Integrierte Analyse pflanzenbiologischer Daten unter besonderer Berücksichtigung der Datenqualität

Dissertation

zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat)

vorgelegt der Technischen Fakultät
der Universität Bielefeld

von Dipl.-Wirtsch.-Inf. Stephan Weise
geb. am 10. November 1975 in Dessau

Stephan Weise:
Integrierte Analyse pflanzenbiologischer Daten unter besonderer Berücksichtigung der Datenqualität

Der Technischen Fakultät der Universität Bielefeld vorgelegt,
am 3. November 2009 verteidigt und genehmigt.

Gutachter:
Prof. Dr. R. Hofestädt, Universität Bielefeld
Prof. Dr. A. Graner, IPK Gatersleben

Promotionskommission:
Prof. Dr. K. Friehs, Universität Bielefeld
Prof. Dr. R. Hofestädt, Universität Bielefeld
Prof. Dr. A. Graner, IPK Gatersleben
Dr. J.M. Risse, Universität Bielefeld

187 Seiten
51 Abbildungen
11 Tabellen

Gedruckt auf alterungsbeständigem Papier (DIN ISO 9706)

Danksagung

Die vorliegende Arbeit entstand während meiner Tätigkeit am Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK) Gatersleben. Sie wurde durch die vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projekte Bioinformatics Centre Gatersleben-Halle (BIC-GH) sowie GABI-GENOBAR ermöglicht.

Mein Dank gilt an erster Stelle Herrn Prof. Dr. Ralf Hofestädt von der Technischen Fakultät der Universität Bielefeld, der mir als betreuender Hochschullehrer die Anfertigung dieser Arbeit überhaupt erst ermöglicht hat. Herrn Prof. Dr. Andreas Graner, geschäftsführender Direktor des IPK Gatersleben, danke ich für die Übernahme des Zweitgutachtens. Weiterhin möchte ich meinem Arbeitsgruppenleiter, Herrn Dr. Uwe Scholz, für die beständige Motivation und die gewährten Freiheiten danken.

Frau Dr. Inge Matthies gilt mein Dank für die hervorragende Zusammenarbeit während der praktischen Umsetzung der in dieser Arbeit entwickelten Konzepte sowie für umfassende fachliche Diskussionen.

Ich danke allen Kolleginnen und Kollegen, insbesondere Steffen Flemming, Dr. Björn Junker, Prof. Dirk Koschützki, Christian Künne, Dr. Matthias Lange, Dr. Marion Röder, Roland Schnee, Karl Spies, Andreas Stephanik, Burkhard Steuernagel und Thomas Thiel, die die Erstellung dieser Arbeit durch Diskussionen, konstruktive Hinweise und Korrekturlesen in besonderer Weise unterstützt haben.

Abschließend möchte ich mich bei meiner Familie und insbesondere bei meiner Frau Daniela und meinem Sohn Elias bedanken, ohne deren Rückhalt und Motivation die Anfertigung dieser Arbeit nicht möglich gewesen wäre.

Kurzfassung

Innerhalb der letzten Jahre hat sich das Datenaufkommen in der Biologie exponentiell vervielfacht. Der Einsatz moderner Hochdurchsatzmethoden verdrängt die traditionelle Art der Forschung in zunehmendem Maße. Der wissenschaftliche Fokus bewegt sich dabei von der Untersuchung einzelner Datendomänen (Bereiche) und problemorientierter Arbeit hin zur domänenübergreifenden und ergebnisoffenen Analyse.

Obwohl bioinformatische Werkzeuge die Datenanalyse in hohem Maße unterstützen, sind umfangreiche, experimentell gewonnene Datensätze nur noch schwer manuell zu handhaben. Dies trifft insbesondere auf genetische Daten zu. Daher ist in den letzten Jahren eine große Anzahl verschiedener Informations- und Analysesysteme entwickelt worden. Der Fokus dieser Systeme liegt häufig nur auf einer Datendomäne; eine integrierte Analyse fehlt vielfach. Unter dem Begriff der integrierten Analyse wird die Zusammenführung (Integration) von Daten aus verschiedenen Domänen mit dem Ziel der gemeinsamen Auswertung verstanden.

Während der Blickpunkt der wissenschaftlichen Gemeinschaft vorrangig auf der Erforschung des Menschen und von Organismen wie der Fruchtfliege oder der Maus liegt, sind Pflanzen vielfach unterrepräsentiert. Dies betrifft sowohl die Gewinnung als auch die Speicherung und Auswertung von Daten. Pflanzen haben jedoch als Nahrungsgrundlage für Mensch und Tier sowie als erneuerbare Energiequellen eine große Bedeutung.

Das Ziel der vorliegenden Arbeit besteht in der Entwicklung eines Konzepts zur Ermöglichung der flexiblen, integrierten Analyse pflanzenbiologischer Daten. Dabei sollen potenzielle Wechselwirkungen zwischen den Datendomänen, beispielsweise Genotyp-Phänotyp-Korrelationen, aufgezeigt werden. Eine wichtige Voraussetzung hierfür ist die Qualität der zugrunde liegenden Daten, insbesondere die Vergleichbarmachung von Daten aus unterschiedlichen Quellen. Hierzu werden spezifische Herausforderungen und Lösungsansätze diskutiert sowie existierende Ansätze betrachtet. Es werden Vorschläge zur integrierten Analyse biologischer Daten unter Berücksichtigung von Spezifika der Pflanzenbioinformatik erarbeitet und erörtert. Besonderes Gewicht wird dabei auf Qualitätsaspekte dieser Daten sowie auf domänenübergreifende Analysemöglichkeiten gelegt. Diese Problematiken werden durch existierende Ansätze in dieser Wissensdomäne häufig nicht zufriedenstellend berücksichtigt.

Als Ergebnis der Arbeit wird der Entwurf eines Konzepts zur integrierten Analyse pflanzenbiologischer Daten unter Verwendung von Datawarehouse-Methoden präsentiert. Anschließend werden die einzelnen Elemente des Konzepts anhand eines Prototypen erläutert. Dieser dient der Integration von phänotypischen, genetischen und Passportdaten von Braugerstensorten mit dem Ziel der flexiblen Durchführung von Assoziationsstudien zur Aufdeckung potentieller Genotyp-Phänotyp-Korrelationen. Besondere Beachtung erfährt dabei die Sicherstellung einer hohen Datenqualität.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Ausgehend von einer Motivation wird in diesem Kapitel die Zielstellung der vorliegenden Arbeit beschrieben. Daran anschließend erfolgt die Vorstellung der Gliederung der Arbeit.

1.1 Motivation und Zielsetzung

Das Datenaufkommen in der Biologie hat sich durch den zunehmenden Einsatz moderner Hochdurchsatzverfahren in den letzten Jahren vervielfacht [Pen05, Kit02, Aug01, Roo01]. Zur Verwaltung von in molekularbiologischen Experimenten gewonnenen Daten wurde eine Vielzahl von Informationssystemen [GC09] geschaffen.

Daneben gibt es große Mengen biologischer Daten, die nicht aus dem molekularen Bereich stammen, beispielsweise phänotypische Daten. Auch diese Daten werden zunehmend elektronisch gehalten und zentral in Datenbanken gespeichert [MPL01, Knü01, MGF98].

Der Fokus der naturwissenschaftlichen Gemeinschaft liegt dabei vorrangig auf der Erforschung des Menschen und anderer Organismen wie der Fruchtfliege oder der Maus. Pflanzen sind oftmals unterrepräsentiert. Davon betroffen sind neben der Gewinnung auch die Speicherung und Analyse von Daten [GBWK$^+$08]. Dieses Defizit steht im Gegensatz zum Nutzwert.

Pflanzen erfahren eine stetig wachsende Bedeutung. Neben ihrer Verwendung als Nahrungsgrundlage für Mensch und Tier kommen Pflanzen zunehmend auch als erneuerbare Energiequellen zum Einsatz [NIB07, THL06]. Weitere Potenziale liegen in der

Nutzung als Grundlage neuer Arzneimittel oder für die chemische Industrie [MB06, SGGC01, Del99].

Die Quellen pflanzenbiologischer Daten zeichnen sich häufig durch Heterogenitäten bezüglich ihrer Anwendbarkeit, Struktur oder ihres Inhaltes aus. Hinzu kommen Daten in proprietären Formaten wie beispielsweise Spreadsheets mit einer großen Vielfalt an Formatierungen. Das Kombinieren dieser Daten erfolgt über Datenintegration. Dabei können und müssen Heterogenitäten aufgelöst werden.

Grundsätzlich bieten sich zur Auswertung solcher Daten zwei Vorgehensweisen an, die hypothesengetriebene und die datengetriebene Analyse. Im ersten Fall steht die Überprüfung von Hypothesen anhand von Stichproben im Mittelpunkt. Dieser Ansatz wird auch als modellgetriebene Analyse bezeichnet. Die zweite Herangehensweise beschäftigt sich mit der explorativen, ergebnisoffenen Untersuchung großer Datenmengen, um neues Wissen zu gewinnen.

Obwohl umfangreiche Datensätze aus verschiedenen Datendomänen, beispielsweise Marker-, Sequenz- und Expressionsdaten, Charakterisierungs- und Evaluierungsdaten, zur Verfügung stehen, erfolgt die Betrachtung dieser Daten häufig isoliert voneinander [PK06, Sea03]. Das sich aus der domänenübergreifenden Untersuchung von Daten ergebende Potenzial (z. B. Genotyp-Phänotyp-Korrelationen) wird in der pflanzlichen Forschung noch nicht hinreichend genutzt.

Abbildung 1.1 zeigt eine schematische Darstellung der potenziellen Verknüpfungspunkte zwischen Datendomänen aus dem pflanzenbiologischen Bereich. Im Mittelpunkt stehen hierbei die so genannten Passportdaten, die Beschreibungen eines (abstrakten) Objektes Pflanze enthalten und als Bindeglied zwischen verschiedenen Datendomänen angesehen werden können.

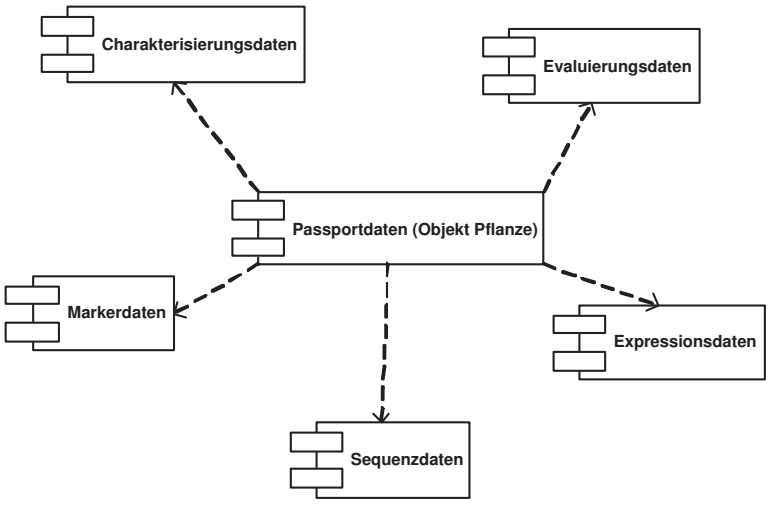


Abbildung 1.1: Beziehungen zwischen verschiedenen Datendomänen

Die vorliegende Arbeit beschäftigt sich mit Herausforderungen und Lösungsansätzen zur integrierten Analyse pflanzenbiologischer Daten. Unter Integration ist die Erlangung des Zugriffs auf eine Anzahl existierender, heterogener Datenquellen und die Verfügbarmachung der Daten über eine zentrale, integrierte Komponente zu verstehen. Der Begriff der integrierten Analyse beschreibt darüber hinaus die Integration und gemeinsame Auswertung von Daten aus verschiedenen Domänen. Um dieses Ziel zu erreichen, wird im Rahmen der Arbeit ein Konzept entwickelt, das die Besonderheiten pflanzenbiologischer Daten berücksichtigt. Einen Schwerpunkt bildet dabei die Sicherstellung hoher Datenqualität. Anhand einer praktischen, biologischen Fragestellung mit wirtschaftlicher Relevanz soll die Anwendbarkeit demonstriert werden.

Gerste (*Hordeum vulgare*) gehört zu den wichtigsten Kulturpflanzen und wurde im Jahr 2000 in Deutschland auf über 18% der landwirtschaftlichen Nutzfläche angebaut [Ref07]. Sie ist sehr anpassungsfähig an verschiedene, auch extreme Umweltbedingungen und findet als Lebens- und Futtermittel sowie als Braugerste vielseitige Verwendung. Deutschland ist nach der Russischen Föderation und Kanada der weltweit drittgrößte Gerstenproduzent [FAO05], wobei ein großer Teil der Sommergersten für die Bierherstellung verwendet wird. Malz- und Brauqualität gehören daher zu den kommerziell wichtigsten Merkmalen von Gerste. Aufgrund der hohen Erzeugerpreise nimmt der Braugerstenanbau zu [ZMP08].

Die gezielte Züchtung von Braugerstensorten mit besserer Malzqualität kann durch die Ermittlung signifikanter Assoziationen von Single-Nucleotide-Polymorphismen (SNP) und Haplotypen-Mustern mit Malz- und Brauqualitätsmerkmalen erheblich erleichtert werden. Dies wird auch als markergestützte Selektion (Marker Assisted Selection, MAS) bezeichnet. Informationen über solche Zusammenhänge ermöglichen eine Zeit- und Kostenersparnis im Zuchtprozess.

Das grundsätzliche Vorgehen dieses Anwendungsfalls ist in Abbildung 1.2 als Workflow dargestellt. Pflanzliche Daten aus verschiedenen Domänen werden experimentell erhoben und in diversen Formaten auf unterschiedlichen Medien gespeichert. Aus diesen Rohdaten werden durch Auswertungen teilweise sekundäre Daten abgeleitet. Um domänenübergreifende Analysen durchzuführen, ist die Integration relevanter Daten erforderlich.

An jeder Position des eben skizzierten Workflows können Fehler auftreten, die die Qualität der Auswertungen nachhaltig negativ beeinflussen. Das Spektrum der Fehler reicht von unzureichend dokumentierten Experimenten, die die Vergleichbarkeit von Daten vermindern [Mem05,Irr05], über Sequenzierfehler [HHM+07,CW92] bis hin zu Problemen bei der Verarbeitung von Rohdaten, die zu Inkonsistenzen in abgeleiteten Daten führen [GBA04, MNF03].

Im Rahmen dieser Arbeit existieren nur begrenzte Möglichkeiten, Einfluss auf die Erhebung von Daten zu nehmen. Dies gilt insbesondere für die Gewinnung von Primärdaten, etwa durch Sequenzierautomaten. Daneben ist davon in gewissem Umfang auch

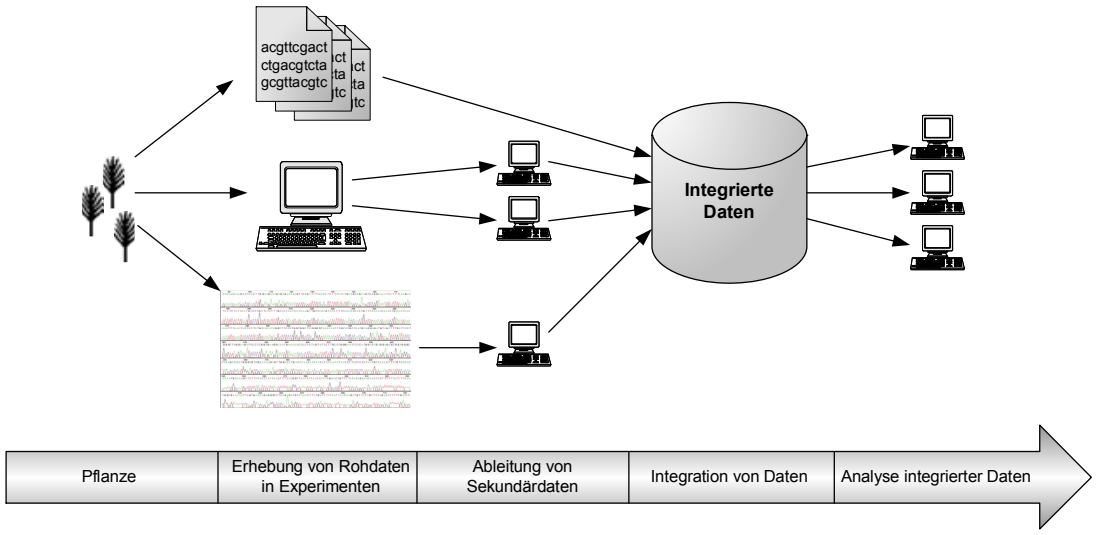


Abbildung 1.2: Schematische Darstellung des Anwendungsfalles

die Generierung sekundärer Daten betroffen. Diese Daten müssen als gegeben hingenommen werden.

Der Fokus wird daher auf die Erkennung und Behandlung von Inkonsistenzen während der Integration von Daten gerichtet sein. Dies kann u. a. das Herausfiltern potenziell fehlerhafter Daten umfassen, um die Qualität von Analyseergebnissen zu verbessern.

1.2 Gliederung der Arbeit

Im folgenden Kapitel 2 der vorliegenden Arbeit werden Grundlagen aus der Informatik und der Biologie vorgestellt. Hierzu zählen insbesondere Datenbanksysteme, Möglichkeiten der Modellierung sowie pflanzenbiologische Datendomänen und Ressourcen, die für die vorliegende Arbeit von Bedeutung sind. Weiterhin wird auf fachübergreifende Grundlagen wie kontrollierte Vokabulare, Taxonomien und Ontologien eingegangen.

Die Vorstellung weiterer für die Arbeit notwendiger Konzepte erfolgt in Kapitel 3. Hier werden neben verschiedenen Ansätzen zur Datenintegration Möglichkeiten der Analyse großer Datenmengen vorgestellt. Dabei wird besonderer Wert auf Methoden des Dataminings gelegt.

In Kapitel 4 erfolgt die Auseinandersetzung mit der oftmals unzureichenden Datenqualität in der Pflanzenbioinformatik. Es wird eine Klassifikation der Ursachen vorgeschlagen und anhand von Beispielen illustriert. Anschließend werden Empfehlungen

zur Verbesserung der Qualität im Rahmen der gesamten, oben beschriebenen Pipeline gegeben.

Eine Analyse verschiedener existierender Systeme wird in Kapitel 5 durchgeführt. Hierzu werden die zehn in [Sch02] entwickelten Kriterien zur Bewertung von Integrationsansätzen um sieben weitere Merkmale ergänzt. Diese dienen dazu, zusätzlich auch das Analysepotenzial der untersuchten Systeme zu bewerten.

Basierend auf dieser Untersuchung und den Überlegungen zur Datenqualität in Kapitel 4 wird in Kapitel 6 ein Konzept zur flexiblen, integrierten Analyse pflanzenbiologischer Daten entwickelt. Im Mittelpunkt stehen dabei ein Datenpool, der Daten aus verschiedenen Domänen hält, sowie analysespezifische Datamarts.

In Kapitel 7 wird ein Prototyp auf der Grundlage des Konzepts präsentiert. Dieser dient der Unterstützung von Assoziationsstudien zur Aufdeckung von Genotyp-Phänotyp-Korrelationen in Braugerstensorten.

Kapitel 8 fasst die Ergebnisse der vorliegenden Arbeit zusammen und schließt mit einem Ausblick ab.

2 Grundlagen

In diesem Kapitel werden grundlegende biologische Konzepte sowie Techniken aus der Informatik beschrieben, die für das Verständnis der vorliegenden Arbeit erforderlich sind. Anschließend wird auf fachübergreifende Konzepte eingegangen. Die Beschreibung der Grundlagen soll vergleichsweise abstrakt erfolgen, um sowohl Lesern mit informatischem als auch mit biologischem Hintergrund den Zugang zu dieser Arbeit zu erleichtern. Für weiterführende Informationen sei auf die referenzierte Literatur verwiesen.

2.1 Grundlagen aus der Informatik

2.1.1 Datenbanksysteme

Ein Datenbanksystem (DBS) dient der elektronischen Datenverwaltung. Es setzt sich aus zwei Teilen zusammen, aus einem Datenbankmanagementsystem (DBMS) und einer oder mehrerer Datenbanken [HS97] (Abbildung 2.1). Die Hauptaufgabe eines DBS besteht in der persistenten Speicherung und der Zurverfügungstellung von Daten für Applikationen und Nutzer.

Als Datenbank wird ein Datenbestand bezeichnet, der strukturiert sowie funktional zusammengehörig ist und von einem Datenbankmanagementsystem verwaltet wird.

Ein Datenbankmanagementsystem ist die Verwaltungssoftware einer Datenbank. Sie speichert, modifiziert und organisiert Daten und beantwortet Anfragen. Hierfür stellt das DBMS eine Datenbanksprache zur Verfügung.

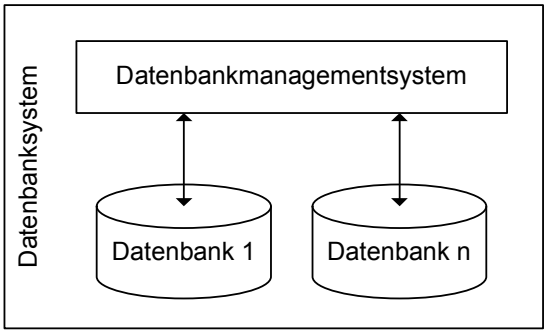


Abbildung 2.1: Schematische Darstellung der Komponenten eines Datenbanksystems nach [HS97]

Die Basisfunktionalität eines Datenbankmanagementsystems wird durch die neun so genannten Codd'schen Regeln [Cod82] beschrieben:

1. **Integration:**

   Alle Daten, die von Applikationen benötigt werden, sollen einheitlich verwaltet werden.

2. **Operationen:**

   Operationen zur Datenspeicherung, Suche und Änderung müssen bereitgestellt werden.

3. **Katalog:**

   Datenbeschreibungen (Metadaten) der Datenbank müssen über einen Katalog (Data Dictionary) zugreifbar sein.

4. **Benutzersichten:**

   Auf den Datenbestand müssen durch das DBMS kontrollierte Sichten erstellbar sein.

5. **Konsistenzüberwachung:**

   Die Korrektheit des Datenbestandes, z. B. bei Änderungen, ist sicherzustellen.

6. **Datenschutz:**

   Es dürfen nur autorisierte Zugriffe auf die Datenbank erlaubt werden.

7. **Transaktionen:**

   Änderungsoperationen sollen zu funktionellen Einheiten zusammengefasst werden können; die Änderungen sollen persistent im Datenbestand gespeichert oder, bei einem Fehler, als Ganzes rückabgewickelt werden.

8. **Synchronisation:**

   Die gegenseitige Beeinflussung paralleler Transaktionen unterschiedlicher Nutzer ist zu vermeiden.

9. **Datensicherung:**

   Im Fall von Systemfehlern muss der Datenbestand wiederherstellbar sein.

Datenunabhängigkeit

Ein Kernkonzept von Datenbankanwendungen bildet die Datenunabhängigkeit. Dies bedeutet, dass die physische Datenbank von Datenbankapplikationen losgelöst ist. Es wird zwischen

- Implementierungsunabhängigkeit und
- Anwendungsunabhängigkeit

unterschieden.

Implementierungsunabhängigkeit heißt, dass die konzeptuelle Sicht auf Daten von der tatsächlichen Speicherung unabhängig ist. Es wird auch von physischer Datenunabhängigkeit gesprochen.

Der Begriff der Anwendungsunabhängigkeit bedeutet, dass die Datenbank von Modifikationen der Anwendungsschnittstellen unabhängig ist. Hier wird auch von logischer Datenunabhängigkeit gesprochen.

In [TK78] wurde zur Realisierung der Datenunabhängigkeit eine 3-Ebenen-Architektur vorgeschlagen, die allgemein akzeptiert ist (Abbildung 2.2).

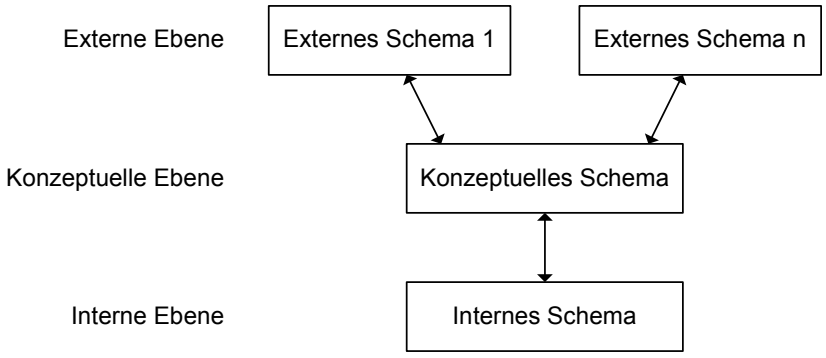


Abbildung 2.2: 3-Ebenen-Schemaarchitektur nach [TK78]

Die unterste Ebene dieser Architektur bildet das interne Schema. Es beschreibt, wie die Datenbank systemspezifisch realisiert wurde (Dateiorganisation, Zugriffspfade etc.). Das interne Schema ist vom genutzten Basissystem abhängig.

Die mittlere Ebene, das konzeptuelle Schema, wird durch die Modellierung der Datenbank in einem systemunabhängigen Datenmodell gebildet. Das konzeptuelle Schema ist von der konkreten Implementierung der internen Ebene unabhängig.

Die oberste Ebene besteht aus einem oder mehreren externen Schemata. Mit diesen werden anwendungs- oder benutzerspezifische Sichten auf den Datenbestand definiert.

Datenbankmodelle

Datenbankmodelle beschreiben Datenbanken. Hierzu steht ihnen ein System von Konzepten zur Verfügung, mit deren Hilfe Syntax und Semantik von Datenbankbeschreibungen (Datenbankschemata) definiert werden [HS97]. Datenbanksysteme können auf verschiedenen Datenbankmodellen basieren.

Die klassischen Vertreter sind

- das hierarchische Datenbankmodell [McG77],
- das Netzwerkdatenbankmodell (CODASYL) [COD71],
- das relationale Datenbankmodell [Cod70] und
- das objektorientierte Datenbankmodell [ABD+89].

Außerdem gibt es Mischformen wie das objektrelationale Datenbankmodell [Cat91]. Das relationale Datenbankmodell hat sich als Standard etabliert und wird auch in der vorliegenden Arbeit verwendet. Daher soll es kurz beschrieben werden.

Der Begriff der Relation, die mathematische Beschreibung einer Tabelle, bildet die Grundlage dieses Konzepts (Abbildung 2.3). Zu modellierende Realweltobjekte (Entitäten) werden dabei durch das so genannte Relationenschema, einer Menge von Attributen, beschrieben. Die Tupel (Zeilen) bilden die Relation über dieses Schema. Eine Menge von Relationenschemata wird als Datenbank bezeichnet.

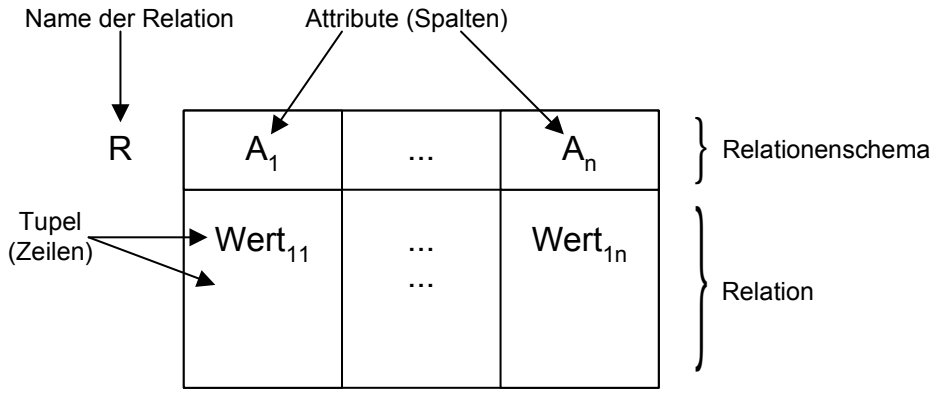


Abbildung 2.3: Veranschaulichung des Relationenmodells nach [Cod70]

Die relationale Algebra definiert die Operationen, die auf einer Menge von Relationen möglich sind [Cod70]. Alle Operationen lassen sich durch sechs Grundoperationen abbilden:

1. **Projektion:**

   Auswählen bestimmter Spalten.

2. **Selektion:**

   Auswählen von Tupeln einer Relation.

3. **Kreuzprodukt (kartesisches Produkt):**

   Kombination aller Tupel der Relation R_1 mit denen der Relation R_2.

4. **Vereinigung:**

   Wenn zwei Relationen R_1 und R_2 das gleiche Relationenschema haben, können sie zu einer einzigen Relation vereint werden. Hierbei werden Duplikate entfernt.

5. **Differenz:**

   Entfernung aller Tupel aus Relation R_1, die auch in Relation R_2 vorhanden sind.

6. **Umbenennung:**

   Umbenennen von Attributen und Relationen für Mengenoperationen zwischen Relationen mit unterschiedlichen Attributen, für Joins über Relationen mit verschiedenen Namen sowie für kartesische Produkte mit identischen Attributbezeichnungen.

Andere Operationen der relationalen Algebra, wie z. B. der Durchschnitt oder der Join, werden durch Kombination dieser sechs Grundoperationen gebildet.

2.1.2 Datenmodellierung

Modellierung ist in der Informatik die abstrakte Abbildung von Objekten der realen Welt zur Beantwortung von Fragestellungen. Die Basis dafür bildet ein in einer formalen Sprache definiertes Modell. Das Ergebnis der Modellierung ist ein Schema.

Eine wichtige Aufgabe im Rahmen dieser Arbeit liegt in der Modellierung von Realweltobjekten und ihren Beziehungen. Daher sollen im Folgenden zwei verbreitete Möglichkeiten der Modellierung vorgestellt werden, die für diese Arbeit relevant sind.

Das Entity-Relationship-Modell (ER-Modell)

Das Entity-Relationship-Modell ist ein Datenmodell, dass von [Che76] zur Beschreibung von Ausschnitten der realen Welt vorgeschlagen wurde. In der konzeptionellen Phase ermöglicht das ER-Modell die Kommunikation zwischen Entwicklern und Nutzern. Das Resultat der Modellierung in Form eines ER-Schemas wird zur Grundlage der späteren Implementierung.

Das ER-Modell enthält drei grundlegende Elemente:

1. **Entität (Entity):**

   Objekte der realen Welt, über die Informationen gespeichert werden sollen, werden als Entitäten bezeichnet, z. B. *Pflanze* oder *Gewächshaus*.

2. **Beziehung (Relationship):**

   Entitäten können miteinander in Beziehung stehen, z. B. Pflanze *wächst in* Gewächshaus.

3. **Attribut:**

   Attribute beschreiben sowohl Eigenschaften von Entitäten als auch Eigenschaften von Beziehungen, z. B. *Wuchshöhe*.

Abbildung 2.4 zeigt ein Beispiel für ein ER-Schema. Entitäten werden als Rechtecke dargestellt, Beziehungen als Rhomben. Attribute (Eigenschaften) sind als Rechtecke mit abgerundeten Ecken mit den jeweiligen Entitäten bzw. Beziehungen verbunden.

Um die Teilnahme von Instanzen von Entitäten an einer Beziehung einzuschränken, wurde das Konzept der Kardinalität entwickelt. Hierfür können ein minimaler und ein maximaler Wert angegeben werden. Im Beispiel aus Abbildung 2.4 würde die Kardinalität *wächst in (Pflanze[1,1], Gewächshaus[0,100])* bedeuten, dass eine Pflanze in

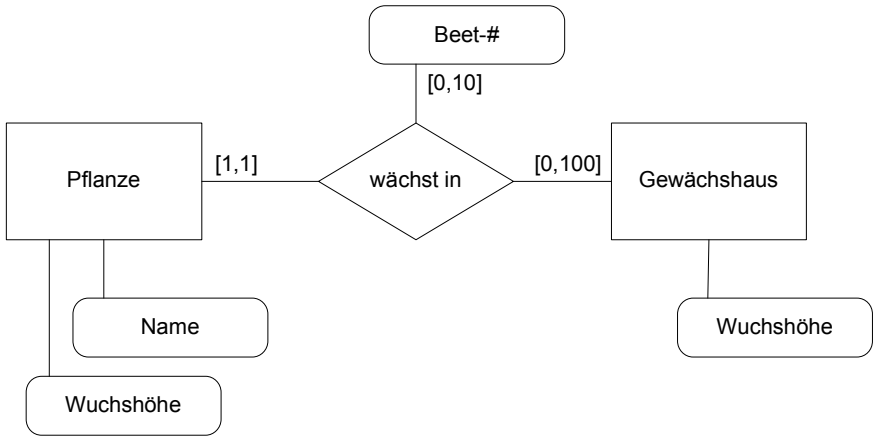


Abbildung 2.4: Ein Beispiel eines Entity-Relationship-Schemas in Chen-Notation [Che76]

genau einem Gewächshaus aufwächst und dass in einem Gewächshaus maximal 100 Pflanzen wachsen.

Bei dem eben vorgestellten ER-Modell wird auch vom klassischen ER-Modell gesprochen. Es wurde durch das erweiterte Entity-Relationship-Modell (EER-Modell) [EGH+92] um eine Reihe von Konstrukten ergänzt. Hierzu zählen insbesondere:

- **Generalisierung:**

  Mit Hilfe einer Generalisierungsbeziehung können Entitäten in einen allgemeineren Kontext gesetzt werden. Beispielsweise sind sowohl eine Gerstenpflanze als auch eine Weizenpflanze eine Pflanze.

- **Partitionierung:**

  Das Gegenstück zur Generalisierung bildet die Partitionierung. Eine Entität kann dadurch in einem spezielleren Kontext betrachtet werden. So kann z. B. ein Gewächshaus sowohl ein kleines als auch ein großes sein.

- **Spezialisierung:**

  Die Spezialisierung ist ein Spezialfall der Partitionierung mit genau einem Eingang und einem Ausgang.

Unified Modelling Language (UML)

Bei der Unified Modelling Language (UML) handelt es sich um eine standardisierte Sprache, die von der Object Management Group (OMG)[1] entwickelt und 1997 vorgestellt wurde. Der Zweck der UML besteht in der Modellierung von Daten, Software und anderen Systemen. Für die zur Modellierung notwendigen Begriffe wurden Bezeichner sowie grafische Notationen definiert und Beziehungen zwischen diesen Begriffen festgelegt. Außerdem können mit der UML Schemata von statischen Strukturen und dynamischen Abläufen erstellt werden. Die aktuelle Version der UML ist die 2.2. [UML07].

Das Ergebnis der Modellierung mit der UML ist ein UML-Diagramm. UML2 unterscheidet sechs verschiedene Typen von Strukturdiagrammen:

- **Klassendiagramm:**

  Klassendiagramme bilden Klassen mit Attributen und Methoden sowie ihre Beziehungen ab.

- **Komponentendiagramm:**

  Mit Komponentendiagrammen werden Abhängigkeiten von Komponenten sowie ihre Beziehungen modelliert.

- **Kompositionsstrukturdiagramm:**

  Kompositionsstrukturdiagramme stellen den Aufbau der Schnittstellen von Klassen oder Komponenten dar.

- **Objektdiagramm:**

  Ein Objektdiagramm hat dieselbe Struktur wie ein Klassendiagramm, jedoch werden die zu einem bestimmten Zeitpunkt existierenden Objekte (=Instanzen von Klassen) und ihre Attributausprägungen dargestellt.

- **Paketdiagramm:**

  Um ein Gesamtmodell in überschaubare Einheiten zu unterteilen, können beliebige Modellelemente (unterschiedlichen Typs) zu Paketen zusammengefasst werden. In einem Paketdiagramm werden die Beziehungen zwischen solchen Paketen oder auch die Komposition eines Paketes aus existierenden Paketen modelliert.

[1] http://www.omg.org [Stand 2009-04-02]

- **Verteilungsdiagramm:**

  Einsatz- oder Verteilungsdiagramme zeigen die Verteilung der Komponenten eines Systems (Hardware, Software) und ihre Kommunikationsbeziehungen.

Weiterhin wird zwischen sieben Verhaltensdiagrammtypen unterschieden:

- **Anwendungsfalldiagramm:**

  Ein Anwendungsfall ist ein Arbeitsablauf in einem System. Ein Anwendungsfall- oder Use-Case-Diagramm modelliert die Interaktionen zwischen Anwendungsfällen und Akteuren.

- **Aktivitätsdiagramm:**

  In Aktivitätsdiagrammen werden anhand verschiedener Knoten Abläufe dargestellt. Diese Knoten sind durch Objekt- und Kontrollflüsse verbunden.

- **Sequenzdiagramm:**

  Mit Sequenzdiagrammen wird der zeitlich begrenzte Austausch von Nachrichten zwischen Akteuren und Objekten modelliert.

- **Kommunikationsdiagramm:**

  Kommunikationsdiagramme dienen der Abbildung der Interaktionen ausgewählter Objekte innerhalb eines bestimmten Kontextes.

- **Interaktionsübersichtsdiagramm:**

  Ein Interaktionsübersichtsdiagramm stellt Teilabläufe durch eingebettete oder aber referenzierte Aktivitäts-, Kommunikations- oder Sequenzdiagramme dar.

- **Zeitdiagramm:**

  Ein Zeitdiagramm stellt den zeitlichen Ablauf von Zustandsänderungen beteiligter Objekte dar.

- **Zustandsdiagramm:**

  Zustandsdiagramme modellieren den Ablauf von Zuständen von Objekten innerhalb ihres Lebenszyklus.

Abbildung 2.5 zeigt eine Auswahl von Notationselementen der UML.

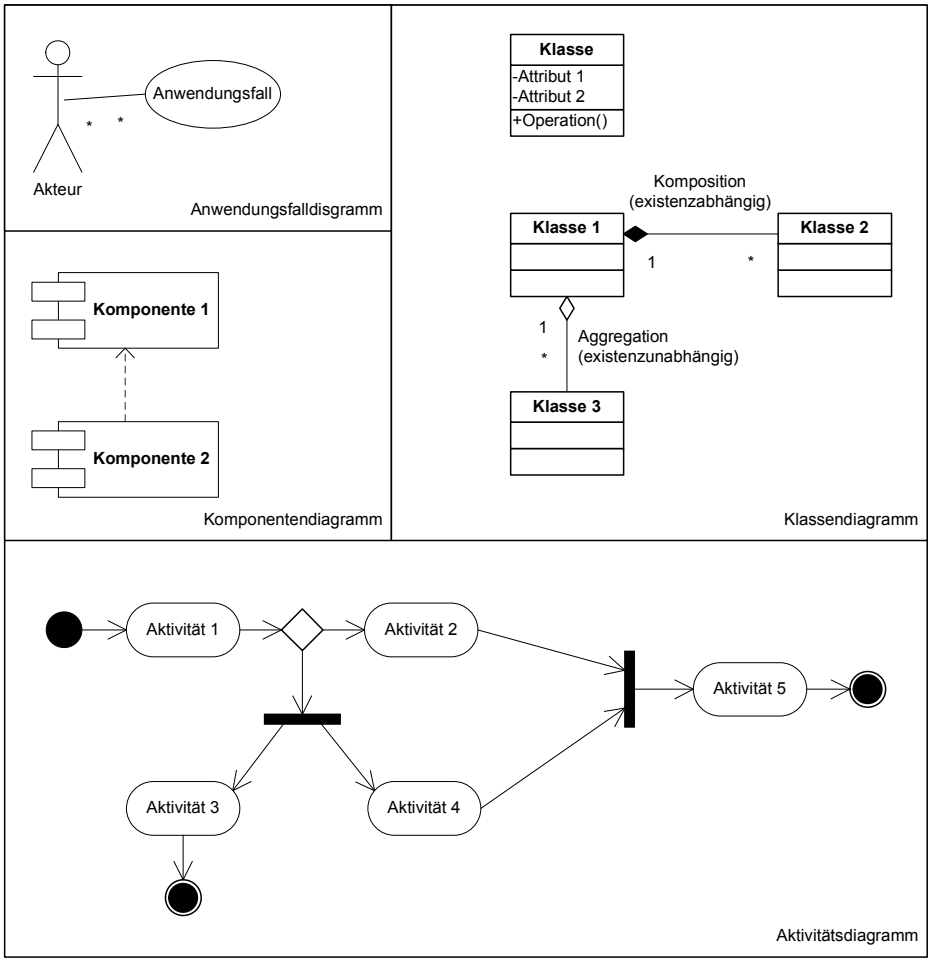


Abbildung 2.5: Auswahl von Notationselementen der UML

2.1.3 Entity-Attribute-Value-Ansatz

Der Entity-Attribute-Value-Ansatz (EAV) ist eine Methode zur generischen Datenbankmodellierung. Hierbei werden Kombinationen aus Realweltobjekt (Entity), Attribut und Ausprägung (Value) als Tupel in einer Tabelle gespeichert. Dieses Vorgehen wird schematisch in Abbildung 2.6 gezeigt.

Ursprünglich wurden Attribut-Wert-Paare (AV) im Bereich der künstlichen Intelligenz eingesetzt [Win84]. Der darauf basierende EAV-Ansatz wurde in den 1990er Jahren im Bereich der Lebenswissenschaften, insbesondere bei Krankenhausinformationssystemen (siehe z. B. [FHJ+90, NB98]), populär.

Eine Erweiterung dieses sehr simplen Ansatzes stellt das so genannte EAV/CR (EAV with classes and relationships) [NMC+99] dar. Hierbei wird versucht, komplexe Objekte (Klassen) und ihre Beziehungen untereinander abzubilden. Die Beziehungen

Beobachtungen		
Entity	Attribute	Value
HOR 1234	Taxon	Hordeum vulgare L.
HOR 1234	Herkunft	Großbritannien
HOR 1234	Jahr	1975
HOR 1234	Wuchshöhe	100
HOR 1234	Ertrag	500
HOR 1234	Blühzeitpunkt	10. Mai
HOR 1234	Resistenz	7
...	...	...

Abbildung 2.6: Speicherung phänotypischer Beobachtungswerte nach dem EAV-Ansatz

selbst werden hierbei auch als EAV-Tripel abgespeichert. Abbildung 2.7 illustriert den EAV/CR-Ansatz am Beispiel von vier Tabellen.

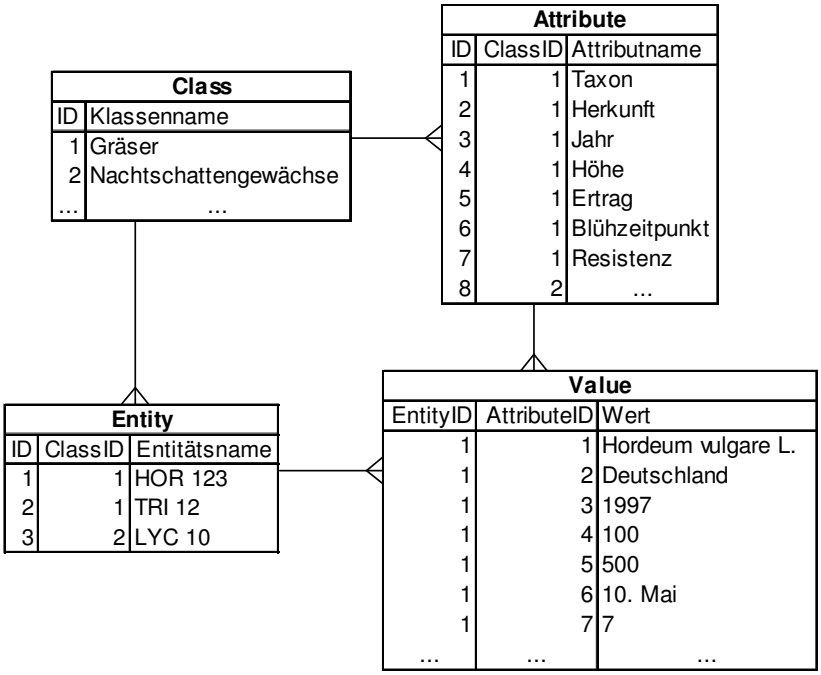


Abbildung 2.7: Speicherung phänotypischer Beobachtungswerte mit dem EAV/CR-Ansatz

Die Vor- und Nachteile einer Verwendung dieses Ansatzes im Rahmen der vorliegenden Arbeit werden in Kapitel 6 ausführlich diskutiert.

2.1.4 Record Linkage

Um Daten aus verschiedenen Quellen integrieren zu können (vgl. Kapitel 3) ist es oftmals erforderlich, Verbindungen über Attributausprägungen herzustellen, die nur ähn-

lich, aber nicht identisch sind. Die Ursachen dafür sind vielfältig, z. B. Schreibfehler, Synonyme etc.

Hier bietet es sich an, die Ähnlichkeit sim zwischen Attributausprägungen zu bestimmen und Records darüber miteinander zu verknüpfen. Ähnlichkeit bedeutet in diesem Kontext, dass die beiden miteinander verglichenen Attributausprägungen eine Reihe gemeinsamer Eigenschaften haben. Das Verhältnis zwischen den gemeinsamen und den unterschiedlichen Eigenschaften bildet dabei den Grad der Ähnlichkeit.

Eine solche Vorgehensweise wird als Record Linkage bezeichnet [NKAJ59, FS69] (vgl. Abschnitt 6.5). Im Fall numerischer Datentypen kann die Ähnlichkeit durch die Verwendung der Abweichung bestimmt werden [MWBL05]. Für alphanumerische Datentypen bieten sich Äquivalenzmethoden und Similarity-Ranking-Methoden an.

Äquivalenzmethoden vergleichen zwei Zeichenketten miteinander und geben im Fall einer Ähnlichkeit TRUE oder im negativen Fall FALSE zurück. Eine Auswahl der verbreitetsten Verfahren soll im Folgenden vorgestellt werden.

- **Lautähnlichkeit:**

  Hierbei erfolgt ein Vergleich zweier Zeichenketten dahingehend, wie sie ausgesprochen klingen. Der bekannteste Vertreter dieses Ansatzes ist der Soundex-Algorithmus [Rus18], der für eine Zeichenkette einen Code aus einem Buchstaben, gefolgt von drei Zahlen generiert (die Gerstensorte *Ingrid* wird hierbei zu I526). Das Funktionsprinzip entspricht dem eines Hash-Verfahrens. Die kodierten Ergebnisse werden verglichen. Nur etwas über ein Drittel der gefundenen Übereinstimmungen sind korrekt [LR96]. Es existieren mehrere Varianten für verschiedene Sprachen. Beispielsweise wurde für die deutsche Sprache das so genannte *Kölner Verfahren* [Pos69] entwickelt.

- **Wortstamm:**

  Bei diesem Verfahren werden zwei Zeichenketten auf der Basis ihrer Wortstämme miteinander verglichen. Hierfür wird ein Suffix-Verzeichnis für mögliche Wortendungen benötigt. Diese Methode ist, wie auch das Lautähnlichkeitsverfahren, sprachabhängig.

- **Groß-/Kleinschreibung:**

  Mit dieser Methode wird überprüft, ob zwei Zeichenketten identisch sind, wenn die Groß-/Kleinschreibung außer Acht gelassen wird, z. B. *Gerste* und *GERSTE*.

- **Synonyme:**

  Anhand kontrollierten Vokabulars kann festgestellt werden, ob zwei Zeichenketten die gleiche Bedeutung haben, z. B. *Gerste* und *Hordeum*.

- **Wildcards:**

  Wildcards sind Platzhalter für andere Zeichen. Hiermit können Zeichenketten dahingehend überprüft werden, ob sie zumindest in Teilen übereinstimmen. Es wird vielfach zwischen Platzhaltern für genau ein Zeichen (z. B. _ oder ?) und Platzhaltern für beliebig viele Zeichen (z. B. % oder *) unterschieden. Beispielsweise würde beim Vergleichen von Sortennamen der Ausdruck *Ingrid%* die Zeichenketten *Ingrid WT* und *Ingrid BC mlo5* für äquivalent befinden, nicht aber die Zeichenkette *Ingrid*.

- **Reguläre Ausdrücke:**

  Reguläre Ausdrücke beschreiben auf der Basis syntaktischer Regeln Zeichenketten. Der reguläre Ausdruck bildet also ein Muster, das mit einer Menge von Zeichenketten verglichen werden kann, um eine Teilmenge herauszufiltern.

Im Gegensatz zu den eben vorgestellten Äquivalenzmethoden vergleichen Similarity-Ranking-Methoden zwar ebenfalls zwei Zeichenketten, geben jedoch zurück, wie groß (Ranking) deren Ähnlichkeit ist.

- **Hamming-Ähnlichkeit:**

  Sie basiert auf der Hamming-Distanz [Ham50]. Die Hamming-Distanz ist ein Maß für die Unterschiedlichkeit digitaler Daten. Zwei Binärdatenblöcke A und B fester Länge werden bitweise verglichen und die Anzahl der verschiedenen Stellen wird gezählt. Haben die verglichenen Datenblöcke unterschiedliche Längen, so ist die Hamming-Distanz unendlich.
  Aus der Hamming-Distanz $hamm(S_A, S_B)$ kann die Hamming-Ähnlichkeit durch $sim_{hamm}(S_A, S_B) = 1 - \frac{hamm(S_A,S_B)}{n}$ berechnet werden. n bezeichnet hierbei die Länge der verglichenen Datenblöcke. Die Hamming-Ähnlichkeit bewegt sich zwischen 1 (identisch) und 0 (keine Ähnlichkeit).

- **Editbasierte Ähnlichkeit:**

  Je weniger Elemente von einer Zeichenkette A substituiert werden müssen, um zu einer Zeichenkette B zu gelangen, desto ähnlicher sind sich A und B. Aus den Kosten von Einfüge-, Lösch- und Ersetzungsoperationen ergibt sich der Editierabstand. Eines der wichtigsten Distanzmaße aus der Gruppe der Editerabstände ist die Levenshtein-Distanz [Lev66]. Ein weiteres wichtiges Maß ist die Damerau-Distanz [Dam64]. Bei dieser wird zusätzlich zu den drei genannten Operationen noch das Vertauschen von Zeichenkettenelementen ermöglicht, um Tipp-/Buchstabierfehlern zu begegnen. Die Levenshtein-Distanz der beiden Zeichenketten *Hordeum* und *Horedum* hat den Wert 2, weil zwei Ersetzungen ausgeführt werden müssen (*e* gegen *d* und *d* gegen *e*). Die Damerau-Distanz hingegen ist 1, weil das Vertauschen von *ed* zu *de* nur eine Operation darstellt.

Die Berechnung der Edit-Ähnlichkeit aus dem Edit-Abstand erfolgt mit der Formel $sim_{ed} = 1 - \frac{ed(S_A,S_B)}{max\{|S_A|,|S_B|\}}$.

- **Longest-Common-Substring-basierte Ähnlichkeit:**

  Das Longest-Common-Substring-Verfahren [Wei73] basiert auf Vergleichen von Teilstrings. Je länger ein Substring ist, den zwei Zeichenketten teilen, desto größer ist ihre Ähnlichkeit. Wenn in einer der zu vergleichenden Zeichenketten ein Fehler (z. B. ein Tippfehler in der Mitte des gemeinsamen Strings) vorliegt, würden die beiden Zeichenketten dann fälschlich als weniger ähnlich bewertet werden.
  Die Berechnung der Ähnlichkeit kann durch $sim_{lcs}(S_A, S_B) = \frac{lcs(S_A,S_B)}{max\{|S_A|,|S_B|\}}$ erfolgen. $lcs(S_A, S_B)$ sei hierbei die Länge des längsten gemeinsamen Teilstrings.

Die drei eben vorgestellten Verfahren reagieren sehr empfindlich auf Vertauschungen von Zeichen und Teilstrings sowie auf Zeichenketten unterschiedlicher Länge. Wird beispielsweise ein Personenname im ersten zu vergleichenden String nach dem Muster „Vorname Nachname“ und im zweiten nach dem Muster „Nachname, Vorname“geschrieben, so ist die Ähnlichkeit bei Verwendung dieser Verfahren relativ gering.

Eine Alternative stellen so genannte tokenbasierte Verfahren dar. Hierbei erfolgt eine Zerlegung der Strings in Token, d. h. Vorkommen von Zeichen. Der Vergleich der Strings erfolgt nun über gemeinsame Token, die Reihenfolge spielt keine Rolle.

Zwei dieser Verfahren, die ebenfalls zu den Similarity-Ranking-Methoden gezählt werden können, sollen im Folgenden vorgestellt werden.

- **Dice-Ähnlichkeitskoeffizient:**

  Der Dice-Ähnlichkeitskoeffizient [Dic45] wird als $D = \frac{2 \cdot |A \cap B|}{|A|+|B|}$ errechnet. Hierbei ist $|A \cap B|$ die Anzahl der Übereinstimmungen und $|A| + |B|$ die Anzahl der Elemente, die verglichen werden; der Koeffizient nimmt einen Wert zwischen 0 und 1 an. Da der Vergleich unabhängig von ihrer Reihenfolge über Token durchgeführt wird, ist der Dice-Ähnlichkeitskoeffizient der beiden Zeichenketten *Hordeum* und *Horedum* $D = \frac{2 \cdot 7}{(7+7)} = 1$.

- **Jaccard-Ähnlichkeitskoeffizient:**

  Der Jaccard-Ähnlichkeitskoeffizient [Jac01] oder Tanimoto-Koeffizient [Tan57] zweier Zeichenketten A und B wird durch $J = \frac{|A \cap B|}{|A \cup B|}$ errechnet, wobei $|A \cap B|$ wieder die Schnittmenge und $|A \cup B|$ die Vereinigung der Elemente von A und B ist. Der Jaccard-Ähnlichkeitskoeffizient der beiden Zeichenketten *Hordeum* und *Horedum* ist $J = \frac{7}{(7+7-7)} = 1$.

In Abhängigkeit des Anwendungsfalles kann es sinnvoll sein, Äquivalenz- und Similarity-Ranking-Methoden zu kombinieren. Beispielsweise kann mit Hilfe von Äquivalenzmethoden eine Vorauswahl getroffen werden (z. B. gleicher Wortstamm). Darauf aufbauend können die Resultate mit Similarity-Ranking-Methoden weiter verfeinert werden. Falls erforderlich, müssen die Ergebnisse manuell verifiziert werden.

2.2 Grundlagen aus der Biologie

2.2.1 Bausteine des Lebens

Organismen bestehen aus Organen, die aus Gewebekomplexen gebildet werden. Gewebe wiederum sind funktionelle Einheiten mehrerer Zellen. Jede Zelle stellt ein eigenständiges und abgegrenztes System dar, das bestimmte Funktionen erfüllen kann. Es haben sich durch die Evolution zwei Arten von Zellen entwickelt, die Prokaryoten und die Eukaryoten. Prokaryoten sind einfach aufgebaut und verfügen nicht über einen Zellkern. Eukaryoten, z. B. Tiere und Pflanzen, sind komplexer aufgebaut und haben einen Zellkern. Sie können sowohl als Einzeller, z. B. Backhefe (*Saccharomyces cerevisiae*), oder auch als Mehrzeller, z. B. Gerste (*Hordeum vulgare*), vorkommen.

Transkription und Translation

Zellen enthalten den Träger der Erbinformation, die Desoxyribonukleinsäure (DNS). In Prokaryoten liegt die DNS in einfacher, geschlossener Form im Cytoplasma vor, während sie in Eukaryoten in linearer Form in den so genannten Chromosomen des Zellkerns gespeichert ist. Die Struktur der DNS, die Doppelhelix, wurde 1953 beschrieben [WC53b, WC53a]. Die Einzelstränge der DNS bestehen aus Desoxyribosemolekülen und Phosphorsäure, die miteinander verbunden sind. Auf jedem Einzelstrang sind die vier organischen Basen Adenin (A), Cytosin (C), Guanin (G) sowie Thymin (T) in einer bestimmten Abfolge angeordnet. Die Basen der beiden Stränge verhalten sich paarweise zueinander komplementär, Adenin auf dem einen Strang ist immer Thymin auf dem anderen Strang zugeordnet, Cytosin immer Guanin. Die beiden sich gegenüber liegenden Basen werden als Basenpaar bezeichnet und sind durch zwei (A=T) bzw. drei (G≡C) Wasserstoffbrücken miteinander verbunden. Die Doppelhelix ist das Resultat einer Drehung dieses leiterförmigen Makromoleküls.

Abfolgen von Basenpaaren, die genetische Informationen kodieren, werden als Gene bezeichnet. Diese enthalten beispielsweise Baupläne für Enzyme (eine Klasse der Proteine), die als Katalysatoren am Stoffwechsel des Organismus beteiligt sind.

DNS verfügt über die Fähigkeit, sich durch die so genannte DNS-Synthese zu replizieren. Dabei wird zuerst die Doppelhelix durch das Enzym Helicase in zwei Einzelsträn-

ge zerlegt, welche jeweils als Vorlage für einen zu synthetisierenden, komplementären Gegenstrang dienen. Dies erfolgt durch DNS-Polymerasen.

Die Synthese von Proteinen auf Basis eines Gens erfolgt über den Mechanismus der Genexpression. Dabei wird zuerst ein Abschnitt der DNS abgeschrieben (transkribiert) und in eine komplementäre Ribonukleinsäuresequenz überführt. Ribonukleinsäure (RNS) ist ein Molekül, bei dem die Base Thymin durch die Base Uracil ersetzt wird. Jeweils drei nebeneinander liegende Basen bilden ein so genanntes Triplett. Dieses definiert eine spezifische Aminosäure und damit ein spezifisches Protein. Hierbei wird vom genetischen Code gesprochen [JM61]. Die Übersetzung dieses Codes in eine Aminosäuresequenz erfolgt durch die Translation und bildet den zweiten Schritt der Genexpression. Das Ergebnis der Translation ist eine Aminosäuresequenz, welche die Basis für ein Protein ist. Abbildung 2.8 veranschaulicht diesen Prozess.

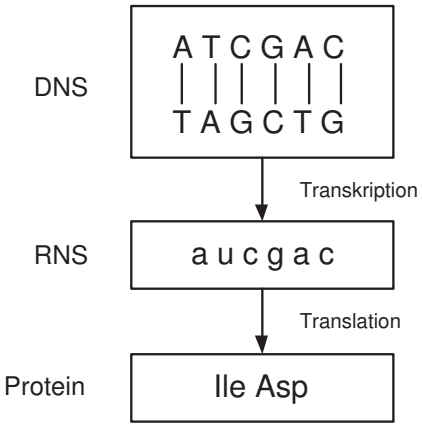


Abbildung 2.8: Schematische Darstellung der Genexpression

Enzyme und biochemische Reaktionen

Proteine, die über die Fähigkeit verfügen, biochemische Prozesse zu katalysieren, heißen Enzyme. Enzyme besitzen ein aktives (katalytisches) Zentrum zur Interaktion mit Metaboliten (Substraten oder Produkten) während biochemischer Reaktionen. Bestimmte Enzyme verfügen darüber hinaus über ein regulatorisches (allosterisches) Zentrum, um ihrerseits durch Metaboliten beeinflusst zu werden (allosterische Regulation).

Eine biochemische Reaktion ist die Umwandlung einer Substratmenge in eine Produktmenge. Die meisten Reaktionen werden durch Enzyme katalysiert. Enzym und Substrat bilden im ersten Schritt der Reaktion einen Enzym-Substrat-Komplex (durch Substratbindung am katalytischen Zentrum des Enzyms), die Aktivierungsenergie wird herabgesetzt, die Reaktionsgeschwindigkeit erhöht. Im zweiten Schritt erfolgt eine

Aufspaltung in Enzym und Produkt. Das Enzym wird bei der Reaktion nicht verbraucht.

Biochemische Reaktionen können durch verschiedene Mechanismen der Zelle reguliert werden. Die Regulation findet dabei auf verschiedenen Ebenen mit diversen Möglichkeiten der Unterscheidung statt. Der am besten erforschte Regulationstyp ist die enzymatische Regulation [Bis00]. Hierbei wird zwischen Inhibitoren und Aktivatoren unterschieden. Inhibitoren sind Moleküle, die auf Enzyme und/oder Substrate einwirken und dadurch eine Reaktion hemmen. Die katalytische Aktivität von Enzymen wird durch Aktivatoren erhöht.

2.2.2 Besonderheiten von Pflanzen

Im Gegensatz zum Tierreich zeichnen sich Pflanzen durch eine Reihe von Besonderheiten aus, die im Folgenden kurz erwähnt werden sollen:

- Die Zellwände pflanzlicher Zellen bestehen hauptsächlich aus Zellulose.
- Zellen enthalten Chloroplasten zur Gewinnung von Energie aus Licht auf dem Wege der Photosynthese. Hierbei wird Licht in Energie in Form von Glucose umgewandelt oder als Stärke gespeichert.
- Zellen enthalten Vakuolen. Das sind kleine Räume innerhalb der Zellen, die Farb- und Duftstoffe etc. enthalten können.
- Genome von Pflanzen sind meist um ein Vielfaches umfangreicher als das des Menschen oder Genome von Tieren (Tabelle 2.1).
- Umweltfaktoren wie Boden, Wetter, biotische (z. B. Schädlingsbefall) und abiotische (z. B. Dürre) Stressfaktoren haben große Einflüsse auf phänotypische Ausprägungen von Pflanzen. Pflanzen sind standortgebunden.
- In Pflanzen ist der so genannte sekundäre Metabolismus stark ausgeprägt. Durch diesen werden Substanzen produziert, die keine besondere Relevanz für die produzierende Zelle haben, jedoch für das Überleben des gesamten Organismus von großem Interesse sind. Hierzu zählen beispielsweise Wachse, die von bestimmten Pflanzen zum Schutz vor UV-Licht produziert werden. Der sekundäre Metabolismus ist für die Forschung besonders interessant (z. B. erhöhter Flavonoid-Gehalt für pharmazeutische Anwendungen).

Tabelle 2.1: Übersicht der Genomgrößen verschiedener Organismen

Organismus	geschätzte Genomgröße
Fruchtfliege (*Drosophila melanogaster*)	ca. $0,12 \cdot 10^9$ Bp [ACH+00]
Mensch (*Homo sapiens*)	ca. $3 \cdot 10^9$ Bp [Int04]
Reis (*Oryza sativa*)	ca. $4,5 \cdot 10^9$ Bp [YHW+02, GRL+02]
Gerste (*Hordeum vulgare*)	ca. $5,5 \cdot 10^9$ Bp [BL95]
Weizen (*Triticum aestivum*)	ca. $17 \cdot 10^9$ Bp [BL95]

2.2.3 Datendomänen

Für die vorliegende Arbeit standen eine Reihe von Inhouse- und öffentlich zugänglichen Datenquellen unterschiedlicher Domänen zur Verfügung, die im Folgenden beschrieben werden. Die Verwendung von Inhouse-Daten wurde bevorzugt, weil bei diesen Daten in den meisten Fällen mehr Informationen über die Qualität der erhobenen Daten und Zusatzinformationen wie z. B. Entwicklungsstadium etc. verfügbar waren. Außerdem kann bei Inhouse-Daten ein größerer Einfluss auf die strukturierte Speicherung genommen werden, der nicht unterschätzt werden sollte (vgl. Kapitel 4).

Sequenzdaten

In der Biologie werden unter dem Begriff der Sequenz Abfolgen von Nukleotiden (DNS-Sequenz) oder Aminosäuren (Aminosäuresequenz) als Ergebnis einer Sequenzierung, d. h. der Ermittlung ihrer Teilbausteine, verstanden.

DNS-Sequenz Eine genomische DNS-Sequenz setzt sich aus einer Abfolge der Nukleotide Adenin (A), Cytosin (C), Guanin (G) und Thymin (T) in unterschiedlicher Häufigkeit auf einem DNS-Strang als Träger von Informationen zusammen. Durch DNS-Sequenzierung wird der Aufbau solcher Sequenzen entschlüsselt. Die DNS-Sequenzierung geht im Wesentlichen auf [MG77] und [SNC77] zurück.

Die Nutzung so genannter Expressed Sequence Tags (ESTs) ist eine weitere Möglichkeit, Sequenzinformationen zu erhalten. Hierzu wird aus einem Organismus extrahierte Boten-RNS (mRNS) zur Stabilisierung durch das Enzym Reverse Transkriptase in komplementäre DNS (cDNS) überführt. Durch Ansequenzieren werden kurze Abschnitte (bis zu 700 Basenpaare) der cDNS-Stränge gewonnen, die Expressed Sequence Tags. Diese einzelnen Teilabschnitte können durch ein Alignment in eine Konsensussequenz überführt werden. Die Sequenzierung mittels der EST-Methode führt nicht zur vollständigen genomischen

DNS, da nur Genabschnitte sequenziert werden, die in einem bestimmten Gewebe, das sich in einem bestimmten Entwicklungsstadium befindet, exprimiert sind. Es fehlen also die Intronbereiche eines Gens. Insofern können nur Aussagen über die Exonstruktur und den transkribierten Bereich eines Gens getroffen werden. Diese Technik wird jedoch, insbesondere im Bereich der Pflanzenbiologie (vgl. [KLF+05]), häufig verwendet, da sie schneller und kostengünstiger als die vollständige Sequenzierung eines Organismus (wie z. B. im Humangenomprojekt[2]) ist. Die EST-Sequenzierung geht auf [AKG+91] zurück.

Aminosäuresequenz Aminosäuresequenzen geben über die Zusammensetzung von Proteinen Auskunft. Sie werden daher auch als Proteinsequenzen bezeichnet. Sammlungen von Proteinsequenzen bilden in der biologischen Forschung eine wichtige Informationsquelle [ABW04].

Markerdaten

In der Genetik wird unter einem Marker ein DNS-Abschnitt verstanden, der in unterschiedlicher Ausprägung in verschiedenen Individuen vorliegt und mit phänotypischen Unterschieden korreliert ist [PV94]. Markerdaten können heutzutage mit modernen Hochdurchsatz-Verfahren schnell und in großen Mengen gewonnen werden. Zur Verwaltung dieser Daten werden zunehmend Datenbanken eingesetzt, z. B. [Sof03,Sei03].

RFLP-Marker Diese Abkürzung steht für Restriction Fragment Length Polymorphism [BWSD80, TYPB89]. Damit werden Unterschiede von DNS-Sequenzen an gleichen Positionen in Chromosomen verschiedener Individuen bezeichnet, die in Form unterschiedlicher Restriktionsfragmentmuster sichtbar gemacht werden können. Die Länge dieser Fragmente wird durch Mutation beeinflusst. Je näher sich RFLPs auf der DNS befinden, desto wahrscheinlicher werden sie gemeinsam vererbt. Daher können RFLPs bei der Genkartierung als genetische Marker verwendet werden.

AFLP-Marker AFLP heißt Amplified Fragment Length Polymorphism [VHB+95]. Bei der AFLP-Markeranalyse wird die DNS mit Hilfe zweier spezifischer Restriktionsenzyme in Fragmente geschnitten. Durch Polymerasekettenreaktionen (PCRs) werden diese Fragmente vervielfältigt. Dabei entstehen, wie auch bei RFLP-Markern, verschieden lange Fragmente, die durch Unterschiede in der Anzahl der Restriktionsschnittstellen bedingt sind. Diese werden durch Gelelektrophorese sichtbar gemacht. Hierbei kann zwischen Individuen unterschieden werden und es können Verwandtschaftsbeziehungen dargestellt werden.

[2] http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml [Stand 2009-04-02]

SNP-Marker Single Nucleotide Polymorphisms (SNPs) [WH02] stellen Variationen von einzelnen Basenpaaren auf der DNS dar, so genannte Punktmutationen, welche zu einem bestimmten Grad im Genpool einer Population vorkommen. Wie häufig genau SNPs auftreten, hängt vom jeweiligen Organismus ab, z. B. tritt beim Mais durchschnittlich alle 60 Basenpaare ein SNP auf [CCJ+02]. SNPs werden u. a. zur Marker-Merkmals-Assoziation in einem sequenzierten Genfragment eingesetzt [MWR09].

In diesem Zusammenhang wird unter dem Begriff eines Haplotypen die Kombination mehrerer SNPs verstanden.

SSR-Marker Simple Sequence Repeats oder Mikrosatelliten sind DNS-Abschnitte, die sich im Genom eines Organismus häufig wiederholen. Sie bestehen aus 1-6 Basenpaaren [TMVG03], die 10-100 mal wiederholt werden. SSRs werden häufig zur Bestimmung von Verwandtschaftsbeziehungen eingesetzt [GLCSF95].

INDEL-Marker Die beiden allelspezifischen genetischen Mutationen Einfügen (Insertion) und Löschen (Deletion) werden gemeinsam als INDEL bezeichnet. INDELs können die Basis für einen DNS-Sequenz-Polymorphismus an einer bestimmten Position im Genom sein, der in einer Anzahl von Genotypen beobachtbar ist. Solche INDELs werden vererbt und können als Marker Anwendung finden [VBJE08].

Expressionsdaten

Unter Genexpression wird die Umsetzung der DNS-Information in Strukturen und Funktionen von Zellen verstanden [EBS+06, RRW+06], z. B. in Enzyme. In Abhängigkeit des Ortes in einem Organismus (z. B. eines bestimmten Gewebes oder Kompartimentes), des Entwicklungsstadiums sowie von Umwelteffekten werden unterschiedlich hohe (oder auch keine) Mengen von Genprodukten hergestellt (exprimiert). Mit Hilfe von Array-Technologien [SSDB95] kann zeitgleich eine Vielzahl von Konzentrationen dieser Produkte ausgewertet werden (Expression Profiling).

Daten über metabolische Netzwerke

Metabolische Netzwerke (Pathways) sind Abfolgen biochemischer Reaktionen, die sich in Abhängigkeit des beobachteten Organismus, seines Entwicklungsstadiums, des Locus innerhalb dieses Organismus sowie externer Faktoren unterscheiden können [WGK+06]. Im pflanzlichen Bereich ist insbesondere der so genannte sekundäre Metabolismus von großem Interesse (vgl. Abschnitt 2.2.2).

Phänotypische Daten

Der Begriff des Phänotyps bezeichnet direkt und indirekt beobachtbare Eigenschaften eines Organismus. Er setzt sich aus einer Vielzahl von Merkmalen zusammen. Nachfolgend sollen zwei Unterkategorien phänotypischer Daten vorgestellt werden.

Charakterisierungsdaten Charakterisierungsdaten sind Merkmale von Pflanzen, die relativ einfach beobachtet werden können. Die gleiche umweltunabhängige und ausschließlich genetisch determinierte Ausprägung ermöglicht eine schnelle phänotypische Unterscheidung. Dazu zählen Merkmale wie die Zeiligkeit bei Getreide oder die Unterscheidung zwischen Sommer- und Wintertyp. Dies wird für die taxonomische Bestimmung von Pflanzen genutzt. Zusätzlich zu einer festen Menge von Charakterisierungsmerkmalen, können für unterschiedliche Fruchtarten noch jeweils eine Reihe weiterer Merkmale betrachtet werden [Knü01].

Evaluierungsdaten Im Gegensatz zu den Charakterisierungsdaten hängen die so genannten Evaluierungsdaten sehr stark von Umwelteinflüssen ab. Dazu zählen Merkmale wie Wuchshöhe, Ertrag oder die Anfälligkeit gegenüber bestimmten abiotischen oder biotischen Stressfaktoren, aber auch Inhaltsstoffe von Pflanzen [Knü01, WKV+06].

Passportdaten

Passportdaten [vHK95] dienen der Identifikation von Genotypen. Sie enthalten Merkmale wie die Akzessionsnummer des Genotypen in einer Genbank, den Fundort oder die Institution, von der das Material bezogen wurde (Donor), den wissenschaftlichen Namen etc. Sie unterliegen in der Regel kaum Änderungen.

Passportdaten bieten die Möglichkeit der Verknüpfung unterschiedlichster Datenquellen. Über den wissenschaftlichen Namen können phänotypische Daten mit molekularen Daten verlinkt werden, beispielsweise mit Sequenz- oder Markerdaten. Durch Informationen wie dem Fundort können zusätzlich noch geografische Daten hinzugezogen werden. Die Passportdaten sind ein wichtiges Bindeglied zwischen verschiedenen Datendomänen (siehe Kapitel 6).

2.2.4 Besondere pflanzliche Datenressourcen

Nachfolgend werden ausgewählte Informationssysteme vorgestellt, die am Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK) gepflegt werden und wertvolle Ressourcen im Rahmen dieser Arbeit darstellen.

Das Genbankinformationssystem GBIS

Das IPK in Gatersleben besitzt eine der weltweit bedeutendsten Lebendsammlungen von Kulturpflanzen und verwandten Wildarten mit ca. 150.000 Pflanzenmustern. Die Erhaltung erfolgt in Form von Saatgut und teilweise auch als vegetatives Material. Für den Betrieb einer Genbank dieser Größenordnung ist es erforderlich, eine Vielzahl von Informationen zu verwalten und ständig verfügbar zu halten. Hierzu wurde am IPK das Genbankinformationssystem GBIS[3] [OK06] entwickelt.

GBIS besteht aus drei Modulen:

- **GBIS/I:**

  GBIS/I ist ein webbasiertes Informationssystem, das hauptsächlich für externe Nutzer von Genbankmaterial konzipiert ist. Hiermit kann auf Passportdaten sowie Charakterisierungs- und Evaluierungsdaten zugegriffen werden. Neben der Recherche verfügt das GBIS/I-Modul über eine Warenkorbfunktionalität und ein Bestellsystem, das es ermöglicht, für ausgewähltes Material Saatgutproben zu ordern.

- **GBIS/M:**

  Das Modul GBIS/M ist die institutsinterne Managementkomponente des Genbankinformationssystems. Es dient der Verwaltung des im Kühllager, in Cryo-Konservierung oder im vegetativen Vermehrungsanbau befindlichen Genbankmaterials. Hierzu zählen insbesondere die Überwachung der Keimfähigkeit der gelagerten Samenproben, die Initiierung des Vermehrungsanbaus und die Verwaltung von Charakterisierungs- und Evaluierungsdaten [WKV+03] aus diesen Anbauten.

- **GBIS/B:**

  Ebenfalls institutsintern ist das Modul GBIS/B. Es dient der Planung des Vermehrungsanbaus von Genbankmaterial sowie der Erhebung von Charakterisierungs- und Evaluierungsdaten mit Hilfe mobiler Erfassungsgeräte.

Abbildung 2.9 zeigt die Nutzerschnittstelle des GBIS/I-Moduls.

[3] http://gbis.ipk-gatersleben.de [Stand 2009-04-02]

Abbildung 2.9: Ein Screenshot der Weboberfläche des GBIS/I-Moduls

Die Europäische Poa-Datenbank (EPDB)

Die Europäische Poa-Datenbank[4] [WHG+07] ist ein zentrales Nutzpflanzeninformationssystem im Rahmen des European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR)[5]. Die EPDB verwaltet Passportdaten (ca. 5.000 Akzessionen) über die Mehrheit der in europäischen Genbanken gehaltenen Rispengras-Akzessionen. Rispengras (*Poa*) ist eine der wertvollsten Futtergrasgattungen.

Die Zielstellungen der Europäischen Poa-Datenbank sind nach [WHG+07]:

- die Katalogisierung der in europäischen Genbanken gehaltenen Poa-Akzessionen,
- das Zurverfügungstellen von Informationen,
- das Identifizieren von Lücken (z. B. durch unterrepräsentierte Arten) innerhalb der europäischen Poa-Sammlungen sowie
- das Identifizieren von Duplikaten.

Der Zweck der Europäischen Poa-Datenbank besteht somit neben der Recherche hauptsächlich in der Verbesserung der europäischen Koordination der Erhaltung von Poa-Akzessionen. Hierzu wird neben den eigentlichen Passportdaten eine Information über

[4] http://poa.ipk-gatersleben.de [Stand 2009-04-02]
[5] http://www.ecpgr.cgiar.org [Stand 2009-04-02]

die Originalität der jeweiligen Akzessionen gepflegt. Über diese Information kann das „Most Original Sample“ (MOS) einer Akzession bestimmt werden. Tabelle 2.2 verdeutlicht die Definition des MOS.

Tabelle 2.2: Most-original-sample-(MOS)-Definition nach [Ano00]

Originalität (EURISCO-Code)	Erklärung
most original sample (1)	Das Institut, welches das Material erhält, ist entweder Sammler oder Züchter der Akzession.
with MOS (2)	Das Institut, welches das Material erhält, ist entweder Sammler oder Züchter der Akzession. Diese Akzession wurde jedoch an mindestens eine andere Institution weitergegeben und wurden von dieser (unter einer anderen Akzessionsnummer) zurückerhalten. Somit ist die betroffene Akzession zwar nicht MOS, befindet sich aber in derselben Sammlung wie das MOS.
one away (3)	Es gibt eine Weitergabe zwischen einer Akzession und dem Original (MOS).
more away (4)	Es wurden zwei oder mehr Weitergaben durchgeführt.
unknown (5)	Es existiert keine Information über den Originalitätsstatus der betroffenen Akzession.

Das Backend der Europäischen Poa-Datenbank bildet das relationale Datenbankmanagementsystem Oracle, das Frontend ist mit JSP/Java implementiert. Ein Ausschnitt der Recherche-Oberfläche wird in Abbildung 2.10 gezeigt.

Die Europäische Poa-Datenbank bildet eine wichtige Schnittstelle zwischen phänotypischen Daten und Markerdaten.

Die Europäische Gerstendatenbank (EBDB)

Ein weiteres zentrales Nutzpflanzeninformationssystem im Rahmen des European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR) ist die Europäische Gerstendatenbank[6] [FWKG06]. Sie wurde 1983 mit den gleichen Zielstellungen wie die Europäische Poa-Datenbank initiiert. Auch hier werden Backend und

[6] http://bic-gh.de/ebdb [Stand 2009-04-02]

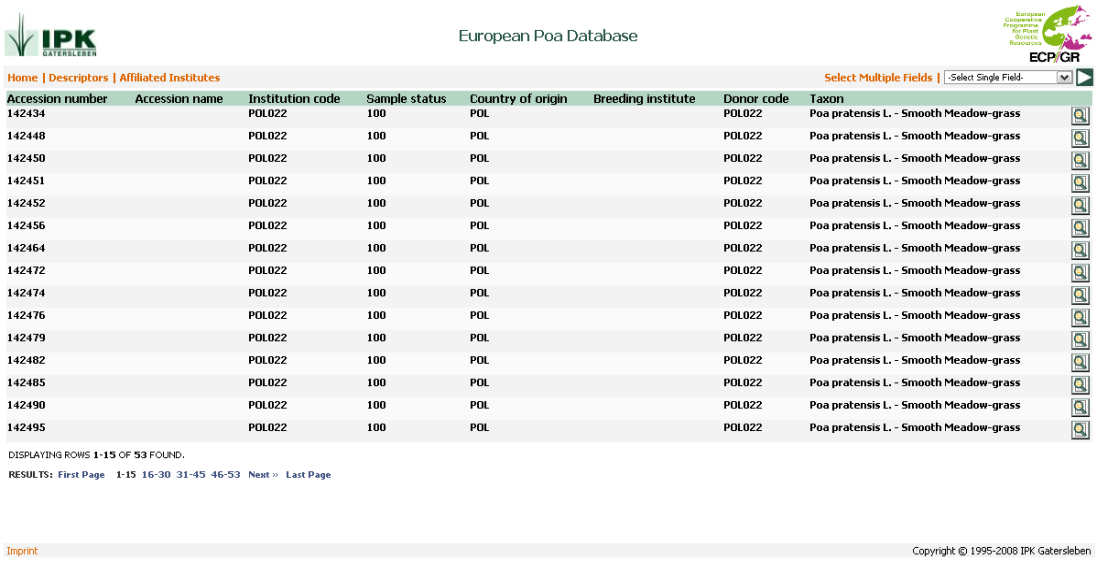


Abbildung 2.10: Ein Screenshot der EPDB-Oberfläche aus [WHG+07]

Frontend durch das relationale Datenbankmanagementsystem Oracle sowie eine JSP/ Java-Oberfläche gebildet.

Die Europäische Gerstendatenbank verwaltet Daten zu ca. 155.000 Gerstenakzessionen, die hauptsächlich in europäischen Genbanken gehalten werden. Hinzu kommen noch Daten zu Pflanzenmustern aus Japan und Australien.

Neben Passportdaten enthält die EBDB auch Charakterisierungsdaten zu allen 155.000 Akzessionen sowie Evaluierungsdaten zu ca. 4.000 Akzessionen, die im Rahmen eines europäischen Projektes[7] gewonnen wurden. Außerdem verwaltet diese Datenbank Informationen über die so genannte Barley Core Collection [KvH95]. Diese Kernkollektion umfasst die vergleichsweise geringe Anzahl von 1.293 Akzessionen. Sie besitzt aufgrund ihrer großen genetischen Diversität eine hohe Variabilität. Für Forschung und Züchtung ist sie eine wichtige Ressource.

Die Europäische Gerstendatenbank bildet ebenso wie die Europäische Poa-Datenbank ein wichtiges Bindeglied zwischen phänotypischen Daten und Markerdaten.

Metabolische Netzwerkdaten aus Meta-All/MetaCrop

Meta-All[8] [WGK+06] ist ein Informationssystem zur Verwaltung metabolischer Netzwerkdaten in Pflanzen. Dieses System speichert sehr detaillierte, feingranulare Daten

[7] Evaluation and Conservation of Barley Genetic Resources to Improve Their Accessibility to Breeders in Europe (EU-GENRES CT98-104)
[8] http://bic-gh.de/meta-all [Stand 2009-04-02]

über Netzwerke und ihre Bestandteile. Dabei wird sehr genau unterschieden, in welchem Organismus sowie an welchem Ort innerhalb dieses Organismus biochemische Prozesse aktiv sind. Weiterhin werden Entwicklungsstadien berücksichtigt. Daten können um eine Vielzahl von Literaturquellen und verschiedene Qualitätstags angereichert werden.

Alle in Meta-All gespeicherten Daten können parallel versioniert werden. Dies stellt einen großen Vorteil gegenüber der weit verbreiteten seriellen Versionierung dar. Zum einen ermöglicht es verteilt arbeitenden Forschergruppen, gemeinsam an Netzwerken zu arbeiten und beispielsweise im Falle abweichender wissenschaftlicher Meinungen parallele Versionen anzulegen. Zum anderen sind auf Basis einer bestimmten Version publizierte Daten immer zugreifbar.

Über eine SBML-Schnittstelle [HFS+03] erfolgt ein Austausch mit Visualisierungs- und Simulationswerkzeugen.

Eine Instanz des Meta-All-Informationssystems ist am IPK in Gatersleben unter der Bezeichnung MetaCrop[9] im Einsatz [GBWK+08, GBJK+08]. Sie enthält manuell kurierte, fein-granulare Stoffwechseldaten über sechs agronomisch bedeutende Kulturpflanzenarten. Abbildung 2.11 zeigt einen Ausschnitt der MetaCrop-Nutzerschnittstelle.

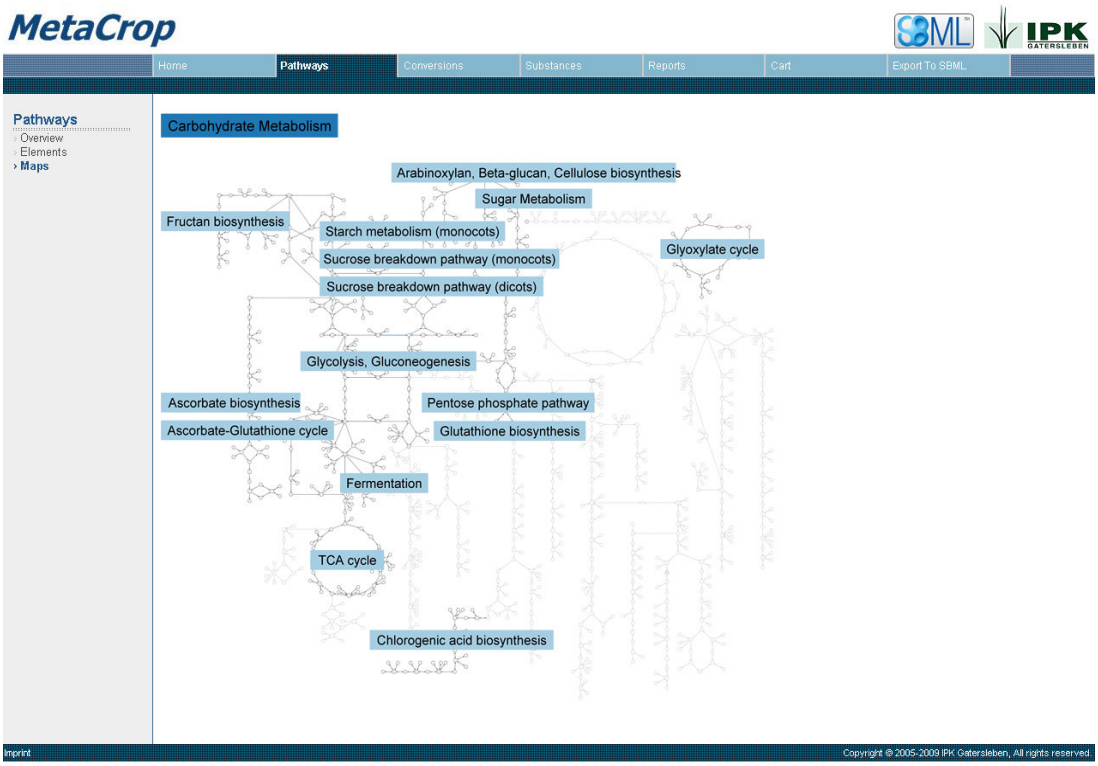


Abbildung 2.11: Ein Screenshot der MetaCrop-Oberfläche aus [GBWK+08]

[9] http://metacrop.ipk-gatersleben.de [Stand 2009-04-02]

2.3 Fachübergreifende Grundlagen

2.3.1 Kontrolliertes Vokabular

Ein kontrolliertes Vokabular ist eine Sammlung von Bezeichnungen mit einheitlicher Syntax, die Begriffen eindeutig zugeordnet sind und damit sicherstellen, dass keine Homonyme auftreten. Der umgekehrte Fall, dass jedem Begriff nur eine Bezeichnung zugeordnet ist, es also keine Synonyme gibt, ist nicht immer gewährleistet.

Eine große Bedeutung haben kontrollierte Vokabulare in der Dokumentationswissenschaft. Beispiele hierfür sind Fachwortverzeichnisse oder Glossare, die Begriffe einer Fachsprache definieren. Ein weiterer Anwendungsbereich kontrollierter Vokabulare ist die Informatik. Hier werden zur eindeutigen Kennzeichnung von Objekten Identifikatoren verwendet.

Der primäre Zweck für die Nutzung kontrollierten Vokabulars besteht darin, die Beschreibung von Objekten/Begriffen konsistent zu machen und die Suche nach diesen zu erleichtern. Zwei Eigenarten der natürlichen Sprache machen die Verwendung solchen Vokabulars notwendig [Ame05]:

- zwei oder mehr Terme beschreiben dasselbe Objekt (Synonyme) oder
- zwei oder mehr Objekte werden durch denselben Term beschrieben (Homonym).

Um ein kontrolliertes Vokabular zu erhalten, werden drei grundlegende Methoden eingesetzt:

- Definition des Anwendungsbereiches der Terme,
- Verknüpfen von synonymen Termen und
- Unterscheidung zwischen homonymen Termen.

Ein biologisches Anwendungsbeispiel kontrollierten Vokabulars ist die Nomenklatur der Enzymnummern[10] [WI92].

2.3.2 Taxonomie

Eine Taxonomie ist eine Sammlung von Termen eines kontrollierten Vokabulars, die hierarchisch strukturiert sind. Jeder einzelne Term befindet sich mit mindestens einem weiteren Term in einer Elter-Kind-Beziehung.

[10] http://www.chem.qmul.ac.uk/iubmb/enzyme [Stand 2009-04-02]

In der Biologie wird der Begriff der Taxonomie für eine hierarchische Klassifikation von Organismen verwendet, um deren Verwandtschaftsbeziehungen darstellen zu können. Wenn durch gemeinsame Merkmale eine Gruppe von Lebewesen unterscheidbar von anderen Gruppen von Lebewesen beschreibbar ist, wird diese Gruppe als Taxon bezeichnet. Die Klassifikation solcher Taxa erfolgt anhand verschiedener Stufen, wie z. B. Ordnung, Familie, Gattung, Art.

2.3.3 Ontologien

Ein großes Problem, gerade im interdisziplinären Bereich, besteht darin, dass vielfach unterschiedliche Vokabularien, Paradigmen, Sprachen und (Modellierungs-)Werkzeuge verwendet werden. Daraus resultieren Schwierigkeiten in der Kommunikation zwischen Individuen und damit auch bei der Identifikation und Spezifikation von Systemen. Dies kann außerdem zu Problemen bei der Interoperabilität und der Wiederverwendbarkeit führen. Beispiele dafür werden in Kapitel 4 erläutert. Ein Lösungsansatz ist die Vereinheitlichung und Spezifikation von Terminologien und Konzepten im Rahmen von Ontologien.

In der *Metaphysica generalis* (abgeleitet vom Werk „Metaphysik“ des Aristoteles), einem speziellen Teil der Philosophie, hat der Begriff der Ontologie (von griech. *on*: Seiendes; *logos*: Wort, Diskurs) als die Lehre vom Seienden eine lange Tradition. Als Begründer gilt der griechische Philosoph Parmenides. In der traditionellen Ontologie steht die Frage im Mittelpunkt, wie sich das Sein zum Seienden verhält. Im Gegensatz dazu ist die moderne analytische Ontologie die Lehre der grundlegenden Kategorien (Entität, Eigenschaft, Ereignis etc.) und ihrem Verhältnis zueinander. Für weitere Informationen vgl. [Wei91, RK98].

In der Informatik wird unter Ontologie die Definition von Klassen (Konzepte, Objekte) und ihren Beziehungen (Attribute, Rollen) untereinander verstanden [Gru93]. Sie ist formal definiert und enthält das Vokabular einer Domäne/eines Bereiches, mit dem Ziel, die Kommunikation zwischen Menschen/Organisationen und die Interoperabilität zwischen Systemen zu verbessern. Ontologien werden nicht von einer einzelnen Person genutzt, sondern durch eine Gruppe vereinbart.

Zur Formulierung von Ontologien werden Ontologiesprachen verwendet. Dies ermöglicht den maschinenverarbeitbaren Austausch zwischen verschiedenen Anwendern/Anwendungen. Als Beispiel sei das Resource Description Framework (RDF)[11] genannt.

Auch in der Biologie fallen sowohl immer größere Datenmengen als auch immer komplexere Daten an, die beispielsweise molekulare Interaktionen beschreiben. Um solche Daten zu verwalten und zueinander in Beziehung setzen zu können, sind vergleichbare Strukturen erforderlich. Hierfür werden auch in diesem Bereich zunehmend Ontologi-

[11] http://www.w3.org/RDF [Stand 2009-04-02]

en genutzt [BR04]. Diese Bioontologien sind formale Repräsentationen verschiedener biologischer Wissensbereiche, deren Objekte zueinander in Beziehung stehen, z. B.

- Gene Ontology (biologische Prozesse, zelluläre Komponenten) [GO 08],
- Trait Ontology (phänotypische Merkmale) [JWN+02] oder
- Plant Ontology (Anatomie und Entwicklungszustände von Pflanzen) [ATI+08]).

Dieses biologische Wissen kann dann mit molekularen Daten, aber auch gleichermaßen mit phänotypischen Daten etc. integriert werden. Abbildung 2.12 zeigt einen Ausschnitt aus der Gene Ontology.

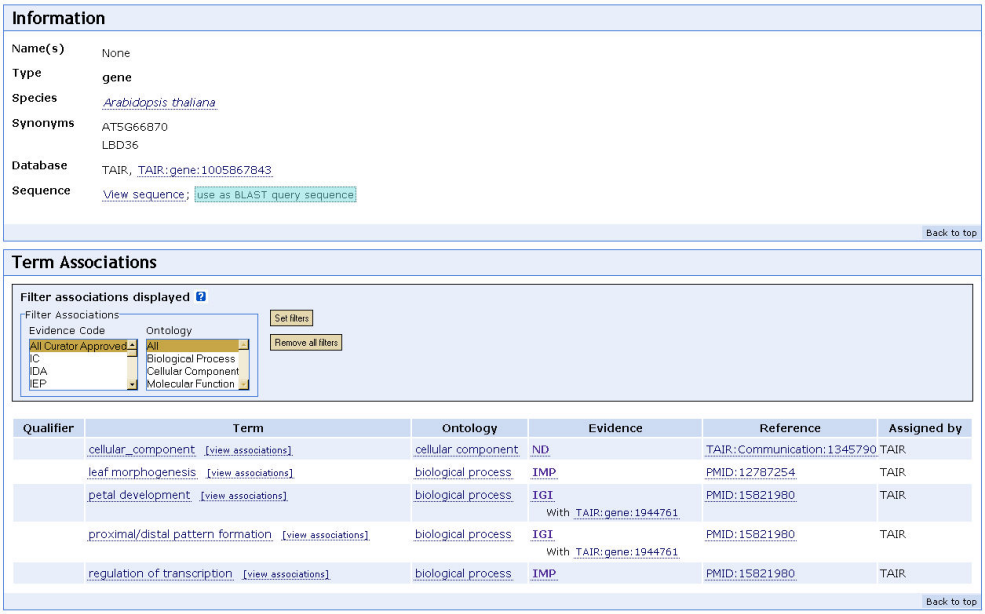


Abbildung 2.12: Ein Ausschnitt aus der Gene Ontology nach [GO 08]

Für weitergehende Informationen sei auf [Gru95, SK02, BR04] verwiesen.

2.3.4 Merkmale und Skalen

Merkmalsausprägungen werden mit Skalen gemessen, welche aus Anordnungen von Werten bestehen [Sch94]. Zum besseren Verständnis der Verfahren zur Datenanalyse, die in Abschnitt 3.2 vorgestellt werden, sollen im Folgenden kurz die wichtigsten Typen von Merkmalen und ihre Merkmalsskalen beschrieben werden.

Nominalskala Von nominal messbaren Werten wird gesprochen, wenn diese nur über die Aussagen gleich oder verschieden geordnet werden können, d. h. es gibt keine natürliche Reihenfolge der Werte. Ein typisches Beispiel ist das Merkmal Geschlecht, das die Ausprägungen weiblich und männlich annehmen kann. Nominal messbare Merkmale heißen daher auch qualitative Merkmale.

Rang- oder Ordinalskala Wenn nominal messbare Werte zusätzlich nach einer natürlichen Reihenfolge geordnet werden können, handelt es sich dabei um ordinal messbare Werte. Ein Beispiel hierfür sind Zensuren. Hierbei kann keine Aussage über den absoluten Wert einer Merkmalsausprägung getätigt werden, z. B. kann nicht gesagt werden, die Zensur 2 ist doppelt so gut wie die Zensur 1 etc. Hier wird auch von intensitätsmäßigen Merkmalen gesprochen. Ein Beispiel aus dem pflanzenbiologischen Bereich bilden Boniturnoten (1 = geringe Merkmalsausprägung, . . . , 9 = hohe Merkmalsausprägung).

Metrische oder Kardinalskala Bei metrisch messbaren Merkmalen sind die Merkmalswerte reelle Zahlen. Solche Merkmale können gemessen werden und verfügen über eine Dimension (Einheit), z. B. Pflanzenhöhe [cm]. Daher werden sie auch als quantitative Merkmale bezeichnet.

2.4 Resümee

In diesem Kapitel wurden Grundlagen für das Verständnis dieser Arbeit präsentiert. Hierzu erfolgte die Vorstellung von Datenbanksystemen sowie von Möglichkeiten der Datenmodellierung. Da die Integration von Daten aus verschiedenen Domänen in dieser Arbeit eine bedeutende Rolle spielt, wurde danach auf Methoden des Record Linkage eingegangen.

Aufgrund der interdisziplinären Ausrichtung der vorliegenden Arbeit, wurden neben Grundlagen der Informatik auch biologische präsentiert. Hierbei lag der Fokus insbesondere auf relevanten biologischen Datendomänen sowie besonderen Datenressourcen. Abschließend erfolgte die Vorstellung fachübergreifender Grundlagen.

3 Datenintegration und -analyse

Auf den allgemeinen Grundlagen aus Kapitel 2 aufbauend sollen auf den folgenden Seiten Konzepte der Datenintegration sowie der Datenanalyse dargestellt werden, die für das Verständnis der vorliegenden Arbeit notwendig sind.

3.1 Datenintegration

Nach [Wie96] ist Integration ein Service, der Inhalte von multiplen, oftmals heterogenen, Datenquellen kombiniert. Dabei besteht das Ziel darin, aus diesen Kombinationen neue Erkenntnisse zu gewinnen.

Bei der Integration von Daten wird zwischen zwei Bereichen unterschieden. Zum einen gibt es die klassische Datenintegration, die vorzugsweise im betrieblichen Umfeld stattfindet. Dabei werden bestehende betriebliche Informationssysteme integriert oder fusioniert, um einen einheitlichen Zugriff auf die darin gehaltenen Daten zu erreichen [Mer97]. Die Datenintegration in diesem Bereich ist Bestandteil des vor einigen Jahren aufgekommenen Begriffes Enterprise Application Integration (EAI) [RMB00], unter dem Technologien zur unternehmensweiten Integration von Geschäftsfunktionen verstanden werden. Zum anderen gibt es die Integration von Daten aus heterogenen Quellen, vielfach aus dem WWW, die mindestens miteinander und nach Möglichkeit zusätzlich mit eigenen Daten verknüpft werden sollen. Hierunter fällt in vielen Fällen die Datenintegration in der Bioinformatik, die einen Schwerpunkt dieser Arbeit bildet.

Datenbanken in der Biologie zeichnen sich oftmals durch eine Reihe von Heterogenitäten aus. Dazu zählen Heterogenitäten bezüglich ihrer Anwendbarkeit, ihrer Struktur,

ihres Inhalts sowie der Plattformen, auf denen sie basieren. Viele dieser Datenbanken wurden ad hoc erstellt und sind im Verlauf der Zeit gewachsen [BK03]; sie sind vielfach ungenügend strukturiert und basieren hauptsächlich auf Flatfiles. Ein Beispiel dafür ist das japanische KEGG-System [KGH+06]. Datenzugriffe sind in diesem Bereich auf verschiedene Arten möglich, über (statische) Web-Seiten, per Web-Interface oder Web-Services, über Datenbankdumps, aber auch über direkte Datenbankzugriffe.

Nach [Sch98] können die verschiedenen Arten der Heterogenität folgendermaßen klassifiziert werden:

- **Heterogenität auf Systemebene:**

  Komponentendatenbanksysteme besitzen unterschiedliche Systemeigenschaften. Hierbei kann es sich z. B. um Optimierungsstrategien, Transaktionsmodelle etc. handeln. Dies kann bereits bei der Entwicklung des Integrationssystems durch spezifische Datenbankadapter o. ä. überwunden werden.

- **Heterogenität auf Datenmodellebene:**

  Hiervon wird gesprochen, wenn unterschiedliche Datenbankmodelle und -sprachen Anwendung finden. Die jeweiligen Datenbankmodelle (hierarchisch, relational etc.; vgl. Abschnitt 2.1.1) müssen vom Integrationssystem unterstützt und in ein einheitliches Datenbankmodell überführt werden.

- **Heterogenität auf Schemaebene:**

  Diese Art der Heterogenität tritt dann auf, wenn gleiche oder ähnliche Daten unterschiedlich repräsentiert werden. Hierzu muss beim Entwurf des Integrationssystems eine Schemaintegration durchgeführt werden. Diese ist z. T. nur manuell möglich.

- **Heterogenität auf Datenebene:**

  Wenn für semantisch äquivalente Datenbankobjekte mit bestimmten Eigenschaften unterschiedliche Daten existieren, wird von Heterogenität auf Datenebene gesprochen. Solche Objekte müssen vom Integrationssystem zur Laufzeit erkannt werden. Dabei sind ebenfalls teilweise manuelle Eingriffe notwendig.

Die Mehrheit der existierenden biologischen Datenquellen steht grundsätzlich zueinander in Beziehung [BK03] und kann daher integriert werden. Bei der Integration

biologischer Daten wird traditionell zwischen zwei Vorgehensweisen unterschieden [DOB95]:

- **Virtuelle oder logische Integration:**

  Dies ist der Standardansatz zur Integration von Internet-Datenquellen. Eine Anfrage wird an verschiedene Datenquellen gesendet und die Ergebnisse werden zu einem Report kombiniert. Sowohl das Absetzen der Anfrage als auch das Erstellen des Ergebnisses erfolgen zur Laufzeit. Daten werden nicht lokal gespeichert.

- **Materialisierte oder physische Integration:**

  Dieses Vorgehen basiert auf der lokalen Speicherung von Daten. Datenquellen werden hierbei regelmäßig auf neue Daten hin durchsucht. Diese werden in einer Datenbank gespeichert und vorhandene Daten entsprechend aktualisiert. Anfragen richten sich direkt an die Datenbank. Ein typischer Vertreter dieses Ansatzes ist ein Datawarehouse-System.

Weiterhin gibt es auch hybride Ansätze (z. B. [KDKR05]). Hierbei werden Datenquellen beispielsweise automatisch durchsucht und es wird ein Index über die vorhandenen Daten erstellt. Anfragen richten sich an diesen Index. Dieser ermöglicht dann den direkten Zugriff auf die Datenquellen.

In den folgenden Abschnitten werden die virtuelle und die materialisierte Integration detaillierter vorgestellt.

3.1.1 Virtuelle Integration

Zur Verdeutlichung der virtuellen Integration wird dieses Konzept anhand der beiden Ansätze der Multidatenbanksysteme und der Mediatorensysteme erläutert.

Multidatenbanksysteme (MDBS)

Multidatenbanksysteme sind Systeme, die aus mehreren separaten Datenbanksystemen zusammengesetzt sind [SH99]. Sie dienen dazu, Daten aus z.T. heterogenen Datenbanken zu extrahieren und sie dem Nutzer in einer homogenen Sicht zu präsentieren. Dabei wird eine logische Integration durchgeführt [Sch98].

Da MDBS selber Datenbanksysteme sind, verfügen sie über Datenbankfunktionalitäten. Sie ermöglichen die Erstellung eines Verbundes über mehrere Datenbanksysteme.

Nach [Sch98] werden zwei Anwendungsszenarien für Multidatenbanksysteme unterschieden:

1. Beim ersten Szenario wird bereits während des Entwurfs festgelegt, dass eine Datenbank von mehreren Datenbankmanagementsystemen verwaltet werden soll. Die Datenbank wird dazu in voneinander unabhängige Partitionen unterteilt. Gründe dafür können Performance und Autonomie sein.

2. Beim zweiten Szenario sollen mehrere bestehende Datenbanksysteme zusammengefasst werden, ohne dabei bereits vorhandene Anwendungen und Datenbestände anzutasten. In diesem Fall kann durch logische Integration eine homogene Sicht auf die gesamte Datenbank erstellt werden.

Abbildung 3.1 zeigt die Unterteilung von Multidatenbanksystemen in Subtypen nach [SL90].

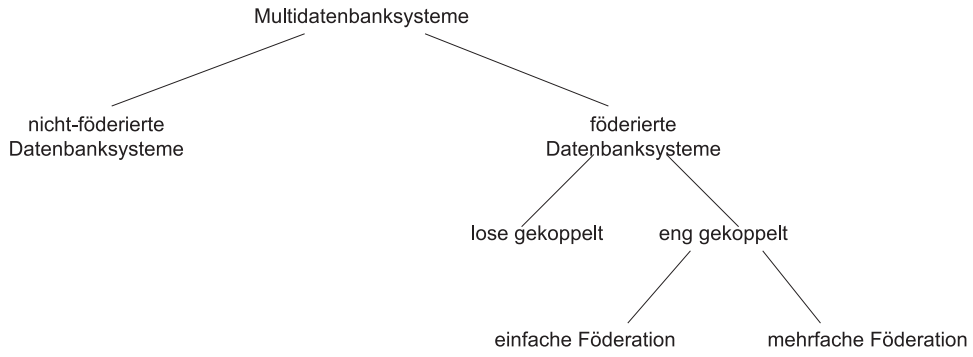


Abbildung 3.1: Unterteilung von Multidatenbanksystemen nach [SL90]

Nicht-föderierte Datenbanksysteme: Nicht-föderierte Datenbanksysteme sind solche Multidatenbanksysteme, bei denen die einzelnen Komponentendatenbanken nicht autonom sind. Daher wird nicht zwischen lokalen und globalen Nutzern unterschieden.

Föderierte Datenbanksysteme (FDBS): Im Gegensatz zu den nicht-föderierten handelt es sich bei den föderierten Datenbanksystemen um Multidatenbanksysteme, bei denen die Autonomie der Komponentendatenbanken erhalten bleibt. Die Daten werden von den jeweiligen Komponentendatenbankmanagementsystemen kontrolliert und es wird zusätzlich zwischen lokalen und globalen Nutzern unterschieden.

Föderierte Datenbanksysteme können weiterhin in lose gekoppelte sowie eng gekoppelte unterteilt werden. Im Fall der losen Kopplung muss jeder Anwender seine spezielle Föderation entwerfen und verwalten. Bei der engen Kopplung wird dies vom Administrator übernommen. Gibt es hierbei nur ein föderiertes Schema, wird von einer einfachen Föderation gesprochen, bei mehreren föderierten Schemata von mehrfacher Föderation.

Abbildung 3.2 illustriert die Architektur eines föderierten Datenbanksystems nach [Con97]. Globale Anwendungen greifen hier über einen gemeinsamen Föderierungsdienst auf die einzelnen, autonomen Komponentendatenbanksysteme zu. Im Gegensatz dazu erhalten lokale Anwendungen direkt über die Datenbankmanagementsysteme der jeweiligen Komponentendatenbanksysteme Zugriff.

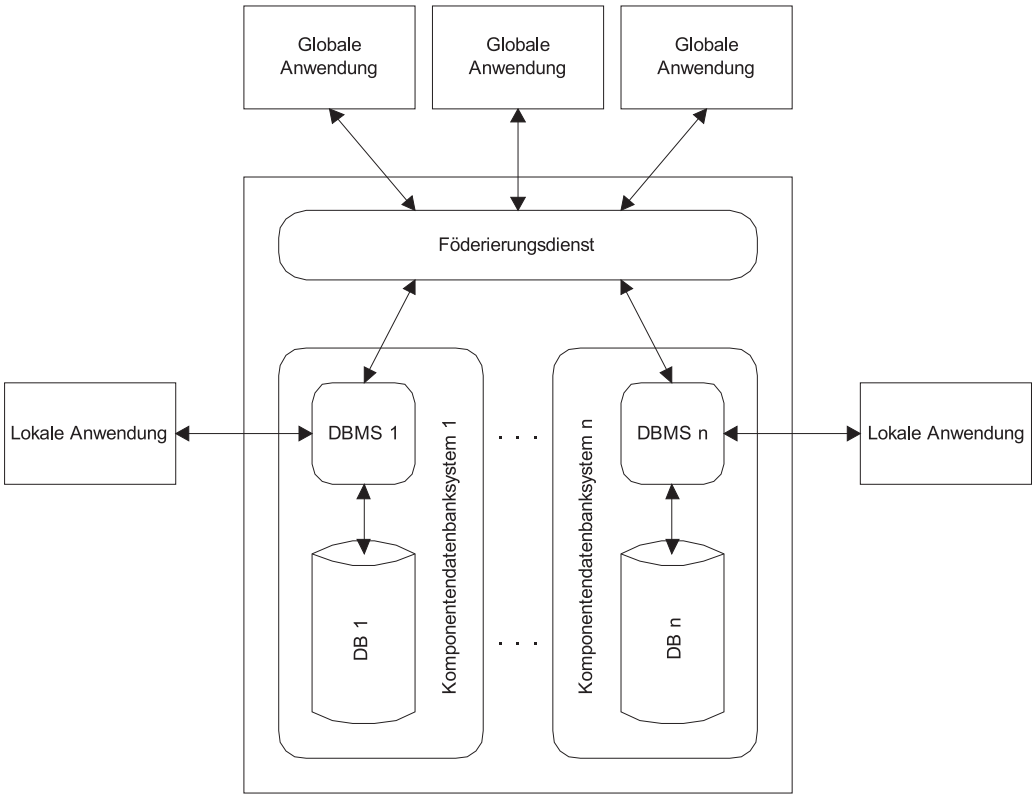


Abbildung 3.2: Schema eines föderierten Datenbanksystems nach [Con97]

Die durch ein föderiertes Datenbanksystem verwalteten Schemata bilden gemeinsam die so genannte Schemaarchitektur. Allgemein üblich ist die 5-Ebenen-Schemaarchitektur nach [SL90], die in Abbildung 3.3 gezeigt wird.

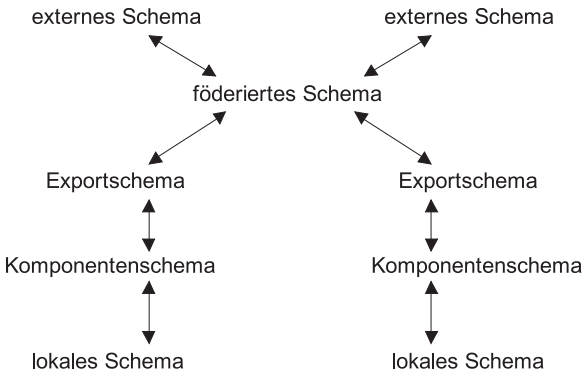


Abbildung 3.3: 5-Ebenen-Schemaarchitektur nach [SL90]

Bei der 5-Ebenen-Schemaarchitektur kommen drei Methoden zum Einsatz:

- **Schematranslation:**

  Hierbei erfolgt die Überführung eines Schemas von einem Datenmodell in ein anderes.

- **Schemaintegration:**

  Mehrere heterogene Schemata werden in ein gemeinsames integriert.

- **Schematransformation:**

  Ein Schema wird modifiziert. Ausgangs- und Zielschema sind dabei nicht identisch.

Im Folgenden werden die Zusammenhänge zwischen diesen drei Begriffen kurz dargestellt. Die lokalen Schemata einer Komponentendatenbank entsprechen den konzeptuellen Schemata. Diese heterogenen Schemata werden im Rahmen einer Translation in das gemeinsame Datenmodell eines föderierten Datenbanksystems (kanonisches Datenmodell) überführt. Die damit vorliegenden Komponentenschemata werden vom Komponentendatenbankmanagementsystem verwaltet, das die Entscheidung darüber trifft, welche der Daten dem föderierten Datenbanksystem zur Verfügung stehen sollen. Aus diesen Daten wird das Exportschema gebildet. Aufgrund der Heterogenitäten der Exportschemata untereinander sind Schemaintegrationen erforderlich, die die Daten in das föderierte Schema (Beschreibung der föderierten Datenbank im kanonischen Schema) überführen. Um verschiedenen Applikationen anwendungsspezifische Sichten zur Verfügung zu stellen, muss im abschließenden Schritt eine Transformation der Daten vom föderierten Schema in externe Schemata durchgeführt werden.

Mediatorensysteme

Während in den vorangegangenen Abschnitten die Integration heterogener Daten auf der Basis von Datenbanksystemen erläutert wurde, gibt es, insbesondere im Bereich der Bioinformatik, große Datenmengen, die nicht in Datenbanken verfügbar sind, sondern in z. T. proprietären Formaten wie HTML-Seiten, Textdateien etc.

Hierzu bietet es sich an, eine zusätzliche Vermittlungsschicht – einen Mediator – zu verwenden, die die unterschiedlichen Formate der jeweiligen Datenquellen versteht. Hierfür wurde von [Wie97] eine 3-Schichten-Architektur (Applikation – Mediator – Datenquelle) vorgeschlagen, die in Abbildung 3.4 gezeigt wird.

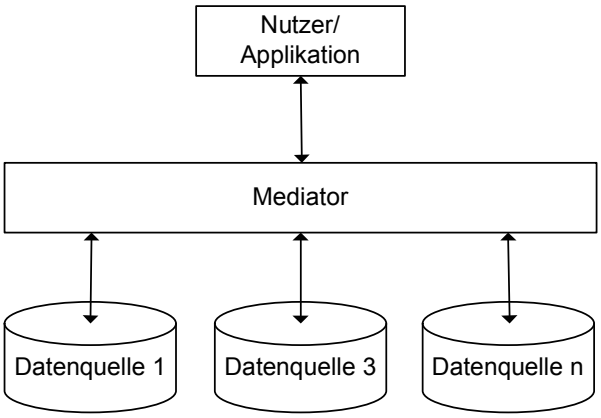


Abbildung 3.4: 3-Schichten-Architektur für die Integration nach [Wie97]

Als Unterstützung von Mediatoren werden oftmals so genannte Wrapper eingesetzt. Dieser Begriff soll vor der eigentlichen Beschreibung von Mediatoren erklärt werden.

Wrapper: Wrapper sind Softwarekomponenten, die Daten und Anfragen von einem Modell in ein anderes konvertieren [CHS+95, PGMW95].

Obwohl für solche Softwarekomponenten keine standardisierten Anforderungen existieren, stimmen die von verschiedenen Autoren im Rahmen von Integrationsprojekten formulierten Anforderungen weitgehend überein.

Nach [HSK+01] und [RS97] hat ein Wrapper vier Aufgaben zu erfüllen:

1. Informationen in ein einheitliches Datenmodell zu überführen,

2. Informationen darüber zurückzugeben, wie eine Datenquelle abgefragt werden kann,

3. Transformation von Abfragen, die an den Wrapper übergeben werden, in ein von der Datenquelle verstandenes Format sowie

4. Abfragen auszuführen und Resultate zurückzugeben.

Sollen Informationen aus verschiedenen Datenbeständen abgefragt werden, erhält ein Mediator diese Anfrage. Dieser verfügt über das Wissen, in welchem Datenbestand welche Information enthalten ist, jedoch nicht, wie sie im Detail abgefragt wird. Deshalb spricht er den Wrapper über eine vordefinierte Schnittstelle an. Der Wrapper antwortet über dieselbe Schnittstelle.

Wenn es identische Mediator-Wrapper-Schnittstellen für mehrere Wrapper gibt, können Wrapper dem Verbergen der Heterogenität von verschiedenen Datenbeständen dienen. Da die jeweilige Abfrageschnittstelle speziell an eine Datenquelle angepasst ist, ist für jede Datenquelle ein spezifischer Wrapper erforderlich.

Zum Erstellen von Wrappern existieren verschiedene Strategien. Einerseits ist es möglich, Wrapper individuell zu programmieren (häufig ad hoc). Dieses Vorgehen kann dazu führen, dass eine Vielzahl von Wrappern existiert, welche sich beispielsweise im verwendeten Datenmodell unterscheiden. Ein weiteres Problem besteht darin, dass die Semantik solcher Wrapper häufig im Quellcode verborgen ist, wodurch die Anpassung und Wiederverwendbarkeit erschwert wird. Andererseits ist es möglich, Wrapper semiautomatisch zu erzeugen. Hierdurch wird die Kompatibilität erhöht und die Einbindung in Mediatoren erleichtert. Als Beispiel seien hier *Garlic*[1] und *TSIMMIS*[2] genannt.

Mediator: Mediatoren sind Software-Module, die zwischen Nutzeranwendungen und Datenquellen vermitteln. Aufgrund ihres Wissens über bestimmte Datenquellen und deren Abfragemöglichkeiten sind sie in der Lage, Informationen für eine darüberliegende Applikationsschicht bereitzustellen [Wie97]. Sie spielen bei der Integration von Daten eine zentrale Rolle.

Mediatoren stellen dynamische Schnittstellen dar und bilden so eine Mittel- oder Zwischenschicht, die Nutzeranwendungen von Datenquellen unabhängig macht. Sie ermöglichen dadurch für den Anwender eine homogene Sicht auf unterschiedliche Arten von Datenquellen.

Ein Mediator erhält Nutzeranfragen in einer standardisierten Anfragesprache. Er verfügt über das Wissen, in welchen Datenquellen die gewünschten Informationen gehalten werden. Üblicherweise werden die Datenquellen von speziell darauf ausgelegten

[1] http://www.almaden.ibm.com/cs/garlic [Stand 2009-04-02]

[2] The Stanford-IBM Manager of Multiple Information Sources, http://www-db.stanford.edu/tsimmis [Stand 2009-04-02]

Wrappern angesprochen, die die Schnittstellen zwischen Mediator und Datenquellen bilden.

Obwohl Mediatoren auch direkt mit Datenquellen interagieren könnten, würde dies für jede Anbindung einer neuen Quelle eine Adaptation des Mediators bedeuten. Anpassungen wären auch bei Änderungen der Datenquellen notwendig. Daher ist es vorteilhafter, jeder Datenquelle einen Wrapper zuzuordnen.

Der Mediator formt an ihn übergebene Anfragen um und verteilt sie an einen oder mehrere Wrapper. Die von diesen zurückgegebenen Resultate werden vom Mediator in das oben erwähnte standardisierte Format transformiert und an den Nutzer übermittelt.

In [SH99] werden Mediatoren nach der Anzahl der Datenquellen, auf die sie zugreifen, unterteilt (siehe Abbildung 3.5). Mediatoren mit Zugriff auf nur eine Datenquelle führen eine Aggregation und Abstraktion von Daten durch, solche mit Zugriff auf mehrere zusätzlich eine Bereinigung und Integration der Daten.

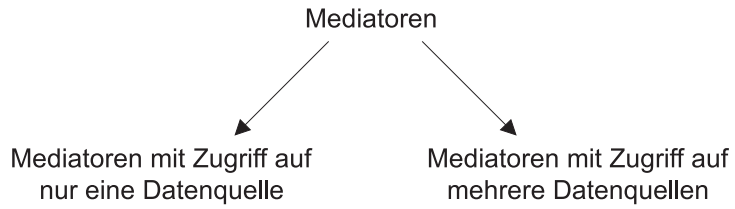


Abbildung 3.5: Unterteilung von Mediatoren nach [SH99]

Eine automatisierte Integration von Daten ist, bedingt durch die Heterogenitäten der Struktur, Schnittstellen und Implementierungen von Mediatoren, sehr schwierig. Daher werden in [Wie97] konzeptuelle Anforderungen vorgeschlagen:

- Mediatoren sollen klein,
- einfach und
- von potentiellen Nutzern überprüfbar sein.

Für kleine und einfache Mediatoren spricht, dass für Entwicklung und Wartung solcher Softwaremodule nur begrenzte Ressourcen erforderlich wären. Böten sie zusätzlich Informationen über sich selbst, wäre es Nutzern möglich, ihre Funktionsweise zu ermitteln, z. B. die verwendeten Regeln.

Zum Betrieb eines Mediators sind zwei Schnittstellen vonnöten. Hierfür wird von [Wie97] die Verwendung von Kommunikationsprotokollen vorgeschlagen.

- **Schnittstelle Nutzer — Mediator:**

  Diese Schnittstelle dient der Weiterleitung der Anfrage vom Nutzer zum Mediator in einer wohldefinierten Anfragesprache wie z. B. XML.

- **Schnittstelle Mediator — Datenquelle:**

  Die Vorgaben dieser Schnittstellen werden durch die zugrunde liegenden Datenquellen gebildet. Der Datenzugriff kann hier etwa über SQL, aber auch über CGI-Aufrufe o. ä. erfolgen. Diese Schnittstelle wird vielfach durch Wrapper realisiert.

Ein Mediator empfängt Anfragen der Applikationsschicht und leitet sie an einen oder mehrere Wrapper weiter. Der Mediator verfügt hierbei über das Wissen, welche Quellen die erforderlichen Daten bereitstellen können und legt die Abfragereihenfolge fest. Die Anfrage wird in Teile zerlegt, in die spezifischen lokalen Schemata der einzelnen Wrapper überführt und an diese übergeben.

Die dem Mediator im jeweiligen Schema der Wrapper übergebenen Resultate werden anschließend in das einheitliche, globale Schema überführt und an die Applikationsschicht weitergeleitet.

3.1.2 Materialisierte Integration

In den vorhergehenden Abschnitten wurden Verfahren zur virtuellen Integration von Daten vorgestellt. Anhand des Datawarehouse-Ansatzes wird nun die materialisierte Integration beschrieben.

Datawarehouses

Operativsysteme dienen der Speicherung und Verwaltung operativer Daten, d. h. Daten, die für den laufenden Geschäftsbetrieb eines Unternehmens erforderlich sind. Ihre Datenstrukturen und Abfragewerkzeuge sind auf Routineaufgaben abgestimmt. Operativsysteme sind transaktionsorientiert, sie speichern keine historischen Daten sondern Momentaufnahmen. Es wird auch von OLTP-Systemen (OnLine Transactional Processing) gesprochen.

Dies ist grundsätzlich auch auf die Bioinformatik übertragbar. Ein Beispiel für ein bioinformatisches Operativsystem ist das in Abschnitt 2.2.4 beschriebene Genbankinformationssystem GBIS.

Ende der 1980er Jahre wurde das Konzept der analytischen Datenbanken oder Datawarehouses eingeführt [DM88]. Die Intention lag in der Trennung der operativen Daten von solchen Daten, die zur Entscheidungsunterstützung oder für das Berichtswesen verwendet werden. Damit spielt die Zeitabhängigkeit von Daten bei Datawarehouses eine große Rolle. Weiterhin dienen Datawarehouses der Integration von Daten aus einer Vielzahl von Quellen.

Es gibt keine eindeutige Definition des Begriffes Datawarehouse. Allgemein akzeptiert ist die Definition eines Datawarehouses nach [Inm05] als Datensammlung mit den folgenden Eigenschaften:

- **subjektorientiert:**

  Modellierung eines spezifischen Anwendungsziels

- **integriert:**

  Nutzung von Daten aus einer Vielzahl von Quellen

- **nicht-flüchtig:**

  Persistente Speicherung von Daten im Warehouse

- **zeitabhängig:**

  Verwendung von Daten, die über einen längeren Zeitraum erhoben wurden, für Zeitreihenanalysen etc.

Architektur: Abbildung 3.6 zeigt schematisch den Prozess der Erstellung eines Datawarehouses.

Durch so genannte Extraktions-Transformations-und-Lade-Prozesse (ETL-Prozesse) werden Daten aus unterschiedlichen, inner- und außerbetrieblichen Quellen in einem Warehouse-Schema vereinigt. Sie laufen in einer Staging Area (Datenbeschaffungsbereich) ab.

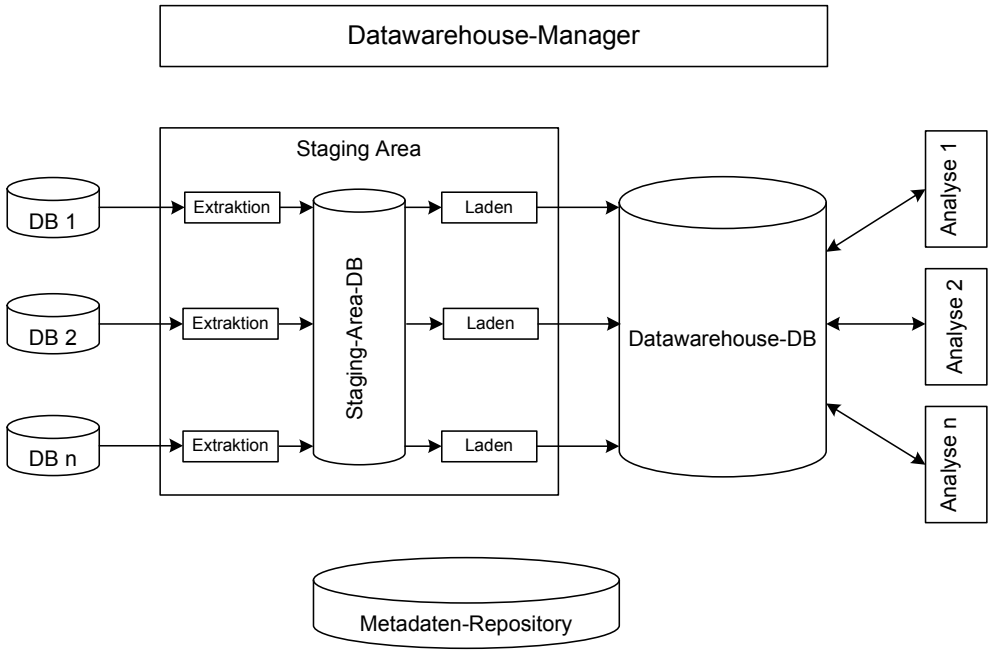


Abbildung 3.6: Schematische Darstellung der Datawarehouse-Erstellung

ETL-Prozesse bestehen aus drei Teilen:

- **Extraktion:**

  Extraktion einer relevanten Teilmenge der Daten aus einer Quelle in die Staging Area. Es erfolgt eine erste Schematransformation (vgl. Abschnitt 3.1.1).

- **Transformation:**

  Transformation der Quelldaten. Dies umfasst u.a. das Mapping unterschiedlicher Wertebereiche aufeinander, die Anpassung von Datentypen und Ausprägungen, die Duplikatentfernung sowie die Aggregation von Werten.

- **Laden:**

  Laden der transformierten Daten aus der Staging Area in das Warehouse-Schema. Hierbei erfolgt eine zweite Schematransformation.

Ein so genanntes Metadaten-Repository dient als Speicher für die für das Datawarehousing erforderlichen Metadaten. Dazu zählen vor allem Informationen über

- angeschlossene Datenquellen,

- ETL-Prozesse und

- die Generierung von Reports etc.

Die Erstellung eines Datawarehouses, insbesondere die Ausführung von ETL-Prozessen, wird durch einen Data-Warehouse-Manager überwacht.

Datamart: Neben dem zentralen Datenlager, der Warehouse-Datenbank, können eine Reihe von Datamarts existieren. Datamarts sind bereichs- oder auswertungsspezifische Sichten auf Daten, die der Beantwortung bestimmter Fragestellungen dienen. Daten in Datamarts verfügen häufig über mehrere Dimensionen und sind aggregiert.

Es gibt verschiedene Gründe, neben dem zentralen Datawarehouse Datamarts zu erstellen:

- es ist eine bestimmte Aufgabe zu erfüllen,
- es gibt eine bestimmte Nutzergruppe, z. B. einen Organisationsbereich oder eine Abteilung,
- es sind spezielle Datenstrukturen erforderlich,
- es gibt Performancegründe, insbesondere bei Abteilungslösungen.

Im Warehousebereich gibt es die sich widersprechenden Ansätze von Inmon [Inm05] und Kimball [Kim98]. Von Inmon wird ein Top-Down-Ansatz propagiert, in dem das Datawarehouse als zentrales Datenlager gesehen wird, von dem alle Datamarts abgeleitet werden. Im Gegensatz dazu beginnt beim Bottom-Up-Ansatz nach Kimball das Datawarehousing mit der Erstellung von Datamarts. Das unternehmensweite Datawarehouse stellt hierbei nur eine virtuelle Kollektion aller Datamarts dar.

Operational Data Store (ODS): Daten liegen in Datawarehouse-Datenbanken und in Datamarts hauptsächlich aggregiert vor. Um auch atomare Daten abspeichern zu können, wurde von [Inm99] das Konzept des Operational Data Stores vorgeschlagen. Auch hier werden im Rahmen des Datawarehouseprozesses aus mehreren Quellen integrierte, bereinigte und subjektorientierte Daten abgespeichert. Im Gegensatz zur Datawarehouse-Datenbank sind im Operational Data Store liegende Daten jedoch

- flüchtig,
- aktuell und
- detailliert.

Operational Data Stores wurden ursprünglich für administrative Berichtsaufgaben entwickelt, d. h. für einfache Anfragen über kleinere Datenmengen. Sie werden regelmäßig aus den angeschlossenen Operativsystemen aktualisiert. Daten eines Operational Data Stores können, aber müssen nicht, vom Datawarehouse weiterverarbeitet werden. Je nach Implementierung bilden sie die Vorstufe zum analytischen Datawarehouse oder werden alternativ bzw. ergänzend zu Datawarehouses eingesetzt. In den letzten Jahren werden Operational Data Stores im Unternehmensbereich verstärkt für das Customer Relationship Management (CRM) genutzt.

3.2 Datenanalyse

Nachdem auf den vorhergehenden Seiten Konzepte der Integration von Daten vorgestellt wurden, sollen im Folgenden Möglichkeiten zur darauf aufbauenden Analyse pflanzenbiologischer Daten beschrieben werden. Hierzu gibt es grundsätzlich zwei Ansätze, die modellgetriebene und die datengetriebene Analyse.

- **Modellgetriebene Analyse:**

  Modellgetriebene Analysen gehen von Hypothesen (von einem Modell der Wirklichkeit abgeleitet) aus und überprüfen diese Hypothesen anhand von Stichproben. Dies ist traditionell das typische Vorgehen in der pflanzenbiologischen Forschung.

- **Datengetriebene Analyse:**

  Datengetriebene oder explorative Analysen hingegen beschreiben und verallgemeinern Muster in Daten (hier: die Gesamtheit der Daten). Dabei werden Hypothesen und Modelle abgeleitet, die dann gezielt überprüft werden können. Dies ist insbesondere im Bereich von High-Throughput-Daten ein viel versprechender Ansatz.

Zur Analyse von in relationalen Datenbankmanagementsystemen vorliegenden pflanzenbiologischen Daten bieten sich die drei Möglichkeiten

- Datenbanksprachen (Abschnitt 3.2.1),
- OnLine Analytical Processing (Abschnitt 3.2.2) und
- Knowledge Discovery in Databases (Abschnitt 3.2.3)

an, die in den folgenden Abschnitten erläutert werden.

3.2.1 Datenbanksprachen

Die am weitesten verbreitete Datenbanksprache ist die Structured Query Language (SQL). Hierbei handelt es sich um eine mächtige und standardisierte Sprache, die sich gut für Ad-hoc-Anfragen und -Analysen eignet.

SQL ist eine Weiterentwicklung von SEQUEL [CB74]. Es existieren mehrere standardisierte Versionen. Daneben gibt es eine Reihe von SQL-Dialekten in verschiedenen Datenbankmanagementsystemen.

SQL ist eine deklarative Anfragesprache für relationale Datenbanken, d. h. mit ihr wird formuliert, was abgefragt werden soll, aber nicht, wie dies geschieht. Sie verfügt über eine vergleichsweise einfache Syntax, die an die englische Sprache angelehnt ist. Mit Hilfe mehrerer Teilsprachen [HS97] bietet SQL eine Vielzahl von Befehlen zur

- Schematadefinition (Data Definition Language, DDL),
- Datenmanipulation (Data Manipulation Language, DML),
- Einrichtung interner Zugriffspfade (Storage Structure Language, SSL) und
- Anfrageformulierung (Interactive Query Language, IQL).

Für gelegentliche Benutzer, insbesondere Experimentatoren mit geringen Informatikkenntnissen, ist sie jedoch zu kompliziert.

Im Rahmen dieser Arbeit werden Datenbanksprachen nicht für die eigentlichen Analysen, sondern nur für vorbereitende Schritte eingesetzt. Daher soll an dieser Stelle nicht weiter auf sie eingegangen werden.

3.2.2 OnLine Analytical Processing (OLAP)

Das OnLine Analytical Processing (OLAP) gehört zu den hypothesengetriebenen Analysesystemen. Der Begriff OLAP wurde 1993 eingeführt [CCS93]. Die ursprüngliche Intention von OLAP-Systemen ist die Entscheidungsunterstützung im Unternehmensumfeld.

Die Grundlage bildet ein OLAP-Würfel (engl. Cube), meist ein Stern- oder Schneeflockenschema, der mit Daten aus operativen oder Datawarehouse-Systemen gespeist wird. Der Würfel besteht aus Indikatoren (=Fakten), Dimensionen (=Kriterien zur Zusammenfassung und Navigation von Indikatoren) und Kategorien (=Wertebereiche von Dimensionen).

OLAP ist abfragezentriert und daher auch für gelegentliche Benutzer geeignet. Die mit solchen Systemen möglichen Analysen zeichnen sich durch eine einfache Komplexität

aus. Aufwendige Berechnungen werden vordefiniert. Neben importierten operativen Daten werden auch abgeleitete, z. B. aggregierte, verwendet.

Typische Anwendungsklassen von OLAP-Systemen sind z. B.

- Visualisierung von Daten,
- Drill down (Detaillierung entlang einer Dimension),
- Drill up (Zusammenfassung entlang einer Dimension),
- Filtern nach bestimmten Kriterien,
- Slicing (Auswahl von Teilmengen der verfügbaren Dimensionen),
- Dicing (Abwahl von Teilmengen der verfügbaren Dimensionen) oder
- Pivoting (Wechseln von Spalten und Zeilen).

In [CCS93] wurden 12 Regeln zur Beschreibung des OLAP-Konzepts vorgestellt. In der Folgezeit fand eine Erweiterung auf 18 statt. Sie wurden kontrovers diskutiert. In [PC95] wurde unter dem Akronym FASMI (Fast Analysis of Shared Multidimensional Information) eine weitere Definition vorgeschlagen, die nur aus 5 Regeln besteht und bis heute Anwendung findet. Diese Regeln lauten:

1. **Fast:**

   Die Antwortzeiten von OLAP-Systemen sollten bei einfachen Anfragen unter einer Sekunde und bei komplexeren unter 30 Sekunden liegen.

2. **Analysis:**

   Das OLAP-System muss in der Lage sein, jegliche Geschäftslogik und Statistik, die für den Anwender relevant ist, zu erfüllen; Anfragen müssen vom Nutzer einfach und schnell zu spezifizieren sein.

3. **Shared:**

   Das System muss einen Mehrbenutzerbetrieb ermöglichen und mit einem entsprechenden Rechtemanagement und Zugriffsschutz ausgestattet sein.

4. **Multidimensional:**

   Das OLAP-System muss eine multidimensionale Sicht auf die zu analysierenden Daten ermöglichen. Die Dimensionen sollen aus verschiedenen Hierarchien bestehen können.

5. **Information:**

   Dem Nutzer müssen alle Daten zur Verfügung stehen. Durch Beschränkungen (z. B. technische) des Analysesystems dürfen keine Daten ausgeblendet sein.

Diese Anforderungen deuten schon darauf hin, dass OLAP für bioinformatische Anforderungen nur bedingt geeignet ist. Analysen pflanzenbiologischer Daten sind oftmals sehr rechenaufwendig, so dass keine Antwortzeiten im Sekundenbereich möglich sind. Dies würde der Definition von OLAP widersprechen. Diesem Problem könnte zwar durch die (teilweise) Vorberechnung begegnet werden, jedoch wird dabei ein weiterer Nachteil sichtbar. Um Vorberechnungen durchzuführen bzw. OLAP generell sinnvoll einsetzen zu können, muss im Vorfeld vollständig spezifiziert werden, welche Analysen auf welchen Daten ausgeführt werden sollen. Dies mag für Zwecke der Entscheidungsunterstützung in einem Wirtschaftsunternehmen möglich sein, jedoch ist zweifelhaft, ob dies in der pflanzenbiologischen Forschung mit häufig wechselnden Fragestellungen sinnvoll ist. Vielmehr besteht die Gefahr, dass nach Spezifikation und Implementierung die Fragestellung bereits überholt ist. Sinnvoll erscheint OLAP dann, wenn (abgesehen vom Finden einer adäquaten Fragestellung) große Mengen von Daten anfallen (z. B. bei Hochdurchsatzmethoden) und die OLAP-Analysen oft ausgeführt werden. Ein Beispiel hierfür ist in [KDR04] beschrieben.

3.2.3 Knowledge Discovery in Databases (KDD)

Der Begriff Knowledge Discovery in Databases (KDD) wurde durch den ersten KDD-Workshop 1989 [PS91] geprägt. KDD wird in [FPSS96a] als nicht-trivialer Prozess der Identifizierung gültiger, neuer und potenziell nützlicher Muster in Daten definiert. Um von Daten zu Wissen zu kommen, bedarf es einer Abfolge von Teilschritten. [BA96] beschreibt 9 Teilschritte, welche in der Literatur vielfach zu 5 Schritten zusammengefasst werden:

1. **Selektion**:

   Entwicklung eines Verständnisses für den Anwendungsbereich, Identifikation der Ziele des KDD-Prozesses sowie Auswahl der Zieldaten aus den Rohdaten.

2. **Vorverarbeitung**:

   Anwendung von einfachen Datenbereinigungsschritten auf die Zieldaten und Entwicklung von Vorgehensweisen zur Anwendung bei fehlenden Daten.

3. **Transformation**:

   Auswahl der für die Zielstellung relevanten Merkmale aus den Zieldaten (feature selection), Reduktion der Dimension der Zieldaten.

4. **Datamining**:

   Hauptbestandteil des KDD-Prozesses. Zur Entdeckung neuer Informationen in großen Datenmengen werden in der Literatur verschiedene Datamining-Aufgaben unterschieden (vgl. z. B. [FPSS96b, Lus02, AZ98]). Zu jeder dieser Aufgaben existiert eine Vielzahl von Methoden. Eine oder mehrere dieser Methoden werden auf die Zieldaten angewandt.

5. **Interpretation/Evaluation**:

   Die entdeckten Muster werden auf verschiedene Arten (z. B. mit Hilfe weiterer Software oder manuell) ausgewertet und interpretiert. Es entsteht neues Wissen.

In den beiden folgenden Abschnitten 3.2.4 und 3.2.5 werden gängige Verfahren zur Vorverarbeitung und Transformation von Daten vorgestellt. Im Anschluss wird in Abschnitt 3.2.6 eine Auswahl gängiger Datamining-Methoden beschrieben. Die ausgewählten Verfahren werden hinsichtlich ihrer Verwendbarkeit bei pflanzenbiologischen Daten diskutiert.

3.2.4 Vorverarbeitung von Rohdaten

Um Datamining-Methoden anwenden zu können, sind eine Reihe von vorgelagerten Schritten erforderlich. Insbesondere die Behandlung von Fehlwerten und Ausreißern führt zu einer Erhöhung der Qualität der zu analysierenden Daten. Potenzielle Verschiebungen in den Ergebnissen können so vermieden oder zumindest verringert werden.

Dies ist im Bereich der pflanzenbiologischen Forschung besonders bedeutsam, da pflanzenbiologische Daten oftmals sehr lückenhaft sind (vgl. Kapitel 4).

Behandlung von fehlenden Werten

Datamining-Algorithmen gehen unterschiedlich mit fehlenden Werten um. Bestimmte Algorithmen, wie sie in Abschnitt 3.2.6 vorgestellt werden, setzen die Behandlung solcher Werte voraus (z. B. k-Means-Clustering), andere Methoden interpretieren sie als fehlende Datenpunkte, die durch den Algorithmus selbst behandelt werden (z. B. Entscheidungsbäume).

Gängige Methoden zur Behandlung von fehlenden Werten sind nach [Sch91]:

- **Eliminierungsverfahren:**

  Entfernung unvollständiger Objekte ganz oder teilweise aus dem zu analysierenden Datensatz. Hierbei können Merkmale, die nicht bei allen Records Werte aufweisen, entfernt werden oder auch einzelne Records, die nicht für alle Merkmale Ausprägungen haben. Dies kann allerdings zu einer Verzerrung führen, so dass der Datensatz u. U. nicht mehr repräsentativ ist.

- **Imputationsverfahren:**

  Ersetzung mit Nicht-Nullwerten. Dazu wird beispielsweise bei kardinalen (metrischen) Werten das arithmetische Mittel und bei klassierten Werten der Modalwert benutzt. Gängig sind auch Median, Minimum, Maximum oder die Ersetzung mit zufällig aus vorhandenen Daten gewählten Werten. Von [Rub78] wurde das Verfahren der mehrfachen Ersetzung (Multiple Imputation) vorgeschlagen. Es sieht vor, für ein fehlendes Datum verschiedene Werte zu ergänzen, um die durch die einfache Ersetzung dieses Datenpunktes hervorgerufene Unsicherheit bezüglich dieses Wertes abzufangen.

- **Schätzverfahren:**

  Schätzung fehlender Daten über Heuristiken. Hierbei kommt häufig der EM-Algorithmus [DLR77] zum Einsatz.

Behandlung von Ausreißern

Als Ausreißer werden Datenpunkte bezeichnet, die weit außerhalb des Streuungsbereiches des Erwartungswertes des jeweiligen Datensets liegen. Ausreißer können auf die Ergebnisse verschiedener Algorithmen signifikante Einflüsse haben (z. B. bei Cluster-Methoden, Bayes'schen Netzen oder Support Vector Machines).

- **Winsorising (Umkodierung):**

  Der Begriff des Winsorisings wurde erstmals 1960 verwendet [Dix60]. Er geht auf Charles P. Winsor zurück, der vorgeschlagen hatte, anstelle von extremen Ausreißern, schlechten oder unbekannten Werten den nächsthöheren respektive nächstniedrigeren beobachteten Wert zu verwenden. Beispielsweise könnten für eine 90-prozentige Winsorisation die unteren 5% der Werte auf das Minimum des 5. Perzentils und die oberen 5% auf das Maximum des 95. Perzentils gesetzt werden.

- **Trimming (Clipping, Entfernung):**

  Eine andere Vorgehensweise ist das Trimming. Dabei werden Werte außerhalb des Streuungsbereiches des Erwartungswertes entfernt und in den Folgeschritten ignoriert [Tuk62].

3.2.5 Transformation von Rohdaten

Im Anschluss an die Vorverarbeitung werden die Daten für die jeweiligen Algorithmen vorbereitet. Welche Schritte genau erfolgen, hängt vom zu verwendenden Algorithmus ab. Im folgenden sollen einige Verfahren zur Reduktion der Dimension von Daten mit dem Ziel der einfacheren Auswertbarkeit vorgestellt werden.

Einteilung in Klassen (Diskretisierung/Binning)

Hierunter wird ein Mapping von (metrisch messbaren) Werten auf ordinale Werte auf Basis von Klassen oder Bereichen von Werten (Bins) verstanden. Dabei werden verwandte Werte gruppiert, um die Anzahl der unterschiedlichen Ausprägungen eines Attributes zu verringern. Der Nutzen besteht in der Erstellung von kompakteren Modellen, die sich leichter berechnen lassen.

Es gibt verschiedene Möglichkeiten für die Umsetzung einer Diskretisierung, z. B. Bins mit gleicher Breite.

Beispiel 3.1 Diskretisierung numerischer Werte:

1 – 20	→	1
21 – 40	→	2
41 – 60	→	3
61 – 80	→	4
81 – 100	→	5
[...]		[...]

Beispiel 3.2 Diskretisierung klassierter Werte:

Hordeum, Triticum, ...	→	Graminae
Lycopersicon, Nicotiana, ...	→	Solanacae
Achillea, Matricaria, ...	→	Asteraceae
Euphorbiaceae, Bromeliaceae, ...	→	Cactacea
[...]		[...]

Normalisierung

Allgemein wird unter Normalisierung eine Abbildung auf einen einheitlichen Wertebereich mit dem Ziel der Vergleichbarmachung von Datensätzen verstanden. Im Falle numerischer Werte kann dies eine Transformation in einen vorgegebenen Wertebereich, z. B. in das Intervall von 0 bis 1, sein.

> **Beispiel 3.3** Normalisierung von Messwerten auf das Intervall $[0, 1]$ mit der Minimum-Maximum-Normalisierung. Wenn x_{min} der kleinste und x_{max} der größte Wert einer Menge X von numerischen Werten ist, so gilt für den auf das Intervall $[0, 1]$ normalisierten Wert x_i: $x'_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$
>
> | | | |
> |---|---|---|
> | 0 | $\rightarrow$ | 0 |
> | 34 | $\rightarrow$ | 0,283 |
> | 75 | $\rightarrow$ | 0,625 |
> | 120 | $\rightarrow$ | 1 |

3.2.6 Datamining

Datamining (Teilprozess des Knowledge Discovery in Databases (KDD) [FPSM92, PSMSU94]) ist ein Prozess, der durch die Anwendung von Methoden auf einen Datenbestand Regeln, Muster oder statistische Auffälligkeiten entdeckt. Die Interpretation der entdeckten Muster ist nicht Bestandteil des Dataminings. Beim Datamining steht im Kontext des KDDs mehr der beschreibende Charakter im Vordergrund, weniger der vorhersagende [FPSS96b]. Datamining wird üblicherweise dann eingesetzt, wenn die Fragestellung nicht genau geklärt ist.

Zur Entdeckung neuer Informationen in großen Datenmengen werden in der Literatur verschiedene Datamining-Aufgaben unterschieden (vgl. z. B. [FPSS96b, Lus02, AZ98]).

Für diese Arbeit wurde eine Einteilung in die folgenden drei Gruppen gewählt:

- Klassifikation,
- Segmentierung und
- Entdeckung von Abhängigkeiten.

Zu jeder dieser Aufgaben existiert eine Vielzahl von Verfahren. Eine oder mehrere dieser Methoden werden auf die Zieldaten angewandt. In den folgenden Ausführungen soll zu jeder dieser Gruppen eine Auswahl verbreiteter Verfahren vorgestellt werden. Diese Präsentation erfolgt nur überblicksweise, um dem Leser ein Verständnis für die Thematik zu vermitteln.

Klassifikation

Das Ziel der Klassifikation besteht in der Zuordnung (Klassifizierung) von Daten in eine Klasse aus einer Menge vordefinierter Klassen. Jedem Datum oder Objekt wird ein Vektor (Featureset) zugeordnet, dessen Dimensionen Eigenschaften des Objekts beschreiben. Damit Objekte klassifiziert werden können, müssen Klassifikatoren trainiert werden. Dazu ist ein bereits zugeordneter Trainingsdatensatz erforderlich. Deswegen werden Klassifikationsverfahren auch als überwacht (supervised) bezeichnet. Je besser ein Klassifikator trainiert ist, desto besser können Datenobjekte klassifiziert werden. Die einzelnen Dimensionen des Featuresets sollen dabei einen möglichst hohen Informationsgehalt für die Klassifikation haben und untereinander keine abhängigen Informationen enthalten. Der Prozess der Auswahl der Features wird als Featureselektion bezeichnet. Eine Auswahl verbreiteter Methoden wird im Folgenden beschrieben.

k-Nearest-Neighbour-Klassifikator: Auf der Basis von bereits eingeordneten Werten trifft der k-Nearest-Neighbour-Klassifikator eine Vorhersage darüber, zu welcher Klasse ein zuzuordnender Wert gehört [CH67]. Es wird dabei entschieden, welche k der zugeordneten Werte dem zu klassifizierenden Wert am nächsten sind. Hierbei wird von den k nächsten Nachbarn gesprochen. Das gewählte Abstandsmaß spielt für die Qualität des Ergebnisses eine entscheidende Rolle.

Entscheidungsbaum: Der Entscheidungsbaum-Algorithmus ist eine Vorhersagemethode, bei der die Zerlegung eines Datensatzes durch eine Serie von Entweder-oder-Entscheidungen erfolgt. Klassen werden hierbei durch bestimmte Attributausprägungen beschrieben. Auf dieser Basis erfolgt die Ableitung von Entscheidungsregeln, z. B.

- wenn Pflanze zur Familie der Gräser gehörig, dann Klasse A, sonst Klasse B,
- wenn Pflanze kleiner als 80cm, dann Klasse C, sonst Klasse D.

Der Datensatz wird anhand seiner Attributwerte schrittweise zerlegt. Dabei wird bei jedem Attribut entschieden, ob ein Kriterium erfüllt ist oder nicht. Diese Zerlegung erfolgt dahingehend, dass die resultierende Partitionierung sukzessive verbessert wird. Zusätzlich kann ein Abbruchkriterium festgelegt werden, z. B. eine minimale Anzahl von Elementen in einem Knoten oder eine maximale Tiefe des Entscheidungsbaumes.

Abbildung 3.7 zeigt das Zustandekommen einer Klassifikation mit einem Entscheidungsbaum.

Entscheidungsbäume unterstützen beliebige Skalen. Die zur Klassifizierung herangezogenen Attribute können dabei sogar unterschiedliche Skalen haben. Weiterhin lassen sich Entscheidungsbäume relativ einfach umsetzen. Damit sind sie gut für die Analyse pflanzenbiologischer Daten geeignet.

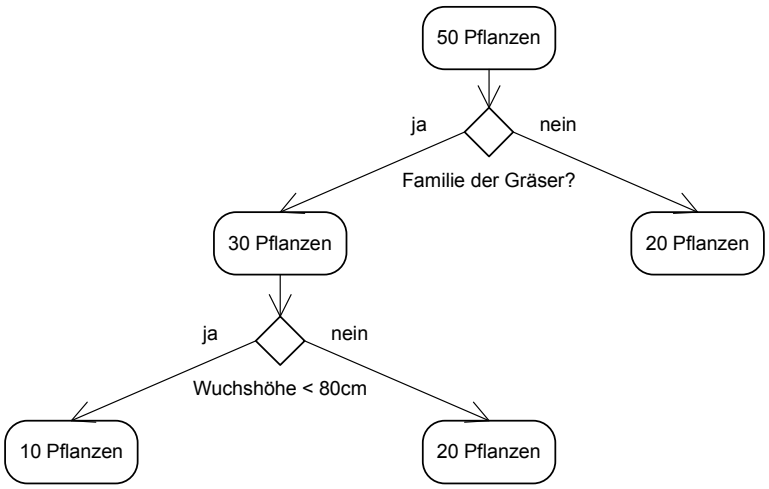


Abbildung 3.7: Klassifikation von Daten mit einem Entscheidungsbaum

Segmentierung (Clustering)

Bei Datensätzen, für die entweder keine Klassen bekannt sind oder aber kein Trainingsset zur Verfügung steht, kann keine klassifizierende Methode angewandt werden. Stattdessen wird versucht, die Daten in sog. Cluster einzuteilen. Dabei gilt, dass die Abstände der Objekte innerhalb eines Clusters untereinander möglichst klein sein sollen, die Abstände zwischen verschiedenen Clustern hingegen möglichst groß. Es existiert eine Vielzahl von Distanzmaßen, z. B. die Euklidische Distanz[3], auf die hier aber nicht näher eingegangen werden soll. Beim Clustering wird auch von unüberwachten Verfahren (unsupervised) gesprochen.

Der Begriff der Clusteranalyse wurde erstmals 1939 verwendet [Try39]. Er vereinigt eine Menge von Algorithmen/Methoden (vgl. [Har75]) zur Gruppierung von vergleichbaren Objekten in verschiedene Kategorien. Damit können Strukturen in Daten erkannt werden, ohne dass dabei die Gründe für die Strukturierungen erklärt werden.

Clustering-Methoden werden genutzt, wenn a priori keine Hypothesen vorliegen (ergebnisoffene Analyse). Im Folgenden sollen ausgewählte, verbreitete Clustering-Methoden kurz vorgestellt werden.

Hierarchisches Clustering: Hierarchisches Clustern [LW67] erfolgt entweder top-down oder bottom-up. Beim Top-down-Ansatz wird ein Basiscluster, der alle Elemente enthält, in mehreren Zyklen sukzessive immer weiter aufgeteilt. Im Gegensatz

[3]Die Euklidische Distanz für zwei Vektoren x und y mit $x = (x_1 \dots x_n)$ und $y = (y_1 \dots y_n)$ ist als $d(x,y) = |x-y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ definiert.

dazu beginnt beim Bottom-up-Ansatz der Clusterprozess mit den einzelnen Elementen, die stückweise zu immer umfangreicher werdenden Gruppen zusammengefasst werden. Das Resultat beider Ansätze ist eine Baumstruktur. Wird der resultierende Baum in einer ausgewählten Ebene geschnitten, ergibt sich eine Clusterung der Ausgangselemente in Abhängigkeit der gewählten Ebene.

Abbildung 3.8 zeigt beispielhaft ein Dendrogramm als Ergebnis des Clusterns von Gerstensorten mit Euklidischem Abstand über zwölf SSR-Marker.

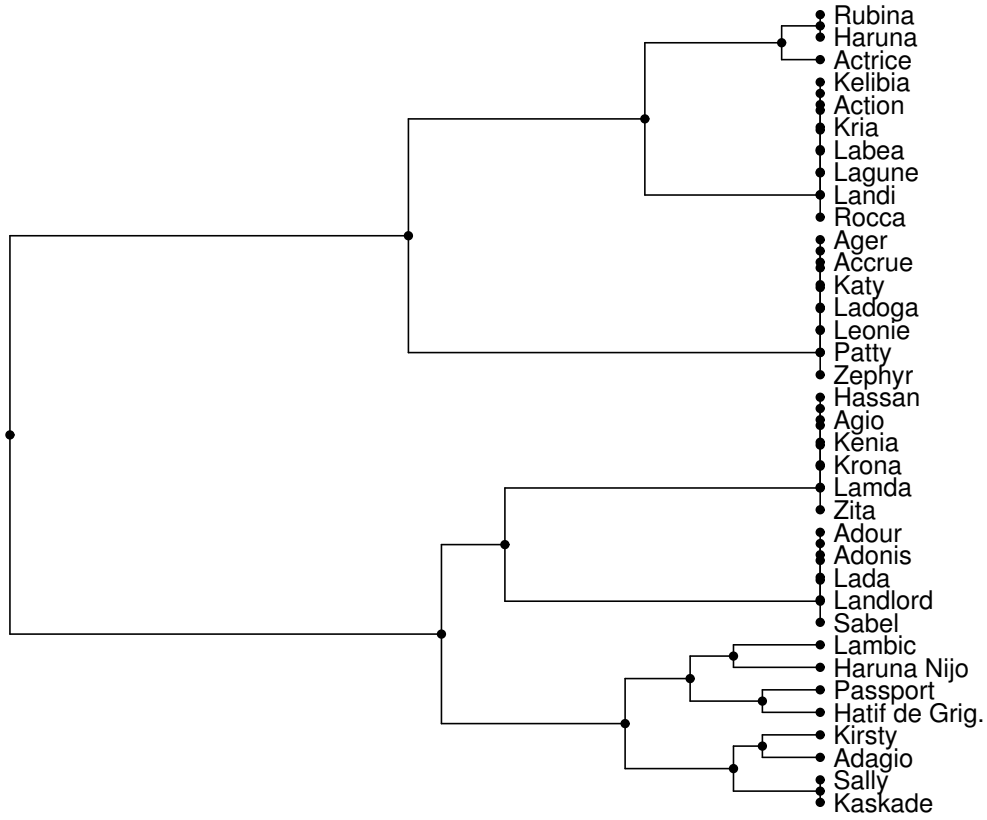


Abbildung 3.8: Dendrogramm mit Gerstensorten als Ergebnis einer hierarchischen Clusterung

k-Means: Ein weiteres, verbreitetes Clustering-Verfahren ist der k-Means-Algorithmus [McQ67, BH67]. Dieser erwartet die Angabe der Anzahl k der Zielcluster. Danach werden k Clusterzentren (Zentroiden) beliebig vergeben. Der Algorithmus ordnet jedem zu clusternden Objekt das nächstliegende Clusterzentrum zu. Für die so entstandenen Cluster werden im nächsten Schritt die Zentren neu berechnet. Unterscheiden sich diese von den ursprünglich (willkürlich) vergebenen, so werden die Objekte auf Basis der neuen Clusterzentren erneut geclustert. Dies wird solange wieder-

holt, bis sich die Zuordnung der Objekte zu den Clusterzentren nicht mehr ändert. Abbildung 3.9 stellt dieses Vorgehen schematisch dar.

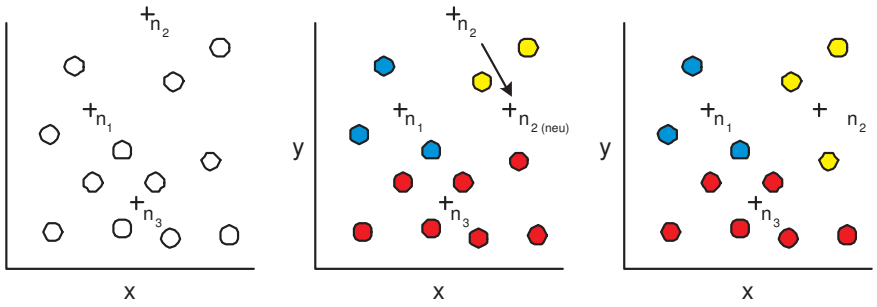


Abbildung 3.9: Schematische Darstellung einer k-Means-Clusterung in drei Schritten

In Abhängigkeit der initial vergebenen Clusterzentren können die mit dem k-Means-Algorithmus ermittelten Cluster erheblich voneinander abweichen. Im pflanzenbiologischen Bereich werden Clustering-Methoden beispielsweise zur Ermittlung der Populationsstruktur von Genotypen genutzt. Dies findet in Kapitel 7 Anwendung. Um zu einem stabilen Ergebnis zu gelangen, muss im Fall des k-Means-Verfahrens mit unterschiedlichen Initialisierungsparametern wiederholt gerechnet werden.

Entdeckung von Abhängigkeiten

In den folgenden Abschnitten soll eine Auswahl von Verfahren zur Aufdeckung von Abhängigkeiten zwischen Merkmalen kurz vorgestellt werden. Solche Verfahren sind ein Forschungsschwerpunkt in der Pflanzenbioinformatik, um Ursache-Wirkungs-Beziehungen beispielsweise zwischen dem Genotyp und dem Phänotyp von Pflanzen zu ergründen. Dieses Beispiel wird im Anwendungskapitel dieser Arbeit (Kapitel 7) aufgegriffen und ausführlich erläutert.

Hauptkomponentenanalyse: Die Hauptkomponentenanalyse (Principal Components Analysis) geht auf [Pea01] zurück und wurde durch [Hot33] verbreitet. Dieses Verfahren ermittelt aus einer Menge von Variablen mit vielen Eigenschaften diejenigen Faktoren, die diese Eigenschaften determinieren.

Daten liegen in einem n-dimensionalen Koordinatensystem als Punktwolke vor. In dieses Koordinatensystem wird nun ein weiteres gelegt und rotiert. Dabei wird die erste Achse derart durch die Punktwolke gelegt, dass die Varianz der Datenpunkte in der Richtung, die die Achse anzeigt, maximal ist. Danach wird eine zweite Achse orthogonal zur ersten so durch die Punktwolke gelegt, dass die Varianz am zweitgrößten ist. Dies wird für alle n Dimensionen wiederholt. Die Varianzen der einzelnen Achsen addieren sich zur Gesamtvarianz. Nun wird betrachtet, welche Faktoren (repräsentiert

durch die Achsen) zusammen den größten Prozentsatz der Gesamtvarianz abdecken. Dies sind die Hauptkomponenten.

Korrelationskoeffizient: In der Statistik wird von einer Korrelation gesprochen, wenn zwei oder mehr Variablen in funktionaler Beziehung zueinander stehen. Hierzu kann ein so genannter Korrelationskoeffizient r mit $r \in [-1, +1]$ berechnet werden. Wenn r gegen $+1$ geht, wird von einer positiven Korrelation zwischen zwei Variablen A und B gesprochen, d. h. wenn der Wert der Variable A größer wird, trifft dies auch auf B zu. Geht r gegen -1, liegt eine negative oder inverse Korrelation vor, d. h. wenn A größer wird, wird B kleiner. Geht r gegen 0, liegt keine Korrelation vor.

Wird eine Korrelation festgestellt, sollte diese im zweiten Schritt auf Signifikanz geprüft werden. Dies ist insbesondere wichtig, wenn der zugrunde liegende Testdatensatz nicht sehr umfangreich ist. Der Signifikanzlevel sagt aus, wie hoch die Wahrscheinlichkeit für eine zufällige Korrelation auf Basis dieses Testdatensatzes ist. Ob eine Korrelation signifikant ist, kann im Rahmen eines Korrelationstestes auf der Basis der t-Verteilung [Stu08] überprüft werden.

Korrelationen können bei metrisch oder ordinal messbaren Merkmalen [Sch94] berechnet werden. Auf nominale Daten kann diese Methode nicht angewandt werden.

Es existieren verschiedene Typen von Korrelationskoeffizienten, z. B. Pearson und Spearman. Der Pearson-Korrelationskoeffizient [Pea96] ist ein dimensionsloses Maß für den Grad des linearen Zusammenhangs. Dabei wird von einer annähernden Normalverteilung ausgegangen. Für zwei metrisch messbare Merkmale A und B gilt:

$$r_{AB} = \frac{COV(A,B)}{\sqrt{VAR(A)}\sqrt{VAR(B)}}$$

Sind die Variablen nicht normalverteilt oder handelt es sich um ordinale Werte, kann der Korrelationskoeffizient nach Spearman verwendet werden. Dieser wird auch als Spearman-Rangkorrelationskoeffizient oder Spearmans Rho [Spe04] bezeichnet. Hierfür werden die Merkmalsausprägungen nach Größe sortiert, danach wird ihnen eine Rangzahl zugewiesen. Mit Hilfe dieser Rangzahlen wird der Pearson-Korrelationskoeffizient berechnet. Ein weiterer Rangkorrelationskoeffizient ist als Kendalls Tau [Ken38] bekannt.

Ein Beispiel sei das Merkmal *Höhe einer Pflanze*, das auf einer Bewertungsskala von 1 bis 3 mit 1=klein (5-20cm), 2=mittel (21-70cm) und 3=groß (71-80cm) angegeben wird. Würde mit diesen Werten der Pearson-Korrelationskoeffizient berechnet, müsste davon ausgegangen werden, dass die Abstände zwischen *klein* und *mittel* sowie zwischen *mittel* und *groß* identisch sind, was aber nicht der Fall ist. Deswegen wird hier ein Rangkorrelationskoeffizient berechnet. Das folgende Beispiel zeigt, wie in diesem Fall Rangzahlen zugewiesen werden.

Beispiel 3.4 Zuweisung von Rangzahlen für ein ordinal skaliertes Merkmal:

Pflanze	Höhe (ordinal)	zugewiesene Reihenfolge	Rangzahl
Pflanze #2	1	1	1,5 $(\frac{1+2}{2})$
Pflanze #1	1	2	1,5
Pflanze #7	2	3	4,5 $(\frac{3+4+5+6}{4})$
Pflanze #3	2	4	4,5
Pflanze #6	2	5	4,5
Pflanze #4	2	6	4,5
Pflanze #5	3	7	7

Assoziationskoeffizient: Bei nominal messbaren Merkmalen kann ein Assoziationskoeffizient A mit $A \in [-1, +1]$ berechnet werden [Sch94]. Es wird von einer Assoziation [Yul00] gesprochen, wenn zwei oder mehr Variablen in Beziehung zueinander stehen.

Für zwei nominal messbare Merkmale X und Y sei $h(\bar{y})$ der häufigste Wert von Y, $h(\bar{y}|x_i)$ der häufigste durch x_i bedingte Wert von Y und n die Gesamtanzahl der Werte. Für den Assoziationskoeffizienten A_{YX} für die Abhängigkeit des Merkmals Y vom Merkmal X gilt dann:

$$A_{YX} = \frac{\sum_{i=1}^{m} h(\bar{y}|x_i) - h(\bar{y})}{n - h(\bar{y})}$$

Nachfolgend wird die Berechnung eines Assoziationskoeffizienten am Beispiel der Merkmale *Gattung* (X) sowie *Ährenfarbe* (Y) verdeutlicht.

Beispiel 3.5 Berechnung des Assoziationskoeffizienten A_{YX} auf Basis der häufigsten Werte der Merkmale X und Y:

Ährenfarbe	Hordeum (x_1)	Triticum (x_2)	Secale (x_3)	Anzahl
grün (y_1)	10	10	30	50
hellbraun (y_2)	30	**50**	**40**	**120**
mittelbraun (y_3)	**40**	30	20	90
dunkelbraun (y_4)	5	20	10	35
Anzahl	85	110	100	**295**

$$A_{YX} = \frac{40 + 50 + 40 - 120}{295 - 120} \approx 0,057$$

Regressionsanalyse: Mit Hilfe der Regressionsanalyse wird versucht, die Tendenz eines Zusammenhangs zwischen einer Variable Y und einer oder mehrerer Variablen $X_1 \dots X_n$ zu beschreiben, wobei Y von $X_1 \dots X_n$ statistisch abhängig ist [Sch94]. Ziel der Regressionsanalyse ist die Bestimmung einer Regressionsfunktion, die die tatsächlich beobachteten Werte möglichst gut abbildet.

Das allgemeine mathematische Modell lautet:

$$Y = f(X_1 \dots X_n, \beta) + \epsilon$$

Hierbei ist β ein Vektor unbekannter Parameter, die mit Hilfe der Regression bestimmt werden sollen, und ϵ ein zufälliger Fehler.

Gibt es einen linearen Zusammenhang zwischen X und Y, kann das Regressionsmodell berechnet werden. Besteht kein linearer Zusammenhang, wird es näherungsweise gelöst. Auf die verschiedenen Verfahren zur Regressionsanalyse soll an dieser Stelle nicht näher eingegangen werden.

Sinnvoll ist der Einsatz der Regressionsanalyse nur bei metrisch messbaren Merkmalen [Sch94]. Damit ist diese Analyseform für eine Vielzahl pflanzenbiologischer Daten nicht einsetzbar. Hierunter fallen z. B. Evaluierungsdaten, die häufig mit einer ordinalen Skala erhoben werden, oder auch SNP-Markerdaten, denen eine nominale Skala zugrunde liegt.

Der Begriff der Regression geht auf Francis Galton zurück [Gal85, Gal86].

Varianzanalyse (ANOVA): Abschließend soll noch ein Verfahren vorgestellt werden, das häufig im pflanzenbiologischen Bereich Anwendung findet – die Varianzanalyse.

Unter dem Begriff der Varianzanalyse (Analysis of variance, abgek. ANOVA) werden statistische Verfahren zusammengefasst, die mit Hilfe der Varianz versuchen, Gesetzmäßigkeiten in Daten zu entdecken. Die Varianzanalyse geht auf den Statistiker und Genetiker Ronald A. Fisher [Fis18] zurück und nutzt die F-Verteilung (auch Fisher-Snedecor-Verteilung) [Sne34] zur Überprüfung auf statistische Signifikanz.

Der Grundgedanke der Varianzanalyse besteht darin, zu versuchen, die Varianz einer abhängigen, metrisch messbaren Zufallsvariable über den Einfluss einer oder mehrerer unabhängiger, kategorialer Gruppenvariablen (Faktoren) zu erklären. Existiert eine abhängige Variable, ist die Varianzanalyse univariat, bei mehreren multivariat. Erfolgt die Gruppeneinteilung der zu untersuchenden Objekte hinsichtlich eines Merkmals, so wird von einfaktorieller Varianzanalyse gesprochen, erfolgt sie hinsichtlich mehrerer Merkmale, von mehrfaktorieller. Die Kombination dieser Unterscheidungsmerkmale führt zu vier Arten der Varianzanalyse:

- der einfaktoriellen univariaten Varianzanalyse,
- der mehrfaktoriellen univariaten Varianzanalyse,
- der einfaktoriellen multivariaten Varianzanalyse sowie
- der mehrfaktoriellen multivariaten Varianzanalyse.

Die Merkmalswerte der abhängigen Variablen werden in Gruppen zerlegt. Dies geschieht auf der Basis der Ausprägungen der unabhängigen Variablen. Danach werden Variationen der Mittelwerte zwischen den so entstandenen Gruppen sowie innerhalb dieser Gruppen betrachtet.

Die Varianz zwischen den Gruppen VAR_{zw} errechnet sich durch die Abweichung der Gruppenmittelwerte $\bar{x}_i$ vom Gesamtmittelwert $\bar{x}$. Wenn m die Anzahl der Gruppen und n_i die Anzahl der Werte pro Gruppe ist, so gilt:

$$VAR_{zw} = \frac{1}{m-1} \sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})^2$$

Die Varianz innerhalb einer Gruppe VAR_{in} wird durch die Abweichungen der Ausprägungen dieser Gruppe x_{ij} vom Gruppenmittelwert $\bar{x}_i$ berechnet. n sei die Gesamtanzahl aller Beobachtungen. Es gilt:

$$VAR_{in} = \frac{1}{n-m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Ist die Varianz zwischen den Gruppen VAR_{zw} größer als die innerhalb der Gruppen VAR_{in}, deutet dies darauf hin, dass ein Zusammenhang zwischen den Ausprägungen der unabhängigen Variable und denen der abhängigen Variable besteht. Dies wird über einen F-Test (Varianzquotiententest) mit

$$F = \frac{VAR_{zw}}{VAR_{in}}$$

auf Signifikanz überprüft.

Es existiert eine Anzahl von Variationen der Varianzanalyse unter Bezeichnungen wie ANCOVA (Analysis of Covariance), MANOVA (Multivariate Analysis of Variance) etc.

Ein Beispiel für den Einsatz dieser Technik in der Pflanzenbiologie ist die Überprüfung auf mögliche Zusammenhänge zwischen Sorten (unabhängige Variable, nominal messbar) und den Ausprägungen eines agronomischen Merkmals (abhängige Variable, metrisch messbar) im Rahmen von Feldversuchen. Die Aufdeckung solcher Zusammenhänge spielt in der Pflanzenzüchtung eine sehr wichtige Rolle.

3.3 Resümee

Im ersten Teil dieses Kapitels wurden mit der virtuellen und der materialisierten Integration zwei Vorgehensweisen präsentiert, die zur Integration pflanzenbiologischer Daten verwendet werden können.

Die virtuelle Integration findet zum Anfragezeitpunkt eines Systems statt. Anfragen werden an verschiedene Datenquellen geleitet, die Resultate werden kombiniert. Hierbei werden durch das Integrationssystem keine Daten lokal gespeichert. Dieses Vorgehen ist bei der Integration von Internet-Datenquellen weit verbreitet. Zur Verdeutlichung wurden die Ansätze der Multidatenbanken und der Mediatoren beschrieben.

Im Gegensatz dazu erfolgt bei der materialisierten Integration eine lokale Speicherung von Daten. Anfragen werden an diesen Datenbestand gerichtet. Als Beispiel für diesen Ansatz wurden Datawarehouse-Systeme erläutert.

Die Verwendbarkeit der beschriebenen Ansätze im Rahmen dieser Arbeit wird in Kapitel 6 diskutiert.

Im zweiten Teil dieses Kapitels wurden drei mögliche Vorgehensweisen zur Analyse pflanzenbiologischer Daten, Datenbanksprachen, OnLine Analytical Processing und Datamining, vorgestellt und in Bezug auf ihre Einsetzbarkeit in der Pflanzenbioinformatik diskutiert. Es wurde festgestellt, dass Datenbanksprachen und OnLine Analytical Processing für experimentelle Anwender nicht oder nur unter besonderen Bedingungen geeignet sind und dass das Hauptaugenmerk im Rahmen dieser Arbeit auf den Komplex des Dataminings gelegt werden sollte.

Im Anschluss wurden verschiedene Verfahren zur Datenanalyse überblicksweise vorgestellt und mit Beispielen aus der Pflanzenbioinformatik untersetzt.

4 Datenqualität in der Pflanzenbioinformatik

Schätzungen von Analysten gehen davon aus, dass allein für US-Unternehmen jährliche Kosten in Höhe von über 600 Milliarden US-Dollar durch schlechte Datenqualität verursacht werden [Eck02]. Dieses Problem beschränkt sich jedoch nicht nur auf den privatwirtschaftlichen Bereich mit seinen Kunden- und Produktdaten, sondern lässt sich auf viele weitere Gebiete, wie das der Lebenswissenschaften, ausweiten.

Wissenschaftler im Bereich der Lebenswissenschaften im Allgemeinen sehen sich häufig mit einer problematischen Datenqualität konfrontiert, die in der Literatur mehrfach diskutiert wurde (vgl. [PK06, MWBL05, RHM04, MNF03, BB96]). Dies gilt auch für pflanzenbiologische Daten im Speziellen (vgl. [Vor02, HMF02]). Die Gründe hierfür sind vielschichtig und sollen in diesem Kapitel erörtert werden.

Zur Durchführung aussagekräftiger Analysen sind qualitativ hochwertige Daten unerlässlich. Insbesondere bei der Integration (siehe Kapitel 3) als notwendiger Voraussetzung für eine domänenübergreifende Datenanalyse wirken sich syntaktische und semantische Heterogenitäten pflanzenbiologischer Daten aber oftmals sehr negativ aus.

Die Syntax betrifft die Struktur oder das Schema von Daten. Von syntaktischer Heterogenität wird gesprochen, wenn gleiche Sachverhalte unterschiedlich modelliert werden oder es strukturelle Unterschiede in der Darstellung gibt. Dies ist beispielsweise der Fall, wenn verschiedene Datenmodelle (z. B. relationales oder hierarchisches Datenmodell, vgl. Abschnitt 2.1.1) Verwendung finden, beim Einsatz unterschiedlicher Modellierungskonzepte oder beim Auftreten von Konflikten auf Datenebene, die aus der Nutzung uneinheitlicher Datentypen, Wertebereiche etc. resultieren. Ebenfalls dazu gehören die Verwendung verschiedener Tabellen-/Attributnamen, abweichende Integritätsbedingungen oder die Nutzung unterschiedlicher Defaultwerte, aber auch län-

dertypische Zeichenkodierungen oder Trennzeichen. Beispielsweise werden im Deutschen Dezimalstellen durch ein Komma getrennt, im Englischen durch einen Punkt.

Die Semantik bezieht sich auf die Bedeutung bzw. den Inhalt von Daten. Von semantischer Heterogenität wird bei unterschiedlicher Bedeutung/Interpretation von Daten gesprochen [KS91]. Dazu zählen Synonyme und Homonyme, die Verwendung unterschiedlicher Einheiten für dasselbe Merkmal, aber auch das Vorhandensein von Duplikaten in Daten, von veralteten oder falschen Daten.

Um Qualitätsprobleme pflanzenbiologischer Daten und ihre Ursachen betrachten zu können, wird die folgende Klassifikation vorgeschlagen:

- Informationstechnische Ursachen,
- Ursachen während der Datengewinnung,
- Konzeptionelle Ursachen und
- Biologisch bedingte Ursachen.

Die folgenden Abschnitte sollen diese vier Kategorien anhand von Beispielen illustrieren. Anschließend werden Vorschläge zur Verbesserung der Datenqualität gemacht.

4.1 Informationstechnische Ursachen für Qualitätsprobleme

4.1.1 Software

In der Bioinformatik ist in den letzten zwei Jahrzehnten eine Vielzahl von Algorithmen und Applikationen entstanden, die weit verbreitet sind. Sie arbeiten mehrheitlich mit unterschiedlichen Ein- und Ausgabeformaten, Schnittstellen und Modellen und sind überwiegend dateibasiert. Standardformate wie SBML [HFS+03] werden noch zu selten eingesetzt. Es werden vielfach keine einheitlichen Vokabulare für verwendete Merkmale genutzt. Häufig werden auch für ein Merkmal unterschiedliche Wertebereiche verwendet oder es bestehen Inkonsistenzen innerhalb der Daten (siehe auch Abschnitt 4.2).

Dies erschwert sowohl die Werkzeug- als auch die Ergebnisdatenintegration in erheblichem Maße. Werkzeuge müssen entweder reimplementiert oder es muss eine zusätzliche Homogenisierungsschicht verwendet werden [BLSS04].

4.1.2 Weiterverbreitung von Daten

Zwei weitere Probleme sind durch die verschiedenen molekularbiologischen Integrationsansätze bedingt. Einerseits können beim Integrationsprozess Daten verloren gehen, weil sie beispielsweise zum Zeitpunkt der Integration vom Integrierenden nicht für wichtig erachtet werden. Andererseits können sich über solche Integrationsprozesse Fehler fortpflanzen. Dies ist insbesondere dann gefährlich, wenn die zu integrierenden Daten durch die Verwendung von Vorhersagemethoden entstanden sind, die aus dem Integrationsprozess resultierenden Daten aber als Fakt angesehen werden.

4.2 Durch die Datengewinnung bedingte Ursachen für Qualitätsprobleme

4.2.1 Rohdaten

Die Gewinnung von Rohdaten (Primärdaten) wird vielfach nicht ausreichend dokumentiert. Insbesondere im molekularbiologischen Bereich werden Zusatzinformationen wie der verwendete Genotyp, das Entwicklungsstadium, Gewebe und Behandlung häufig nicht oder nur unstrukturiert als Fließtext in Bemerkungsfeldern abgelegt. Genau diese Zusatzinformationen haben jedoch im pflanzenbiologischen Bereich einen hohen Einfluss (vgl. Abschnitt 4.4).

Neben den durch die Dokumentation bedingten Problemen gibt es auch methodenbedingte. So können bei der Sequenzierung von DNS nicht immer alle Nukleotide eindeutig identifiziert werden und es kommt zu Sequenzierfehlern [CW92, HHM+07]. Dies kann seine Ursache sowohl in der eingesetzten Hardware (Sequenzierautomaten) als auch in der verwendeten Auswertungssoftware haben. Werden solche Fehler nicht erkannt und (manuell) bereinigt, können sie sich von der Gewinnung der Daten bis hin zu den Analysen fortpflanzen.

4.2.2 Abgeleitete Daten

Die Gewinnung von abgeleiteten Daten (Sekundärdaten) als Ergebnis von Analysen der Rohdaten und/oder weiteren Sekundärdaten wird ebenfalls häufig schlecht oder gar nicht dokumentiert. Insbesondere fehlen oftmals Aussagen über die Verlässlichkeit der abgeleiteten Daten bzw. diese Aussagen sind nur umgangssprachlich formuliert, was wiederum Probleme nach sich zieht. Die Verwendung einer einheitlichen Bewertungsskala, ähnlich der Schulnoten, würde die Vergleichbarkeit abgeleiteter Daten verbessern.

Weiterhin kommt es häufig zum inkorrekten Einsatz statistischer Verfahren [GBA04]. So wird teilweise versucht, Boniturdaten mit Hilfe des einfachen arithmetischen Mittels zu verrechnen. Da es sich dabei aber um ordinal messbare Werte handelt, ist ein arithmetisches Mittel per definitionem nicht erlaubt.

4.2.3 Zeitlich begrenzte Projekte

Daten werden vielfach im Rahmen zeitlich begrenzter Projekte gewonnen. Forschungsergebnisse sollen schnellstmöglich publiziert werden. Die Daten selbst werden aber oft nicht persistent und zentral gespeichert. Die dezentrale Speicherung in Spreadsheet-Dateien mit diversen Formatierungen wird häufig als einfache Lösung bevorzugt. Dies führt zu einem erheblichen zeitlichen und personellen Mehraufwand für die syntaktische Datenaufbereitung vor einer Integration. In vielen Fällen wird dieses Problem durch unzureichende Dokumentation noch verschärft.

4.2.4 Manuelle Erfassung von Daten

Bedingt durch den Faktor Mensch liegt im manuellen Erfassen von Informationen ein großes Potenzial für fehlerhafte Daten. Dies umfasst auch das Kopieren von Daten bzw. Teilmengen von Daten zwischen verschiedenen Spreadsheet-Dateien o. ä. (vgl. Abschnitt 4.2.3).

4.3 Konzeptionelle Ursachen für Qualitätsprobleme

4.3.1 Bewertungssysteme

Die Erfassung von Daten erfolgt oftmals nicht einheitlich. Beispielsweise werden phänotypische Eigenschaften von Pflanzen häufig mit Hilfe so genannter Boniturskalen erfasst. Das sind Bewertungsskalen, die verschiedene Ausprägungen eines Merkmals mit diskreten Werten (Klassen) kodieren.

Die folgenden zwei Beispiele sollen die Verwendung unterschiedlicher Boniturskalen am Beispiel des phänotypischen Merkmals *Echter Mehltau (Blumeria graminis)* für die Gerste (*Hordeum vulgare*) illustrieren[1].

[1] http://pgrc.ipk-gatersleben.de/eval/hor/hor.html [Stand 2009-04-02]

Beispiel 4.1 Bonitурskala Nr. 1 für das Merkmal *Echter Mehltau*:

0 → kein Befall
1 → geringer Befall
2 → mittlerer Befall
3 → starker Befall
4 → sehr starker Befall

Beispiel 4.2 Bonitурskala Nr. 2 für das Merkmal *Echter Mehltau*:

a → anfällig
r → resistent

Diese Beispiele verdeutlichen zwei Probleme. Zum einen werden für die Merkmalsausprägungen unterschiedliche Schlüssel verwendet, obwohl es hierfür Empfehlungen gibt, z. B. [IPG94]. Zum anderen wird eine abweichende Anzahl von Ausprägungen verwendet (im ersten Beispiel 5, im zweiten 2).

Das erste Problem kann durch Metainformationen, die die jeweiligen Schlüssel aufeinander abbilden, leicht gelöst werden. Das zweite Problem hingegen ist nicht trivial. Da in der Mehrheit der Fälle keine Informationen über die Verteilung der beobachteten Werte innerhalb der Merkmalsklassen (Boniturwerte) erfasst werden, können die beiden Skalen nicht korrekt aufeinander abgebildet werden. Teilweise kann dies durch Fachwissenschaftler durchgeführt werden, die auf langjährige Erfahrungen mit solchen Daten zurückgreifen können. Allerdings kommt es dabei fast zwangsläufig zum Auftreten von Fehlern und damit auch zur Fortpflanzung dieser Fehler durch nachfolgende Analyseschritte.

Ein weiteres Problem der Bonitурskalen zeigt sich am Beispiel des Züchtungsfortschritts. Beispielsweise bekam eine im Jahr 1980 zugelassene Gerstensorte für das Merkmal *Ertrag* die Bestnote 9. Bedingt durch den Fortschritt in der Pflanzenzüchtung ist der Ertrag dieser Sorte heutzutage nicht mehr erstklassig und würde z.B. nur noch mit der Boniturnote 7 bewertet. Daten dieser Sorte sind somit nicht mit aktuell erhobenen Daten anderer Sorten vergleichbar. Dadurch, dass Boniturnoten Bereiche von Ausprägungen zusammenfassen, ist im nachhinein auch keine Aktualisierung der alten Daten möglich.

4.3.2 Informationssysteme

Verschiedene Informationssysteme erlauben nur den Autoren eines Datensatzes, diesen zu ändern. Dadurch wird eine Vielzahl von falschen Daten nicht aktualisiert bzw. es werden weitergehende Erkenntnisse nicht eingepflegt [BB96].

Viele Informationssysteme arbeiten ohne oder nur mit linearer Versionierung von Daten. Dies ist in zweierlei Hinsicht problematisch. Zum einen sind Erkenntnisse in der

biologischen Forschung häufig in der Fachwelt umstritten, so dass es sinnvoll wäre, verschiedene, parallele Versionen desselben biologischen Sachverhalts verfügbar zu haben. Zum anderen wird häufig basierend auf einer bestimmten Version eines Informationssystems publiziert. Wird dieses linear versioniert, sind ältere Versionen von Informationen z. T. nicht oder nur mit Schwierigkeiten zugreifbar [WGK+06, Wei05].

4.3.3 Vorhersagemethoden

Ein weiteres Problem liegt in der Verwendung von Vorhersagemethoden. Dies ist grundsätzlich nicht negativ zu bewerten, insbesondere dann, wenn bestimmte biologische Sachverhalte noch nicht ausreichend erforscht sind. Problematisch ist dieses Vorgehen jedoch, wenn solche Vorhersagen als Fakt interpretiert oder sogar noch generalisiert und auf weitere Daten (z. B. von anderen Organismen) übertragen werden [BB96]. In [MNF03] sind verschiedene Studien angeführt, die bei abgeleiteten Informationen über strukturelle oder funktionale Eigenschaften von Sequenzen von Fehlerraten bis zu über 40% ausgehen.

4.3.4 Nichteinheitliche Vokabulare / Methoden

Oftmals werden in der Pflanzenbiologie keine einheitlichen Vokabulare verwendet. Zwar werden in der Molekularbiologie zunehmend sowohl quasi-standardisierte Verfahren als auch kontrolliertes Vokabular in Form von Ontologien eingesetzt, beispielsweise die GeneOntology [GO 08], die PlantOntology [ATI+08] oder die TraitOntology [JWN+02], und dadurch sukzessive die Vokabulare in diesem Bereich vereinheitlicht. Jedoch werden im Bereich der Daten über pflanzengenetische Ressourcen (z. T. durchaus vorhandene[2]) Vokabulare deutlich seltener eingesetzt, wodurch die Vergleichbarkeit solcher Daten erheblich gemindert wird.

Vielfach werden bei Experimenten keine standardisierten Methoden verwendet. Dadurch kommt es zu erheblichen Problemen, wenn Daten aus vergleichbaren Versuchen, die aber in unterschiedlichen Laboratorien erhoben wurden, verglichen werden sollen [Mem05, Irr05].

[2] Z. B. FAO/IPGRI Multi-Crop Passport Descriptors (MCPD) [ADM01]

4.4 Biologisch bedingte Ursachen für Qualitätsprobleme

Neben den vom Menschen verursachten Qualitätsproblemen gibt es auch solche, die in der Natur des biologischen Objektes (Pflanze) begründet sind. Beispielsweise sind je nach Entwicklungsstadium eines Organismus verschiedene Gene aktiv (Genexpression) und damit unterschiedliche Stoffwechselwege und -funktionen. Dies hat natürlich auch Auswirkungen auf phänotypische Ausprägungen. Hinzu kommen Organe, Gewebe, Zelltypen und Kompartimente (subzelluläre Einheiten), die in einzelnen Entwicklungsstadien noch nicht existieren. Biochemische Reaktionen finden nicht immer zur selben Zeit an denselben Orten im Organismus statt. Diese Aufzählung ließe sich beliebig weiter fortsetzen.

Wie schon in Abschnitt 4.2 beschrieben, gilt es, die Erhebung von Daten feingranular zu dokumentieren. Jedoch verbleiben, bedingt durch die Anzahl von Einflussfaktoren, z. T. große Unsicherheiten bezüglich der Aussagekraft von Daten. Dies gilt es, bei der Analyse solcher Daten zu berücksichtigen. Diese Problematik soll am Beispiel von Umwelteffekten illustriert werden. Sie wird während der Beschreibung des Anwendungsfalls in Kapitel 7 erneut aufgegriffen werden.

Umwelteinflüsse (z. B. Licht, Boden, Klima), biotische Stressfaktoren (z. B. Krankheitserreger, Nährstoffmangel bzw -überschuss) und abiotische Stressfaktoren (z. B. Trockenheit, Toxizität) haben erhebliche Auswirkungen auf Pflanzen (z. B. [HNFH06, IPS05]). In Abbildung 4.1 wird anhand mehrerer Kornqualitätsmerkmale gezeigt, wie groß der Einfluss von Umwelt- und sonstigen Merkmalen auf phänotypische Ausprägungen bei Pflanzen ist. Oft sind weniger als 30% der phänotypischen Merkmalsausprägung genetisch determiniert. Besonders deutlich wird dies am Merkmal *Rohprotein Gerste (RP)* mit ca. 10% genetischer Varianz.

4.5 Lösungsvorschläge

Das aus den vorhergehenden Abschnitten resultierende primäre Ziel ist die Vergleichbarkeit von Daten. Ist das Zustandekommen der Daten nicht durchgängig und strukturiert dokumentiert, werden Integration und weitere Analyse z. T. erheblich erschwert oder sogar gänzlich unmöglich gemacht.

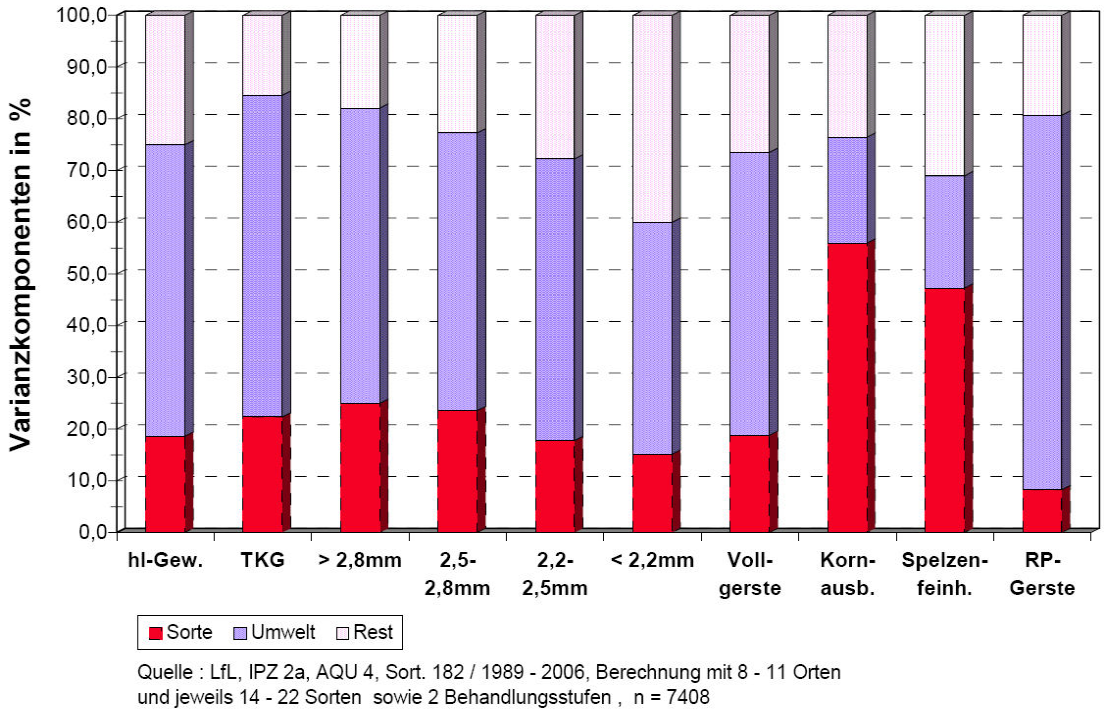


Abbildung 4.1: Einflüsse von Umwelt- und sonstigen Merkmalen auf mehrjährige Kornqualitätsparameter bei Sommergersten, Quelle: [HNFH06]

Zur Verbesserung der Datenqualität werden daher folgende Vorschläge gemacht:

- **Automatische Erfassung von Daten:**

  Direkte Einspeisung von Daten (z. B. von Sequenzierautomaten) in Informationssysteme. Das manuelle Erfassen und Kopieren von Daten sollte nach Möglichkeit vermieden werden, um das damit einhergehende Fehlerpotenzial zu verringern.

- **Protokollierung:**

  Jeder Schritt, egal ob Gewinnung von Roh- oder abgeleiteten Daten, sollte ausführlich und vor allem strukturiert dokumentiert werden, um eine Nachvollziehbarkeit zu gewährleisten.

- **Messverfahren:**

  Keine Verwendung von Boniturnoten, sondern Einsatz von Messverfahren. Werden dabei unterschiedliche Einheiten verwendet, so können diese, sofern sie als Metadaten verfügbar sind, problemlos abgeglichen werden.

- **Mittelwertvermeidung:**

  Keine Verwendung von Mittelwerten, da dabei das Problem der Gewichtung von Merkmalsausprägungen auftreten kann; einzelne Messwerte sind zu bevorzugen. Werden dennoch Mittelwerte oder diskrete Werte (Boniturnoten) genutzt, wären zusätzlich noch Informationen über die Verteilung der einzelnen Datenpunkte wünschenswert.

- **Datenbanknutzung:**

  Experimentell gewonnene Daten sollten in jedem Fall sowohl zentral als auch persistent gespeichert werden. Hierzu bieten sich Datenbankmanagementsysteme an.

- **Vermeidung von Ad-hoc-Statistik:**

  Keine Ad-hoc-Anwendung von statistischen Verfahren zur Datenauswertung. Dabei kommt es durch fehlerhafte Anwendung oft zu abgeleiteten Daten, deren Weiterverarbeitung bedenklich ist. Das Vorgehen sollte auf jeden Fall ausführlich dokumentiert werden, damit das Zustandekommen der Daten nachvollzogen werden kann.

- **Kennzeichnung abgeleiteter Daten:**

  Abgeleitete Daten sollten unbedingt auch als solche gekennzeichnet werden. Zusätzlich sollten (mit Unsicherheiten behaftete) vorhergesagte Daten mit einem entsprechenden Label versehen werden (Quality-Tagging). Ebenfalls sollten Daten gekennzeichnet werden, die im Rahmen einer Integration aus anderen Quellen importiert werden.

- **Parallele Versionierung:**

  Im Gegensatz zu der in der Literatur häufig geäußerten Meinung, dass für naturwissenschaftliche Daten eine sequenzielle Versionierung ausreichend sei [RL05], wird vorgeschlagen, im pflanzenbiologischen Bereich eine parallele Versionierung zu verwenden. Einerseits bietet dieses Vorgehen Unterstützung beim Auftreten abweichender fachlicher Meinungen. Andererseits wird dadurch auch das Veröffentlichen von wissenschaftlichen Ergebnissen erleichtert, da auch ältere Daten direkt zugreifbar sind.[3]

[3] Der Einsatz von Qualitäts-Tags in Verbindung mit paralleler Versionierung wurde im Rahmen dieser Arbeit in [WGK+06] vorgeschlagen.

- **Metadatennutzung:**

  Nutzung von Metadaten über Experimentaufbau, verwendete Methoden, Parameter, Einheiten, Entwicklungsstadien und Zytologie von Pflanzen etc. Hierzu gehören insbesondere:

  - die zu verwendenden Merkmalsbezeichnungen (möglichst auf Basis kontrollierten Vokabulars),
  - das Mapping von Merkmalen aus mehreren Quellen aufeinander, d. h. das Auflösen von Synonymen und Homonymen,
  - die erlaubten Wertebereiche, im Falle diskreter Werte auch die erlaubten einzelnen Ausprägungen,
  - die verwendeten Methoden zur Merkmalserfassung (die u. U. schon Rückschlüsse auf die Verlässlichkeit der Daten erlauben),
  - die verwendeten Einheiten der Merkmale (zuzüglich Informationen zur Umrechnung zwischen den Einheiten),
  - die externen Einflüsse / Umweltbedingungen bei der Erhebung der Daten sowie
  - Qualitätstags.

  Hierfür sollten, soweit möglich, kontrollierte Vokabulare oder biologische Ontologien zum Einsatz kommen.

Semantische und syntaktische Probleme von Daten müssen im Rahmen einer Datenintegration gelöst werden. Dafür unbedingt notwendig sind Expertenwissen und die Definition von Integritätsbedingungen. Besonders letzteres wird durch unvollständiges Wissen über die einzelnen Datendomänen und eine Vielzahl von biologischen Ausnahmen erschwert [MWBL05]. Probleme können oftmals nur durch Fachwissenschaftler der jeweiligen Datendomänen gelöst werden. Danach kann die Bereinigung der eigentlichen Daten zumindest semi-automatisch ablaufen; evtl. auftretende Konflikte müssen meist manuell aufgelöst werden.

Die Messung der Qualität pflanzenbiologischer Daten ist nur unter Zuhilfenahme geeigneter Stichproben möglich. Hierzu müssen die Daten der Stichprobe mit nachweislich korrekten Daten (z. B. aus einer externen Quelle oder manuell aufbereitet) verglichen werden. Beispielsweise kann die Korrektheit eines Stichprobeneintrages über die Distanz zum korrespondierenden Eintrag der externen Quelle bestimmt werden. Bei alphanumerischen Daten können hierfür die in Abschnitt 2.1.4 vorgestellten Methoden Anwendung finden. Eine weitere Verwendungsmöglichkeit dieser Methoden, die Verknüpfung von Records unterschiedlicher Datendomänen, wird in Abschnitt 6.5.2 ausführlich diskutiert. Für einen Überblick verschiedener Metriken zur Bewertung von Datenqualität sei auf [Kli08] verwiesen.

Im Rahmen dieser Arbeit wird der Fokus auf die Sicherstellung einer hohen Datenqualität während der Integration von Daten gelegt.

4.6 Resümee

In diesem Kapitel wurde sich mit der Qualität pflanzenbiologischer Daten auseinandergesetzt. Hierzu erfolgte eine Klassifikation der Ursachen von Qualitätsproblemen. Die vorgeschlagenen vier Kategorien wurden anhand von Beispielen diskutiert.

Am Ende des Kapitels wurden Vorschläge zur Verbesserung der Datenqualität gemacht und erläutert. Diese fließen in den Entwurf eines Konzepts in Kapitel 6 sowie in die Entwicklung eines Prototypen in Kapitel 7 ein.

5 Untersuchung existierender Integrations- und Analyseansätze

In diesem Kapitel sollen ausgewählte bioinformatische Integrations- und Analyseansätze vorgestellt und anhand verschiedener Bewertungskriterien untersucht werden. Da keine geeigneten Integrations- und Analyseansätze speziell für den pflanzlichen Bereich zur Verfügung stehen, erfolgt die Betrachtung allgemeiner bioinformatischer Ansätze. Ein Schwerpunkt liegt dabei auf ihrer Anwendbarkeit auf pflanzenbiologische Daten. Die Ergebnisse dieser Untersuchung werden als Anforderungen in das in Kapitel 6 zu entwickelnde Konzept einfließen.

5.1 Bewertungskriterien

Einheitliche Kriterien sind die Voraussetzung für den Vergleich von Integrations- und Analyseansätzen. Hierzu soll im ersten Schritt auf die Arbeit von [Sch02] zurückgegriffen werden. Darin wurden verschiedene bioinformatische Integrationsansätze analysiert und anhand von zehn Kriterien mit $+$ bzw. $-$ bewertet, die hier kurz vorgestellt werden:

- **Grad der Integration (G):**

  Hierbei wird festgestellt, ob die Schemata der einzelnen zu integrierenden Datenquellen in einem gemeinsamen, globalen Schema zusammengeführt werden $(+)$ (enge Kopplung) oder ob das globale Schema nur aus einem Nebeneinander von Teilschemata besteht $(-)$ (lose Kopplung).

- **Materialisierung der Integration (M):**

  Dieses Kriterium bewertet, zu welchem Grad die integrierten Daten materialisiert bzw. verlinkt sind. Materialisierte Daten sind performanter zugreifbar, allerdings unter Umständen nicht aktuell $(-)$. Verlinkte oder virtuell integrierte Daten sind ständig aktuell $(+)$, allerdings ist diese Aktualität mit Kosten in Form von Netzwerkzugriffen etc. verbunden.

- **Realisierungsstand (R):**

  Dieses Merkmal unterscheidet, ob ein zu bewertender Ansatz implementiert ist $(+)$ oder ob es sich dabei nur um einen theoretischen Ansatz handelt $(-)$.

- **Plattformunabhängigkeit (P):**

  Hier wird unterschieden, inwieweit ein Integrationssystem an eine bestimmte Architektur gebunden $(-)$ oder auf verschiedenen einsetzbar ist $(+)$.

- **Internetfähigkeit (I):**

  Dieses Merkmal gibt an, ob entfernte Zugriffe (Internet/Intranet) auf das Integrationssystem möglich sind $(+)$ oder ob eine lokale Installation erforderlich ist $(-)$.

- **Schnittstelle, Anfragesprachen (SA):**

  Dieses Kriterium ist erfüllt, wenn mit Hilfe von Standardanfragesprachen wie z. B. SQL (siehe Abschnitt 3.2.1) auf die integrierten Daten zugegriffen werden kann.

- **Schnittstelle, Programmiersprachen (SP):**

  Das mit dem vorigen eng verwandte Kriterium gibt über die Möglichkeiten des Datenzugriffs über Application Programming Interfaces (APIs) wie beispielsweise JDBC [Sun09] Auskunft.

- **Schnittstelle, Datenausgabeformate (SF):**

  Hier wird bewertet, ob die integrierten Daten mit Standardaustauschformaten zugreifbar gemacht werden können $(+)$. Im bioinformatischen Umfeld kommen hierzu insbesondere SBML [HFS+03] oder auch MAGE-ML [SMS+02] in Betracht.

- **Flexibilität (F):**

  Dieses Merkmal unterscheidet, ob das zu bewertende Integrationssystem hinsichtlich neuer Anforderungen anpassbar ist $(+)$ oder ob es sich um eine statische Lösung handelt $(-)$.

- **Unterstützung von Informationsfusion (U):**

  Mit dem zehnten Merkmal wird bewertet, ob ein System Informationsfusion unterstützt $(+)$. Informationsfusion bedeutet in diesem Kontext die Kombination von Daten heterogener Quellen mit dem Ziel der Ableitung neuer Informationen.

Um neben der Integration auch die Analyse von pflanzenbiologischen Daten bewerten zu können, ist die Verwendung weiterer Kriterien erforderlich. Hierzu wurden auf Basis von in verschiedenen pflanzenbioinformatischen Forschungsprojekten gewonnenen Erkenntnissen sieben weitere Kriterien entwickelt. Diese werden nun erläutert:

- **Gleichzeitige Verwendung verschiedener Datendomänen (D):**

  Zur Verschiebung des Fokus der pflanzenbiologischen Forschung vom hypothesengetriebenen zum datengetriebenen Arbeiten wird es als notwendig erachtet, Daten unterschiedlicher Domänen nicht nur separat zu betrachten. Mit diesem Kriterium soll analysiert werden, inwieweit das zu bewertende System die gleichzeitige Analyse verschiedener Datendomänen ermöglicht $(+)$ oder ob es auf bestimmte Datendomänen beschränkt ist $(-)$.

- **Unterstützung ergebnisoffener Analysen (E):**

  Hier wird bewertet, ob ein System nur die Überprüfung von Hypothesen ermöglicht $(-)$ oder darüber hinaus auch die ergebnisoffene Untersuchung von integrierten Datenbeständen $(+)$. Letzteres ist dann gegeben, wenn durch die Anwendung von Analysemethoden neue Hypothesen generiert werden können.

- **Beschränkung auf eine Klasse von Analysen (A):**

  Es soll untersucht werden, ob Auswertungssysteme auf eine Klasse von Analysen oder sogar Daten fokussieren $(-)$ oder ob grundsätzlich eine Vielzahl von Datendomänen und Analysen berücksichtigt werden kann $(+)$.

- **Beschränkung auf ein festes Zielschema (Z):**

  Bezüglich des Zielschemas soll unterschieden werden, ob das zu untersuchende System ein Schema für die integrierten Daten zwingend vorgibt $(-)$ oder ob dieses variabel ist, insbesondere hinsichtlich von Analysen, die zum Zeitpunkt des Entwurfs des Systems noch nicht angedacht waren $(+)$. Dieses Merkmal steht in engem Zusammenhang mit den Kriterien F und A.

- **Verwendbarkeit bei proprietären Datenformaten (V):**

  Im Bereich der pflanzenbiologischen Forschung liegen große Datenmengen nur in Form proprietärer, z. T. sehr heterogener Dateien vor. Daher ist die Verwendbarkeit eines Analyseansatzes auch für diese Daten sehr wichtig $(+)$.

- **Berücksichtigung der Datenqualität (Q):**

  Es soll untersucht werden, ob in einem Integrations- und Analysesystem die Qualität der zugrundeliegenden Daten in angemessener Weise berücksichtigt wird und ob Mechanismen zur Qualitätsverbesserung vorgesehen sind $(+)$.

- **Nutzung von Metadaten (N):**

  Mit diesem Kriterium wird bewertet, ob ein System die Nutzung von Metadaten ermöglicht, um die Nachvollziehbarkeit und die Vergleichbarkeit von sowohl integrierten Daten als auch Analyseergebnissen zu gewährleisten $(+)$.

Nachfolgend wird eine Auswahl aktueller Systeme zur Integration und Analyse biologischer Daten vorgestellt und anhand der siebzehn oben beschriebenen Kriterien bewertet. Erfüllt das System das jeweilige Kritierium, wird dies mit einem $+$ bewertet, im negativen Fall mit $-$. Sind keine Informationen vorhanden, wird dies mit $\circ$ gekennzeichnet. Die Ergebnisse werden am Ende des Kapitels in Form einer Tabelle zusammengefasst.

5.2 Gene-EYe

In [RHM04] wird die Integrations- und Analyseplattform Gene-EYe vorgestellt. Gene-EYe implementiert eine 3-Schichten-Architektur, die aus den Ebenen *Genome Data Store*, *Genome Database* und *Genome Data Warehouse* besteht (Abbildung 5.1).

In der untersten Ebene, dem *Genome Data Store*, können Daten aus verschiedenen Quellen (V mit $+$) relational und materialisiert abgespeichert werden (M mit $-$). Die daraus resultierenden Nachteile bezüglich der Aktualität der Daten werden bewusst akzeptiert. Das relationale Schema der jeweiligen Daten wird direkt von ihrer Quelle abgeleitet. Die Daten werden in dieser Ebene noch nicht integriert, für jede Datenquelle gibt es ein eigenes Schema. Weitere Quellen können eingebunden werden (F mit $+$).

Die zweite Ebene, die *Genome Database*, dient der eigentlichen Integration der Daten. Dies erfolgt über die Definition problemspezifischer Sichten. Dabei werden Daten aus verschiedenen Quellen zusammengeführt (G mit $+$). Es gibt keine Einschränkung in Form eines festen Zielschemas (Z mit $+$). Analysen werden auf den integrierten Daten dieser Ebene durchgeführt (U mit $+$). Der Fokus des Systems liegt dabei auf molekularbiologischen Daten, speziell auf Sequenzdaten (D mit $-$). Dies gilt auch für die angebotenen Analysen (A mit $-$), die im Rahmen des so genannten *Genomic Toolkits* verfügbar sind.

Die Datenanalyse ist im beschriebenen Anwendungsfall auf ein konkretes Problem fokussiert, es wird jedoch davon ausgegangen, dass sich in gewissem Umfang auch neue Hypothesen ableiten lassen (E mit $+$).

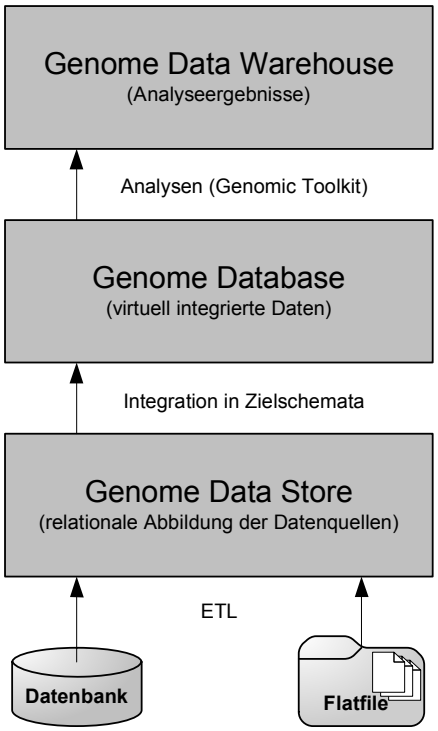


Abbildung 5.1: Schematische Darstellung des Gene-EYe-Ansatzes nach [RHM04]

In der obersten Ebene, dem *Genome Data Warehouse*, werden Analyseergebnisse abgespeichert. Solche Ergebnisse können wiederum als neue Datenquellen für weitere Analysen in den *Genome Data Store* eingebunden werden.

Über Datenaustauschformate liegen keine Informationen vor (SF mit $\circ$). Gene-EYe ist implementiert (R mit $+$), über eine Plattformunabhängigkeit wird jedoch keine Aussage getroffen (P mit $\circ$). Dasselbe gilt für die Internetfähigkeit des Systems (I mit $\circ$).

Gene-EYe unterstützt die Nutzung von Metadaten (N mit $+$) und die Bereinigung von Daten im Rahmen von ETL-Prozessen (Q mit $+$).

5.3 Columba

Columba [RMT$^+$04, TRM$^+$05] ist ein System zur multidimensionalen Integration von Proteinstrukturdaten (D mit $-$). Im Fokus dieses Ansatzes stehen Datenobjekte aus der Protein Data Bank (PDB) [BWF$^+$00]. Als Dimensionen sind um diese Objekte Daten verschiedener Quellen gruppiert, die der Beschreibung von Proteindaten dienen (Z mit $-$). Diese Daten werden materialisiert (M mit $-$) in jeweils eigenen Schemata (G mit $-$) gespeichert.

Columba integriert Daten aus verschiedenen heterogenen Quellen und ermöglicht damit Informationsfusion (U mit $+$). Die zu integrierenden Daten werden hauptsächlich aus Flat-Files oder HTML-Dateien extrahiert, daher kann die Verwendbarkeit bei proprietären Datenformaten (V) mit $+$ bewertet werden.

Columba ist implementiert (R mit $+$). Resultate werden als XML-Dateien angeboten (SF mit $+$). Es existiert eine zentrale Columba-Instanz. Daher wird das Kriterium Plattformunabhängigkeit (P) mit $-$ bewertet.

Columba hat Zugriff auf eine eingeschränkte Anzahl von Datenquellen[1]. Die mit Columba durchführbaren Auswertungen sind sehr eng umrissen (A und E mit $-$), hinsichtlich neuer Anforderungen erscheint das System unflexibel (F mit $-$).

Auf das eigentliche Integrationssystem ist kein Internetzugriff möglich, sondern nur auf die integrierten Daten (I mit $-$).

Über die Nutzung von Metadaten wird keine Aussage getroffen (N mit $\circ$). Mechanismen zur Verbesserung der Qualität der integrierten Daten werden nicht eingesetzt (Q mit $-$).

5.4 GeWare

GeWare [KDR04, RKL07] ist ein datawarehousegestützter Ansatz zur Integration und Analyse von microarraybasierten Genexpressions- und Annotationsdaten (D mit $-$) im biomedizinischen Anwendungsbereich. GeWare wird in Abbildung 5.2 schematisch dargestellt.

GeWare verwendet für die integrierten Daten ein festes Zielschema (Z mit $-$). Es handelt sich dabei um ein multidimensionales Datenmodell mit Expressionsdaten als Fakten sowie Annotationen, Proben, Experimenten und Methoden als Dimensionen (G mit $+$). Zu integrierende Daten werden in einer Staging Area vorverarbeitet und im eben genannten Zielschema materialisiert gespeichert (M mit $-$). Daten aus verschiedenen Quellen können integriert und gemeinsam analysiert werden. Somit unterstützt GeWare Informationsfusion (U mit $+$). Die Integration neuer Daten kann über Webseiten gesteuert werden (I mit $+$).

Während die zu integrierenden Expressionsdaten durch die Fokussierung auf Affymetrix-Chips[2] standardisiert sind, bietet GeWare eine Reihe von Parsern, um Annotationsdaten aus verschiedenen Quellen zu importieren (V mit $+$).

[1] Auf der Columba-Webseite (http://www.columba-db.de) sind 19 Datenquellen gelistet [Stand 2009-04-02].

[2] http://www.affymetrix.com [Stand 2009-04-02]

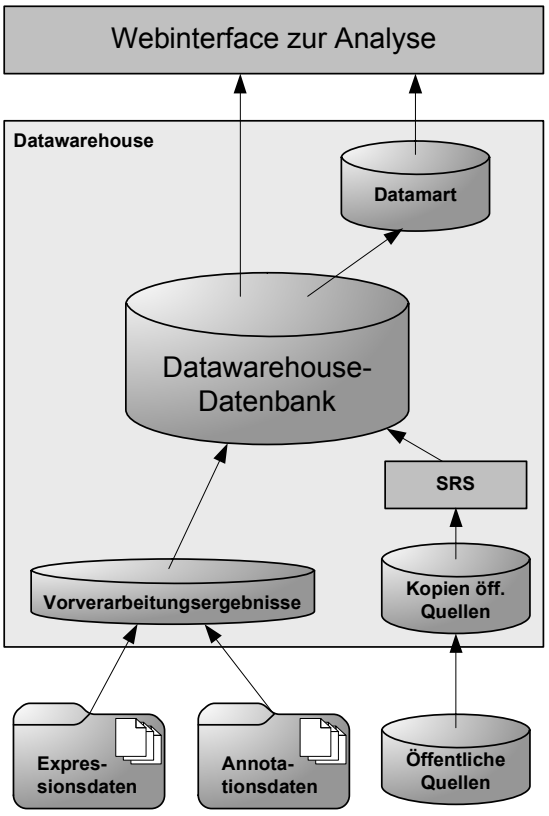


Abbildung 5.2: Schematische Darstellung des GeWare-Ansatzes nach [RKL07]

Es werden verschiedene Normalisierungs- und Auswertungsmethoden angeboten, die aber auf die Domäne der Expressionsdaten beschränkt sind (A mit $-$). Ebenso wird die Möglichkeit, neue Arten von Daten zu verarbeiten oder neue Analysemethoden anzuwenden, vom festen Zielschema des GeWare-Ansatzes eingeschränkt (F mit $-$).

GeWare verwendet das Konzept des OnLine Analytical Processings (OLAP) zur Analyse, d. h. die durchführbaren Analysen müssen im Vorfeld spezifiziert werden (vgl. Abschnitt 3.2.2). Ergebnisoffene Analysen sind damit nicht möglich (E mit $-$).

Analysierte Daten können als Flat-Files mit verschiedenen Separatoren exportiert werden. Zur Unterstützung eines Standardaustauschformates wie MAGE-ML liegen keine Informationen vor (SF mit $\circ$).

GeWare ist implementiert (R mit $+$). Über eine Plattformunabhängigkeit kann keine Aussage getroffen werden (P mit $\circ$). Dies betrifft auch die Nutzung von Metadaten (N mit $\circ$). Eine Verbesserung der Datenqualität wird durch verschiedene Normalisierungsmethoden unterstützt (Q mit $+$).

5.5 Atlas

Mit Atlas [SHX+05] wird ein Datawarehouse-Ansatz zur materialisierten Integration (M mit −) biologischer Daten mit der Intention eines Inhouse-Repositorys (I mit −) verfolgt (Abbildung 5.3).

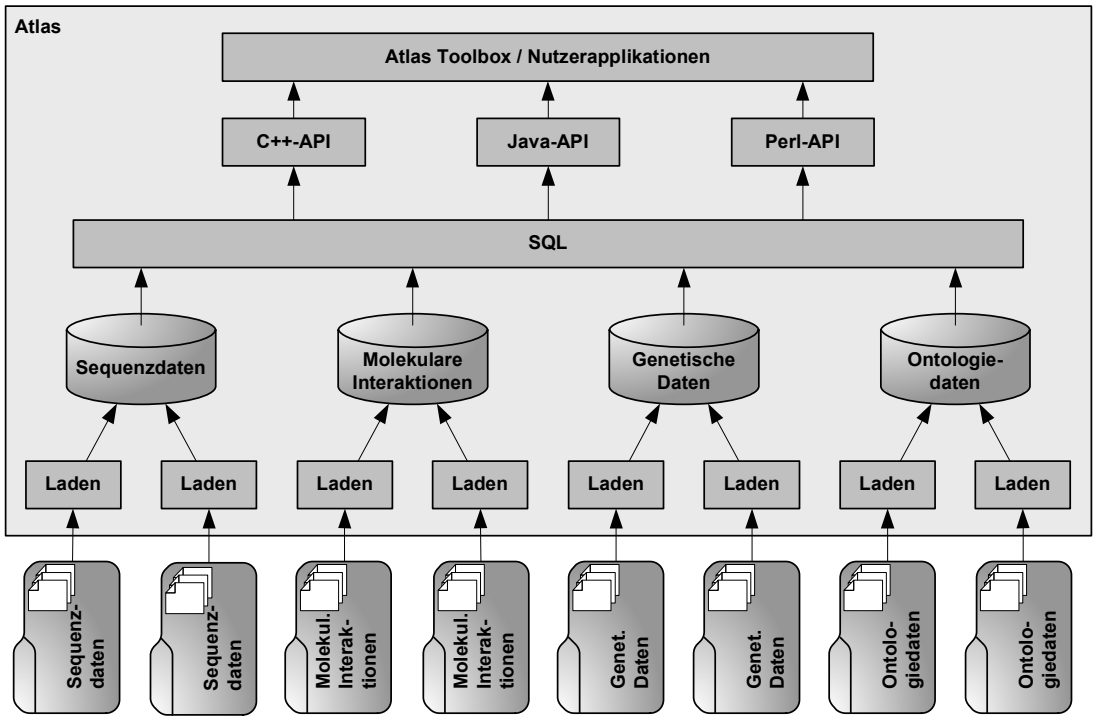


Abbildung 5.3: Schematische Darstellung des Atlas-Ansatzes nach [SHX+05]

Obwohl sich das System auf molekulare Daten stützt, ermöglicht es dennoch die Integration von vier Gruppen von Daten: Sequenzen, molekulare Interaktionen, Daten mit Beziehungen zu Genen sowie Ontologiedaten (D mit +).

Mit Atlas ist nur die Integration einer eingeschränkten Anzahl von Datenquellen möglich. Die Daten werden in ein MySQL-System integriert, es existiert kein globales Schema, sondern für jede der vier Datengruppen ein separates (G mit −). Die Teilschemata sind dabei jeweils fix (Z mit −). Auf diese Daten kann durch Application Programming Interfaces (API) mit verschiedenen Programmiersprachen, z. B. C++, Java und Perl, zugegriffen werden. Die festen Zielschemata schränken eine flexible Erweiterung mit neuen Anforderungen ein (F mit −).

Es existiert mit der so genannten *Atlas Toolbox* eine Anzahl Kommandozeilen-APIs für einfache Zugriffe (z. B. Sequence Retrieval) auf die integrierten Daten (SP mit +) einschließlich einer Reihe einfacher Applikationen. Weitere Anwendungen können auf Basis dieser APIs sowie der oben erwähnten Programmiersprachen-APIs belie-

big hinzugefügt werden. Die Art der möglichen Analysen wird nur durch die vier integrierbaren Gruppen von Daten eingeschränkt; weitere Einschränkungen bestehen nicht (A mit +). Ergebnisoffene Analysen sind damit grundsätzlich möglich (E mit +).

Atlas ist implementiert (R mit +) und für eine Unix-Basis vorgesehen (P mit −). Zum Export wird das General Feature Format (GFF)[3] unterstützt (SF mit +). Da der Fokus von Atlas auf der Integration von Daten aus etablierten bioinformatischen Quellen liegt, wird die Verwendbarkeit bei proprietären Datenformaten (V) mit − bewertet. Informationsfusion wird unterstützt (U mit +). Atlas nutzt Metadaten im Rahmen von Ontologien (N mit +), Mechanismen zur Qualitätsverbesserung werden nicht eingesetzt (Q mit −).

5.6 BioWarehouse

BioWarehouse [LPW+06] ist ein weiterer Datawarehouse-Ansatz zur Integration und Analyse bioinformatischer Daten.

Der Fokus dieses Systems liegt auf pathwayorientierten Daten (A und D mit −). Zur Speicherung der integrierten Daten dient ein festes Zielschema (Z mit −). Dadurch wird die Erweiterung um neue Anforderungen eingeschränkt (F mit −).

BioWarehouse ist nicht für die Verwendung von Daten mit proprietären Formaten konzipiert (V mit −), sondern integriert Daten aus existierenden Datenbanken. Ein Export integrierter Daten in Standardaustauschformaten ist nicht möglich (SF mit −).

Integrierte Daten werden beim BioWarehouse-Ansatz in einem gemeinsamen, globalen Schema zusammengeführt (G mit +) und dort materialisiert (M mit −).

Die Intention von BioWarehouse liegt in einer lokalen Installation. Auf integrierte Daten kann über das Internet zugegriffen werden, auf die Integrationsroutinen selbst nicht (I mit −).

Obwohl in [LPW+06] nur sehr einfache Datenabfragen beschrieben sind, werden ergebnisoffene Analysen mit dem BioWarehouse-Ansatz grundsätzlich als möglich eingeschätzt (E mit +).

BioWarehouse ist realisiert (R mit +). Über eine Plattformunabhängigkeit kann keine Aussage getroffen werden (P mit ○). Informationsfusion wird unterstützt (U mit +). Über die Nutzung von Metadaten und die Berücksichtigung der Datenqualität wird keine Aussage getroffen (N und Q mit ○).

[3] http://www.sanger.ac.uk/Software/formats/GFF [Stand 2009-04-16]

5.7 BioMart

BioMart (vormals EnsMart) [KKS+04] ist ein Framework zur Integration von Daten aus verschiedenen Quellen und zur Erzeugung von Nutzerschnittstellen zur Abfrage der integrierten Daten (Abbildung 5.4).

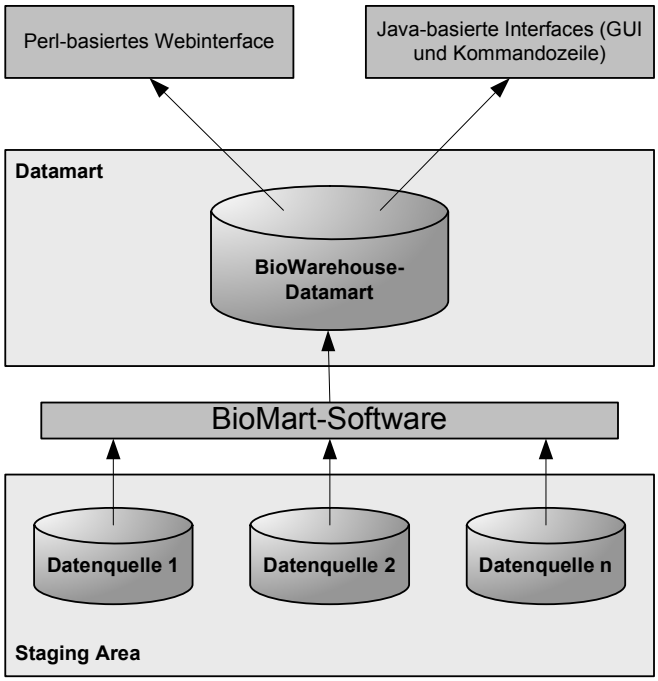


Abbildung 5.4: Schematische Darstellung des BioMart-Ansatzes nach [KKS+04]

Die BioMart-Software erzeugt aus einer Anzahl von Datenquellen, die zuvor spezifiziert wurden, einen Datamart, dem ein Sternschema mit Fakten und Dimensionen zugrunde liegt (G und Z mit $+$). Daten werden in diesem Datamart materialisiert (M mit $-$).

Die zu integrierenden Quellen müssen in einer Staging Area relational vorliegen, proprietäre Formate, wie sie in der Pflanzenbioinformatik sehr häufig vorkommen, werden nicht unterstützt (V mit $-$). Die Daten können beliebigen Datendomänen angehören (D mit $+$).

Auf die integrierten Daten kann über das Internet zugegriffen werden, auf die zur Integration verwendeten Prozeduren nicht (I mit $-$).

Obwohl Datamart-Schemata von der BioMart-Software auf der Grundlage der spezifizierten Datenquellen generiert werden, wird die Flexibilität dieses Ansatzes hinsichtlich der Adaptierbarkeit bei neuen Anforderungen negativ bewertet (F mit $-$). Die Ursache liegt in den sehr eingeschränkten Analysemöglichkeiten. Das Nutzerinterface

bietet ausschließlich Reports an, deren Treffermenge durch das Setzen von Filtern verringert werden kann (A mit $-$). Damit sind auch keine ergebnisoffenen Analysen möglich (E mit $-$).

Abfrageergebnisse können u. a. im Microsoft-Excel-Format, das in der biologischen Forschung weit verbreitet ist, heruntergeladen werden (SF mit $+$).

BioMart ist realisiert (R mit $+$). Die BioMart-Software ist in Perl bzw. Java implementiert und es existieren vorkompilierte Pakete für verschiedene Betriebssysteme (P mit $+$). Informationsfusion wird vom BioMart-Ansatz unterstützt (U mit $+$). Über die Berücksichtigung der Datenqualität und die Nutzung von Metadaten kann keine Aussage getroffen werden (Q und N mit $\circ$).

5.8 Resümee

In diesem Kapitel wurden sechs bioinformatische Integrations- und Analyseansätze anhand eines einheitlichen Sets von Kriterien bewertet. Tabelle 5.1 fasst die Bewertung der vorgestellten Ansätze zusammen. Merkmale, zu denen keine ausreichenden Informationen vorlagen, wurden mit $\circ$ gekennzeichnet. Da die vorgestellten Ansätze relationale Datenbankmanagementsysteme verwenden, kann davon ausgegangen werden, dass bei allen Systemen ein Datenzugriff sowohl über Anfrage- als auch über Programmiersprachen (z. B. mit JDBC oder ODBC) möglich ist (SA und SP mit $+$).

Bei der Analyse hat sich gezeigt, dass die Mehrheit der Ansätze nicht für die datendomänenübergreifende Arbeit, sondern nur für ausgewählte Domänen konzipiert ist. Ebenfalls fokussieren diese Ansätze mehrheitlich auf eine eingeschränkte Menge von Analysemöglichkeiten. Die Flexibilität hinsichtlich der Adaptation eines Systems an neue Anforderungen wurde, mit Ausnahme des Gene-EYe-Systems, negativ bewertet. Diese Aussage wird noch dadurch unterstützt, dass mehrheitlich fixe Zielschemata für die integrierten Daten verwendet werden, wodurch eine Anpassung erschwert wird.

Die während dieser Analyse gewonnenen Erkenntnisse, insbesondere die aufgezeigten fehlenden Eigenschaften der vorgestellten Ansätze, sollen bei der Entwicklung eines Konzepts zur integrativen Analyse pflanzenbiologischer Daten in Kapitel 6 berücksichtigt werden.

Tabelle 5.1: Überblick der bewerteten Integrations- und Analyseansätze

	G	M	R	P	I	SA	SP	SF	F	U	D	E	A	Z	V	Q	N
Gene-EYe	+	−	+	○	○	+	+	○	+	+	−	+	−	+	+	+	+
Columba	−	−	+	−	−	+	+	+	−	+	−	−	−	−	+	−	○
GeWare	+	−	+	○	+	+	+	○	−	+	−	−	−	−	+	+	○
Atlas	−	−	+	−	−	+	+	+	−	+	+	+	+	−	−	−	+
BioWarehouse	+	−	+	○	−	+	+	−	−	+	−	+	−	−	−	○	○
BioMart	+	−	+	+	−	+	+	+	−	+	+	−	−	+	−	○	○

Legende:

G	–	Grad der Integration	enge Kopplung (+)	lose Kopplung (−)
M	–	Materialisierung der Integration	nicht materialisiert (+)	materialisiert (−)
R	–	Realisierungsstand	implementiert(+)	theoretischer Ansatz (−)
P	–	Plattformunabhängigkeit	unabhängig (+)	plattformgebunden (−)
I	–	Internetfähigkeit	entfernter Zugriff (+)	lokale Installation (−)
SA	–	Schnittstelle, Anfragesprachen	unterstützt (+)	nicht unterstützt (−)
SP	–	Schnittstelle, Programmiersprachen	unterstützt (+)	nicht unterstützt (−)
SF	–	Schnittstelle, Datenausgabeformate	verschiedene Formate (+)	nur ein Format (−)
F	–	Flexibilität	anpassbar (+)	statisch (−)
U	–	Unterstützung von Informationsfusion	unterstützt (+)	nicht unterstützt (−)
D	–	Gleichzeitige Verwendung verschiedener Datendomänen	mehrere Domänen (+)	nur eine Domäne (−)
E	–	Unterstützung ergebnisoffener Analysen	unterstützt (+)	nicht unterstützt (−)
A	–	Beschränkung auf eine Klasse von Analysen	keine Beschränkung (+)	Beschränkung (−)
Z	–	Beschränkung auf ein festes Zielschema	keine Beschränkung (+)	Beschränkung (−)
V	–	Verwendbarkeit bei proprietären Formaten	verwendbar (+)	nicht verwendbar (−)
Q	–	Berücksichtigung der Datenqualität	berücksichtigt (+)	nicht berücksichtigt (−)
N	–	Nutzung von Metadaten	möglich (+)	nicht möglich (−)

6 Entwicklung eines Konzepts

In diesem Kapitel soll ein Konzept zur integrierten Analyse von Daten im pflanzenbiologischen Umfeld entworfen werden. Teile der hier beschriebenen Architektur wurden bereits in [KGM+07] vorgestellt.

In Abschnitt 3.1 wurden biologische Datenbanken und ihre Besonderheiten, insbesondere die Heterogenitäten der Struktur, der Semantik und der Zugriffsmöglichkeiten, beschrieben. Wenn Daten aus mehreren Systemen integriert und analysiert werden sollen, sind Transformationen notwendig, welche nicht immer automatisch ausgeführt werden können. Insbesondere ist es im Bereich pflanzenbiologischer Daten vielfach notwendig, eine manuelle Datenkuration durchzuführen [ZFT+05]. Der Zugriff auf Daten über Webinterfaces impliziert lange Antwortzeiten. Daten, die in keinem Informationssystem vorliegen, sondern nur als z. B. einfache Flatfiles, verfügen über keine Anfrageschnittstellen. Dies ist in der Pflanzenbioinformatik der Standardanwendungsfall. Daher muss eine Möglichkeit geschaffen werden, solche Daten für Analysen integriert und performant zur Verfügung zu stellen.

Aus den zuvor genannten Gründen scheidet der Einsatz der virtuellen Integration (siehe Abschnitt 3.1) aus. Am vielversprechendsten erscheint die Verwendung materialisierter Integration in Verbindung mit Datawarehouse-Methoden.

Der Standard-Datawarehouse-Ansatz, wie er im Unternehmensbereich Anwendung findet, ist nur bedingt auf biologische Daten übertragbar [Aug01]. Einerseits sind Datenquellen im Life-Science-Bereich häufiger Schema- und Systemevolution unterworfen und andererseits ist der Prozess der Warehouse-Erstellung per se zeitaufwändig und kostspielig. Andere Merkmale eines Datawarehouses wie z. B. die zeitliche Dimension von Daten, die insbesondere im Unternehmensbereich für die Betrachtung

von Umsatzentwicklung etc. sehr wichtig ist, spielen bei biologischen Daten z. T. nur eine untergeordnete Rolle (z. B. bei Sequenzdaten) [SKSB00].

Dessen ungeachtet erfreuen sich Datawarehouse-Methoden in der bioinformatischen Forschung wachsender Beliebtheit. Im Folgenden soll ein Prozessfluss zur Integration und Analyse pflanzenbiologischer Daten unter Verwendung von Datawarehouse-Methoden entworfen werden.

Hierbei müssen vier Hauptaufgaben berücksichtigt werden: Analysen sollen (1) flexibel und zeitnah durchführbar sein und es müssen sowohl (2) datengetriebene Analysen (im Hinblick auf Massendaten) als auch gleichzeitig (3) problemorientiertes Arbeiten (hypothesengetrieben) ermöglicht werden. Darüber hinaus soll (4) eine hohe Datenqualität sichergestellt werden.

Es wird eine mehrschichtige Architektur, die aus den Schritten

1. Quelldaten,
2. Extraktion, (Transformation,) Laden,
3. Datenpool,
4. Transformation und Laden,
5. analysespezifische Datamarts und
6. Analyse

besteht, vorgeschlagen. Abbildung 6.1 illustriert die vorgeschlagene Architektur. Die einzelnen Schichten werden in den folgenden Abschnitten diskutiert.

6.1 Schicht 1: Quelldaten

Die erste Schicht umfasst die Quelldaten. Diese können über Operativsysteme, proprietäre Formate, XML-Dateien oder Flatfiles mit vielfältigen Formatierungen verfügbar sein. Auf diese Daten wird in der hier vorgeschlagenen Architektur nur lesend zugegriffen.

Wenn Möglichkeiten der Einflussnahme auf die Erhebung der Quelldaten bestehen, sollten die in Abschnitt 4.5 gemachten Vorschläge zur Verbesserung der Datenqualität berücksichtigt werden. Hierbei sind insbesondere die automatische und strukturierte Erfassung von Daten in Informationssystemen sowie die Verwendung von Metadaten hervorzuheben.

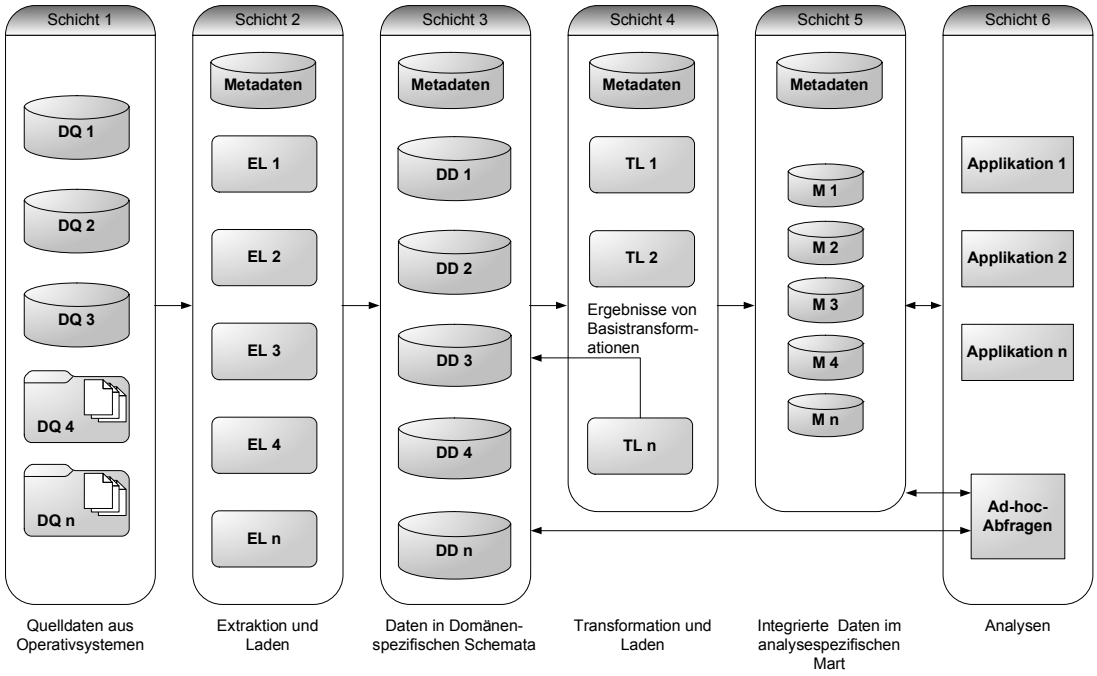


Abbildung 6.1: Architekturüberblick des Systems

6.2 Schicht 2: Extraktion, (Transformation,) Laden

Schicht 2 beinhaltet die Extraktions-, Transformations- und Ladeschritte, die für den Datenimport aus den Quellen notwendig sind. Hierfür bietet sich der Einsatz etablierter Technologien wie des Oracle Warehouse Builders (OWB) [Ora09c] oder (im bioinformatischen Bereich) des Sequence-Retrieval-Systems (SRS) [EHB03] an.

Es wird vorgeschlagen, den Fokus dieser Schicht auf Extraktion und Laden zu legen, um so viel Quelldaten wie möglich im Datenpool (Schicht 3) zu speichern. Unabhängig davon soll es möglich sein, in dieser Schicht (Basis-)Transformationen durchzuführen. Die eigentlichen Transformationen sollten jedoch möglichst erst in Schicht 4 stattfinden, um flexibel hinsichtlich verschiedener Transformationsmethoden oder zukünftiger Analyseansätze zu sein. Dies liegt in der Datenlage der Pflanzenbioinformatik begründet, in der Daten oftmals nur in proprietären Formaten und in unzureichender Qualität vorliegen (vgl. Kapitel 4) und in der Fragestellungen häufig wechseln. Mit Hilfe von Transformations- und Ladeschritten (Schicht 4) sollen analysespezifische Datamarts (Schicht 5) befüllt werden.

Es kann jedoch sinnvoll sein, auf Basis von Metadaten nur bestimmte Datensätze zur Speicherung zuzulassen (z. B. einheitliche Benennung von Merkmalen, korrekte Wertebereiche), um die später erfolgende Datenbereinigung zu vereinfachen. Dies muss

im Kontext des jeweiligen Anwendungsgebietes diskutiert werden. Auch hier sei auf die Vorschläge aus Abschnitt 4.5 verwiesen.

6.3 Schicht 3: Datenpool

Das zentrale Element der hier vorgeschlagenen Architektur wird von einem Datenpool gebildet, der Eigenschaften von sowohl Datawarehouse-Datenbanken als auch Operational Data Stores kombiniert.

Wie in Abschnitt 3.1.2 erläutert wurde, besteht die Intention einer Datawarehouse-Datenbank darin, vor allem aggregierte Daten über einen längeren Zeitraum persistent zu speichern [Inm05]. Im Gegensatz dazu ist ein Operational Data Store in der Datawarehouse-Terminologie ein Datenbankschema, dessen Zweck darin besteht, nicht-aggregierte, flüchtige Daten über einen kurzen Zeitraum zu halten [Inm99]. Des Weiteren können die Daten im Operational Data Store um abgeleitete (sekundäre) Daten ergänzt werden.

In der Pflanzenbiologie wird häufig sehr problemorientiert geforscht und Fragestellungen sind oftmals nur für eine bestimmte Zeit von Interesse. Daher stellt sich die Frage, inwieweit es sinnvoll ist, jeweils zu versuchen, komplexe Datawarehouse-Schemata zu entwickeln. Zum einen ist der Entwicklungsprozess zeitaufwendig und zum anderen sind solche Schemata, vor allem Stern- oder Schneeflockenschemata, auf bestimmte Analysen zugeschnitten und müssen im Falle der Nutzung für andere Analysen neu aus den Quellen in andere Schemata etc. importiert werden.

Daher wäre es vorteilhaft, einen Basisdatenpool aufzubauen und zu pflegen. Zur Sicherstellung einer hohen Qualität erscheint es sinnvoll, Werkzeuge zur manuellen Kurierung der Daten zu schaffen. Die manuelle Kurierung von Daten findet in der Pflanzenbiologie zunehmend Verbreitung [ZFT+05]. Im Rahmen dieser Arbeit wurde ein solches Vorgehen in [WGK+06] und [GBWK+08] beschrieben.

Der Basisdatenpool kann die Grundlage späterer Analysen bilden. Die Analysen in der Pflanzenbioinformatik sind unterschiedlicher Natur; das in Unternehmens-Datawarehouses häufig eingesetzte OLAP stellt hier nur einen kleinen Anteil.

Für den Basisdatenpool wird die Verwendung datendomänenspezifischer Schemata vorgeschlagen. Um Daten aus neuen Quellen zeitnah und mit nur geringem Aufwand in den Datenpool zu importieren, müssen diese Schemata nach Möglichkeit generisch sein. Hierfür bietet es sich an, je nach Datendomäne, zwischen einer klassischen relationalen Modellierung sowie der Verwendung des so genannten Entity-Attribute-Value-Ansatzes (EAV) abzuwägen. Die Vor- und Nachteile des jeweiligen Vorgehens sollen im Folgenden diskutiert werden.

Klassische relationale Modellierung

Die klassische relationale Modellierung ist die etablierte Technik. Hierbei werden Eigenschaften von Realweltobjekten auf Attribute von Modellobjekten abgebildet. Dadurch kann das Ausmaß der Logik, die in der Applikationsebene benötigt wird, relativ niedrig gehalten werden. Jedes Attribut stellt eine Spalte in einer Tabelle dar. Referentielle Integritäten etc. können durch herkömmliche DBMS-Mechanismen sichergestellt werden. Ein solches Schema kann auch ohne tieferes Hintergrundwissen gelesen werden. Allerdings impliziert dieses Vorgehen eine Reihe von Nachteilen. Die Art der Modellierung ist vergleichsweise statisch. Kommen weitere Attribute hinzu, muss eine Schemaanpassung durchgeführt werden. Fallen Attribute weg, muss ebenfalls das Schema geändert werden respektive diese Attribute altern aus. Das Hinzufügen eines Attributes bedeutet nicht zwangsläufig nur die Erweiterung einer Tabelle. In vielen Fällen ist damit das Erstellen einer neuen Tabelle (1:n-Beziehung) verbunden, um der Normalform zu entsprechen. Bei m:n-Beziehungen kommen noch Brückentabellen hinzu. Je nach Frequenz der Schemaanpassungen kann es damit zu einer schwer überschaubaren Anzahl von Tabellen kommen. Abbildung 6.2 zeigt ein Beispiel für diese Art der Modellierung.

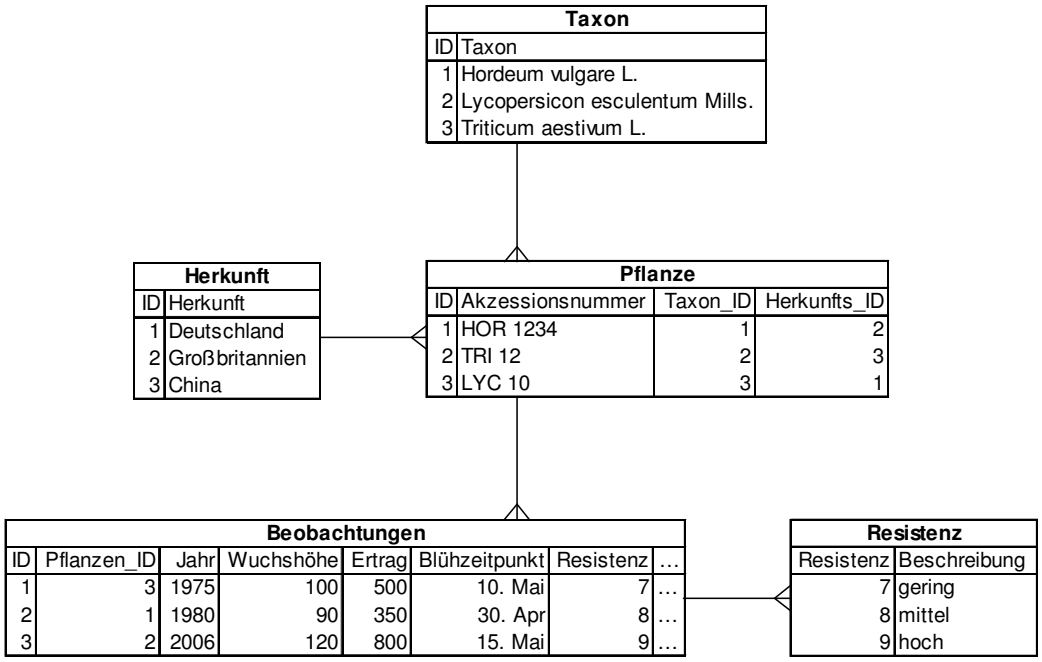


Abbildung 6.2: Speicherung phänotypischer Beobachtungswerte mit klassischer relationaler Modellierung

Entity-Attribute-Value-Ansatz

Der in Abschnitt 2.1.3 vorgestellte Entity-Attribute-Value-Ansatz (EAV) speichert, im Gegensatz zur klassischen Modellierung, Kombinationen aus Realweltobjekt (Entity), Attribut und Ausprägung (Value) als Tupel in einer Tabelle.

Die Vorteile dieses Ansatzes sollen nachfolgend erläutert werden:

- **Flexibilität:**

  Attribute werden nicht zu Spalten einer Tabelle, sondern werden als Zeilen in einer Tabelle gehalten. Auf diese Weise ist das beliebige Hinzufügen von Attributen möglich, ohne das zugrunde liegende Schema anpassen zu müssen.

- **Kompaktheit des Schemas:**

  Ein nach dem EAV-Ansatz entworfenes Schema ist sehr kompakt und unanfällig für Änderungen. Dies kann sinnvoll sein, wenn große Mengen an Datensätzen mit heterogenen Attributen abgespeichert werden sollen.

- **Effizienz:**

  Daten können sehr speicherplatzeffizient gehalten werden. Dies betrifft vor allem lückenhafte Daten sowie Datensätze mit sich ändernden Attributen, wie sie für die Pflanzenbioinformatik charakteristisch sind. Bei der klassischen relationalen Modellierung würden sehr viele Attribute unbesetzt bleiben.

Neben den eben beschriebenen Vorteilen sind mit dem EAV-Ansatz auch Nachteile verbunden:

- **Eingeschränkte Integritätssicherung:**

  Bei der Verwendung des EAV-Ansatzes kommen integritätssichernde Mechanismen relationaler Datenbankmanagementsysteme nur eingeschränkt zum Tragen. Tritt beispielsweise bei der Implementierung von Anfragen ein Schreibfehler beim Namen eines Attributes auf, so führt dies nicht zu einer Fehlermeldung des DBMS. Eine Abfrage liefert in diesem Fall keine Resultate, was aber für das DBMS keinen Fehler darstellt, da der Attributname keine Tabellenspalte ist, sondern in einer Tabellenzeile gespeichert ist. Das Debuggen wird dadurch erheblich erschwert.
  Checkconstraints, referentielle Integrität etc. können teilweise über Datenbanktrigger nachgebildet werden.

- **Verlagerung von Logik / erhöhter Entwicklungsaufwand:**

  Die Verwendung des EAV-Ansatzes bedingt die Verlagerung von Logik in die Anwendungsebene, da das Schema aufgrund seiner kompakten und flexiblen Natur Realweltobjekte und ihre Beziehungen nur eingeschränkt widerspiegelt. Dies trifft auch auf die Erweiterung EAV/CR zu.
  Ein Entwickler kann nicht auf der Grundlage der in einem herkömmlichen relationalen Schema modellierten Abhängigkeiten und Beziehungen aufbauen, sondern muss sich, um die Logik implementieren zu können, zuerst umfassend in das jeweilige Anwendungsgebiet einarbeiten. Hierbei muss beachtet werden, dass sich der Gesamtaufwand vervielfacht, wenn mehrere Entwickler beteiligt sind.

- **Umfangreiche Verwendung von Metadaten:**

  Werden nur große Mengen verschiedener Attribute verwendet, die voneinander unabhängig sind, kann die Nutzung dieses Ansatzes sinnvoll sein. Sollen allerdings ebenfalls Beziehungen zwischen Attributen abgebildet werden, kann dies, je nach Komplexität der Beziehungen und Umfang des Schemas, zu Schwierigkeiten führen. Auch in der Erweiterung EAV/CR müssen hierfür eine Reihe zusätzlicher Attribute und/oder Tabellen als Metadaten eingeführt werden, um die Beziehungen abzubilden, was die Vorteile der flexiblen Modellierung relativieren kann und/oder wiederum zusätzliche Logik in die Anwendungsschicht verlagert.

- **Transponierung von Daten / erhöhter Rechenaufwand:**

  Attribute sind beim EAV-Ansatz nicht, wie bei relationalen Tabellen üblich, nebeneinander angeordnet, sondern als Zeilen einer Tabelle untereinander. Um für Darstellungszwecke oder zur Weiterverarbeitung mit externen Werkzeugen eine konventionelle Präsentation zu erhalten, muss eine rechenaufwendige Transponierung durchgeführt werden. Dies kann mit Hilfe von (PL/SQL-)Prozeduren oder auch mit einer Vielzahl von Selbstverbünden (Selfjoins) der Tabelle über Aliase realisiert werden. Wird versucht, die Tabelle, die die Attributausprägungen (untereinander) enthält, über Aliase mit sich selbst zu verbinden, kann dies u. U. fehlschlagen, da es bei einer Reihe von Datenbankmanagementsystemen Limitierungen hinsichtlich der Anzahl von Tabellen/Aliasen pro Verbund gibt. Dieses Problem ist in [NB98] ausführlich beschrieben. Wird versucht, Daten mit einer (PL/SQL-)Prozedur (ohne Selfjoins) zu transponieren[1], kann dies nicht in einer einzigen Abfrage erfolgen, sondern diese Prozedur muss mehrere Teilabfragen ausführen.

[1] Das im Rahmen dieser Arbeit verwendete DBMS Oracle lässt dies ohne weiteres zu.

- **Erhöhte Fehleranfälligkeit:**

  Werden sehr viele Attribute verwendet und wird der Datenbestand dadurch unübersichtlich, ist es möglich, dass Attribute z. B. in ähnlicher Schreibweise doppelt angelegt werden. Dies fällt bei einer klassischen Modellierung viel schneller auf.

- **Performanceeinbußen:**

  Die Nutzung von EAV-Schemata kann mit einer Reihe von Performanceeinbußen verbunden sein, insbesondere bei sehr großen Datenmengen oder bei Verwendung sehr komplexer SQL-Anfragen [CNM+00].

Schlussfolgerung

Wie eben aufgeführt, haben sowohl die klassische relationale Modellierung als auch der EAV-Ansatz Vor- und Nachteile. Daher sollten die Vorteile beider vorgenannter Techniken miteinander kombiniert werden. Es erscheint hierbei sinnvoll, diejenigen Entitäten, deren Attribute vollständig beschrieben werden können, klassisch zu modellieren. Ist dies nicht möglich, kann um flexible Konstrukte ergänzt werden. Es wird vorgeschlagen, für jede der in Abschnitt 2.2.3 beschriebenen Datendomänen ein eigenes Schema zu entwerfen. Diese Schemata müssen einerseits den Spezifika der jeweiligen Domänen Rechnung tragen, aber andererseits auch so flexibel sein, dass problemlos neue Datenquellen angeschlossen werden können, ohne dass das Schema einer Adaptation bedarf.

Alle domänenspezifischen Schemata bilden gemeinsam den Datenpool aus Schicht 3 des hier beschriebenen Konzepts.

6.4 Schicht 4: Transformation und Laden

Schicht 4 umfasst Transformations- und Ladeschritte (ergänzend zu denen aus Schicht 2), die notwendig sind, um analysespezifische Datamarts, die in Schicht 5 beschrieben werden, zu befüllen. Der Schwerpunkt liegt hier auf den beiden Aufgaben

- Verbesserung der Datenqualität (Abschnitt 6.4.1) und
- Vorbereitung / Vorverarbeitung von Daten (Abschnitt 6.4.2).

Diese beiden Punkte spiegeln sich in Abbildung 6.3 wider, mit deren Hilfe die für Schicht 4 vorgeschlagenen Prozesse erläutert werden sollen.

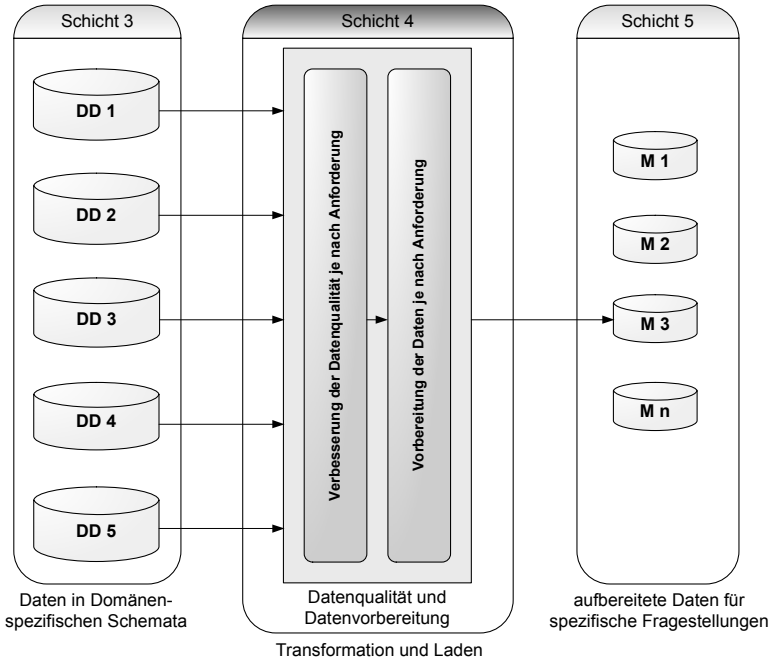


Abbildung 6.3: Detailansicht von Schicht 4 des Konzepts

6.4.1 Verbesserung der Datenqualität

Grundsätzlich sollen zu analysierende Daten eine möglichst hohe Qualität aufweisen. Dieser Anspruch ist nicht trivial. Auf die oftmals unzureichende Datenqualität in der Bioinformatik sowie auf verschiedene Vorschläge zur Verbesserung der Qualität insbesondere pflanzenbiologischer Daten wurde bereits ausführlich in Kapitel 4 eingegangen. Obwohl die Datenqualität in jeder der hier beschriebenen Schichten berücksichtigt werden sollte, wird vorgeschlagen, das Hauptaugenmerk auf Schicht 4 zu legen.

Aus der Sicht eines Datenbankentwicklers lässt sich das Problem qualitativ schlechter Daten durch die konsequente Verwendung von Konsistenzregeln, passenden Datentypen[2], Fremdschlüsseln etc. lösen. Ist ein Datenbankschema derart angelegt, können darin nur qualitativ hochwertige Datensätze abgespeichert werden. Hier wird aber sofort sichtbar, dass dadurch lediglich die Sicherstellung der Datenqualität vollständig in die Ebene der Datengewinnung bzw. der Operativsysteme verlagert wird. Wenn ohne Einflussmöglichkeiten auf externe Daten zugegriffen werden muss, könnten daher im schlechtesten Fall keine Daten verfügbar sein, die den jeweiligen Qualitätskriterien entsprechen.

Als pragmatischer Ansatz wird deshalb vorgeschlagen, im Datenpool (Schicht 3) weitgehend auf die oben genannten Constraints etc. zu verzichten, um die Befüllung des Datenpools mit großen Mengen von Daten zu ermöglichen. In der darauffolgenden

[2] z. B. keine Zahlen in Feldern vom Datentyp *Character*

Schicht 4 können verschiedene Schritte zur Verbesserung der Datenqualität unternommen werden. Welche Operationen dabei durchgeführt werden und in welchem Umfang dies geschieht, hängt von den Analyseaufgaben ab, für die der jeweilige Datamart (Schicht 5) entwickelt wird.

Denkbar sind hier z. B. folgende Möglichkeiten[3]:

- Sicherstellung, dass für ein zu analysierendes Merkmal die Ausprägungen innerhalb eines bestimmten Wertebereichs liegen; andere Daten werden dabei herausgefiltert,
- Datentypkonvertierungen entsprechend des Zielschemas,
- Umrechnung von Ausprägungen, die in verschiedenen Einheiten vorliegen,
- Abbildung von Bewertungsskalen, die unterschiedliche Werte verwenden (z. B. Bonitурskalen), aufeinander,
- Normalisierung zur Vergleichbarmachung von Daten aus verschiedenen Experimenten etc.

Die entsprechend behandelten Daten stehen danach dem analysespezifischen Datamart-Schema in Schicht 5 zur Verfügung. Die Ergebnisse von Basistransformationen, z. B. die Normalisierung von Expressionsdaten, können im Rahmen von Ergänzungen die Daten in Schicht 3 erweitern und damit sukzessive zu einer Verbesserung der Daten im Datenpool beitragen. Dies ist schematisch auch in Abbildung 6.1 dargestellt. Es wird vorgeschlagen, hierbei auf Updates der Daten in Schicht 3 zu verzichten, so dass die ursprünglichen Daten erhalten bleiben. Stattdessen könnten die transformierten Daten als qualitätsverbesserte Version gespeichert werden. Dies ermöglicht es, auch zukünftig weitere Methoden auf die Originaldaten anzuwenden, z. B. ein neuartiges Normalisierungsverfahren.

Eine weitere Aufgabe von Schicht 4 ist das Identifizieren von Duplikaten innerhalb von Datendomänen. Werden pflanzenbiologische Daten aus verschiedenen Quellen integriert, ist das Auftreten von Duplikaten möglich. Diese sollten eliminiert werden. Hierzu bietet es sich an, die Suche nach Duplikaten in den Passportdaten zu beginnen. Passportdaten nehmen diesbezüglich eine Schlüsselstellung ein (vgl. Abschnitt 2.2.3).

Für die Duplikatsuche in Passportdaten können u. a. die in Abschnitt 2.1.4 vorgestellten Methoden eingesetzt werden. Beispielsweise können Sorten- oder wissenschaftliche Namen mit Hilfe des Editierabstandes dahingehend überprüft werden, ob sie denselben Term definieren und sich nur durch einen Buchstabierfehler o. ä. unterscheiden. Hierbei wird von internem Data Linkage oder Record Linkage gesprochen.

[3] jeweils unter Nutzung von Metadaten

Noch genauere Ergebnisse können erzielt werden, wenn zusätzlich noch auf Informationen wie bestimmte Charakterisierungs- oder Markerdaten zurückgegriffen werden kann. Hierbei sollten aber keine Daten wie z. B. Evaluierungsdaten, die stark von externen Einflüssen abhängen können, verwendet werden. Außerdem sind die oben beschriebenen Metadaten und Expertenwissen unerlässlich.

Mit einem ähnlich gelagerten Problem befasst sich Abschnitt 6.5.2. Dort werden verschiedene Methoden ausführlicher diskutiert.

6.4.2 Vorbereitung / Vorverarbeitung von Daten

Eng verbunden mit der Verbesserung der Datenqualität ist auch die Vorbereitung von Daten für Analysen, wobei sich die beiden Aufgaben z. T. nicht klar voneinander abgrenzen lassen.

Damit Daten mit den in Kapitel 3.2 beschriebenen Methoden analysiert werden können, müssen bestimmte Voraussetzungen erfüllt sein. Dazu gehören je nach anzuwendender Methode beispielsweise

- die Durchführung von Diskretisierungen,
- die Behandlung von fehlenden Werten sowie
- das Erkennen von Ausreißern.

Diese vorbereitenden Schritte sollen in dieser Schicht durchgeführt werden. Die dafür notwendigen Methoden wurden in Abschnitt 3.2.4 beschrieben. Soll OLAP verwendet werden, müssen hier beispielsweise auch Aggregationen durchgeführt werden.

6.5 Schicht 5: Analysespezifische Datamarts

Schicht 5 umfasst Datamarts, die entsprechend den Spezifikationen von Nutzern entwickelt werden und den Anforderungen der jeweiligen Analysen angepasst sind. An dieser Stelle soll noch einmal hervorgehoben werden, dass die Intention, die dem hier beschriebenen Konzept zugrunde liegt, nicht darin besteht, ein allumfassendes Datawarehouse als Basis für sämtliche Analysen zu entwickeln. Dieses Vorgehen erscheint im Bereich der Pflanzenbioinformatik nicht angemessen. Neben dem Datenpool aus Schicht 3 (in dem die unterschiedlichen Domänen noch nicht miteinander verknüpft sind) und den Datamarts aus Schicht 5 sollte es keine weitere Materialisierung geben. Hierdurch wird auch der Datenqualität Rechnung getragen, die per se subjektiv ist.

Es wird vorgeschlagen, kein festes Zielschema zu verwenden, um sich nicht auf bestimmte Auswertungen festzulegen. Daher muss es ermöglicht werden, Zielschemata mit beliebigen Inhalten aus dem Datenpool (Schicht 3) zusammenzustellen. Hierzu sollten so weit wie möglich etablierte Werkzeuge Verwendung finden. Im Rahmen dieser Arbeit wurde vielfach der Oracle-Warehouse-Builder eingesetzt.

Um domänenübergreifende, analysespezifische Datamarts zu erstellen, müssen zuerst die Schemata oder Teile der Schemata verschiedener Datendomänen miteinander verbunden werden. Im zweiten Schritt können die eigentlichen Daten über identifizierende Attributwerte verknüpft werden. Dieses Vorgehen ist in Abbildung 6.4 dargestellt.

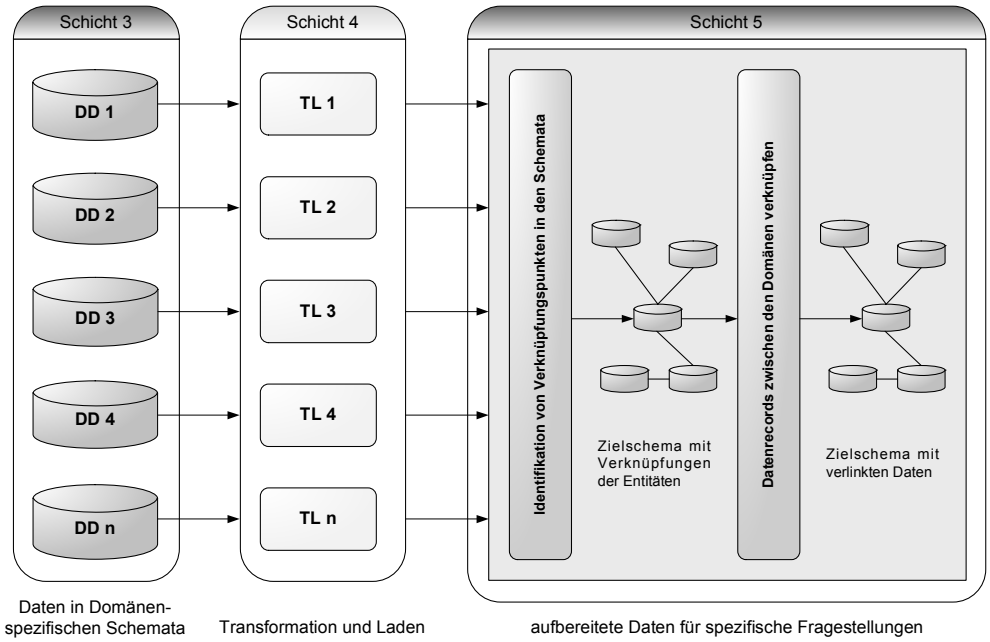


Abbildung 6.4: Detailansicht von Schicht 5 des Konzepts

6.5.1 Verknüpfen von Schemata unterschiedlicher Domänen

Um Schemata von Datendomänen oder Teile von Datendomänen miteinander zu verbinden, ist in jedem Fall Expertenwissen notwendig. Jedoch bietet es sich an, Verknüpfungspunkte bei gleichen oder ähnlichen Tabellen- und Attributbezeichnungen und auch bei gleichen Datentypen zu suchen. Dies kann semi-automatisch (mit anschließender Korrektur) über das Data-Dictionary des DBMS oder mit Hilfe kontrollierten Vokabulars ablaufen.

Grundsätzlich wird das Bindeglied zwischen verschiedenen Datendomänen in der Pflanzenbioinformatik durch das Objekt Pflanze gebildet. Solche Objekte sind am besten

durch Passportdaten abzubilden. Über Attribute wie Akzessionsnummer, Sortenname oder Taxonomie lassen sich vielfach Daten unterschiedlicher Domänen miteinander verbinden. Unabhängig davon hat die praktische Erfahrung im Rahmen dieser Arbeit gezeigt, dass physisch oftmals auch Sequenzdaten im Fokus von Untersuchungen stehen und sich Datendomänen auf diese Weise über Sequenzidentifikatoren verbinden lassen. Ein großer pflanzenbiologischer Forschungsschwerpunkt ist die Analyse von Genstruktur und -funktion. Sequenzierungen sind dabei ein häufig genutzter Ansatz.

Spezifika des jeweiligen Zielschemas müssen in Abhängigkeit der Anforderungen zusätzlich modelliert werden. Dies gilt insbesondere im Hinblick auf die Vorverarbeitung von Daten (Abschnitt 6.4.2). Das Ergebnis ist ein analysespezifisches Datamartschema.

6.5.2 Verknüpfen der Records unterschiedlicher Domänen

Nach der Identifikation von Verknüpfungspunkten zwischen verschiedenen Datendomänen und der Erstellung eines analysespezifischen Datamartschemas, müssen im zweiten Schritt die Daten der jeweiligen Domänen integriert werden. Hierzu ist es erforderlich, Pflanzenobjekte zu identifizieren, die in mehr als einer Domäne vorkommen, und darüber deren Daten zu verbinden. Dabei kann häufig nicht der Gleichheitsoperator verwendet werden (deterministisches Record Linkage). Deswegen ist es bei der Verwendung von Daten aus verschiedenen Quellen oftmals notwendig, ähnlich wie beim Finden von Duplikaten (Abschnitt 6.4.1), Attribute von Pflanzenobjekten auf ihre Ähnlichkeit hin zu überprüfen. Im Fokus steht jedoch in diesem Fall nicht das Finden von Duplikaten innerhalb einer Datendomäne, sondern das Suchen nach Identifikatoren, mit denen sich Daten aus verschiedenen Domänen mit großer Wahrscheinlichkeit einander zuordnen lassen.

Im Falle numerischer Werte ist dies mit Hilfe von Abweichungen vergleichsweise einfach ($d(x, y) = |x - y|$), kann aber bei alphanumerischen Werten sehr kompliziert werden. Wie in Kapitel 4 ausgeführt, wird die Datenqualität in der Pflanzenbioinformatik oftmals nicht den Erwartungen gerecht. So wird häufig kein kontrolliertes Vokabular verwendet, sondern es gibt unterschiedliche Schreibweisen von Sortennamen etc. Als mögliche Lösung wird der Einsatz von Similarity-Ranking-Methoden oder Äquivalenzmethoden vorgeschlagen (vgl. Abschnitt 2.1.4).

Während im Bereich der pflanzengenetischen Ressourcen als Identifikatoren meist Akzessionsnummern verwendet werden, denen ein relativ vollständiger wissenschaftlicher Name einer Pflanze zugeordnet ist (z. B. „*Hordeum vulgare* convar. *intermedium* (Körn.) Mansf.“), wird darauf bei genetischen oder molekularen Daten oftmals nicht so viel Wert gelegt. Der Nutzer solcher Daten ist häufig damit konfrontiert, dass wis-

senschaftliche Namen nur gekürzt verwendet werden (z. B. nur „*Hordeum vulgare*“ oder die englische Bezeichnung „*Barley*“). Ebenfalls wird mit Sortennamen gearbeitet. Das gelegentlich auftretende Problem mehrfach vergebener Sortennamen soll hier vernachlässigt werden. Schwerwiegender ist, dass im Falle der Sortennamen, die zentrale Identifikatoren darstellen, diese oftmals um zusätzliche Informationen angereichert werden. Beispielsweise finden sich zur Gerstensorte „Ingrid“ auch Einträge wie „Ingrid WT“, „Zweizeilige Gerste Ingrid“, „Ingrid BC mlo5“ oder „Ingrid MLG“. Um Daten mit Hilfe solcher Identifikatoren aufeinander abbilden zu können, muss versucht werden, Ähnlichkeiten zu berechnen.

Da es sich um unterschiedlich lange Zeichenketten handelt, würde die Hamming-Distanz gegen unendlich gehen und scheidet somit als Grundlage für eine Ähnlichkeitsberechnung aus. Vergleichbar sind auch die Ergebnisse bei der Verwendung von Editdistanzen. Während sich bei einfachen Schreibfehlern (einschließlich eines Buchstabens zu viel oder zu wenig) gute Resultate erzielen lassen, werden die Editdistanzen bei deutlich verschieden langen Zeichenketten wie im vorgestellten Fall sehr groß. Ähnlich unbefriedigend verhält es sich auch beim Soundex-Algorithmus.

Die Anwendung dieser Methoden soll exemplarisch am zuletzt genannten Soundex-Algorithmus gezeigt werden. Während dieser Algorithmus für die Sortennamen „Ingrid BC mlo5“ und „Ingrid MLG“ denselben Lautähnlichkeitswert berechnet (I526), ergeben die Bezeichnungen „Ingrid WT“ und „Zweizeilige Gerste Ingrid“ die beiden unterschiedlichen Werte I526 und Z242. Dies deckt sich mit den Untersuchungen aus [LR96], in denen gezeigt wird, dass nur ungefähr ein Drittel der identifizierten Mappings korrekt ist. Weiterhin muss hier in Betracht gezogen werden, dass die Sortenbezeichnungen einschließlich ihrer jeweiligen Erweiterungen in verschiedenen Sprachen vorliegen können[4]. Der Soundex-Algorithmus ist ursprünglich für die englische Sprache entwickelt worden. Zwar gibt es zwischenzeitlich Implementierungen für verschiedene Sprachen, jedoch ist es als wahrscheinlich anzusehen, dass die notwendige Kombination dieser Implementierungen die Ergebnisse weiter verschlechtert.

Im hier beschriebenen Fall erbrachten ein Local-Alignment-Algorithmus [SW81] sowie ein Longest-Common-Substring-Algorithmus die besten Ergebnisse. Unabhängig davon wird es aufgrund der Datenlage in vielen Fällen notwendig sein, die Ergebnisse manuell zu überprüfen.

An dieser Stelle soll darauf hingewiesen werden, dass es eine Vielzahl interessanter und viel versprechender Ansätze für das Record-Linkage gibt. Die Intention dieser Arbeit besteht jedoch darin, einen variablen Prozessfluss vorzuschlagen, der eine flexible Datenintegration und -analyse ermöglicht, ohne jedes Mal substanzielle Anpassungen vornehmen zu müssen. Aus diesem Grund wird vorgeschlagen, den Fokus auf eine beschränkte Menge von Algorithmen zu legen, die als so genannte nutzerdefinierte

[4] Die in dieser Arbeit betrachteten Passportdaten enthalten Informationen über Genotypen aus einer Vielzahl von Ländern einschließlich länderspezifischer Sortenbezeichungen.

Funktionen (User Defined Functions, UDFs) implementiert werden können, um eine Anwendung innerhalb eines Datenbankmanagementsystems zu ermöglichen[5].

6.6 Schicht 6: Analyse

Da die biologische Forschung sehr vielfältig ist, soll der Fokus nicht auf eine Analysemethode oder eine Klasse von Methoden gelegt werden. Die Flexibilität des vorgeschlagenen Konzepts (insb. der analysespezifischen Datamarts) ermöglicht es, sowohl Datamining-Methoden (z. B. für datengetriebene Analysen) als auch statistische Methoden zu verwenden.

In einigen Fällen erscheint auch der Einsatz von OnLine Analytical Processing (OLAP) [CCS93] vielversprechend (z. B. bei Expressionsdaten), allerdings wird diese Methode als Standardanwendung für die Pflanzenbioinformatik als nicht sinnvoll erachtet. Dies liegt zum einen an dem nur geringen Potenzial der mit OLAP durchführbaren Analysen und zum anderen an der oftmals hypothesengetriebenen pflanzenbiologischen Forschung. OLAP dient im Unternehmensbereich der Entscheidungsfindung und nicht komplexen Analysen. Es bietet die Möglichkeit, durch verschiedene Ebenen der Dimensionen von Sternschemata zu navigieren und Aggregationen von Fakten durchzuführen. Daher kann OLAP eher dazu dienen, einen Überblick über Daten zu bekommen. Soll OLAP im pflanzenbiologischen Bereich eingesetzt werden, besteht die Gefahr, dass nach Spezifikation und Implementierung die (oftmals einfache) Fragestellung bereits überholt ist. Sinnvoll ist OLAP, wenn (abgesehen vom Finden einer adäquaten Fragestellung) große Mengen von Daten anfallen (z. B. bei Hochdurchsatzmethoden) und die OLAP-Analysen oft ausgeführt werden. Ein Beispiel hierfür ist in [KDR04] beschrieben.

Für das hier beschriebene Konzept wird die Verwendung von Metadaten für die Sicherstellung hoher Datenqualität als essentiell angesehen. Es wird vorgeschlagen, in den einzelnen Schichten geeignete Strukturen zu schaffen, um solche Daten zu verwalten (z. B. Informationen über Merkmale, Experimente etc.). Außerdem sollten alle Elemente nachvollziehbar dokumentiert werden, insbesondere die Lade- und Transformationsprozeduren der Schichten 2 und 4.

[5] Im Rahmen dieser Arbeit wurden verschiedene Methoden als nutzerdefinierte Funktionen implementiert und stehen während der Integration zur Verfügung.

6.7 Bewertung des Konzepts

Abschließend soll das in den vorhergehenden Abschnitten beschriebene Konzept anhand der in Kapitel 5 vorgestellten 17 Kriterien bewertet werden.

- **Grad der Integration (G):**

  Der Datenpool in Schicht 3 des beschriebenen Konzepts verwendet datendomänenspezifische Schemata, in die Daten aus verschiedenen Quellen importiert werden. Darüber hinaus werden die Daten aus diesen Schemata in den analysespezifischen Datamarts in Schicht 5 domänenübergreifend miteinander integriert. Daher kann der Grad der Integration mit $+$ bewertet werden.

- **Materialisierung der Integration (M):**

  Im Rahmen dieser Arbeit werden Daten materialisiert integriert. Damit muss die Materialisierung der Integration mit $-$ bewertet werden. Dies wird hier bewusst akzeptiert, um den in Schicht 3 propagierten Datenpool zu schaffen.

- **Realisierungsstand (R):**

  In diesem Kapitel wurde ein Konzept erörtert. Daher wird der Realisierungsstand an dieser Stellle mit $-$ bewertet. Eine Implementierung im Rahmen eines Anwendungsfalls wird in Kapitel 7 beschrieben.

- **Plattformunabhängigkeit (P):**

  Das Konzept kann, in Abhängigkeit des verwendeten Datenbankmanagementsystems und genutzter Analysemethoden, auf beliebigen Betriebssystemplattformen umgesetzt werden. Die Plattformunabhängigkeit kann mit $+$ bewertet werden.

- **Internetfähigkeit (I):**

  Der Zugriff auf das Integrationssystem via Internet/Intranet ist mit dem in diesem Kapitel beschriebenen Konzept implementierbar. Deshalb wird die Internetfähigkeit mit $+$ bewertet.

- **Schnittstelle, Anfragesprachen (SA):**

  Integrierte Daten werden im beschriebenen Konzept in einem Datenbankmanagementsystem relational abgespeichert und sind über Standardanfragesprachen wie SQL zugreifbar $(+)$.

- **Schnittstelle, Programmiersprachen (SP):**

  Dasselbe gilt für den Datenzugriff über Application Programming Interfaces wie JDBC $(+)$.

- **Schnittstelle, Datenausgabeformate (SF):**

  Datenaustauschformate können mit dem beschriebenen Konzept ebenfalls umgesetzt werden $(+)$.

- **Flexibilität (F):**

  Das beschriebene Konzept ist, insbesondere durch die in Schicht 5 propagierten analysespezifischen Datamarts, flexibel auf neue Anforderungen anpassbar. Das Kriterium Flexibilität kann daher mit $+$ bewertet werden.

- **Unterstützung von Informationsfusion (U):**

  Die Kombination pflanzenbiologischer Daten aus heterogenen Quellen mit dem Ziel der Ableitung neuer Informationen ist die Kernaufgabe, die mit dem in dieser Arbeit beschriebenen Konzept erfüllt werden soll. Damit kann auch dieses Kriterium mit $+$ bewertet werden.

- **Gleichzeitige Verwendung verschiedener Datendomänen (D):**

  Dieselbe Aussage ist auch auf die gleichzeitige Verwendung verschiedener Datendomänen zutreffend. Deshalb wird dieses Kriterium mit $+$ bewertet.

- **Unterstützung ergebnisoffener Analysen (E):**

  Durch das Vorhandensein verschiedener datendomänenspezifischer Schemata in Schicht 3 sowie durch die flexible Erstellung von Datamarts sind beliebige (ergebnisoffene) Analysen möglich $(+)$.

- **Beschränkung auf eine Klasse von Analysen (A):**

  Dadurch, dass im erarbeiteten Konzept grundsätzlich keine Datendomänen ausgeschlossen werden und Analyseschemata flexibel erstellbar sind, gibt es keine Beschränkung auf eine Klasse von Analysen $(+)$.

- **Beschränkung auf ein festes Zielschema (Z):**

  Mit den analysespezifischen Datamarts in Schicht 5 können flexible Zielschemata erstellt werden. Damit kann dieses Kriterium mit $+$ bewertet werden.

- **Verwendbarkeit bei proprietären Datenformaten (V):**

  Durch den Datenpool in Schicht 3 des Konzepts wird sichergestellt, dass alle Daten, unabhängig von proprietären Ausgangsformatierungen, für Analysen relational vorliegen. Aus diesem Grund wird die Verwendbarkeit bei proprietären Datenformaten mit $+$ bewertet.

- **Berücksichtigung der Datenqualität (Q):**

  Die Qualität der integrierten Daten bildet einen der Schwerpunkte der vorliegenden Arbeit und wurde beim hier beschriebenen Konzept entsprechend berücksichtigt. Deshalb wird dieses Merkmal mit $+$ bewertet.

- **Nutzung von Metadaten (N):**

  Im Rahmen dieser Arbeit wird die verstärkte Verwendung von Metadaten propagiert. Diese Forderung floss auch in den Entwurf des Konzepts ein. Dieses Kriterium kann mit + bewertet werden.

Tabelle 6.1 fasst die Bewertung des Konzepts zusammen. In Kapitel 7 wird die Implementierung des dort beschriebenen Anwendungsfalls noch einmal, insbesondere in Bezug auf das Kriterium Realisierungsstand, überprüft.

Tabelle 6.1: Bewertung des Konzepts

	G	M	R	P	I	SA	SP	SF	F	U	D	E	A	Z	V	Q	N
Konzept	+	−	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Legende:

G	–	Grad der Integration	enge Kopplung (+)	lose Kopplung (−)
M	–	Materialisierung der Integration	nicht materialisiert (+)	materialisiert (−)
R	–	Realisierungsstand	implementiert(+)	theoretischer Ansatz (−)
P	–	Plattformunabhängigkeit	unabhängig (+)	plattformgebunden (−)
I	–	Internetfähigkeit	entfernter Zugriff (+)	lokale Installation (−)
SA	–	Schnittstelle, Anfragesprachen	unterstützt (+)	nicht unterstützt (−)
SP	–	Schnittstelle, Programmiersprachen	unterstützt (+)	nicht unterstützt (−)
SF	–	Schnittstelle, Datenausgabeformate	verschiedene Formate (+)	nur ein Format (−)
F	–	Flexibilität	anpassbar (+)	statisch (−)
U	–	Unterstützung von Informationsfusion	unterstützt (+)	nicht unterstützt (−)
D	–	Gleichzeitige Verwendung verschiedener Datendomänen	mehrere Domänen (+)	nur eine Domäne (−)
E	–	Unterstützung ergebnisoffener Analysen	unterstützt (+)	nicht unterstützt (−)
A	–	Beschränkung auf eine Klasse von Analysen	keine Beschränkung (+)	Beschränkung (−)
Z	–	Beschränkung auf ein festes Zielschema	keine Beschränkung (+)	Beschränkung (−)
V	–	Verwendbarkeit bei proprietären Formaten	verwendbar (+)	nicht verwendbar (−)
Q	–	Berücksichtigung der Datenqualität	berücksichtigt (+)	nicht berücksichtigt (−)
N	–	Nutzung von Metadaten	möglich (+)	nicht möglich (−)

6.8 Resümee

In diesem Kapitel wurde ein aus sechs Schichten bestehendes Konzept zur Integration und Analyse pflanzenbiologischer Daten vorgeschlagen. Die Kernelemente werden aus einem Datenpool (Schicht 3) sowie aus analysespezifischen Datamarts (Schicht 5) gebildet.

Für den Datenpool wurde die Verwendung datendomänenspezifischer Datenbankschemata empfohlen. Es wurden die Vor- und Nachteile statischer sowie generischer Modellierung diskutiert und angeregt, einen Hybridansatz zu verwenden, der die Berücksichtigung der Spezifika der einzelnen Datendomänen vorsieht.

Maßnahmen zur Verbesserung der Datenqualität in den einzelnen Schichten des beschriebenen Konzepts wurden unter Verweis auf Abschnitt 4.5 vorgeschlagen.

Abschließend wurde das beschriebene Konzepts anhand der in Abschnitt 5.1 beschriebenen Kriterien bewertet.

7 Anwendung

In den folgenden Abschnitten wird die Anwendbarkeit des in Kapitel 6 entworfenen Konzepts demonstriert. Anhand eines praktischen Anwendungsfalls aus der Pflanzengenetik wird hierzu die Entwicklung eines Prototypen zur Unterstützung von Assoziationsstudien zur Aufdeckung von signifikanten Genotyp-Phänotyp-Korrelationen in Braugerstensorten beschrieben. Die Ermittlung von Assoziationen ist für die praktische Pflanzenzüchtung von großem, wirtschaftlichem Interesse.

Der Fokus des hier vorgestellten Anwendungsfalls lag auf der bioinformatischen Unterstützung bei der Durchführung von Assoziationsstudien im Rahmen des GABI-MALT-Forschungsverbundes[1].

7.1 Beschreibung des Anwendungsfalls

Zu den kommerziell wichtigsten Eigenschaften der Gerste gehören die Malz- und die Brauqualität. Diese werden durch viele Merkmale bestimmt, die auch zueinander in Wechselbeziehungen stehen. Aufgrund der unterschiedlichen Heritabilität sind hierbei auch Umweltfaktoren zu berücksichtigen. Ihnen muss durch die Betrachtung vieler Einzelwerte aus verschiedenen Jahren und Orten Rechnung getragen werden. Außerdem sind Merkmalsausprägungen nicht immer monokausal bzw. monogenetisch bedingt[2]. Genetische Marker wie SNPs können die Identifikation von Sorten oder Zuchtstämmen mit günstigen Eigenschaften wesentlich erleichtern. Zur Aufdeckung von Marker-Merkmals-Beziehungen wurde hier der Kandidatengenansatz gewählt.

[1] http://www.gabi.de/projekte-alle-projekte-neue-seite-348.php [Stand 2009-04-02]

[2] Merkmale wie z. B. diastatische Kraft sind von den Aktivitäten verschiedener Enzyme sowie von Umwelteinflüssen abhängig

Im Rahmen des GABI-MALT-Forschungsverbundes waren die Forschungsziele in Teilprojekt 4 [MFR07]:

1. die Identifikation von Kandidatengenen, die einen Einfluss auf den Mälzungsprozess haben,
2. die Analyse der allelischen Diversität der Kandidatengene sowie
3. die Assoziation von Haplotypen- und SNP-Mustern mit Malz- und Brauqualitätsmerkmalen.

Dieses Teilprojekt wurde am IPK Gatersleben von der Arbeitsgruppe Gen- und Genomkartierung bearbeitet. Hierbei war die Durchführung von Assoziationsstudien (Forschungsziel 3) bioinformatisch zu unterstützen. Voraussetzung für die Assoziation sind zum einen die Generierung von SNPs und Haplotypenmustern für bestimmte Kandidatengene und zum anderen die Ermittlung von relevanten Brauqualitätsmerkmalen in einem repräsentativen Sortenset.

Die hier entwickelten SNP- und INDEL-Marker stoßen auf großes Interesse in der praktischen Pflanzenzüchtung, wo sie in der markergestützten Selektion (MAS) Anwendung finden [McC04, RH98].

Entwicklung von SNP- und INDEL-Markern

In der Arbeitsgruppe Gen- und Genomkartierung des IPKs wurden bisher 48 Gene ausgewählt, deren Einflüsse auf den Mälzungsprozess aus der Literatur bekannt sind (vgl. z. B. [FPL+03, HCMC+03]). Zusätzlich wurden 16 weitere interessante Kandidatengene aufgrund ihrer Expressionsdaten (ESTs) herangezogen. Von diesen Kandidatengenen wurden für Referenzsets genetisch diverser Sorten Genfragmente (400–700 bp) mittels Polymerasekettenreaktion (PCR) amplifiziert und sequenziert. Geeignete SNPs und INDELs zur Entwicklung von Markerassays wurden anhand von Polymorphiegrad und -häufigkeit ausgewählt.

Die untersuchten Genotypen repräsentieren europäische Gerstensorten sowie Züchtungsmaterial der am GABI-MALT-Forschungsverbund beteiligten Unternehmen.

Zusammenstellung phänotypischer Daten

Für die Durchführung von Assoziationsstudien sind neben Markerdaten phänotypische Daten für Malz- und Brauqualität von Gerstensorten, wie z. B. Keimfähigkeit, Eiweißlösungsgrad oder Rohproteingehalt, notwendig. Hierzu mussten Daten zusammengetragen werden, die im Verlauf der letzten 20 Jahre in Deutschland über wirtschaftlich

bedeutsame Sorten erhoben worden waren. Dabei wurden hauptsächlich Daten des Bundessortenamtes (BSA) und aus Landessortenversuchen (LSV) sowie statistische Jahrbücher der Deutschen Braugerstengemeinschaft[3] (BGJB) herangezogen.

Zusätzlich wurden im Teilprojekt 3 des GABI-MALT-Forschungsverbundes an der Bayerischen Landesanstalt für Landwirtschaft[4] (LfL) über drei Jahre an zwei Orten Feldversuche und anschließende Analysen zur Malzqualität mittels Mikromälzung durchgeführt.

Ziel des Anwendungsfalls

Ziel war es, signifikante Assoziationen von Haplotypen- sowie SNP- und INDEL-Mustern mit Malz- und Brauqualitätsmerkmalen von Gerstensorten zu finden.

Durch den Einsatz von Hochdurchsatzverfahren wie der Pyrosequenzierung kann sehr effizient eine große Anzahl von Sorten auf SNP-Polymorphismen untersucht werden. Die durch die Genotypisierung gewonnenen SNP-Markerdaten können durch Assoziationsberechnungen in Beziehung zu phänotypischen Merkmalen gesetzt werden. Dies führt zu einer deutlichen Reduzierung der zeitintensiven phänotypischen Selektion (z. B. auf Ertragsparameter) im Feld und kostenintensiver Laboranalysen und damit zu einer Optimierung der Sortenentwicklung (markergestützte Selektion).

7.2 Anforderungen

Die Umsetzung des eben beschriebenen biologischen Anwendungsfalls ließ sich aufgrund des Umfangs der dabei verwendeten Datensätze nicht gut manuell bewältigen. Die Bearbeitung sollte daher durch den Einsatz von Informatikmethoden unter Zuhilfenahme des in Kapitel 6 entwickelten Konzepts unterstützt und effizienter gemacht werden.

Es wurden allgemeine sowie Anforderungen zur Integration und Analyse von Daten formuliert. Diese werden im Folgenden beschrieben und durch Use-case-Diagramme dargestellt.

[3] Arbeitsgemeinschaft zur Förderung des Qualitätsgerstenbaus im Bundesgebiet e.V. (Braugersten-Gemeinschaft), http://www.braugerstengemeinschaft.de [Stand 2009-04-02]
[4] http://www.lfl.bayern.de [Stand 2009-04-02]

7.2.1 Allgemeine Anforderungen

Die zur Bearbeitung des Anwendungsfalles erforderlichen Daten standen teilweise in proprietären Formaten, davon mehrheitlich in MS-Excel-Dateien, zur Verfügung. Dies betraf alle Markerdaten. Die phänotypischen Daten zur Malz- und Brauqualität lagen primär nur in gedruckter Form (statistische Jahrbücher der Deutschen Braugerstengemeinschaft) vor.

Zur Speicherung phänotypischer und Markerdaten mussten daher geeignete Datenbankstrukturen entwickelt werden.

Außerdem mussten Strukturen zur Speicherung von Passportdaten geschaffen werden. Diese verbinden phänotypische und Markerdaten miteinander. Hinzu kam die Verwaltung von Charakterisierungsdaten wie z. B. die Unterscheidung zwischen Sommer- und Winterform, Zeiligkeit etc.

Die zu verarbeitenden Daten sollten mit Hilfe von Bulk-Uploads in die Datenbank importiert werden. Dazu war es notwendig, Vorlagen (MS-Excel) zu entwerfen, die von Experimentatoren zur Zusammenstellung der Daten verwendet werden können. Dies war insbesondere für die strukturierte Erfassung der bisher in statistischen Jahrbüchern vorliegenden phänotypischen Daten erforderlich.

Weiterhin waren Werkzeuge zu entwickeln, um die in den Dateien durch Experimentatoren erfassten Daten in die entwickelten Datenbankstrukturen zu importieren. Besonderes Augenmerk war hierbei auf eine grafische Benutzerführung, insbesondere zum Aufzeigen von Formatierungsfehlern und fehlenden bzw. inkonsistenten Daten, zu legen.

Wie in Kapitel 4 beschrieben, ist es oftmals notwendig, importierte naturwissenschaftliche Daten manuell zu kurieren. Dies trägt zur Verbesserung der Qualität bei. Hierfür waren für die Experimentatoren geeignete Werkzeuge zu schaffen.

Die Anforderungen sind in Abbildung 7.1 zusammengefasst.

7.2.2 Anforderungen zur Integration

Im Rahmen des hier beschriebenen Anwendungsfalles sollte eine Integration von phänotypischen und Markerdaten durchgeführt werden. Dazu mussten geeignete Datenbankstrukturen zur Durchführung der Integration entworfen werden.

Assoziationsstudien führen (häufig) auch zu falsch-positiven Ergebnissen [MBV05, CMDS03, Nat99]. Gründe dafür sind u. a. hohe Umweltvarianzen untersuchter Merkmale, zu geringe Stichprobenumfänge oder Fehlstellen. Als größter Einflussfaktor auf das Assoziationsergebnis wurde im Rahmen dieses Anwendungsfalls die Umweltvarianz pro Sorte und Merkmal betrachtet (Abschnitt 4.4). Zum Verifizieren von Ergeb-

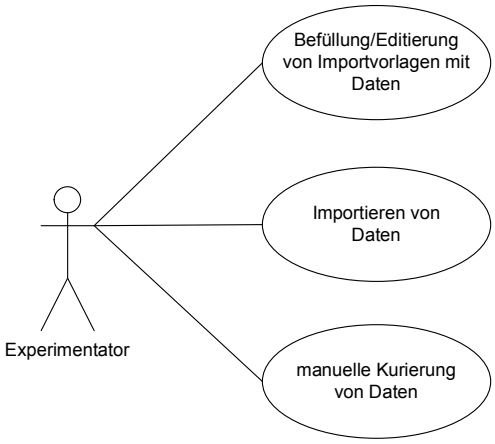


Abbildung 7.1: Allgemeine Anforderungen

nissen war es erforderlich, eine Vielzahl verschiedener Varianten zu betrachten, z. B. entweder Sommer- oder Winterformen, eine Auswahl nach Standortfaktoren der Versuchsanbauten, über Orte oder Jahre gemittelte Werte [CWL+08, MP07].

Dieses Vorgehen machte ein wiederholtes Neuimportieren von Subsets der Ausgangsdaten mit verändertem Datenumfang sowie die Aufbereitung des vorhandenen Datenmaterials nötig. Daher mussten Lade- und Bereinigungsprozeduren entwickelt werden, die flexible Adaptationsmöglichkeiten bieten. Die durchzuführenden Bereinigungsschritte werden im Detail in Abschnitt 7.3.4 besprochen.

Abbildung 7.2 fasst die Aufgaben der Integration zusammen.

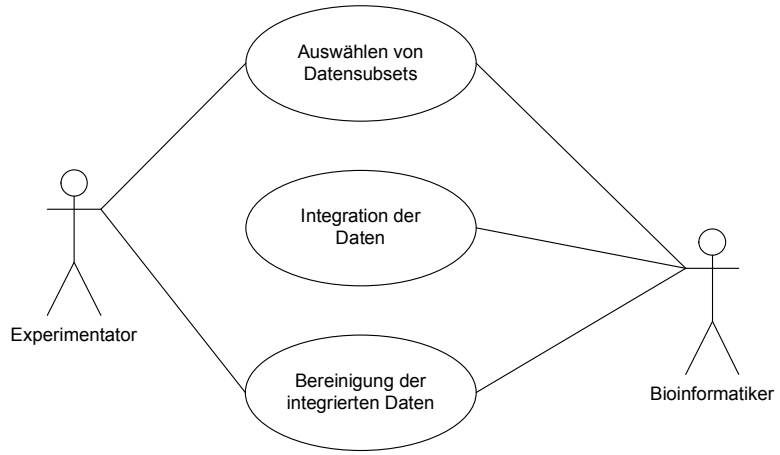


Abbildung 7.2: Anforderungen der Integration

7.2.3 Anforderungen zur Analyse

Assoziationsstudien sollten auf Basis der integrierten und bereinigten Daten mit dem in der Pflanzengenetik etablierten Softwarewerkzeug TASSEL[5] [BZK+07] durchgeführt werden.

TASSEL importiert phänotypische und genetische Daten in jeweils unterschiedlichen, proprietären Formaten. Dazu war es erforderlich, Prozeduren zu entwerfen, die es ermöglichen, generische Anfragen an den integrierten Datenbestand zusammenzustellen. Diese generisch erzeugten Anfragen sollten ausgeführt und die abgefragten Daten in von TASSEL lesbare Dateiformate exportiert werden. Hierbei waren fünf Typen von Datendateien zu unterscheiden:

- phänotypische Daten,
- Markerdaten,
- Daten über Populationsstrukturen (Q-Matrix),
- Daten über Sortenähnlichkeiten (Kinship-Matrix) sowie
- 1-2-Matrizen mit Haplotypenmustern.

Die letztgenannten 1-2-Matrizen bestehen aus Haplotyp- und Genotypbezeichnungen. Das Vorhandensein eines bestimmten Haplotypenmusters bei einem Genotypen wird in der Matrix mit 1, das Fehlen mit 2 gekennzeichnet.

Der Export von genetischen, phänotypischen und Haplotypendaten sollte jeweils für ein Gen oder Genfragment erfolgen. Nicht zu allen Genotypen, für die Markerdaten verfügbar sind, existieren phänotypische Daten. Dies trifft auch im umgekehrten Fall zu. Daher mussten die zu diesem Gen bzw. Genfragment gehörenden Markerdaten mit phänotypischen Daten abgeglichen werden. Damit sollte gewährleistet werden, dass nur der Export der Schnittmenge von Genotypen zugelassen wird, für die sowohl genetische als auch phänotypische Daten vorliegen. Hierbei war zu beachten, dass für die einzelnen Genotypen eine Vielzahl phänotypischer Messwerte existiert, die in verschiedenen Jahren, Regionen, Versuchen und auch bei sich unterscheidenden Bodenbeschaffenheiten (Standortgruppen) erfasst wurden. Für den Export war für diese Werte pro Genotyp und Merkmal je ein Mittelwert zu bilden. Welche Werte in den Export einfließen sollen, musste vom Anwender zu spezifizieren sein (z. B. nur Berücksichtigung von Genotypen, zu denen für ein bestimmtes phänotypisches Merkmal mindestens 80 Messwerte aus unterschiedlichen Jahren und verschiedenen Orten

[5] TASSEL-Version 2.0.1, http://www.maizegenetics.net/tassel [Stand 2009-04-09]

Deutschlands vorliegen). Es mussten daher geeignete Prozeduren entworfen werden, die interaktiv durch die Experimentatoren bedienbar sein sollten.

Bei der Durchführung von Assoziationsstudien sind zusätzliche Informationen zur Populationsstruktur sowie zur Sortenähnlichkeit in Form einer Q-Matrix und einer Kinship-Matrix zu berücksichtigen. Dies dient der Eliminierung von unspezifischen Ergebnissen. Die Generierung dieser Informationen sollte auf der Basis von SSR-Markerdaten erfolgen. Daraus sollten mit der Software STRUCTURE [FSP07] Q-Matrizen und mit der Software SPAGeDi [HV02] Kinship-Matrizen erstellt werden. Q-Matrizen bestehen aus Clusterinformationen, die Genotypen zu Gruppen zusammenzufassen. Kinship-Matrizen enthalten Ähnlichkeitskoeffizienten, welche Ausprägungen zwischen 0 und 1 annehmen und Verwandschaftsbeziehungen der untersuchten Sorten repräsentieren. Populationsstruktur und Verwandschaftsgrad des untersuchten Sortensets haben einen Einfluss auf die Assoziationsergebnisse.

In Abbildung 7.3 sind die Aufgaben der Analyse zusammengefasst.

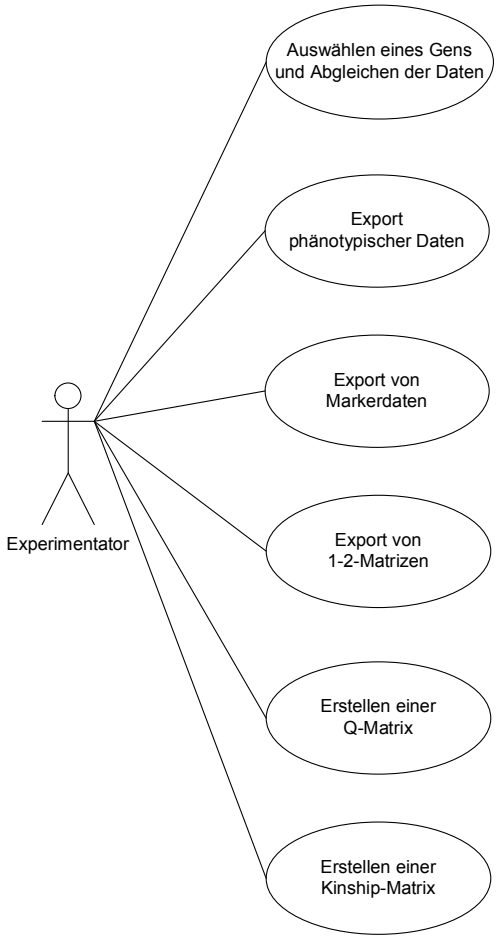


Abbildung 7.3: Anforderungen der Analyse

7.3 Prototyp

Die Umsetzung des Anwendungsfalles erfolgte im Rahmen eines Prototypen unter Verwendung des in Kapitel 6 beschriebenen Konzepts. Abbildung 7.4 zeigt die Nutzung dieses Konzepts für den Prototypen. Die einzelnen Schichten des Prototypen werden im Folgenden beschrieben.

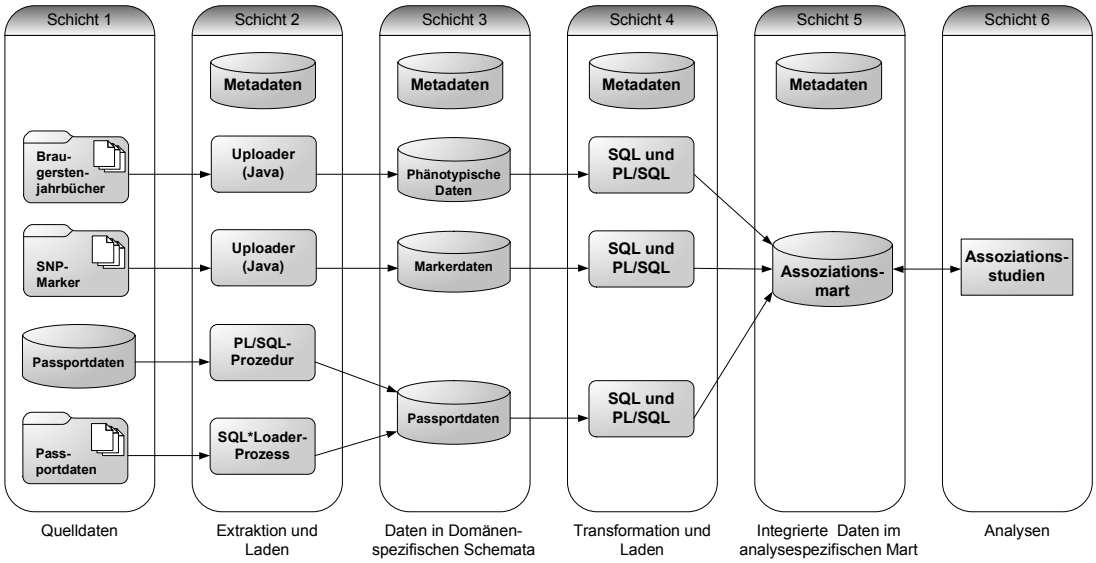


Abbildung 7.4: Anwendung des in Kapitel 6 entworfenen Konzepts

7.3.1 Schicht 1: Quelldaten

Die zur Bearbeitung des Anwendungsfalls erforderlichen Quelldaten waren phänotypische, Marker- und Passportdaten.

Phänotypische Daten

Die verwendeten phänotypischen Daten lagen überwiegend in Form statistischer Jahrbücher der Deutschen Braugerstengemeinschaft vor. Zur Erfassung dieser Daten wurde eine MS-Excel-Vorlage entwickelt. Die Daten wurden durch Experimentatoren und mit Hilfe studentischer Hilfskräfte mit dieser Vorlage erfasst. Abbildung 7.5 zeigt einen Ausschnitt einer solchen Datendatei. Insgesamt sind phänotypische Daten von ca. 170 Sorten aus Braugerstenversuchen, die in einem Zeitraum von 20 Jahren stattfanden, zusammengetragen worden. Diese Daten waren nicht in elektronischer Form verfügbar, wodurch die zeitintensive, manuelle Datenerfassung nicht vermieden werden konnte.

		Sorte	MQ-RP [%]	MQ-lösl. N [mg/100g MTrS]	MQ-ELG [%]	MQ-VZ 45 [%]	MQ-Vergärb. Extrakt [%]	MQ-Extr. [%]	MQ-Visk. [mPas]	MQ-Brab. [HE]	MQ-Friab. [%]	MQ-MQI
14	M	Alexis	10,1	672,0	41,7	45,9	84,8	81,8	1,5	96,0	86,7	8,7
18	M	Baronesse	10,1	570,0	35,5	38,0	82,0	79,9	1,5	125,0	64,7	5,2
19	M	Steffi	10,4	634,0	38,3	40,2	83,5	80,8	1,5	118,0	75,5	6,8
22	M	Krona	10,4	754,0	45,5	44,1	84,4	82,3	1,5	102,0	81,2	8,3
26	M	Sissy	10,5	804,0	48,2	48,9	83,7	81,8	1,5	109,0	79,1	8,3
31	M	Halla	10,4	729,0	43,5	52,9	83,2	81,8	1,5	124,0	74,0	8,3
32	M	Scarlett	10,1	698,0	43,5	49,6	83,9	82,4	1,5	116,0	75,3	8,5
33	M	Barke	10,3	657,0	40,3	45,8	84,2	82,9	1,5	103,0	82,6	8,8
34	M	Orthega	9,7	521,0	33,6	35,2	80,8	80,8	1,6	142,0	64,9	4,8
84	R	Thuringia	10,1	718,0	44,4	49,5	84,7	82,2	1,5	117,0	78,7	8,7
34	G	Brenda	10,3	732,0	44,8	44,5	84,4	82,7	1,5	102,0	80,1	8,5
41	G	Mentor	10,1	691,0	43,2	42,5	82,8	81,8	1,5	110,0	79,3	7,5
52	G	Maresi	10,3	740,0	45,0	44,6	84,1	82,0	1,5	108,0	80,0	8,1

Abbildung 7.5: Ausschnitt einer formatierten Inputdatei mit phänotypischen Daten (Schicht 1)

Grundsätzlich wäre das Einscannen der statistischen Jahrbücher verbunden mit automatischer Texterkennung besser geeignet, um Fehler bei der (manuellen) Dateneingabe zu vermeiden. Aufgrund der sehr heterogenen Strukturierungen der Jahrbücher wurde sich jedoch für die manuelle Erfassung der Daten in den oben beschriebenen Vorlagen entschieden. Die Korrektheit der so zusammengestellten Daten wurde durch umfangreiche Stichproben untersucht.

Markerdaten

Die zur Verfügung stehenden SNP- und INDEL-Informationen wurden am IPK in der Arbeitsgruppe Gen- und Genomkartierung generiert und von Experimentatoren um Metainformationen wie z. B. Markerart oder genetische Region angereichert. Abbildung 7.6 zeigt einen Ausschnitt einer solchen Datei. Zur besseren Unterscheidbarkeit wurden SNP-Ausprägungen und Haplotypen farbig markiert und Fehlstellen durch die Kodierung *-999* ersetzt. Insgesamt wurden für 19 Kandidatengene Markerdaten bei jeweils 450 – 700 Sorten und Zuchtstämmen erfasst und werden mit Daten zu weiteren Kandidatengenen laufend ergänzt.

Widersprüchliche Daten wie z. B. fehler- oder lückenhafte SNP- und INDEL-Informationen bei gleichen Sorten oder aufgetretene Sequenzierfehler wurden in Schicht 4 automatisch bereinigt.

Zusätzlich wurden alle Sorten in der Arbeitsgruppe Gen- und Genomkartierung mit 24 Mikrosatellitenmarkern (SSRs) untersucht [MWR09]. Diese sind repräsentativ über das gesamte Genom der Gerste verteilt (Tabelle 7.1) und sollten als Basis für die Generierung von Q- und Kinship-Matrizen verwendet werden.

lfd. Nr	Programm	Sorte	SNP1	SNP2	SNP4	3bp-INDEL	SNP5	Haplotyp
			Pyro	Pyro	Pyro		Pyro	
			Intron1	Exon2	Exon4	Exon4	Exon4	
				AGT = S	GTC = V	GCG = A	GCG = A	
			A/G	AGC = S	GTT = V	1=Ins 2=Del	GCA = A	
64	O	Accrue	G	T	T	2	G	H3_GM200_D
51	O	Action	G	C	T	2	G	H1_GM200_D
60	O	Actrice	G	C	T	2	G	H1_GM200_D
171	G	Adagio	A	C	C	1	A	H2_GM200_I
100	G	Adonis	G	C	C	1	-999	-999
92	O	Adonis	G	C	C	1	G	H4_GM200_I
176	G	Adour	G	C	T	2	G	H1_GM200_D
290	G	Ager	G	C	T	2	G	H1_GM200_D
324	G	Ager	G	C	T	2	G	H1_GM200_D
253	G	Agio	G	C	T	2	G	H1_GM200_D
29	O	Alexis	G	C	T	2	G	H1_GM200_D
1	G	Alexis	G	C	T	2	G	H1_GM200_D
93	O	Alissa	G	T	T	2	G	H3_GM200_D
92	G	Alissa	G	T	T	2	G	H3_GM200_D
94	O	Allegra	G	T	T	2	G	H3_GM200_D
93	G	Allegra	G	T	T	2	G	H3_GM200_D
159	G	Alliot	G	C	T	2	-999	-999
35	G	Alpaka	G	T	T	2	G	H3_GM200_D
188	G	Amarine	G	T	T	2	G	H3_GM200_D
276	G	Amsel	A	C	C	2	A	H2_GM200_D
58	O	Anaconda	G	C	T	2	G	H1_GM200_D
187	G	Angela	G	C	T	2	-999	-999
30	O	Angora	G	T	T	2	G	H3_GM200_D
7	G	Angora	G	T	T	2	G	H3_GM200_D

Abbildung 7.6: Ausschnitt einer formatierten Inputdatei mit Markerdaten sowie Metainformationen wie der Lokalisation im Gen oder der Translation in Aminosäuren (Schicht 1)

Tabelle 7.1: Genomweite Verteilung der 24 Mikrosatellitenmarker (SSRs) auf den Chromosomen der Gerste [Mat08]

1H	2H	3H	4H	5H	6H	7H
Bmag0211, Bmag0579, Bmag0718, HVM20	Bmag0518, Bmag0749, GBMS0160, GBMS0247, HVM36	Bmag0013, Bmag0225, Bmag0603	Ebmac0701, GBMS0087, HVM40	Ebmac0684, GBMS0032	Ebmac0602, GBMS0083, GBMS0125	Ebmac0755, GBMS0035, GBMS0111, GBMS0192

Passportdaten

Passportinformationen zu den untersuchten Gerstensorten stammten hauptsächlich von den Züchtern und wurden von Experimentatoren in einer durch Semikola separierten Datei zusammengetragen. Zusätzlich verwendete Daten befanden sich bereits in Datenbankmanagementsystemen. Dazu gehören Daten aus dem Genbankinformationssystem GBIS sowie der Barley Core Collection (vgl. Abschnitt 2.2.4).

7.3.2 Schicht 2: Extraktion und Laden

Phänotypische Daten

Zum Importieren der in Schicht 1 zusammengestellten Daten in das domänenspezifische Schema für phänotypische Daten in Schicht 3 wurden verschiedene Java-Klassen entwickelt. Diese Klassen sind über eine grafische Oberfläche auf Basis der Oracle-Application-Express-Technologie (APEX) [Ora09a] von Experimentatoren einsetzbar. Abbildung 7.7 zeigt einen Ausschnitt der Benutzeroberfläche.

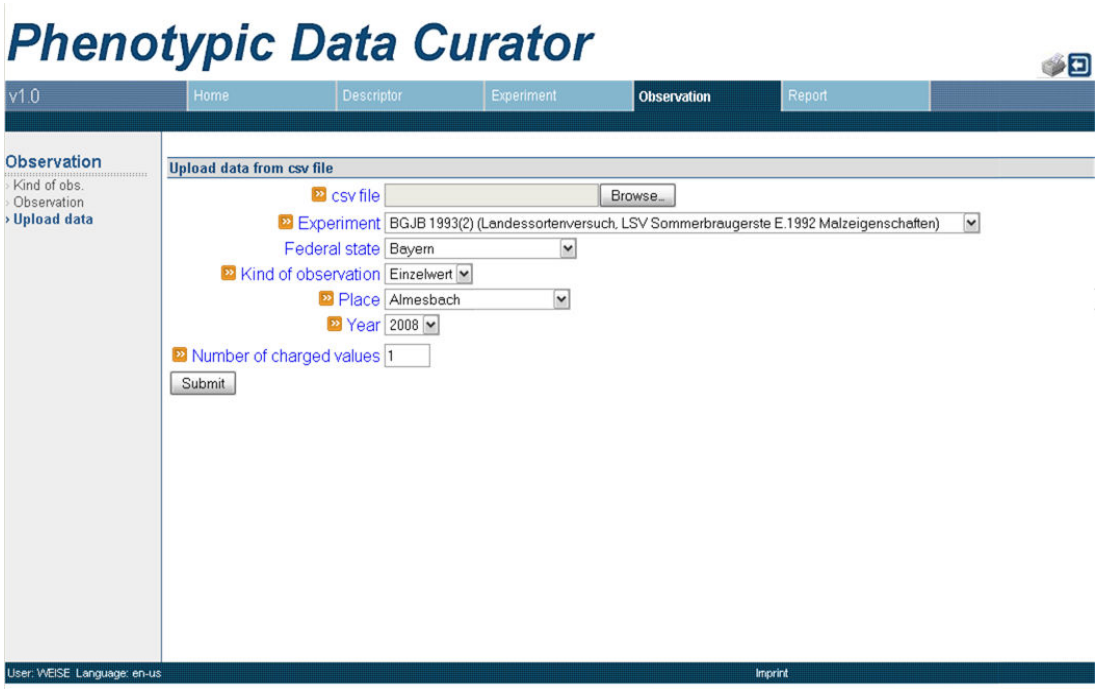


Abbildung 7.7: Screenshot der Datenimport-Applikation für phänotypische Daten (Schicht 2)

Während des Ladevorgangs wurden erste Datenüberprüfungen auf Basis von Metadaten automatisch durchgeführt. Damit konnte sichergestellt werden, dass nur Daten zu Merkmalen und Sorten erfasst wurden, die bereits in der Datenbank hinterlegt waren, und die bei der manuellen Erfassung aufgetretenen Schreibfehler entdeckt wurden.

Markerdaten

Für die Markerdaten wurde eine vergleichbare Lösung wie für die phänotypischen Daten unter Verwendung von Java-Klassen und einer APEX-basierten Oberfläche realisiert. In Abbildung 7.8 ist die Benutzeroberfläche dargestellt. Auch hier wurden während des Importvorgangs Überprüfungen durchgeführt (siehe oben).

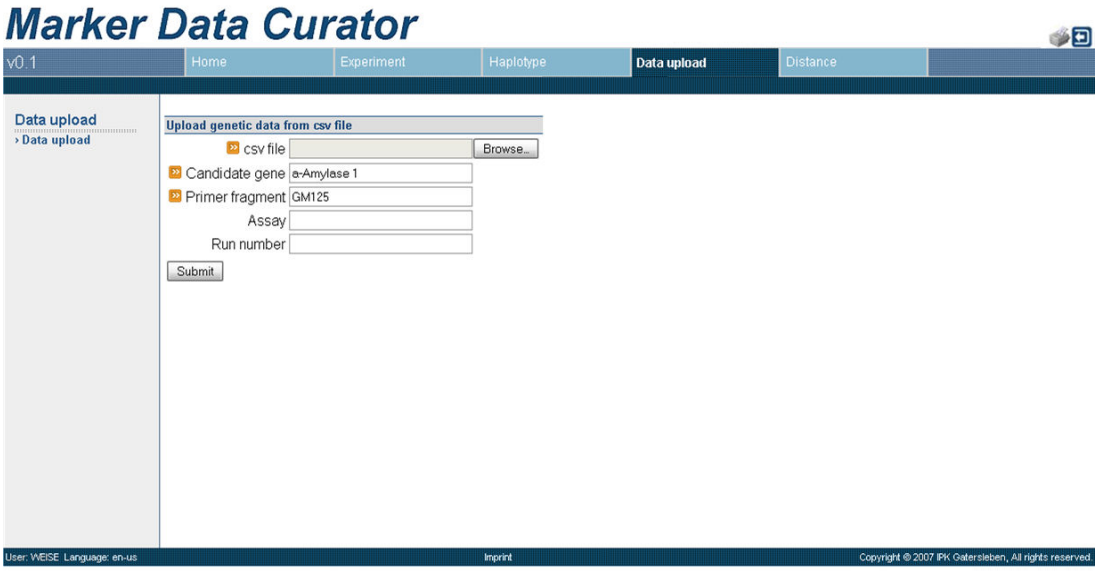


Abbildung 7.8: Screenshot der Datenimport-Applikation für Markerdaten (Schicht 2)

Passportdaten

Die zur Durchführung der Assoziationsstudien benötigten Passportdaten wurden durch die Experimentatoren in einer durch Semikola separierten Datei zusammengetragen, welche über einen Oracle-SQL*Loader-Prozess [Ora09b] importiert wurde. Die bereits in Datenbankmanagementsystemen vorliegenden Daten wurden über eine PL/SQL-Prozedur importiert.

7.3.3 Schicht 3: Datenpool

Um den für Schicht 3 propagierten Datenpool zu erstellen, wurden domänenspezifische Schemata für phänotypische und Markerdaten sowie für die Passportdaten als Bindeglied entworfen.

Phänotypische Daten

Das Schema für die phänotypischen Daten verwendet den Ansatz des so genannten Single-Observation-Konzepts [vHH92, MPL01]. Dabei wird jede Merkmalsausprägung als eine Beobachtung unter bestimmten Bedingungen betrachtet. Bedingungen sind beispielsweise das beobachtete Merkmal selbst, die zur Beobachtung angewandte Methode, das Experiment, innerhalb dessen die Beobachtung stattgefunden hat, oder das Entwicklungsstadium, in dem sich die Pflanze zum Beobachtungszeitraum befand.

Die Bedingungen sind in eigenen Datenbanktabellen modelliert. Ihre Kombination erfolgt völlig flexibel über Fremdschlüssel beim Abspeichern der beobachteten Merkmalsausprägungen. Das Single-Observation-Konzept ist eine Spezialisierung des EAV-Ansatzes.

Die Struktur zum Abspeichern der phänotypischen Daten wurde vereinfacht bereits in [WKV+06] dargestellt.

Markerdaten

Eine vergleichbare Struktur auf Basis des EAV-Ansatzes wurde auch zur Speicherung der Markerdaten entwickelt. Hierüber erfolgt die Verwaltung von SNP- und SSR-Markerdaten sowie INDEL-Informationen.

Passportdaten

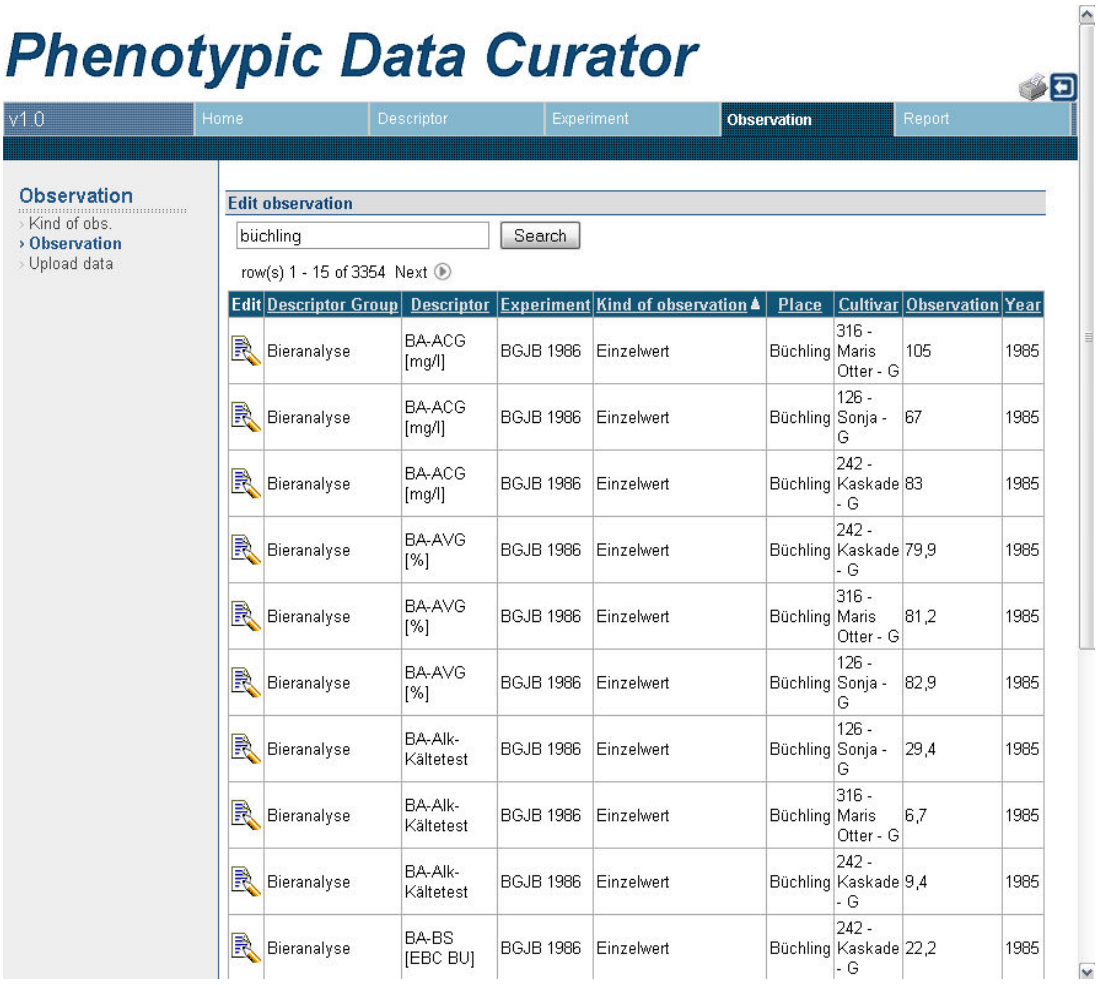
Im Vergleich zu den generischen Schemata für phänotypische und Markerdaten sind die im Kontext der Passportdaten erforderlichen Attribute eindeutig definiert. Hier war daher der Entwurf eines einfach strukturierten und statischen Schemas ausreichend.

Wie in Abschnitt 7.2.1 gefordert, wurden Werkzeuge zur manuellen Kuration der in den Datenpool importierten Informationen entwickelt. Abbildung 7.9 zeigt einen Screenshot des Kurationswerkzeuges für phänotypische Daten. Weitere Screenshots sind in Anhang A abgebildet. Diese Werkzeuge wurden zur Überprüfung der importierten Daten anhand von Stichproben verwendet.

7.3.4 Schicht 4: Transformation und Laden

Um Daten aus den eben beschriebenen domänenspezifischen Schemata in den Assoziationsmart (siehe Abschnitt 7.3.5) zu laden, wurden SQL-Anweisungen und PL/SQL-Prozeduren entwickelt. Der Entwurf erfolgte dahingehend, dass Anpassungen leicht durchführbar sind. Dies ist notwendig, um eine Vielzahl von individuellen Teildatenmengen in den Mart zu laden. In diesem Schritt wurde die eigentliche Integration der phänotypischen und der Markerdaten ausgeführt.

Um nach erfolgreichem Laden die Datenbereinigung durchführen zu können, wurden verschiedene PL/SQL-Funktionen und -Prozeduren entworfen. Das Vorgehen wird im Folgenden beschrieben.



Edit	Descriptor Group	Descriptor	Experiment	Kind of observation	Place	Cultivar	Observation	Year
	Bieranalyse	BA-ACG [mg/l]	BGJB 1986	Einzelwert	Büchling	316 - Maris Otter - G	105	1985
	Bieranalyse	BA-ACG [mg/l]	BGJB 1986	Einzelwert	Büchling	126 - Sonja - G	67	1985
	Bieranalyse	BA-ACG [mg/l]	BGJB 1986	Einzelwert	Büchling	242 - Kaskade - G	83	1985
	Bieranalyse	BA-AVG [%]	BGJB 1986	Einzelwert	Büchling	242 - Kaskade - G	79,9	1985
	Bieranalyse	BA-AVG [%]	BGJB 1986	Einzelwert	Büchling	316 - Maris Otter - G	81,2	1985
	Bieranalyse	BA-AVG [%]	BGJB 1986	Einzelwert	Büchling	126 - Sonja - G	82,9	1985
	Bieranalyse	BA-Alk-Kältetest	BGJB 1986	Einzelwert	Büchling	126 - Sonja - G	29,4	1985
	Bieranalyse	BA-Alk-Kältetest	BGJB 1986	Einzelwert	Büchling	316 - Maris Otter - G	6,7	1985
	Bieranalyse	BA-Alk-Kältetest	BGJB 1986	Einzelwert	Büchling	242 - Kaskade - G	9,4	1985
	Bieranalyse	BA-BS [EBC BU]	BGJB 1986	Einzelwert	Büchling	242 - Kaskade - G	22,2	1985

Abbildung 7.9: Verwaltung von Merkmalsausprägungen mit dem Kurationswerkzeug für phänotypische Daten (Schicht 3)

Um eine ausreißerbedingte Veränderung des Analyseergebnisses zu vermeiden, mussten die im Rahmen des Anwendungsfalls genutzten phänotypischen Daten dahingehend untersucht und behandelt werden. Hierzu erfolgte ein Entfernen (Trimming) von Merkmalsausprägungen von Sorten, die in mehreren Jahren und an unterschiedlichen Standorten evaluiert wurden, wenn die Abweichung vom Mittelwert für dieses Merkmal für die entsprechende Sorte mehr als 20% betrug. Dieser Grenzwert wurde zusammen mit den Experimentatoren festgelegt. Als Alternative zu diesem Vorgehen wurde eine Prozedur zur Ersetzung von Ausreißerwerten (Winsorising) entwickelt. Dabei werden Ausreißer durch die nächsthöheren bzw. nächstniedrigeren beobachteten Werte ersetzt.

Bei den Markeranalysen wurden stichprobenweise Sortenduplikate mit getestet, wodurch eine interne Verifikation bzgl. der korrekten SNP- und INDEL-Muster je Sorte

gewährleistet wurde. Außerdem traten auch verschiedene Sorten mit dem gleichen Namen, aber unterschiedlichen Eltern auf. Dies spiegelte sich auch in nicht-identischen Haplotypenmustern wider. Bei diesen war keine eindeutige Zuordnung zu den entsprechenden phänotypischen Daten möglich.

Es wurden daher Prozeduren entwickelt, die SNP-Markerinformationen auf Basis des Sortennamens vergleichen. Trat eine Sorte mehrfach auf, erfolgte ein Vergleich aller verfügbaren Markerdaten dieser Sorte. Bei Unterschieden an einer Markerposition wurden alle Daten dieser Position durch die Kodierung *-999* ersetzt, die von der Software TASSEL als Fehlstelle interpretiert wird. Waren alle Positionen unterschiedlich, erfolgte die Entfernung der betreffenden Sorte. Sorten mit unterschiedlicher Herkunft und Stammbaum, aber gleichem Zulassungsnamen wurden ebenfalls eliminiert.

Nicht eindeutig auszuwertende oder fehlende Markerdaten wurden durch die Kodierung *-999* ersetzt. Dies betrifft insbesondere Sequenzierfehler, d. h. Markerpositionen, denen nicht eindeutig ein Nukleotid wie C oder T zugeordnet werden konnte.

In Anhang B.1 werden die für die Bereinigung der Markerdaten entworfenen Funktionen und Prozeduren beschrieben.

7.3.5 Schicht 5: Analysespezifischer Datamart

Zur Verwaltung der integrierten Daten und zur Erfüllung der in Abschnitt 7.2 beschriebenen Anforderungen wurde ein Datamart-Schema entworfen. Dieses hält die in Schicht 4 bereinigten Daten. Zur Erhöhung der Zugriffsgeschwindigkeit wurden die in diesem Schema gespeicherten Daten teilweise denormalisiert. Abbildung A.3 in Anhang A zeigt das Datenbankschema des Assoziationsmarts.

7.3.6 Schicht 6: Analyse

Zur Durchführung von Assoziationsstudien sollte das etablierte Softwarewerkzeug TASSEL verwendet werden. Hierzu mussten sowohl phänotypische Daten als auch Markerdaten und Haplotypenmuster in Form von 1-2-Matrizen in Dateien mit einem von TASSEL lesbaren Format exportiert werden (Abschnitt 7.2.3).

Dafür war es erforderlich, verschiedene PL/SQL-Funktionen und -Prozeduren zu entwickeln, die es Experimentatoren ermöglichen, das entsprechende Kandidatengen und das zugehörige Genfragment (z. B. Proteindisulfidisomerase, GM082), die Herkunft der Gerstensorten sowie die zu berücksichtigenden Daten zu spezifizieren.

Auf Basis dieser Spezifikation wurde ein Abgleich von phänotypischen und Markerdaten durchgeführt. Damit konnte die ausschließliche Verwendung von Sorten sichergestellt werden, für die Daten beider Domänen vorlagen. Dies war notwendig, weil die verwendeten Daten aus verschiedenen Quellen stammen und nicht orthogonal sind. Das bedeutet, dass nicht zu allen Sorten, zu denen genetische Daten vorliegen, auch phänotypische existieren und umgekehrt. Des Weiteren wurden die phänotypischen Daten im Verlauf von 20 Jahren erfasst. Dabei wurden nicht in jedem Jahr Daten zu denselben Sorten erhoben. Um die Anzahl der Fehlstellen in den Datensätzen für die folgenden Assoziationsstudien zu verringern, war es daher erforderlich, genetische und phänotypische Daten abzugleichen.

Fehlstellen der Markerdaten wurden bereits während der Bereinigung in Schicht 4 mit der Kodierung *-999* ersetzt. Das eventuelle Nichtvorhandensein von phänotypischen Daten wird erst während des jeweiligen Abgleichs mit den Markerdaten für ein Gen oder Genfragment offenbar. Daher erfolgte eine Ersetzung mit *-999* erst während der Ausführung der PL/SQL-Prozeduren. Als Alternative zur bloßen Markierung als Fehlstelle wurde eine weitere Prozedur entwickelt, die Fehlstellen mit verschiedenen Methoden ersetzt (Value Imputation), z. B. durch den Mittelwert für das jeweilige Merkmal.

Bei der Generierung von 1-2-Matrizen für die Haplotypenmuster von Kandidatengenen, die das Vorhandensein eines Haplotypenmusters mit 1 und das Fehlen mit 2 kodieren, wurden nur Gerstensorten berücksichtigt, die vollständige Markerdatensätze aufweisen, da sonst aufgrund fehlender Markerdaten der entsprechende Haplotyp nicht generiert werden kann.

Die für den Abgleich und Export von phänotypischen, Marker- und Haplotypendaten entwickelten Funktionen und Prozeduren werden in Anhang B.2 beschrieben.

Der Export der abgeglichenen Daten erfolgte in speziellen, von TASSEL lesbaren, proprietären Dateiformaten. Abbildung 7.10 zeigt die Darstellung von Haplotypenmustern mit der auf dem Assoziationsmart basierenden Anwendung. Weitere Abbildungen dieser Applikation werden in Anhang A gezeigt.

Die oben beschriebenen Prozeduren bildeten ebenfalls die Grundlage zum Export der im Assoziationsmart enthaltenen SSR-Markerdaten. Basierend auf diesen Daten wurden mittels der Software STRUCTURE Q-Matrizen generiert und Kinship-Matrizen wurden entsprechend mittels der Software SPAGeDi erstellt.

Die exportierten Daten konnten danach durch das Softwarepaket TASSEL geladen und zur Berechnung von Assoziationen genutzt werden. Abbildung 7.11 zeigt die entsprechende Verwendung dieses Werkzeugs am Beispiel von Daten zum Gen Proteindisul-

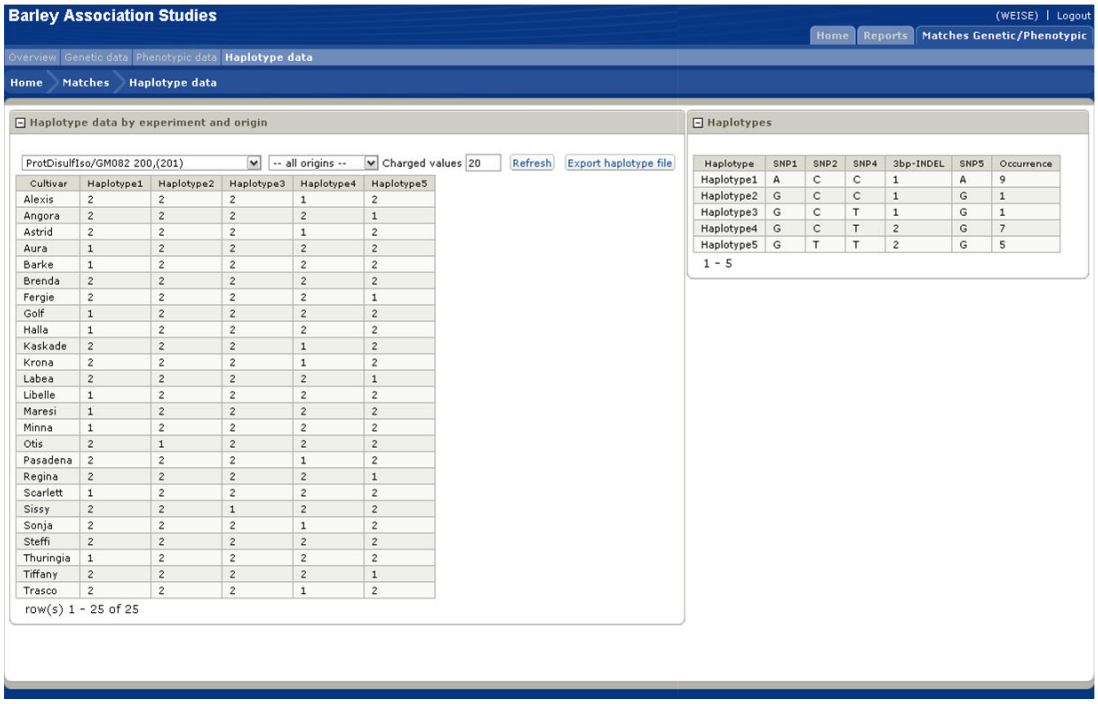


Abbildung 7.10: Anzeige von Haplotypenmustern mit der Assoziationsmart-Anwendung (Schicht 6)

fidisomerase; es wird das Ergebnis einer Assoziationsberechnung unter Berücksichtigung der Populationsstruktur dargestellt.

7.4 Einschätzung des Prototypen

Die Evaluation des vorgestellten Prototypen fand im Rahmen seiner Verwendung im Projekt GABI-MALT statt. Dabei konnte festgestellt werden, dass der Prototyp eine deutliche Reduzierung von Arbeitszeit bei gleichzeitig gesteigerter Qualität der zu analysierenden Daten ermöglicht hat. Zusätzlich konnten bei der manuellen Handhabung/Generierung auftretende Fehler durch die Automatisierung vollständig vermieden werden.

Im Folgenden soll zuerst der Zeitgewinn bei Einsatz des Prototypen gegenüber dem manuellen Vorgehen am Beispiel von fünf Subsets für Daten über 19 Kandidatengene veranschaulicht werden. Im Anschluss wird die Verbesserung der Datenqualität mit Hilfe verschiedener Beispiele verdeutlicht.

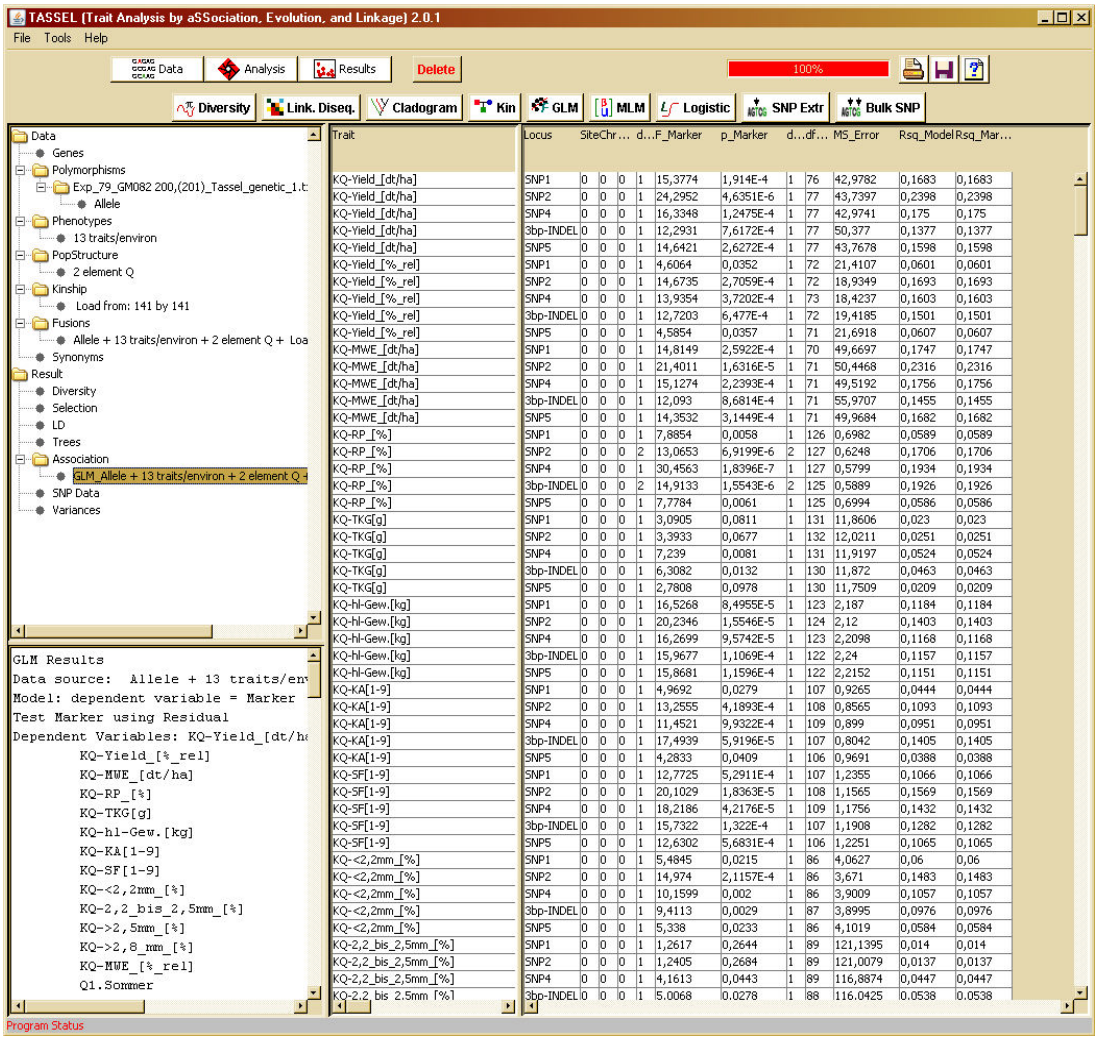


Abbildung 7.11: Verwendung des Softwarewerkzeugs TASSEL [BZK+07] (Schicht 6). Dargestellt sind die Ergebnisse einer Assoziationsstudie.

7.4.1 Zeitbedarf

Im Rahmen des hier beschriebenen Anwendungsfalls wurden Assoziationen mit verschiedenen Datensets gerechnet. Diese umfassten

- alle Sorten,
- alle Sommergersten,
- alle Wintergersten,

- alle zweizeiligen Gersten,
- alle sechszeiligen Gersten

sowie Kombinationen zwischen diesen Datensets (z. B. zweizeilige Sommergersten, sechszeilige Wintergersten) mit und ohne Berücksichtigung der Populationsstruktur.

Die Verrechnung verschiedener Datensets erfolgte einerseits mit dem Ziel, falschpositive Ergebnisse sukzessive auszuschließen (Abschnitt 7.2.2) und andererseits, um die Subsets zu identifizieren, in denen mit diagnostischen Markern auf positive Merkmale selektiert werden kann [CWL+08, MP07].

Es wurden zwei Abschätzungen zum Zeitaufwand durchgeführt:

1. Aufwand der manuellen Berechnung von Assoziationen (Fall 1),
2. Aufwand der Berechnung mit dem Prototypen (Fall 2).

Der Aufwand wird in benötigten Arbeitstagen im Umfang von 8h/Tag angegeben. Verschiedene Arbeitsschritte wie z. B. das Erfassen von Daten sind bei den beiden Vorgehensweisen identisch und werden daher hier nicht weiter betrachtet. Andere Arbeitsschritte, beispielsweise das Bereinigen von Daten oder das Abgleichen genetischer und phänotypischer Informationen, wurden manuell nur stichprobenweise durchgeführt. Der dabei ermittelte Aufwand wurde in Abhängigkeit von der Zahl der zu untersuchenden Kandidatengene etc. hochgerechnet.

Zeitbedarf bei manueller Durchführung von Assoziationsstudien (Fall 1)

1. **Verifizieren der Daten:**

   Die erfassten phänotypischen und genetischen Daten mussten auf Korrektheit überprüft werden. Hierbei lag das Hauptaugenmerk auf der Eliminierung von Duplikaten und Widersprüchen innerhalb der genetischen Daten. [32 Tage]

2. **Abgleichen genetischer und phänotypischer Daten pro Subset:**

   Die genetischen und phänotypischen Daten des Anwendungsfalls stammen aus verschiedenen Quellen und sind nicht orthogonal; nicht zu allen Sorten liegen genetische und phänotypische Daten gleichermaßen vor. Die phänotypischen Daten sind von der Verfügbarkeit der jeweiligen Sorte im betrachteten Zeitraum von 20 Jahren abhängig. Zur Verringerung der Anzahl der Fehlstellen war ein Abgleich von genetischen und phänotypischen Daten notwendig. Dieser Sortenabgleich erfolgte für jedes bei der Assoziationsstudie betrachtete Kandidatengen. [14 Tage]

3. **Erstellen von Input-Dateien mit genetischen Daten pro Subset:**

   Für die Nutzung des Softwarewerkzeuges TASSEL war es erforderlich, Inputdateien für genetische Daten in einem speziellen, proprietären Format zu erstellen. [4 Tage]

4. **Erstellen von Input-Dateien mit phänotypischen Daten pro Subset:**

   Dasselbe Vorgehen war für das Erstellen von Input-Dateien mit phänotypischen Daten notwendig. [4 Tage]

5. **Erstellen von Input-Dateien mit Daten über Haplotypenmuster pro Subset:**

   Weiterhin mussten Input-Dateien mit Informationen über Haplotypenmuster erstellt werden. Hierbei wurden für jedes Gen alle auftretenden Haplotypenmuster bestimmt. Danach erfolgte für jede Sorte die Kodierung des Vorhandenseins eines bestimmten Haplotypenmusters mit 1 und des Fehlens mit 2. [4 Tage]

Der Aufwand für die anschließende Berechnung von Assoziationen mit dem Softwarewerkzeug TASSEL ist beim manuellen Vorgehen und der Verwendung des Prototypen identisch und soll daher in beiden Fällen nicht in die Berechnung des Aufwandes einfließen. Dieselbe vereinfachende Annahme soll auch auf die Berechnung von Q- und Kinship-Matrizen mittels bestehender Softwarewerkzeuge Anwendung finden.

Schritt 1 musste einmal ausgeführt werden. Die Schritte 2–5 beziehen sich auf den Abgleich von genetischen Daten zu 19 Kandidatengenen mit phänotypischen Daten einschließlich des Erstellens der beschriebenen Input-Dateien und mussten für jedes der fünf Subsets wiederholt werden.

Zeitbedarf bei Durchführung von Assoziationsstudien mit dem Prototypen (Fall 2)

1. **Entwicklung des Prototypen:**

   Zur Automatisierung der nachfolgenden Schritte wurde ein Prototyp auf Basis des in Kapitel 6 entwickelten Vorgehensmodells implementiert. [60 Tage]

2. **Laden von Daten in den Assoziationsmart pro Subset:**

   Das Laden der importierten Daten in den Assoziationsmart erfolgte semi-automatisch über vorgefertigte Prozeduren bzw. Abfragen. Dabei wurde automatisch eine Fehlerbereinigung durchgeführt. Dies betrifft u. a. die Beseitigung von Widersprüchen in den Markerausprägungen sowie die Ausreißer- und Fehlstellenbehandlung bei phänotypischen Daten. Hierauf wird noch einmal detailiert in Abschnitt 7.4.2 eingegangen. [10 min $\approx$ 0,02 Tage bei 8 h/Tag]

3. **Abgleichen genetischer und phänotypischer Daten pro Subset:**

   Wie oben dargestellt, stammen die genetischen und phänotypischen Daten des Anwendungsfalls aus verschiedenen Quellen und sind nicht orthogonal. Daher mussten genetische und phänotypische Daten anhand der Sorten abgeglichen werden. Dieser Abgleich erfolgte in Abhängigkeit eines vom Nutzer ausgewählten Gens sowie phänotypischer Merkmale automatisch. [10 min $\approx$ 0,02 Tage bei 8 h/Tag]

4. **Erstellen von Input-Dateien mit genetischen Daten pro Subset:**

   Das Erstellen von Input-Dateien mit genetischen Daten für das Softwarewerkzeug TASSEL erfolgte auf Basis der abgeglichenen Daten automatisch. [10 min $\approx$ 0,02 Tage bei 8 h/Tag]

5. **Erstellen von Input-Dateien mit phänotypischen Daten pro Subset:**

   Das Erstellen von Input-Dateien mit phänotypischen Daten erfolgte ebenfalls automatisch. Dabei war ein vergleichbarer zeitlicher Aufwand zu verzeichnen. [10 min $\approx$ 0,02 Tage bei 8 h/Tag]

6. **Erstellen von Input-Dateien mit Daten über Haplotypenmuster pro Subset:**

   In gleicher Weise wurden Input-Dateien mit Daten über Haplotypenmuster erstellt. [10 min $\approx$ 0,02 Tage bei 8 h/Tag]

Wie oben festgestellt, werden Assoziationen mit dem Softwarewerkzeug TASSEL gerechnet. Daher ist der Aufwand für das Berechnen von Assoziationen sowohl beim manuellen Vorgehen als auch bei Verwendung des Prototypen identisch und soll in beiden Fällen nicht in den Vergleich einfließen.

Auch im Fall des Prototypen mussten die Schritte 2–6 mehrfach durchgeführt werden. Hierzu wurden die fünf oben genannten Subsets und Daten für 19 Kandidatengene verwendet. Im Gegensatz zur manuellen Durchführung wurde auch der Schritt der Datenverifizierung mehrfach ausgeführt. Dies war möglich, weil die Verifizierung in Vorbereitung auf die Verrechnung von Daten über 36 weitere Kandidatengene vollständig automatisiert wurde. Der Aufwand für die Entwicklung des Prototypen (60 Tage) fiel nur einmalig an.

Beim manuellen Vorgehen nimmt die benötigte Arbeitszeit mit jedem weiteren Subset deutlich zu (Summe: 162 Tage), während sich der Gesamtaufwand bei Verwendung des Prototypen nur sehr geringfügig verändert (Summe: 60,5 Tage). Tabelle 7.2 fasst die Ergebnisse beider Aufwandsschätzungen zusammen.

Zusätzlich zu den bisher betrachteten 19 Kandidatengenen sind Daten zu 36 weiteren Genen verfügbar, mit denen Assoziationsstudien durchgeführt werden sollen. Zur Verwaltung dieser Daten und zur Berechnung von Assoziationen im Rahmen des nach

Tabelle 7.2: Zusammenfassung der Aufwandsschätzungen für das manuelle Vorgehen und die Nutzung des Prototypen. Neben dem variablen Aufwand für ein einzelnes Subset ist auch jeweils die Summe des Aufwands für die fünf verwendeten Subsets alle Sorten, alle Sommergersten, alle Wintergersten, alle zweizeiligen Gersten und alle sechszeiligen Gersten angegeben. Der fixe Aufwand bezieht sich im Fall des manuellen Vorgehens auf das Verifizieren von Daten (32 Tage) und bei der Nutzung des Prototypen auf dessen Entwicklung (60 Tage). Hierbei ist hervorzuheben, dass sich der Aufwand für das Verifizieren von Daten erhöht, wenn die verwendete Datenbasis um neue phänotypische und/oder genetische Daten erweitert wird. Bei Verwendung des Prototypen sind keine Anpassungen erforderlich; der Aufwand wird daher für diesen Fall konstant bleiben.

Arbeitsschritt	Aufwand [Tage]		Aufwand für 5 Subsets [Tage]	
	manuell	Prototyp	manuell	Prototyp
Verifizieren der Daten	32	-	32	-
Entwicklung des Prototypen	-	60	-	60
Summe des fixen Aufwands [Tage]	*32*	*60*	*32*	*60*
Laden in den Assoziationsmart	-	0,02	-	0,1
Abgleichen genetischer und phänotypischer Daten	14	0,02	70	0,1
Erstellen von Input-Dateien mit genetischen Daten	4	0,02	20	0,1
Erstellen von Input-Dateien mit phänotypischen Daten	4	0,02	20	0,1
Erstellen von Input-Dateien mit Daten über Haplotypenmuster	4	0,02	20	0,1
Summe des variablen Aufwands [Tage]	*26*	*0,1*	*130*	*0,5*
Summe gesamt [Tage]	***58***	***60,1***	***162***	***60,5***

GABI-MALT anschließenden Forschungsprojektes GABI-GENOBAR[6] soll der Einsatz des Prototypen erfolgen. Hierfür müssen die genetischen Daten um die Informationen für die 36 Gene erweitert werden. Der Import weiterer phänotypischer Daten ist bereits abgeschlossen.

[6]GABI-GENOBAR: A genome wide approach to associate genetic diversity to agronomically important traits in barley, http://www.gabi.de/projekte-alle-projekte-neue-seite-170.php [Stand 2009-04-02]

Mit Hilfe dieses Vergleichs konnte gezeigt werden, dass der Prototyp bei der Verwaltung und Vorbereitung von großen Datensätzen für Assoziationsstudien zu einer deutlichen Zeitersparnis geführt hat und, insbesondere im Hinblick auf die weiteren zu analysierenden Daten, eine enorme Steigerung der Arbeitseffizienz ermöglicht.

7.4.2 Erhöhung der Datenqualität

Im Folgenden soll die mit dem Einsatz des Prototypen erfolgte Verbesserung der Datenqualität an verschiedenen Beispielen verdeutlicht werden. Hierzu wurde ein Subset mit Daten über 58 zweizeilige Sommergersten in den Assoziationsmart geladen. Es umfasste genetische Informationen zu 19 Kandidatengenen sowie phänotypische Daten zu 57 Merkmalen.

Wie in Abschnitt 7.3.4 beschrieben, wurden die zur Genotypisierung verwendeten Gerstensorten aus verschiedenen Quellen bezogen. Darunter befand sich eine größere Anzahl von Sortenduplikaten. Dies diente zum einen der internen Verifikation der jeweiligen Sorte. Zum anderen traten verschiedene Sorten mit identischem Namen, aber unterschiedlichen Elternsorten auf. Dies spiegelte sich in widersprüchlichen SNP- und Haplotypenmustern wider. Abbildung 7.12 zeigt am Beispiel des Kandidatengens α-Amylase 1 verschiedene Sorten, deren Nukleotidausprägungen sich für einzelne SNP-Marker unterscheiden.

Zur Bereinigung dieser Daten kamen verschiedene Prozeduren zum Einsatz, die für jedes Kandidatengen Sortenduplikate ermittelten. Dabei wurden die Markerausprägungen positionsweise verglichen. Bei Übereinstimmung wurden Duplikate eliminiert, bei Widersprüchen alle Daten der jeweiligen Position für die entsprechende Sorte entfernt. Die dadurch auftretenden Fehlstellen wurden mit *-999* kodiert. Im Beispiel des Kandidatengens α-Amylase 1 wurden 24 Widersprüche bereinigt und insgesamt 61 Duplikate entfernt.

Wie in Abschnitt 7.2.3 beschrieben, sollten die phänotypischen Daten zur Assoziationsberechnung pro Sorte und Merkmal zu einem Mittelwert aggregiert werden. Zur Beschleunigung der Abfragen des Assoziationsmarts wurden diese Aggregationen während des Ladens (Schicht 4) vorberechnet und materialisiert. Hierbei wurden Merkmalsausprägungen von Sorten, die an verschiedenen Standorten und/oder in mehreren Jahren untersucht wurden, als Ausreißer behandelt, wenn die Abweichung vom Mittelwert dieses Merkmals für die entsprechende Sorte mehr als 20% betrug. Auf diesen Grenzwert wurde sich in Absprache mit den Experimentatoren geeinigt (vgl. Abschnitt 7.3.4). Ausreißer wurden entweder entfernt (Trimming) oder durch den nächsthöheren bzw. nächstniedrigeren Wert ersetzt (Winsorising). Im Fall des Subsets der zweizeiligen Sommergersten fand die Behandlung von 54 Ausreißern statt.

Sorte	SNP-Bezeichnung	Nukleotid
Bolivia	SNP1	G
Bolivia	SNP3	G
Bolivia	SNP4	C
Bolivia	SNP4	G
Bolivia	SNP5	A
Bolivia	SNP6	C
Danuta	SNP3	G
Danuta	SNP4	G
Danuta	SNP4	C
Danuta	SNP5	A
Danuta	SNP6	C
Delta	SNP1	C
Delta	SNP3	G
Delta	SNP4	G
Delta	SNP5	A
Delta	SNP5	G
Delta	SNP6	C
Delta	SNP6	G
Extract	SNP1	C
Extract	SNP3	A
Extract	SNP4	G
Extract	SNP5	G
Extract	SNP6	G
Extract	SNP6	C
Havanna	SNP1	G
Havanna	SNP3	G
Havanna	SNP4	C
Havanna	SNP5	A
Havanna	SNP6	C

Abbildung 7.12: SNP-Markerausprägungen am Beispiel des Gens α-Amylase 1. Widersprüche wurden farblich hervorgehoben

Da die verwendeten phänotypischen Daten nicht orthogonal sind (vgl. Abschnitt 7.3.6), trat weiterhin eine größere Anzahl von Fehlstellen auf, die vor der Verwendung des Softwarewerkzeugs Tassel zu behandeln waren. Hierfür wurden Prozeduren entwickelt, die Fehlstellen entweder mit *-999* kodierten oder Ersetzungen (Value Imputation), beispielsweise mit dem Mittelwert, durchführten. Im Subset der zweizeiligen Sommergersten wurden 864 Fehlstellen auf diese Art behandelt (26% der aggregierten Daten bei 57 Merkmalen und 58 Sorten). Tabelle 7.3 fasst die hier beschriebenen Schritte zusammen.

Tabelle 7.3: Schritte zur Erhöhung der Qualität von Daten zur Berechnung von Assoziationen für das Kandidatengen α-Amylase 1

Vorgehen	bereinigte Datenpunkte
Entfernung von Duplikaten	61
Bereinigung von Widersprüchen	24
Behandlung von Ausreißern	54
Behandlung von Fehlstellen	864

Anhand dieser Beispiele wurde gezeigt, dass der entwickelte Prototyp neben einer deutlichen Zeitersparnis einen wichtigen Beitrag zur Vorverarbeitung von genetischen und phänotypischen Daten für Assoziationsstudien geleistet hat. Durch die Automatisierung aller Arbeitsschritte vom Laden über das Bereinigen bis hin zum Exportieren von Daten für externe Softwarewerkzeuge können bei der manuellen Handhabung auftretende Fehler vermieden werden und es wird eine gleichbleibend hohe Datenqualität garantiert. Somit ist es möglich, Assoziationsstudien im größeren Maßstab ohne erhöhten Personalaufwand durchzuführen.

7.4.3 Ergebnisse

Ergebnisse von Assoziationsstudien für Gerste und Malzqualität liegen für eine Reihe von Kandidatengenen vor. Zwei Artikel wurden hierzu bereits veröffentlicht [MWR09, MWFR09]. Tabelle 7.4 stellt ausschnittsweise Ergebnisse aus [MWR09] vor. Darin wurden für das auf Chromosom 6H der Gerste kodierende Gen für die α-Amylase 1 (*amy1*) sechs SNP-Marker, die vier Haplotypen definieren, ermittelt. Die Berechnung von Assoziationen dieser SNPs und Haplotypen mit 14 Malzqualitätsmerkmalen erfolgte mit Daten über 117 Europäische Sommer- und Wintergerstensorten. Dabei wurden die Subpopulationen der Winter- und Sommerformen getrennt betrachtet und bei der Gesamtpopulation die Populationsstruktur berücksichtigt.

Tabelle 7.4: Auswahl signifikanter Marker-Merkmals-Beziehungen (gekennzeichnet durch x) für die Malzqualitätsparameter Kochfarbe und Friabilimeter und sechs SNP-Markern aus jeweils 72 Winter- und 45 zweizeiligen Sommergerstensorten unter Berücksichtigung der Populationsstruktur bei der α-Amylase 1 [MWR09]

Merkmal	SNP	Winter	Sommer (nur zweizeilig)
Kochfarbe [EBC]	SNP1	x	x
	SNP2		x
	SNP3	x	
	SNP4		x
	SNP5	x	x
	SNP6	x	x
Friabilimeter [%]	SNP2	x	
	SNP3	x	
	SNP5	x	
	SNP6	x	

7.4.4 Bewertung

Abschließend wird der Prototyp anhand der in Kapitel 5 vorgestellten Kriterien bewertet.

- **Grad der Integration (G):**

  Der Datenpool in Schicht 3 des Prototypen verwendet datendomänenspezifische Schemata, in die phänotypische, Marker- und Passportdaten importiert wurden. Die domänenübergreifende Integration dieser Daten erfolgte im analysespezifischen Assoziationsmarts in Schicht 5. Der Grad der Integration kann daher mit $+$ bewertet werden.

- **Materialisierung der Integration (M):**

  Daten wurden im Prototypen materialisiert integriert. Damit muss die Materialisierung der Integration mit $-$ bewertet werden. Wie in Abschnitt 6.7 ausgeführt, wurde dies an dieser Stelle bewusst akzeptiert, um den in Schicht 3 propagierten Datenpool zu schaffen.

- **Realisierungsstand (R):**

  Abweichend von der Bewertung des Konzepts in Abschnitt 6.7 wird das Kriterium Realisierungsstand (R) des Prototypen mit $+$ bewertet, da eine durchgängige Implementierung erfolgt ist.

- **Plattformunabhängigkeit (P):**

  Der Prototyp wurde unter Verwendung des Datenbankmanagementsystems Oracle sowie der Oracle-Application-Express-Technologie implementiert. Für beide existieren Versionen für verschiedene Betriebssystemplattformen. Die entwickelten Datenimport-Werkzeuge und die zur Berechnung der Assoziationen eingesetzten Softwarewerkzeuge basieren auf der Programmiersprache Java. Die Plattformunabhängigkeit kann deshalb mit $+$ bewertet werden.

- **Internetfähigkeit (I):**

  Der Zugriff auf das Integrationssystem via Internet/Intranet ist mit dem Prototypen implementiert. Daher wird die Internetfähigkeit mit $+$ bewertet.

- **Schnittstelle, Anfragesprachen (SA):**

  Integrierte Daten sind mit dem Datenbankmanagementsystem Oracle relational abgespeichert und über Standardanfragesprachen wie SQL zugreifbar $(+)$.

- **Schnittstelle, Programmiersprachen (SP):**

  Auf den Datenzugriff über Application Programming Interfaces wie JDBC trifft dasselbe zu $(+)$.

- **Schnittstelle, Datenausgabeformate (SF):**

  Das Kriterium Schnittstelle, Datenausgabeformate (SF) wird im Falle des Prototypen mit $-$ bewertet. Obwohl eine Exportfunktionalität für Daten implementiert ist, ist das Exportformat, bedingt durch die verwendete Software TASSEL, proprietär. Aufgrund der strukturierten Speicherung der verwendeten Daten ist die Implementierung beliebiger weiterer Datenaustauschformate mit vertretbarem Aufwand durchführbar.

- **Flexibilität (F):**

  Der hier vorgestellte Prototyp ist flexibel auf neue Anforderungen anpassbar. Das Kriterium Flexibilität kann deshalb mit $+$ bewertet werden.

- **Unterstützung von Informationsfusion (U):**

  Der Prototyp unterstützt die Kombination von Daten aus drei Domänen mit dem Ziel der Ableitung neuer Informationen. Daher wird auch dieses Merkmal mit $+$ bewertet.

- **Gleichzeitige Verwendung verschiedener Datendomänen (D):**

  Der Prototyp verwendet phänotypische, Marker- und Passportdaten. Aus diesem Grund kann dieses Kriterium mit $+$ bewertet werden.

- **Unterstützung ergebnisoffener Analysen (E):**

  Die generischen, datendomänenspezifischen Schemata in Schicht 3 ermöglichen die flexible Erstellung von Datamarts, wie des hier vorgestellten Assoziationsmarts. Beliebige (ergebnisoffene) Analysen werden dadurch ermöglicht $(+)$.

- **Beschränkung auf eine Klasse von Analysen (A):**

  Obwohl der vorgestellte Prototyp der Berechnung von Assoziationen dient, gibt es grundsätzlich keine Beschränkung auf eine Klasse von Analysen $(+)$.

- **Beschränkung auf ein festes Zielschema (Z):**

  Das Datenbankschema des Prototypen enthält generische Komponenten zur Verwaltung von phänotypischen und Markerdaten und kann flexibel erweitert werden. Deshalb wird dieses Kriterium mit $+$ bewertet.

- **Verwendbarkeit bei proprietären Datenformaten (V):**

  Der Datenpool in Schicht 3 des Prototypen stellt sicher, dass alle Daten in relationale Strukturen überführt werden. Die Verwendbarkeit bei proprietären Datenformaten wird daher mit $+$ bewertet.

- **Berücksichtigung der Datenqualität (Q):**

  Der entwickelte Prototyp ist dahingehend implementiert, eine hohe Qualität der integrierten Daten sicherzustellen. Dieses Merkmal kann mit $+$ bewertet werden.

- **Nutzung von Metadaten (N):**

  Wie in Abschnitt 4.5 diskutiert, wurden bei der Entwicklung des Prototypen Strukturen geschaffen, um Metadaten zu verwalten (z. B. Informationen über Merkmale, Experimente etc.) Weiterhin sind alle implementierten Funktionen und Prozeduren nachvollziehbar dokumentiert worden (+).

Tabelle 7.5 fasst die Bewertung des Prototypen zusammen.

Tabelle 7.5: Bewertung des Prototypen

	G	M	R	P	I	SA	SP	SF	F	U	D	E	A	Z	V	Q	N
Prototyp	+	−	+	+	+	+	+	−	+	+	+	+	+	+	+	+	+

Legende:

G	–	Grad der Integration	enge Kopplung (+)	lose Kopplung (−)
M	–	Materialisierung der Integration	nicht materialisiert (+)	materialisiert (−)
R	–	Realisierungsstand	implementiert(+)	theoretischer Ansatz (−)
P	–	Plattformunabhängigkeit	unabhängig (+)	plattformgebunden (−)
I	–	Internetfähigkeit	entfernter Zugriff (+)	lokale Installation (−)
SA	–	Schnittstelle, Anfragesprachen	unterstützt (+)	nicht unterstützt (−)
SP	–	Schnittstelle, Programmiersprachen	unterstützt (+)	nicht unterstützt (−)
SF	–	Schnittstelle, Datenausgabeformate	verschiedene Formate (+)	nur ein Format (−)
F	–	Flexibilität	anpassbar (+)	statisch (−)
U	–	Unterstützung von Informationsfusion	unterstützt (+)	nicht unterstützt (−)
D	–	Gleichzeitige Verwendung verschiedener Datendomänen	mehrere Domänen (+)	nur eine Domäne (−)
E	–	Unterstützung ergebnisoffener Analysen	unterstützt (+)	nicht unterstützt (−)
A	–	Beschränkung auf eine Klasse von Analysen	keine Beschränkung (+)	Beschränkung (−)
Z	–	Beschränkung auf ein festes Zielschema	keine Beschränkung (+)	Beschränkung (−)
V	–	Verwendbarkeit bei proprietären Formaten	verwendbar (+)	nicht verwendbar (−)
Q	–	Berücksichtigung der Datenqualität	berücksichtigt (+)	nicht berücksichtigt (−)
N	–	Nutzung von Metadaten	möglich (+)	nicht möglich (−)

7.5 Resümee

In diesem Kapitel wurde die Entwicklung eines Prototypen des in Kapitel 6 entworfenen Konzepts beschrieben. Ziel dieses Prototypen war die Unterstützung der Durchführung von Assoziationsstudien zur Aufdeckung von Genotyp-Phänotyp-Korrelationen bei Gerstensorten.

Dazu wurden ca. 80.000 phänotypische Datensätze und ca. 120.000 Markerdatensätze (SNPs, INDELs und SSRs) importiert und aufbereitet. Die Sicherstellung einer hohen Datenqualität erfolgte in mehreren Schritten. Es wurden MS-Excel-Vorlagen zur strukturierten Erfassung von Daten durch Experimentatoren entwickelt. Eine erste Überprüfung fand während des Imports dieser Daten mit Hilfe Java-basierter Ladeprozeduren statt. Auf der Grundlage von Metadaten wurden dabei nur Datensätze zugelassen, für die Informationen über Merkmale und Experimente im Datenpool hinterlegt waren. Grafische Benutzeroberflächen unterstützten die manuelle Kurierung der importierten Daten. Für die Integration der importierten Daten aus dem Datenpool in den Assoziationsmart wurden mehrere zusätzliche Funktionen und Prozeduren, insbesondere zum Auffinden und Eliminieren von Duplikaten sowie von Sequenzierfehlern, aber auch zum Behandeln von Ausreißern und Fehlstellen entwickelt. Es erfolgten Abgleiche genetischer und phänotypischer Daten für jedes einzelne Kandidatengen. Diese konnten in speziellen Exportformaten für externe Softwarewerkzeuge bereitgestellt werden.

Der Prototyp ist vollständig implementiert. Er steht für Assoziationsstudien zur Verfügung und wurde ausschnittsweise in [KGM+07] publiziert.

Vor der Entwicklung des Prototypen erfolgte das Vorbereiten von phänotypischen und Markerdaten manuell und beanspruchte sehr viel Zeit. Dies betraf insbesondere das Zusammenstellen der Daten aus einer Vielzahl von Quellen, den Abgleich der Daten zwischen den beiden Domänen sowie die Erstellung von Inputdateien für die verwendeten externen Softwarewerkzeuge. Für das Berechnen von Assoziationen auf der Basis von Teilmengen dieser Daten (z. B. nur Sommergersten) mussten die manuellen Schritte jeweils erneut ausgeführt werden.

Der Prototyp hat in erheblichem Ausmaß zur Steigerung der Effizienz bei der Verwaltung und der Vorbereitung zur Verrechnung großer Datensätze in der Assoziationsgenetik beigetragen. Insbesondere der Abgleich von Daten unterschiedlichen Umfangs und die Erstellung von Inputdateien für externe Softwarewerkzeuge sowie das Auffinden und Bereinigen von Widersprüchen und Duplikaten innerhalb der Markerdaten und von Ausreißern und Fehlstellen innerhalb der phänotypischen Daten wurde vollständig automatisiert. Gleichzeitig konnte damit eine gleichbleibend hohe Qualität der zu analysierenden Daten sichergestellt werden; der Faktor Mensch wurde als Fehlerquelle weitestgehend ausgeschlossen.

Resultate von Assoziationsberechnungen für eine Reihe von Kandidatengenen mit Malz- und Brauqualitätsmerkmalen liegen vor. Hierüber wurden bereits zwei Arbeiten veröffentlicht [MWR09,MWFR09]. Der hier entwickelte Prototyp wird im Anschlussforschungsprojekt GABI-GENOBAR auf Daten einer großen Anzahl von Zuchtstämmen angewandt und weiterentwickelt werden.

Aufgrund großen Interesses von Züchtungsunternehmen und Forscherkollegen wurde die im Datenpool des Prototypen erstellte Sammlung phänotypischer Daten eigenständig publiziert [WSRM09].

8 Zusammenfassung und Ausblick

Auf den nächsten Seiten sollen die einzelnen Teile der vorliegenden Arbeit zusammengefasst werden. Anschließend wird ein Ausblick auf potenzielle Erweiterungen der in dieser Arbeit besprochenen Konzepte sowie auf zusätzliche Anwendungsfälle erfolgen.

8.1 Zusammenfassung

Bedingt durch den Einsatz von Hochdurchsatzmethoden hat in den letzten Jahren eine Vervielfachung des Datenaufkommens in der Biologie stattgefunden. Damit einhergehend bewegt sich der wissenschaftliche Fokus von der Untersuchung einzelner Datendomänen und problemorientierter Arbeit hin zur domänenübergreifenden und ergebnisoffenen Analyse.

Trotz der Unterstützung durch bioinformatische Werkzeuge wird das individuelle und manuelle Arbeiten mit umfangreichen Datensätzen zunehmend komplizierter. Diesem Problem wurde in den letzten Jahren durch die Entwicklung einer großen Anzahl von Informations- und Analysesystemen Rechnung getragen. Jedoch steht häufig nur eine Datendomäne im Blickpunkt dieser Systeme; integrierte Analysen werden vernachlässigt.

Während große Datenmengen und bioinformatische Infrastruktur im Bereich der Humanmedizin, Taxonomie und molekulargenetischen Modellorganismen wie der Fruchtfliege oder der Maus vorhanden sind, sind die landwirtschaftlichen Nutzpflanzen trotz ihrer stetig wachsenden Bedeutung als Nahrungsgrundlage für Mensch und Tier sowie als erneuerbare Energiequelle vielfach unterrepräsentiert. Davon sind einerseits

die Gewinnung sowie andererseits auch die Speicherung und Auswertung von Daten betroffen. Daher bestand das Ziel dieser Arbeit in der Übertragung moderner Informatikmethoden auf diesen Teil der biologischen Forschung.

Der Fokus der Arbeit war auf die Entwicklung eines Konzepts zur integrierten Analyse pflanzenbiologischer Daten gerichtet. Dieses sollte es ermöglichen, potenzielle Wechselwirkungen zwischen den Datendomänen, beispielsweise Genotyp-Phänotyp-Korrelationen, aufzuzeigen. Eine wichtige Voraussetzung für die integrierte Analyse bildet die Qualität der zugrunde liegenden Daten. Dies gilt insbesondere für die Vergleichbarmachung von Daten aus unterschiedlichen Quellen.

Hierzu wurden spezifische Herausforderungen und Lösungsansätze diskutiert. Des weiteren erfolgte die Betrachtung existierender Ansätze und die Bewertung dieser nach einheitlichen Kriterien. Dabei wurden bestehende Kriterien zur Bewertung von Integrationssystemen aufgegriffen und um Kriterien zur Bewertung domänenübergreifender Datenanalysen erweitert. Es wurden Vorschläge zur integrierten Analyse biologischer Daten unter Berücksichtigung von Spezifika der Pflanzenbioinformatik erarbeitet und diskutiert. Besonderes Gewicht kam dabei Qualitätsaspekten dieser Daten sowie domänenübergreifenden Analysemöglichkeiten zu. Dies war insbesondere wichtig, da diese Problematiken durch existierende Ansätze häufig nicht zufriedenstellend berücksichtigt werden. Zur Sicherstellung hoher Datenqualität nehmen in diesem Kontext Metadaten eine besondere Stellung ein.

Basierend auf diesen Untersuchungen wurde ein Konzept zur flexiblen, integrierten Analyse pflanzenbiologischer Daten unter Verwendung von Datawarehouse-Methoden entwickelt. Dieses besteht aus einer sechsschichtigen Architektur, deren Kernelemente ein Datenpool sowie analysespezifische Datamarts bilden. Das Hauptaugenmerk lag dabei auf der Bereitstellung qualitativ hochwertiger Daten für Analysen. Die vorgeschlagene Architektur wurde ausführlich diskutiert. Anschließend erfolgte die Erläuterung der einzelnen Elemente anhand eines Prototypen.

Dieser Prototyp dient der Integration von phänotypischen, genetischen und Passportdaten von Braugerstensorten mit dem Ziel der flexiblen Durchführung von Assoziationsstudien zur Aufdeckung potentieller Genotyp-Phänotyp-Korrelationen. Besonderes Gewicht wurde dabei auf die Sicherstellung einer hohen Datenqualität gelegt.

Der Prototyp leistet einen unterstützenden Beitrag zur Assoziationsgenetik, indem er in erheblichem Ausmaß zur Steigerung der Effizienz bei der Verwaltung und der Vorbereitung zur Verrechnung großer Datensätze geführt hat. Ergebnisse von Assoziationsstudien liegen vor und sind Gegenstand aktueller Publikationen.

Obwohl der Fokus der vorliegenden Arbeit auf die pflanzenbiologische Forschung gerichtet war, ist das entwickelte Konzept neben dem Prototypen grundsätzlich auch auf andere Bereiche der biologischen Forschung anwendbar.

8.2 Ausblick

Auf dem präsentierten Prototypen aufbauend sind verschiedene Erweiterungsmöglichkeiten vorstellbar. Es wäre wünschenswert, zusätzliche Prozeduren zu entwickeln, die es ermöglichen, eine Vorauswertung der im Assoziationsmart befindlichen Daten über statistische Verfahren durchzuführen und Vorhersagen darüber zu treffen, welche Kombinationen phänotypischer und Markerdaten für die Berechnung von Assoziationen besonders geeignet erscheinen. Dies würde der Steigerung der Effizienz solcher Analysen dienen.

Die im Rahmen dieser Arbeit entwickelten Methoden werden auch im kürzlich begonnenen, vom BMBF geförderten Forschungsprojekt GABI-GENOBAR Verwendung finden. Wie im beschriebenen Anwendungsfall soll auch in diesem Projekt die Untersuchung von Marker-Merkmals-Assoziationen erfolgen. Das Ziel besteht in der genomweiten Assoziation genetischer Vielfalt mit agronomisch bedeutsamen Merkmalen der Gerste, d. h. in der Identifikation von genomischen Regionen mit Einflüssen auf Merkmalsausprägungen. Sie können für die Auswahl geeigneter Elternkombinationen für die Züchtung neuer Sorten verwendet werden. Langfristig betrachtet könnte dadurch eine zeitnahe Zulassung neuer bzw. verbesserter Sorten gewährleistet werden. Dies kann dadurch ermöglicht werden, dass die zu berechnenden Assoziationen einen wirksamen Einsatz markergestützter Selektion erlauben und damit zu einem schnelleren Züchtungsfortschritt führen. Zur Unterstützung dieser Arbeiten wird in diesem Projekt die Erstellung des so genannten GENOBAR-WAREHOUSES auf Basis der in dieser Arbeit entwickelten Methoden erfolgen. Hier können insbesondere die während der Realisierung des Prototypen gewonnenen Erfahrungen in der qualitätsberücksichtigenden Integration und Aufbereitung von Daten für Assoziationsstudien zum Tragen kommen. Das GENOBAR-WAREHOUSE wird die Grundlage für die Anwendung statistischer Verfahren bilden.

Im Rahmen der Entwicklung des Prototypen wurde eine umfangreiche Sammlung phänotypischer Daten erstellt (vgl. Kapitel 7), die ebenso für das GENOBAR-Projekt eine hohe Relevanz hat. Weiterhin wurde am IPK Gatersleben ein Operativsystem, die Pyrosequencing Database (PSQDB)[1], entwickelt. Der Zweck der PSQDB bestand in der Verwaltung von Pyrosequenzierungsdaten zur Unterstützung der Genotypisierung einer Kollektion von Lolium-Akzessionen (Weidelgras) der Gaterslebener Genbank. Die Strukturen und z. T. auch die Daten der beiden genannten Systeme stehen als Vorarbeiten für das GENOBAR-Projekt zur Verfügung.

Die wirtschaftliche Relevanz des Projektes wird durch die große Anzahl daran beteiligter Gerstenzüchter verdeutlicht. Eine Übertragung der gewonnenen Erkenntnisse auf andere Getreidearten ist denkbar. Dies ist durch orthologe Bereiche der Genome

[1] http://bic-gh.de/psqdb [Stand 2009-04-02]

verschiedener Spezies, etwa bei verwandten Arten, möglich. Ebenso denkbar ist die Anwendung von Teilen des GENOBAR-WAREHOUSES auf andere Getreidearten.

Neben der Verwendung für Assoziationsstudien könnte eine vielversprechende Anwendung des entwickelten Konzepts im Vergleich unveränderter und genetisch veränderter Pflanzen bestehen. Zur Erfassung experimenteller Daten werden zunehmend Laborinformationsmanagementsysteme (LIMS) [Gib96] eingesetzt. Dabei handelt es sich um Softwaresysteme zur Verwaltung und persistenten Speicherung von chemischen, physikalischen, biologischen oder medizinischen Labordaten sowie Metadaten. In einem LIMS wird die gesamte Prozesskette vom Probeneingang über die Kontrolle des Messprozesses bis hin zur Verwertung zusammenhängend erfasst. Dies dient sowohl der Rationalisierung als auch der Erfüllung von Dokumentationspflichten (insbesondere im Kontext der ISO-Normen 9001 und 17025[2]). Da in etablierten LIMS Daten in relationalen Strukturen vorliegen, wird eine potenzielle Integration von Informationen aus solchen Systemen vereinfacht. Weiterhin wirkt sich eine strukturierte Speicherung von Daten in diesen Systemen positiv auf die Datenqualität aus. Auch in LIMS werden identifizierende Daten nach dem Prinzip der Passportdaten verwaltet. Hierüber ist unter Verwendung des vorgestellten Konzepts die Integration der Daten mit denen anderer Quellen derselben Domäne und zusätzlich auch mit Daten anderer Domänen möglich.

Eine weitere Anwendung des Konzepts könnte in der Unterstützung systembiologischer Forschung liegen. Es existieren sehr große Sammlungen von Daten verschiedener Domänen, z. B. genomische Daten aus GeneOntology [GO 08] oder KEGG [KAG+08], metabolische Daten aus AraCyc [ZFT+05] oder MetaCrop [GBWK+08] bzw. proteomische Daten aus BRENDA [BEC+07], die in einen aussagekräftigen Zusammenhang gebracht werden müssen. Aufgrund des Umfangs solcher Daten wird es zunehmend schwerer, alle relevanten Experimente mit konventionellen Methoden durchzuführen. Die Systembiologie versucht, komplexe biologische Prozesse in ihrer Gesamtheit zu verstehen und auf mathematische Modelle abzubilden [Kit02]. Solche Modelle dienen der Durchführung von Simulationen (In-silico-Experimente) und können damit die Grundlage für Vorhersagen und anschließende experimentelle Überprüfungen bilden (Abbildung 8.1).

Zur Bereitstellung relevanter Daten für systembiologische Fragestellungen ist die Verwendung des Konzepts aus Kapitel 6 denkbar. Dabei ist das Konzept dem Bereich der Modellerstellung in Abbildung 8.1 zuzuordnen. Abbildung 8.2 zeigt hierzu die potenzielle Anwendung.

[2] http://www.iso.org [Stand 2009-04-02]

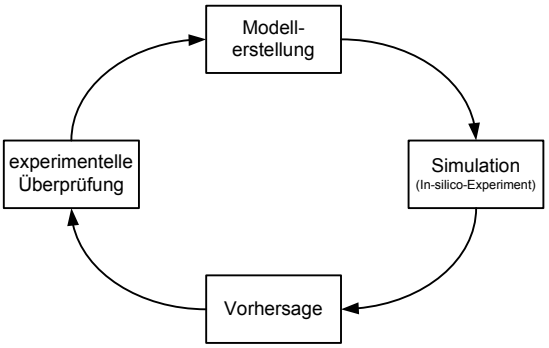


Abbildung 8.1: Zyklus der Systembiologie in Anlehnung an [Kit02]

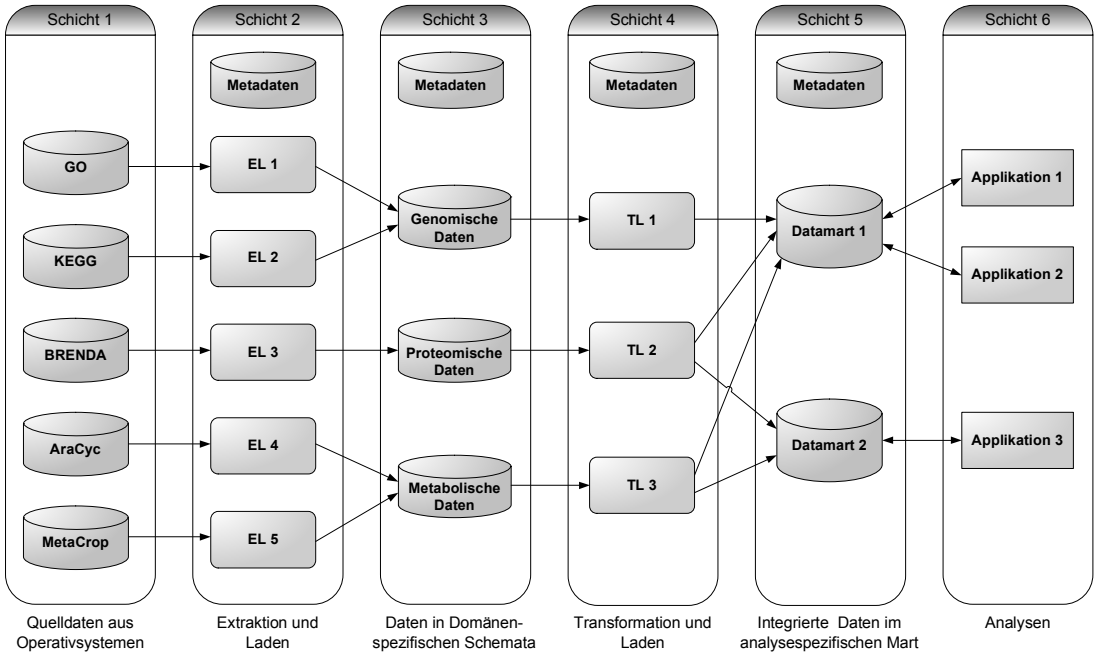


Abbildung 8.2: Potenzielle Anwendung des in Kapitel 6 entworfenen Konzepts für die Unterstützung systembiologischer Forschung

Für den Bereich metabolischer Daten von Kulturpflanzen sind am IPK Gatersleben Vorarbeiten aus [GBWK+08, WGK+06] vorhanden. Eine Auswahl der systembiologischen Fragestellungen, die die Grundlage dieser Arbeiten bildeten, soll abschließend kurz aufgeführt werden:

- Existieren Abweichungen bei der Verwendung unterschiedlicher Reaktionskinetiken zur Simulation?
- Wie verhält sich die Konzentration eines Metaboliten, wenn die Aktivität eines eine Reaktion katalysierenden Enzyms verdoppelt wird?

- Gibt es Differenzen der Metabolitkonzentrationen bei Aktivierung oder Inhibierung eines Enzymes in einem Genotypen im Vergleich zum Wildtyp?
- Bestehen Zusammenhänge zwischen Pathways? Dies ist aufgrund unvollständigen Wissens über solche Netzwerke keine triviale Aussage. Die Beantwortung einer solchen Fragestellung könnte auf Basis unterschiedlicher Metabolitkonzentrationen erfolgen.

Zusammenfassend ist nochmals festzustellen, dass das in der vorliegenden Arbeit entwickelte Konzept gegenwärtig in einem weiteren Forschungsprojekt zur Unterstützung der Erforschung von Marker-Merkmals-Assoziationen (GABI-GENOBAR) Anwendung findet. Zusätzliche Einsatzpotenziale wurden skizziert.

A Screenshots des Prototypen

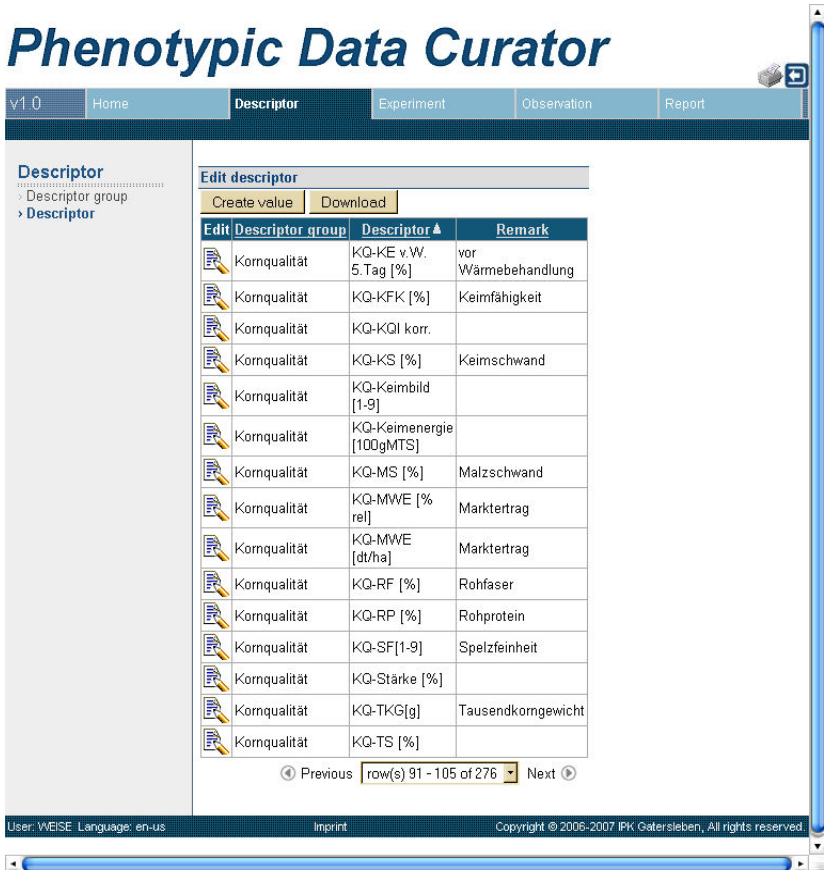


Abbildung A.1: Verwaltung phänotypischer Merkmale mit dem Kurationswerkzeug für phänotypische Daten (Schicht 3)

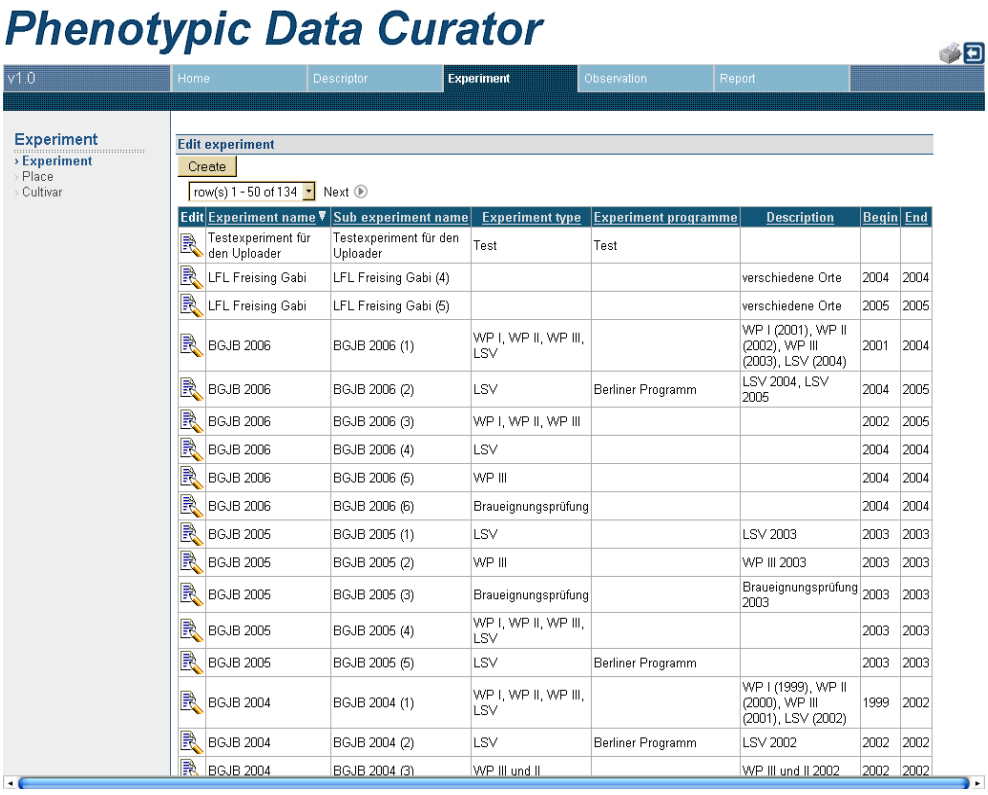
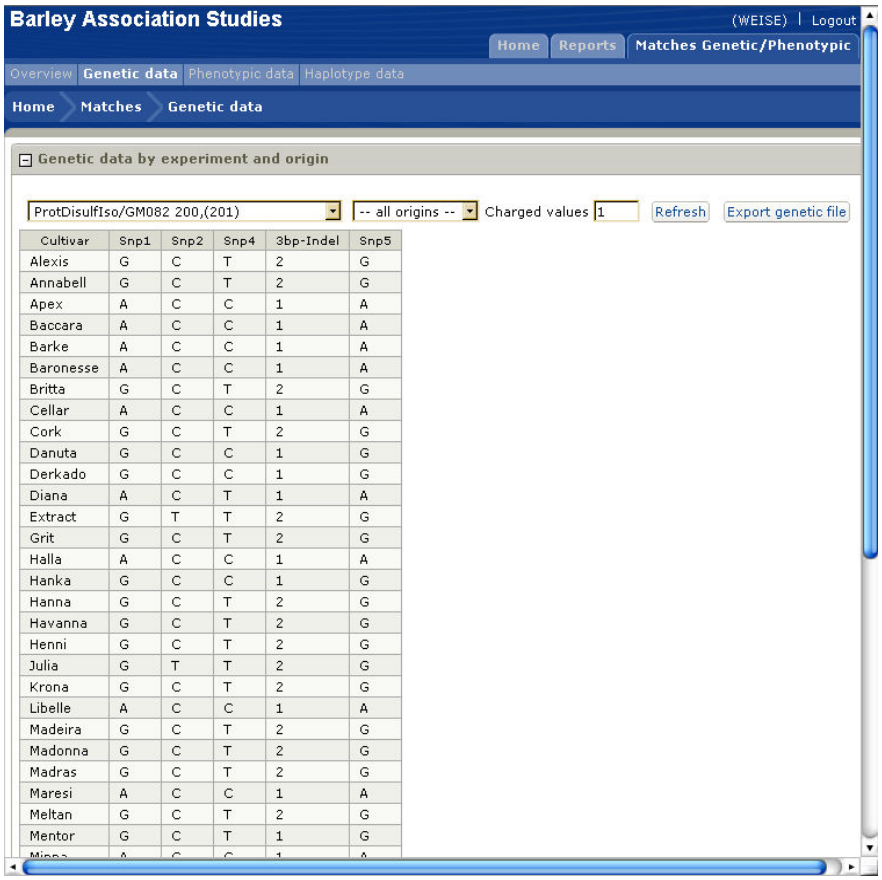


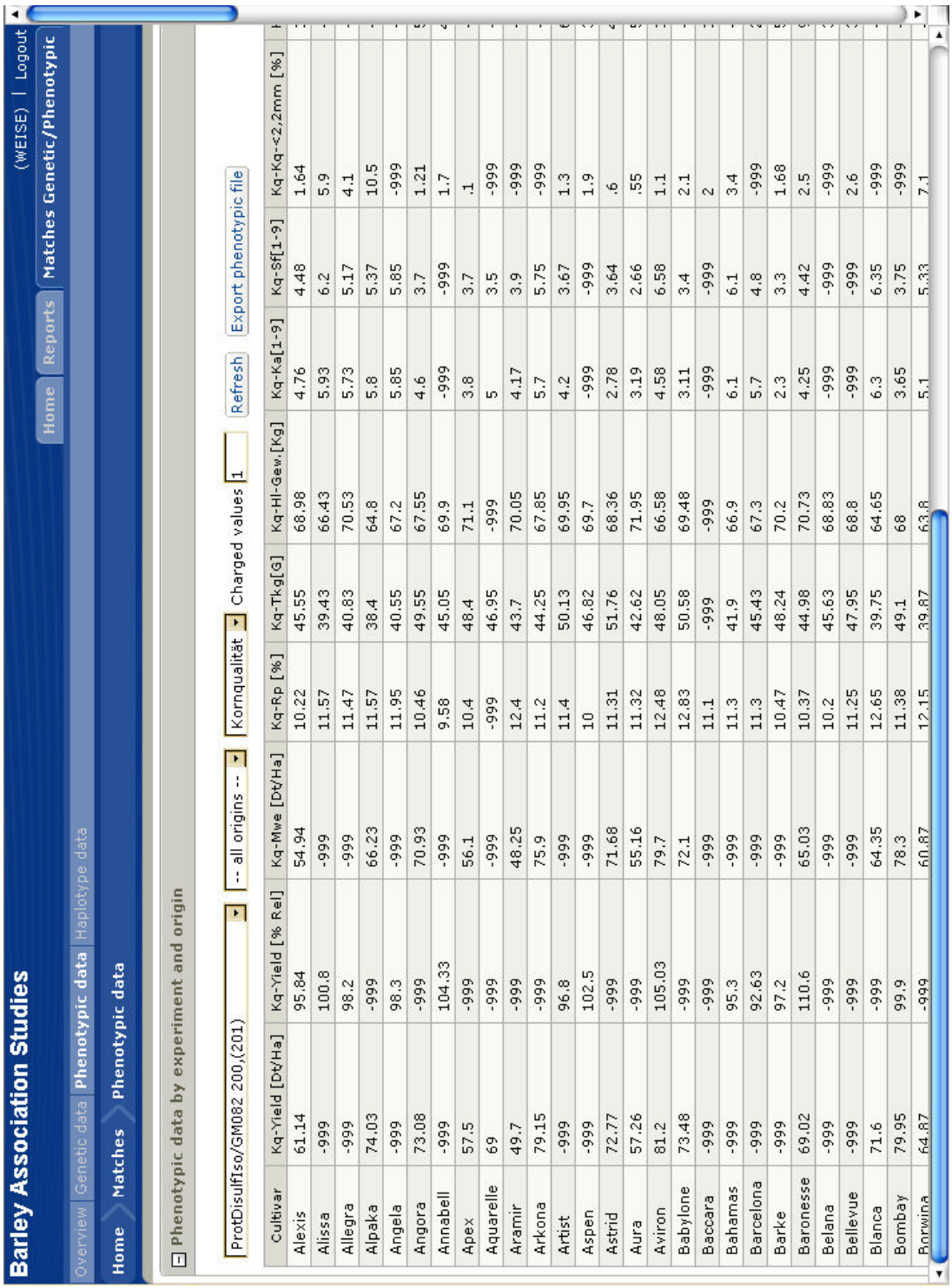
Abbildung A.2: Verwaltung von Experimenten, in denen phänotypische Daten erhoben wurden (Schicht 3)

Abbildung A.3: Kerntabellen des Datenbankschemas des Assoziationsmarts (Schicht 5); hinzu kommen 37 logische und materialisierte Sichten, die hier nicht dargestellt sind



Cultivar	Snp1	Snp2	Snp4	3bp-Indel	Snp5
Alexis	G	C	T	2	G
Annabell	G	C	T	2	G
Apex	A	C	C	1	A
Baccara	A	C	C	1	A
Barke	A	C	C	1	A
Baronesse	A	C	C	1	A
Britta	G	C	T	2	G
Cellar	A	C	C	1	A
Cork	G	C	T	2	G
Danuta	G	C	C	1	G
Derkado	G	C	C	1	G
Diana	A	C	T	1	A
Extract	G	T	T	2	G
Grit	G	C	T	2	G
Halla	A	C	C	1	A
Hanka	G	C	C	1	G
Hanna	G	C	T	2	G
Havanna	G	C	T	2	G
Henni	G	C	T	2	G
Julia	G	T	T	2	G
Krona	G	C	T	2	G
Libelle	A	C	C	1	A
Madeira	G	C	T	2	G
Madonna	G	C	T	2	G
Madras	G	C	T	2	G
Maresi	A	C	C	1	A
Meltan	G	C	T	2	G
Mentor	G	C	T	1	G

Abbildung A.4: Anzeige von Markerinformationen mit der Assoziationsmart-Anwendung (Schicht 6)



Cultivar	Kq-Yield [Dt/Ha]	Kq-Yield [% Rel]	Kq-Mwe [Dt/Ha]	Kq-Rp [%]	Kq-Tkg[G]	Kq-Hl-Gew.[Kg]	Kq-Ka[1-9]	Kq-Sf[1-9]	Kq-Kq-<2,2mm [%]
Alexis	61.14	95.84	54.94	10.22	45.55	68.98	4.76	4.48	1.64
Alissa	-999	100.8	-999	11.57	39.43	66.43	5.93	6.2	5.9
Allegra	-999	98.2	-999	11.47	40.83	70.53	5.73	5.17	4.1
Alpaka	74.03	-999	66.23	11.57	38.4	64.8	5.8	5.37	10.5
Angela	-999	98.3	-999	11.95	40.55	67.2	5.85	5.85	-999
Angora	73.08	-999	70.93	10.46	49.55	67.55	4.6	3.7	1.21
Annabell	-999	104.33	-999	9.58	45.05	69.9	-999	-999	1.7
Apex	57.5	-999	56.1	10.4	48.4	71.1	3.8	3.7	.1
Aquarelle	69	-999	-999	-999	46.95	-999	5	3.5	-999
Aramir	49.7	-999	48.25	12.4	43.7	70.05	4.17	3.9	-999
Arkona	79.15	-999	75.9	11.2	44.25	67.85	5.7	5.75	-999
Artist	-999	96.8	-999	11.4	50.13	69.95	4.2	3.67	1.3
Aspen	-999	102.5	-999	10	46.82	69.7	-999	-999	1.9
Astrid	72.77	-999	71.68	11.31	51.76	68.36	2.78	3.64	.6
Aura	57.26	-999	55.16	11.32	42.62	71.95	3.19	2.66	.55
Aviron	81.2	105.03	79.7	12.48	48.05	66.58	4.58	6.58	1.1
Babylone	73.48	-999	72.1	12.83	50.58	69.48	3.11	3.4	2.1
Baccara	-999	-999	-999	11.1	-999	-999	-999	-999	2
Bahamas	-999	95.3	-999	11.3	41.9	66.9	6.1	6.1	3.4
Barcelona	-999	92.63	-999	11.3	45.43	67.3	5.7	4.8	-999
Barke	-999	97.2	-999	10.47	48.24	70.2	2.3	3.3	1.68
Baronesse	69.02	110.6	65.03	10.37	44.98	70.73	4.25	4.42	2.5
Belana	-999	-999	-999	10.2	45.63	68.83	-999	-999	-999
Bellevue	-999	-999	-999	11.25	47.95	68.8	-999	-999	2.6
Blanca	71.6	-999	64.35	12.65	39.75	64.65	6.3	6.35	-999
Bombay	79.95	99.9	78.3	11.38	49.1	68	3.65	3.75	-999
Borwina	64.87	-999	60.87	12.15	39.87	63.8	5.1	5.33	7.1

Abbildung A.5: Anzeige von phänotypischen Daten mit der Assoziationsmart-Anwendung (Schicht 6)

B Quellcodes

B.1 Bereinigung importierter Daten im Assoziationsmart

Tabelle B.1: Übersicht der Funktionen und Prozeduren, die zur Bereinigung importierter Daten im Assoziationsmart entwickelt wurden

Name	Beschreibung
Funktion *cleanCultivar*	Entfernt doppelte Zeilen der Markerdaten einer mehrfach auftretenden Sorte.
Funktion *cleanMarker*	Überprüft, ob für eine mehrfach auftretende Sorte alle Zeilen der Markerdaten identisch sind.
Funktion *getMultipleCultivarsPerExp*	Findet mehrfach auftretende Sorten innerhalb des Datensatzes.
Prozedur *cleanseGeneticData*	Liest die in den Assoziationsmart importierten Markerdaten und überprüft auf mehrfach auftretende Sorten und widersprüchliche Informationen.

Fortsetzung auf der nächsten Seite

Fortsetzung von Tabelle B.1

Name	Beschreibung
Prozedur *cleansePhenotypicData*	Liest die in den Assoziationsmart importierten phänotypischen Daten und berechnet Aggregationen für die Assoziationsmartanwendung vor. Hierbei werden zusätzlich Ausreißer behandelt.

Listing B.1 zeigt beispielhaft eine der in Tabelle B.1 aufgeführten Funktionen für die Bereinigung der Markerdaten.

Listing B.1: Funktion *cleanMarker*

```
1  -- überprüfen, ob für eine mehrfach auftretende Sorte alle Zeilen identisch sind
2  create or replace function cleanMarker (p_cultivar IN varchar2, p_seq_experiment_id
       IN number, p_marker_no IN number)
3  return number
4  as
5
6  -- Variablen
7  ret number;
8  p_differences number := 0;
9
10 begin
11
12 -- In einer Schleife alle Marker der übergebenen Sorte für alle Zeilen dieser Sorte
       durchgehen.
13 -- Dabei die Ausprägungen des jeweiligen Markers über alle Zeilen der Sorte
       vergleichen. Wenn Unterschiede
14 -- auftreten, ALLE Ausprägungen dieses Markers auf '-999' setzen.
15 -- Danach die mehrfachen Zeilen der Sorte bis auf eine löschen.
16 for i in 1..p_marker_no
17 loop
18
19   -- auf Unterschiede in der Markerausprägung überprüfen
20   select count(distinct marker) as differences
21   into p_differences
22   from
23   (
24     select d.t020_id_pk as seq_experiment_id,
25       e.T007_CULTIVAR_NAME as cultivar,
26       a.T022_marker_ORDER as marker_order,
27       a.T022_MARKER as marker,
28       b.T021_HAPLOTYPE as haplotype
29     from ((t022_marker a left outer join t021_haplotype b on a.
           T022_T021_ID_HAPLOTYPE_FK = b.t021_id_pk)
30                 left outer join t020_seq_experiment d on b.
                       T021_T020_ID_SEQ_EXPERIMENT_FK = d.t020_id_pk)
31                 left outer join t007_cultivar e on b.T021_T007_ID_CULTIVAR_FK = e.
                       t007_id_pk
32     where d.t020_id_pk = p_seq_experiment_id
33       and
34       e.T007_CULTIVAR_NAME = p_cultivar
35       and
36       a.T022_marker_ORDER = i
```

```
37    );
38
39    -- im Fall von Unterschieden auf '-999' setzen
40    if ( p_differences > 1) then
41
42    DBMS_OUTPUT.PUT_LINE('---->Marker no. ' || i || ', differences: ' || p_differences
          );
43
44    -- die Markerausprägung auf '-999' setzen
45    update t022_marker
46    set t022_marker = '-999'
47    where t022_id_pk in (
48                select a.T022_ID_PK
49                from ((t022_marker a left outer join t021_haplotype b on a.
                      T022_T021_ID_HAPLOTYPE_FK = b.t021_id_pk)
50                            left outer join t020_seq_experiment d on b.
                                T021_T020_ID_SEQ_EXPERIMENT_FK = d.t020_id_pk)
51                            left outer join t007_cultivar e on b.T021_T007_ID_CULTIVAR_FK
                                = e.t007_id_pk
52                where d.t020_id_pk = p_seq_experiment_id
53                  and
54                  e.T007_CULTIVAR_NAME = p_cultivar
55                  and
56                  a.T022_marker_ORDER = i
57              );
58
59    -- die zugehörigen Haplotypeneinträge als modifiziert kennzeichnen
60    update t021_haplotype
61    set t021_haplotype = 'modified'
62    where t021_id_pk in (
63                select b.T021_ID_PK
64                from ((t022_marker a left outer join t021_haplotype b on a.
                      T022_T021_ID_HAPLOTYPE_FK = b.t021_id_pk)
65                            left outer join t020_seq_experiment d on b.
                                T021_T020_ID_SEQ_EXPERIMENT_FK = d.t020_id_pk)
66                            left outer join t007_cultivar e on b.T021_T007_ID_CULTIVAR_FK
                                = e.t007_id_pk
67                where d.t020_id_pk = p_seq_experiment_id
68                  and
69                  e.T007_CULTIVAR_NAME = p_cultivar
70                  and
71                  a.T022_marker_ORDER = i
72              );
73
74    end if;
75
76    p_differences := 0;
77
78    -- die mehrfachen Zeilen bis auf eine löschen
79    ret := cleanCultivar (p_cultivar, p_seq_experiment_id, i);
80
81  end loop;
82
83
84
85  return 0;
86
87  end;
88  /
```

B.2 Abfrage und Export von Daten aus dem Assoziationsmart

Tabelle B.2: Übersicht der Funktionen und Prozeduren, die zum Abgleich und zum Export von Daten aus dem Assoziationsmart entwickelt wurden

Name	Beschreibung
Prozedur *imputeValues*	Überprüft die in den Assoziationsmart geladenen phänotypischen Daten für alle Kombinationen Sorte/Merkmal hinsichtlich Fehlstellen und führt Ersetzungen (Value Imputation) mit verschiedenen Methoden durch (z. B. Durchschnitt, Minimum, Maximum etc.).
Funktion *buildPhenotypicQuery*	Stellt eine SQL-Abfrage über phänotypische Daten zusammen.
Funktion *getDescriptorNumber*	Bestimmt die Anzahl der zur Verfügung stehenden Deskriptoren.
Funktion *getDescriptorName*	Formatiert die Deskriptorennamen für die Exportdatei.
Prozedur *generatePhenotypicFile*	Erstellt die phänotypische Exportdatei.
Funktion *buildGeneticQuery*	Stellt eine SQL-Abfrage über genetische Daten zusammen.
Funktion *getMarkerNumber*	Bestimmt die Anzahl der zur Verfügung stehenden Marker.
Funktion *getMarkerName*	Formatiert die Markernamen für die Exportdatei.
Prozedur *generateGeneticFile*	Erstellt die genetische Exportdatei.
Funktion *buildHaplotypeQuery*	Stellt eine SQL-Abfrage über Haplotypendaten zusammen.
Funktion *getHaplotypeNumber*	Bestimmt die Anzahl der zur Verfügung stehenden Haplotypen.

Fortsetzung auf der nächsten Seite

Fortsetzung von Tabelle B.2

Name	Beschreibung
Funktion *getHaplotypeName*	Formatiert die Haplotypennamen für die Exportdatei.
Prozedur *generateHaplotypeFile*	Erstellt die Exportdatei für die Haplotypendaten.

Listing B.2 zeigt beispielhaft eine der in Tabelle B.2 aufgeführten Funktionen für die Abfrage von Haplotypeninformationen. Die Funktionen zur Abfrage phänotypischer und Markerdaten sind vergleichbar aufgebaut.

Listing B.2: Funktion *buildHaplotypeQuery*

```
1  -- Anhand der Seq_Experiment_ID und des Origins die SQL-Abfrage zusammenstellen
2  create or replace function buildHaplotypeQuery (p_seq_experiment_id IN number,
       p_origin IN varchar2, p_charged_values IN number)
3  return varchar2
4  as
5  -- Variablen deklarieren
6  query           varchar2(32000 char) := null;
7  query2          varchar2(32000 char) := null;
8  query3          varchar2(32000 char) := null;
9  cid             integer;
10 ignore          integer;
11 p_haplotype_no  number               := 0;
12 p_haplotype     varchar2(4000)       := null;
13 p_marker_no     number               := 0;
14 p_counter       number               := 1;
15
16 -- Alle Markernamen abfragen, die zur übergebenen Seq_Experiment_ID gehören
17 cursor c_marker_names is
18   select distinct a.t022_marker_name as m_name, a.T022_marker_ORDER as m_order
19   from (t022_marker a left outer join t021_haplotype b on a.T022_T021_ID_HAPLOTYPE_FK
         = b.T021_ID_PK)
20              left outer join t020_seq_experiment c on b.T021_T020_ID_SEQ_EXPERIMENT_FK
                  = c.T020_ID_PK
21   where c.T020_ID_PK = p_seq_experiment_id
22   order by a.T022_marker_ORDER asc;
23
24 begin
25
26 -- Abfrage der genetischen Daten erstellen
27 query := buildGeneticQuery (p_seq_experiment_id, p_origin, p_charged_values);
28
29 -- einen neuen Cursor erstellen und dessen ID zurückgeben
30 cid := DBMS_SQL.OPEN_CURSOR;
31
32 -- die eben erstellte Abfrage als innere Abfrage verwenden und alle
33 -- vorkommenden Haplotypen distinct abfragen
34 query2 := 'select distinct';
35
36 for my_marker in c_marker_names
37 loop
38
```

```
39    query2 := query2 || ' "' ||  my_marker.m_name || '", ';
40
41    p_marker_no := p_marker_no + 1;
42
43  end loop;
44
45  query2 := substr ( query2, 1, length(query2)-2);
46
47  query2 := query2 || ' from ( ' || query || ' ) where ';
48
49  for my_marker in c_marker_names
50  loop
51    query2 := query2 || '"' ||  my_marker.m_name || '" <> ''-999'' and ';
52  end loop;
53
54  query2 := substr (query2, 1, length(query2)-5);
55
56  query2 := query2 || ' order by ';
57
58  for my_marker in c_marker_names
59  loop
60    query2 := query2 || '"' ||  my_marker.m_name || '" asc, ';
61  end loop;
62
63  query2 := substr (query2, 1, length(query2)-2);
64
65  -- die Anzahl der zurückgelieferten Zeilen der eben zusammengestellten Abfrage
       bestimmen
66  DBMS_SQL.PARSE (cid, 'select count(1) as no_rows from ( ' || query2 || ')', dbms_sql.
       native);
67  DBMS_SQL.DEFINE_COLUMN (cid, 1, p_haplotype_no);
68  ignore := DBMS_SQL.EXECUTE_AND_FETCH (cid);
69  DBMS_SQL.COLUMN_VALUE (cid, 1, p_haplotype_no);
70
71  -- Cursor schließen
72  DBMS_SQL.CLOSE_CURSOR(cid);
73
74  query3 := 'select cultivar';
75
76  -- einen neuen Cursor erstellen und dessen ID zurückgeben
77  cid := DBMS_SQL.OPEN_CURSOR;
78
79  DBMS_SQL.PARSE (cid, query2, dbms_sql.native);
80
81  for i in 1..p_marker_no
82  loop
83    DBMS_SQL.DEFINE_COLUMN (cid, i, p_haplotype, 4000);
84  end loop;
85
86  ignore := DBMS_SQL.EXECUTE (cid);
87
88  for i in 1..p_haplotype_no
89  loop
90
91    ignore := dbms_sql.fetch_rows(cid);
92
93    query3 := query3 || ', case when ';
94
95    for my_marker in c_marker_names
96    loop
97
98    DBMS_SQL.COLUMN_VALUE (cid, p_counter, p_haplotype);
99
100   if ( p_counter = p_marker_no ) then
101     query3 := query3 || '"' || my_marker.m_name || '" = ''' || p_haplotype || ''' ';
```

```
102     else
103       query3 := query3 || '" ' || my_marker.m_name || '" = ''' || p_haplotype || ''' and
                ';
104     end if;
105
106     p_counter := p_counter + 1;
107
108     end loop;
109
110     p_counter := 1;
111
112     query3 := query3 || ' then 1 else 2 end Haplotype' || i;
113
114   end loop;
115
116   query3 := query3 || ' from ( ' || query || ' )';
117
118   -- Cursor schließen
119   DBMS_SQL.CLOSE_CURSOR(cid);
120
121   return query3;
122
123   end;
124   /
```

Glossar

Akzession: Unter einer Akzession wird ein Sammlungsmuster einer bestimmten Art verstanden, das in einer Genbank gelagert und erhalten wird.

Allelische Diversität: Als Allel wird eine Form eines Gens bezeichnet, die an einem Genlocus eines Chromosoms in einer charakteristischen Basenkombination auftritt [Ibe92]. Variationen der Zusammensetzung von Allelen, beispielsweise innerhalb einer Art, werden als allelische Diversität bezeichnet.

Bioinformatik: Nach [RHM04] ist die Bioinformatik eine Disziplin der Angewandten Informatik, deren Ziel darin besteht, mit Hilfe der Informatik große Mengen biologischer Daten zu verwalten und zu analysieren. Oftmals wird unter Bioinformatik nur die Unterstützung der biologischen Forschung auf molekularer Ebene verstanden (vgl. [BBC+99]). Im Gegensatz zur Bioinformatik (nach klassischer Definition) beschäftigt sich die Biodiversitätsinformatik mit Daten über pflanzengenetische Ressourcen auf nicht-molekularer Ebene. Dabei handelt es sich um den Zweig der Pflanzenbiologie, der sich mit dem Schutz vor Extinktion und Generosion und dem Erhalt der Vielfalt pflanzengenetischer Ressourcen (Biodiversität) beschäftigt. Die Verwendung des Begriffes Bioinformatik soll in dieser Arbeit immer einschließlich des Bereichs der Biodiversitätsinformatik verstanden werden.

Datendomäne: Eine Domäne (von lat. *dominium*: Herrschaftsbereich) definiert allgemein ein Fach- oder Wissenschaftsgebiet. Mit dem Begriff Datendomäne ist hier die Gesamtheit der Daten eines bestimmten Bereichs gemeint, z. B. die Gesamtheit der Sequenzdaten.

Genbank: Die Hauptaufgabe einer Genbank besteht in der Erhaltung und Reproduktion von Kulturpflanzen mit dem Ziel der Minderung des Aussterbens (Extinktion) und des Rückgangs der biologischen Vielfalt (Generosion) von Kulturpflanzen und ihren wildwachsenden Verwandten. Hierzu werden Pflanzen und/oder reproduktive Teile von Pflanzen (in der Regel Samen) einer Vielzahl von Arten mit Hilfe unterschiedlicher Verfahren, z. B. Kühllager, In-vitro-Kulturen oder Kryokonserven, gelagert.

Genotyp: Der Genotyp ist die Gesamtheit der Allele eines Organismus.

Heritabilität: Die Heritabilität (Erblichkeit) gibt an, inwieweit die Ausprägung eines phänotypischen Merkmals genetisch determiniert ist [Fel92].

Integrierte Analyse: Der Begriff der integrierten Analyse bezeichnet die Zusammenführung (Integration) von Daten aus verschiedenen Bereichen (Domänen) mit dem Ziel der gemeinsamen Auswertung.

Kandidatengen: Unter diesem Begriff werden Gene verstanden, die möglicherweise im Zusammenhang mit einer bestimmten phänotypischen Ausprägung stehen.

Markergestützte Selektion: Das Ziel der markergestützten Selektion besteht in der Identifikation der einzelnen Gene, die direkt oder indirekt mit einem quantitativen Merkmal gekoppelt sind [Bec93]. Dazu werden für diese Gene funktionale und diagnostische Marker entwickelt, um den genetischen Beitrag an einer phänotypischen Ausprägung zu ermitteln. In der Pflanzenzüchtung ist es damit möglich, Elternlinien mit gewünschten Eigenschaften zu kombinieren und innerhalb der spaltenden F2-Populationen anhand der Marker die Individuen zu selektieren, die die gewünschten Allele besitzen.

Pflanzengenetische Ressource: Als pflanzengenetische Ressource wird generativ oder vegetativ vermehrungsfähiges Material von Pflanzen bezeichnet, das einen aktuellen oder zumindest potentiellen Wert für Ernährung, Land- oder Forstwirtschaft hat. Dazu zählen auch Landrassen, verwandte Wildarten und Wildformen von Kulturpflanzen [OBB95].

Phänotyp: Der Phänotyp ist die Gesamtheit aller beobachtbaren Merkmale eines Organismus.

Sorte: Eine Sorte (engl. Cultivar) ist eine Variante einer Kulturpflanzenart, die sich in mindestens einem Merkmal deutlich von anderen Sorten dieser Art unterscheidet. Sorten werden von Behörden, in Deutschland dem Bundessortenamt, zugelassen und tragen einen Sortennamen.

Literaturverzeichnis

[ABD+89] ATKINSON, M. P., F. BANCILHON, D. J. DEWITT, K. R. DITTRICH, D. MAIER, and S. B. ZDONIK: *The Object-Oriented Database System Manifesto*. In *Proceedings of the 1st International Conference on Distributed and Object-Oriented Databases (DOOD'89), Kyoto, Japan*, pages 223–240, December 1989.

[ABW04] APWEILER, R., A. BAIROCH, and C.H. WU: *Protein sequence databases*. Curr. Opin. Chem. Biol., 8(1):76–80, 2004.

[ACH+00] ADAMS, M.D., S.E. CELNIKER, R.A. HOLT, and OTHERS: *The Genome Sequence of Drosophila melanogaster*. Science, 287(5461):2185–2195, March 2000.

[ADM01] ALERCIA, A., S. DIULGHEROFF, and T. METZ: *FAO/IPGRI Multi-Crop Passport Descriptors (MCPD)*. FAO (Food and Agriculture Organization of the United Nations) - IPGRI (International Plant Genetic Resources Institute), 2001.

[AKG+91] ADAMS, M.D., J.M. KELLEY, J.D. GOCAYNE, M. DUBNICK, M.H. POLYMEROPOULOS, H. XIAO, C.R. MERRIL, A. WU, B. OLDE, R.F. MORENO, A.R. KERLAVAGE, W.R. MCCOMBIE, and J.C. VENTER: *Complementary DNA sequencing: expressed sequence tags and human genome project*. Science, 252(5013):1651–1656, 1991.

[Ame05] AMERICAN NATIONAL STANDARDS INSTITUTE (ANSI): *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies (ANSI/NISO Z39.19-2005)*. National Information Standards Organization, published by NISO Press, Bethesda, Maryland, U.S.A., July 2005.

[Ano00] ANON.: *Appendix I. The identification of most original samples (MOS)*. In MAGGIONI, L., P. MARUM, N.R. SACKVILLE HAMILTON, M. HULDÉN, and E. LIPMAN (editors): *Report of a Working Group on Forages. Seventh Meeting. 18-20 November 1999, Elvas, Portugal*,

pages 214–217. International Plant Genetic Resources Institute, Rome, Italy, 2000.

[ATI+08] AVRAHAM, S., C.-W. TUNG, K. ILIC, P. JAISWAL, E.A. KELLOGG, S. MCCOUCH, A. PUJAR, S.Y. REISER, L.AND RHEE, M.M. SACHS, M. SCHAEFFER, L. STEIN, P. STEVENS, L. VINCENT, F. ZAPATA, and D. WARE: *The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations*. Nucleic Acids Research, 36(suppl_1):D449–D454, January 2008.

[Aug01] AUGEN, J.: *Information technology to the rescue!* Nature Biotechnology, 19(6s):BE39–BE40, June 2001.

[AZ98] ADRIAANS, P. and D. ZANTINGE: *Data Mining*. Addison Wesley, 1998.

[BA96] BRACHMAN, R. J. and T. ANAND: *The process of knowledge discovery in databases: A human centered approach*. In FAYYAD, U. M., G. PIATETSKY-SHAPIRO, P. SMYTH, and R. UTHURUSAMY (editors): *Advances in Knowledge Discovery and Data Mining*, chapter 2, pages 37–57. AAAI/MIT Press, 1996.

[BB96] BORK, P. and A. BAIROCH: *Go hunting in sequence databases but watch out for the traps*. Trends in Genetics, 12(10):425–427, 1996.

[BBC+99] BACKOFEN, R., F. BRY, P. CLOTE, H.-P. KRIEGEL, T. SEIDL und K. SCHULZ: *Bioinformatik - Aktuelles Schlagwort*. Informatik-Spektrum, 22(5):376–378, 1999.

[Bec93] BECKER, H.: *Pflanzenzüchtung*. Ulmer, Stuttgart, 1993.

[BEC+07] BARTHELMES, J., C. EBELING, A. CHANG, I. SCHOMBURG, and D. SCHOMBURG: *BRENDA, AMENDA and FRENDA: the enzyme information system in 2007*. Nucleic Acids Research, 35(suppl_1):D511–D514, January 2007.

[BH67] BALL, G.H. and D.J. HALL: *A clustering technique for summarizing multivariate data*. Behavioral Science, 12(2):153–155, March 1967.

[Bis00] BISSWANGER, H.: *Enzymkinetik*. WILEY-VCH, Weinheim et al., 3. Auflage, 2000.

[BK03] BRY, F. and P. KRÖGER: *A Computational Biology Database Digest: Data, Data Analysis, and Data Management*. Distributed and Parallel Databases, 13(1):7–42, January 2003.

[BL95] BENNETT, M.D. and I.J. LEITCH: *Nuclear DNA Amounts in Angiosperms*. Annals of Botany, 76(2):113–176, August 1995.

[BLSS04] BALKO, S., M. LANGE, R. SCHNEE, and U. SCHOLZ: *BioDataServer: an Applied Molecular Biological Data Integration Service*. In RAHM, E. (editor): *Data Integration in the Life Sciences*, volume 2994 of *Lecture Notes in Bioinformatics*, pages 140–155, Berlin, Heidelberg, 2004. Springer-Verlag.

[BR04] BARD, J.B.L and S.Y. RHEE: *Ontologies in Biology: Design, Applications and Future Challenges*. Nature Reviews Genetics, 5:213–222, March 2004.

[BWF+00] BERMAN, H.M., J. WESTBROOK, Z. FENG, G. GILLILAND, T.N. BHAT, H. WEISSIG, I.N. SHINDYALOV, and P.E. BOURNE: *The Protein Data Bank*. Nucleic Acids Research, 28(1):235–242, 2000.

[BWSD80] BOTSTEIN, D., R.L. WHITE, M. SKOLNICK, and R.W. DAVIS: *Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms*. American Journal of Human Genetics, 32(3):314–331, May 1980.

[BZK+07] BRADBURY, P.J., Z. ZHANG, D.E. KROON, T.M. CASSTEVENS, Y. RAMDOSS, and E.S. BUCKLER: *TASSEL: software for association mapping of complex traits in diverse samples*. Bioinformatics, 23(19):2633–2635, October 2007.

[Cat91] CATTELL, R.G.G.: *Object Data Management: Object-Oriented and Extended Relational Database Systems*. Addison-Wesley, 1991.

[CB74] CHAMBERLIN, D.D. and R.F. BOYCE: *SEQUEL: A Structured English Query Language*. In ALTSHULER, G., R. RUSTIN, and B. PLAGMAN (editors): *Proceedings of the 1974 ACM SIGFIDET (now SIGMOD) Workshop on Data Description, Access and Control*, volume 1, pages 249–264, Ann Arbor, Michigan, USA, May 1–3 1974. ACM, New York, NY, USA.

[CCJ+02] CHING, A., K. CALDWELL, M. JUNG, M. DOLAN, O. SMITH, S. TINGEY, M. MORGANTE, and A. RAFALSKI: *SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines*. BMC Genetics, 3(1):e19, October 2002.

[CCS93] CODD, E.F., S.B. CODD, and C.T. SALLEY: *Providing OLAP to User-Analysts: An IT Mandate*. White paper, E.F. Codd & Associates, 1993.

[CH67] COVER, T. and P. HART: *Nearest neighbor pattern classification*. IEEE Transactions on Information Theory, 13(1):21–27, January 1967.

[Che76] CHEN, P.: *The Entity-Relationship Model – Toward a Unified View of Data*. ACM Transactions on Database Systems, 1(1):9–36, March 1976.

[CHS+95] CAREY, M.J., L.M. HAAS, P.M. SCHWARZ, M. ARYA, W.F. CODY, R. FAGIN, M. FLICKNER, A.W. LUNIEWSKI, W. NIBLACK, D. PETKOVIC, J. THOMAS, J.H. WILLIAMS, and E.L. WIMMERS: *Towards heterogeneous multimedia information systems: The garlic approach*. In *Proceedings of the Fifth International Workshop on Research Issues in Data Engineering: Distributed Object Management (RIDE-DOM 1995)*, pages 124–131, Taipei, Taiwan, March 6–7 1995.

[CMDS03] COLHOUN, H.M., P.M. MCKEIGUE, and G. DAVEY SMITH: *Problems of reporting genetic associations with complex outcomes*. The Lancet, 361(9360):865–872, March 2003.

[CNM+00] CHEN, RS, P NADKARNI, L MARENCO, F LEVIN, J ERDOS, and PL MILLER: *Exploring Performance Issues for a Clinical Database Organized Using an Entity-Attribute-Value Representation*. Journal of the American Medical Informatics Association, 7(5):475–487, October 2000.

[Cod70] CODD, E.F.: *A Relational Model of Data for Large Shared Data Banks*. Communications of the ACM, 13(6):377–387, June 1970.

[COD71] CODASYL DATA BASE TASK GROUP: *Report of the CODASYL Data Base Task Group*. ACM, April 1971.

[Cod82] CODD, E.F.: *Relational database: A practical foundation for productivity*. Communications of the ACM, 25(2):109–117, February 1982.

[Con97] CONRAD, S.: *Föderierte Datenbanksysteme - Konzepte der Datenintegration*. Springer-Verlag, Berlin, Heidelberg, 1. Auflage, 1997.

[CW92] CLARK, A.G. and T.S. WHITTAM: *Sequencing Errors and Molecular Evolutionary Analysis*. Molecular Biology and Evolution, 9(4):744–752, July 1992.

[CWL+08] COCKRAM, J., J. WHITE, F. LEIGH, V. LEA, E. CHIAPPARINO, D. LAURIE, I. MACKAY, W. POWELL, and D. O'SULLIVAN: *Association mapping of partitioning loci in barley*. BMC Genetics, 9(1):e16, February 2008.

[Dam64] DAMERAU, F.J.: *A technique for computer detection and correction of spelling errors*. Communications of the ACM, 7(3):171–176, March 1964.

[Del99] DELLAPENNA, D.: *Nutritional Genomics: Manipulating Plant Micronutrients to Improve Human Health*. Science, 285(5426):375–379, July 1999.

[Dic45] DICE, L.R.: *Measures of the Amount of Ecologic Association Between Species*. Ecology, 26(3):297–302, 1945.

[Dix60] DIXON, W.J.: *Simplified Estimation from Censored Normal Samples*. The Annals of Mathematical Statistics, 31(2):385–391, June 1960.

[DLR77] DEMPSTER, A., N. LAIRD, and D. RUBIN: *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, Series B, 39(1):1–38, 1977.

[DM88] DEVLIN, B.A. and P.T. MURPHY: *An architecture for a business and information system*. IBM Systems Journal, 27(1):60–80, 1988.

[DOB95] DAVIDSON, S.B., C. OVERTON, and P. BUNEMAN: *Challenges in Integrating Biological Data Sources*. Journal of Computational Biology, 2(4):557–572, 1995.

[EBS+06] EICHMANN, R., S. BIEMELT, P. SCHÄFER, U. SCHOLZ, C. JANSEN, A. FELK, W. SCHÄFER, G. LANGEN, U. SONNEWALD, K.H. KOGEL, and R. HÜCKELHOVEN: *Macroarray expression analysis of barley susceptibility and nonhost resistance to blumeria graminis*. Journal of Plant Physiology, 163(6):657–670, April 2006.

[Eck02] ECKERSON, W.W.: *Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data*. TDWI Report Series, The Data Warehousing Institute, Seattle, USA, February 2002.

[EGH+92] ENGELS, G., M. GOGOLLA, U. HOHENSTEIN, K. HÜLSMANN, P. LÖHR-RICHTER, G. SAAKE, and H.-D. EHRICH: *Conceptual modelling of database applications using an extended ER model*. Data & Knowledge Engineering, 9(2):157–204, 1992.

[EHB03] ETZOLD, T., H. HARRIS, and S. BEAULAH: *SRS: An Integration Platform for Databanks and Analysis Tools in Bioinformatics*. In LACROIX, Z. and T. CRITCHLOW (editors): *Bioinformatics: Managing Scientific Data*, chapter 5, pages 109–145. Morgan Kaufmann Publishers, 2003.

[FAO05] FAO (FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS): *Major Food and Agricultural Commodities and Producers*, 2005. http://www.fao.org/es/ess/top/commodity.html?lang=en&item=44 &year=2005 [Stand 2009-04-03].

[Fel92] FELDMAN, M.W.: *Heritability: some theoretical ambiguities*. In KELLER, E.F. and E.A. LLOYD (editors): *Keywords in Evolutionary Biology*, pages 151–157. Harvard Universtity Press, Cambridge, USA, 1992.

[FHJ+90] FRIEDMAN, C., G. HRIPCSAK, S. JOHNSON, J. CIMINO, and P. CLAYTON: *A generalized relational schema for an integrated clinical patient database*. In *Proc 14th Symp Comput Appl Med Care*, pages 335–339, 1990.

[Fis18] FISHER, R.A.: *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*. Transactions of the Royal Society of Edinburgh, 52:399–433, 1918.

[FPL+03] FOX, G.P., J.F. PANOZZO, C.D. LI, R.C.M. LANCE, P.A. INKERMAN, and R.J. HENRY: *Molecular basis of barley quality*. Australian Journal of Agricultural Research, 54(12):1081–1101, 2003.

[FPSM92] FRAWLEY, W. J., G. PIATETSKY-SHAPIRO, and C. J. MATHEUS: *Knowledge discovery in databases - an overview*. AI Magazine, 13:57–70, 1992.

[FPSS96a] FAYYAD, U. M., G. PIATETSKY-SHAPIRO, and P. SMYTH: *From data mining to knowledge discovery: An overview.* In FAYYAD, U., G. PIATETSKY-SHAPIRO, P. SMYTH, and R. UTHURUSAMY (editors): *Advances in Knowledge Discovery and Data Mining (AKDDM)*, pages 1–30, Menlo Park, Calif., 1996. AAAI Press.

[FPSS96b] FAYYAD, U.M., G. PIATETSKY-SHAPIRO, and P. SMYTH: *Knowledge discovery and data mining: Towards a unifying framework.* In *Proc. 2nd Int'l. Conf. on Knowledge Discovery and Data Mining (KDD-96)*, pages 82–88, Menlo Park, CA, 1996. AAAI Press.

[FS69] FELLEGI, I.P. and A.B. SUNTER: *A theory of record linkage*. Journal of the American Statistical Association, 64(328):1183–1210, December 1969.

[FSP07] FALUSH, D., M. STEPHENS, and J.K. PRITCHARD: *Inference of population structure using multilocus genotype data: dominant markers and null alleles*. Molecular Ecology Notes, 7(4):574–578, 2007.

[FWKG06] FUNKE, T., S. WEISE, H. KNÜPFFER und I. GROSSE: *Ein neues Gesicht für die Europäische Gerstendatenbank (EBDB)*. In: *Ausgewählte Vorträge aus GPZ-Arbeitsgemeinschaften*, Band 70 der Reihe *Vorträge für Pflanzenzüchtung*, Seiten 79–80, Göttingen, März 2006. Gesellschaft für Pflanzenzüchtung e. V. (GPZ).

[Gal85] GALTON, F.: *The British Association, Section H, Anthropology, Opening Address by Francis Galton*. Nature, 32(830):507–510, September 1885.

[Gal86] GALTON, F.: *Regression Towards Mediocrity in Hereditary Stature*. The Journal of the Anthropological Institute of Great Britain and Ireland, 15:246–263, 1886.

[GBA04] GARCÍA-BERTHOU, E. and C. ALCARAZ: *Incongruence between test statistics and P values in medical papers*. BMC Medical Research Methodology, 4:e13, May 2004.

[GBJK+08] GRAFAHREND-BELAU, E., B.H. JUNKER, D. KOSCHÜTZKI, C. KLUKAS, S. WEISE, U. SCHOLZ, and F. SCHREIBER: *Towards systems biology of developing barley grains: A framework for modeling metabolism*. In AHDESMÄKI, M., K. STRIMMER, N. RADDE, J. RAHNENFÜHRER, K. KLEMM, H. LÄHDESMÄKI, and O. YLI-HARJA (editors): *Proceedings of the Fifth International Workshop on Computational Systems Biology, WCSB 2008, June 11–13 2008, Leipzig, Germany*, pages 41–44, 2008.

[GBWK+08] GRAFAHREND-BELAU, E., S. WEISE, D. KOSCHÜTZKI, U. SCHOLZ, B.H. JUNKER, and F. SCHREIBER: *MetaCrop: a detailed database of crop plant metabolism*. Nucleic Acids Research, 36(suppl_1):D954–D958, January 2008.

[GC09] GALPERIN, M.Y. and G.R. COCHRANE: *Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009*. Nucleic Acids Research, 37(suppl_1):D1–D4, 2009.

[Gib96] GIBBON, G.A.: *A brief history of LIMS*. Laboratory Automation and Information Management, 32(1):1–5, May 1996.

[GLCSF95] GOLDSTEIN, D.B., A.R. LINARES, L.L. CAVALLI-SFORZA, and M.W. FELDMAN: *An Evaluation of Genetic Distances for Use With Microsatellite Loci*. Genetics, 139(1):463–471, January 1995.

[GO 08] GO (THE GENE ONTOLOGY CONSORTIUM): *The Gene Ontology project in 2008*. Nucleic Acids Research, 36(suppl_1):D440–D444, January 2008.

[GRL+02] GOFF, STEPHEN A., DARRELL RICKE, TIEN-HUNG LAN, GERNOT PRESTING, RONGLIN WANG, MOLLY DUNN, JANE GLAZEBROOK, ALLEN SESSIONS, PAUL OELLER, HEMANT VARMA, and OTHERS: *A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica)*. Science, 296(5565):92–100, 2002.

[Gru93] GRUBER, T.R.: *A Translation Approach to Portable Ontology Specifications*. Knowledge Acquisition, 5(2):199–220, 1993.

[Gru95] GRUBER, T. R.: *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. International Journal Human-Computer Studies, 43(5/6):907–928, November 1995.

[Ham50] HAMMING, R.W.: *Error-detecting and error-correcting codes*. Bell System Technical Journal, 29(2):147–160, 1950.

[Har75] HARTIGAN, J.: *Clustering Algorithms*. Wiley & Sons, New York, 1975.

[HCMC+03] HAYES, P.M., A. CASTRO, L. MARQUEZ-CEDILLO, A. COREY, C. HENSON, B.L. JONES, J. KLING, D. MATHER, I. MATUS, C. ROSSI, and K. SATO: *Genetic diversity for quantitatively inherited agronomic and malting quality traits*. In BOTHMER, R. VON, T.J.L. VAN HINTUM, H. KNÜPFFER, and K. SATO (editors): *Diversity in Barley (Hordeum vulgare)*, volume 7 of *Developments in Plant Genetics and Breeding*, chapter 10, pages 201–226. Elsevier, 3 July 2003.

[HFS+03] HUCKA, M., A. FINNEY, H. M. SAURO, H. BOLOURI, J. C. DOYLE, H. KITANO, A. P. ARKIN, B. J. BORNSTEIN, D. BRAY, A. CORNISH-BOWDEN, A. A. CUELLAR, S. DRONOV, E. D. GILLES, M. GINKEL, V. GOR, I. I. GORYANIN, W. J. HEDLEY, T. C. HODGMAN, J.-H. HOFMEYR, P. J. HUNTER, N. S. JUTY, J. L. KASBERGER, A. KREMLING, U. KUMMER, N. LE NOVÈRE, L. M. LOEW, D. LUCIO, P. MENDES, E. MINCH, E. D. MJOLSNESS, Y. NAKAYAMA, M. R. NELSON, P. F. NIELSEN, T. SAKURADA, J. C. SCHAFF, B. E. SHAPIRO, T. S. SHIMIZU, H. D. SPENCE, J. STELLING, K. TAKAHASHI, M. TOMITA, J. WAGNER, and J. WANG: *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models*. Bioinformatics, 19(4):524–531, 2003.

[HHM+07] HUSE, S., J. HUBER, H. MORRISON, M. SOGIN, and D. WELCH: *Accuracy and quality of massively parallel dna pyrosequencing*. Genome Biology, 8(7):R143, 2007.

[HMF02] HECKENBERGER, M., A.E. MELCHINGER und M. FRISCH: *Verwendung von Datenbankprogrammen in der Pflanzenzüchtung am Beispiel MS Access*. In: *Bericht der 53. Jahrestagung der Vereinigung der Pflanzenzüchter und Saatgutkaufleute Österreichs*, Seiten 47–53, 26.-28. November 2002.

[HNFH06] HERZ, M., U. NICKL, K. FINK und G. HENKELMANN: *Landessortenversuch Gerste, Faktorielle Sortenversuche, Brauqualität und Kornphysikalische Untersuchungen*. Bayerische Landesanstalt für Landwirtschaft, Institut für Pflanzenbau und Pflanzenzüchtung, 2006.

[Hot33] HOTELLING, H.: *Analysis of a complex of statistical variables into principal components*. Journal of Educational Psychology, 24:417–441, 1933.

[HS97] HEUER, A. und G. SAAKE: *Datenbanken - Konzepte und Sprachen*. International Thomson Publishing, Bonn, Albany, Attenkirchen, 1997.

[HSK+01] HAAS, L. M., P. M. SCHWARZ, P. KODALI, E. KOTLAR, J. E. RICE, and W. C. SWOPE: *Discoverylink: A system for integrated access to life sciences data sources*. IBM Systems Journal, 40(2):489–511, 2001.

[HV02] HARDY, O.J. and X. VEKEMANS: *SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels*. Molecular Ecology Notes, 2(4):618–620, December 2002.

[Ibe92] IBELGAUFTS, H.: *Gentechnologie von A bis Z*. VCH-Verlagsgesellschaft, Weinheim, 1992.

[Inm99] INMON, W.H.: *Building the operational data store*. John Wiley & Sons, 2nd edition, 1999.

[Inm05] INMON, W.H.: *Building the Data Warehouse*. John Wiley & Sons, 4th edition, October 2005.

[Int04] INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM: *Finishing the euchromatic sequence of the human genome*. Nature, 431(7011):931–945, October 2004.

[IPG94] IPGRI: *Descriptors for Barley (Hordeum vulgare L.)*. International Plant Genetic Resources Institute, Rome, Italy, 1994.

[IPS05] IPS (INSTITUTE FOR PLANT PROTECTION): *Monitoring of the Environmental Effects of the Bt Gene*. Final report, Bayerische Landesanstalt für Landwirtschaft, August 2005.

[Irr05] IRRIZARRY, R.A. ET AL.: *Multiple-laboratory comparison of microarray platforms*. Nature Methods, 2(5):345–349, May 2005.

[Jac01] JACCARD, P.: *Étude comparative de la distribution florale dans une portion des Alpes et des Jura. [Comparative study of the distribution of flora in a region of the Alps and the Jura]*. Bulletin del la Société Vaudoisedes Sciences Naturelles, 37:547–549, 1901.

[JM61] JACOB, F. and J. MONOD: *Genetic regulatory mechanisms in the synthesis of proteins*. Journal of Molecular Biology, 3:318–356, 1961.

[JWN+02] JAISWAL, P., D. WARE, J. NI, K. CHANG, W. ZHAO, S. SCHMIDT, X. PAN, K. CLARK, L. TEYTELMAN, S. CARTINHOUR, L. STEIN, and S. MCCOUCH: *Gramene: development and integration of trait and gene ontologies for rice*. Comparative and Functional Genomics, 3(2):132–136, 2002.

[KAG+08] KANEHISA, M., M. ARAKI, S. GOTO, M. HATTORI, M. HIRAKAWA, M.AND ITOH, T. KATAYAMA, S. KAWASHIMA, S. OKUDA, T. TOKIMATSU, and Y. YAMANISHI: *KEGG for linking genomes to life and the environment*. Nucleic Acids Research, 36(suppl_1):D480–D484, January 2008.

[KDKR05] KIRSTEN, T., H.-H. DO, C. KÖRNER, and E. RAHM: *Hybrid Integration of Molecular-Biological Annotation Data*. In LUDÄSCHER, B. and L. RASCHID (editors): *Data Integration in the Life Sciences, Second International Workshop, DILS 2005, San Diego, CA, USA, July 20-22, 2005. Proceedings*, volume 3615 of *Lecture Notes in Bioinformatics*, pages 208–223. Springer-Verlag Berlin Heidelberg, 2005.

[KDR04] KIRSTEN, T., H.-H. DO, and E. RAHM: *A Data Warehouse for Multidimensional Gene Expression Analysis*. Leipzig Bioinformatics Working Paper No. 1, Interdisciplinary Centre for Bioinformatics, University of Leipzig, November 2004.

[Ken38] KENDALL, M.: *A New Measure of Rank Correlation*. Biometrika, 30:81–89, 1938.

[KGH+06] KANEHISA, M., S. GOTO, M. HATTORI, K.F. AOKI-KINOSHITA, M. ITOH, S. KAWASHIMA, T. KATAYAMA, M. ARAKI, and M. HIRAKAWA: *From genomics to chemical genomics: new developments*

in KEGG. Nucleic Acids Research, 34(suppl_1):D354–D357, January 2006.

[KGM+07] KUENNE, C., I. GROSSE, I. MATTHIES, U. SCHOLZ, T. SRETENOVIC-RAJICIC, N. STEIN, A. STEPHANIK, B. STEUERNAGEL, and S. WEISE: *Using Data Warehouse Technology in Crop Plant Bioinformatics*. Journal of Integrative Bioinformatics, 4(1):e88, 2007.

[Kim98] KIMBALL, R.: *Bringing Up Supermarts – A step-by-step approach to building a data warehouse from granular data*. DBMS and Internet Systems, 11(1):47–53, January 1998.

[Kit02] KITANO, H.: *Systems biology: A brief overview*. Science, 295(5560):1662–1664, March 2002.

[KKS+04] KASPRZYK, A, D KEEFE, D SMEDLEY, D LONDON, W SPOONER, C MELSOPP, M HAMMOND, P ROCCA-SERRA, T COX, and E BIRNEY: *EnsMart: A Generic System for Fast and Flexible Access to Biological Data*. Genome Research, 14(1):160–169, January 2004.

[KLF+05] KÜNNE, C., M. LANGE, T. FUNKE, H. MIEHE, T. THIEL, I. GROSSE, and U. SCHOLZ: *CR-EST: a resource for crop ESTs*. Nucleic Acids Research, 33(suppl_1):D619–D621, January 2005.

[Kli08] KLIER, M.: *Metriken zur Bewertung der Datenqualität – Konzeption und praktischer Nutzen*. Informatik-Spektrum, 31(3):223–236, 2008.

[Knü01] KNÜPFFER, H.: *Handling of characterization and evaluation data in crop databases*. In MAGGIONI, L. and O. SPELLMAN (editors): *Report of a Network Coordinating Group on Cereals. Ad hoc meeting, 7-8 July 2000, Radzików, Poland*, pages 58–65, Rome, Italy, 2001. International Plant Genetic Resources Institute (IPGRI).

[KS91] KIM, W. and J. SEO: *Classifying Schematic and Data Heterogeneity in Multidatabase Systems*. IEEE Computer, 24(12):12–18, December 1991.

[KvH95] KNÜPFFER, H. and T.J.L. VAN HINTUM: *The Barley Core Collection - an international effort*. In HODGKIN, T., A.H.D. BROWN, T.J.L. VAN HINTUM, and E.A.V. MORALES (editors): *Core Collections of Plant Genetic Resources*, pages 171–178. Wiley & Sons, Chichester, U.K., 1995.

[Lev66] LEVENSHTEIN, V: *Binary Codes of Correcting Deletions, Insertions and Reversals*. Soviet Physics Doklady, 10(8):707–710, 1966.

[LPW+06] LEE, TJ, Y POULIOT, V WAGNER, P GUPTA, DWJ STRINGER-CALVERT, JD TENENBAUM, and PD KARP: *Biowarehouse: a bioinformatics database warehouse toolkit*. BMC Bioinformatics, 7:e170, March 2006.

[LR96] LAIT, A. and B. RANDELL: *An assessment of name matching algorithms*. Technical Report No. 550, Department of Computing Science, University of Newcastle upon Tyne, 1996.

[Lus02] LUSTI, M.: *Data Warehousing und Data Mining - Eine Einführung in entscheidungsunterstützende Systeme*. Springer-Verlag, Berlin, Heidelberg, 2. Auflage, 2002.

[LW67] LANCE, G.N. and W.T. WILLIAMS: *A general theory of classificatory sorting strategies. I. Hierarchical systems*. Computer Journal, 9:373–380, 1967.

[Mat08] MATTHIES, I., 2008. personal communication, 2008-10-23.

[MB06] METZGER, J.O. and U. BORNSCHEUER: *Lipids as renewable resources: current state of chemical and biotechnological conversion and diversification*. Applied Microbiology and Biotechnology, 71(1):13–22, June 2006.

[MBV05] MATULLO, G., M. BERWICK, and P. VINEIS: *Gene–Environment Interactions: How Many False Positives?* Journal of the National Cancer Institute, 97(8):550–551, April 2005.

[McC04] MCCOUCH, S.: *Diversifying Selection in Plant Breeding*. PLoS Biology, 2(10):e347, October 2004.

[McG77] MCGEE, W.C.: *The information management system IMS/VS - Part I: General structure and operation*. IBM Systems Journal, 16(2):84–95, 1977.

[McQ67] MCQUEEN, J.: *Some methods for classification and analysis of multivariate observations*. In LE CAM, L.M. and J. NEYMAN (editors): *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Statistical Laboratory of the University of California, Berkeley, June 21 – July 18, 1965 and December 27, 1965 – January 7, 1966, 1967.

[Mem05] MEMBERS OF THE TOXICOGENOMICS RESEARCH CONSORTIUM: *Standardizing global gene expression analysis between laboratories and across platforms*. Nature Methods, 2(5):351–356, May 2005.

[Mer97] MERTENS, P.: *Integrierte Informationsverarbeitung*. Betriebswirtschaftlicher Verlag Dr. Th. Gabler GmbH, Wiesbaden, 11. Auflage, 1997.

[MFR07] MATTHIES, I.E., K. FOERSTER, and M.S. RÖDER: *GABI-MALT: An integrated approach to the genetic and functional dissection of malting quality in barley Subproject 4: SNP-detection and haplotype analysis in candidate genes for malting*. In *GABI-PROGRESS-REPORT*, pages 30–32. 2007.

[MG77] MAXAM, A.M. and W. GILBERT: *A new method for sequencing DNA*. PNAS, 74(2):560–564, February 1977.

[MGF98] MORICO, G., F. GRASSI, and C. FIDEGHELLI: *Horticultural Genetic Diversity: Conservation and Sustainable Utilization and Related International Agreements*. In *World Conference on Horticultural Research, 17-20 June 1998*, Rome, Italy, 1998.

[MNF03] MÜLLER, H., F. NAUMANN, and J.-C. FREYTAG: *Data quality in genome databases*. In *Proceedings of the Conference on Information Quality (IQ 03)*, Boston, October 2003.

[MP07] MACKAY, I. and W. POWELL: *Methods for linkage disequilibrium mapping in crops*. Trends in Plant Science, 12(2):57–63, February 2007.

[MPL01] MCLAREN, C. G., A. PORTUGAL, and J. G. F. LIESHOUT: *Design of the data management system (DMS)*. Technical report, International Crop Information System, 2001.

[MWBL05] MÜLLER, H., M. WEIS, J. BLEIHOLDER und U. LESER: *Erkennen und Bereinigen von Datenfehlern in naturwissenschaftlichen Daten*. Datenbank-Spektrum, 15:26–35, 2005.

[MWFR09] MATTHIES, I.E., S. WEISE, J. FÖRSTER, and M.S. RÖDER: *Association mapping and marker development of the candidate genes (1→3),(1→4)-β-D-Glucan-4-glucanohydrolase and (1→4-β-Xylan-endohydrolase 1 for malting quality in barley.* Euphytica, 2009. DOI: 10.1007/s10681-009-9915-6.

[MWR09] MATTHIES, I.E., S. WEISE, and M.S. RÖDER: *Association of haplotype diversity in the α-amylase gene amy1 with malting quality parameters in barley*. Molecular Breeding, 23(1):139–152, January 2009.

[Nat99] *Freely associating*. Nature Genetics, 22(1):1–2, May 1999. editorial.

[NB98] NADKARNI, P.M. and C. BRANDT: *Data Extraction and Ad Hoc Query of an Entity-Attribute-Value Database*. Journal of the American Medical Informatics Association, 5(6):511–527, December 1998.

[NIB07] *News in brief*. Nature, 447(7147):897, June 2007.

[NKAJ59] NEWCOMBE, H.B., J.M. KENNEDY, S.J. AXFORD, and A.P. JAMES: *Automatic linkage of vital records*. Science, 130(3381):954–959, October 1959.

[NMC+99] NADKARNI, P.M., L. MARENGO, R. CHEN, S. SKOUFOS, G. SHEPHERD, and P. MILLER: *Organization of Heterogeneous Scientific Data Using the EAV/CR Representation*. Journal of the American Medical Informatics Association, 6(6):478–493, December 1999.

[OBB95] OETMANN, A., R. BROCKHAUS, and F. BEGEMANN: *Deutscher Bericht zur Vorbereitung der 4. Internationalen Technischen Konferenz der FAO über pflanzengenetische Ressourcen (4. ITKPGR) vom 17.-23. Juni 1996 in Leipzig*. Bericht, Sekretariat des Nationalen Komitees zur Vorbereitung der 4. ITK-PGR bei der Zentralstelle für Agrardokumentation und -information (ZADI), Informationszentrum für Genetische Ressourcen (IGR), Bonn, 1995.

[OK06] OPPERMANN, M. und H. KNÜPFFER: *GBIS – Das neue Genbankinformationssystem am IPK*. In: *Ausgewählte Vorträge aus GPZ-Arbeitsgemeinschaften*, Band 70 der Reihe *Vorträge für Pflanzenzüchtung*, Seiten 47–49, Göttingen, März 2006. Gesellschaft für Pflanzenzüchtung e. V. (GPZ).

[Ora09a] ORACLE CORPORATION: *Oracle Application Express*, 2009. http://www.oracle.com/technology/products/database/application_express/index.html [Stand 2009-04-03].

[Ora09b] ORACLE CORPORATION: *Oracle SQL*Loader*, 2009. http://www.oracle.com/technology/products/database/utilities/htdocs/sql_loader_overview.html [Stand 2009-04-03].

[Ora09c] ORACLE CORPORATION: *Oracle Warehouse Builder*, 2009. http://www.oracle.com/technology/products/warehouse/index.html [Stand 2009-04-03].

[PC95] PENDSE, N. and R. CREETH: *Fast Analysis of Shared Multidimensional Information (FASMI)*. The OLAP Report, 1995. http://www.olapreport.com/fasmi.htm [Stand 2009-04-03].

[Pea96] PEARSON, K.: *Mathematical contributions to the theory of evolution. iii. regression, heredity and panmixia*. Philosophical Transactions of the Royal Society of London, 187:253–318, 1896.

[Pea01] PEARSON, K.: *On lines and planes of closest fit to systems of points in space*. The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, 6(2):559–572, 1901.

[Pen05] PENNISI, E.: *How Will Big Pictures Emerge From a Sea of Biological Data?* Science, 309(5731):94, July 2005.

[PGMW95] PAPAKONSTANTINOU, Y., H. GARCIA-MOLINA, and J. WIDOM: *Object exchange across heterogeneous information sources*. In YU, P. S. and A. L. P. CHEN (editors): *11th Conference on Data Engineering*, pages 251–260, Taipei, Taiwan, 1995. IEEE Computer Society.

[PK06] PHILIPPI, S. and J. KÖHLER: *Addressing the problems with life-science databases for traditional uses and systems biology*. Nature Reviews Genetics, 7(6):482–488, June 2006.

[Pos69] POSTEL, H.J.: *Die Kölner Phonetik – Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse*. IBM-Nachrichten, 19:925–931, 1969.

[PS91] PIATETSKY-SHAPIRO, G.: *Knowledge Discovery in real Databases: A Report on the IJCAI-89 Workshop*. AI Magazine, 11(5):68–70, 1991.

[PSMSU94] PIATETSKY-SHAPIRO, G., C. MATHEUS, P. SMYTH, and R. UTHURUSAMY: *KDD-93 - progress and challenges in knowledge discovery in databases*. AI Magazine, 15:77–82, 1994.

[PV94] PHILLIPS, R.L. and I.K. VASIL (EDS.): *DNA-Based Markers in Plants*, volume 1 of *Advances in Cellular and Molecular Biology of Plants*. Kluwer Academic Publishers, Dordrecht, Boston, London, 1994.

[Ref07] REFERAT 226 (KOORDINATION DER UMWELTANGELEGENHEITEN, BIOLOGISCHE VIELFALT, GENETISCHE RESSOURCEN): *Agrobiodiversität erhalten, Potenziale der Land-, Forst- und Fischereiwirtschaft erschließen und nachhaltig nutzen*. Strategiepapier, Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz, Bonn, Dezember 2007.

[RH98] RIBAUT, J.-M. and D. HOISINGTON: *Marker-assisted selection: new tools and strategies*. Trends in Plant Science, 3(6):236–239, June 1998.

[RHM04] RIEGER, P., S. HEYMANN und H. MÜLLER: *Datenbankgestützte Wissensakquisition in den Lebenswissenschaften*. Datenbank-Spektrum, 10:14–21, 2004.

[RK98] RUNGGALDIER, E. und C. KANZIAN: *Grundprobleme der Analytischen Ontologie*. Schöningh, Paderborn - München - Wien - Zürich, 1998.

[RKL07] RAHM, E., T. KIRSTEN, and J. LANGE: *The GeWare data warehouse platform for the analysis of molecular-biological and clinical data*. Journal of Integrative Bioinformatics, 4(1):e47, 2007.

[RL05] RIECHE, S. und U. LESER: *Versionierung in relationalen Datenbanken*. In: *Studierendenprogramm der 11. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW 2005)*, Karlsruhe, März 2005. Gesellschaft für Informatik e.V.

[RMB00] RUH, W. A., F. X. MAGINNIS, and W. J. BROWN: *Enterprise Application Integration: A Wiley Tech Brief*. John Wiley & Sons, Chichester, England, October 2000.

[RMT+04] ROTHER, L., H. MÜLLER, S. TRISSL, I. KOCH, T. STEINKE, R. PREISSNER, C. FRÖMMEL, and U. LESER: *COLUMBA: Multidimensional Data Integration of Protein Annotations*. In RAHM, E. (editor): *Data Integration in the Life Sciences*, volume 2994 of *Lecture Notes in Bioinformatics*, pages 156–171, Berlin, Heidelberg, 2004. Springer-Verlag.

[Roo01] ROOS, D.S.: *Computational Biology: Bioinformatics – Trying to Swim in a Sea of Data*. Science, 291(5507):1260–1261, February 2001.

[RRW+06] RADCHUK, R., V. RADCHUK, W. WESCHKE, L. BORISJUK, and H. WEBER: *Repressing the expression of the SUCROSE NONFERMENTING-1-RELATED PROTEIN KINASE gene in pea embryo causes pleiotropic defects of maturation similar to an abscisic acid-insensitive phenotype*. Plant Physiology, 140(1):263–278, January 2006.

[RS97] ROTH, M. T. and P. SCHWARZ: *Don't scrap it, wrap it! - a wrapper architecture for legacy data sources*. In JARKE, M., ;.J. CAREY, K.R. DITTRICH, F.H. LOCHOVSKY, P. LOUCOPOULOS, and M.A. JEUSFELD (editors): *Proceedings of 23rd International Conference on Very Large Data Bases (VLDB'97)*, Athen, August 25-29 1997.

[Rub78] RUBIN, D.B.: *Multiple Imputation in Sample Surveys – a phenomenological Bayesian approach to nonresponse*. In *Proceedings of the*

Section on Survey Research Methods, American Statistical Association, pages 20–34, 1978.

[Rus18] RUSSEL, R.C.: *Index*. US patent 1,261,167, April 2 1918. pp. 1–4, http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=1,261,167 [Stand 2009-04-06].

[Sch91] SCHWAB, G.: *Fehlende Werte in der angewandten Statistik*. Deutscher Universitätsverlag, 1991.

[Sch94] SCHWARZE, J.: *Grundlagen der Statistik I – Beschreibende Verfahren*. Verlag Neue Wirtschafts-Briefe, Herne/Berlin, 7. Auflage, 1994.

[Sch98] SCHMITT, R.I.: *Schemaintegration für den Entwurf Föderierter Datenbanken*. Dissertation, Otto-von-Guericke-Universität, Fakultät für Informatik, Magdeburg, 1998.

[Sch02] SCHOLZ, U.: *FRIDAQ – Ein Framework zur Integration molekularbiologischer Datenbestände*. Berichte aus der Medizinischen Informatik und Bioinformatik. Shaker-Verlag, Aachen, 2002.

[Sea03] SEARLS, D.B.: *Data integration – connecting the dots*. Nature Biotechnology, 21(8):844–845, August 2003.

[Sei03] SEITZ, A.: *MARKER-DB - ein Informationssystem zur Verwaltung von molekularen Markern: Anwendungsentwicklung*. Studienarbeit, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2003.

[SGGC01] SCHELLER, J., K. H. GUHRS, F. GROSSE, and U. CONRAD: *Production of spider silk proteins in tobacco and potato*. Nature Biotechnology, 19(6):573–577, June 2001.

[SH99] SAAKE, G. und A. HEUER: *Datenbanken: Implementierungstechniken*. MITP-Verlag GmbH, Bonn, 1. Auflage, 1999.

[SHX+05] SHAH, SP, Y HUANG, T XU, MMS YUEN, J LING, and BFF OUELLETTE: *Atlas - a data warehouse for integrative bioinformatics*. BMC Bioinformatics, 6:e34, February 2005.

[SK02] SCHULZE-KREMER, S.: *Ontologies for molecular biology and bioinformatics*. In Silico Biology, 2(3):179–193, 2002.

[SKSB00] SCHÖNBACH, C., P. KOWALSKI-SAUNDERS, and V. BRUSIC: *Data warehousing in molecular biology*. Briefings in Bioinformatics, 1(2):190–198, 2000.

[SL90] SHETH, A. and J.A. LARSON: *Federated database systems for managing distributed, heterogenous, and autonomous databases*. ACM Computing Surveys, 22(3):183–236, 1990.

[SMS+02] SPELLMAN, P., M. MILLER, J. STEWART, C. TROUP, U. SARKANS, S. CHERVITZ, D. BERNHART, G. SHERLOCK, C. BALL, M. LEPAGE, M. SWIATEK, W.L. MARKS, J. GONCALVES, S. MARKEL, D. IORDAN, M. SHOJATALAB, A. PIZARRO, J. WHITE, R. HUBLEY, E. DEUTSCH, M. SENGER, B. ARONOW, A. ROBINSON, D. BASSETT, C. STOECKERT, and A. BRAZMA: *Design and implementation of microarray gene expression markup language (MAGE-ML)*. Genome Biology, 3(9):research0046.1–research0046.9, 2002.

[SNC77] SANGER, F., S. NICKLEN, and A.R. COULSON: *DNA sequencing with chain-terminating inhibitors*. PNAS, 74(12):5463–5467, December 1977.

[Sne34] SNEDECOR, G.W.: *Calculation and Interpretation of Analysis of Variance and Covariance*. Collegiate Press, Inc., Ames, Iowa, USA, 1934.

[Sof03] SOFFNER, M.: *MARKER-DB - ein Informationssystem zur Verwaltung von molekularen Markern: Entwicklung der Datenhaltungskomponente*. Studienarbeit, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2003.

[Spe04] SPEARMAN, C.: *The proof and measurement of association between two rings*. American Journal of Psychology, 15:72–101, 1904.

[SSDB95] SCHENA, M., D. SHALON, R.W. DAVIS, and P.O. BROWN: *Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray*. Science, 270(5235):467–470, October 1995.

[Stu08] STUDENT (W.S. GOSSETT): *The Probable Error of a Mean*. Biometrika, 6(1):1–25, 1908.

[Sun09] SUN MICROSYSTEMS: *JDBC Overview*, 2009. http://java.sun.com/products/jdbc/overview.html [Stand 2009-04-06].

[SW81] SMITH, T. and M. WATERMAN: *Identification of Common Molecular Subsequences*. Journal of Molecular Biology, 147(1):195–197, March 1981.

[Tan57] TANIMOTO, T.T.: *Internal report*. IBM Technical Report Series, November 1957.

[THL06] TILMAN, D., J. HILL, and C. LEHMAN: *Carbon-Negative Biofuels from Low-Input High-Diversity Grassland Biomass*. Science, 314(5805):1598–1600, December 2006.

[TK78] TSICHRITZIS, D. and A. KLUG: *The ANSI/X3/SPARC DBMS Framework Report of the Study Group on Database Management Systems*. Information Systems, 3(3):173–191, 1978.

[TMVG03] THIEL, T., W. MICHALEK, R.K. VARSHNEY, and A. GRANER: *Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.)*. Theoretical and Applied Genetics, 106(3):411–422, February 2003.

[TRM+05] TRISSL, S., L. ROTHER, H. MÜLLER, T. STEINKE, I. KOCH, R. PREISSNER, C. FRÖMMEL, and U. LESER: *Columba: An integrated database of proteins, structures, and annotations*. BMC Bioinformatics 2005, 6:e81, 2005.

[Try39] TRYON, R. C.: *Cluster Analysis*. Edwards Brothers, Ann Arbor, 1939.

[Tuk62] TUKEY, J.W.: *The Future of Data Analysis*. The Annals of Mathematical Statistics, 33(1):1–67, March 1962.

[TYPB89] TANKSLEY, S.D., N.D. YOUNG, A.H. PATERSON, and M.W. BONIERBALE: *RFLP Mapping in Plant Breeding: New Tools for an Old Science*. Nature Biotechnology, 7(3):257–264, March 1989.

[UML07] *UML Version 2.2*. Object Management Group (OMG), 2007. [http://www.omg.org/spec/UML/2.2] [Stand 2009-04-06].

[VBJE08] VALI, Ü., M. BRANDSTRÖM, M. JOHANSSON, and H. ELLEGREN: *Insertion-deletion polymorphisms (indels) as genetic markers in natural populations*. BMC Genetics, 9(1):e8, 2008.

[VHB+95] VOS, P., R. HOGERS, M. BLEEKER, M. REIJANS, T. VAN DE LEE, M. HORNES, A. FRITERS, J. POT, J. PALEMAN, M. KUIPER, and M. ZABEAU: *AFLP: a new technique for DNA fingerprinting*. Nucleic Acids Research, 23(21):4407–4414, October 1995.

[vHH92] HINTUM, T. J. L. VAN and T. HAZEKAMP: *Genis Data Dictionary*. Centre for Plant Breeding and Reproduction Research (CPRO-DLO), Centre for Genetic Resources, The Netherlands (CGN), Wageningen, The Netherlands, 1992.

[vHK95] HINTUM, TH.J.L. VAN and H. KNÜPFFER: *Duplication within and between germplasm collections. i. tracing duplication on the basis of*

passport data. Genetic Resources and Crop Evolution, 42(2):127–133, 1995.

[Vor02] VORWALD, J.: *Verfügbarmachung von Evaluierungsdaten im Genbankinformationssystem (GBIS) des IPK Gatersleben*. In: *Bericht der 53. Jahrestagung der Vereinigung der Pflanzenzüchter und Saatgutkaufleute Österreichs*, Seiten 41–46, 26.-28. November 2002.

[WC53a] WATSON, J. D. and F. H. C. CRICK: *Genetical implications of the structure of deoxyribonucleic acid*. Nature, 171(4361):964–967, May 1953.

[WC53b] WATSON, J. D. and F. H. C. CRICK: *Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid*. Nature, 171(4356):737–738, April 1953.

[Wei73] WEINER, P.: *Linear pattern matching algorithms*. In *Proceedings of the 14th IEEE Annual Symposium on Switching and Automata Theory*, pages 1–11, October 1973.

[Wei91] WEISSMAHR, B.: *Ontologie*. Kohlhammer, Stuttgart, 2. Auflage, 1991.

[Wei05] WEISE, S.: *Development of the plant-specific database MetaCrop*. In *Abstracts of the Mini Symposium on Metabolome Analysis in Plants*, page 5, Gatersleben, Germany, 17–18 March 2005. Leibniz Institute of Plant Genetics and Crop Plant Research (IPK).

[WGK+06] WEISE, S., I. GROSSE, C. KLUKAS, D. KOSCHÜTZKI, U. SCHOLZ, F. SCHREIBER, and B.H. JUNKER: *Meta-All: a system for managing metabolic pathway information*. BMC Bioinformatics, 7(1):e465, October 2006.

[WH02] WEINER, M.P. and T.J. HUDSON: *Introduction to SNPs: discovery of markers for disease*. Biotechniques, 32(Supplement):S4–S13, June 2002.

[WHG+07] WEISE, S., S. HARRER, I. GROSSE, H. KNÜPFFER, and E. WILLNER: *The European Poa Database (EPDB)*. FAO–Bioversity Plant Genetic Resources Newsletter, No. 150:64–70, 2007.

[WI92] WEBB, E.C. and INTERNATIONAL UNION OF BIOCHEMISTRY AND MOLECULAR BIOLOGY (IUBMB). NOMENCLATURE COMMITTEE: *Enzyme Nomenclature 1992*. Academic Press Inc., San Diego, USA, September 1992.

[Wie96] WIEDERHOLD, G.: *Intelligent Integration of Information: A Special Double Issue of the Journal of Intelligent Information Systems, Volume 6, Numbers 2/3*. Kluwer Academic Publishers, Boston, Dordrecht, London, 1996.

[Wie97] WIEDERHOLD, G.: *Mediators in the Architecture of Future Information Systems*. In HUHNS, MICHAEL N. and MUNINDAR P. SINGH (editors): *Readings in Agents*, pages 185–196. Morgan Kaufmann, San Francisco, CA, USA, 1997.

[Win84] WINSTON, P.H.: *Artificial Intelligence*. Addison-Wesley, Reading, Mass, 2nd edition, 1984.

[WKV+03] WEISE, S., H. KNÜPFFER, J. VORWALD, U. SCHOLZ, and I. GROSSE: *Reorganisation of Characterisation Data and Evaluation Data at the IPK Genebank*. In MEWES, H.-W., V. HEUN, D. FRISHMAN, and S. KRAMER (editors): *Proceedings of the German Conference on Bioinformatics 2003 (GCB'03)*, volume II: Posters, pages 103–104, Garching, Germany, 12–13 October 2003.

[WKV+06] WEISE, S., H. KNÜPFFER, J. VORWALD, U. SCHOLZ und I. GROSSE: *Integration von phänotypischen Daten in das Plant Data Warehouse*. In: *Ausgewählte Vorträge aus GPZ-Arbeitsgemeinschaften*, Band 70 der Reihe *Vorträge für Pflanzenzüchtung*, Seiten 84–86, Göttingen, März 2006. Gesellschaft für Pflanzenzüchtung e. V. (GPZ).

[WSRM09] WEISE, S., U. SCHOLZ, M.S. RÖDER, and I.E. MATTHIES: *MetaBrew: A comprehensive database of malting quality traits in brewing barley*. Barley Genetics Newsletter, 39:1–4, 2009.

[YHW+02] YU, JUN, SONGNIAN HU, JUN WANG, GANE KA-SHU WONG, SONGGANG LI, BIN LIU, YAJUN DENG, LI DAI, YAN ZHOU, XIUQING ZHANG, and OTHERS: *A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica)*. Science, 296(5565):79–92, 2002.

[Yul00] YULE, G.U.: *On the Association of Attributes in Statistics*. Philosophical Transactions of the Royal Society of London, Ser. A, 194:257–319, 1900.

[ZFT+05] ZHANG, P., H. FOERSTER, C.P. TISSIER, L. MUELLER, S. PALEY, P.D. KARP, and S.Y. RHEE: *MetaCyc and AraCyc. Metabolic Pathway Databases for Plant Research*. Plant Physiology, 138(1):27–37, May 2005.

[ZMP08] ZMP (ZENTRALE MARKT- UND PREISBERICHTSTELLE FÜR ERZEUGNISSE DER LAND-, FORST- UND ERNÄHRUNGSWIRTSCHAFT

GMBH): *Erzeugerpreise für Braugerste weiterhin auf Höhenflug*, 2008. http://www.zmp.de/presse/agrarwoche/marktgrafik/2008_02_20_zmp-marktgrafik.asp [Stand 2009-04-06].

Index

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Insbesondere habe ich nicht die Hilfe eines kommerziellen Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

Magdeburg, den 22. April 2009