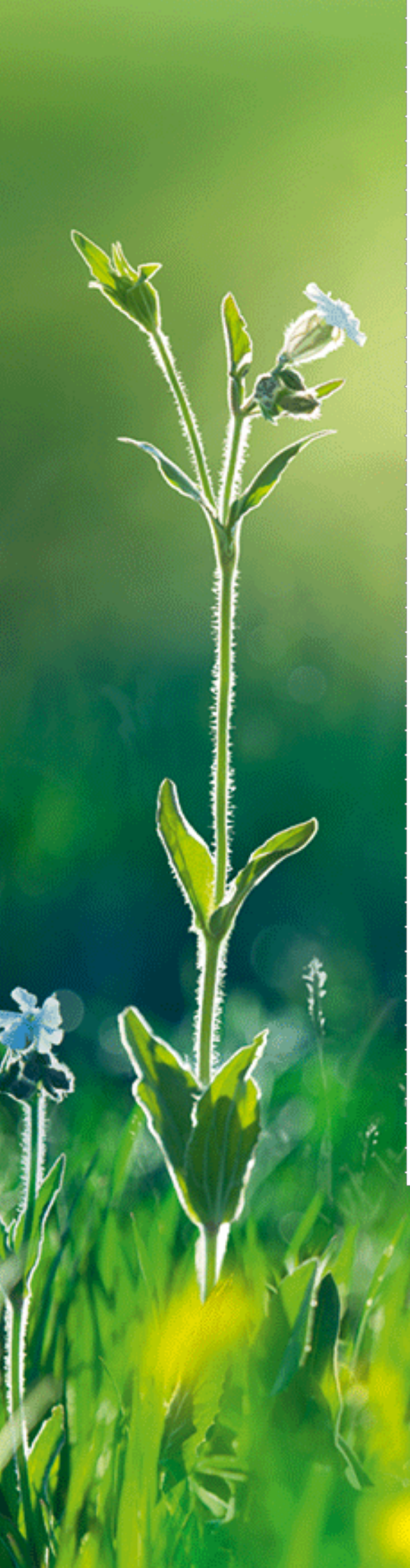

MINISTERIUM
FÜR EIN
LEBENSWERTES
ÖSTERREICH

HBLFA RAUMBERG - GUMPENSTEIN
LANDWIRTSCHAFT

Abschlussbericht
DATALYS

Projekt Nr. 100852

Entwicklung eines Prüf- und Analysesystems für kontinuierliche Datenströme sowie deren Überführung in relationale Datenbanken

Development of a system for test and analysis of continuous data streams and their transformation into relational databases

Projektleitung:

Mag. Dr. Andreas Schaumberger, MSc
HBLFA Raumberg-Gumpenstein

Projektmitarbeiter:

Manuel Adelwöhrer, HBLFA Raumberg-Gumpenstein
Martina Schink, HBLFA Raumberg-Gumpenstein

Projektlaufzeit:

2012 – 2015

raumberg-gumpenstein.at

Impressum

Herausgeber
Höhere Bundeslehr- und Forschungsanstalt für Landwirtschaft
Raumberg-Gumpenstein, A-8952 Irdning
des Bundesministeriums für Land- und
Forstwirtschaft, Umwelt und Wasserwirtschaft

Für den Inhalt verantwortlich
Andreas Schaumberger

Druck, Verlag und © 2016
Höhere Bundeslehr- und Forschungsanstalt für Landwirtschaft
Raumberg-Gumpenstein, A-8952 Irdning

Dieses Forschungsprojekt wird vom Bundesministerium für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft finanziert.

Inhaltsverzeichnis

Zusammenfassung

Im Rahmen des Projektes DATALYS wurde eine Anwendung entwickelt, die geeignet ist, Daten aus verschiedenen Quellen wissenschaftlicher Experimente (kontinuierliche Datenströme aus Sensoren, manuelle Eingabe, Laborergebnisse, Computermodelle, usw.) in eine Datenbank mit relationalem Datenmodell zu überführen. Das Datenmodell kann beliebig erweitert werden und passt sich so unterschiedlichen Fragestellungen bzw. Anwendungsbereichen ohne Mehraufwand an. Die Flexibilität des Datenmodells erlaubt eine interdisziplinäre Anwendung sowie eine optimale Anpassung an Veränderungen und Erweiterungen der Forschungsschwerpunkte. Im Rahmen eines Datenbankmanagementsystems (DBMS) wird die Funktionalität des Datenmodells mit den Standardfunktionen wie Leistungsfähigkeit, Zugriffsmanagement, Datensicherheit und Qualitätssicherung kombiniert und bietet auf diese Weise die Grundlage für eine effiziente wissenschaftliche Auswertung der erfassten Daten. Neben einer automatisierten Migration von Sensordaten ermöglicht eine Benutzerschnittstelle über Web-Browser die benutzergesteuerte Datenein- und Datenausgabe. Diese Datenbankschnittstelle ist gleichzeitig die Grundlage für die Zusammenarbeit von Projektpartnern, welche Daten verteilt in das System einspeisen und zentral abfragen können. Werden Daten aus unterschiedlichen Disziplinen in ein gemeinsames System integriert, ergibt sich daraus die Möglichkeit, mit geeigneten Analysen und Synthesen explorativ an neuen Fragestellungen zu arbeiten, auch über einzelne Projektlaufzeiten hinaus. Das System bietet zusätzlich die Möglichkeit, Zeitreihendaten durch entsprechende Visualisierung zu kontrollieren und Probleme wie Datenlücken oder Ausreißer zu beseitigen.

Summary

In DATALYS we developed an application to migrate data from different sources of scientific experiments (continuous data streams from sensors, manual input, laboratory data, computer models) into a database with relational data model. The data model can be extended for any purposes and adjusted to different scientific issues and scopes without any additional effort. The flexibility of the developed model supports interdisciplinary approaches as well as optimal adjustments to changes and extensions of scientific priorities. Within a database management System (DBMS) the functions of our data model is combined with standard DBMS functions like high performance, access management, security and quality management as the background for efficient scientific analysis of all stored data. Besides the computer-aided migration of sensor data, a Web Browser user interface supports the manual input and query of data. This database interface is also the platform of co-operation between project partners, who feed the system in a distributed way and query it centrally. If data of different disciplines are integrated in one system, there is a big chance to discover new scientific issues with explorative analysis and synthesis even beyond project life times. The system additionally supports visually data control mechanism to find and solve problems with missing data and outliers.

1 Einleitung

Die wichtigste Grundlage für die Bearbeitung wissenschaftlicher Fragestellungen sind Daten (Brunt, 2000). Sie stehen am Beginn eines Prozesses, in dem mittels Selektion, Transformation und anderer Formen der Verarbeitung Information generiert wird, die bei einer entsprechenden Analyse und Interpretation zu neuem Wissen führt (Fayyad *et al.*, 1996a).

Bedingt durch den technologischen Fortschritt der letzten Jahrzehnte ist die Anzahl digitaler Daten in allen Bereichen des Lebens nahezu explodiert. Insbesondere in den Naturwissenschaften ist die Sammlung und Auswertung riesiger Datenmengen die Voraussetzung für das Verständnis komplexer Zusammenhänge. Umweltparameter werden dabei oft kontinuierlich mit Sensorsystemen beobachtet und führen zu massiven Datenströmen, welche permanent von Sensoren zu physikalischen Speicherstrukturen fließen (Deelman und Chervenak, 2008). Eine weitere Quelle für die Entstehung enormer Datenmengen in der Wissenschaft sind Modelle. Mit geeigneter Software und leistungsfähiger Hardware werden meist auf der Basis realer Daten Szenarien und Projektionen gerechnet, welche die Menge des Originaldatenbestandes um ein vielfaches übertreffen können. Ein Beispiel dafür ist die Klimamodellierung.

Im Rahmen der Versuchsanlage ClimGrass an der HBLFA Raumberg-Gumpenstein werden die Auswirkungen von Klimaveränderungen (Temperaturzunahme, erhöhte CO_2-Konzentration in der Atmosphäre und Wasserstress) auf einen Grünlandmischbestand untersucht (vergleiche Projekt ClimGrassEco). Zahlreiche Sensoren (Wetterstationen, Lysimeter, usw.) erfassen die verschiedensten Umweltbedingungen in hoher zeitlicher Auflösung. Dabei entstehen viele Daten, die zum Zweck einer sinnvollen, interdisziplinären Auswertung zu einem zentralen Datenbestand zusammengeführt werden müssen. Neben den Sensordaten fließen manuell generierte Daten wie die Ergebnisse aus Versuchsernten, Bodenprobennahmen, nicht-invasiven Pflanzenbestandsbeobachtungen, usw. ein. Wie bei allen wissenschaftlichen Experimenten bilden auch hier Daten die Schnittstelle zwischen der Versuchsanlage mit all den dort stattfindenden Aktivitäten und den daraus gewonnenen wissenschaftlichen Erkenntnissen. Das hier vorgestellte Datenmodell basiert auf einer mehrjährigen, iterativen Entwicklung und folgt den in der Literatur empfohlenen Richtlinien zur Konzeption von „Scientific Databases“. Während der Entwicklungsphase konnten Erfahrungen gesammelt werden, die für ein funktionierendes Zusammenspiel von Datenerfassung, -verarbeitung und -nutzung von großer Bedeutung sind. Neben einer allgemeinen Beschreibung der Anforderungen für das Datenmanagement wissenschaftlicher Anwendungen erfolgt in diesem Bericht auch eine detaillierte Vorstellung des Datenmodells und der dafür implementierten Benutzerschnittstelle zur Ein- und Ausgabe der Daten.

Da Rohdaten für wissenschaftliche Auswertungen nur bedingt geeignet sind, muss zumindest eine Qualitätsprüfung sicherstellen, dass Eingabefehler oder Sensorprobleme keine negativen Auswirkungen auf die weitere Analyse der Daten haben. Fehler in den Daten dürfen die Aussagekraft wissenschaftlicher Auswertungen keinesfalls beeinträchtigen. Während bei manuell eingegebenen Daten die Prüfung aufgrund der geringen Datenmenge noch relativ einfach zu bewältigen ist, steigt die Komplexität von Prüfalgorithmen mit der Anzahl von Daten deutlich an, da neben der Fehlerlokalisierung auch die Fehlerbehebung im Sinne einer Versionierung realisiert werden muss. Manuelles Eingreifen wird mit zunehmender Datenmenge immer schwieriger und zeitaufwendiger. Deshalb muss die Analyse und Prüfung von Datenströmen zwangsläufig mit Hilfe automatischer bzw. halbautomatischer Prozesse vorgenommen werden.

2 Datenmanagement in der Wissenschaft

2.1 Die Rolle des Datenmanagements in der Forschung

Bei der Abwicklung von Forschungsprojekten konzentrieren sich viele der beteiligten WissenschaftlerInnen in erster Linie auf fachliche Aspekte. Das Datenmanagement steht zu Beginn der Arbeit meist nicht im Fokus des Interesses, da noch keine Daten vorliegen und deshalb auch nicht die dringliche Notwendigkeit besteht, sich mit Strukturen und Arbeitsabläufen auseinanderzusetzen, welche eine effiziente Auswertung betreffen. Wie die Erfahrung zeigt, ist dies meist der Start einer provisorischen Datenhaltung, die dann aus Zeit- und Ressourcenmangel bis zum Projektende beibehalten wird. Oft fehlt auch das entsprechende Know-how zur Bedienung von Datenbankmanagementsystemen sowie für die Entwicklung von komplexen Datenmodellen. Solange die Datenmenge nicht allzu groß ist, werden die Daten dann in der Praxis gern in vielen kleinen Dateien gehalten und mit völlig ungeeigneter Software verwaltet (Hunt *et al.*, 2001). Vielfach kommt hier Microsoft Excel zum Einsatz, das neben der Datenanalyse oft auch zur Datenhaltung missbraucht wird. Fehler und Datenverluste aufgrund mangelnder Übersichtlichkeit und eingeschränkter Abfragemöglichkeiten sind damit vorprogrammiert.

Falls im Laufe der Projektarbeit erkannt wird, dass die Kapazitäten des „Provisoriums“ nicht ausreichen und eine nachträgliche Datenmodellierung unumgänglich ist, können massive Probleme entstehen. Das Datenmanagement greift gewöhnlich tief in die täglichen Arbeitsabläufe ein und alle damit verbundenen Änderungen stellen eine Herausforderung dar. Wenn sich beispielsweise die anfangs geplanten Datenstrukturen für eine Auswertung als unbrauchbar erweisen, die Strukturen zu erweitern sind oder die Daten mit Partnern geteilt werden müssen, braucht es eine intensive Phase der Reorganisation, die auch mit hohen Kosten verbunden sein kann (Van den Eynden *et al.*, 2011).

Bei Anträgen zu größeren Forschungsprojekten ist es mittlerweile notwendig, einen Datenmanagementplan sowie ein Konzept für eine gemeinsame Datennutzung vorzulegen (vgl. zum Beispiel die Anforderungen an das Datenmanagement beim EU Framework Programme for Research and Innovation Horizon 2020). Lediglich kleinere Projektförderschienen verzichten zurzeit noch auf die Einforderung expliziter Vorschläge zum Datenmanagement und zur Datennutzung (Diekmann, 2012). Das Ziel obligatorischer Datenmanagementpläne ist es, den Lebenszyklus wissenschaftlicher Daten zu verlängern und deren Bestand und interdisziplinäre Nutzung über die Projektdauer hinaus zu gewährleisten. Dabei spielt der Austausch von Daten zwischen Personen oder Institutionen eine herausragende Rolle, denn nur standardisierte Datenstrukturen mit entsprechenden Metadaten können von Dritten sinnvoll genutzt werden.

Datenmanagement beginnt bereits bei der Konzeption eines Forschungsprojektes, erfährt seine intensivste Nutzungsphase während der Sammlung und Analyse von Daten und dient anschließend als Grundlage für Veröffentlichungen und für eine gemeinsame Nutzung (data sharing), auch über die Projektlaufzeit hinaus. Je komplexer und umfangreicher Forschungsprojekte konzipiert sind, desto größer ist der Bedarf an einem zentral organisierten Datenmanagementsystem, um die vorgenannten Aufgaben auch effizient durchführen zu können (Brunt, 2000). Hine (2006) bezeichnet Datenbanken in diesem Zusammenhang auch als „scientific instruments“.

Datenbankmanagementsysteme (DBMS) spielen eine herausragende Rolle bei der Bewältigung umfangreicher Daten mit heterogenen Strukturen. Sie bieten gegenüber dateibasier-

tem Management große Vorteile in der Administration, dem Datenaustausch, der Zusammenstellung für Auswertungen und in der Qualitätssicherung. In der vorliegenden Arbeit wird das Datenmanagement ausschließlich im Zusammenspiel mit DBMS betrachtet, da gerade in komplexen Forschungsprojekten jede andere Datenverwaltung ungeeignet ist. Neben einer langfristigen Archivierung der Daten spielen DBMS vor allem auch bei der kurzfristigen Unterstützung von Datensicherheit eine wichtige Rolle, da Benutzerzugriffsverwaltung, Backuproutinen und weitere Sicherungsmaßnahmen meist voll in das Datenmanagement integriert sind. Zusätzlich wird in diesen Systemen eine laufende Prüfung der Datenintegrität vorgenommen, sodass ein Mindestmaß an Datenqualität gewährleistet wird, welche bei einer unkoordinierten Datenhaltung nur sehr schwer erreicht werden kann.

Die Nutzung der Daten während eines Forschungsprojektes, insbesondere der Austausch innerhalb der beteiligten Forscher, wird durch ein entsprechendes Datenmanagement nicht nur unterstützt, sondern stellt vielfach die Basis für die gemeinsame Arbeit dar. Mit möglichst einfachen und weitgehend systemunabhängigen Schnittstellen für den Datenzugriff, wie sie beispielsweise für Internet-Browser-Seiten programmiert werden können, ist die Bedienung und Administration einer gemeinsamen Datenbank für alle Beteiligten mit geringem Einarbeitungsaufwand machbar. Individuelle und auf das Projekt abgestimmte Abfrage- und Eingabeformulare als Schnittstellen zur Datenbank erleichtern den Zugang beteiligter WissenschaftlerInnen zur vergleichsweise anspruchsvollen DBMS-Technologie. Die administrative Arbeit bleibt dabei auf wenige Administratoren beschränkt, die dafür aber auch das notwendige Know-how hinsichtlich Programmierung und Datenmodellierung mitbringen müssen.

2.2 Wissenschaftliche Datenbanken und Datenmodellierung

Bei Forschungsprojekten mit interdisziplinärem Ansatz ist die Verfügbarkeit von Daten aus allen in einem Projekt beteiligten Fachbereichen das entscheidende Kriterium für eine explorative Analyse, die als Voraussetzung für fachübergreifende Erkenntnisse gilt. Die Heterogenität der Anforderungen verschiedener Disziplinen an das Datenmanagement ist dabei eine große Herausforderung, die in erster Linie mittels geeigneter Datenmodelle bewältigt werden muss. Vor allem die Agrarforschung ist durch Multidisziplinarität und der dadurch notwendigen Integration heterogener Datenbestände gekennzeichnet (Diekmann, 2012). Die Forderung nach uneingeschränkter Skalierbarkeit und Anpassungsfähigkeit der Strukturen in wissenschaftlichen Datenbanken ist bei der Datenmodellierung, insbesondere in der Agrarforschung, unbedingt zu berücksichtigen, da spätere Anpassungen in der Regel enormen Aufwand verursachen. Eine Datenbank für wissenschaftliche Anwendungen soll die Integration von Datenstrukturen unterschiedlichster Disziplinen unterstützen und die Forderung nach fachübergreifender Auswertung bedienen können. Anwender aus der Wissenschaft haben unterschiedliche Anforderungen an das Datenmanagement, die meist davon abhängen, wie groß die Datenmenge eines Projektes und deren Diversität bzw. Komplexität ist. Liegt der Fokus auf einer langfristigen Verfügbarkeit, bei denen aktuelle Anforderungen bzw. persönliche Präferenzen eine untergeordnete Rolle spielen, führt kaum ein Weg an der Verwendung einer Datenbank vorbei, aus der sich nach Porter (2000)) folgende Vorteile ergeben:

- Datenbanken führen zu einer signifikanten Verbesserung der Datenqualität.
- Datenbanken gewährleisten eine langfristige und fachübergreifende Datenverfügbarkeit und können in manchen Fällen zu enormen Einsparungen führen, da die Speicherung von Daten wesentlich geringere Kosten verursacht als neuerliche Datenerhebungen.

Vor allem bei Langzeitstudien und interdisziplinären Projekten fallen Daten an, die über den eigentlichen Projektzweck selbst einer breiteren Nutzung zugänglich gemacht werden könnten. Allein die Verfügbarkeit wissenschaftlicher Daten kann zu neuen Fragestellungen anregen und in Kombination mit eigenen Datenerhebungen Antworten bei vergleichsweise geringen Kosten liefern (Porter, 2000). Voraussetzung für diese Nutzung ist allerdings die Strukturierung nach allgemeinen Standards (z. B. in Form eines relationalen Datenmodells mit referentieller Integrität) sowie die Integration von Metadaten. Ohne Beschreibung, zumindest hinsichtlich Erhebungsmethoden und weiterer Verarbeitungsschritte, sind Rohdaten für eine wissenschaftliche Nutzung nicht geeignet.

Da die zeitlichen und finanziellen Ressourcen eines Forschungsprojektes in der Regel limitiert sind, wird sich ein systematisches Datenmanagement dann durchsetzen, wenn es möglichst einfach zu bedienen ist und auf spezielle Anforderungen flexibel reagieren kann.

Ein erfolgreiches Datenmanagement reflektiert in erster Linie die konzeptionelle Auseinandersetzung mit Forschungsfragen, insbesondere innerhalb eines Teams, und ist nur in weiterer Folge das Ergebnis einer technologischen Umsetzung. Die Entwicklung derartiger Systeme sollte deshalb aus der Perspektive der Anwender, in diesem Fall der WissenschaftlerInnen, erfolgen (Brunt, 2000). Es handelt sich um einen evolutionären Prozess, bei dem es nicht Ziel ist, in einem ersten Wurf bereits die „ultimative Datenbank“ zu erstellen (Porter, 2000). Meist entwickeln sich Benutzeranforderungen mit Zunahme des Wissens um die Möglichkeiten, die ein Datenbankmanagementsystem bietet. Im Laufe des Projektfortschrittes werden die Anforderungen meist mehr und vor allem immer konkreter. Evolutionäre Entwicklung und eine Ausdehnung des Anforderungsspektrums setzt ein Datenmodell voraus, welches Anpassungen ohne Mehraufwand unterstützt.

Professionelles Management wissenschaftlicher Daten erfährt gerade im Zuge einer umfassenden Technologisierung bei der Datenerhebung zunehmend an Bedeutung (Borgman *et al.*, 2006). Höhere Präzision und immer kleinere Zeitintervalle bei der Erfassung der Umwelt über Sensoren führen zu riesigen Datenmengen, welche nur mit anwendungsspezifischen DBMS-Adaptionen effizient bewältigt werden können. Dazu kommt die Generierung von Daten aus zunehmend komplexer werdenden Simulationsmodellen. Aus diesen Gründen umfasst zeitgemäßes Datenmanagement neben der eigentlichen Datenverwaltung auch die technische und organisatorische Unterstützung des gesamten Workflows (Ailamaki *et al.*, 2010). Die technischen Fähigkeiten vieler WissenschaftlerInnen als Voraussetzung einer adäquaten Datenanalyse können mit den rasant wachsenden Datenmengen in vielen Fällen nicht Schritt halten. Eine explorative Analyse mit einer effektiven Datenmanipulation stellt damit die größte Hürde für eine umfassenden Nutzung großer Datenbestände dar (Fayyad *et al.*, 1996b). Eine bestmögliche Unterstützung in allen Bereichen des Workflows wird zu einer Schlüsselfunktion bei der Generierung neuen Wissens. Ausgewählte Schritte dieses Workflows sind beispielsweise:

- Datenerfassung mittels Sensoren, Beobachtungen mit manuellen Aufzeichnungen, Laboruntersuchungen, Computersimulationsmodelle, usw.
- Interpretation und Transformation von Rohdaten: Bei Sensordaten müssen beispielsweise elektrische Signale in Zieleinheiten umgewandelt werden.
- Automatische Migration von Daten aus Datenloggern bzw. manuellen Aufzeichnungen in eine Datenbank unter Berücksichtigung der Strukturen des Datenmodells.
- Datenkontrolle und Qualitätssicherung.
- Zugriff auf Daten mittels Abfragen und deren Ausgabe in Standardformaten.
- Analyse und Synthese der Daten sowie deren Interpretation.

Neben den Daten selbst ist die Erfassung von Metadaten zu jedem Verarbeitungsschritt eine Voraussetzung für die fachübergreifende bzw. langfristige Nutzung. Daten finden nur dann Wiederverwendung in anderen Projekten, wenn sie anhand der Metadaten vollständig verstanden werden können und sämtliche Manipulationen an den Rohdaten dokumentiert wurden (Arzberger *et al.*, 2004, Heidorn, 2009).

2.3 Implementierung eines Datenmanagementsystems

Das im Rahmen von DATALYS implementierte Managementsystem wurde für alle im Experiment ClimGrass anfallenden Daten entwickelt. Die Arbeiten an diesem System haben parallel mit der Installation der Versuchsanlage begonnen, also zu einem Zeitpunkt, wo die konkreten Anforderungen noch bei weitem nicht klar definiert waren. Im Zuge des Aufbaus am Feld, der Installation von sensorbestückten Messinstrumenten, wie beispielsweise der Wiegelysimeter sowie der Instrumente für Beheizung und CO_2-Begasung wurde der Bedarf an einem umfangreichen Datenmanagementsystem immer klarer und notwendiger. Die anfangs überschaubare Anzahl von Tabellen und Beziehungen wurde ständig größer und komplexer. Dieser evolutionäre Entwicklungsprozess wurde von vielen Diskussionen mit allen Beteiligten begleitet und erst Ende 2015, nach der ersten vollständigen Vegetationsperiode mit Anlagen-Echtbetrieb, zu einem vorläufigen Abschluss gebracht.

Im Laufe der Entwicklung hat sich gezeigt, dass es notwendig ist, ein Modell bereitzustellen, das in allen Teilen frei skalierbar ist. Beispielsweise muss es damit möglich sein, jederzeit zusätzliche Parameter verarbeiten zu können. Da es immer wieder zu Anpassungen und Erweiterungen gekommen ist, musste von Beginn an das Modell so konzipiert werden, dass es maximale Flexibilität erlaubt. Aus dieser Notwendigkeit, bedingt durch den langen Entwicklungsprozess, ist ein Datenmanagementsystem entstanden, dass, abgesehen von seinem Einsatz in ClimGrass, nun für sämtliche Forschungsprojekte ohne Mehraufwand adaptierbar ist. Es kann sozusagen als Framework zur Realisierung eines Datenmanagementsystems für beliebige Fragestellungen herangezogen werden, wobei der Fokus auf naturwissenschaftlichen Experimenten liegt.

Um während der Entwicklung die ständigen Änderungen am Datenmodell effizient umsetzen zu können, wurden Scripts in Data Manipulation Language (DML) geschrieben, mit denen auf Knopfdruck die Datenbankstruktur inklusive dem statischen Inhalt (z.B. Bezeichnungen von Einheiten) neu erzeugt werden kann. Die ersten beiden Entwicklungsjahre waren dadurch geprägt, dass nach jeder Erweiterung und Anpassung des Modells die Datenbank gelöscht und mittels geändertem DML-Script wieder neu erzeugt wurde.

Das zentrale Element des Datenmodells ist die konsequente Trennung zwischen Mess- bzw. Beobachtungswerten und den dazugehörigen Metadaten. Das geht so weit, dass sogar die Information über den Parameter vom Wert getrennt wird und nur über eine entsprechende Beziehung zur Parametertabelle rekonstruiert werden kann. Redundante Informationen werden im gesamten Modell vermieden, wodurch eine entsprechende Performance gewährleistet bleibt, auch wenn im Laufe vieler Jahre eine sehr große Datenmenge entstehen sollte. Die starke relationale Aufgliederung der Tabellen erfordert allerdings auch eine intensive technische Unterstützung bei der Dateneingabe und Datenabfrage, da bei jedem Datenbankzugriff die Daten so zusammengestellt werden müssen, dass daraus lesbare und interpretierbare Information entsteht.

ClimGrass ist eine Versuchsanlage, welche Antworten auf viele Fragen zu den Folgen der Klimaveränderung für das Grünland beantworten wird. In den nächsten Jahren werden sich WissenschaftlerInnen aus unterschiedlichen Institutionen und Disziplinen im Rahmen von

Forschungsprojekten diesen Fragen stellen. Das Modell muss demnach eine weitere, sehr wichtige Eigenschaft besitzen: die Unterstützung einer verteilten Datenerfassung. Die Schnittstelle zur Datenbank muss von jedermann und von überall bedienbar sein, und zwar schreibend (Datenimport) und über Lesezugriffe (Datenbankabfragen). Dies erfordert zwangsläufig auch eine detaillierte Rechtevergabe für alle Projektpartner, ebenso auch für interessierte Dritte.

Zudem ist zum jetzigen Zeitpunkt noch nicht bekannt, wer für welche Projekte das System im Rahmen von ClimGrass in Zukunft nutzen wird. Auch die genaue Zielsetzung zukünftiger Forschungsprojekte ist unbekannt, d.h. die dafür notwendigen Parameter, Untersuchungsgegenstände, Methoden und Geräte sind ebenfalls nicht definiert. Da allerdings bereits jetzt in laufenden Projekten Daten erzeugt und mit dem ClimGrass-Datenmodell verwaltet werden, würden strukturelle Änderungen als Reaktion auf neue Projekte das bisherige System und die dort gespeicherten Daten in ein mehr oder minder großes Chaos stürzen. Aus diesem Grund sind alle Strukturen so angelegt, dass sie für die oben erwähnten Erweiterungen bereits gerüstet sind und keine strukturellen Veränderungen notwendig sind. Diese Art der Konzeption hat den großen Zusatznutzen, dass die vorliegende Implementierung des Datenmanagements nicht nur für neue Projekte innerhalb von ClimGrass erweiterbar ist, sondern auch für Projekte in gänzlich anderem Kontext geeignet ist.

In den folgenden Kapiteln wird auf die technische Umsetzung des Datenmodells bzw. der Datenbankkonfiguration eingegangen und anhand der Entitäten und Relationen die Funktionalität gezeigt.

3 Konfiguration des Datenbankmanagementsystems

3.1 Datenbankserver und Datenbank

Für die Speicherung sämtlicher Daten der ClimGrass-Versuchsanlage wird Microsoft SQLServer als Datenbankmanagementsystem verwendet. Im Netzwerk der HBLFA Raumberg-Gumpenstein befindet sich dafür ein eigener Server mit dieser Datenbank-Software. Die Daten sowie Datenbank-Backups werden auf einer mit diesem Server verbundenen Storage-Einheit gespeichert.

Die Datenbank dient der Speicherung sämtlicher Daten unterschiedlicher Partner, die im Rahmen der ClimGrass-Anlage wissenschaftlich aktiv sind. Dies betrifft insbesondere kontinuierliche Datenströme aus mehreren Sensorsystemen. Dazu gehören sowohl fix eingerichtete Messeinrichtungen wie die agrarmeteorologische Wetterstation (BOKU), die Lysimeter und die Temperaturerhöhungs- und CO_2-Begasungsanlagen, als auch temporär eingesetzte Instrumente wie Sensoren zur Messung der Bodenatmung. Die Ergebnisse aus verschiedenen Beobachtungen (Versuchsernten, Bodenproben, Wasserproben, Bonituren, usw.) sowie Metainformationen zu Versuchsmanagement und Beprobungen werden ebenfalls in die Datenbank durch manuelle Eingaben aufgenommen.

Die Nutzung der Datenbankinfrastruktur ist für alle Beteiligten fakultativ. Werden Daten aus Experimenten und Proben in der bereitgestellten ClimGrass-Datenbank abgelegt, sind festgelegte Regeln und Definitionen einzuhalten, um die Konsistenz bzw. die referentielle Integrität innerhalb des gesamten Datenmodells aufrecht zu erhalten. Eine Aufnahme von Daten in die Datenbank ist demnach von der Abstimmung auf die vorgegebene Datenstruktur abhängig, welche im Folgenden beschrieben wird.

3.2 DBMS-Wartung mit Zugriffsberechtigungen und Sicherung

Für die Erstellung der ClimGrass-Datenbank mit ihren Tabellen sowie für das Einfügen der statischen Daten (Parameterdefinitionen, Einheiten, Versuchsvarianten, verwendete Sensoren und Geräte, usw.) werden SQL Query Files genutzt. Der gesamte Definitionsprozess ist so mittels Programmcode festgelegt und beliebig oft reproduzierbar. Für die Datenbankerstellung sowie für die Durchführung von Sicherungs- und Wartungsaufgaben muss ein Benutzer mit Administratorrechten ausgestattet sein.

Im Zuge der Datenbankerstellung wurden zwei Benutzer festgelegt, die eine unterschiedliche Rechtestruktur aufweisen. Es handelt sich dabei um

- *ClimGrassAdmin:* Die SQL-Datenbank ClimGrass wird ausschließlich über diesen User administriert. Die Erstellung von Tabellen und Views sowie alle datenbezogenen Maßnahmen wie Import oder Update werden über diesen User vorgenommen.
- *ClimGrassUser:* Dieser User hat lediglich Leserechte und wird dann verwendet, wenn interne oder externe Anwender auf die Daten zugreifen.

Die Datenbankfiles und die dazugehörigen Log-Daten werden direkt am SQLServer bzw. auf der mit diesem Server verbundenen Storage-Einheit abgelegt.

Das Recovery Model für die Datenbank wurde mit „Simple“ festgelegt, d.h. die Logdaten enthalten nur die zuletzt durchgeführten Transaktionen. Für die ClimGrass-Datenbank ist ein Wartungsplan eingerichtet, in dem verschiedene Aufgaben von der Integritätsprüfung bis hin zum Backup in festgelegten Zeitintervallen abgearbeitet werden.

Die Sicherungen werden im ClimGrass-Backup-Verzeichnis abgelegt. Über *Database Mail* erfolgt eine automatische Verständigung an den Operator, falls die Ausführung einer der Sicherungs-Jobs fehlerhaft ist.

Differentielle Backups werden wöchentlich durchgeführt, die vollständige Sicherung erfolgt einmal pro Monat, wobei auch eine Reihe von Managementaufgaben zur Erhaltung der Effizienz durchgeführt werden. Die einzelnen Aufgaben sowie deren Ablauf wurden auf Best-Practice-Beispiele abgestimmt und sind in *Abbildung 1* dargestellt. Sie zeigt eine Übersicht der Wartungsaufgaben für den Job FULL Backup.

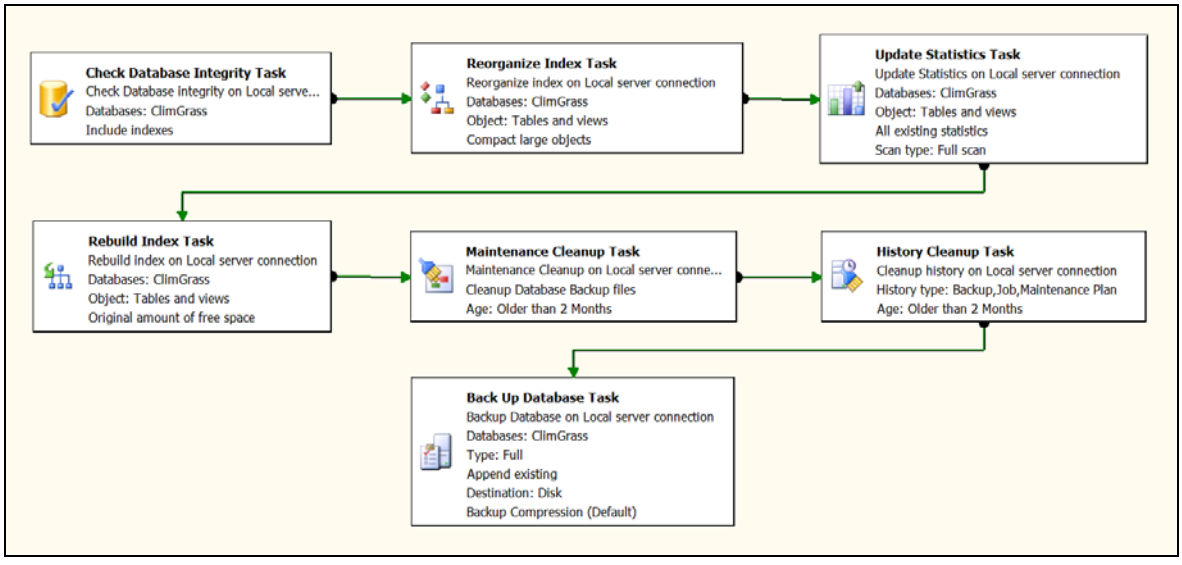

Abbildung 1: **Wartungsplan und Sicherung der Datenbank ClimGrass**

4 Datenbankstruktur

4.1 Das Datenmodell im Überblick

In ClimGrass ist geplant, verschiedenste Fragestellungen unterschiedlicher Projektpartner behandeln zu können. Im Laufe der Projektdauer wird sich der Kreis aktiv an den Experimenten Beteiligter wahrscheinlich vergrößern. Die Anforderungen an die Datenhaltung werden damit zunehmend vielfältiger und sind dabei ebenso unterschiedlich wie die involvierten Fachdisziplinen. Unter diesen Voraussetzungen ist ein gemeinsames Datenmodell erforderlich, dass in jeder Hinsicht skalierbar ist und dabei eine effiziente Grundlage für das Data Mining aller Partner darstellt. Alle in ClimGrass anfallenden Daten sollen zentral gespeichert werden und nach einem festzulegenden Berechtigungssystem für eine fachübergreifende Datenanalyse bereitgestellt werden (vgl. Janssen *et al.*, 2012).

Das Datenkonzept folgt dabei einem einfachen Aufbau, wie er in *Abbildung 2* dargestellt ist. Im Mittelpunkt steht ein aus verschiedensten Quellen (Lysimeter, Wetterstation, Labor, FACE, usw.) stammender Beobachtungswert, welcher in zentralen Tabellen, hier als *Measurement* zusammengefasst, abgespeichert wird. Die semantische Bedeutung des Beobachtungs- bzw. Messwertes selbst ist nur durch Zusatzinformationen wiederherstellbar. So wird jedem Wert ein Parameter, eine Versuchseinheit (*Experimental Unit*), ein Eigentümer/Erzeuger (*Partner*) und eine Information zur Probennahme (*Sampling*) bzw. zu einem zeitlich abgrenzbaren Ereignis (*Event*) wie beispielsweise die Ernte eines Grünlandaufwuchses zugeordnet. Damit wird sichergestellt, was mit dem Wert beschrieben wird (*Parameter*), wer dafür verantwortlich ist (*Partner*), auf welche Versuchseinheit er zu beziehen ist (*Experimental Unit*) und ob er das Ergebnis einer Beprobung (*Sampling*) ist.

Da ein zu jedem Wert definierter Zeitstempel grundsätzlich nichts darüber aussagt, welchem temporären Ereignis er zuzuordnen ist, wird eine Verknüpfung mit *Event* realisiert. Hier kann festgehalten werden, dass es sich um Messwerte handelt, die beispielsweise innerhalb eines bestimmten Aufwuchses angefallen sind, oder die zum Eintritt einer phänologischen Phase erhoben wurden.

Jedem Messwert können beliebig viele Versionen zugeordnet werden, die in *Versioning* erfasst sind. Implizit wird davon ausgegangen, dass die letzte Version die korrekte ist und für eine Weiterverwendung bereitgestellt wird. Damit können beispielsweise Datenlücken rechnerisch ersetzt werden, ohne die Information über die Lücke selbst zu verlieren.

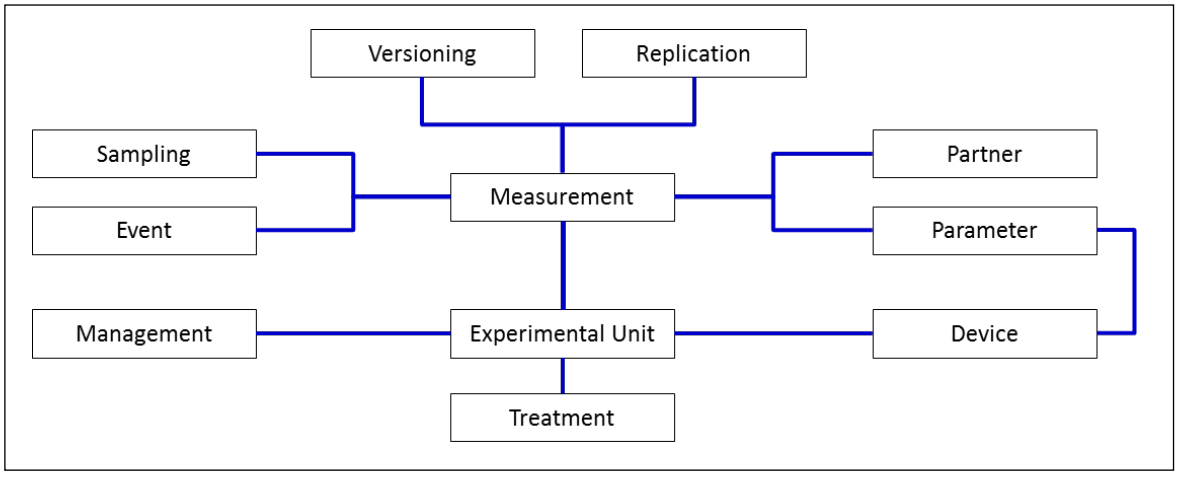

Abbildung 2: **Hauptkomponenten des Datenmodells ClimGrass**

In der Tabelle *Replication* sind Wiederholungen von Beobachtungen bzw. Messungen enthalten, aus denen ein Wert für *Measurement* generiert wird. Werden beispielsweise innerhalb einer Versuchsparzelle (*Experimental Unit*) mehrere Bodenproben genommen, so können die einzelnen Ergebnisse zusammen mit der genauen räumlichen Position innerhalb der Parzelle in *Replication* festgehalten werden. Die Charakteristik des Bodens innerhalb einer Zelle ergibt sich dann aus einem Mittelwert der Wiederholungen, welcher in *Measurement* abgespeichert wird.

Die Versuchseinheit (*Experimental Unit*) muss einer gewissen Betreuung und Pflege unterzogen werden. Alle dafür notwendigen Maßnahmen werden in *Management* erfasst. Die Beschreibung der Variante, mit deren Faktoren eine Versuchseinheit behandelt wird, findet sich in *Treatment*. Viele Messungen werden mit Hilfe von Sensoren durchgeführt. Einem definierten Parameter kann deshalb ein Sensor (*Device*) zugeordnet werden, um damit die Herkunft des Messwertes zweifelsfrei bestimmen zu können. Alle fix installierten Messinstrumente beziehen sich auf eine bestimmte räumliche Position, also auf eine *Experimental Unit*. Jene Instrumente, welche variabel eingesetzt werden, sind keiner Versuchseinheit zugeordnet und deshalb Bestandteil der gesamten Versuchsanlage.

4.2 Das Datenmodell im Detail

Während der Übersichtsplan in *Abbildung 2* lediglich die Hauptkomponenten zeigt, sind in *Abbildung 3* die konkreten Tabellen mit ihren Verknüpfungen und deren Multiplizitäten unter Wahrung der referentiellen Integrität zum aktuellen Entwicklungsstand angeführt. Im relationalen Datenmodell sind nach den Regeln der Datennormalisierung die vereinfacht dargestellten Komponenten von *Abbildung 2* in mehrere Relationen aufgefächert. So ist beispielsweise der einem Messwert (*MeasurementData, WeatherData, LysimeterData* oder *FaceData*) zugeordnete Parameter (*DataUnit*) eine Sammlung von Informationen aus den Tabellen *ParameterUnit, Parameter, Unit, UnitBasis*, *TimeScale* und *Device*.

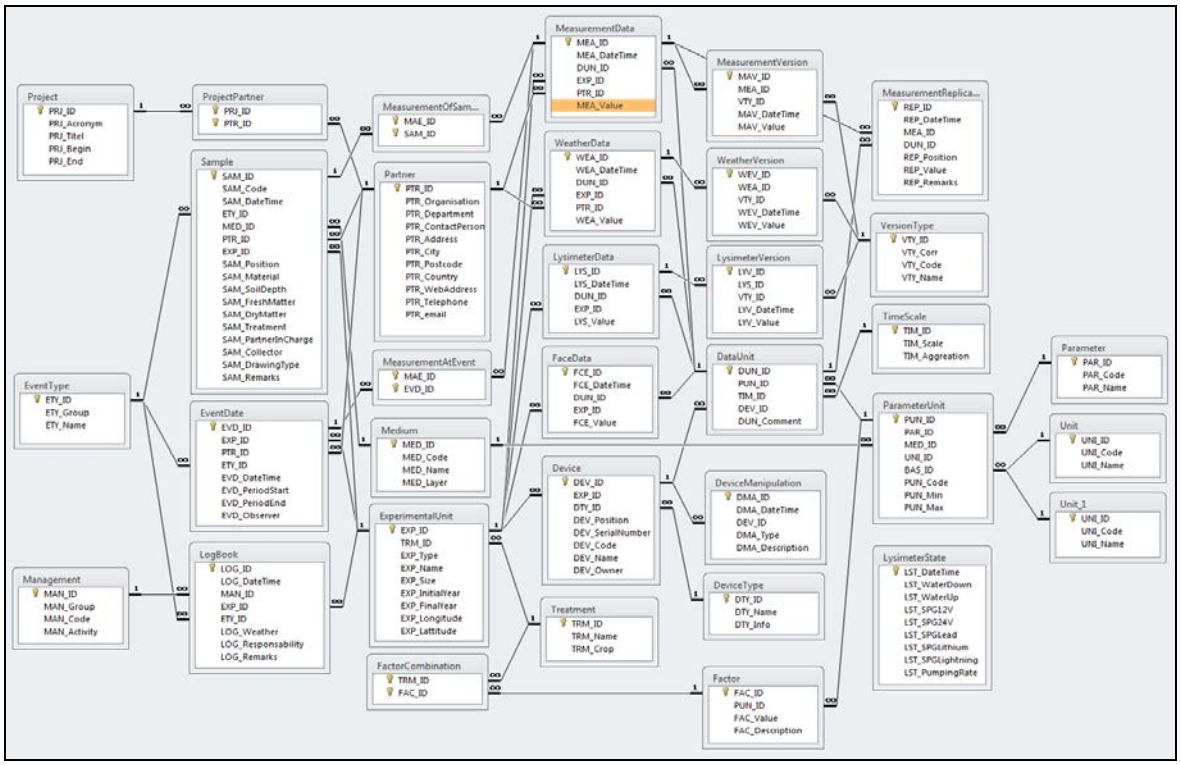

Abbildung 3: **Relationen des Datenmodells ClimGrass**

4.3 Tabellenbeschreibungen

4.3.1 Parameter, Einheiten und Bezugsgrößen

Mit jedem neuen Projekt bzw. Projektpartner werden die Anforderungen an die zu speichernden Beobachtungen und Messwerte größer. Neue Parameter mit entsprechenden Einheiten müssen deshalb ohne Änderung der Datenbankstruktur hinzugefügt werden können.

Das in *Abbildung 4* dargestellte Konzept ist Teil des gesamten Datenmodells und berücksichtigt eine strukturunabhängige Erweiterung der zur Speicherung von Daten erforderlichen Parameterinformationen.

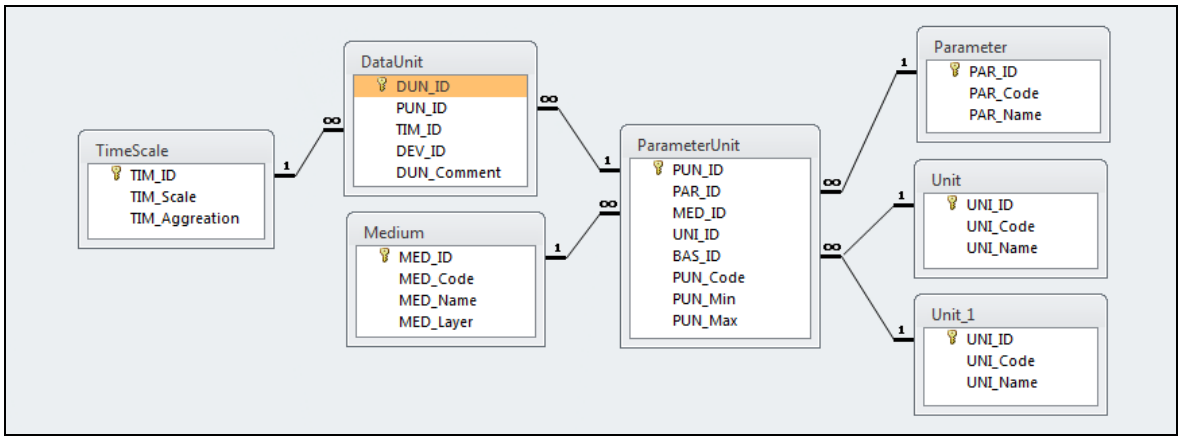

Abbildung 4: **Teilausschnitt mit Relationen zur Parameterbeschreibung**

Tabelle: **Parameter** *(PAR_ID*, PAR_Code, PAR_Name)
Beziehungen: keine

Die Parameter werden mit ihrem vollständigen Namen aufgelistet. Diese Tabelle enthält im Feld PAR_Code entsprechende Abkürzungen, jedoch keine Hinweise auf Einheiten oder Bezugsgrößen. Parameter können beliebig erweitert werden, wobei jedoch darauf zu achten ist, dass auf Einheiten oder Bezugsgrößen basierende Unterschiede nicht zu einer Mehrfachdeklaration von Parametern führt.

Tabelle: **Unit** und **Unit_1** *(UNI_ID*, UNI_Code, UNI_Name)
Beziehungen: keine

Mit der hier vorgenommenen Trennung von *Parameter* und *Unit* ist es möglich, einem Parameter unterschiedliche Einheiten zuzuordnen. In der Spalte UNI_Code ist die Abkürzung der Einheit gespeichert, unter UNI_Name die vollständige Bezeichnung. Erfasst sind die für ClimGrass relevanten Dimensionen in unterschiedlichen Auflösungen.

Um das Schlüsselattribut UNI_ID der Tabelle *Unit* ein zweites Mal mit der Tabelle *ParameterUnit* und der dort benötigten Variable BAS_ID verknüpfen zu können, wird die Tabelle *Unit* mit der Erweiterung *„_1“* nochmal dargestellt. Durch die zweifache Referenzierung der Datensätze in der Tabelle *Unit* ist eine beliebige Kombination von Einheit (UNI_ID) und Bezugsgröße (BAS_ID) aus derselben Datensammlung möglich.

In der Tabelle *ParameterUnit* stehen damit Einheiten zusammen mit deren Basis zur Beschreibung von Messdaten zur Verfügung (z. B. Milligramm pro Liter, Milligramm pro Kilo, Stickstoff pro Hektar, usw.).

Tabelle: **Medium** *(MED_ID*, MED_Code, MED_Name, MED_Layer)
Beziehungen: keine

Da sich bestimmte Parameter mit gleicher Einheit und Basis auf unterschiedliche Objekte beziehen können, ist eine zusätzliche Kombination mit einem als *Medium* bezeichneten Datenbankeintrag notwendig. Gemessene Bodentemperatur mit der Einheit °C können sich beispielsweise auf mehrere Bodenhorizonte beziehen, wobei die Horizontdefinition als Bezugsobjekt in der Tabelle Medium definiert ist. Jedes Medium wird zusätzlich einem Layer (Soil, Plant oder Atmosphere) zugeordnet.

Tabelle: **ParameterUnit** *(PUN_ID*, PAR_ID, MED_ID, UNI_ID, BAS_ID, PUN_Code, PUN_Min, PUN_Max)
Beziehungen: Parameter (PAR_ID), Medium (MED_ID), Unit (UNI_ID), UnitBasis (BAS_ID)

Mit der Zusammensetzung einer Parameterdefinition aus vier Tabellen ist eine maximale Flexibilität mit beliebiger Erweiterungsmöglichkeit gegeben. Jeder neue Parameter kann mit bestehenden oder neu definierten Einheiten, Bezugsgrößen und -objekten beliebig kombiniert und mit Angabe eines Wertebereiches ergänzt werden. Die Information über den Wertebereich ermöglicht eine aus den Daten selbst generierte Plausibilitätsprüfung. Der Parameter wird auch mit einer entsprechenden Abkürzung (PUN_Code) versehen.

Tabelle: **TimeScale** *(TIM_ID*, TIM_Scale, TIM_Aggregation)
Beziehungen: keine

Parameter und deren Einheiten (zusammengefasst in der Tabelle *DataUnit*) können sich auf unterschiedliche Zeitskalen mit verschiedenen Aggregationsniveaus beziehen. In der Tabelle *TimeScale* ist eine große Zahl von Zeiteinheiten, beginnend mit Sekunden bis hin zum Jahr definiert. Für zeitlich unabhängige Messgrößen steht eine Verknüpfung mit „No-Scale“ zur Verfügung. Jedes Skalenniveau wird zusätzlich über die Art und Weise der Aggregation differenziert.

Zur Auswahl stehen: None, Average, Minimum, Maximum, Sum. Alle in der Datenbank erfassten Beobachtungs- und Messwerte werden über die TIM_ID mit der zeitlichen Skala und dem dazugehörigen Aggregationsniveau kombiniert. Mit Hilfe dieser Tabelle ist es deshalb möglich, Messwerte für unterschiedliche Zeitschritte mehrfach aufzubereiten und anschließend in einer gemeinsamen Tabelle abzuspeichern.

Tabelle: **DataUnit** *(DUN_ID*, PUN_ID, TIM_ID, DEV_ID, DUN_Comment)
Beziehungen: ParameterUnit (PUN_ID), TimeScale (TIM_ID), Device (DEV_ID)

Jeder Beobachtungs- und Messwert in der Datenbank erhält über eine Verknüpfung mit der Tabelle *DataUnit* seine inhaltliche Definition. Eine *DataUnit* beschreibt über die PUN_ID den Parameter, seine Einheit mit Basis und Bezugsobjekt (*Medium*), die zeitliche Skala über die TIM_ID und kann sich auf einen zur Generierung des Messwertes verwendeten Sensor (Messgerät) aus der Tabelle *Device* beziehen.

Wird ein bestimmter Parameter von mehreren Sensoren erfasst, erfolgt über die DEV_ID eine eindeutige Zuordnung. DUN_Comment enthält grundsätzlich eine verbale Zusammenfassung aller Relationen (PUN_ID, TIM_ID und DEV_ID). Dieses Feld kann aber auch für allfällige Anmerkungen genutzt werden, speziell dann, wenn auf eine bestimmte Methode der Beobachtung hingewiesen werden soll, z. B. die angewandte Methodik einer Pflanzenbonitur. In diesem Fall bezieht sich die DEV_ID auf den Eintrag „NoDevice“ und DUN_Comment übernimmt die Rolle einer genauen Spezifikation.

4.3.2 Messinstrumente

In *Abbildung 5* sind jene Tabellen dargestellt, in denen die Sensoren und Messinstrumente verwaltet werden, die im Rahmen von ClimGrass zum Einsatz kommen. Wie bei den Parametern (Abschnitt 4.3.1) muss auch bei diesen Daten die Möglichkeit bestehen, jederzeit neue Geräte hinzufügen zu können. Neben der Erfassung unterstützt die Datenstruktur auch die Aufzeichnung diverser Manipulationen an den Geräten.

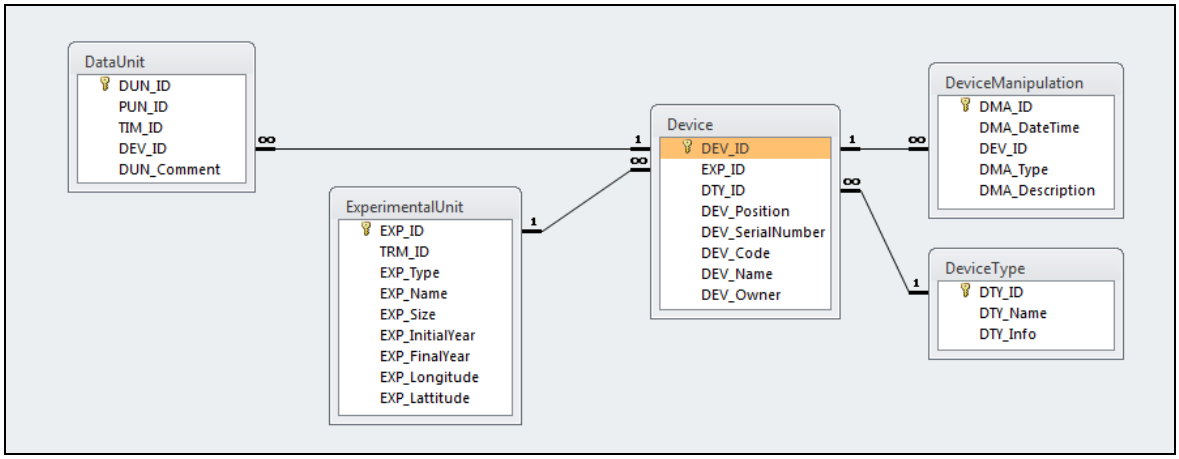

Abbildung 5: **Teilausschnitt mit Relationen zur Beschreibung von Messinstrumenten**

Tabelle: **DeviceType** *(DTY_ID*, DTY_Name, DTY_Info)
Beziehungen: keine

Alle in ClimGrass verwendeten Messinstrumente werden in verschiedene Klassen unterteilt, deren Definition in der Tabelle *DeviceType* festgehalten wird. Der Name spezifiziert die Geräteeinheit und kann noch um detailliertere Informationen erweitert werden. Mit Hilfe dieser Klassifizierung ist es später möglich, Datenbankabfragen über eine ganze Gruppe von Sensoren durchführen zu können.

Tabelle: **Device** *(DEV_ID*, EXP_ID, DTY_ID, DEV_Position, DEV_SerialNumber, DEV_Code, DEV_Name, DEV_Owner)
Beziehungen: ExperimentalUnit (EXP_ID), DeviceType (DTY_ID)

Zu den in der Tabelle *Device* erfassten Einträgen gehören sämtliche Sensoren, analoge und digitale Messgeräte sowie diverse Instrumente, die dem Erfassen von Daten dienen. Alle Geräte müssen einer Versuchseinheit (ExperimentalUnit) zugeordnet werden. Die Verknüpfung mit der Tabelle *DeviceType* erlaubt die Klassifizierung von ganzen Messeinheiten. So können beispielsweise alle meteorologischen Erfassungssysteme einer bestimmten Wetterstation zugeordnet werden. Für jedes Gerät kann die genaue Position innerhalb der Versuchseinheit spezifiziert werden. Handelt es sich um eine Versuchsparzelle, wird die Lage gemäß einer Quadtree-Codierung, wie sie in *Abbildung 9* schematisch dargestellt ist, vorgenommen. Statt des vorgegebenen Positionscodes kann auch ein bis zu 10 Zeichen langer Eintrag erfolgen. Die Tabelle *Device* hat auch den Charakter eines Inventarverzeichnisses für die Versuchsanlage ClimGrass. Aus diesem Grund ist das Feld DEV_SerialNumber nach Möglichkeit auszufüllen. Zu leichteren Identifikation des Gerätes im Rahmen diverser Datenbankabfragen wird der Inhalt von DEV_Code verwendet. Der Code kann und soll auch einen Hinweis zum Installationsort bzw. zur übergeordneten Messeinheit enthalten. Mit einer vergebenen und in *DataUnit* verwendeten DEV_ID wird eine wichtige Verknüpfung zwischen Messwert und Erfassungssystem realisiert. Nur so

bleibt eine eindeutige Identifizierung und Zuordnung eines Messwertes gewährleistet (beispielsweise wird die Globalstrahlung von drei unterschiedlichen Sensoren erfasst – sie kann nur durch das jeweilige Erfassungsinstrument eindeutig unterschieden werden).

Falls ein Sensor zu ersetzen ist, muss entschieden werden, ob ein neuer Eintrag erfolgt, oder ob der bisherige Eintrag aktualisiert wird (der bestehende Datensatz darf unter keinen Umständen gelöscht werden, da mit ihm möglicherweise historische Messdaten verknüpft sind). Bei einem neuen Eintrag ist auch eine Neudefinition in *DataUnit* zwingend erforderlich. Die Messwerterfassung des neuen Sensors kann damit vom alten unterschieden werden. Wird die Zeile lediglich aktualisiert, d. h. die DEV_ID bleibt auch beim neuen Sensor unverändert, wird die Erfassung neuer Messdaten mit derselben Verknüpfung fortgeführt.

Da über die nachfolgend beschriebene Tabelle *DeviceManipulation* aber eine Rekonstruktion des alten Sensors mit dem Datum der Änderung möglich ist, wird dringend empfohlen, bei einem Austausch bestehender Systeme die vorhandenen Einträge unter der jeweiligen DEV_ID beizubehalten und lediglich die dazugehörigen Felder zu ändern.

Tabelle: **DeviceManipulation** (*DMA_ID*, DMA_DateTime DEV_ID, DMA_Type, DMA_Description
Beziehungen: Device (DEV_ID)

Werden Messinstrumente in ihrer Konfiguration und räumlichen Lage verändert, sollte dies mit dem jeweiligen Datum und mit dem Manipulationstyp aufgezeichnet werden. Über DEV_ID wird eine Beziehung zu einem bestimmten Messinstrument hergestellt und die Manipulation zusammen mit dem Datum festgehalten. Als Typen sind folgende Einträge fest vorgegeben: *Installation*, *Demounting*, *PositionChange* und *Modification*.

Das Feld DMA_Description bietet Raum für zusätzliche Informationen. Als Konvention hat im Falle von „PositionChange“ in diesem Feld folgender Eintrag zu gelten: „From: 42341 To: 22131“ (Quadtree-Codes gemäß der grafischen Definition in *Abbildung 9*).

4.3.3 Versuchseinheit, Varianten und Versuchsmanagement

Alle Beobachtungen und Messungen im Rahmen von ClimGrass beziehen sich auf Untersuchungsobjekte, die im Datenmodell als *ExperimentalUnits* abgebildet werden. Eine Versuchseinheit ist ein beliebiges Objekt, auf das sich Experimente beziehen können; dies ermöglicht eine flexible Verwendung der Datenstruktur – von Labor- bis hin zu Feldexperimenten. Jeder Versuchseinheit sind Varianten zuordnen, die in der Regel aus Faktorkombinationen bestehen und sich auf einzelne Parameter beziehen. In *Abbildung 6* sind die dafür benötigten Tabellen und ihre Verknüpfungen dargestellt.

Neben den Tabellen zur Erfassung der Messwerte (*MeasurementData*, *LysimeterData*, *FaceData*) nimmt die Aufstellung der Versuchsobjekte (*ExperimentalUnit*) einen zentralen Stellenwert im gesamten Modell ein. Vielfach beschreibt dieses Objekt (z. B. Versuchsparzelle) auch den räumlichen Bezug. Messwerte und Beobachtungen, Probennahmen (*Sample*), Ereignisse (*EventDate*), Managementaufgaben (*LogBook*), Messinstrumente (*Device*) und Versuchsvarianten (*Treatment*) beziehen sich auf die in *ExperimentalUnit* definierten Datensätze.

Mit dieser Vielzahl an Verknüpfungen bekommt der Identifikationswert EXP_ID einen hohen Stellenwert, der nicht nur innerhalb des Datenmodells relevant ist, sondern auch für alle „extern“ gehaltenen Informationen und Programme die elementare räumliche Bezugsgröße darstellt.

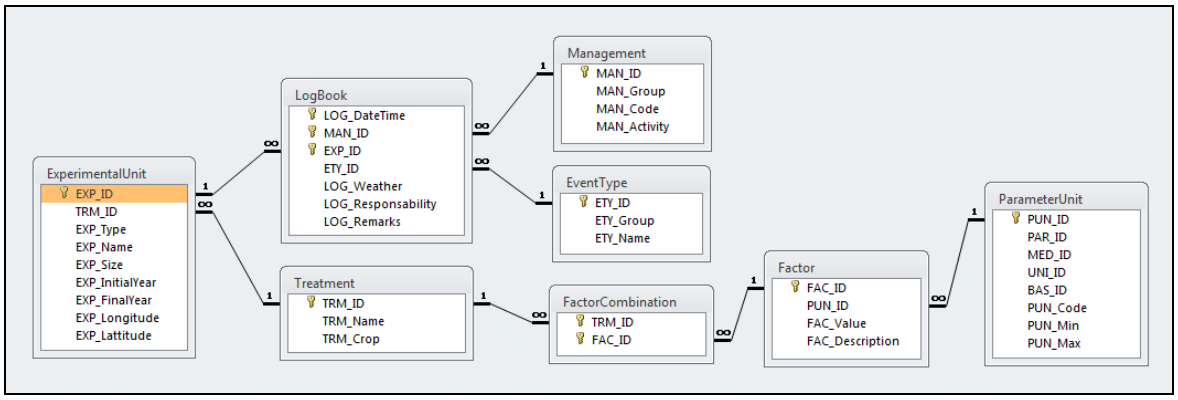

Abbildung 6: **Teilausschnitt mit Relationen für Versuchsvarianten und Versuchsmanagement**

Tabelle: **ExperimentalUnit** *(EXP_ID*, TRM_ID, EXP_Type, EXP_Name, EXP_Size, EXP_InitialYear, EXP_FinalYear, EXP_Longitude, EXP_Lattitude)
Beziehungen: Treatment (TRM_ID)

Jeder Versuchseinheit ist ein Treatment zugeordnet, da nur durch die Definition einer bestimmten Behandlung das Konzept eines Versuchsobjektes realisierbar ist. Um die Zusammenfassung von Gruppen zu unterstützen, ist im Feld EXP_Type eine Abkürzung mit vier Zeichen einzutragen. Gegenwärtig enthält die Tabelle drei unterschiedliche Klassen, wobei der Eintrag „FACI“ mit der EXP_ID 0 eine Sonderstellung einnimmt: er bezieht sich auf die gesamte ClimGrass-Versuchsanlage. Daneben sind noch die Gruppen „PLOT“ für die Versuchsparzellen (ID 1 bis 54) und „MESO“ für die noch zu installierenden Mesokosmen vertreten. Die IDs der Versuchsparzellen sind zur räumlichen Identifikation aller Aktvitäten innerhalb von ClimGrass maßgeblich, so zum Beispiel für sämtliche Übersichtspläne oder auch für die FACE-Regelungsalgorithmen in LabView.

Die weiteren Felder dieser Tabelle dienen der näheren Beschreibung des Objektes mit optionaler Angabe von Größe und geographischer Lage. Mit den beiden Feldern EXP_InitalYear und EXP_FinalYear kann die „aktive“ Periode eines Versuchsobjektes eingeschränkt werden. Wird beispielsweise ein Experiment auf einer Parzelle beendet, kann durch eine neue Erfassung derselben Parzelle mit Vergabe einer weiteren EXP_ID ein Folgeexperiment im Datenmodell so abgebildet werden, dass die Datenerfassung für beide Experimente deutlich voneinander zu trennen ist.

Tabelle: **Treatment** *(TRM_ID*, TRM_Name, TRM_Crop)
Beziehungen: keine

Im Rahmen eines Experiments ist die Behandlung (Variante) ein wichtiger Bestandteil des Versuchsobjektes und gemäß dem Versuchsplan mit ihm verbunden. Die Tabelle *Treatment* enthält eine verbale Beschreibung dieser Varianten, meist eine Kombination mehrerer Faktoren sowie eine Klassenzuordnung über das Feld TRM_Crop (im Fall von ClimGrass ist dies „Grassland“).

Tabelle: **FactorCombination** *(TRM_ID*, *FAC_ID)*
Beziehungen: Treatment (TRM_ID), Factor (FAC_ID)

Eine Variante besteht meist aus mehreren Faktoren, welche wiederum in mehreren Varianten enthalten sein können. Zur Auflösung dieser n:m-Beziehung zwischen *Factor* und *Treatment* muss die Tabelle *FactorCombination* eingeführt werden.

Tabelle: **Factor** *(FAC_ID*, PUN_ID, FAC_Value, FAC_Description)
Beziehungen: ParameterUnit (PUN_ID)

Die einzelnen Elemente einer Versuchsvariante sind als Faktoren definiert. Neben einer Beschreibung ist die Verknüpfung zu einem Parameter die Grundlage der Definition. Das Feld FAC_Value enthält die Information darüber, ob und wie der Parameter für die Versuchsvariante variiert wird. Beispielsweise wird eine Parzelle mit den Faktoren einer moderaten Temperatur- und CO_2-Erhöhung durch die beiden Parameter Temperatur und CO_2 bestimmt, aber erst mit dem FAC_Value von 1.5 °C und 150 ppm quantifiziert.

Der Wert 0 zeigt keine Veränderung gegenüber einer 0-Variante an; bei logischen Ja/Nein-Informationen enthält FAC_Value den Wert 1 für die Berücksichtigung des Parameters (Ja-Wert). Beispiel dafür ist der Parameter „Rainout Shelter" – beim Faktor „Activated Rainout Shelter" ist in FAC_Value 1 eingetragen. Eine Variante (*Treatment*) kann aus beliebig vielen Faktoren zusammengesetzt werden. Dies wird über Verknüpfungen in der Tabelle *FactorCombination* realisiert.

Der Sinn einer Auflösung von Treatments in einzelne Faktoren besteht darin, dass sämtliche Daten auf ihren Bezug zu einem bestimmten Faktor hin abgefragt werden können.

Tabelle: **Management** *(MAN_ID*, MAN_Group, MAN_Code, MAN_Activity)
Beziehungen: keine

Die Betreuung der Versuchsparzellen und Instrumente erfordert verschiedene Aktivitäten, welche in der Tabelle *Management* gelistet und in vier Gruppen unterteilt werden: „Harvest" (Ernte), „Maintain" (Erhaltungsmaßnahmen), „Checkup" (Kontrolle) und „Fertiliz" (Düngung). Jeder Aktivität, welche in MAN_Activity verbal beschrieben wird, ist ein zweistelliger Code zugeordnet, der für die Aufzeichnungen am Feld verwendet wird.

Tabelle: **LogBook** *(LOG_ID*, LOG_DateTime, MAN_ID, EXP_ID, ETY_ID, LOG_Weather, LOG_Responsability, LOG_Remarks)
Beziehungen: Management (MAN_ID), ExperimentalUnit (EXP_ID), EventType (ETY_ID)

Die Tabelle *LogBook* dient der Erfassung sämtlicher Versuchsmanagementaktivitäten zusammen mit Informationen, welche für die Interpretation von Messdaten und Beobachtungen unter Umständen maßgeblich sein können (z. B. Wettersituation). Über die Verknüpfung mit Management wird die Aktivitätsgruppe spezifiziert, mit EXP_ID erfolgt eine explizite Zuordnung zu einer bestimmten Versuchseinheit (ExperimentalUnit) und mit dem EventType wird die Aktivität einem Zeitpunkt oder einer Periode zugeordnet, welche im Versuchsablauf eine wichtige Rolle spielt (z. B. die Zuordnung von Aktivitäten zu einem bestimmten Aufwuchs).

Die Personen, welche mit der Durchführung beauftragt sind, werden in LOG_Responsibility eingetragen, mögliche Anmerkungen in LOG_Remarks. Mit den Logbuchaufzeichnungen sollte es unter anderem möglich sein, die Ursache von ungewöhnlichen Datenaufzeichnungen mit Managementaktivitäten in Verbindung zu bringen.

4.3.4 Beobachtungen und Messungen, Zeitbezug, Versionierung und Wiederholungen

Alle im Rahmen von ClimGrass durchgeführten Beobachtungen und Messungen werden in den Tabellen *MeasurementData*, *LysimeterData* und *FaceData* abgelegt. Diese Tabellen weisen die gleiche Struktur auf und stehen im Zentrum des Datenmodells; sie sind mit allen anderen Tabellen direkt und indirekt verknüpft.

In *Abbildung 7* sind jene Verknüpfungen zu den drei Beobachtungs- und Messdatentabellen dargestellt, welche die Zuordnung zu Ereignissen, Versionen und Wiederholungen betreffen.

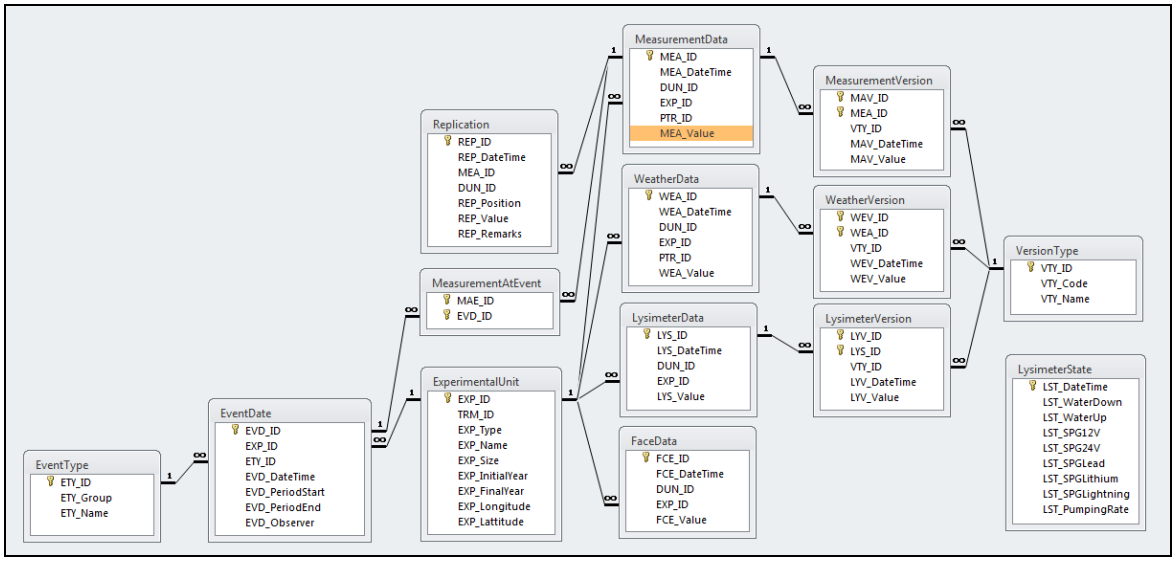

Abbildung 7: **Teilausschnitt mit Relationen für Beobachtungen und Messungen, deren zeitliche Zuordnung, Versionierung und Wiederholung**

Tabelle: **EventType** *(ETY_ID*, ETY_Group, ETY_Name)
Beziehungen: keine

Wenn Ereignisse für eine Selektion von Beobachtungs- und Messdaten relevant sind, muss zunächst eine Kategorisierung vorgenommen werden. Ereignistypen können sich beispielsweise auf Managementaktivitäten oder auf phänologische Ereignisse beziehen. So sind Schnitttermine oder die Periode eines bestimmten Grünlandaufwuchses für die Abgrenzung von Beobachtungsdaten wichtig. Einträge in *EventType* können sich sowohl auf Zeitpunkte als auch auf Zeitperioden beziehen.

Tabelle: **EventDate** *(EVD_ID*, EXP_ID, ETY_ID, EVD_DateTime, EVD_PeriodStart, EVD_PeriodEnd, EVD_Observer)
Beziehungen: ExperimentalUnit (EXP_ID), EventType (ETY_ID)

Voraussetzung für die konkrete Definition von Ereignissen ist die Existenz eines *EventTypes*. Ereignisse desselben Typs können beliebig oft wiederkehren; sie unterscheiden sich primär in der laufend vergebenen EVD_ID. Jedem konkreten Ereignis kann entweder ein explizites Datum (EVD_DateTime) oder ein Zeitraum zwischen zwei Datumseinträgen (EVD_PeriodStart und EVD_PeriodEnd) zugewiesen werden. Im Gegensatz zu den Einträgen der Tabelle *TimeScale*, wo Parameter mit einer zeitlichen Skala kombiniert werden, ist hier eine beliebige zeitliche Abgrenzung definierbar. Sie ermöglicht damit eine ereignisspezifische Selektion von Beobachtungs- und Messdaten und geht über den Dokumentatscharakter der Daten in *Management* und *LogBook* hinaus. Jedes Ereignis bezieht sich auf eine konkrete Versuchseinheit. Findet dasselbe Ereignis (z. B. die Periode des ersten Aufwuchses) auf mehreren Versuchseinheiten statt, muss es für jede *ExperimentalUnit* einen eigenen Ereigniseintrag geben. Damit ist eine flexible Zuweisung und eine Differenzierung innerhalb gleichartiger Versuchseinheiten (Parzellen) möglich, erfordert allerdings mehr Erfassungsaufwand. Falls ein Ereignis auf alle Versuchseinheiten allerdings gleichermaßen wirkt, kann die EXP_ID 0 gesetzt werden – diese ID bezieht sich auf die gesamte Versuchsanlage.

Tabelle: **MeasurementData** *(MEA_ID*, MEA_DateTime, DUN_ID, EXP_ID, PTR_ID, MEA_Value
Beziehungen: DataUnit (DUN_ID), ExperimentalUnit (EXP_ID), Partner (PTR_ID)

Alle Beobachtungen und Messungen der Projektpartner werden in dieser Tabelle gespeichert. Dabei wird jeder erfasste Wert zusammen mit zeitlicher, räumlicher und semantischer Information abgelegt. Der Zeitstempel in MEA_DateTime enthält Datum und Uhrzeit, wobei Daten mit ausschließlichem Datumsbezug die Uhrzeit 00:00:00 zugewiesen wird. Die semantische Bedeutung eines Wertes (MEA_Value) wird über die Verknüpfung der Spalte DUN_ID mit der Beschreibung in *DataUnit* definiert. Die räumliche Zuordnung zu einer bestimmten Versuchseinheit (*ExperimentalUnit*) erfolgt über EXP_ID. Damit eine Beobachtung dem Projektpartner direkt zugerechnet werden kann, enthält der Datensatz über PTR_ID eine Verknüpfung zum Partner. Die Erfassung der Daten erfolgt dezentral, wird extern auf Konsistenz mit dem Datenmodell geprüft und erst dann in die zentrale Datenbank übernommen. Beim Übertrag erhält jeder Datensatz eine fortlaufende Identifikationsnummer (MEA_ID), die für alle weiteren Verknüpfungen maßgeblich ist.

Tabelle: **MeasurmentReplication** *(REP_ID*, REP_DateTime, MEA_ID, DUN_ID, REP_Position, REP_Value, REP_Remarks)
Beziehungen: MeasurementData (MEA_ID), DataUnit (DUN_ID)

Messwerte können aus einer Kombination mehrerer Einzelwerte generiert werden. Diese Wiederholungen werden in der Tabelle *MeasurmentReplication* erfasst. Mit der Verknüpfung zu einem Messwert (MEA_ID) findet eine eindeutige Zuordnung zu dem sich aus mehreren Einzelmessungen ergebenden Wert statt. Damit eine Datenerfassung in *MeasurmentReplication* möglich ist, muss zunächst ein Eintrag in *MeasurementData* vorhanden sein. Damit ergibt sich folgender Workflow:

a) Die Einzelmessungen müssen aggregiert werden.
b) Der aggregierte Wert muss in *MeasurementData* eingetragen werden – es wird eine neue MEA_ID vergeben.
c) Die einzelnen Messungen werden in *MeasurmentReplication* zusammen mit der zuvor erstellten MEA_ID eingetragen.

Die Einzelmessungen können sich von dem ihnen zugeordneten (aggregierten) Messwert (MEA_ID) sowohl in zeitlicher (REP_DateTime) als auch in räumlicher (REP_Position) Hinsicht unterscheiden. Die räumliche Differenzierung wird im Fall von Versuchsparzellen gemäß der Quadtree-Codierung, wie sie in *Abbildung 9* schematisch dargestellt ist, vorgenommen. Mit der Spalte DUN_ID kann jedem Wiederholungswert eine vom MEA_Value abweichende *DataUnit* zugeordnet werden. Dies ist dann notwendig, wenn die Einzelwerte in Replication auf Parameter basieren, die sich vom Messwert in *MeasurementData* unterscheiden (z. B. eine andere zeitliche Skala oder unterschiedliche Einheiten). In *REP_Remarks* kann beispielsweise die Methode eingetragen werden, mit der ein aggregierter Wert in *MeasurementData* berechnet wird. Üblicherweise wird hier für mehrere Werte das arithmetische Mittel als Zentralwert verwendet, könnte aber auch ausnahmsweise der Median sein – dies sollte dann vermerkt werden.

Tabelle: **MeasurementVersion** *(MAV_ID*, MEA_ID, VTY_ID, MAV_DateTime, MAV_Value)
Beziehungen: MeasurementData (MEA_ID), VersionType (VTY_ID)

Messwerte können falsch sein oder überhaupt fehlen. In diesem Fall wird in *MeasurementData* ein Fehlercode abgespeichert. Um Auswertungen zu ermöglichen, für die eine lückenlose und/oder fehlerfreie Datenreihe erforderlich ist, kann jeder Wert mittels geeig-

neter Methoden (Interpolation, Plausibilitätsprüfung, Mittelwertbildung, manuelle Korrektur usw.) beliebig oft versioniert werden. Es ist lediglich eine Verknüpfung mit dem originalen Messwert über das Feld MEA_ID herzustellen. Als eindeutiger Schlüssel wird ein Autoinkrementwert verwendet, der von Auswerteprogrammen in Abhängigkeit der MEA_ID hinsichtlich der letzten Version ausgewertet wird. Dabei wird der größte MAV_ID einer bestimmten Messung (MEA_ID) als die aktuelle Version eines Messwerte ausgelesen. Alle Zugriffe auf *MeasurementData* sind nur dann zulässig, wenn vorab geprüft wurde, ob für einen bestimmten Messwert eine Version existiert. Jener Versionswert mit der höchsten MAV_ID zu einer bestimmten MEA_ID ist jener Wert, welcher für Dritte anstelle des in *MeasurementData* gespeicherten Wertes, der implizit die Version 0 darstellt, zugänglich gemacht wird. Ist kein versionierter Wert vorhanden ist, stellt der Messwert in *MeasurementData* die Erst- und gleichzeitig die Letztversion mit der implizit vorhandenen und somit höchsten Versionsnummer 0 dar.

Tabelle: **VersionType** *(VTY_ID*, VTY_Code, VTY_Name)
Beziehungen: keine

Jede Versionierung eines Messwertes muss exakt begründet sein und deshalb einer vordefinierten Kategorie zugeordnet werden. Die Tabelle VersionType enthält eine Liste der Versionierungsgründe, aus der ein Eintrag zu näheren Beschreibung der Version in MeasurementVersion oder LysimeterVersion ausgewählt werden muss. Mit dem Typ wird auch die Information zur Methodik der Versionierung direkt in der Datenbank abgebildet. Somit enthält jeder Versionswert (MAV_Value bzw. LYS_Value) über die Verknüpfung mit VTY_ID Angaben zur Methodik seiner Entstehung.

Tabelle: **WeatherData** *(WEA_ID*, WEA_DateTime, DUN_ID, EXP_ID, PTR_ID, WEA_Value)
Beziehungen: DataUnit (DUN_ID), ExperimentalUnit (EXP_ID), Partner (PTR_ID)

Wetterdaten bilden gegenüber anderen Erhebungen gut abgrenzbare Beobachtungen. Sie werden über mehrere Sensoren einer Wetterstation gesammelt und dienen der Beurteilung des atmosphärischen Einflusses auf die ClimGrass-Experimente. Mittlerweile stammen die Daten aus unterschiedlichen Wetterstationen, die über die PTR_ID den jeweiligen Partnern zugeordnet sind.

WeatherData weist die gleiche Struktur und Verknüpfungsinformationen wie *MeasurementData* auf. Für *WeatherData* ist auch die Option einer Versionierung vorgesehen und kann in gleicher Weise wie bei *MeasurementData* umgesetzt werden.

Tabelle: **WeatherVersion** *(WEV_ID*, WEA_ID, VTY_ID, WEV_DateTime, WEV_Value)
Beziehungen: WeatherData (WEA_ID), VersionType (VTY_ID)

WeatherVersion verhält sich zu *WeatherData* exakt gleich wie *MeasurementVersion* zu *MeasurementData*. Alle Möglichkeiten und Einschränkungen für die Measurement-Kombination gelten in gleicher Weise auch für die Wetterdaten.

Tabelle: **LysimeterData** *(LYS_ID*, LYS_DateTime, DUN_ID, EXP_ID, LYS_Value)
Beziehungen: DataUnit (DUN_ID), ExperimentalUnit (EXP_ID)

Die sechs in ClimGrass verbauten Lysimeter mit ihren Sensoren produzieren Tag für Tag in der Regel 32.112 Werte. Da diese Datenmenge nicht im Verhältnis zu den übrigen Beobachtungen steht, werden diese Daten in einer eigenen Tabelle abgespeichert. *LysimeterData* weist die gleiche Struktur wie *MeasurementData* auf, enthält allerdings keine Verknüpfung zu Partnern, da die Zugehörigkeit der Daten zur HBLFA ohnehin unveränderlich und für alle Werte gleichermaßen gültig ist. Alle anderen Verknüpfungen entsprechen je-

nen der Tabelle *MeasurementData*; sämtliche dort vorgenommenen Abfragen sind damit auf *LysimeterData* übertragbar. Eine Zusammenführung beider Tabellen ist aufgrund der gleichen Struktur möglich. Für *LysimeterData* ist auch die Option einer Versionierung vorgesehen und kann in gleicher Weise wie bei *MeasurementData* umgesetzt werden. Da es sich bei diesen Daten um automatische Sensormessungen handelt, ist das Konzept einer Wiederholung (*Replication*) hier nicht vorgesehen und auch nicht sinnvoll.

Tabelle: **LysimeterVersion** *(LYV_ID*, LSY_ID, VTY_ID, LYV_DateTime, LYV_Value)
Beziehungen: LysimeterData (LYS_ID), VersionType (VTY_ID)

LysimeterVersion verhält sich zu *LysimeterData* exakt gleich wie *MeasurementVersion* zu *MeasurementData*. Alle Möglichkeiten und Einschränkungen für die Measurement-Kombination gelten in gleicher Weise auch für die Lysimeterdaten.

Tabelle: **FaceData** *(FCE_ID*, FCE_DateTime, DUN_ID, EXP_ID, FCE_Value)
Beziehungen: DataUnit (DUN_ID), ExperimentalUnit (EXP_ID)

Neben *MeasurementData* und *LysimeterData* beinhaltet das ClimGrass-Datenmodell eine dritte Datentabelle, welche die CO_2- und die Temperaturmessungen (FACE) sowie den Status der Rainout-Shelter auf den Versuchsparzellen beinhaltet. Da diese Gruppe von Daten zu anderen Versuchsfragen hin deutlich abgrenzbar ist, wurde dafür eine eigene Tabelle geschaffen, welche allerdings die gleiche Struktur wie *LysimeterData* aufweist.

Wie schon bei *MeasurementData* und *LysimeterData* sind auch hier sämtliche Verknüpfungen zu *DataUnit* und *ExperimentalUnit* für Abfragen mit Einbeziehung sämtlicher Tabellen des Datenmodells möglich. Im Gegensatz zu den beiden anderen Datentabellen gibt es hier keine Möglichkeit der Versionierung oder Replikation.

Tabelle: **MeasurementAtEvent** *(MEA_ID, EVD_ID)*
Beziehungen: MeasurementData (MEA_ID), EventDate (EVD_ID)

Beobachtungen und Messungen werden vielfach mit einem Ereignis in Beziehung gesetzt. Beispielsweise fallen Erhebungen in der Dauer eines bestimmten Grünlandaufwuchses an. Eine mögliche Auswertung basiert genau darauf, nur für diesen Aufwuchs relevante Daten heranzuziehen. In einem solchen Fall ist es notwendig, die in MeasurementData abgelegten Werte der in EventDate definierten Periode zuzuordnen und daraufhin zu selektieren. Vor allem Bonituren und Erntedaten im Rahmen von Grünlandversuchen müssen zwangsläufig mit einem bestimmten Aufwuchs verknüpft werden, um eine aufwuchsbezogene Auswertung zu ermöglichen.

Ein Beobachtungs- oder Messwert kann aber auch für mehrere Ereignisse relevant sein. So bezieht sich ein Temperaturwert beispielsweise auf eine Auswertung für die Dauer eines Aufwuchses, auf eine bestimmte Vegetationsperiode, oder auch auf ein bestimmtes Erntedatum. Die Tabelle *MeasurementAtEvent* dient hauptsächlich dazu, die m:n-Multiplizität zwischen *EventDate* und *MeasurementData* entsprechend aufzulösen und eine Mehrfachzuordnung in beide Richtungen zu unterstützen.

Tabelle: **LysimeterState** *(LYS_DateTime,* LYS_WaterDown, LYS_WaterUp, LYS_SPG12V, LYS_SPG24V, LYS_SPGLead, LYS_SPGLithium, LYS_SPGLightning, LYS_PumpingRate)
Beziehungen: keine

Diese Tabelle steht als einzige zu keiner anderen Tabelle im ClimGrass-Datenmodell in Beziehung. Sie enthält lediglich Statusmeldungen der Lysimeter und dient zur Überprüfung der ordnungsgemäßen Funktionalität der in den Lysimetern verbauten Sensorsysteme.

4.3.5 Probenziehung und Projektpartner

Viele Messdaten basieren auf Proben, die einer Versuchseinheit entnommen und analysiert werden. Die daraus gewonnenen Daten werden in der Tabelle *MeasurementData* abgespeichert. Die Metainformationen einer Probennahme sind für die Bewertung und Einordnung der daraus gewonnen Daten wichtig und werden deshalb im ClimGrass-Datenmodell festgehalten. Nach Eingabe der Analysedaten kann eine Verbindung zu den Beprobungsinformationen hergestellt und damit jedem Wert die Eckdaten seiner Entstehung zugeordnet werden. Die Versuchsanlage ClimGrass soll verschiedenen Projektpartnern und Projektfragestellungen eine Grundlage ihrer Arbeit bieten. Alle Projekte und Partner werden in der Datenbank erfasst und zur Dokumentation der Verantwortlichkeiten und der Zuweisung von Rechten an Daten genutzt. Dazu gibt es eine Verknüpfungen zwischen Partnern und deren Messdaten bzw. Beprobungsdaten. Zu jedem in *MeasurementData* erfassten Datensatz ist der Besitzer direkt mit angegeben. Die Liste der Partner bildet in weiterer Folge die Grundlage für die Erstellung eines Datenbankzugriffs- und Rechtesystem.

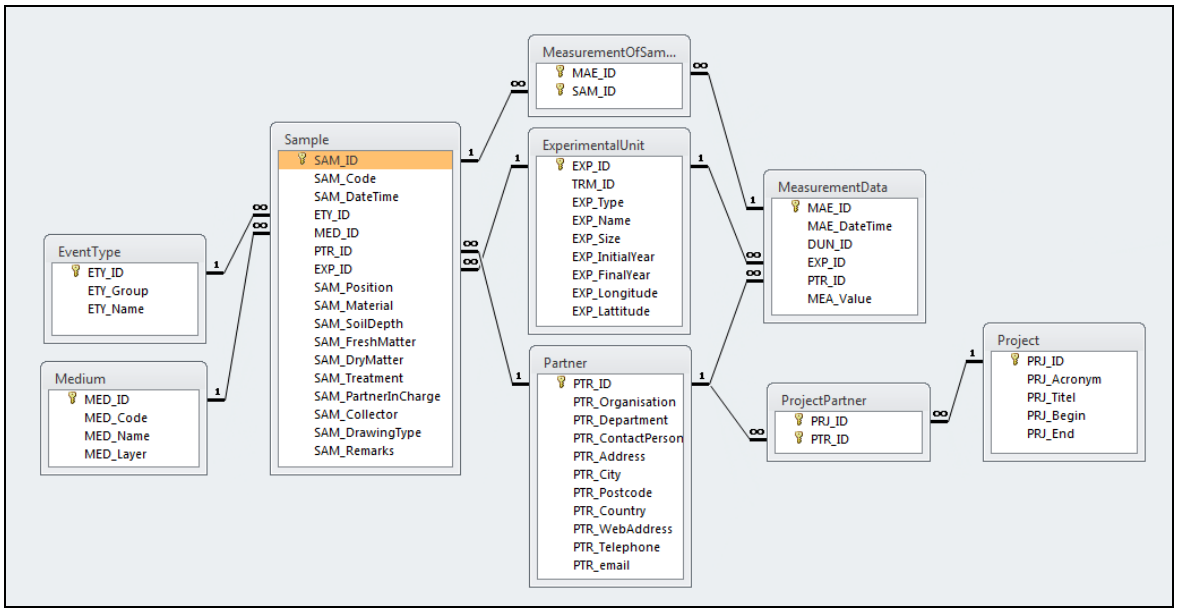

Abbildung 8: **Teilausschnitt mit Relationen für Probenziehung und Partnerverwaltung**

Tabelle: **Sample** *(SAM_ID*, SAM_Code, SAM_DateTime, MED_ID, PTR_ID, EXP_ID, SAM_Position, SAM_Material, SAM_SoilDepth, SAM_FreshMatter, SAM_DryMatter, SAM_Treatment, SAM_PartnerInCharge, SAM_Collector, SAM_DrawingType, SAM_Remarks)

Beziehungen: Medium (MED_ID), Partner (PTR_ID), ExperimentalUnit (EXP_ID)

In der Tabelle *Sample* werden die Metadaten einer Probennahme unter fortlaufender ID (Autowert) gespeichert. Zusätzlich besteht die Möglichkeit, jede Probe mit einer eigenen Codierung zu kennzeichnen. Das Datum bezieht sich auf den Vorgang der Probennahme, die Verknüpfung zu *EventType* beschreibt die Zugehörigkeit zu einem bestimmten Ereignistyp (z. B. Ernte eines bestimmten Aufwuchses, Eintritt einer phänologischen Phase, usw.). Im Feld MED_ID kann das Bezugsmedium definiert und mit der Tabelle *Medium* verknüpft werden; dies kann beispielsweise eine Entnahme des Bodens in einer bestimmten Tiefe, eine Biomasseprobe aus dem Pflanzenbestand, eine Einzelpflanze, usw. sein. Jede Probe wird auf einer bestimmten Versuchseinheit (EXP_ID) entnommen, wo zusätzlich die genaue Position (SAM_Position) definiert werden kann. Im Fall von Versuchspar-

zellen kann die in *Abbildung 9* dargestellte Codierung zur näheren Bestimmung der räumlichen Position herangezogen werden.

Das Feld SAM_Material enthält wie bereits MED_ID eine Verknüpfung zur Tabelle *Medium*. Beispiel: Wenn im Feld MED_ID das Knaulgras referenziert wird, so kann im Feld SAM_Material zusätzlich spezifiziert werden, dass sich die Knaulgrasprobe lediglich auf die Entnahme von Blättern beschränkt. In beiden Fällen werden Daten aus der Tabelle *Medium* zur genauen Beschreibung des Bezugsobjektes herangezogen, einmal die ID des Knaulgraseintrages und dann die ID des Eintrages Blatt.

Das Feld SAM_SoilDepth ist fakultativ und dient der genauen Beschreibung von Bodentiefen bzw. Horizonten, auf die sich die Entnahme einer Bodenprobe bezieht. Zwar sind in der Tabelle *Medium* eine ganze Reihe von Bodentiefen vordefiniert, allerdings kann aus diesen Daten nicht auf den bezugnehmenden Horizont geschlossen werden. Wenn sich eine Probe beispielsweise auf den Horizont von 10 bis 30 cm bezieht, so muss in MED_ID die Referenz auf Soil030cm gesetzt werden, SAM_SoilDepth enthält den Wert „10-30“. Als Konvention wird also festgesetzt, dass als *Medium* (MED_ID) die unterste Horziontgrenze referenziert wird und die Mächtigkeit (von-bis) das Feld SAM_SoilDepth enthält. Handelt es sich bei den Proben um Pflanzenproben bleibt dieses Feld leer.

Die beiden Felder SAM_FreshMatter bzw. SAM_DryMatter enthalten die entnommene Probenmenge in Gramm. Für das Feld SAM_Treatment sind fünf Parameter vorgesehen: Fresh, Cooled, Frozen, DeepFrozen und (Air)Dried. Die Verantwortlichkeit eines bestimmten Partners für die Probennahme wird mit Hilfe einer Verknüpfung zur Tabelle Partner und des Eintrages der PTR_ID im Feld SAM_PartnerInCharge festgelegt. Im Gegensatz dazu steht im Feld PTR_ID ebenfalls ein Verweis auf einen bestimmten Partner, hier ist es nicht der Verantwortliche für die Entnahme, sondern für die Probe selbst. Beispiel: Ein externer Partner beauftragt die HBLFA zur Probenziehung. Im Feld PTR_ID wird auf den externen Partner verwiesen, im Feld SAM_PartnerInCharge wird die HBLFA referenziert. Um die Verantwortlichkeit der Probennahme an einzelnen Personen festzumachen, werden im Textfeld SAM_Collector die Personen genannt, welche die Beprobung durchgeführt haben. Für den Typ der Beprobung kann in SAM_DrawingType entweder 1 für Mischprobe oder 2 für Einzelprobe eingetragen werden. Das Feld SAM_Comment schafft die Möglichkeit, Kommentare zur Probe zu erfassen.

Tabelle: **MeasurementOfSample** *(MEA_ID, SAM_ID)*
Beziehungen: Measurement (MEA_ID), Sample (SAM_ID)

Um Beobachtungs- und Messdaten mehrfach mit den Metadaten der Probennahme zu verknüpfen, ist eine eigene Tabelle für die Umsetzung der m:n-Multiplizität erforderlich. Mit dieser Tabellen- und Verknüpfungsstruktur ist es möglich, einzelne Messwerte mehreren Probennahmen zuzuordnen. Das ist allerdings nur dann gegeben, wenn der Messwert das Produkt einer Aggregation von Wiederholungen (*Replication*) mit jeweils eigener Probennahme ist. Damit ist nachvollziehbar, welche Proben die Grundlage für die Bildung des aggregierten Messwertes liefern. Eine direkte Verknüpfung zwischen *Replication* und *Sample* ist in diesem Fall nicht notwendig, da in *Replication* das Feld REP_Position die räumliche Lage der jeweiligen Probennahme beschreibt und mit der Positionsangabe in Sample (SAM_Position) ident sein muss.

In allen Fällen, in denen der Messwert direkt und ohne Wiederholung aus dem Untersuchungsmaterial bestimmt wird, kann dieser nur einer einzigen Probennahme zugeordnet werden. Umgekehrt können sich aus einer Probe mehrere Messwerte ergeben, die dann in der Tabelle *MeasurementOfSample* entsprechend zu verknüpfen sind. Ein Vorteil der Aus-

lagerung einer Zuordnung zwischen *MeasurementData* und *Sample* in eine eigene Tabelle liegt auch darin, dass nur dann Verknüpfungen generiert werden, wenn Messdaten aus Probenziehungen hervorgehen. In allen anderen Fällen, wie beispielsweise für Wetterdaten, ist diese Beziehung nicht relevant.

Tabelle: **Partner** *(PTR_ID*, PTR_Organisation, PTR_Department, PTR_ContactPerson, PTR_Address, PTR_City, PTR_Postcode, PTR_Country, PTR_WebAddress, PTR_Telephone, PTR_email)
Beziehungen: keine

Die Tabelle enthält die Namen und Kontaktdaten aller im Rahmen der ClimGrass-Versuchsanlage beteiligten Personen und ihrer Institutionen.

Tabelle: **Project** *(PRJ_ID*, PRJ_Acronym, PRJ_Titel, PRJ_Begin, PRJ_End)
Beziehungen: keine

Es wird angenommen, dass ClimGrass die Grundlage für mehrere interne und externe Projekte sein wird. Eine Liste der Projekte mit ihrem zeitlichen Beginn und Ende ermöglicht über eine Verknüpfung mit den Partnern eine Abfrage aller zu einem bestimmten Projekt gehörenden Daten. Eine projektbezogene Datenselektion kann auch die Grundlage für die Vergabe von differenzierten Datenbankberechtigungen sein.

Tabelle: **ProjectPartner** *(PRJ_ID*, *PTR_ID)*
Beziehungen: Project (PRJ_ID), Partner (PTR_ID)

Eine Projektpartnerschaft wird im Sinne des ClimGrass-Datenmodells damit definiert, dass ein Datensatz der Tabelle *Partner* mit einem Datensatz in *Project* verknüpft wird. Die Beziehung findet ihre Abbildung in der Tabelle *ProjectPartner*. Damit ist eine indirekte Beziehung zwischen *MeasurementData* und *Project* hergestellt, sodass projektspezifische Daten eindeutig einem in *Project* gelisteten Projekt zugeordnet werden können.

4.3.6 Räumliche Referenzierung innerhalb der Versuchsparzellen

Auf den Parzellen der ClimGrass-Anlage werden an unterschiedlichen Stellen Messungen und Probennahmen durchgeführt. Um exakt definieren zu können, wo welche Beobachtungen stattfanden, wird ein System eingeführt, dass jeden Ort der Parzellen in einer räumlichen Auflösung von 12.5 x 12.5 cm mittels Code beschreiben kann. Dazu eignet sich ein Quadtree-Ansatz, der eine rekursive Teilung in immer kleinere Einheiten unterstützt (vgl. Samet, 1984). Mit entsprechend angefertigten Rahmen kann diese Aufteilung nicht nur theoretisch auf Datenbankebene, sondern auch in der Natur, also direkt auf der Parzelle nachvollzogen werden.

Die Codierung weist gemäß schematischer Darstellung in *Abbildung 9* eine fünfstufige, hierarchische Gliederung auf. Mit diesem fünfstelligen Code lassen sich sämtliche Positionen innerhalb der Parzelle beschreiben. Nicht beanspruchte Hierarchiestufen weisen den Wert 0 auf – so hat beispielsweise eine 1 x 1 Meter große Fläche im unteren, rechten Eck der Parzelle die Codierung 33000, die entsprechende 0.5 x 0.5 Meter große Fläche, ebenfalls ganz unten rechts 33300. Eine exakte Lokalisierung ist beispielsweise dann notwendig, wenn Messungen und/oder Probennahmen an denselben Positionen in bestimmten Zeitintervallen wiederholt werden müssen.

Die Voraussetzung für eine spätere räumliche Zuordnung ist die Verwendung des Quadtree-Codes in den verschiedenen Tabellen des ClimGrass-Datenbankschemas. Die Positionierung von Sensoren bleibt mit dieser Information auch noch nach Jahren der Versuchstä-

tigkeit nachvollziehbar. Standortänderungen von Sensoren bieten dann möglicherweise Erklärungen für systematische Shifts in den Daten, falls solche auftreten sollten. Auch bestimmte Managementtätigkeiten können mit Hilfe dieser räumlichen Zuordnung genau spezifiziert werden.

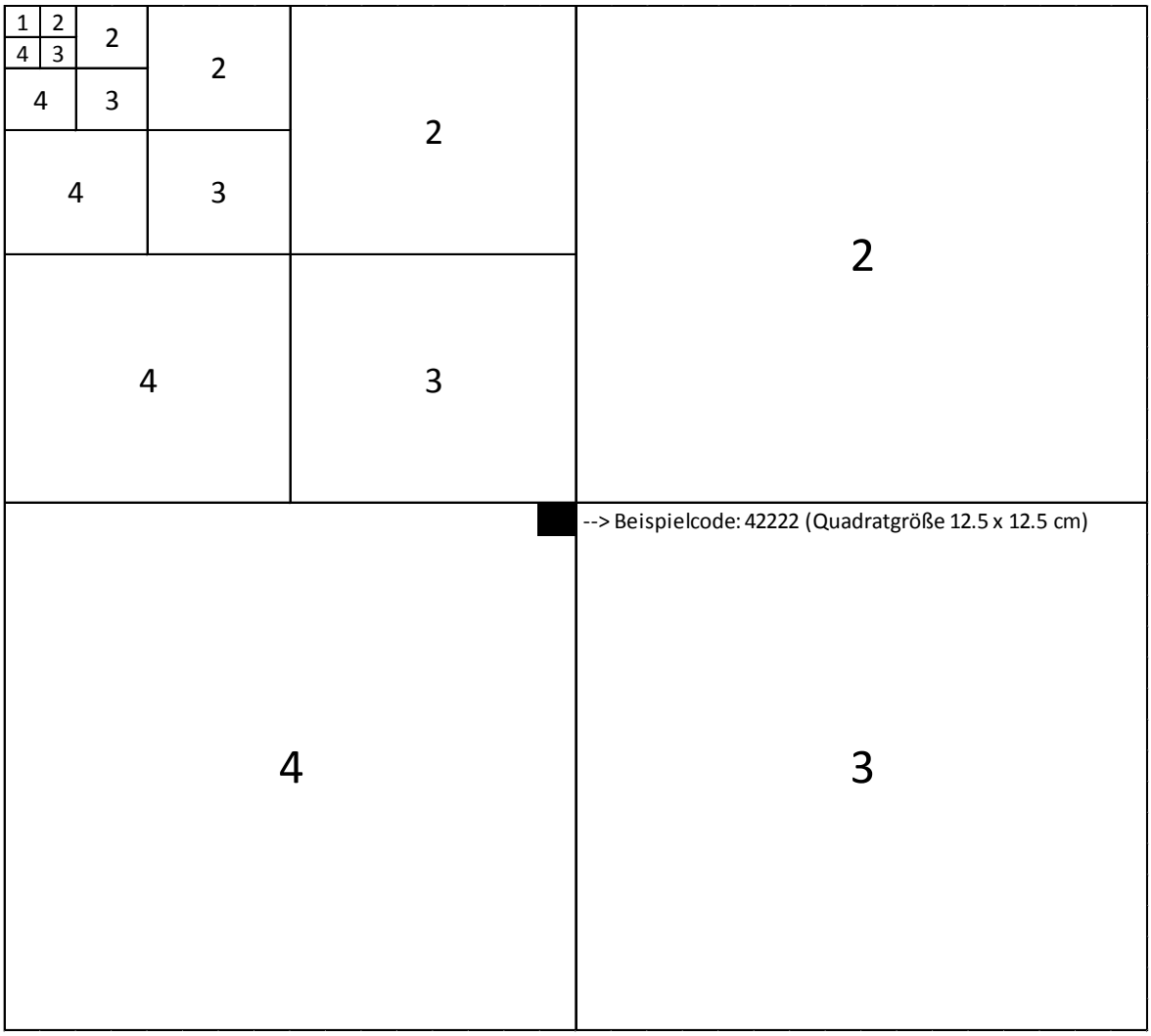

Abbildung 9: **Positionen innerhalb der Versuchsparzelle mit Hilfe einer Quadtree-Codierung (Ausrichtung am Versuchsfeld von Norden nach Süden – 11111 befindet sich in südöstlicher Himmelsrichtung)**

5 Datenmigrations- und Benutzerschnittstelle

5.1 Automatisierte Datenmigration

Sensorsysteme zur laufenden Beobachtung von Umweltparametern liefern in der Regel einen Datenstrom, der zu definierten Zeitintervallen Werte in vorbereitete Datenstrukturen speichert. Meist handelt sich um große Datenmengen, vor allem, wenn die Zeitintervalle sehr kurz sind. So werden beispielsweise 252 Temperatur- und CO_2-Regelungsdaten für ClimGrass rund um die Uhr im 10-Sekunden-Takt gespeichert. Lysimeterdaten werden im Minutentakt erfasst. Zunächst werden die Werte über Sensoren gemäß den Definitionen der Datenlogger-Software in Textdateien gespeichert. Diese Textdateien fassen meist alle

Daten eines Tages zusammen und werden einmal täglich, kurz nach Mitternacht, über eine in C# programmierte Zugriffsroutine mit File Transfer Protocol (FTP) vom ursprünglichen Speicherort abgeholt, datenmodellkonform umstrukturiert und in die Datenbank gespeichert. Der ganze Vorgang findet automatisiert statt und wird lediglich in einer Logdatei vermerkt, die von Zeit zu Zeit überprüft werden muss. Zurzeit werden in die ClimGrass-Datenbank täglich folgende Datenmengen übertragen:

- Lysimeterdaten: 32.112
- Wetterdaten der agrarmeteorologischen Station: 2.208
- Temperatur- und CO_2-Messwerte an den Versuchsparzellen: 9.771

Da durch eine automatisierte Erfassung die Datenmenge in kurzer Zeit generell sehr stark zunimmt, ist eine performanceoptimierte Strukturierung der Daten ein wichtiger Aspekt der Datenmodellierung für wissenschaftliche Applikationen.

5.2 Benutzerschnittstelle über Web-Browser

Die manuelle Eingabe bzw. die Übertragung von im Vorfeld aufbereiteten Massendaten erfolgt über eine Web-Browser-Benutzerschnittstellte. *Abbildung 10* zeigt die Hauptseite mit den Menüpunkten. Die ersten vier Funktionen dienen der Ein- und Ausgabe von Daten, mit den weiteren sechs Tasten werden Favoriten festgelegt und am Ende befinden sich Funktonen zur Datenbankadministration. Die Programmierung erfolgte in ASP.NET mit Zugriff auf die in einem SQL-Server gespeicherten Daten.

Abbildung 10: **Startseite des ClimGrass Data Warehouse mit Hauptmenü**

Die ClimGrass-Datenbank bietet die Möglichkeit, Daten aus den verschiedensten Fachbereichen oder über mehrere Projekte hinweg zu speichern (vgl. *Abbildung 11*). Dafür ist eine umfangreiche Liste an Parametern, Versuchseinheiten (z.B. Parzellen), Metadaten zur Probennahme, usw. notwendig. Damit einzelne Nutzer die mit der Zeit unübersichtlich großen Datensammlungen für ihre meist spezifischen Zwecke effizient nutzen können, bietet das System die Möglichkeit, die jeweils relevanten Tabelleneinträge zu selektieren und damit für die weitere Verwendung bei der Dateneingabe und -abfrage auf die notwendige Anzahl zu reduzieren. Bodenparameter werden beispielsweise zur Dateneingabe eines Bodenkundlers eher verwendet werden als die Parameter zur Bestimmung von Futterqualitäten oder anderer pflanzenbaulicher Kenngrößen.

Die ausgewählten Favoriten werden auf Dauer mit dem eingeloggten Benutzer verknüpft, sodass bei jedem neuerlichen Login auch die vormals definierte Favoritenliste automatisch geladen wird. Mit dieser Art „Customizing“ beeinträchtigt auch eine massive Ausdehnung der Datenbank auf andere Anwendungen und Projekte den eigenen Wirkungsbereich nicht.

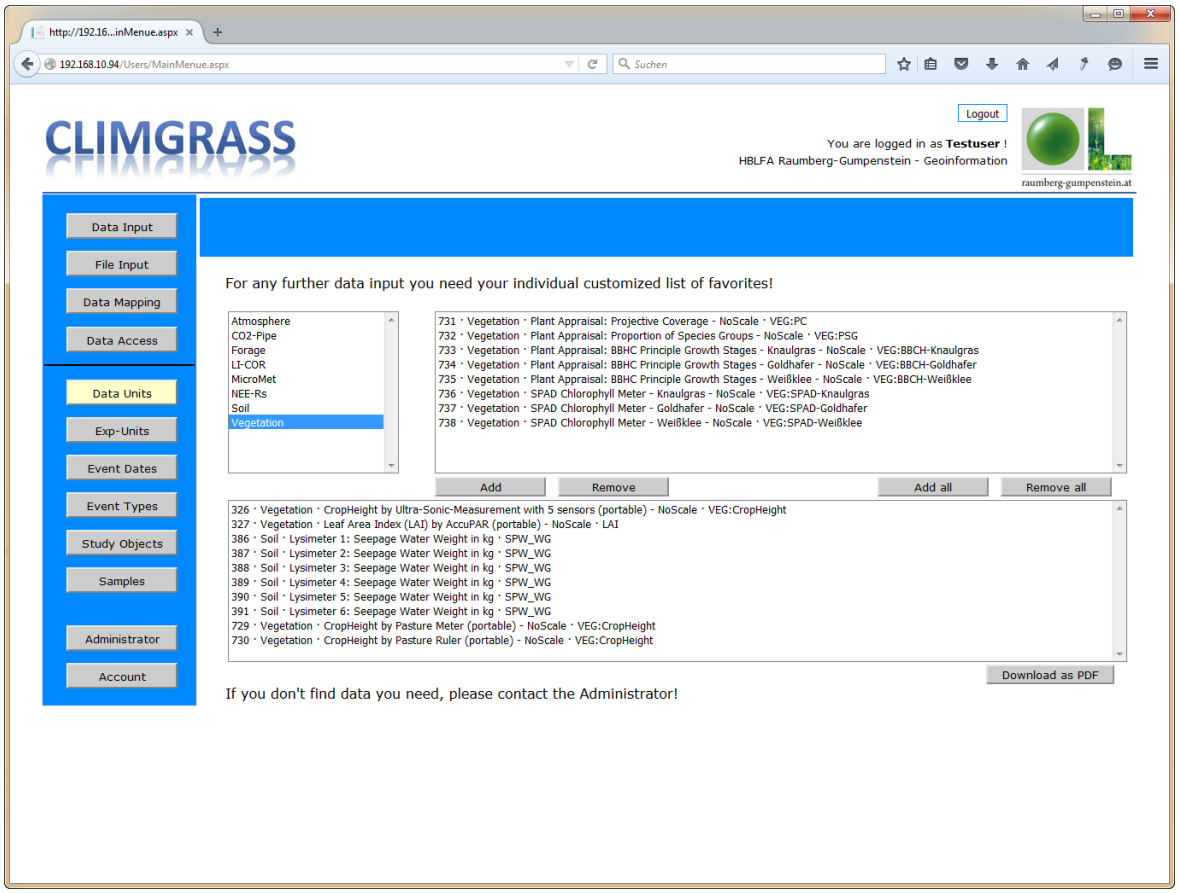

Abbildung 11: **Definition von benutzerangepassten Favoriten für statische Tabelleneinträge**

Bei der Dateneingabe, wie sie beispielhaft in *Abbildung 12* dargestellt ist, werden die Favoritenlisten den einzelnen Eingabefeldern hinterlegt und erleichtern damit das Auffinden und Selektieren gewünschter Listeneinträge. *Abbildung 12* zeigt die Erfassung von Metadaten einer Probennahme. Drop-Down-Listen wie „Exp.-Unit“ oder „Event Type“ enthalten die auf die Favoriten eingeschränkten Inhalte der Tabellen „ExperimentalUnit“ oder „EventType“.

Bereits eingegebene Daten werden in einer Übersichtstabelle angezeigt, in der Daten bearbeitet oder auch gelöscht werden können. Die Anzeige beschränkt sich allerdings auf jene

Datensätze, für die ein Nutzer die entsprechenden Rechte besitzt bzw. die von ihm selbst eingegeben wurden. Wird ein Datensatz zur Bearbeitung ausgewählt, werden alle Felder mit den bereits gespeicherten Daten gefüllt und können nun korrigiert werden.

Neben der manuellen Eingabe von „Samples“ in dem gezeigten Eingabeformular, wird auch eine Batch-Eingabe unterstützt. Dazu kann mittels „Export Template“ Button die für eine gültige Speicherung notwendige Datenstruktur in Form einer CSV-Datei exportiert werden. Mit Unterstützung durch Excel-Funktionen (z.B. automatisches Füllen) kann die Tabelle dann befüllt werden und durch den „Import Template“-Befehl in die Datenbank übertragen werden. Mit jeder Tabellenzeile wird ein neuer Datensatz erzeugt und gespeichert. Informationen über die korrekte Eingabe in den jeweiligen Feldern bietet eine PDF-Datei, welche mit dem Button „Instructions“ jederzeit aufgerufen werden kann.

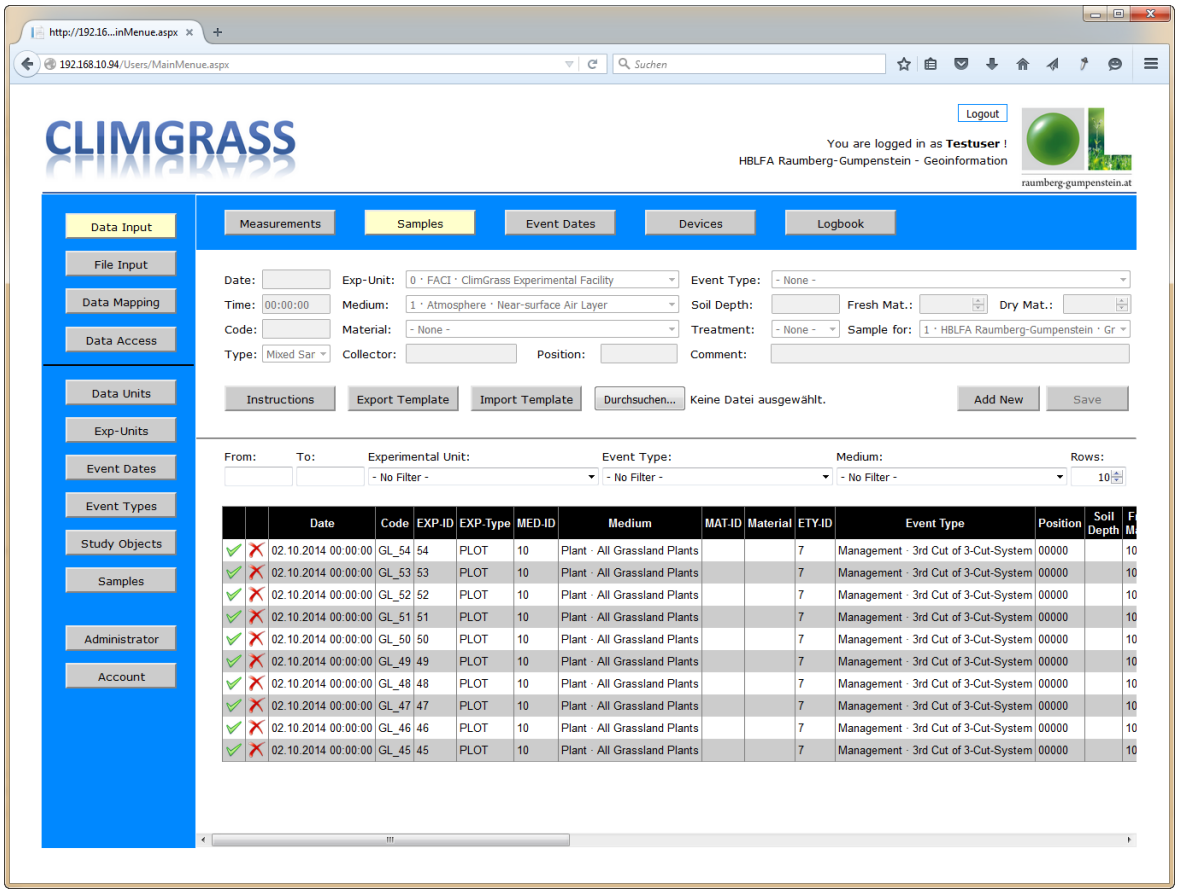

Abbildung 12: **Beispiel für die Dateneingabe mit einer Übersicht bereits gespeicherter Datensätze**

Eingaben zu „Measurements“, „Samples“ und „Event Dates“ können von sämtlichen Nutzern durchgeführt werden. Die Eingabefunktion „Devices“ und „Logbook“ lässt sich nur von eigens dafür berechtigten Personen aktivieren. Mit „Devices“ können beispielsweise alle verwendeten Geräte und Sensoren verwaltet werden. Das „Logbook“ dient der Aufzeichnung von Managementaktivitäten rund um die Experimente.

Die linke Menüspalte beinhaltet neben der manuellen Eingabe „Data Input“ auch eine Funktion mit „File Input“. Damit können beliebig struktierte CSV-Tabellen so angepasst werden, dass sie den Anforderungen des ClimGrass-Datenmodells entsprechen und so mit einem Arbeitsschritt in die Datenbank importiert werden. Dies ist die zu bevorzugende Art der Dateneingabe, da sie wesentlich effizienter ist, als die formularunterstützte manuelle Eingabe über „Data Input“ – „Measurements“.

Die Funktion „Data Mapping“ ermöglicht die Zuordnung von Mess- bzw. Beobachtungsdaten, die in der Tabelle „Measurements“ gespeichert sind, zu Events (Tabelle „Event Dates“) oder zu den Metadaten einer Probennahme (Tabelle „Samples“). Bei Events handelt es sich um Zeitpunkte und Perioden, die für bestimmte Daten maßgeblich ist. Beispielsweise ist die Aufwuchsdauer eines bestimmten Schnitts als Event definierbar und kann mit allen Daten verlinkt werden, die in diesem Zeitraum entstanden sind. Ebenso können beispielsweise die Laborergebnisse einer Bodenuntersuchung mit den Daten der Probennnahme verbunden werden. Jeder Analysewert kann damit auf die Methode, den Zeitpunkt oder den Verantwortlichen der Probennahme zurückgeführt werden.

Die produktive Nutzung der ClimGrass-Datenbank besteht darin, Daten für bestimmte Auswertungen aus der Datenbank abrufen zu können. *Abbildung 13* zeigt das Formular für die Datenauswahl. Nutzer können entsprechend ihrer Rechte auf die gespeicherten Daten unter Verwendung zahlreicher Filter zugreifen und die Abfrageergebnisse in unterschiedlichem Format als CSV-Dateien speichern. Als Filter dienen die auf die jeweiligen Favoriten eingeschränkten Tabelleneinträge bzw. Inhalte, für die ein Nutzer Zugriffsrechte hat. Parameter können dabei beliebig zusammengestellt und kombiniert werden. Alle ausgewählten Filter werden im Hintergrund zu einem entsprechenden SQL Select-Statement zusammengestellt, an die Datenbank übergeben und dort ausgeführt.

Da sämtliche Daten des ClimGrass-Experimentes zentral gespeichert werden, kann mit diesem Zugriffsinstrument eine Verknüpfung unterschiedlichster Informationen realisiert werden. Damit wird ein lückenloses Data Mining über den gesamten ClimGrass-Datenbestand für alle beteiligten Personen und Institutionen auf der Basis einer gemeinsamen Plattform unterstützt.

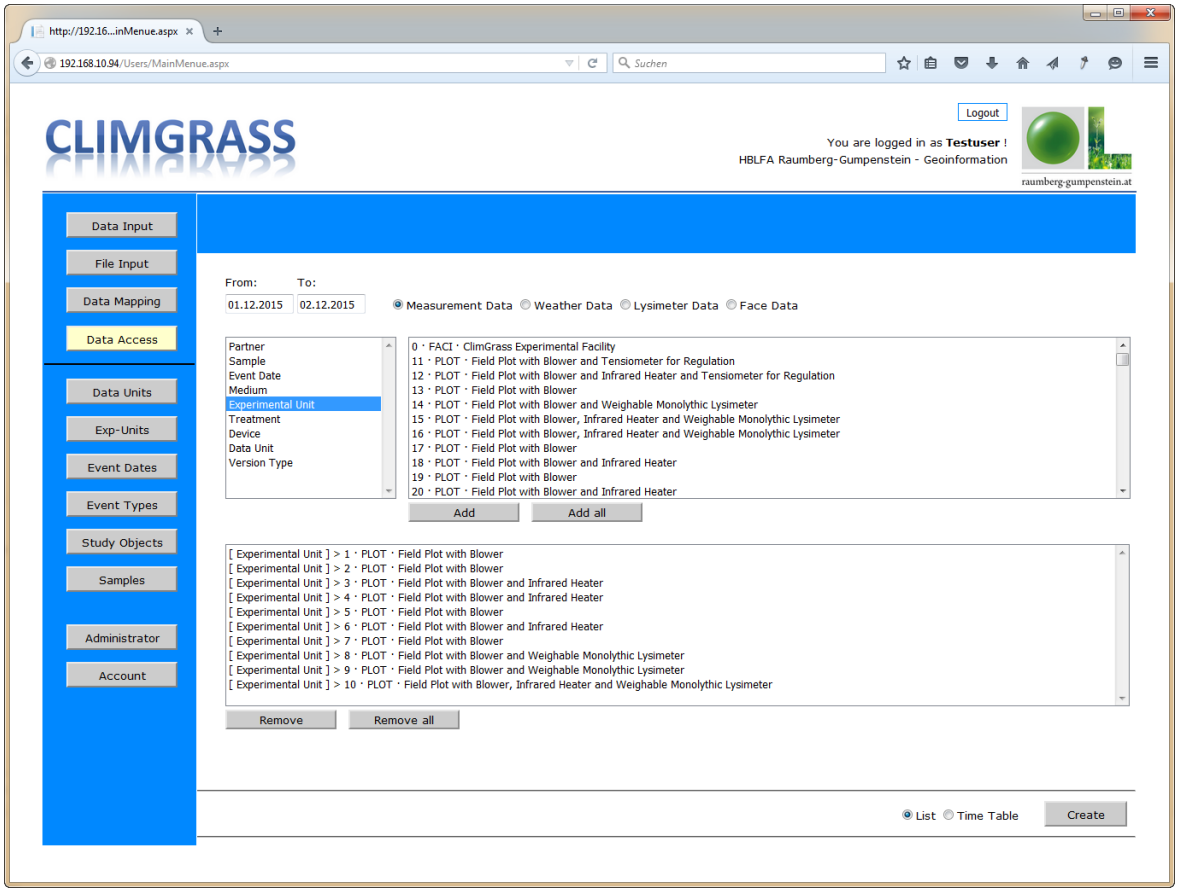

Abbildung 13: **Datenzugriff über menügesteuerte Selektion von Auswahlkriterien**

Jeder gespeicherte Wert ist mit der Kennung eines Projektpartners verbunden, sodass der gesamte Datenbestand dem jeweiligen Besitzer eindeutig zugeordnet werden kann. Benutzer bzw. Projektpartner als Datenbesitzer können aber auch die Rechte auf ihre eigenen Daten mit anderen teilen. Diese Rechtevergabe zwischen Datenbesitzern und anderen Datennutzern kann über die in *Abbildung 14* gezeigte Seite administriert werden. Dabei ist eine Aufgliederung der Zugriffsrechte bis auf Parameterebene möglich.

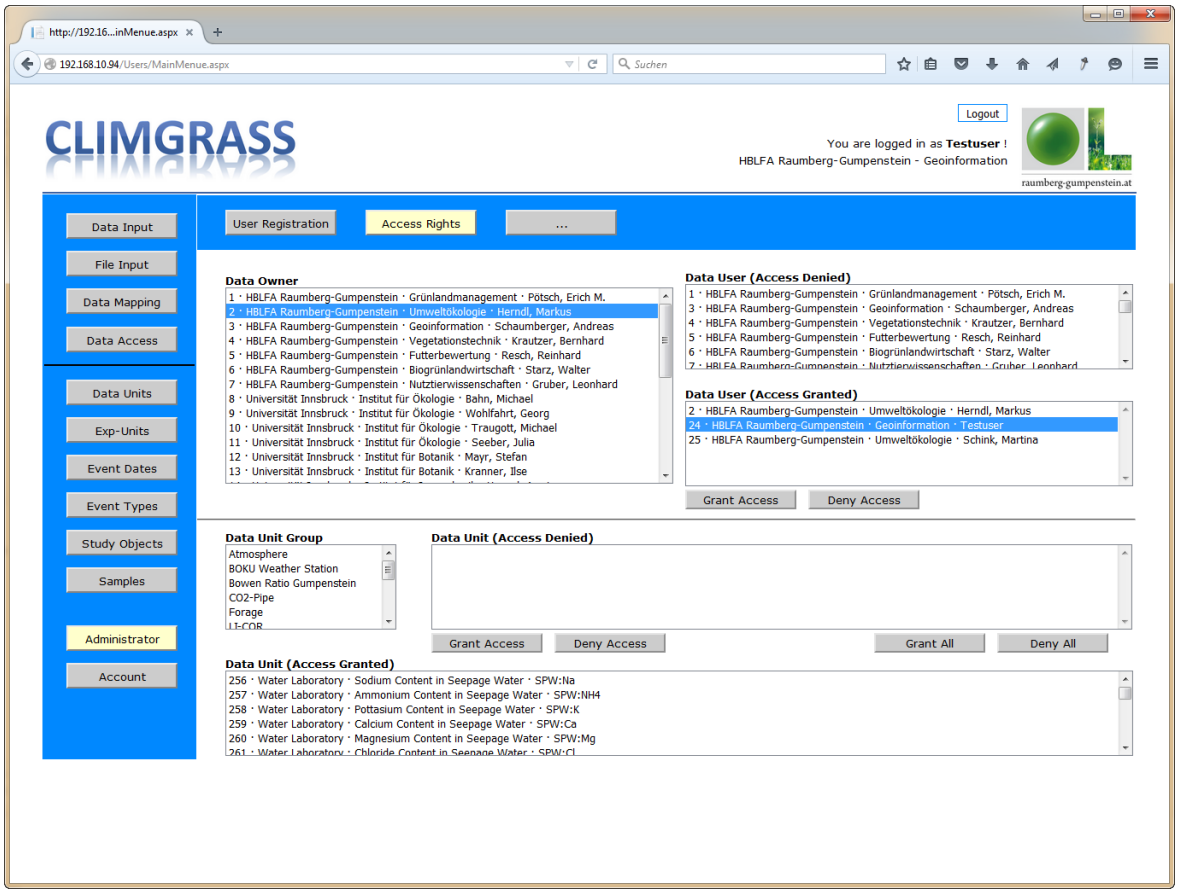

Abbildung 14: **Administrationsseite zur Definition von Benutzerzugriffen**

Die in diesem Abschnitt vorgestellten Abbildungen stellen keine vollständige Beschreibung der Benutzerschnittstelle dar, sondern zeigen lediglich das Konzept der Datenerfassung und -nutzung über eine Web-basierte Plattform. Die Komplexität des Datenmodells und seine streng relationale Implementierung erfordert zwangsläufig ein Bedienungswerkzeug. Für das verteilte Datenmanagement stellt eine Web-Oberfläche das optimale Instrument dar. Mit der Nutzung eines Web-Browsers muss keine eigene Software auf den Client-Computern installiert werden. Ebenso ist die Funktionalität auf den tatsächlichen Bedarf abgestimmt und deshalb auch einfach zu bedienen – eine Forderung, die im Zusammenhang von Scientific Databases immer wieder zur Sprache kommt.

Das ClimGrass Data Warehouse wurde in ASP.NET implementiert und dient als Schnittstelle zur ClimGrass SQL-Server-Datenbank. Die Benutzerdaten (Username, Passwort, Login-Daten, Benutzergruppe) und die jedem Benutzer zugeordnete Liste der Favoriten werden in eine SQL-Server-Express-Datenbank gespeichert, welche direkt am ClimGrass Internet Information Server gespeichert wird. Die Bearbeitung der Registrierung und die Zuordnung von Benutzern zu Projektpartner werden ebenso wie die Administration der Zugriffsrechte ausschließlich vom ClimGrass Administrator auf Anweisung der Datenbesitzer bzw. Projektpartner durchgeführt.

6 Datenanalyse und Datenkontrolle

Kontinuierliche Datenströme, wie sie mit Hilfe von Sensoren in oft sehr kurzen Zeitintervallen erfasst werden, führen in der Regel zu großen Datenmengen. Eine Sicherung der Datenqualität durch manuelle Prüfung ist nicht mehr möglich, da diese zum einen sehr viele Daten umfasst und zum anderen regelmäßig wiederholt werden müsste. Zeitreihen mit vielen Einzeldaten zeigen problematisches oder fehlerhaftes Verhalten sehr gut in grafischen Darstellungen (Trendlinien). Wenn beispielsweise Sensoren defekt sind und falsche Daten liefern, ist dies sofort im Kurvenverlauf ersichtlich. Die Aufbereitung der Daten in entsprechenden Diagrammen ist allerdings sehr zeitaufwendig.

Aus diesem Grund wird die technische Datenkontrolle bzw. die Beaufsichtigung der ClimGrass-Versuchsanlage mit täglich bzw. wöchentlich automatisch aufbereiteten Grafiken, wie sie beispielhaft in *Abbildung 15* dargestellt sind, vorgenommen. Die Implementierung und Visusalisierung erfolgt mit Windows Scripts und der Software Diadem von National Instruments (vgl. National Instruments, 2011). Die Daten werden aus der ClimGrass-Datenbank für den gewünschten Prüfzeitraum geladen und entsprechend den definierten Vorgaben (Report-Template) als PDF ausgegeben und abgespeichert. Zur Begutachtung wird an die Verantwortlichen eine Nachricht mit Link auf die erstellte PDF-Datei per mail geschickt. Auf diese Weise ist eine Kontrolle von vielen Daten in kurzer Zeit möglich. Oft kann ein Fehlverhalten eines einzelnen Sensors nur im direkten Vergleich mit anderen Zeitreihen erkannt werden. Eine Zusammenstellung mehrerer Parameter ist deshalb unbedingt notwendig.

Eine Kontrolle der Daten aus automatischen Erfassungssystemen ist bei einer Verwendung für wissenschaftliche Auswertungen zwingend notwendig und deshalb äußerst wichtig.

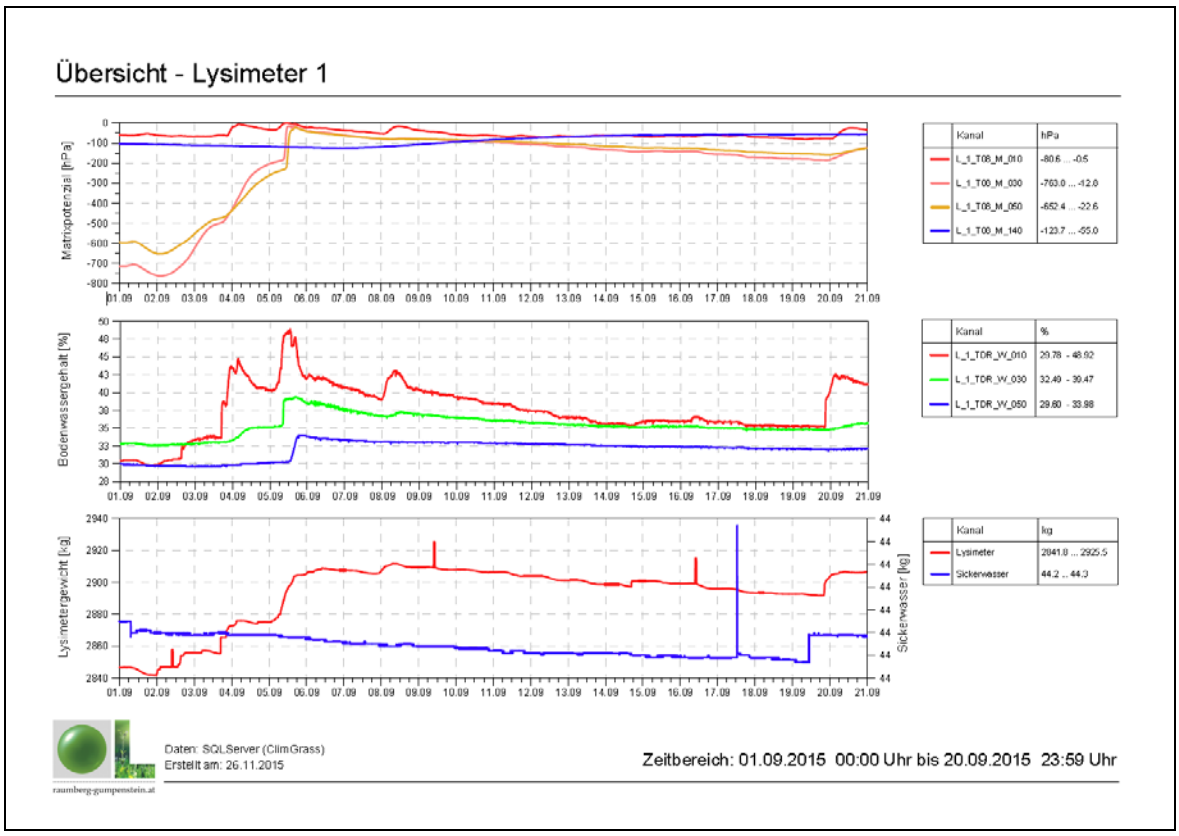

Abbildung 15: **Beispiel für eine Auswertung kontinuierlicher Datenströme zur Qualitätskontrolle**

7 Schlussfolgerungen und Ausblick

Datenintensive Forschungsprojekte verlangen zwingend nach einer intensiven Auseinandersetzung mit Datenmodellen und Datenbanken. Die vielfach in der Praxis verwendeten Ansätze, Daten dateibasiert, oft sogar mit Microsoft Excel oder Statistikpaketen, zu organisieren ist, dann vollkommen ungeeignet, wenn beim Datenmanagement für wissenschaftliche Zwecke unter anderem folgende Ziele verfolgt werden sollen:

- Vollständig automatisierte Workflows für Sensordaten.
- Sicherung der Datenqualität durch technische Kontrollmaßnahmen und Erhaltung der referentiellen Integrität im Rahmen relationaler Datenmodelle.
- Integrierte Speicherung von Metadaten zur Erhebung und Verarbeitung aller Daten.
- Kosteneinsparung durch Erhaltung des Datenbestandes über Projektlaufzeiten hinaus.
- Explorative Forschungsansätze auf der Grundlage standardisierter Datenverfügbarkeit.
- Verteiltes Datenmanagement für mehrere Projektpartner.
- Entwicklung von Software für einen bedarfsoptimierten Datenzugriff (schreibend und lesend) als Schnittstelle zur Datenbank.
- Nutzung von integrierten Funktionen eines Datenbankmanagementsystems (Sicherung, Benutzerrechte, standardisierte Zugriffe (SQL) und Routinen (DML).

Die Besonderheiten wissenschaftlicher Datenbanken müssen bei der Konzeption von dafür geeigneten Datenmodellen berücksichtigt werden. Forschungsprojekte sind in der Regel keine statischen Konstrukte, sondern flexibel in Ausrichtung und Umsetzung. Neue und zusätzliche Forschungsansätze im Laufe eines Projektes sind ebenso üblich wie Änderungen bei bereits laufenden Experimenten. Dazu kommen noch fallweise Veränderungen der Projektpartnerstruktur. Wenn Datenmodelle nicht in der Lage sind, auf diese Dynamik so zu reagieren, dass die bis dahin gespeicherten Daten unangetastet bleiben, ist mit jeder kleinen Änderung ein sehr hoher Aufwand verbunden. Datenmodelle müssen also sowohl hinsichtlich struktureller Anpassungen als auch hinsichtlich der gespeicherten Datenmenge frei skalierbar sein. Die Performance muss unabhängig von der Datenmenge auf einem konstant hohen Niveau bleiben.

WissenschaftlerInnen im Agrarsektor sind mit fachlichen Themen naturgemäß besser vertraut als mit technischen Belangen. Vor allem der Einsatz relationaler Datenmodelle setzt viel Know-how im Bereich Datenmodellierung und Datenbankmanagement voraus. Dies stellt definitiv eine Hürde für die Arbeit mit derartigen Systemen dar. Bei Projekten, in welchen unter anderem auch mit einer automatisierten Datenerhebung über Sensoren gearbeitet wird, sollte das unbedingt berücksichtigt werden. Dateibasierte Datenhaltung in heterogenen Strukturen ist in solchen Fällen zum Scheitern verurteilt. Leider stellen sich die Nachteile einer strukturlosen Datenhaltung erst ab einer kritischen Datenmenge ein. Während wenige Daten noch mit gewohnten Instrumenten und Softwaretools (z.B. Excel) handhabbar sind, ist die Weiterverarbeitung der kontinuierlich mit Sensoren generierten Daten in Abhängigkeit ihrer Menge mit zunehmender Ineffizienz zu bewerkstelligen. Strukturelle Änderungen sind dann aufgrund der bereits aufgelaufenen großen Datenmenge kaum oder nur mit größten Schwierigkeiten durchzusetzen und erfordern unter Umständen auch tiefgreifende Einschnitte in die tägliche Arbeit.

Fehlende Datenmanagementkonzepte erfordern bei datenintensiven Forschungsansätzen einen überproportional hohen Ressourceneinsatz, um bei der Verarbeitung und Auswertung mit der wachsenden Datenmenge schritthalten zu können. Zudem ist ein geeignetes Datenmodell, eingebettet in einem Datenbankmanagementsystem, die Voraussetzung für eine signifikante Verlängerung der „Lebensdauer" von wissenschaftlichen Daten sowie einer verteilten Datenerfassung und -abfrage.

Je mehr ein Datenmodell die dort gespeicherten Daten im Sinne einer relationalen Datenhaltung aufteilt, desto schwieriger wird die Deutung von Daten in den einzelnen Relationen. Die Spalten solcher Tabellen sind oft nur noch mit „Unique Identifier" gefüllt, welche die Verknüpfung mit Inhalten in anderen Tabellen repräsentieren. Nur mittels Beziehungen zwischen den Relationen erschließt sich so die Semantik der dort verteilt gespeicherten Information. Dies ist zwar die Voraussetzung für eine hohe Performance von Datenzugriffen und gewährleistet auch die von der Datenmenge unabhängige Skalierbarkeit, führt jedoch dazu, dass eine Lesbarkeit und Interpretation der Daten nur dann möglich ist, wenn diese in Abfragen und Sichten über mehrere Tabellen hinweg wieder zusammengeführt werden. Um Daten effizient eingeben und abfragen zu können, ist deshalb eine benutzerfreundliche Schnittstelle unumgänglich, welche alle notwendigen und zusammenhängenden Daten im Hintergrund entsprechend aufbereitet.

Eine geeignete Software muss dafür sorgen, dass alle Daten, sowohl jene, welche automatisch von Sensoren in die Datenbank überführt werden müssen, als auch jene, welche manuell eingegeben werden, in die richtigen Strukturen fließen. Ohne Softwareunterstützung ist dies bei komplexen Datenmodellen nicht oder nur mit unvertretbar hohem Aufwand möglich. Eine weitere technische Unterstützung ist bei der Kontrolle von Datenreihen notwendig. Automatische Datenabfragen für definierte Prüfzeiträume und eine übersichtliche grafische Aufbereitung sind für eine laufende Qualitätssicherung unverzichtbar.

Die Arbeit an dem hier vorliegenden Datenmodell hat mehrere Jahre in Anspruch genommen und ist immer noch einer ständigen Weiterentwicklung unterworfen. Die Erfahrung hat gezeigt, dass im Laufe eines Projektes immer neue Anforderungen an das Datenmodell gestellt werden, die entsprechende Erweiterungen notwendig machen. Der evolutionäre Charakter eines Datenmodells im Rahmen von Forschungsprojekten setzt die Möglichkeit der Anpassung bzw. der Adaption voraus. Entscheidend ist dabei die grundsätzliche Konzeption – Beziehungen zwischen Relationen sind von wesentlich größerer Bedeutung als die Felddefinitionen und Datentypen in den einzelnen Tabellen.

Während die Erhebung von Daten nach methodischen Vorgaben erfolgt, ist die Datenauswertung und -aufbereitung ein explorativer Prozess, der die Grundlage für die Generierung von wissenschaftlichen Ergebnissen bildet. Er kumuliert in der Ausarbeitung von Publikationen und Vorträgen. Die Agrarwissenschaft ist mit unterschiedlichen Disziplinen besetzt, die unterschiedliche Daten erzeugen. Systeme, welche Daten über einzelne Fachbereiche hinweg in homogene Strukturen integrieren, bilden eine hervorragende Möglichkeit, interdisziplinäre Zusammenhänge zu entdecken. Mit der Möglichkeit einer Analyse und Synthese vieler Informationen, welche sehr effizient aus einem zentralen Datenbestand nahezu beliebig zusammengestellt werden können, kommt einem ausgereiften Datenmanagement ein hoher Stellenwert zu und ist neben der Publikationstätigkeit als zentrales Instrument wissenschaftlicher Arbeit anzusehen.

8 Literatur

Ailamaki, A.; Kantere, V. und Dash, D. (2010): Managing scientific data. *Communications of the ACM* **53** (6), *ACM*, New York, USA, 68-78.

Arzberger, P.; Schroeder, P.; Beaulieu, A.; Bowker, G.; Casey, K.; Laaksonen, L.; Moorman, D.; Uhlir, P. und Wouters, P. (2004): An International Framework to Promote Access to Data. *Science* **303** (5665), 1777-1778.

Borgman, C.; Wallis, J. und Enyedy, N. (2006): Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology. In Gonzalo *et al.* (Eds.): Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science, 4172, *Springer Berlin Heidelberg*, 170-183.

Brunt, J.W. (2000): Data Management Principles, Implementation and Admninistration. In Michener und Brunt (Eds.): Ecological Data: Design, Management and Processing, Methods in Ecology, *Blackwell Science Ltd*, 25-47.

Deelman, E. und Chervenak, A. (2008): Data Management Challenges of Data-Intensive Scientific Workflows. 8th IEEE International Symposium on Cluster Computing and the Grid, Lyon, France, May 19-22, 2008, 687-692.

Diekmann, F. (2012): Data Practices of Agricultural Scientists: Results from an Exploratory Study. *Journal of Agricultural & Food Information* **13** (1), 14-34.

Fayyad, U.; Piatetsky-Shapiro, G. und Smyth, P. (1996a): From Data Mining to Knowledge Discovery in Databases. *AI Magazine* **17** (3), 37-54.

Fayyad, U.; Haussler, D. und Stolorz, P. (1996b): Mining scientific data. *Communications of the ACM* **39** (11), 51-57.

Heidorn, P.B. (2009): Shedding Light on the Dark Data in the Long Tail

of Science. *Library Trends* **57** (2, Fall 2008), 280-299.

Hine, C. (2006): Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work. *Social Studies of Science* **36** (2), 269-298.

Hunt, L.A.; White, J.W. und Hoogenboom, G. (2001): Agronomic data: advances in documentation and protocols for exchange and use. *Agricultural Systems* **70** (2–3), 477-492.

Janssen, S.; Van Kraalingen, D.W.G.; Boogaard, H.L.; De Wit, A.J.W.; Franke, J.; Porter, C. und Athanasiadis, I.N. (2012): A generic data schema for crop experiment data in food security research. International Congress on Environmental Modelling and Software, Managing Resources of a Limited Planet: Pathways and Visions under Uncertainty, Sixth Biennial Meeting, Leipzig, Germany, International Environmental Modelling and Software Society (iEMSs), July 1-5, 2012, 2447-2454.

National Instruments (2011): DIAdem: Daten finden, analysieren und dokumentieren, *National Instruments Ireland Resources Limited*, no.com, 131 S.

Porter, J.R. (2000): Scientific Databases. In Michener und Brunt (Eds.): Ecological Data: Design, Management and Processing, Methods in Ecology, *Blackwell Science Ltd*, 48-69.

Samet, H. (1984): The Quadtree and Related Hierarchical Data Structures. *ACM Computing Surveys* **16** (2), 187-260.

Van den Eynden, V.; Corti, L.; Woollard, M.; Bishop, L. und Horton, L. (2011): Managing and Sharing Data: Best Practice for Researchers. Essex, UK Data Archive, University of Essex, Essex, 35 S.