Verfügbarmachung von Evaluierungsdaten im Genbankinformationssystem (GBIS) des IPK Gatersleben

J. VORWALD

1 Ziele der Entwicklung von GBIS

Im Zuge der Vereinigung der deutschen Genbanken (Genbank der Bundesanstalt für Züchtungsforschung - BAZ - in Braunschweig, Institut für Pflanzengenetik und Kulturpflanzenforschung - IPK - in Gatersleben) am Standort Gatersleben bzw. an den Außenstellen des IPK in Malchow/Poel und Groß Lüsewitz ist der Neuaufbau eines Genbankinformationssystems (GBIS) geplant. Die bestehenden Datenbanken in Braunschweig (Anfänge aus den frühen 1970er Jahren, Neuaufbau ab 1999) und Gatersleben (Anfänge aus der Mitte der 1980er Jahre, Neuaufbau ab 1992, vgl. FREYTAG & KNÜPFFER, 1994) bzw. MALCHOW (1995/96) genügen nicht mehr dem Stand der Technik. Eine Ist-Standsanalyse (EPORTAS, 2002) kommt zu dem Schluss, dass eine Neuentwicklung deutliche Vorteile gegenüber einer Integration aller Daten in eines der vorhandenen Systeme erbringt. Darüber hinaus können neue Funktionalitäten implementiert werden, die in keinem der bestehenden Systeme integriert sind (z. B. Web-Schnittstelle für die Einbindung der Außenstellen des IPK). Ein Fortbestehen eines verteilten Systems würde den Administrationsaufwand über Gebühr erhöhen.

GBIS soll modular aufgebaut werden (*Abbildung 1*). Kernsystem wird wie bei allen bestehenden Systemen (vgl. auch das System des CGN, Wageningen/Niederlande, MENTING & VAN HINTUM, 2001) ein Passport-Modul für die in der Genbank zu verwaltenden pflanzengenetischen Ressourcen sein. Daran angegliedert werden zunächst weitere Module zur Verwaltung von Adress-, Anbau-, (Saatgut-)Management-, Transfer-, Taxonomie- und Evaluierungsdaten.

Das Modul zur Handhabung der Evaluierungsdaten soll im Mittelpunkt dieser Betrachtung stehen. Unter Evaluierungsdaten sollen Primär- und Sekundär-Evaluierungsdaten verstanden werden, die im Gegensatz zu den weitestgehend genetisch fixierten Charakterisierungsdaten hoch umweltabhängig sind.

Primärevaluierungsdaten werden in diesem Sinne meist von Genbanken beim Erstanbau bzw. der Folgereproduktion unter konstanten Standort- und Klimabedingungen erhoben, während Sekundärevaluierungsdaten einer weiteren Variabilität der Anbaubedingungen unterworfen sind und außerhalb der Genbanken von verschiedenen Institutionen (Züchter, Landwirte usw.) erhoben werden (KNÜPFFER, 2001).

Das Modul für Evaluierungsdaten des GBIS ist auch geeignet, Charakterisierungsdaten zu verwalten, da im datentechnischen Sinne keine Unterschiede auszumachen sind.

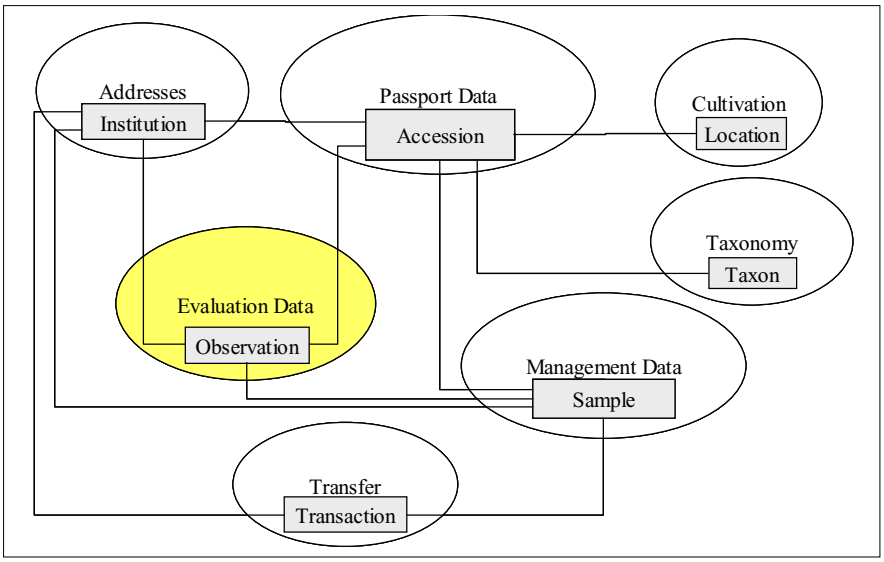


Abbildung 1: **Module in GBIS**

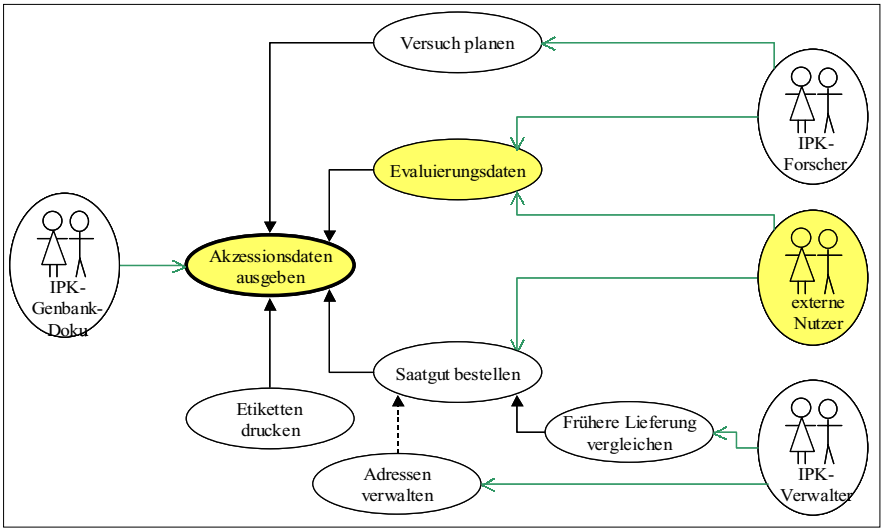


Abbildung 2: **Anwendungsfall „Akzessionsdaten ausgeben“**

Autor: Jörn VORWALD, IPK Gatersleben, Corrensstraße 3, D-06466 GATERSLEBEN

2 Datenmodell

Eines der Ziele von GBIS ist es, der Aufgabe der Informationsbereitstellung in der Hinsicht nachzukommen, dass externe Nutzer des Systems (außerhalb der Genbank des IPK, z. B. andere IPK-Arbeitsgruppen, Züchter und Landwirte) alle zu einer Genbankakzession verfügbaren Daten anfordern können. Zu diesem Zweck werden verschiedene Anwendungsfälle (use cases, vgl. BOOCH, 1991, RUMBAUGH et al., 1991) modelliert, von denen einer beispielhaft dargestellt wird (*Abbildung 2*).

Die dargestellten Akteure („IPK-Genbankdokumentation“, „IPK-Forscher“, „externe Nutzer“, „IPK-Verwalter“) haben Zugriff auf unterschiedliche Teile des zentralen Anwendungsfalls „Akzessionsdaten ausgeben“. Der Akteur „IPK-Genbankdokumentation“ muss als Administrator fungieren und auf alle untergeordneten Anwendungsfälle Zugriff haben.

Der „externe Nutzer“ soll Saatgut bestellen und Evaluierungsdaten sehen aber keine Adressen verwalten können, was neben den Administratoren den „IPK-Verwaltern“ vorbehalten bleibt.

Alle Anwendungsfälle, die bisher auszumachen sind, wurden in einem Lastenheft niedergelegt (EPORTAS, 2002), welches laufend fortgeschrieben wird.

Die Ausführungen des Lastenheftes sind in einem Pflichtenheft zu ergänzen und v. a. zu vertiefen (BALZERT, 2000), um einerseits in einem mehrköpfigen Entwicklerteam eine Verteilung der Aufgaben zu ermöglichen und andererseits die Grundlage für eine Dokumentation des Gesamtsystems zu liefern.

Daneben erfolgt eine Modellierung auf Datenebene mit Hilfe von Entity-Relationship-Diagrammen (CHEN, 1976). Der Ansatz des ‚single obeservation concept’ (VAN HINTUM & HAZEKAMP, 1992, MCLAREN et al., 2001) wird auf das Modell der Evaluierungsdaten im GBIS bezogen dargestellt (*Abbildung 3*). Dieses Konzept fokussiert auf das Zusammentreffen von experimentbezogenen Metadaten (‚experiment data’; Kontext des Versuchs, Versuchsansteller, Ort des Versuchs usw.), einer angewandten Methode (‚method’) und einem Merkmal (‚descriptor’) in einer Beobachtung (‚observation’). Diese Beobachtung umfasst einen merkmalsbezogenen Datensatz (‚observation data’; den eigentlich beobachteten Wert) mit dessen Metadaten (‚data value’; Wertebereich).

Auf diese Weise sind sehr flexibel die verschiedensten Beobachtungen mit ihren Merkmalen, Wertebereichen usw. in einer einheitlichen Struktur innerhalb der Datenbank abbildbar. Der Nachteil des single observation concept besteht in dem vergleichsweise hohen Aufwand der wissenschaftlichen Administration. Es sollen nur endlich viele Merkmal-Begriffe verwaltet werden, so dass Synonyme separat verwaltet müssen. Auch die Wertebereiche müssen innerhalb der Datenbank einheitlich sein, so dass eine Abbildung des von einem Versuchsansteller verwendeten Wertebereichs auf einen anderen bezogen auf das gleiche Merkmal möglich sein muss.

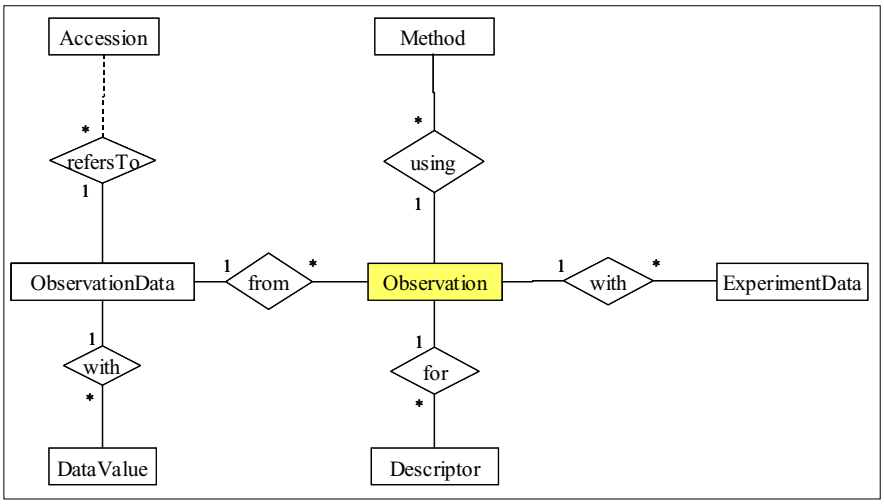


Abbildung 3: **Ausschnitt des ER-Modells von GBIS zur Abbildung des ‚single observation concept’**

3 Zugang zu Evaluierungsdaten

3.1 Gegenwärtiger Zugang

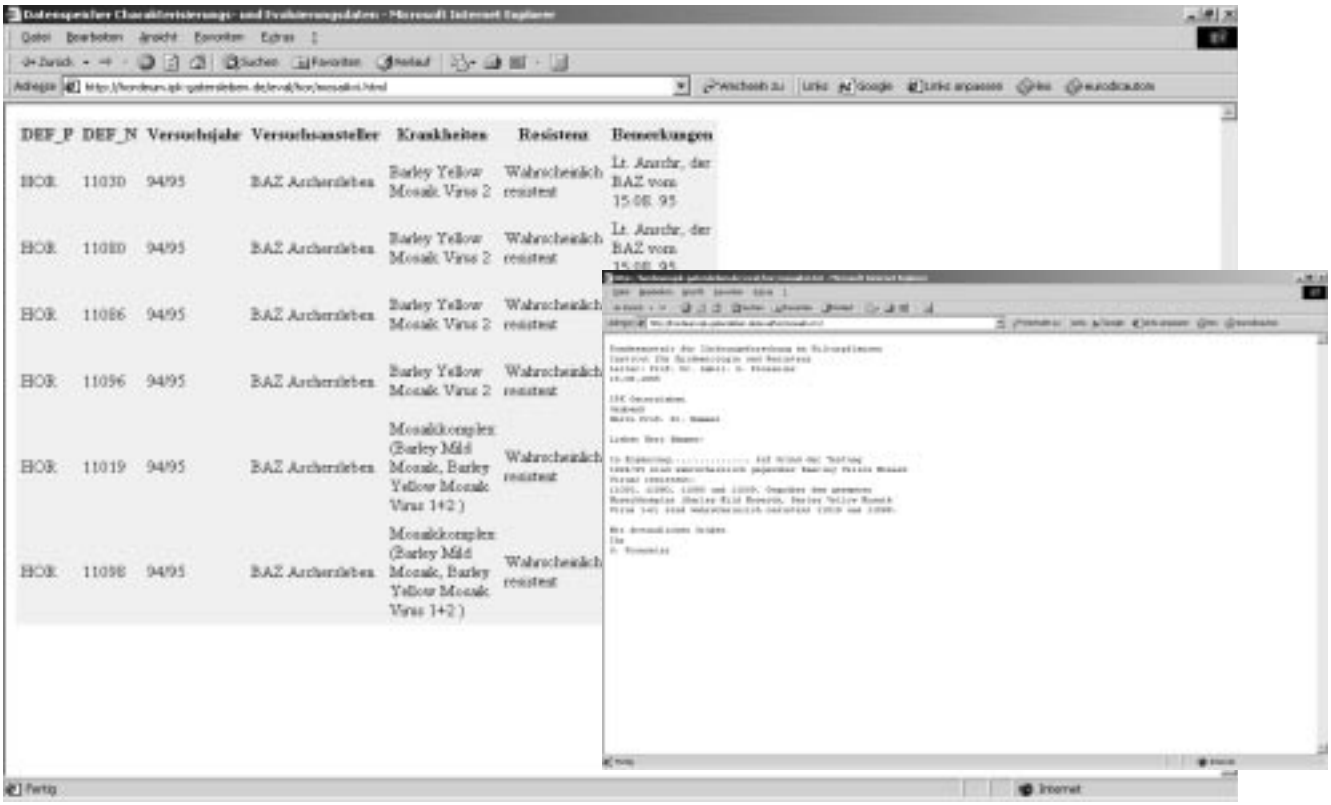
Zur Zeit stehen Evaluierungsdaten nur eingeschränkt zur Verfügung. Die größte Einschränkung stellt dabei eine fehlende Suchmöglichkeit dar. Nach dem Einstieg über das Portal http://hordeum.ipk-gatersleben.de/eval/eval.html (28.11.2002) ergibt sich für den Nutzer die Auswahlmöglichkeit des Sortiments (auf Gattungs-Ebene). Er erhält eine Übersicht über die verfügbaren Evaluierungsdaten, weiterhin das Angebot zu Ansicht und download.

In den meisten Fällen sind Metadaten zu den Versuchen verfügbar. Die Qualität dieser Metadaten differiert jedoch: Ausführliche Versuchsbeschreibungen (z. B. KRÜGER & HAMMER, 1995), Korrespondenzen über Versuchsergebnisse (z. B. PROESELER, 1995) oder lediglich ein Literaturzitat zu den Ergebnissen (z. B. NOVER & LEHMANN, 1973) bilden die Spannweite.

In der Ansicht werden html-basierte oder Excel-Tabellen geboten (*Abbildung 4*). Diese Tabellen offenbaren die heterogene Struktur, in der die Daten vorliegen: Meist sind die Ergebnisse nach Gaterslebener Akzessionsnummern zeilenweise geordnet. Die Spalten geben Auskunft über die Deskriptoren. Ein Abgleich der Terminologie fehlt, d. h. die in Abschnitt 2 beschriebenen Synonyme werden nicht verwaltet. Metadaten über die Tabellenstruktur sind in einer gesonderten Tabelle an die Übersicht über die zur Verfügung stehenden Untersuchungsergebnisse aufgeführt (*Abbildung 4*). Das ist notwendig, da für die Deskriptoren und andere Tabellenspalten häufig Abkürzungen verwendet wurden, deren Sinn sich dem Nutzer nicht unmittelbar erschließt.

3.2 Künftiger Zugang

Die Unzulänglichkeiten des gegenwärtigen Systems sollen künftig beseitigt werden. Insbesondere soll eine komfortable Suchfunktion dem Nutzer einen gezielten Zugriff auf die vorhandene Daten ermöglichen. Darüber hinaus soll das Informationsangebot dahingehend erweitert werden, dass alle zu den Evaluierungen verfügbaren Metadaten (v. a. auch Bilddaten) angeboten werden können.



DEF_P	DEF_N	Versuchsjahr	Versuchsansteller	Krankheiten	Resistenz	Bemerkungen
HOR	11030	94/95	BAZ Aschersleben	Barley Yellow Mosaic Virus 2	Wahrscheinlich resistent	Lt. Anschr. der BAZ vom 15.08.95
HOR	11080	94/95	BAZ Aschersleben	Barley Yellow Mosaic Virus 2	Wahrscheinlich resistent	Lt. Anschr. der BAZ vom 15.08.95
HOR	11086	94/95	BAZ Aschersleben	Barley Yellow Mosaic Virus 2	Wahrscheinlich resistent	
HOR	11096	94/95	BAZ Aschersleben	Barley Yellow Mosaic Virus 2	Wahrscheinlich resistent	
HOR	11019	94/95	BAZ Aschersleben	Mosaikkomplex (Barley Mild Mosaic, Barley Yellow Mosaic Virus 1+2)	Wahrscheinlich resistent	
HOR	11098	94/95	BAZ Aschersleben	Mosaikkomplex (Barley Mild Mosaic, Barley Yellow Mosaic Virus 1+2)	Wahrscheinlich resistent	

Abbildung 4: **Gegenwärtiger Zugang zu Evaluierungsdaten des IPK**

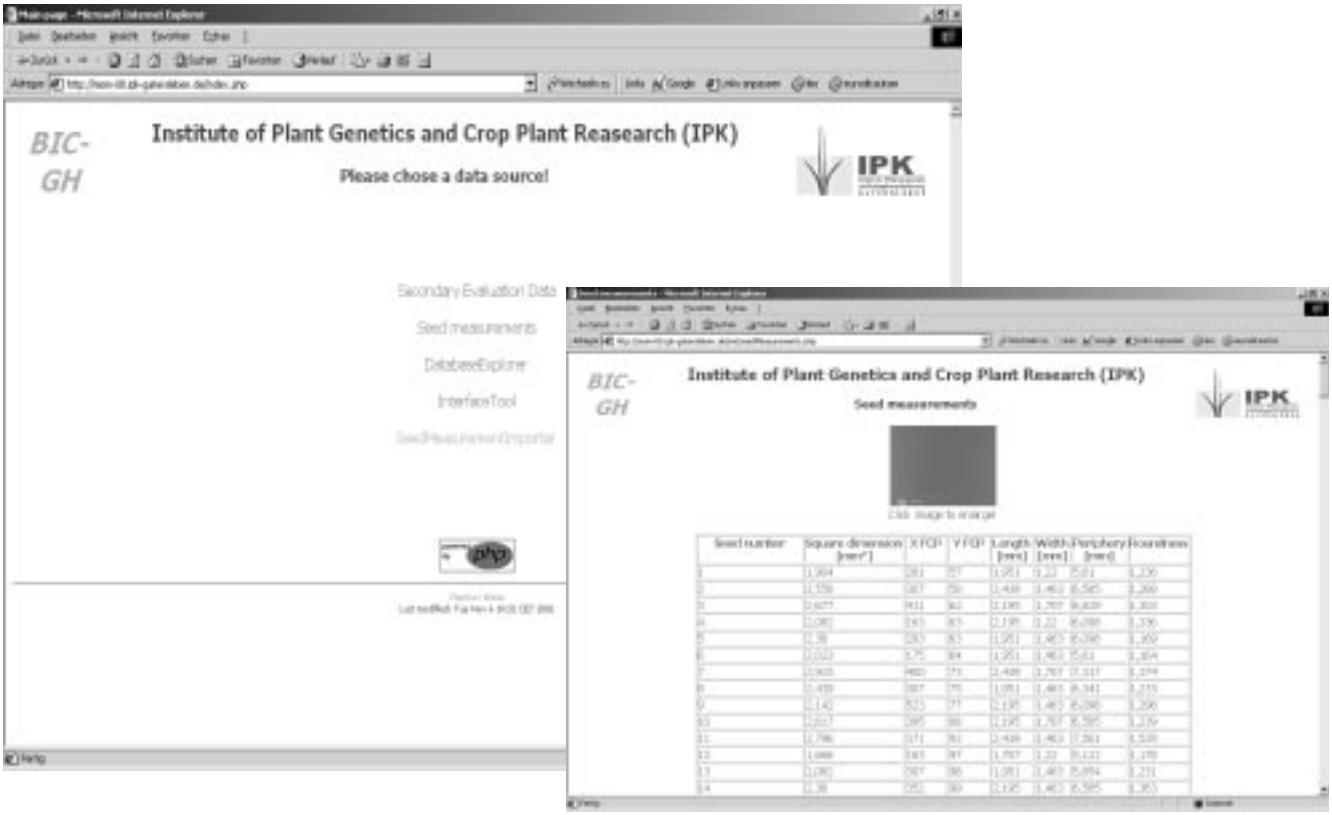


Abbildung 5: **Künftiger Zugang zu Evaluierungsdaten des IPK**

Im Projekt PDW (Bioinformatikzentrum Gatersleben-Halle (BIC-GH): Entwicklung eines Plant Data Warehouse für genotypische, phänotypische, taxonomische und Expressionsdaten von Kulturpflanzen) wurde das im Abschnitt 2 erwähnte Datenmodell federführend entwickelt und implementiert. Es existiert ein Prototyp zur Abfrage der vorhandenen und zur Eingabe neuer Evaluierungsdaten (*Abbildung 5*; Portal: http://mom-08.ipk-gatersleben.de/; 28.11.2002). Die-

ser Prototyp ist prinzipiell nicht ständig online verfügbar, da noch Entscheidungen zur wissenschaftlichen Administration ausstehen.

4 Probleme der Datenbereitstellung

4.1 Wissenschaftliche Administration

Neben der Datenbankadministration im engeren Sinne, also dem Verfügbarmachen der Daten auf der technischen Ebene, kann die wissenschaftliche, d. h. inhaltliche Administration im jetzigen Zustand als Kernproblem ausgemacht werden. Die von den Nutzern der genetischen Ressourcen (Genbankakzessionen) übermittelten Evaluierungsdaten sind bisher nahezu ohne Überarbeitung in ihrem Urzustand online verfügbar gemacht worden. Aus einer solchen Arbeitsweise resultierte ein Problem verteilter Daten, indem zwar innerhalb eines Evaluierungsvorgangs eine Struktur verwendet wurde, diese aber mit der Struktur eines anderen Evaluierungsvorgangs fast nie vergleichbar ist.

Häufig tritt das Problem auf, verschieden skalierter Wertebereiche vergleichen zu müssen. Neben metrisch skalierten Ergebnissen (Längenangaben, Datumsangaben usw.) sind ordinal skalierte Ergebnisse (v. a. Boniturnoten) die Regel. Nicht ohne Bedeutung ist dabei, dass gelegentlich unstandardisierte Bonitурskalen verwendet wurden (z. B. Noten 1 - 2 - 3 - 4 - 5), deren Abbildung auf eine standardisierte Bonitурskala (Ganzzahlen im Intervall [0,9]) von Merkmal zu Merkmal unterschiedlich ausfallen kann.

Da die Genbank in Gatersleben seit 1945 besteht, ist im Lauf der Zeit ggf. mehrfach ein Wechsel der Skalen zu beobachten. Anfang der 1990er Jahre gab es mindestens im Getreidesortiment einen Wechsel der Wertebereiche und Skalen hin zu deutschlandweit üblichen Standards.

Schließlich stellt die Heterogenität der zugelieferten Daten ein nicht unerhebliches Problem dar. Da ggf. jeder Versuchsansteller eine eigene Nomenklatur für die Deskriptoren benutzt, ist wenigstens eine „Inflation der Synonyme“ zu beobachten. Darüber hinaus kann auch von einer „Inflation der Wertebereiche“ gesprochen werden, was unmittelbar auch mit dem oben beschriebenen IPK-internen Problem des Wechsels der Skalen im Laufe der Zeit zusammenhängt.

4.2 „Datenformat“ der Zulieferer

Ein zweites wichtiges Problem im Zusammenhang mit der Bereitstellung von Evaluierungs- (und anderen) Daten stellt das Format dieser Daten dar. Besonders historische Daten liegen auf Papierlisten vor. Im erwähnten Zeitraum ab 1945 sind notgedrungen die Anfänge der Dokumentation auf dem Informationsträger Papier zu suchen. Papier macht nach wie vor einen nicht unbedeutenden Anteil an den verwendeten Informationsträgern aus (vgl. MEYER, 2002, in diesem Band). Dieses Papier unterliegt einem Alterungsprozess; seine Grundfarbe kann variieren; die Daten sind mit verschiedenen Schreibgeräten festgehalten worden, deren „outputs“ gleichfalls einem Alterungsprozess ausgesetzt sind. Darüber hinaus sind die Informationen handschriftlich festgehalten. Aus diesen und weiteren Gründen verbietet sich die automatische Digitalisierung. Selbst das Festhalten der Listenansichten in Scans erbringt nur bedingt brauchbare Ergebnisse. Eine Suchfunktion ist in solcherart Dokumenten nicht implementierbar. In diesen Fällen führt an der manuellen digitalen Erfassung kein Weg vorbei. Solcherart wurden beispielhaft von 1996 bis 1999 im Rahmen des EVA-Projektes am IPK Evaluierungsdaten von ca. 11.500 Gersten-Akzessionen erfasst (http://www.genres.de/eva/; 28.11.2002).

Selbst das Vorliegen von Information auf historischen EDV-Datenträgern ist heutzutage schon zum Problem geworden, da die Lesegeräte nur noch eingeschränkt verfügbar sind. Acht-Zoll- und selbst 5,25-Zoll-Disketten sind längst nicht mehr Standard, Magnetbänder wurden mehrfach im Laufe der Zeit hinsichtlich der Speicherungs-Modi verändert, so dass selbst prinzipiell elektronisch verfügbare Information real nicht mehr oder nur noch eingeschränkt zur Verfügung steht (vgl. HENDRIKS, 1997).

Ist das Problem des Informationsträgers lösbar, stellen die eigentlichen Dateiformate oder auch ihre Verwendung eine weitere Hürde dar. Als Beispiel sei darauf hingewiesen, dass weniger geübte Computeranwender gelegentlich dazu neigen, in Excel-Arbeitsblättern zusammengehörige Informationen (z. B. Spaltenüberschriften) auf mehrere Zellen zu verteilen, was ein automatisches Einlesen in eine andere Struktur erschwert oder unmöglich macht.

4.3 Unvollständige Information

Um die Evaluierungsdaten angemessen beurteilen zu können, sind alle Nutzer auf das Vorliegen von Metainformationen angewiesen. Eine verzögerte Entwicklung von Pflanzen ist ggf. im Kontext der zur Vegetationsperiode herrschenden Witterung erklärbar. Fehlen die Daten, ist eine Fehlinterpretation nicht ausgeschlossen.

Bodenfaktoren haben erheblichen Anteil am physiologischen Status der Pflanzen, so dass das Fehlen von Informationen darüber gleichfalls Fehlschlüsse über die Evaluierungsergebnisse zulässt. Diese Reihe ließe sich nahezu beliebig fortsetzen. Entscheidend ist lediglich, dass Versuchsergebnisse um so besser validierbar oder falsifizierbar sind, je genauere Informationen über die Versuchsbedingungen und -zusammenhänge vorliegen.

5 Zusammenfassung

Im Zuge der Zusammenführung der deutschen Genbanken (Genbank der Bundesanstalt für Züchtungsforschung - BAZ - in Braunschweig; Genbank des Instituts für Pflanzengenetik und Kulturpflanzenforschung - IPK - in Gatersleben) ist der Neuaufbau eines Genbank-Informationssystems erforderlich. Dieses GBIS hat die Bereitstellung von Information über die Genbanksortimente als wichtige externe Aufgabe. Damit verbunden ist die Erweiterung des Informationsangebotes für alle Nutzer, indem u. a. die Ergebnisse von Evaluierungen einzelner Genbankakzessionen verfügbar gemacht und mit einer komfortablen Suchfunktion komplettiert werden sollen. Zur Beherrschung der Komplexität wird bei der Realisierung auf das „single observation concept“ (vgl. van HINTUM & HAZEKAMP, 1992, MCLAREN et al., 2001) abgestellt.

6 Danksagung

An dieser Stelle möchte ich herzlich meinen Kollegen am IPK, Herrn Dipl.-In-

form. Stephan WEISE (Entwicklung von Datenmodell und Prototyp), Herrn Dr. Helmut KNÜPFFER (Diskussionen, Bereitstellung von Hintergrundinformationen) sowie Herrn Dipl.-Inform. (FH) Steffen FLEMMING (Diskussionen zum Datenmodell) danken, ohne die dieser Beitrag nicht hätte realisiert werden können. Das Projekt „Aufbau einer bundeszentralen ex situ Genbank für landwirtschaftliche und gartenbauliche Kulturpflanzen: Zusammenführung der Genbanken des IPK und der BAZ Braunschweig, Arbeitspaket 1: Fusion der Genbank-Informationssysteme (GBIS)“ (http://www.verteiltes-projektmanagement.de/gbis; 28.11.2002) wird vom Bundesministerium für Bildung und Forschung - BMBF - unter dem Förderkennzeichen 0312830 gefördert. Das „Verbundprojekt Bioinformatikzentrum Gatersleben-Halle (BIC-GH): Entwicklung eines Plant Data Warehouse für genotypische, phänotypische, taxonomische und Expressionsdaten von Kulturpflanzen“ (PDW) (http://bic-gh.ipk-gatersleben.de/index.php; 28.11.2002) wird unter dem Förderkennzeichen 0312706A gleichfalls vom BMBF gefördert.

7 Literatur

BALZERT, H., 2000: Lehrbuch der Software-Technik. Bd. I Software-Entwicklung. Spektrum Akademischer Verlag, Heidelberg/Berlin, 2. Aufl. 1136 S.

BOOCH, G., 1991: Object-oriented design with applications, Redwood City: The Benjamin/ Cummings Publishing Company.

CHEN, P., 1976: The Entity-Relationship model - towards a unified view of data, in: ACM transactions on database systems, Vol. 1, No. 1, March 1976: 9-36.

EPORTAS, 2002: Studie Genbankdokumentation im Hinblick auf die Entwicklung von GBIS. Dokumente I - IV. Düsseldorf, Polykopie. 15 + 68 + 11 + 26 S.

FREYTAG, U. & H. KNÜPFFER, 1994: Aspekte der Datenverarbeitung für das interne Genbankmanagement in Gatersleben. Bericht über die Arbeitstagung 1994 der Arbeitsgemeinschaft der Saatzuchtleiter im Rahmen der Vereinigung österreichischer Pflanzenzüchter, BAL Gumpenstein, 22.-24.11.1994: 195 - 202.

HENDRIKS, K.B., 1997: Der Endogene Zerfall von Archivgut - ein zwangsläufiges Phänomen? In: Weber, H. (Hrsg.) Bestandserhaltung. Herausforderung und Chancen, Veröff. der Staatlichen Archivverwaltung Baden-Württemberg, Bd. 47, Stuttgart: 21-44.

KNÜPFFER, H., 2001: Handling of characterization and evaluation data in crop databases. In: MAGGIONI, L. and O. SPELLMAN, (comp.) Report of a Network Coordinating Group on Cereals. Ad hoc meeting, 7-8 July 2000, Radzików, Poland. International Plant Genetic Resources Institute, Rome, Italy: 58-65.

KRÜGER, H. und K. HAMMER, 1995: Evaluierung der (Anethum graveolens)-Kollektion der Genbank Gatersleben (chemische Zusammensetzung der Fruchtöle). http://hordeum.ipk-gaters-leben.de/eval/anet/anethmkr.html (28.11.2002)

MCLAREN, C.G., A. PORTUGAL and J.G.F. LIESHOUT, 2001: Design of the data management system (DMS). International Crop Information System. Technical Development Manual. http://www.icis.cgiar.org/ICIS07k.pdf (28.11.2002)

MENTING, F. B. J. and T. J. L. VAN HINTUM, 2001: Genis Data Dictionary. Draft July 2001. Wageningen, 31 S.

NOVER, I.O. und CH. LEHMANN, 1973: Resistenzeigenschaften im Gersten- und Weizensortiment Gatersleben, 17. Prüfung von Sommergersten auf ihr Verhalten gegen Mehltau (Erysiphe graminis DC. f. sp. hordei Marchal). Kulturpflanze XXI: 275-294. (http://hordeum.ipk-gatersleben.de/eval/hor/hor_eg.txt) (28.11.2002)

PROESELER, G., 1995: http://hordeum.ipk-gatersleben.de/eval/hor/mosaikvi.txt (28.11.2002)

RUMBAUGH, J., M. BLAHA, W. PREMERLANI, F. EDDY, F. and W. LORENSEN, 1991: Object-oriented modelling and design. Englewood Cliffs: Prentice Hall.

VAN HINTUM, Th.J.L. and T. HAZEKAMP, 1992: Genis Data Dictionary, July 1992. Centre for Plant Breeding and Reproduction Research (CPRO-DLO), Centre for Genetic Resources, The Netherlands (CGN), Wageningen, The Netherlands.