dass noch Angaben fehlen (s. Abb. 2).

Bei Seen, Küstengewässern und Grundwasserkörpern ist die Vorgehensweise weniger komplex. Maßnahmen an den Seen wurden aus einer zentralgehaltenen Microsoft Access-Datenbank übernommen. Die Maßnahmen der übrigen Gewässerkategorien werden interaktiv in die Maßnahmendatenbank eingegeben. Alle erforderlichen Angaben zu einer Maßnahme sowie die Zuordnung zu einer Maßnahmengruppe im Maßnahmenkatalog erfolgt über die Web-Formulare, mit denen auch eine Bearbeitung bereits erfasster Maßnahmen durchgeführt wird (s. Kap. 4.4). Aufgrund der verschiedenen Verfahren zur Planung und Bewertung der Maßnahmen sind je nach Gewässerkategorie unterschiedliche Pflichtangaben vorzunehmen.

#### **4.4 Bearbeitung der Maßnahmendaten**

Alle Eigenschaften der Maßnahmen können in der MDB nachträglich bearbeitet werden. Räumliche Angaben wie Länge und Fläche werden beim GIS-Import automatisch übernommen, können aber nachträglich modifiziert werden. Neue Maßnahmen können eingegeben, ebenso können Maßnahmen geteilt und gelöscht werden.

Einige Angaben, wie beispielsweise die Einschätzung der Zielerreichung mit Hilfe der geplanten Maßnahmen, beziehen sich auf die Gesamtheit aller an einem Wasserkörper geplanten Maßnahmen und nicht auf einzelne Maßnahmen.

#### **4.5 Auswertungen und Berichte**

Die Erstellung von Berichten für die EU ist ein wichtiges Ziel der MDB. Für die Erzeugung der Flussgebietsbewirtschaftungspläne und der Maßnahmenprogramme werden Listen aller umzusetzenden Maßnahmen benötigt.

Mithilfe der MDB können Auflistungen nach unterschiedlichen Kriterien mit räumlichen Aggregationen auf verschiedenen Ebenen generiert werden. Der zugrunde liegende Maßnahmenkatalog ist mit allen relevanten Klassifizierungsschemata verknüpft, so dass die MDB Berichte sowohl für die EU als auch für nationale und regionale Bedürfnisse erstellen kann.

Insbesondere liefert die MDB Einzelheiten zur Abschätzung der Wirkungen (z.B. Nährstoffreduktion) gemäß der Ziele und zur Kosteneffizienz der Maßnahmen, um so die Planung auf Landesebene zu ermöglichen: Fließgewässer- und See-Wasserkörper wurden mit Prioritätsfaktoren versehen. Für jeden Wasserkörper wird ein Kosteneffizienzwert auf der Basis des Prioritätsfaktors, der Kostensumme der geplanten Maßnahmen und der Länge bzw. Fläche des Wasserkörpers berechnet (s. Abb. 3). Gemäß dieses Effizienzwertes wird eine Rangfolge der Gewässer generiert, die die Entscheidungen unterstützt, welche Maßnahmen im Rahmen eines vorgegebenen Kostenbudgets umgesetzt werden sollten. Durch manuelle Veränderung der Effizienzwerte kann die berechnete Rangfolge beeinflusst werden.

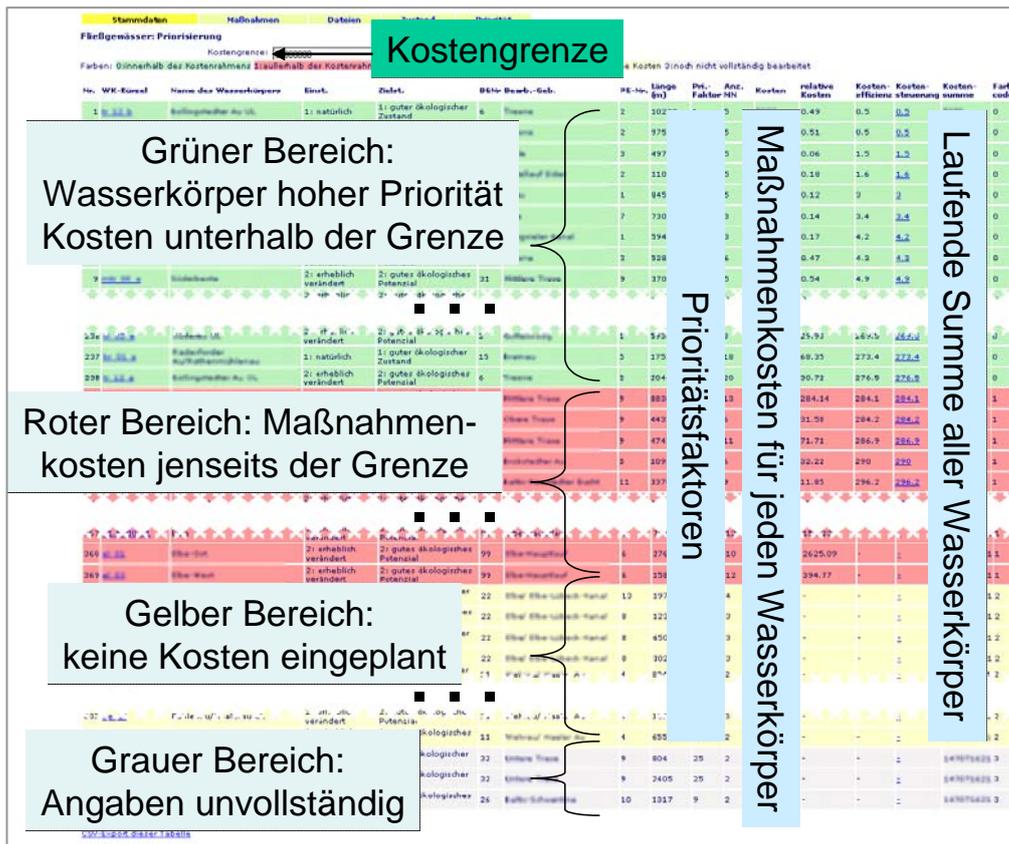


Abbildung 3: Beispiel für die Auswertung von Kosten und Prioritäten der Wasserkörper mit den geplanten Maßnahmen: Die Sortierung der Fließgewässer-Wasserkörper erfolgt nach ihrem Kosteneffizienzwert basierend auf den Kosten der Maßnahmen, der Wasserkörperlänge und einem Prioritätsfaktor.

## 5 Zusammenfassung und Ausblick

Bisher konnten die mit der Maßnahmendatenbank (MDB) angestrebten Ziele erreicht werden. Die erste Ausbaustufe ging nach kurzer Entwicklungszeit in den Produktivbetrieb, so dass die erforderlichen Maßnahmen-Informationen zur Erfüllung der Berichtspflichten jetzt zur Verfügung stehen.

Gewässerkategorie	Anzahl Wasserkörper	Anzahl Maßnahmen
Flüsse / Übergangsgewässer	607	2030
Seen	73	227
Küstengewässer	40	70
Grundwasser	64	404

Tabelle 1: Zahl der in der Maßnahmendatenbank verwalteten Maßnahmen (ungefähre Angaben, Mai 2008).

Als Web-Applikation implementiert bietet die Maßnahmendatenbank allen beteiligten Behörden und Organisationen den Zugang zum WRRL-Maßnahmenplanungsprozess, so dass sich die Entscheidung zum Einsatz von Web-Technologien als richtig erwiesen hat. Mehr als 2700 Maßnahmen werden zur Zeit in der Maßnahmendatenbank verwaltet (s. Tabelle 1).

Als negative Folge des zeitlich eng beschränkten Entwicklungsprozesses musste auf die direkte Integration der räumlichen Informationen aus vorhergehenden Maßnahmenplanungsschritten verzichtet werden. Diese Geodaten müssen in der zweiten Ausbaustufe ergänzt und mit dem dann aktuellen Maßnahmenbestand verknüpft werden.

Parallel zur Inbetriebnahme der ersten Ausbaustufe erfolgte eine Ist-Analyse der Anforderungen zur Maßnahmenbearbeitung als Voraussetzung für die Konzeption der zweiten Ausbaustufe, die alle relevanten Phasen der Maßnahmenplanung und –verwaltung unterstützen soll. Beabsichtigt ist die Umsetzung der zweiten Ausbaustufe als Modul in K3-Umwelt, da diese Software-Umgebung in vielen Kreisen Schleswig-Holsteins eingesetzt wird und die Kreise in Zukunft eine zunehmend stärkere Rolle im Prozess der WRRL-Maßnahmenplanung übernehmen werden. Die zweite Ausbaustufe der Maßnahmendatenbank wird eine enge Kopplung mit dem GIS realisieren, so dass auch komplexere räumliche Auswertungen möglich sein werden und die exakte Verortung der Maßnahmen in der Anwendung zugänglich sein wird.

Aus administrativer Sicht bedeutet die Entwicklung der Web-basierten Datenbank mit Wasserkörper-Informationen einen wichtigen Fortschritt in der Umsetzung der Wasserrahmenrichtlinie in Schleswig-Holstein. Die zentrale Datenbank ermöglicht zum ersten Mal den Zugriff auf Daten zu Wasserkörper-Zustand und den geplanten Maßnahmen für alle beteiligten Partner in jeder Phase des Planungsprozesses.

Auf diese Weise verbessert das vorgestellte Werkzeug die Transparenz der Berichtsinhalte sowie die Entscheidungen der Partner und minimiert durch mangelnde Daten verursachte Missverständnisse und Diskussionsbedarf. Zudem verwaltet die Anwendung geeignete Abfrageroutinen zur Erfüllung der nationalen Berichtspflichten und verringert damit signifikant Zeit und Aufwand für das Datenmanagement.

## 6 Literatur

- Directive 2000/60/EC (2000): Directive 2000/60/EC of the European parliament and of the council of 23 October 2000 establishing a framework for Community action in the field of water policy (EU Water Framework Directive), in: Official Journal of the European Communities, pp. 327/1-327/72.
- DWA (2008): Der Prozess der Bewirtschaftungsplanung gemäß Europäischer Wasserrahmenrichtlinie – Handlungsempfehlungen aus Sicht der DWA. Korrespondenz Wasserwirtschaft 1(4): pp. 212-214.
- Kallis, G. & Butler D. (2001): The EU water framework directive: measures and implications. Water Policy 3: pp. 125-142.
- Kaika, M. (2003): The Water Framework Directive: A New Directive for a Changing Social, Political and Economic European Framework. European Planning Studies 11: pp. 299 – 316.
- Rammert, U. & Hosenfeld, F. (2003): Dynamic and Interactive Presentation of Environmental Information. In: Gnauck, A. & Heinrich, R. (eds.): The Information Society and Enlargement of the European Union, 17th International Symposium Informatics for Environmental Protection, Cottbus 2003, pp. 517-524.
- Spitzer A. (2008): Öffentlichkeitsbeteiligung im Sinn der Europäischen Wasserrahmenrichtlinie – eine Annäherung. Korrespondenz Wasserwirtschaft 1(5): pp. 243-247.
- Willems, P.& de Lange, W.J. (2007): Concept of technical support to science–policy interfacing with respect to the implementation of the European water framework directive. Environmental Science & Policy 10: pp. 464-473.



# ETOX: Informationssystem Ökotoxikologie und Umweltqualitätsziele - Ein Werkzeug zur Entwicklung von Umweltqualitätsanforderungen

Dieter Schudoma, Umweltbundsamt, [dieter.schudoma@uba.de](mailto:dieter.schudoma@uba.de)

## Abstract

The information system ETOX <http://webetox.uba.de/webETOX> is used as a tool in the process developing science based reference values (quality criteria) for the protection of aquatic and terrestrial ecosystems as well as an instrument for the assessment of substances. ETOX enables effective documentation of evaluated effects data and environmental quality requirements for the media water, soil, sediment and biota. A quick retrieval of this information provides a basis for the assessment of measured data within the scope of environmental monitoring and the further development of quality criteria. The documentation of quality requirements for these media is a fundamental basis for the development of environmental quality criteria across media and objects of protection, which establish a basis for the determination (definition) of mandatory environmental quality standards. ETOX is mainly used in the Federal Environment Agency (Umweltbundesamt - UBA) for the development of environmental quality standards for the implementation of the European Water Framework Directive 60/2000/EC. The exchange of information about effect data and proposals for environmental quality criteria on a European level is important for this process. ETOX offers project orientated data input and is open for a participation of other institutions. ETOX can be used as a site for the exchange of information about effect data and quality requirements on national and international levels. Further, a data exchange with the GSBL <http://www.gsbl.de> exists via an interface formate.

## Einleitung

Das Informationssystem ETOX <http://webetox.uba.de/webETOX> dient als Grundlage zur Entwicklung von wissenschaftlich begründeten Referenzwerten zum Schutz der aquatischen und terrestrischen Lebensgemeinschaften sowie zur Bewertung von Stoffen. ETOX ermöglicht neben validierten Wirkungsdaten, Umweltqualitätsanforderungen für Stoffe und Wirkparameter in den Medien Wasser, Boden, Sedimente und Biota in einer übersichtlichen Form zu dokumentieren. Die schnelle Verfügbarkeit dieser Daten ist eine wichtige Grundlage für die Bewertung von Umweltdaten im Rahmen der Umweltbeobachtung und bei der Weiterentwicklung von Qualitätskriterien. Die Dokumentation von medialen Qualitätsanforderungen ist eine wesentliche Grundlage für die Entwicklung von medien- und schutzgutübergreifenden Umweltqualitätskriterien, die die Basis zur Festlegung von Umweltqualitätsstandards bilden.

ETOX wird im Umweltbundesamt (UBA) vor allem für die Entwicklung von Umweltqualitätsnormen zur Umsetzung der EU Wasserrahmenrichtlinie 60/2000/EG

genutzt. Hierbei ist der Informationsaustausch auf europäischer Ebene von Wirkungsdaten und Vorschlägen für Umweltqualitätskriterien von besonderem Interesse. ETOX ermöglicht eine projektorientierte Datenerfassung und ist für die Beteiligung von Projekten anderer Institutionen offen. ETOX kann als Austauschform für Wirkungsdaten und Qualitätsanforderungen in nationalen und internationalen Bereich genutzt werden. Weiterhin besteht durch verfügbarer Schnittstellenformate ein Datenaustausch mit dem Gemeinsamen Stoffdatenpool Bund/ Länder (GSBL) <http://www.gsbl.de> .

## **1 Ziele der Entwicklung**

Das Informationssystem ETOX ist aus mehreren Forschungsprojekten entstanden. Ziel der Entwicklung war es, Wirkungsdaten, Qualitätsanforderungen für Stoffe zu den Medien Wasser, Boden und Biota (Rückstandswerte) in einem System zusammenzufassen, um zum Beispiel den Prozess der Entwicklung von Umweltqualitätsnormen für die Umsetzung der Wasserrahmenrichtlinie (WRRL) zu optimieren. Das Informationssystem hat folgende Hauptmerkmale:

- Online Datenrecherche und -Eingabe über das Internet
- Zweisprachige Webseite (Deutsch – Englisch)
- Dateneingabe organisiert über Projekte
- Schnittstellen zum Daten-Export/Import mit dem GSBL (Gemeinsamer Stoffdatenpool Bund / Länder)
- Datenaustausch zwischen Umweltbehörden und Forschungsinstituten
- Information der Fachöffentlichkeit

## **2 Systemarchitektur**

ETOX besteht im Kern aus einer SQL-Serverdatenbank (Tabellen und Views). Die einzelnen Tabellen sind mit referentieller Integrität verbunden, um Inkonsistenzen zu verhindern. Der Zugriff der Web-Clients auf die Datenbank erfolgt über Active Server Pages, die auf einen Web-Server im UBA (Tomcat Applikation) gehostet sind (Abbildung 1).

# Informationssystem ETOX

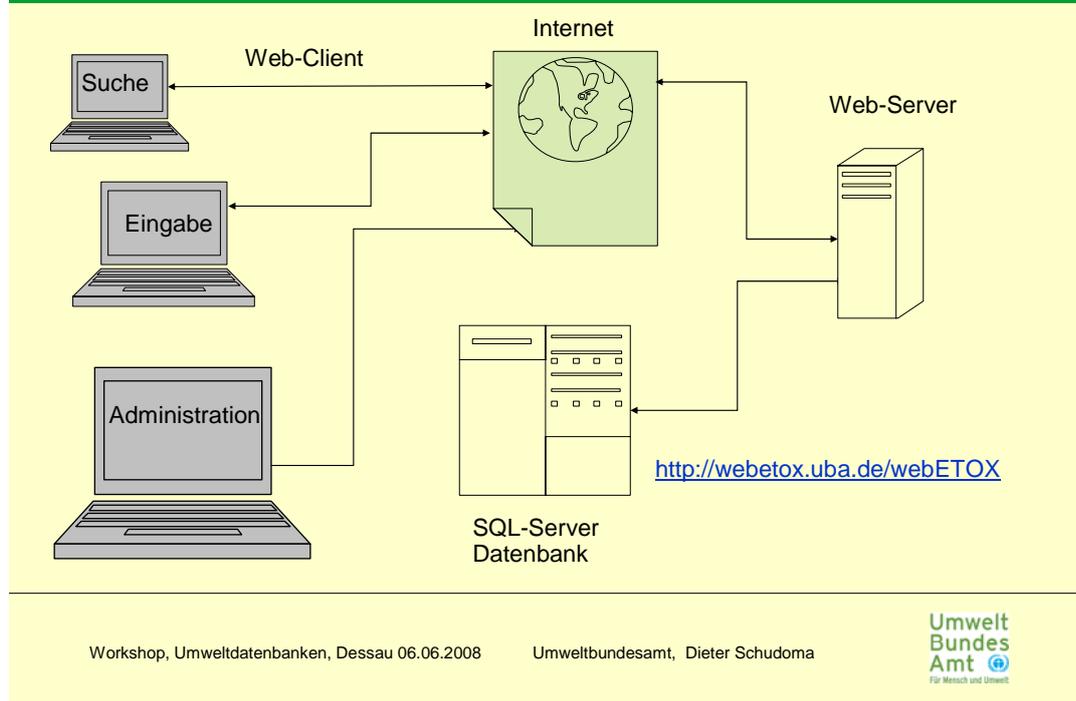


Abbildung 1 Überblick zur Systemarchitektur

### 3 Recherche und Datenbestand

Die Suche von freigegebenen Wirkungsdaten und Qualitätszielen in ETOX erfordert keine Anmeldung (Abbildung 2). Über die Menüpunkte „Organismus“, „Stoff“ und „Literatur“ ist eine Suche der verfügbaren Einträge des Organismen-, Stoff- und Literaturkatalogs möglich. Weiterhin sind alle verfügbaren Einträge der Auswahlboxen der Such- und Eingabemasken unter Stammdaten einsehbar. Im Handbuch, das unter der Rubrik „downloads“ abgerufen werden kann, sind die einzelnen Suchfunktionen und Eingabekonventionen ausführlich beschrieben. Unter „Links“ sind ausgewählte Publikationen zum Thema Entwicklung von Qualitätsanforderungen für Oberflächengewässer gelistet. Unter „weitere Seiten“ sind beispielhaft Links zu Seiten zusammengestellt, die Informationen im Bereich Stoffbewertung oder die zur Recherche von ökotoxikologischen Daten z.B. des Informationssystems US EPA ECOTOX und Umweltqualitätszielen dienen.

Für registrierte und angemeldete Nutzer bietet die Web-Applikation eine online Erfassung von Wirkungsdaten und Qualitätszielen. Der Zugang und die Datenfreigabe werden über ein Rechte- und Nutzerkonzept geregelt.

Abbildung 2 Startseite des Informationssystems ETOX

## 4 Ausblick

Es ist geplant den Datenbestand von ETOX in den nächsten Jahren durch folgende Projekte und Arbeiten ergänzt werden:

- Wasserqualitätskriterien für weitere Pflanzenschutzmittel, Biozide, Arzneimittel und andere prioritäre Stoffe (LAWA-Projekt 2008/2009)
- Aufnahme von ausgewählten Ergebnissen aus Prüfungen des UBA Fachgebiets IV 2.4 Wassergefährdende Stoffe - Ökotoxikologielabor
- Up date der GSBL-Daten für das Merkmal Ökotoxikologie
- Up date der Canadian Environmental Quality Guidelines
- Aufnahme von Predicted No Effect Concentrations (PNEC) aus EU Risk Assessment Reports
- Kooperation mit Umweltbehörden in Europa (F, NL, UK) bei der Erfassung nationaler Wasserqualitätsanforderungen
- Integration einer internen Accessdatenbank der Bundesanstalt für Gewässerkunde („BfG-Ökotoxdatenbank“)

Weitere Institutionen oder Forschungseinrichtungen, die sich am Informationssystem ETOX als Projektpartner beteiligen möchten, sind willkommen.