

Dietrich Göttlicher und Jörg Bendix

Eine modulare Multi-User Datenbank für eine ökologische Forschergruppe mit heterogenem Datenbestand

Der reibungslose Austausch von Datensätzen innerhalb einer interdisziplinär ausgerichteten ökologischen Forschergruppe kann nur durch eine geeignete Datenbankstruktur garantiert werden. Die Konzeption, Struktur und Implementierung einer geeigneten Datenhaltung wird am Beispiel der komplexen, relationalen Datenbank der DFG-Forschergruppe 402 vorgestellt.

1 Einführung

Die DFG Forschergruppe 402 (FOR 402) „*Funktionalität in einem tropischen Bergregenwald Südecuadors: Diversität, dynamische Prozesse und Nutzungspotentiale unter ökosystemaren Gesichtspunkten*“ ist ein Gemeinschaftsprojekt von Biologen, Ökologen, Geographen, Geowissenschaftlern, Forstwissenschaftlern und Wissenschaftlern anderer Disziplinen. Das zentrale Arbeitsgebiet umfasst die Reserva Biósfera de San Francisco um die ECSF Forschungsstation (Estación Científica de San Francisco), die sich im Tal des Rio San Francisco (3°58'18''S, 79°4'45''W, 1860 m über Meer) östlich der Stadt Loja (Südecuador) befindet (Beck & Müller-Hohenstein 2001, Bendix et al. 2004). Die Untersuchungen beziehen sich auf den tropischen Bergregenwald entlang von Höhen transekten zwischen 800 bis 3600 m über Meer im erweiterten Projektgebiet. Grundlegendes Ziel der Forschergruppe ist eine floristische und faunistische Inventarisierung verschiedener Organismengruppen sowie die umfassende Geländeaufnahme der abiotischen Lebensbedingungen (Relief, Klima, Wasserhaushalt, Boden) im natürlichen Bergregenwald und auf Flächen verschiedener Degradationsstufen (z.B. Sekundärwald, bewirtschaftete und aufgelassene Agrarflächen). Damit soll ein umfassendes Verständnis des gesamten Ökosystems erreicht werden, das als Grundlage zur Entwicklung von Konzepten für eine nachhaltige Forstnutzung mit einheimischen Baumarten unter Erhalt der natürlichen Biodiversität dienen soll (zur weiteren Information sei auf www.bergregenwald.de verwiesen). Da die meisten Analysen interdisziplinär angelegt sind, ist ein reibungsloser Austausch von aufbereiteten und qualitätsgeprüften Daten über die Datenbank unbedingte Voraussetzung. Zum besseren Prozessverständnis, aber auch zur Simulation der Auswirkung verschiedener Nutzungsstadien auf die abiotische und biotische Umwelt ist weiterhin der Einsatz statistischer und numerischer Modelle für Teilkomplexe des Ökosystems vorgesehen. Hier spielt die Datenbank eine zentrale Rolle, da zur Initialisierung, Eichung und Validierung der Modelle ein normierter Zugriff auf den interdisziplinären Datenbestand notwendig ist.

Die gesamte Forschergruppe besteht aus mehr als dreißig einzelnen Arbeitsgruppen, die sich in vier Projektbereiche

gliedern. Insgesamt sind über 70 Mitarbeiter deutscher und ecuadorianischer Forschungseinrichtungen in der Forschergruppe tätig.

Zur zielgerichteten Verarbeitung von spezifischen Datensätzen eines Mitarbeiters in anderen Teilprojekten reicht ein Austausch der eigentlichen Datensätze alleine nicht aus. Vielmehr müssen dem Nutzer vielfältige Zusatzinformationen zu den einzelnen Datensätzen zugänglich gemacht werden. Diese sogenannten Metadaten sind beispielsweise Informationen über Ort und Zeit der Erhebung, aber auch genauere Angaben mit welcher Methodik die Daten erhoben wurden bzw. welche Datenqualität zu erwarten ist (DENZER & GÜTTLER 1997, Krasemann 1996). Ein einfaches, gemeinsam zu nutzendes Dateisystem, auf das z.B. über das File Transfer Protokoll (ftp) zugegriffen werden könnte, wird diesem erweiterten Informationsbedarf nicht gerecht. Darüber hinaus macht die Heterogenität der gesammelten Datentypen (Datentabellen, Bilddaten, GIS-Layer etc.) die Nutzung einer handelsüblichen, vorkonfigurierten Datenbanksoftware nahezu unmöglich.

Andere interdisziplinäre Verbundprojekte haben gezeigt, dass besonders individuelle angepasste Lösungen bei der Implementierung der notwendigen Metadatenbank ein probates Mittel darstellen, die Interdisziplinarität und den Austausch von Forschungsergebnissen innerhalb der Forschergruppen nachhaltig zu verbessern (s. z.B. Informationssystem PANGAEA, Diepenbroeck et al. 1997, SPECCHIO, Bojinski et al. 2003, CERA-2, Toussaint et al. 1999, KERIS, HOSENFELD 1999, MÜCKSCHEL et al. 2004).

Im vorliegenden Artikel wird das Metadatenbankkonzept der DFG-Forschergruppe 402 (FOR402meta) vorgestellt. Das System ist als klassische Client-Server Umgebung ausgerichtet. Der Nutzer greift über eine gewöhnliche Internetverbindung mit einem Standard-Webbrowser auf die Datenbank zu. Eine graphische Benutzerführung erleichtert Datensuche und Zusammenstellung der Datensätze. Grundsätzliche Entwicklungsmerkmale der Metadatenbank orientieren sich an den Vorgaben der einzelnen Arbeitsgruppen. Die Struktur wird im laufenden Entwicklungsprozess neuen Anforderungen kontinuierlich angepasst.

2 Vorbereitungsphase und Konzeption

2.1 Entwicklung der Forschergruppe

Aus der Historie der Forschergruppe ergibt sich ein Entwicklungskonzept für das Datenbank- und Informationssystem, das von herkömmlichen Herangehensweisen mit z.B. vereinheitlichten Datenformaten und integrierten Datenmodellen deutlich abweicht. Der Forschergruppe vorgeschaltet war bereits eine 4-jährige Forschungsphase (im Rahmen eines gebündelten Normalverfahren der DFG von 1997-2000), in der datenintensive Kernarbeitsgruppen (z.B. Meteorologie) mit der Messwerterfassung begannen und darüber hinaus wichtige Grundlagendaten (GIS-Layer) erhoben wurden, die für das gesamte Projekt auch heute noch von großer Bedeutung sind. Über ein Konzept zur einheitlichen Datenhaltung wurde damals nicht nachgedacht, vielmehr wurden alle Daten in generischen Formaten und nach der spezifischen Vorgehensweise der einzelnen Arbeitsgruppen dezentral abgelegt. Auch mit Einrichtung der Forschergruppe 2001 änderte sich diese Herangehensweise nicht grundsätzlich. Allerdings wurde hier bereits eine spezifische botanische Bilddatenbanken mit Bestimmungsfunktion entwickelt (Visual Plants, Dalitz & Homeier 2004). In der ersten zweijährigen Phase (2001-2002) konnte die interdisziplinäre Forschung deutlich verstärkt werden, basierte datentechnisch aber v.a. auf bilateralem Austausch mit der zugehörigen Information, wie die Daten der jeweils anderen Gruppe zu handhaben bzw. zu interpretieren sind. Auf Anregung der Gutachter und mit der Einsicht, dass zukünftig v.a. Modellierungsvorhaben auf multidisziplinäre Datenbestände zurückgreifen müssen, wurde beschlossen, für die zweite Phase der Forschergruppe (2003-2004) ein zentrales Datenbankprojekt einzurichten das nunmehr auch die in sechs Jahren gewachsenen Datenstrukturen integrieren sollte. Bei der Planung wurde auf eine möglichst kostengünstige Struktur (Open Source Software, Technikerstelle) Wert gelegt und die Zeitbelastung der einzelnen Arbeitsgruppen zur Aufbereitung ihrer Datensätze sollte nicht ansteigen. In der Planungsphase wurde eine Befragung der Mitglieder durchgeführt, wie und in welchen Formaten die Daten besonders zum Austausch vorliegen sollen. Es ergab sich hier ein klares Votum z.B. für Tabellenverarbeitungsprogramme (EXCEL), die aus Datenbank technischer Sicht wenig komfortabel erscheinen, aber für die Mitglieder der Forschergruppe als außerordentlich wünschenswert gekennzeichnet wurden. Als größeres Problem ergab sich ein mit der Datenmenge und Anzahl der Arbeitsgruppen ansteigendes Informationsdefizit, welche Daten in welcher Qualität etc. vorhanden sind. Auf der Basis der eingeschränkten Ressourcen für das Zentralprojekt und den Ergebnissen der Befragung wurde das Konzept für ein zentrales Datenbank- und Informationssystem entwickelt, das die folgenden Kernpunkte enthält:

(a) Die Datenerstellung durch die einzelnen Arbeitsgruppen soll aus Gründen der Arbeitsminimierung mit den generisch bevorzugten Programmen weitergeführt und die Dateien in einem einfachen Filesystem zentral vorgehalten werden. Allerdings wurden Normierungen z.B. der Dateinamen, der Tabellenstruktur bei gleichartigen Datentypen (z.B. Sammel-listen) und bestimmte Aggregierungsstufen (sowohl zeitlich als auch räumlich) vereinbart, die den Datenaustausch und die Einbindung in die Metadatenbank deutlich vereinfachen.

Von den betroffenen Arbeitsgruppen mussten damit aber auch die historischen Daten rückwirkend bearbeitet werden. Auf die Entwicklung und Implementierung eines einheitlichen Datenmodells wurde v.a. aufgrund der eingeschränkten Ressourcen und den Wünschen der Arbeitsgruppen verzichtet.

(b) Um das Informationsdefizit abzubauen und um eine besser Planbarkeit für multidisziplinäre Fragestellungen (z.B. kann ich diesen Datensatz für meine Modellierungsvorhaben verwenden?) zu gewährleisten, wurde der Schwerpunkt auf die Entwicklung einer normierten Metadatenstruktur gelegt. Dies hebt vor allem den Informationscharakter des Systems hervor.

2.2 Inventarisierung

Um eine zielorientierte Ausarbeitung der Datenbankkonzeption vornehmen zu können, wurden in einer Vorbereitungs- und Inventarisierungsphase (2002) alle Arbeitsgruppen zu den von ihnen erhobenen Datensätzen befragt. Folgende Parameter, die für die konzeptionelle Gestaltung der Datenbank wichtig sind, wurden abgefragt:

1. Welche Art von Daten werden erhoben?
2. In welchem Dateiformat werden die Daten gespeichert?
3. In welchen Zeitabschnitten werden die Daten erhoben?
4. Wieviel Speicherplatz ist für die erhobenen Daten notwendig?
5. Welche Datensätze werden von anderen Arbeitsgruppen benötigt?
6. In welchem Dateiformat sollten diese Datensätze vorgehalten werden?
7. Welche zusätzlichen Suchkriterien können für die Daten geliefert werden? Welche Suchkriterien werden für Datensätze anderer Arbeitsgruppen benötigt?

Die Ergebnisse der Befragung zeigten, dass fast 90 % der Daten in Tabellenform vorliegen. Hier sind in erster Linie verschiedene Messdaten, die zeitlich regelmäßig an einem Ort erhoben werden, zu nennen. Circa 70 % der Arbeitsgruppen erheben zudem Daten in einem Digitalbild oder GIS-Format. Hierbei handelt es sich weitestgehend um digitale Photographien von Pflanzen, Wurzeln, Bodenprofilen, etc. bzw. beschreibende Vektorebenen. Allgemein lässt sich festhalten, dass die Inhalte der erhobenen Daten sehr heterogen sind. Die verwendeten bzw. bevorzugten Dateiformate sind dementsprechend vielfältig. Der größte Teil der tabellarischen Daten liegt im Microsoft Excel Format vor. Digitale Bilder sind größtenteils als JPEG, aber auch in anderen Bildformaten (z.B. TIFF, Bitmap) gespeichert. Die GIS-Daten liegen in der Regel im generischen ArcView Shapeformat, die Fernerkundungsdaten im binären IDRISI-Format vor. Insgesamt ist bei den GIS- und Satellitendaten die Unsicherheit verfügbarer und verbreiteter Formate besonders groß. Abgeschlossene Dokumente wie Poster und Zwischenberichte liegen hauptsächlich als PDF vor.

Besonders wichtig für die weitere Planung ist es, einen Überblick des zu erwartenden Speichervolumens zu erlangen. Die Bild- und GIS-Daten einschließlich der Fernerkundungsdaten weisen erwartungsgemäß den größten Speicherbedarf auf. Tabellarische Daten und andere Dokumente treten in Bezug auf den benötigten Speicherplatz in den Hintergrund. Auf der Grundlage der eingegangenen Fragebögen fallen für die 6-jährige Projektdauer in der gesamten Forschergruppe fast 300 GB Daten an.

2.3 Datenvereinbarung

Eine bedeutende Frage innerhalb einer großen Forschergruppe sind verbindliche Regeln zu den Nutzungs- und Verwertungsrechten spezifischer Datensätze. Jedes Mitglied der Forschergruppe muss eine eigens für die Forschergruppe entwickelte Datenvereinbarung unterzeichnen, in der die Details der Verwertung geregelt sind. So muss beispielsweise bei der Verwendung von Daten der Urheber in Kenntnis gesetzt werden und Veröffentlichungen der Daten bedürfen einer genaueren Absprache zwischen den einzelnen Parteien. Der Zugriff auf die Datenbank ist nur nach der Unterzeichnung der Datenvereinbarung möglich.

2.4 Konzeption

Die konzeptionellen Anforderungen der zu entwickelnden Metadatenbank ergeben sich aus der Befragung der einzelnen Arbeitsgruppen sowie den weiteren Vorgaben hinsichtlich der finanziellen projektinternen Förderung. Im Einzelnen müssen folgende Punkte bei der Entwicklung beachtet werden:

- Die Datensätze müssen vor dem Zugriff Dritter gesichert werden.
- Die Datenbank muss über eine normale Internetverbindung zu erreichen sein.
- Eine nutzerfreundliche Schnittstelle soll die Handhabung auch für Datenbank ungeübte Mitarbeiter ermöglichen.
- Das Budget legt den Einsatz von freier (Open Source) Software nahe.
- Das logische, konsistente, relationale Datenmodell muss den mit der Laufzeit anwachsenden Anforderungen genügen.
- Neue Datensatztypen sollen flexibel integriert werden können.
- Die Datenbank muss mehrsprachig aufgebaut sein (deutsch, englisch, spanisch).

Die Entwicklung der Datenbankstruktur folgt einem evolutionären Zyklus (Sommerville 2001). Erste Basisversionen werden innerhalb ausgewählter Arbeitsgruppen getestet und weiterentwickelt. Die Nachteile einer evolutionären Entwicklung bestehen oft in einer schlechten Systemarchitektur und mangelhafter Dokumentation, da verschiedene Systemteile kontinuierlich korrigiert bzw. neu eingefügt werden. Kompensiert werden diese Nachteile durch die Tatsache, dass die vorliegende Metadatenbank von einer Person programmiert wird. Der Entwickler behält die gesamte Übersicht und kann alle neuen Anforderungen in das bestehende System integrieren. Die schnelle, praxisnahe Umsetzung und die Bereitstellung des Systems ist somit gewährleistet.

Beim Aufbau der Datenbankstruktur werden die oben angeführten Anforderungskriterien umgesetzt. Dabei wird ein Internet basiertes Client-Server System auf der Basis von frei verfügbarer Software ausgewählt. Die Anmeldung an die Datenbank erfolgt über eine geschützte Benutzernamen- und Passwortabfrage. Das System basiert auf einer relationalen Datenbank, die modular erweitert werden kann und Mehrsprachigkeit unterstützt.

Die grundlegenden Regeln bei der Erstellung der relationalen Datenbank zur Vermeidung von Redundanzen und

Prozessdaten werden eingehalten. Die Entwicklung folgt weitgehend dem Einsatz der Normalformen bis zur 3. Stufe. Damit die Komplexität der Datenbankstruktur überschaubar bleibt und in einigen Bereichen eine besserer Abfrageeffizienz erreicht werden kann, wird dies jedoch nicht bis zur letzten Konsequenz verfolgt und ein kleine Teile bestehender Information nicht extra in eigene Tabellen ausgelagert (Kofler 2003). Das ausgewählte Datenbanksystem unterstützt einfache relationale Verknüpfungen wie 1:1 oder 1:n Beziehungen durch Vergabe von Fremdschlüsseln. Viele-zu-Viele (n:m) Verknüpfungen werden über den Einsatz einer extra dafür angelegten Tabelle realisiert. Die referentielle Integrität des Datenbestandes wird über Abfragen aus dem entwickelten Datenbanksystem geprüft.

3 Datenbankdesign und Implementierung

3.1 Systemaufbau

Das gesamte System wird nach den Projektvorgaben in einer klassischen Client-Server Umgebung realisiert (Abb. 1). Die Konfiguration entspricht bekannten Systemen zur Generierung dynamischer HTML-Seiten (z.B. Kofler 2003). Als

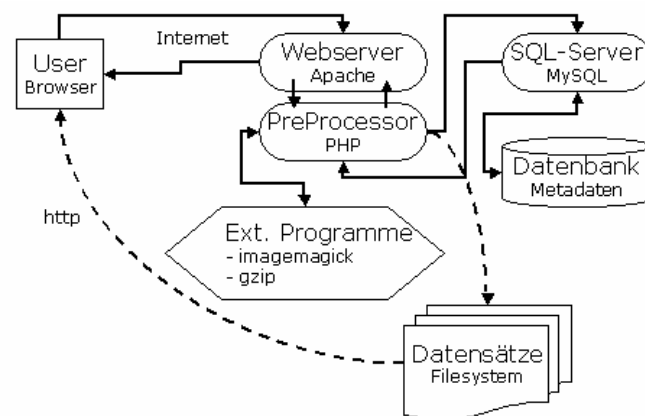


Abbildung 1: Schematischer Aufbau des Datenbanksystems „FOR402meta“

Software kommt ein normaler Webserver (Apache, www.apache.org), der durch einen Präprozessor (PHP, www.php.net) zur Erzeugung dynamischer HTML-Seiten ergänzt wird zum Einsatz. Als Datenbankserver wird auf eine MySQL Datenbank zurückgegriffen (www.mysql.com). Der Server besteht aus einem Standard PC, der über das Hochschulrechenzentrum mit dem Internet verbunden ist. Als Dateispeicher dienen zum einen handelsübliche Festplatten, aber auch ein DVD-Juke-Box System, in dem 105 DVDs je nach Bedarf automatisch gelesen werden können. Als Besonderheit führt das PHP-Modul bei Bedarf eigenständige Programme auf dem Server aus. So ist eine Verbindung zu einem Programm zur Manipulation von Bildern (imagemagick, www.imagemagick.org) und zur Komprimierung / Zusammenfassung von Dateien (gzip, www.gzip.org) implementiert. Die Software zur Bildbearbeitung wird für die Erstellung von Karten mit Stationsdaten verwendet, die Komprimierung und Zusammenfassung von Dateien kommt beim Herunterladen der Datensätze zum Einsatz.

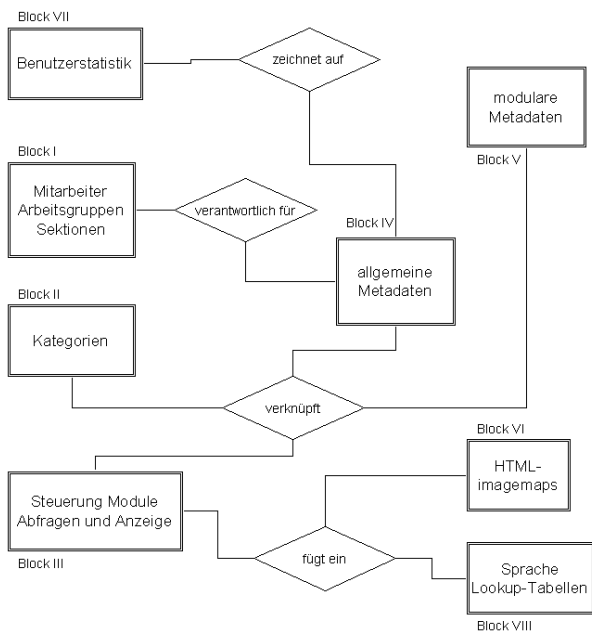


Abbildung 2: Schematische Darstellung der relationalen Datenbankstruktur „FOR402meta“.

Die eigentlichen Dateien mit den erhobenen Daten werden in einem normalen Dateisystem vorgehalten. In erster Linie soll der Nutzer in die Lage versetzt werden, die für seine Untersuchung benötigten Daten über die Metainformationen zu finden. Beim Herunterladen eines ausgewählten Datensatzes wird letztlich auf die tatsächliche Datei im Dateisystem zurückgegriffen und diese über die Internetverbindung übertragen. Der Nutzer erhält allerdings nicht nur die Datei mit den eigentlichen Daten (z.B. eine Excel-Tabelle), sondern auch alle verfügbaren Metainformationen aus der Metadatenbank in einer separaten Textdatei.

3.2 Datenbankdesign

Die relationale Struktur der gesamten Datenbank „FOR402meta“ ist in Abbildung 2 schematisch dargestellt. Grundlage für die Darstellung ist ein Entity-Relationship-Modell, das alle Gegenstände und Beziehungen der zu modellierenden Realität abstrahiert (Kemper & Eickler 2004). Die Datenbank besteht aus 45 einzelnen Tabellen, die relational miteinander verknüpft sind (s. Abb. 3). Es lassen sich mehrere funktionale Blöcke zusammenfassen, die nachfolgend erläutert werden.

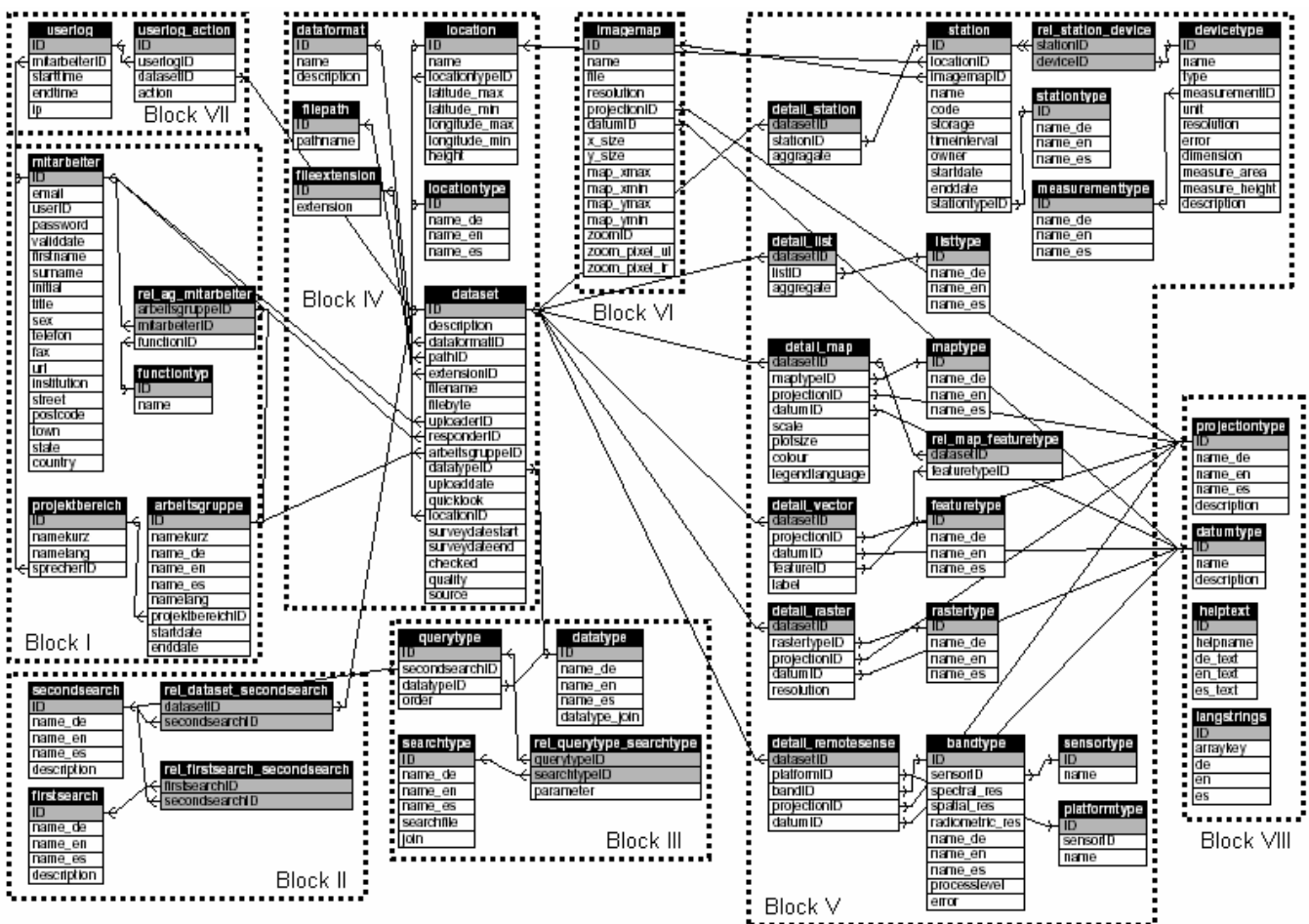


Abbildung 3: Schema der relationalen Datenbankstruktur „FOR402meta“ mit allen Fremdschlüsselbeziehungen (Pfeillinien). Die Tabellennamen sind schwarz hinterlegt, die Primärschlüssel grau, die gepunkteten Linien umschließen funktionale Blöcke.

3.2.1 Datenstruktur für Mitarbeiter, Arbeitsgruppen und Projektbereiche (Block I)

Informationen über jeden einzelnen Mitarbeiter (Name, Anschrift, usw.) und die einzelnen Arbeitsgruppen in den vier Projektbereichen sind in einzelnen Tabellen abgelegt. Die Zugehörigkeit der Mitarbeiter zu den jeweiligen Arbeitsgruppen bzw. Projektbereichen wird ebenso abgebildet wie die Funktion, die ein Mitarbeiter in der jeweiligen Arbeitsgruppe besitzt. In der Mitarbeitertabelle sind zusätzlich auch die relevanten Felder zur Anmeldung an die Datenbank (Benutzername und Passwort) verschlüsselt gespeichert. Die Tabellen „Mitarbeiter“ und „Arbeitsgruppe“ sind mit der Datensatztabelle verknüpft, um die Zugehörigkeit zu dokumentieren.

3.2.2 Datenstruktur für zwei Kategorieebenen (Block II)

Neben der Suche nach Datensätzen aus einer bestimmten Arbeitsgruppe ist auch alternativ die generelle Suche nach Datensätzen innerhalb von hierarchisch angeordneten Kategorien vorgesehen. Hier werden zwei Ebenen angeboten. Jede untere Kategorie kann in verschiedenen Hauptkategorien vorkommen. Jeder einzelne Datensatz ist mit mindestens einer Unterkategorie verbunden, kann aber auch in mehreren Kategorien aufgelistet werden.

3.2.3 Datenstruktur zur Steuerung der Detailinformationen und Auswahlmodule (Block III)

Alle Datensätze werden einem Datentyp zugeordnet. Dieser Datentyp entscheidet darüber, welche zusätzlichen Informationen zu jedem Datensatz aufgenommen werden und welche weiteren Funktionen zur Verfügung stehen. Die erweiterten Suchmöglichkeiten werden je nach Datentyp und Kategorie modular in die generelle Suchseite eingebaut. Alle Suchfunktionen werden nur einmal abgelegt und können mehreren Abfragetypen zugeordnet werden. In den Tabellen werden neben optionalen Parametern auch die für eine Erweiterung der SQL-Anweisungen benötigten Textbausteine gespeichert.

3.2.4 Datenstruktur zur Beschreibung der generellen Datensatzinformationen (Block IV)

Der Kern der Datenbank bildet die Tabelle „Dataset“. Hier werden alle generellen Informationen zu einem Datensatz gespeichert, wie Angaben über den Dateinamen, -größe und -pfad, Datentyp, geographische Lage, Zeitraum, Qualität und Urheber. Es wird auf verschiedene Lookup-Tabellen zurückgegriffen, um gleiche Angaben auszulagern. Die geographische Lage ist ebenfalls in einer gesonderten Tabelle aufgelistet. Die hier eingetragenen Angaben müssen für jeden Datensatz erhoben werden.

3.2.5 Datenstruktur zur Beschreibung der modularen Datensatzinformationen (Block V)

Je nach Datentyp werden weitere Parameter für jeden Datensatz aufgenommen, um verfeinerte Suchmöglichkeiten anzubieten. Für jeden Datentyp wird ein Detailmodul konzi-

piert, das alle relevanten Informationen aufnehmen kann. Folgende Datenarten liegen in Modulen vor, in denen die aufgelisteten Informationen gespeichert werden:

- Fernerkundungsdaten (Informationen zu Plattform, Sensor, detaillierte Bandbeschreibung und kartographische Merkmale)
- Rasterdaten (GIS) (Rastertyp, Auflösung und kartographische Merkmale)
- Vektordaten (GIS) (dargestelltes Merkmal, Beschriftung und kartographische Merkmale)
- Kartenlayout (Kartentyp, Karteninhalt, Größe, Maßstab, Sprache und kartographische Merkmale)
- Ortsgebundene Zeitreihe (Stationsmerkmale wie Stationsart, Ort, Name, Art der Speicherung, Zeitintervall der Messung, Betreiber, Laufzeit der Station und Merkmale aller eingesetzter Untersuchungsverfahren / Messgeräte wie Typ, Messgröße, Einheit, Auflösung, Fehler)
- Sammel Listen (Inhalt, räumliche Aggregation).

Indem alle Detailmodule auf verschiedene Lookup-Tabellen zurückgreifen, werden redundante Einträge in der Datenbank vermieden.

3.2.6 Datenstruktur zur Erzeugung von HTML-Imagemaps (Block VI)

In einer Tabelle werden Informationen über Basiskarten des Untersuchungsgebietes gespeichert, um die Lokalitäten der ortsgebundenen Daten in einer HTML-Imagemap anzeigen zu können. Bei einem Aufruf von Stationsdaten werden dann die jeweiligen Orte mit Hilfe der abgelegten Informationen auf diesen Basiskarten eingezeichnet und dem Nutzer als interaktive HTML-Imagemap angeboten. Eine geographisch ausgerichtete Suche ist damit ebenfalls möglich. Die gespeicherten Informationen in der Tabelle „imagemap“ sind nötig, um die geographischen Koordinaten der Stationen auf die Pixelkoordinaten der Basiskarten zu konvertieren.

3.2.7 Datenstruktur zur Aufzeichnung von Nutzerinformationen (Block VII)

Sobald ein Nutzer die Datenbank aufgerufen hat und sich durch seinen Benutzernamen und sein Passwort verifizieren kann, werden seine Aktionen in der Tabelle „userlog_action“ aufgezeichnet. Dadurch kann im Nachhinein festgestellt werden, welches Mitglied welche Datensätze heruntergeladen hat und eine allgemeine Nutzungsstatistik der Datenbank erzeugt werden.

3.2.8 Datenstruktur weiterer Lookup-Tabellen und Sprachsteuerung (Block VIII)

Für die kartographischen Angaben in einer Vielzahl von Einzeltabellen wurden Lookup-Tabellen mit Informationen zu geodätischem Datum und Projektion ausgelagert. Diese sind in der Regel durch 1:n Verbindungen mit den jeweiligen Tabellen verbunden und vermeiden so eine redundante Datenhaltung.

Da das gesamte Datenbanksystem mehrsprachig aufgebaut wird, sind hier alle vorkommenden Textblöcke in den jeweiligen Sprachen abgelegt. Je nach Sprachauswahl werden die entsprechenden Textblöcke in die HTML-Seite einbezogen.

3.3 Modulare Metadatenstruktur

Es ist zu erwarten, dass im Laufe des Projektes weitere Datentypen in die Datenbank integriert werden müssen. Eine Aufnahme aller relevanten Metadaten in nur einer einzigen Tabelle würde jedoch unweigerlich zu vielen leeren Einträgen in den Datenfeldern führen, weil immer nur ein Teil des Datensatzes logisch zu füllen wäre. Abfragen auf eine derartige Tabelle würden viele leere Ergebnisse ermitteln und eine weitere Verarbeitung erschweren bzw. eine benutzerfreundliche Anzeige nur nicht befriedigen. Eine modulare Ausrichtung der vorliegenden Metadatenbank ist daher aus Sicht der relationalen Datenbankstruktur unumgänglich. Die modulare Ausrichtung sowohl innerhalb der Metadatenstrukturen als auch in der Bereitstellung von detaillierten Suchfunktionen macht eine Erweiterung im vorhandenen Datenbanksaufbau möglich. Desweiteren lassen sich neue modulare Metadatenstrukturen nach Vorgaben aus den einzelnen Arbeitsgruppen entwickeln und in das laufende System integrieren.

Die Kategorien sind ebenfalls im bisherigen System erweiterbar. Durch die Aufnahme von neuen Datensätzen kann es zu einer Ausweisung von neuen Kategorien kommen. Diese werden in die relevanten Tabellen eingefügt und mit den jeweiligen Datensätzen verknüpft. Die detaillierten Suchfunktionen sind jeweils in einzelnen PHP-Skripten abgelegt. In der Datenbank werden die spezifischen Suchmuster den Datentypen zugeordnet. Für neue Suchfunktionen werden die Abfrageformulare in neuen Skripten formuliert und in die Datenbank aufgenommen. Aus diesem Grund lassen sich sowohl völlig neue Datentypen, aber auch einzelne neue Suchfunktionen für bestehende Datentypen ohne großen Aufwand einarbeiten.

Die Darstellung der Ergebnisse in tabellarischer Form wird ebenfalls über die zugeordneten Datentypen eines jeden Datensatzes bestimmt. Je nach Typ werden unterschiedliche Informationen in einer kleinen Zusammenfassung angezeigt. Der generell modulare Aufbau des Systems ist die Grundlage für eine flexible Handhabung der Datenbankstruktur. Dies erweist sich als sehr praktikabel und hält den wachsenden Anforderungen stand. Die Einteilung der Datensätze auf der einen Seite in thematische Kategorien und andererseits in beschreibende Datentypen machen eine detaillierte Suche und eine differenzierte Darstellung möglich.

3.4 Graphische Nutzerschnittstelle

Die graphische Nutzerschnittstelle (vgl. Abbildung 4) begleitet den Nutzer auf der Suche nach relevanten Datensätzen. Sie wird vollständig aus den serverseitigen PHP-Skripten in HTML erstellt. Der Nutzer hat an verschiedenen Stellen über Formulare die Möglichkeit, individuelle Eingaben an den Server und damit an die Datenbank zu übermitteln.

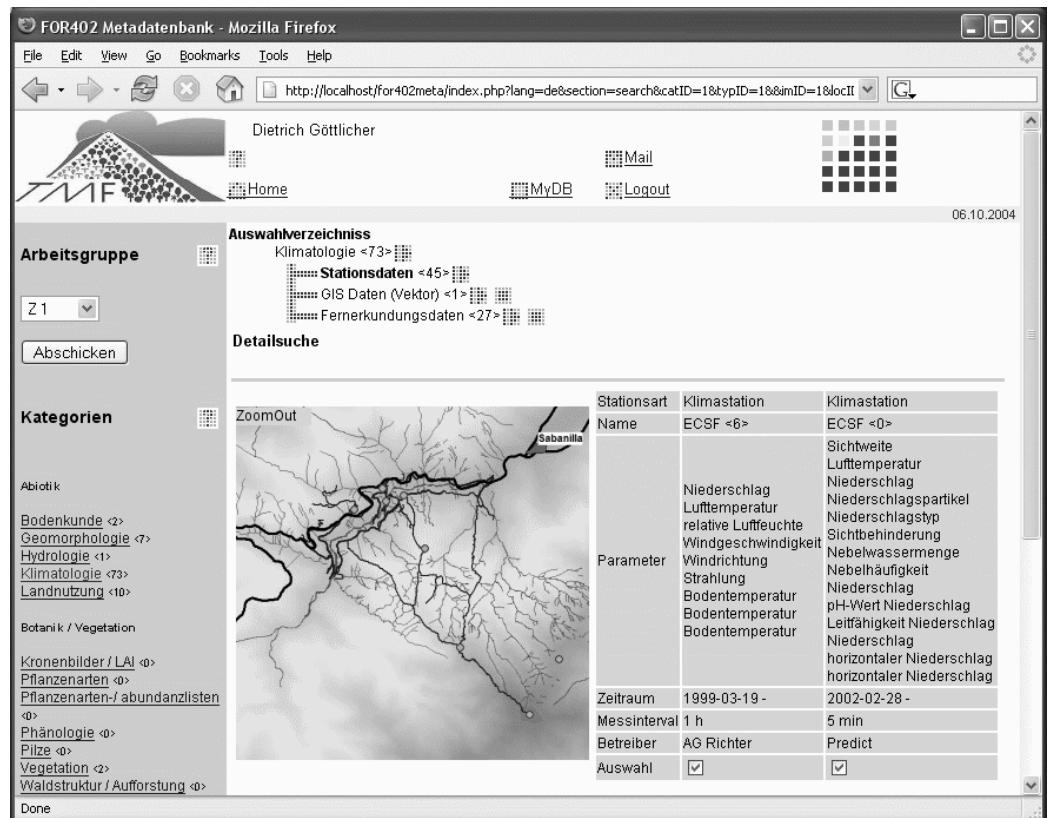


Abbildung 4: Die graphische Nutzerschnittstelle mit dem Auswahlverzeichnis aller Datensätze einer Kategorie und der Detailsuche nach ortsgebundenen Messdaten.

Bei der Suche nach einzelnen Datensätzen kann der Nutzer zwischen einer bestimmten Arbeitsgruppe oder einer thematischen Kategorie wählen. Für die erste Auswahl werden alle vorhandenen Datensätze nach Datentypen geordnet in einem Auswahlverzeichnis angezeigt (Abb. 4). Es lassen sich alle Datensätze einer Arbeitsgruppe bzw. einer Kategorie anzeigen oder aber detailliertere Suchmasken aufrufen, die je nach Datentyp weitere Auswahlkriterien festlegen lassen. Suchabfragen sind in der Regel nach allen in der Metadatenbank aufgenommenen Parametern möglich. Klassische eingrenzende Abfragen beziehen sich auf den Zeitpunkt bzw. den Ort. Zusätzlich kommen in verschiedenen Datentypen spezifische Abfragen vor. Fernerkundungsdaten können nach Sensor und Spektralkanal eingegrenzt werden. GIS-Daten und Kartenlayouts sind nach Karten- und Layerinhalten differenzierbar. Ortsgebundene Daten lassen sich nach verschiedenen Messparametern auswählen.

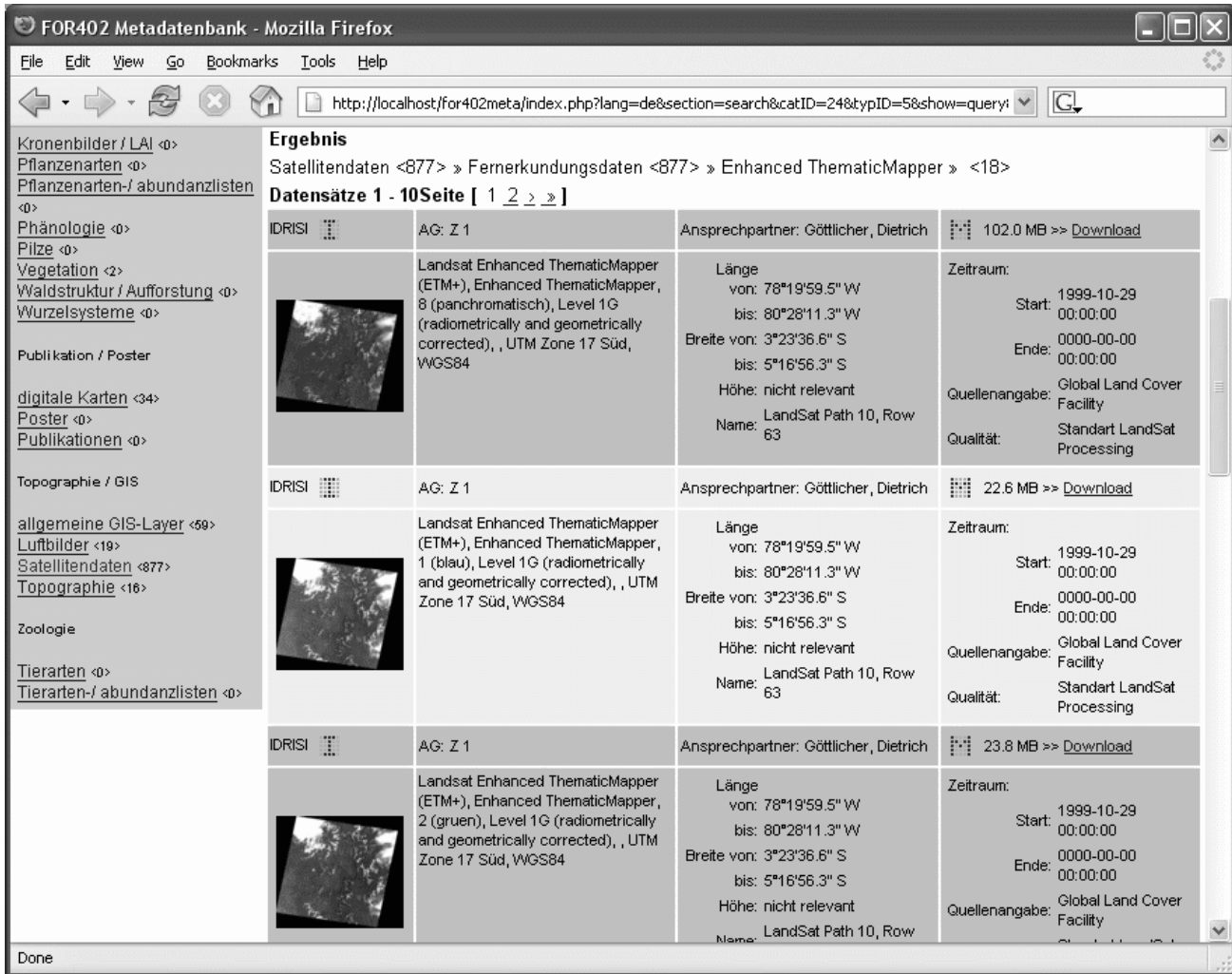


Abbildung 5: Ergebnisanzeige in tabellarischer Form mit Quicklook und Metainformation.

Die Anzeige der ausgewählten Datensätze erfolgt in einer tabellarischen Zusammenfassung (Abb. 5). Bei bildhaften Daten werden neben den wichtigsten Informationen auch eine Miniaturansicht (Quicklook) angezeigt. Zusätzlich kann sich der Nutzer die gesamten Metadaten in einem neuen Fenster komplett auflisten lassen.

Ein Hilfesystem innerhalb der Nutzerschnittstelle erzeugt zu jedem einzelnen Bereich kurze Erläuterungen, sodass sich auch ein ungeübter Benutzer in dem System zurecht finden kann. Desweiteren ist es möglich, seine persönlichen Daten wie Anschrift etc. zu ändern.

4 Praxiseinsatz

Nach einer mehrmonatigen Implementierungsphase wurde die erste Version der Metadatenbank im Sommer 2003 für die Forschergruppe freigeschaltet. Seitdem wurden laufend Erweiterungen und Veränderungen am Datenbanksystem vorgenommen. Inzwischen besteht der Quelltext aus ~5000 Zeilen PHP-Code. Seit Mitte 2004 liegt der Schwerpunkt auf der Integration von prioritären Datensätzen der einzelnen Arbeitsgruppen. Mitarbeiter aller Teilprojekte haben durch die Unterzeichnung der Datenvereinbarung einen Zugang zur Metadatenbank erhalten. Die letzten Monate zeigen einen verstärkten Zugriff auf das System und vermehrtes Herunterladen einzelner Datensätze.

Die ersten in die Metadatenbank aufgenommenen Datensätze beschränken sich zum größten Teil auf die Themengebiete Klimatologie, allgemeine GIS-Daten und vorgefertigte Karten bzw. Luftbilder. Dies sind allerdings alles Datensätze, die von einem großen Teil der Arbeitsgruppen für ihre eigenen Forschungen bzw. zur Darstellung von Ergebnissen benötigt werden. Insgesamt sind mittlerweile über 1000 einzelne Datensätze in die Datenbank aufgenommen worden.

Als praktisches Beispiel zur Nutzung der Metadatenbank sei exemplarisch die Ermittlung eines flächendeckenden Blattflächenindex (LAI) angeführt, der als Basis für eine SVAT-Modellierung im Untersuchungsgebiet benötigt wird. Die Ableitung des LAI soll über verschiedene Fernerkundungsdaten operationalisiert werden. Dazu erfolgt in der Datenbank die Suche nach geeigneten Satellitendaten und Luftbildern, die von verschiedenen Arbeitsgruppen zur Verfügung gestellt werden. Ebenfalls kann aus der Datenbank ein digitales Geländemodell bezogen werden, das zur Kalibrierung der Satellitendaten (Geländekorrektur) benötigt wird. Zur Abgrenzung und Einordnung einzelner Vegetationseinheiten in den Fernerkundungsdaten liegen punktuelle Beschreibungen (Vegetationsstruktur und pflanzensoziologische Aufnahmen) in tabellarischer Form vor. In verschiedenen Arbeitsgruppen vektorisierte GIS-Layer (z.B. Detailkartierungen der Vegetationseinheiten) kommen ebenfalls in Betracht. Diese räumlichen Informationen wird

mit den Fernerkundungsdaten im Rahmen einer Landbedeckungsklassifikation verschnitten. Aus den ebenfalls in der Datenbank vorhandenen Messdaten des LAI verschiedener botanischer Arbeitsgruppen werden für die einzelnen Landbedeckungseinheiten jeweils spezifische Transferfunktionen zwischen LAI und aus Satellitendaten ermittelten NDVI-Werten (Vegetationsindex) abgeleitet.

Grundsätzlich können für alle interdisziplinären Fragestellungen die jeweils benötigten Daten in der Metadatenbank gesucht und gefunden werden. Zeitliche und räumliche Abfragen beschleunigen den Auswahlprozess. Alle mitgelieferten Metainformationen über Ort, Zeit, Aufnahmemethode und Qualität helfen, die heruntergeladenen Daten raumzeitlich korrekt einzuordnen, inhaltlich zu verstehen und zielgerichtet einzusetzen. Die Angabe einer verantwortlichen Person für jeden Datensatz ermöglicht darüber hinaus eine schnelle Kontaktaufnahme, um ggf. auftretende Fragen auszuräumen.

Letztlich wird der Erfolg und die Nutzungsintensität der Metadatenbank an ihrem inhaltlichen Angebot gemessen.

5 Schlussfolgerung

Bei der Konzeption und Implementierung einer Datenbank zur Unterstützung für eine interdisziplinär arbeitende Forschergruppe sind immanente Schwierigkeiten zu überwinden. Die Heterogenität der erhobenen Datensätze macht eine ausgereifte Datenbankstruktur notwendig, wenn detaillierte Suchmöglichkeiten zu den einzelnen Datensätzen angeboten werden sollen. Das hier vorgestellte Metadatenbanksystem zeigt modulare Strukturen in verschiedenen Bereichen. Dadurch lassen sich für die unterschiedlichen Datentypen individuelle Abfrage- und Anzeigemöglichkeiten realisieren. Dieser Ansatz ermöglicht auch die laufende Erweiterung der Datenbank für neue Datentypen.

Nur eine vorausschauende Konzeption und Strukturierung der Datenbank ermöglicht die unproblematische Erweiterung des Systems. Sinnvoll wäre es sicherlich, direkt zu Beginn der Forschungsaktivitäten mit dem Aufbau der Datenbank unter Verwendung eines integrierten Datenmodells zu beginnen, damit alle Daten von Anfang an in die Datenbank aufgenommen werden können. Die Rekonstruktion „alter“ Datensätze trifft häufig auf Schwierigkeiten, da die Bearbeiter nicht mehr zu Verfügung stehen um die notwendigen Metadaten bereitzustellen. Die Bereitstellung der Daten setzt die Mitarbeit der einzelnen Arbeitsgruppen voraus. Eine Regelung über eine projektinterne Datenvereinbarung hat sich bewährt.

Das Konzept hat sich nach der Implementierungsphase als sehr tragfähig erwiesen (hohe Zugriffszahlen). Der Datenaustausch ist deutlich intensiviert worden und die vornehmlich bilateralen Arbeiten der ersten Phase werden zunehmend durch multilaterale Ansätze ergänzt. Dabei hat sich gezeigt, dass manchmal eine Nachbearbeitung der Daten bezogen auf die äußerst spezifischen Anforderungen der einzelner Vorhaben notwendig ist und auch nicht durch ein einheitliches Datenmodell abzufangen wäre. Gerade im Modellierungsbereich ergeben sich äußerst spezifische Eingabeformate, die bedient werden müssen. Die Arbeiten laufen in den betroffenen Gruppen aber sehr effektiv, für bestimmte Fälle bietet das Datenbankprojekt auch eine Konvertierungshilfe (z.B. zwischen generischen GIS-/DBV-Formaten) an.

Der Einsatz von freier Software hat sich auch für ein Projekt dieser Größenordnung als unproblematisch erwiesen. Durch die aktive Nutzergemeinschaft der diversen Open Source Programme werden dem Entwickler bei auftretenden Problemen auf Nachfrage zeitnah Lösungen vorgeschlagen.

Danksagung

Die Autoren danken der Deutschen Forschungsgemeinschaft für die großzügige Förderung des Projekts Z1 im Rahmen der Forschergruppe 402 (BE 1780/9-1). Nature and Culture International (NCI, www.natureandculture.org) sei für die logistische Unterstützung vor Ort gedankt, ebenso wie den Mitgliedern der Forschergruppe, ohne deren konstruktive Diskussionsbeiträge z.B. im Rahmen des Marburger Datenbankworkshops und ohne deren Datenlieferungen ein solches Projekt nicht möglich wäre.

6 Literatur

- BECK, E., MÜLLER-HOHENSTEIN, K. (2001): Analysis of undisturbed and disturbed tropical mountain forest ecosystems in Southern Ecuador. *Die Erde* 132, S. 1-8.
- BENDIX, J., ROLLENBECK, R., PALACIOS, W.E. (2004): Cloud detection in the Tropics – a suitable tool for climate-ecological studies in the high mountains of Ecuador. *Int. J. Remote Sensing* 10, S. 4521-4540.
- BOJINSKI, S., SCHAEPMANN, M., SCHÄPFER, D. & ITTEN, K. (2003): SPECCHIO: a database for remote sensing applications. *Computers & Geosciences* 29, S. 27-38.
- DALITZ, H. & HOMEIER, J. (2004): Visual Plants - Image based tool for plant diversity research. *Lyonia* 6 (1), S. 49-62.
- DENZER, R. und GÜTTLER, R. (1997): Rolle von Metainformationen in Umweltinformationssystemen – Konzept für Metainformationssysteme. In: Arndt, H.-K., Günther, O., Hilty, L. M., Rautenstrauch, C. (Hg.): *Metainformation und Datenintegration in betrieblichen Umweltinformationssystemen (BUIIS)*, *Umweltinformatik aktuell*, 14, S. 25-37, Marburg.
- DIEPENBROEK, M., GROBE, H., REINKE, M. (1997): PANGAEA – Ein Informationssystem für die Meeresforschung. *Deutsche Gesellschaft für Meeresforschung Mitteilungen* 1-2, S. 94-96.
- HOSENFELD, F. (1999): Integration heterogener ökologischer Informationen und deren Präsentation im WWW. In: Kramer, R. & Hosenfeld, F. (Hg.): *Heterogene, aktive Umweltdatenbanken*, *Praxis der Umwelt-Informatik*, 8, S. 53-72, Marburg.
- KEMPER, A., EIKLER, A. (2004): *Datenbanksysteme*. Oldenbourg, München.
- KOFLER, M. (2003): *MySQL – Einführung, Anwendung, Referenz*. Addison-Wesley, München.
- KRASEMANN, H. L. (1996): *Meta-Information – Was ist das?*. In: Kremers, H. & Krasemann, H. L. (Hg.): *Umweltdaten verstehen durch Metainformation*, *Praxis der Umwelt-Informatik*, 6, S. 9-14, Marburg.
- MÜCKSCHEL, C., NIESCHULZE, J., SCHACHTEL, G. A., LI, S., SLOBODA, B., KÖHLER, W. (2004): *Web-basierte Informationssysteme in interdisziplinären Umwelt-Forschungsprojekten – am Beispiel der beiden DFG-Sonderforschungsbereiche 299 (Giessen) und 552 (Göttingen)*.

tingen/Kassel). Zeitschrift für Agrarinformatik, 3 (12), S. 46-55.

SOMMERVILLE, I. (2001): Software Engineering. Pearson Studium, München.

TOUSSAINT, F., LAUTENSCHLAGER, M., REINKE, M. (1999): CERA-2 – Ein raumbezogenes Daten- und Metadatenmodell. In: Kramer, R. & Hosenfeld, F. (Hg.): Heterogene, aktive Umweltdatenbanken, Praxis der Umwelt-Informatik, 8, S. 107-120, Marburg.

Kurzfassung

Für eine multidisziplinär arbeitende ökologische Forschergruppe mit mehr als 30 Arbeitsgruppen wird eine Datenbank konzipiert, um allen Mitgliedern Zugang zu den in den einzelnen Teilprojekten erhobenen Datensätzen auch in aggregierter Form zu ermöglichen. Anforderungen an das Datenbanksystem sind u.a. die Sicherheit des Datenbestands und eine durchgängige Verfügbarkeit über ein nutzerfreundliches Web-Interface. Die Implementierung erfolgt kosteneffektiv mit Open-Source Software. Das entwickelte System basiert auf einer klassischen Client-Server-Architektur, wobei die relationale Datenbank in über 40 einzelne Tabellen aufgeteilt ist. Da die Datensätze aus verschiedenen Disziplinen (z.B. Klimatologie, Pflanzenökologie) mit unterschiedlicher Datenorganisation stammen, werden für die Beschreibung der einzelnen Datentypen modulare Einheiten vorgesehen. Die Organisation in modulare Einheiten ist absolute Voraussetzung für die flexible Handhabung des Gesamtsystems.

Stichworte: Metadatenbank, Datenbankkonzeption, Forschergruppe, MySQL, PHP

Summary

A database is designed for a multidisciplinary ecology research group consisting of more than 30 individual teams. Its main purpose is to provide access to all data raised in the sub-projects in an aggregated form. Data security and 24-hour accessibility via a user-friendly web interface are among the core requirements on the database. The implementation is based on cost-effective open source software. It is designed in a classical client-server architecture with a relational database of more than 40 individual tables. Since the data records originate from different disciplines (e.g. climatology, plant ecology) with different data organization, the descriptions of the individual data types are stored in modular units. This modularity is a critical precondition for the flexible handling of the overall system.

Keywords: Metadatabase, database design, research group, MySQL, PHP

Autoren

Dietrich Göttlicher, Dipl.-Geograph

Fachbereich Geographie, Laboratory for Climatology and Remote Sensing (LCRS)

Philipps-Universität Marburg

Deutschhausstr. 10

35032 Marburg

Fon: +49 (0)6421/28 25910, Fax.: +49 (0)6421/28 28950

Email: dietrich.goettlicher@staff.uni-marburg.de

Prof. Dr. Jörg Bendix

Fachbereich Geographie, Laboratory for Climatology and Remote Sensing (LCRS)

Philipps-Universität Marburg

Deutschhausstr. 10

35032 Marburg

Fon: +49 (0)6421/28 24266, Fax.: +49 (0)6421/28 28950

Email: bendix@staff.uni-marburg.de