

Martin Brändli und Marcel Frehner

# Die Virtuelle Datenbank – Ein Werkzeug für die integrierte Datenverarbeitung räumlich verteilter Datenbestände

Verteilt vorliegende, räumliche Datenbanken und Datenbestände erfordern spezifische Techniken und Vorgehensweisen um einen effektiven Umgang mit den Daten zu ermöglichen. Die Integration, Visualisierung und Analyse der Daten spielt dabei eine wesentliche Rolle.

## 1 Einleitung

Die effektive Handhabung von Daten, welche zu einer erfolgreichen Modellierung, Analyse und Herleitung adäquater Informationen führen soll, verlangt ein umfassendes Datenmanagement. Brunt (2000) präsentiert eine idealisierte Sicht des Managements ökologischer Daten als einen Prozess, der sich durch verschiedene, nacheinander ablaufende Arbeitsschritte auszeichnet (siehe Abbildung 1). Die einzelnen Schritte dieses Prozesses bilden untereinander verkoppelte Komponenten eines umfassenden Systems für das Management von Daten aus den Bereichen Ökologie und Umwelt.

Innerhalb der Umwelt- und Ökologieforschung besitzt dieses System des Datenmanagements nach wie vor weite Verbreitung. Dabei wird häufig nur eine kleine Zahl von Versuchsflächen mit Datenerhebungen, die nur kurze Zeitspannen abdecken, untersucht (Michener 2000). Neue Herausforderungen, der Fortschritt der Informationstechnologie sowie andere, sich stark ändernde Bedingungen, verlangen nach einer substantiellen Anpassung dieses traditionellen Schemas des Datenmanagements:

- *Langfristiges Monitoring:* Ökologische Phänomene werden in zunehmenden Mass in Beobachtungsprogrammen untersucht, welche auf lange Zeiträume ausgerichtet sind.
- *Zunehmende Datenmenge:* Der Umfang ökologischer Daten steigt stetig an. Daten, die in grösseren Projekten anfallen, werden nicht mehr in einer einzigen Datenbank gehalten, sondern als verteilte Datenbestände verwaltet. Oftmals finden Aufteilungen so statt, dass unterschiedliche Typen von Beobachtungen in getrennten Datenbanken gespeichert werden. Dabei muss die Möglichkeit zur

- *Ausdehnung des Beobachtungsraumes:* Der Massstab für die Untersuchung räumlicher Phänomene verschiebt sich tendenziell von einer lokalen hin zu einer globalen Betrachtungsweise, da verschiedene umweltbedingte Herausforderungen wie *Global Change*, Nachhaltigkeit und Biodiversität grossmassstäblich angegangen werden müssen.

Als wichtigste Konsequenz aus diesen Veränderungen und neuen Anforderungen schliessen wir, dass dem Management ökologischer Daten und im Besonderen der Datenintegration und der gemeinsamen Datennutzung (*Sharing*) heute und in nächster Zukunft herausragende Bedeutung zugemessen werden muss. Die Datenintegration beinhaltet Techniken und Methoden, die die kombinierte Nutzung von Datenbeständen aus verschiedensten Quellen ermöglichen (Vckovski et al. 1999, Lutz et al. 2003). Für die gemeinsame Nutzung von Daten müssen diese für alle in einem Projekt involvierten Partner und andere interessierte Kreise in einer einheitlichen und standardisierten Form verfügbar gemacht werden.

Dieser Beitrag zeigt mit der *Virtuellen Datenbank* eine Architektur auf, die die oben genannten Herausforderungen für den Managementprozess ökologischer Daten annimmt und es ermöglicht, Datenbestände aus unterschiedlichsten Quellen über Internet mit einem einheitlichen Zugriff zu nutzen. Ziel der Virtuellen Datenbank ist es, verteilt vorliegende Umweltdatenbanken so zu integrieren, dass *transparent* auf Datenbestände zugegriffen werden kann, um zielgerichtete Analysen durchführen zu können. Unter dem Konzept der Transparenz wird verstanden, dass die Herkunft der einzelnen Datenbestände in einer verteilten Umgebung verborgen bleibt bzw. das System als eine Einheit und nicht als Sammlung einzelner Bestandteile wahrgenommen wird

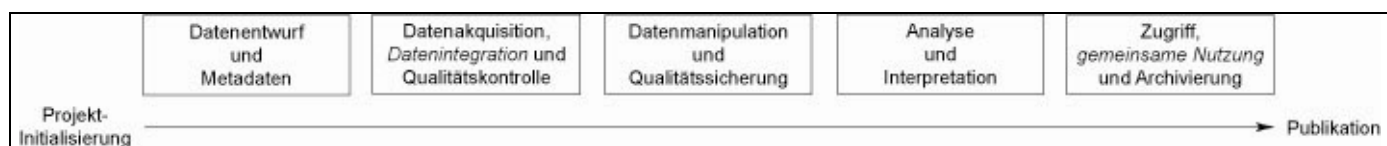


Abbildung 1: Komponenten des Managements ökologischer Daten (adaptiert nach Brunt 2000). Kursiv gesetzte Aufgaben wurden zu Brunt (2000) hinzugefügt und werden als Hauptherausforderungen des heutigen Datenmanagements betrachtet.

nachträglichen Datenintegration für Analyse und Interpretation gewährleistet bleiben.

(Coulouris et al. 2001, S. 23). Als Ausgangspunkt dienen verschiedene Datenbanken des Schweizerischen Bundesam-

tes für Umwelt, Wald und Landschaft (BUWAL), welche Natur- und Landschaftsdaten zu unterschiedlichen Themen beinhalten. Das BUWAL ist verantwortlich für die Akquisition und die Verwaltung solcher Daten auf nationaler Ebene, weist die Umsetzung dieser Aufgaben aber unterschiedlichen Institutionen zu. Die sektorübergreifende Visualisierung und Analyse der Daten erfordert daher eine Plattform, welche die Daten zu integrieren vermag.

Im folgenden Kapitel werden die wichtigsten Grundlagen dieser Arbeit beleuchtet. Anschliessend werden wir auf das Konzept und die Architektur der Virtuellen Datenbank eingehen und deren technische Realisierung aufzeigen. Abschliessend wird der gewählte Ansatz anhand von drei ausgewählten Kriterien diskutiert und bewertet.

## 2 Internetbasierte Nutzung räumlicher Daten

Der Aufbau eines Datenmanagements für die integrierte Handhabung ökologischer Daten, wie es in der Einleitung skizziert wurde, profitiert von Forschungsanstrengungen sowohl in der IT- als auch in der GIS-Gemeinde. Geographische Informationssysteme (GIS) verfügen über Funktionen für die Erfassung, die Speicherung, die Extraktion, das Management, die Analyse und die Ausgabe raumbasierter Daten (Burrough and McDonnell 1998). Die Funktionalität wird in den überwiegenden Fällen durch monolithische Systeme zur Verfügung gestellt. Die jeweils benötigten Daten werden meist in einer einzigen direkt mit dem GIS verknüpften Datenbank gespeichert. Mit dem Aufkommen des Internets und den damit verbundenen neuen Möglichkeiten der Datenverarbeitung hat sich ein Paradigmenwechsel von einer geschlossenen Systemarchitektur zu einem verteilten geographischen Informations-Service-Paradigma vollzogen (Tsou und Buttenfield 2002). Verteilung meint hier sowohl die Aufteilung von Daten auf unterschiedliche Datenbanksysteme als auch die Verteilung von Verarbeitungsressourcen mit der Bildung so genannter Geo-Services (Peng und Tsou 2003). Verschiedene Forschungsanstrengungen und Technologieentwicklungen haben dazu beigetragen, dass dieser Paradigmenwechsel zunehmend umgesetzt wird:

- *Entwicklung von Softwareprodukten:* Die wichtigsten kommerziellen Anbieter von GIS-Produkten haben so genannte Map-Server entwickelt, welche es ermöglichen, Karten in digitaler Form über das Web zu verbreiten. Neben kommerziellen Produkten wurden in OpenSource-Projekten Softwarelösungen entwickelt, die heute erfolgreich eingesetzt werden, wie zum Beispiel der UMN-MapServer (mapserver.gis.umn.edu). Diesen Lösungen ist gemein, dass sie sich vor allem auf Aspekte des Web-Mappings beschränken. Dabei handelt es sich um Techniken zur Nutzung, Verbreitung und Erzeugung von Karten mittels Internet sowie um Techniken für einfache räumliche Abfragen (Kraak 2001, Orthofer und Loibl 2004). Die Manipulier-, Abfrage- und Analyse-möglichkeiten sind jeweils eingeschränkt, was auf den Umstand zurückzuführen ist, dass in den meisten Fällen keine Daten, sondern nur digitale Karten in Form von Bildern zur Verfügung stehen.
- *Definition und Verbreitung offener Schnittstellen:* Die Definition, Spezifikation, Implementierung und Verbreitung offener Schnittstellen für die gemeinsame Nutzung und den Transfer von Geodaten wurde durch das

1994 gegründete OpenGIS Consortium (OGC, heute Open Geospatial Consortium) wesentlich vorangetrieben. Herring (1999) und Kottmann (1999) präsentieren eine vertiefte Diskussion des Datenmodells und des Spezifikationsprozesses, die der Definition der offenen Schnittstellen zugrunde liegen. Für den Internet-/Intranetbasierten Vertrieb von Daten spielen vor allem die OGC-Spezifikationen im Bereich der Webservices eine Rolle. Im Vordergrund steht dabei erstens die vom OGC verabschiedete Spezifikation des Web Map Service (WMS, OGC 2001), die Regeln über die Datenabfrage und die Art der Datenbeschreibung durch die Datenserver enthält. Den WMS charakterisieren dieselben Eigenschaften, wie oben bereits beschrieben: Es lassen sich nur Karten in Form von Bilddaten austauschen, so dass die analytische Nutzung zugrunde liegender Daten versagt bleibt. Die zweite Spezifikation von Interesse, die Web Feature Service Spezifikation (WFS, OGC 2002), definiert Schnittstellen, die der Manipulation und dem Austausch geographischer Objekte dienen. Im Gegensatz zum WMS werden dabei nicht Rastergraphiken, sondern Geodaten in Vektorform übers Internet transportiert.

- *Etablierung der Extensible Markup Language (XML) für den Datentransfer:* XML hat sich heute in verschiedensten Anwendungsgebieten als Format für den Datentransfer über das Web etabliert. In der Geodatenverarbeitung wurde mit der Geography Markup Language (GML, OGC 2003) ein XML-Schema entwickelt, das den Austausch von Geodaten auf der Basis von XML und den Spezifikationen des OGC ermöglicht, so zum Beispiel für den oben erwähnten Datenaustausch mittels WFS.
- *Erweiterung der Funktionalität von webbasierten Applikationen hin zu Analysewerkzeugen:* Softwarehersteller, wie ESRI mit dem *ArcGIS Server*, erweitern ihre Map-Server Produkte um Verarbeitungsfunktionen, die neben einfachen Visualisierungs- und Abfragemöglichkeiten auch die Anwendung fortgeschrittener Funktionen ermöglichen. Neben kommerziellen Entwicklungen zielen auch Forschungsarbeiten in die Richtung der Einbettung verbesserter räumlicher Operatoren ab. Tsou (2004) beschreibt eine auf Java-Applets basierende Applikation, die GIS- und Fernerkundungswerkzeuge integriert, um Probleme der Adressenzuordnung, der Netzwerkanalyse, der Klassifikation von Bilddaten und der Erkennung darin enthaltener Änderungen zu lösen. Anderson und Moreno-Sanchez (2003) demonstrieren die Anwendung offener Spezifikationen und OpenSource-Software in der räumlichen Datenverarbeitung. Die Resultate ihrer Arbeit zeigen, dass Software-Bibliotheken solcher OpenSource-Projekte einen Reifegrad besitzen, der es erlaubt, sie in WebGIS-Projekten einzusetzen.

## 3 Entwurf der Architektur der Virtuellen Datenbank

Die Architektur der Virtuellen Datenbank folgt dem Trend der webbasierten Datennutzung und Datenanalyse. Sie baut auf der Verwendung von standardisierten und offenen Schnittstellen auf. Dabei gelten die folgenden Zielsetzungen:

- Integration von verteilt vorliegenden Datenbanken und Datenbeständen (in der Folge als Datenkomponenten be-

zeichnet) auf unterschiedlichsten Systemen ohne Autonomieeinschränkungen der einzelnen involvierten Komponenten,

- Beschränkung der Funktionalität auf verteilte Datenabfragen,
- Datenzugriff über einheitlich definierte, offene und standardisierte Schnittstellen,
- integrierte Darstellung, Abfrage und Analyse der Daten mittels Web-Browsern und Web-Mapping Software.

Der Entwurf der Virtuellen Datenbank folgt dem Prinzip der losen Kopplung individueller Daten- sowie Softwarekomponenten (Sheth und Larson 1990, Geppert und Dittrich 2001) und ist in klar separierte und zueinander in Beziehung stehende so genannte *Tiers* strukturiert. Abbildung 2 zeigt die notwendigen Daten- und Softwarekomponenten als Elemente von drei einzelnen *Tiers*:

1. Enterprise Information System Tier (EIS tier): Die EIS-Tier besteht aus den verteilten Datenbeständen, die es zu integrieren gilt. Diese liegen räumlich verteilt vor und sind Bestandteil der jeweiligen Institutionen (Unternehmen). Datenbestände können sowohl in unterschiedlichen Datenbankensystemen als auch in unterschiedlichen Dateiformaten vorliegen.

2. Middle Tier: Die Middle-Tier besteht aus mehreren Bestandteilen. Die Datenbestände werden über *Zugriffsschichten* integriert, die Schnittstellen vorgeben, über welche die Daten abgefragt werden können und festlegen, wie die Daten zur Verfügung gestellt werden müssen. Neben Daten kann auch auf beschreibende Metadaten zugegriffen werden. Die Softwarekomponenten der Zugriffsschichten werden jeweils an den Standorten der einzelnen Datenbestände implementiert und installiert. Der *Integrationsschicht* kommen die Aufgaben der Steuerung des verteilten Zugriffs und der Vereinigung der Daten zu, so dass eine transparente Sicht auf die kombinierten Daten ermöglicht wird. Für die Aufgabe der Steuerung muss die Integrationsschicht einerseits die „Adressen“ und Datenbestände der einzelnen Datenkomponenten kennen und andererseits Anfragen, die von Benutzerseite gestellt werden, an die Komponenten verteilen. Die so genannte *Spatial Analysis Engine* stellt Methoden für die Durchführung räumlicher Analysen an den durch die Integrationsschicht zur Verfügung gestellten Daten bereit. Die graphische Aufbereitung vorhandener und analysierter Daten in Form von Karten, Graphiken oder Tabellen übernimmt die *Map-Server*-Komponente.

3. Client Tier: Mit einem Web-Browser wird den BenutzerInnen eine Oberfläche zur Verfügung gestellt, mit der räumliche und nicht-räumliche Daten ausgewählt, abgefragt, analysiert und visualisiert werden können. Für die BenutzerInnen soll dabei verborgen bleiben, aus welcher Datenkomponente die dargestellten Daten stammen.

#### 4 Implementierung der Virtuellen Datenbank

Die Implementierung der einzelnen Softwarekomponenten, insbesondere diejenigen der Middle-Tier, basiert hauptsächlich auf der Programmiersprache Java und den Webtechnologien der Java 2 Plattform, wobei hauptsächlich Java Servlets und JavaServer Pages zum Einsatz kommen. Ein Java-Servlet ist ein Computerprogramm, das innerhalb eines so genannten Servlet-Containers (z.B. Tomcat) auf einem Webserver läuft. JavaServer Pages dienen der dynamischen, serverseitigen Generierung von HTML-Seiten. Die Verwendung dieser Technologien bedeutet, dass sowohl die verteilt vorliegenden Zugriffsschichten als auch die anderen drei Komponenten (Integrationsschicht, Spatial Analysis Engine und Map-Server) an einen Webserver angebunden sein müssen. Als Kommunikationsprotokoll für den Transfer von Abfragen und Daten dient HTTP.

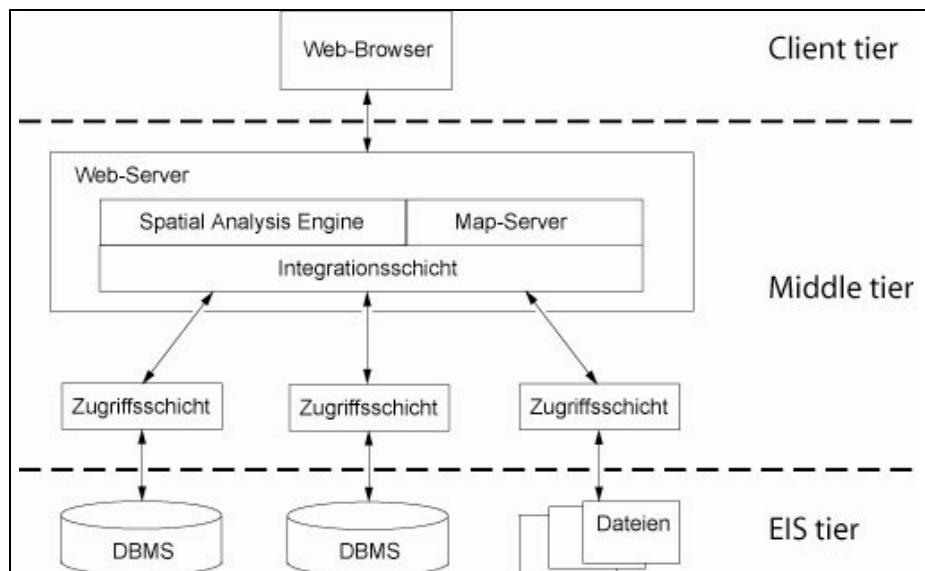


Abbildung 2: Die Architektur der Virtuellen Datenbank

##### 4.1 EIS Tier

Die EIS-Tier der Virtuellen Datenbank besteht aus räumlich verteilt vorliegenden heterogenen Datenbeständen. Die Heterogenität betrifft sowohl die Datenstrukturen als auch die Speicher- und Datenbank Management Systeme (DBMS). In der aktuellen Implementierung eines Prototypen der Virtuellen Datenbank sind Datenbestände von drei Institutionen Bestandteil der EIS-Tier: Beim ersten Datenbestand handelt es sich um das *Datenzentrum für Natur und Landschaft* (DNL, Baltensweiler und Brändli 2004) der Eidgenössischen Forschungsanstalt WSL. Die DNL-Datenbank speichert hauptsächlich Inventardaten von in der Schweiz geschützten Biotopen (Moore, Auen, Amphibienlaichgebiete, etc.). Als DBMS wird Oracle verwendet, worauf die Spatial Database Engine (SDE) von ESRI für die Handhabung räumlicher Datentypen aufgesetzt ist. Die zweite Datenbank ist in Neuenburg am Centre Suisse de Cartographie de la Faune installiert und speichert Fundorte bedrohter Tierarten. Für die Speicherung dient ebenfalls Oracle als DBMS, wobei die x-

und y-Koordinaten der Fundorte als Datenfelder in normalen Oracle-Tabellen abgelegt sind. Die dritte Datenkomponente, die Fundorte von bedrohten und seltenen Moosarten beinhaltet, wird durch das Institut für Systematische Botanik an der Universität Zürich betrieben. Die Attributdaten der Fundorte werden ebenfalls in Oracle-Tabellen abgespeichert. Die Geometriedaten werden hingegen in ESRI-Shapefiles abgelegt, weil es sich bei den Fundorten nicht immer um einzelne Punkte handelt, sondern auch grössere Gebiete umschrieben werden müssen, was eine Speicherung als Polygone notwendig macht

## 4.2 Middle Tier

Die Middle-Tier setzt sich aus den Softwarekomponenten der Zugriffsschichten, der Integrationsschicht, des Map-Servers und der Spatial Analysis Engine zusammen, deren Implementierungen in der Folge einzeln besprochen werden.

### 4.2.1 Zugriffsschicht

Jede Datenkomponente der EIS-Tier erfordert die Implementierung einer Zugriffsschicht, welche auf die entsprechende Datenhaltung zugeschnitten ist. Um eine möglichst hohe Flexibilität und die in der Zielsetzung angestrebte Autonomie der einzelnen Datenbanken zu gewährleisten, wird bei der Implementierung ein Ansatz verfolgt, der auf der Verwendung von Schnittstellen basiert. Dabei fiel die Wahl auf die Spezifikation des Web Feature Service WFS (OGC 2002). Die Spezifikation definiert Schnittstellen für das Abfragen, Generieren, Löschen und Aufdatieren von geographischen Objekten, die je nach Bedürfnis nur teilweise implementiert werden müssen. Für die Virtuelle Datenbank wird das vorgeschriebene Minimum implementiert, das sich auf das Abfragen von Daten und Metadaten mit den drei folgenden Schnittstellen beschränkt:

- *GetCapabilities*: Beinhaltet die Angaben zu den Daten, welche über den Service bezogen und zu den Operationen, die darauf ausgeübt werden können. Die Funktionalität einer über die Zugriffsschicht verfügbaren Datenkomponente wird mit einem XML-Request angefordert und über eine in XML formulierte Antwort zurückgegeben.
- *DescribeFeatureType*: Beinhaltet die Beschreibung der Struktur der zur Verfügung gestellten Daten. Die Struktur wird mittels XML-Schema beschrieben, das sich nach den Elementen der vom OGC spezifizierten Geography Markup Language (GML) richten muss.
- *GetFeature*: Erlaubt den Zugriff auf die Daten über räumliche und nicht-räumliche Abfragekriterien. Wie bei den anderen beiden Operationen erfolgt die Anfrage mit einem XML-Dokument, welches Anfragen und Einschränkungen mit speziellen Filtern definiert. Als Antwort resultiert ein GML-Dokument, dessen Struktur mit dem vorher angefragten XML-Schema übereinstimmen muss.

Auf die unterschiedlichen Formen der Datenhaltung wurde bei der Besprechung der EIS-Tier eingegangen. Die entsprechende Implementierung der Zugriffsschichten soll hier für den Fall der Datenhaltung mit SDE sowie für den Datenzugriff auf Oracle-Tabellen kurz erläutert werden. Für den Zugriff auf SDE kann auf das von ESRI zur Verfügung

gestellte SDE-API für Java zurückgegriffen werden. Dieses ermöglicht den javabasierten Zugriff auf räumliche Datentypen. Demgegenüber kann für den Zugriff auf Oracle das von Java bereitgestellte JDBC-API (Java Database Connectivity) verwendet werden, wofür eine Oracle-spezifische Implementierung vorhanden sein muss. In beiden Fällen erfolgt die Konvertierung der individuell extrahierten Daten in das GML-Format durch eigene Java-Klassen.

### 4.2.2 Integrationsschicht

Analog zu den Zugriffsschichten wird auch die Integrationsschicht mit Java und Java-Servlets auf einem Webserver implementiert. Die Integrationsschicht verteilt Anfragen an die Zugriffsschichten und sammelt deren Antworten in Form von GML-basierten XML-Schemen und GML-Daten ein. XML-Parser aus OpenSource-Projekten helfen bei der Interpretation und beim Zusammenfügen der empfangenen Daten.

Während der Implementierung hat sich gezeigt, dass beim Transfer von GML-Daten im Gegensatz zu den bei Webapplikationen üblicherweise im JPEG-Format transferierten Bilddaten grosse Datenmengen anfallen. Dies hat zu Datenübertragungszeiten geführt, die weit über tolerierbaren Werten lag. Deshalb wurde für die Integrationsschicht zusätzlich ein Caching-Mechanismus implementiert, der grosse Mengen an Geometriedaten repliziert hält. Das Problem, die Daten aktuell zu halten, wird dadurch gelöst, dass die Zugriffsschichten den Status ihrer Daten regelmässig überprüfen und der Integrationsschicht Änderungen signalisieren. Bei Bedarf löst die Integrationsschicht ihrerseits die Aufdatierung des entsprechenden Replikats aus.

### 4.2.3 Map-Server

Zur Visualisierung von Daten, Abfragen und Analysere-sultaten in Form von Karten, Graphiken und Tabellen, wird der Internet-Map-Server *ArcIMS* von ESRI verwendet. *ArcIMS* bietet verschiedene Java-Klassen und so genannte Tag-Libraries für die Programmierung von JavaServer Pages an, welche die Gestaltung von HTML-Seiten und den serverseitigen Einbau von graphischen Kartenelementen und Legenden erlauben. Die Verwendung von *ArcIMS* hat hauptsächlich institutionelle und historische Gründe, es könnten auch andere Map-Server-Produkte wie beispielsweise der *UMN MapServer* zum Einsatz kommen.

### 4.2.4 Spatial Analysis Engine

Mit der Verwendung von *ArcIMS* als Map-Server stehen für die Virtuelle Datenbank verschiedene einfache räumliche Datenverarbeitungsfunktionen wie räumliche Abfragen und Pufferberechnungen bereits zur Verfügung. Allerdings reichen diese nicht aus, um komplexeren Fragestellungen und Modellierungsaufgaben, die eine Verknüpfung verschiedener Datensätze voraussetzen, nachzugehen. Die wichtigste Funktion der Verknüpfung ist die Verschneidung (Overlay) geometrischer Daten, die die Grundlage für räumliche Modellierungsaufgaben und viele weitere Operationen bildet. Mit den drei in der EIS-Tier beschriebenen Datensätzen könnte mit einer Overlay-Funktion u. a. der Frage nachgegangen werden, welche bedrohten Tierarten oder seltenen

Moosarten sich innerhalb bestimmter Schutzperimeter von Auenflächen befinden. Wegen der grossen Bedeutung dieser räumlichen Verschneidungsoperationen wurde deshalb für die *Spatial Analysis Engine* ein Overlay-Tool implementiert, das die Verschneidungen von beliebigen Geometrien ermöglicht. Verwendet wurde dazu die Java-Version der MapObjects-Programmbibliothek von ESRI (Version 2.0).

### 4.3 Client Tier

Für die Client-Tier kommt ein einfacher Web-Browser zum Einsatz, der als dünner Klient für die Benutzerinteraktion und Darstellung der Daten konzipiert ist. Klientenseitiger Code beschränkt sich bei dünnen Klienten auf HTML und JavaScript, welche für die Anforderungen der Virtuellen Datenbank völlig genügen. Von der Middle-Tier her werden Karten nur als Bilddaten - beispielsweise im JPEG-Format - übermittelt und dargestellt. Als Alternative wurde auch mit Scalable Vector Graphics (SVG) experimentiert. Jedoch haben Datenschutzprobleme die Verwendung von Vektordaten durch den Web-Browser bisher verhindert.

### 5 Diskussion

Für die Virtuelle Datenbank ist im Moment ein Prototyp implementiert, der die oben erwähnten drei unterschiedlichen Datenbestände der EIS-Tier zu einem Datenverbund integriert und der im Moment ausschliesslich für die drei involvierten Institutionen sowie für das BUWAL zugänglich ist. Die ersten Erfahrungen, die mit dem Prototypen gemacht wurden, sollen hier kurz diskutiert werden, wobei wir uns auf die drei Themen Flexibilität, Skalierbarkeit und Metadaten beschränken.

Der schnittstellenbasierte Ansatz für den Zugriff auf die einzelnen Datenbestände hat sich als hoch flexibel erwiesen. Dank der Verwendung von Schnittstellen müssen für die einzelnen Datenkomponenten nicht Datenbankschemen oder Speicherformen umstrukturiert, sondern lediglich Anpassungen der Software für den Zugriff durchgeführt werden, um die Daten schnittstellenkonform abfragen zu können. Die Datenbanken behalten so weitestgehend ihre Autonomie. Zwar wurden zur Vereinfachung des Datenzugriffs vereinzelt Datenbank-Views angelegt, was aber lediglich als Ergänzung der einzelnen Datenbankschemen gewertet werden muss.

Die realisierte Architektur der Virtuellen Datenbank zeichnet sich auch durch eine hohe Skalierbarkeit aus. Die oben identifizierte Herausforderung grosser Datenmengen, die sich insbesondere bei Geometrien räumlicher Daten und der Umwandlung in GML ergibt, konnte durch Datenreplikation entschärft werden. Wird die Datenmenge grösser oder kommt eine neue Datenkomponente hinzu, fällt zwar in der Integrationsschicht ein leicht erhöhter Aufwand bezüglich Datenintegration (Parsen, Erstellen des Replikats) an, der hinsichtlich der allgemeinen Leistungsfähigkeit jedoch kaum ins Gewicht fällt. Zudem können diese Aktionen im Hintergrund der Webapplikation ausgeführt werden, so dass sie dem Benutzer verborgen bleiben. Probleme wirft im Moment die Leistung der *Spatial Analysis Engine* bezüglich gegenseitiger Verschneidung grösserer Datenmengen auf. Eine Leistungssteigerung erwarten wir durch den geplanten Einbau von alternativen und durch OpenSource-Projekte frei verfügbaren Software-Bibliotheken wie zum Beispiel Geotools

([www.geotools.org](http://www.geotools.org)) und die JTS Topology Suite ([www.vividsolutions.com/jts](http://www.vividsolutions.com/jts)).

Bei der Integration von Daten aus unterschiedlichen Quellen spielt die Möglichkeit der Beurteilung der Daten - speziell der Datenqualität - für die Eignung hinsichtlich einer spezifischen Anwendung eine grosse Rolle. Zudem können durch unterschiedliche Bearbeitungsgenauigkeiten der einzelnen Datenbestände bei der Kombination von Daten Inkonsistenzen auftreten. Eigenschaften eines Datensatzes, worunter auch Angaben zur Datenqualität fallen, werden durch begleitende Metadaten beschrieben. Die angebotenen Schnittstellen des WFS bieten eine eingeschränkte Möglichkeit, auf solche Metadaten zuzugreifen. Dabei handelt es sich um Angaben wie den Titel, eine kurze Inhaltsbeschreibung und Schlüsselwörter. Zusätzlich kann eine URL-Adresse angegeben werden, über die weitere in standardisierter Form abgelegte Metadaten zu finden sind. Solche erweiterten Metadaten können Informationen über fehlerhafte Daten oder Datenunsicherheiten liefern und auf mögliche Datenheterogenitäten innerhalb der Datenintegration hinweisen. Die aktuelle Implementierung der Virtuellen Datenbank greift zwar auf solche Metadaten zu, berücksichtigt sie aber nicht für die Integration. Dies liegt vor allem daran, dass für die dem Prototypen zur Verfügung stehenden Datenkomponenten mit Ausnahme des DNLs (Baltensweiler und Brändli 2004) keine Datenqualitätsangaben vorhanden sind.

Auf einige Grundvoraussetzungen bezüglich der Beschreibung der Datenqualität hat man sich innerhalb des Datenverbundes einigen können. Diese betreffen im Moment ausschliesslich Angaben zu Raum und Zeit vorhandener Daten. So müssen Koordinatenwerte zur räumlichen Lokalisierung, beispielsweise von Fundorten untersuchter Tierarten, mit einer Genauigkeitsangabe versehen werden. Dies gilt ebenfalls für die Angaben zum Datum eines Fundes oder der Erhebung eines Datensatzes. Für die kombinierte Analyse von Daten sollten Genauigkeitsangaben neben Raum und Zeit allerdings auch die Thematik einschliessen, so dass auch weitere in einem Datensatz enthaltene Attribute näher charakterisiert werden müssen.

Ein wichtiger Schritt für die Zukunft wird daher einerseits sein, den involvierten Institutionen eines solchen Datenverbundes die Bedeutung von Metadaten bezüglich der Datenqualität sämtlicher in einem Datensatz enthaltener Merkmale aufzuzeigen, damit die Datensätze entsprechend charakterisiert werden können. Andererseits muss die methodische Forschung um den Einbau und die Entwicklung existierender und neuer Ansätze zur Berücksichtigung von Datenheterogenitäten, Datenunsicherheiten und fehlerhaften Daten erweitert werden. Dies gilt insbesondere für die Aufgabe der Integration der Daten, aber auch für Analyseschritte mit der Fähigkeit, Datenqualitätsangaben zu berücksichtigen, zu verfolgen und weiterzugeben.

### 6 Schlussfolgerungen

Wir haben eine Architektur für eine integrierte webbasierte Plattform zur Integration und Analyse von verteilt vorliegenden Umweltdatenbanken präsentiert. Der grosse Vorteil der Virtuellen Datenbank liegt darin, dass jedermann, der einen Webbrowser besitzt und Zugriff zum Internet hat, transparent auf räumliche Daten zugreifen und diese umfassend visualisieren und explorieren kann. Zugleich können die Daten für

fortgeschrittene räumliche Analysen verwendet werden. Der gewählte Ansatz der Datenintegration hat sich dank des Zugriffs mit offenen und standardisierten Schnittstellen als äusserst flexibel erwiesen. Die Leistungsfähigkeit der Plattform äussert sich durch eine hohe Skalierbarkeit, welche vor allem durch das Replizieren von grossen Datenmengen, wie sie bei der Verarbeitung räumlicher Daten häufig auftreten, erreicht wird. Der aktuelle Prototyp der Virtuellen Datenbank muss in Zukunft vor allem durch Methoden ergänzt werden, welche sowohl bei der Integration als auch bei der Analyse verteilt vorliegender Datenbestände auf Datenqualitätsmerkmale und auftretende Datenheterogenitäten eingehen.

## 7 Literatur

- ANDERSON, G., MORENO-SANCHEZ, R. (2003): Building Web-Based Spatial Information Solutions around Open Specifications and Open Source Software. *Transactions in GIS*, 7:447-466.
- BALTENSWEILER, A., BRÄNDLI, M. (2004): Web-based Exploration of Environmental Data and Corresponding Metadata, in Particular Lineage Information. In: Scharl, A. (Hg.): *Environmental Online Communication. Advanced Information and Knowledge Processing Series*, Springer, London, S. 127-132.
- BRUNT, J. W. (2000): Data Management Principles, Implementation and Administration. In: Michener, W. K., Brunt, J. W. (Hg.): *Ecological Data: Design, Management and Processing*. Blackwell Science, Oxford, S. 25-47.
- BURROUGH, P. A., MCDONNELL, R. A. (1998): *Principles of Geographical Information Systems*. Oxford University Press, Oxford.
- COULOURIS, G., DOLLIMORE, J., KINDBERG, T. (2001): *Distributed Systems, Concepts and Design*. Third edition, Pearson Education Ltd, Essex.
- GEPPERT, A., DITTRICH, K. R. (2001): Component Database systems: Introduction, Foundations, and Overview. In: Dittrich, K. R., Geppert, A. (Hg.): *Component Database Systems*. Morgan Kaufmann Publishers, San Francisco, S. 1 - 28.
- HERRING, J. (1999): The OpenGIS data model. *Photogrammetric Engineering and Remote Sensing*, 65: 585-588.
- KOTTMAN, C. (1999): The Open GIS Consortium and progress toward interoperability in GIS. In: Goodchild, M. F., M. Egenhofer, R. Fegeas und C. Kottman (Hg.): *Interoperating Geographic Information Systems*, Kluwer, Boston.
- KRAAK, M.-J. (2001): Settings and needs for web cartography. In: Kraak, M.-J., Brown, A. (Hg.): *Web Cartography. Developments and prospects*. Taylor and Francis, London und New York.
- LUTZ, M., RIEDMANN, C., PROBST, F. (2003): A Classification Framework for Approaches to Achieve Semantic Interoperability between GI Web Services. In: Kuhn, W., M. F. Worboys, M. F., Timpf, S. (Hg.): *COSIT 2003. Lecture Notes in Computer Sciences*, 2825, Springer, Heidelberg, S. 186 – 2003.
- MICHENER, W. K. (2000): Research Design: Translating Ideas to Data. In: Michener, W. K., Brunt, J. W. (Hg.): *Ecological Data: Design, Management and Processing*. Blackwell Science, Oxford, S. 1-24.
- OGC (2001): *Web Map Service Implementation Specification*. Version: 1.1.1. Open GIS Consortium, Inc.
- OGC (2002): *Web Feature Service Implementation Specification*. Version: 1.0.0. Open GIS Consortium, Inc.
- OGC (2003): *OpenGIS® Geography Markup Language (GML) Implementation Specification*. Version: 3.00. Open GIS Consortium, Inc.
- ORTHOFFER, R., LOIBL, W. (2004): Sharing Environmental Maps on the Web: The Austrian EnviroMap System. In: Scharl, A. (Hg.): *Environmental Online Communication. Advanced Information and Knowledge Processing Series*, Springer, London, S. 133-144.
- PENG, Z., TSOU, M. (2003). *Internet GIS: Distributed geographic information services for the internet and wireless networks*. John Wiley & Sons, New Jersey.
- SHETH, A. P., LARSON, J. A. (1990): Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys*, 22: 183-236.
- TSOU, M.-H., BUTTENFIELD, B. P. (2002): A Dynamic Architecture for Distributing Geographic Information Services. *Transactions in GIS*, 6:355-381.
- TSOU, M.-H. (2004): Integrating Web-based GIS and image processing tools for environmental monitoring and natural resource management. *Journal of Geographical Systems*, 6:155-174.
- VCKOVSKI, A., K. E. BRASSEL, K. E., SCHECK, H.-J. (1999): Interoperating Geographic Information Systems. *Proceedings Second International Conference, INTEROP'99. Lecture Notes in Computer Science*, Vol. 1580, Springer, Berlin, Heidelberg.

## Kurzfassung

*Mit der Virtuellen Datenbank wird eine Architektur präsentiert, die auf die Integration, Visualisierung und Analyse von verteilt vorliegenden räumlichen Datenbanken und Datenbeständen abzielt. Die Architektur belässt die Daten zentral an ihren jeweiligen Standorten und greift internetbasiert und über einheitlich definierte offene Schnittstellen auf die Daten zu. Berücksichtigt werden dabei die OpenGIS-Spezifikationen des Web Feature Service (WFS) und der Geography Markup Language (GML). Mit der Virtuellen Datenbank lassen sich durch die eingegliederte Kopplung an die Software von Map-Servern und einer erweiterten räumlichen Analysefunktionalität verteilte Daten umfassend visualisieren, abfragen und für räumliche Analysen verwenden.*

**Stichworte:** Datenintegration, Verteilte Datenbanken, Web-Services, Offene Schnittstellen, Umweltdaten

## Summary

*We present the Virtual Database - an architecture aiming at the integration of heterogeneous distributed spatial data repositories for comprehensive data visualization, exploration and analysis. Different databases are brought together in order to build an integrated data federation. Data access is based on uniformly defined, standardized and open interfaces conforming to the OpenGIS specifications for Web Feature Services (WFS) and the Geography Markup Language (GML), and selected interoperable technologies proposed*

*and ratified by the W3C Consortium. Visualization, exploration and analysis functionality are provided by the integrated coupling of map server software and extended spatial data handling methods. First experiences with the Virtual Database show its high flexibility and scalability resulting from the interface-based approach and a caching mechanism based on data replication.*

**Keywords:** *Data integration, distributed data bases, Web services, open interfaces, environmental data*

## **Autoren**

### **Dr. Martin Brändli**

Eidgenössische Forschungsanstalt für Wald, Schnee und  
Landschaft WSL  
Zürcherstrasse 111  
CH-8903 Birmensdorf  
Fon: +41 1 739 23 92, Fax.: +41 1 739 22 15  
Email: martin.braendli@wsl.ch

### **Dipl. geogr. Marcel Frehner**

Eidgenössische Forschungsanstalt für Wald, Schnee und  
Landschaft WSL  
Zürcherstrasse 111  
CH-8903 Birmensdorf  
Fon: +41 1 739 26 83, Fax.: +41 1 739 22 15  
Email: marcel.frehner@wsl.ch