

Eckhard K. Groll

Probleme und Erfahrungen beim Aufbau nachhaltig konsistenter Datenbanken

Vorliegende Arbeit soll einige Probleme des Aufbaues nachhaltig konsistenter Datenbanken aufzeigen und Erfahrungen bei der Anwendung entsprechend implementierter Programme zur Erfassung einer Bibliographie alter, entomologisch relevanter Literatur bis einschließlich 1863 am Deutschen Entomologischen Institut im Zentrum für Agrarlandschafts- und Landnutzungsforschung (ZALF) e.V. vermitteln.

1 Einführung

Seit 1886 dient das Deutsche Entomologische Institut (DEI) als Informationszentrum für die Entomologen der Welt. Traditionell wurden und werden im DEI bzw. in Kooperationsprojekten unter Mitwirkung der Wissenschaftler des DEI zahlreiche Informationen mit entomologischem Bezug gesammelt und herausgegeben, z. B. Bibliographien der entomologischen Weltliteratur (HORN und SCHENKLING, 1928-1929; DERKSEN und SCHEIDING, 1963 - 1975; GAEDIKE und SMETANA, 1978, 1984), Verbleib entomologischer Sammlungen (HORN, KAHLE et al., 1935, 1990), Abkürzungsverzeichnis taxonomischer

Autoren (SCHMITT, HÜBNER und GAEDIKE, 1998) und zahlreiche Typenkataloge.

Weiterhin verfügt das DEI über eine bedeutende Insekten-sammlung und eine der größten entomologischen Spezialbibliotheken. Der Bestand unserer Bibliothek beläuft sich auf ca. 25.000 Bände Monographien und ca. 46.000 Bände Zeitschriften von 1.300 Zeitschriftentiteln. Die Monographien umfassen Ausgaben seit dem 16. Jahrhundert. Alle genannten Ressourcen werden nach und nach in relationalen Datenbanken mittels dem DBMS PARADOX (Datenbankmanagementsystem) gespeichert und für in- und externe Nutzung verfügbar gemacht. Ziel ist die Publikation solcher Datenbanken im Internet als sich ständig vervollkommnende und langfristig konsistente Referenzen für die systematische und

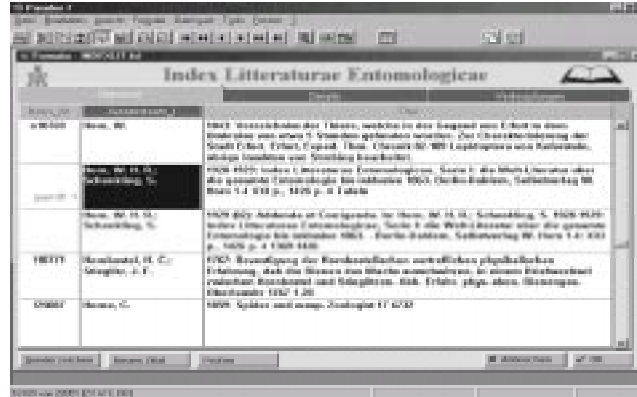


Abb. 1: Eingabeseite für Index Litteraturae Entomologicae

angewandte Entomologie. Das aktuellste Projekt ist gegenwärtig die verbesserte Neuauflage des Index Litteraturae Entomologicae Serie I.

2 Konzeption

Beim Aufbau nachhaltig konsistenter Datenbanken wurden einige Probleme herausgearbeitet und als bei der Konzeption zu berücksichtigende Forderungen formuliert. Im Gegensatz zu den sogenannten ACID-Forderungen (atomicity, consistency, isolation, durability) die ein relationales DBMS erfüllen muß, geht es dabei jedoch um sach- und nutzerbezogene Probleme.

2.1 Vertrauen in die Daten

Anwender einer gemeinsam genutzten Datenbank müssen sich einerseits auf die Daten der Mitbenutzer verlassen können und andererseits Daten verlässlich beisteuern. Zu Beginn des Projektes "Neuauflage des Index Litteraturae Entomologicae" bezogen wir alle Mitarbeiter des DEI in die

Datenerfassung ein. Durch die Schwierigkeiten der Thematik wurde die Zahl der Mitarbeiter später auf wenige Experten reduziert. Die frühe Phase führte jedoch zu wertvollen Erkenntnissen. Z. B. steigt die Qualität der Daten, wenn der Name des Eingebenden eines Datensatzes bekannt ist. Fragen, die ein Mitarbeiter äußert, existieren meist auch bei anderen. Die Antworten in Form von Konventionen, Rezepten und Beispielen wurden deshalb für alle verfügbar, schriftlich fixiert. Aus diesen und weiteren Erfahrungen leiten sich allgemeingültige vertrauensbildende Forderungen ab:

- Der Verantwortliche (Besitzer, Eingebener) für jeden Datensatz muss erkennbar sein;
- Daten müssen vor unbefugter Modifikation geschützt sein;
- Andererseits ist die Zusammenarbeit ausdrücklich gewollt, d. h. Daten dürfen von befugten Mitarbeitern modifiziert werden und alle anderen Datenbankbenutzer können Modifikationen veranlassen;
- Die Geschichte jedes Datensatzes von der Anlage bis zum aktuellen Zeitpunkt muss transparent sein;
- Das Schlüsselssystem (die Verbindung) der Teildatenbanken muss während der gesamten Lebenszeit der Daten stabil sein;

- Eine Kommunikation über die der Zusammenarbeit zugrundeliegenden Regeln muss möglich sein.

2.2 Genaue Widerspiegelung des Sachverhaltes

Jede Implementierung ist ein Kompromiss zwischen Wirklichkeit und Abbild. Als Beispiel soll die Datenbank "Orthopteran Species File" der Autoren D. OTTE und P. NASKRECKI (1997) dienen. Sie präsentiert nahezu alle Typen der Heuschrecken der Welt mit Systematik, Synonymen, Bild und Ton. Die Daten dafür stammen u.a. aus gedruckten Typenkatalogen von Museen. Im alphanummerischen Feld DEPOSITORY wird der Aufbewahrungsort jedes Typus gespeichert. Diese Konstruktion kann jedoch Namensänderungen eines Museums nicht korrekt abbilden. So sind die Namen der Typen des DEI unter mindestens 6 synonymen Orten abgespeichert (Deutsches Ent. Inst., Eberswalde / Deutsches Ent. Inst., Berlin-Dahlem / Deutsches Ent. Mus., Dahlem-Berlin / Berlin-Dahlem / Deutsches Ent. Inst. Berlin-Dahlem / Deutsches Entomologisches Institut, Eberswalde). Bei einer Abfrage, welche Typen im DEI verfügbar sind, enthält die Antwort nicht einmal die Hälfte der Namen des tatsächlich vorhandenen Materials. Die Autoren sind sich dieses Mangels bewusst und arbeiten an einer Version, die solche Synonyme widerspiegeln kann.

In Anlehnung an Datenmodelle, wie ASC (1993), BIOTA (1997) u.a. liegen unseren Datenbanken Überlegungen zugrunde, wie man stabile, konsistente Daten erreichen kann. Sie werden im Abschnitt Datenmodelle erläutert.

2.3 Änderungen und Wissensfortschritt

Können mehrere Mitarbeiter gleichzeitig auf eine Datenbank schreibend zugreifen, gewährleisten geeignete DBMS die Konsistenz der Daten durch Sperren. Lehrbücher benutzen gern dieses Beispiel: Verkäufer B kann erst dann eine bestimmte Anzahl Hämmer aus dem Lager abrufen (Datensatz ist gesperrt), nachdem Verkäufer A seinen Abruf von Hämmern beendet (den Datensatz freigegeben) hat, mit anderen Worten, wenn Verkäufer B den neuen Wissensstand erkennen kann. Was aber, wenn die Artikelbezeichnung "Hammer" in "Schlosserhammer" und "Putzhammer" geändert werden muss? Ohne Information über diese Modifikation finden die Verkäufer den gewünschten Artikel nicht wieder.

Dieser banale Fall ist im in Rede stehenden Projekt jedoch die Regel. Beispielsweise hat die Zeitschrift Preußische Provinzial-Blätter im Zeitraum bis 1863 mehrere Titeländerungen erfahren (Vaterländisches Archiv für Wissenschaft, Kunst, Industrie und Agrikultur, oder Preuss. Provinzial-Blätter / Archiv für vaterländische Interessen oder Preußische Provinzial-Blätter. Neue Folge / Neue Preußische Provinzial-Blätter). Durch zahlreiche Abkürzungen oder fehlerhafte Sekundärquellen wird die Verwirrung noch potenziert. Naturgemäß werden solche Probleme erst nach und nach sichtbar. Hat aber ein Mitarbeiter die korrekten Namen recherchiert, so sollen sie auch sofort für die anderen verfügbar sein. Deshalb wurden die Datenbanken so aufgebaut, dass sich das Wissen der Mitarbeiter anhäufen und untereinander mitteilen kann.

- Regel 1: Datensätze dürfen nicht verschwinden;
- Regel 2: Daten dürfen nur ausnahmsweise, z. B. bei Eingabefehlern korrigiert (syntaktisch geändert) werden;
- Regel 3: Änderungen, die auf neuen Erkenntnissen beruhen, sind keine Korrekturen sondern führen zu Ableitungen;
- Regel 3a: Ändert sich die Syntax eines Datensatzes bei gleicher Semantik, wird ein neuer Datensatz angelegt und der alte dem neuen als Synonym nachgeordnet;
- Regel 3b: Ändert sich die Semantik eines Datensatzes, werden neue Datensätze angelegt und verweisen auf ihren Ursprungsdatensatz.

2.4 Wiederfinden von Daten

Im o. g. Beispiel müssten Datensätze für neue Werkzeuge zentral im Lager (wo die realen Objekte aufbewahrt sind) angelegt werden. Eine wissenschaftliche Datenbank ist jedoch gleichzeitig Lager und Abbild, d. h. jeder schreibberechtigte Mitarbeiter darf neue Datensätze anlegen, aber erst, wenn er sicher sein kann, dass entsprechende Daten fehlen. Besondere Aufmerksamkeit muß deshalb dem Wiederfinden von Daten gewidmet werden (GROLL, E. K.; TAEGER, A., 1997).

- Daten in Fremdsprachen (Einschränkung: auf der Basis lateinischer Buchstaben) müssen sprachgenau wiedergegeben werden und sich ohne genaue Kenntnis der Fremdsprache wiederfinden lassen;
- unterschiedliche Transkriptionen von Namen müssen berücksichtigt werden;
- reguläre, schnelle Suchstrategien müssen durch alternative ergänzt werden können (Einbeziehung weiterer Daten, Darstellung im Kontext); eine Kommunikation zwischen den Nutzern muss möglich sein (Erklärungen, Sammlung von Beispielen und Ausnahmen, Nachrichten an Verantwortliche).

3 Datenmodelle

Ziel unserer Datenmodelle war die schnelle Umsetzung in eine arbeitende Datenbank auf der Basis von PARADOX (ANONYM 1994a, b und c). Deshalb wurden einerseits zahlreiche Erkenntnisse von ASC (1993) und BIOTA (1997) übernommen, andererseits praktische Erfahrungen eingebracht. Am Beispiel der Darstellungen für die Neufassung des Index Litteraturae Entomologicae Serie I soll dies nachfolgend erläutert werden.

3.1 Datenmodell für Publikationen

Die Datenbank speichert Autor(en), Jahr, Titel und bibliographische Angaben einer Publikation. Die Datensätze dienen zahlreichen andern Teildatenbanken, wie Nomenklatur, Systematik und Bionomie von Insekten, Bibliographien und Biographien als Datenquelle. Dabei stehen die Erfordernisse der wissenschaftlichen Zitierung gegenüber einer umfassenden bibliographischen Notation im Vordergrund. In der Abbildung 2 sind die Attribute von Publikationen und die Beziehungen zu Autoren, Quellen usw. dargestellt. Die Ellipsen in hervorgehobener Darstellung mit Namen im Singular repräsentieren ein Objekt (Entität; im Beispiel eine Publikation), das meist als Datensatz implementiert ist. Ellipsen in regulärer Darstellung mit Namen im Plural sind Mengen von Objekten (im Beispiel die Autorenteam, Quellen, Nutzer usw.), die als Tabellen von Datensätzen implementiert sind. Hinter den Rechtecken verbergen sich die Eigenschaften (Attribute) der Objekte, die später meist als Felder der Datensätze erscheinen. So sind Publikationen durch Autoren geschrieben, in einem bestimmten Jahr unter einem Titel erschienen usw. Die Linien zeigen entweder die Zugehörigkeit der Attribute oder die Verbindung zu anderen Objektmengen an. Man kann die Zugehörigkeit von beiden Seiten aus lesen, z.B. "Der Titel einer Publikation ist ..." bzw. "Die Publikation hat den Titel ...". Bei den Verbindungen zu anderen Objektmengen spielen hier die 1:1- (3, kein oder genau ein Objekt ist verbunden) und die 1:n-Verbindung (4, kein oder beliebig viele Objekte sind verbunden) eine Rolle. So ist eine Publikation von genau einem Autorenteam geschrieben worden. Für unbekannte Autoren steht das Team "Anonym". Demgegenüber kann ein Autorenteam (Abb. 3) aus beliebig vielen Autoren bestehen.

Schließlich sind einige Attribute mit "P" oder "F" gekennzeichnet. "P" steht für Primärschlüssel, eine eindeutige Nummer für jedes Objekt einer Objektmenge. "F" ist ebenfalls ein Primärschlüssel, jedoch eines Objektes aus einer anderen Objektmenge. Ist er Attribut eines Objektes, so wird

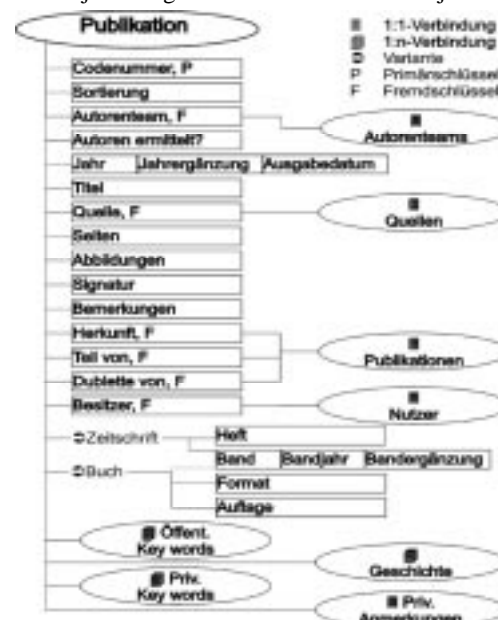


Abb. 2: Datenmodell für Publikationen

der Fremdschlüssel genannt. Beispielsweise wird das Autorenteam einer Publikation nicht direkt im Datensatz gespeichert, sondern nur die Schlüsselnummer eines Autorenteams, die mittels Nachschlagen aus der Referenztafel der Autorenteams übernommen werden muss.

Erläuterungen zu einigen Attributen:

- AUTOREN ERMITTELT? wird benutzt, wenn die Autoren nicht in der Originalpublikation genannt, sondern recherchiert wurden. Bei der Ausgabe kann ein ermitteltes Team in eckigen Klammern ([]) eingeschlossen dargestellt werden. In der Sortierung unterscheidet es sich jedoch nicht vom regulären Team.
- JAHR, JAHRERGÄNZUNG und AUSGABEDATUM tragen der großen Bedeutung des exakten Ausgabedatums einer Publikation, z.B. einer Originalbeschreibung, aufgrund der Prioritätsregel in der Nomenklatur Rechnung. JAHR repräsentiert das Erscheinungsjahr oder ein Intervall zweier Jahre (z.B. 1855-1857) und sorgt für die chronologisch korrekte Sortierung. JAHRERGÄNZUNG gestattet Zusätze zum Erscheinungsjahr (z.B. "vor 1855"). AUSGABEDATUM schließlich ermöglicht die genaue Speicherung von Tag, Monat und Jahr.
- TITEL: Das stark vereinfachte Objekt PUBLIKATION hat nur drei Varianten, den Zeitschriftenartikel, die Monographie (vereinfacht Buch) und den Artikel innerhalb einer Monographie. TITEL repräsentiert dementsprechend den Titel des Artikels, den Buchtitel bzw. die Kapitelüberschrift. Für diese Varianten (Ü) stehen weiterhin die Attribute HEFT, BAND und ergänzende Bandangaben sowie FORMAT und AUFLAGE zur Verfügung.
- HERKUNFT wird benutzt, wenn gezeigt werden soll, dass die Angaben zur Publikation aus dem Literaturverzeichnis einer andern Publikation stammen. Es enthält in diesem Fall die CODENUMMER dieser Publikation. Dieses rekursive Verfahren wird von uns vielfach angewendet und ergänzt das relationale Modell durch die Möglichkeit, hierarchische Strukturen abzubilden.
- DOUBLETTE VON: Durch den Import von Daten, Programmfehler beim Sortieren und Suchen und durch Unachtsamkeit der Mitarbeiter werden gelegentlich Dubletten angelegt. Wegen der Regel 1 dürfen sie nicht gelöscht werden, können jedoch über dieses Attribut mit einer als gültig definierten Publikation verbunden werden (Regel 3a). So entsteht keine Unterbrechung der Schlüsselverbindung zu anderen Datenbanken (z.B. Bibliographie), in denen diese Publikation möglicherweise verwendet wird. Mehr noch, bei einer Revision kann die Dublette durch die gültige Publikation automatisch ersetzt werden.
- TEIL VON: Ein Datensatz, der mehrere Objekte repräsentiert, wie z.B. "Ahrens, A. 1812-1816: Fauna Insectorum Europae. Halle, Kümmerl", kann in die beiden

Datensätze "Ahrens, A. 1812" und "Ahrens, A. 1816" aufgegliedert werden. Das Attribut TEIL VON verweist auf das ursprüngliche Objekt. Auch hier bleibt die Verbindung zwischen Datensätzen einer nachgeordneten und der ursprünglichen Datenbank erhalten. Bei einer Revision muss der Wissensfortschritt jedoch durch fallweise Entscheidung für jeden Teildatensatz eingearbeitet werden.

- BESITZER bezeichnet den Verantwortlichen des Datensatzes.
- ÖFFENTLICHE KEY WORDS gestatten die Zuordnung von allgemeingültigen Stichwörtern zur Publikation (s. Abb. 6).
- PRIVATE KEY WORDS: Wie oben, jedoch nutzerspezifisch.
- PRIVATE ANMERKUNGEN gestattet die Zuordnung von nutzerspezifischen Bemerkungen, Sonderdrucknummern usw. zur Publikation (s. Abb. 7) und ist, wie PRIVATE KEY WORDS, nur für den jeweils angemeldeten Nutzer sichtbar.
- GESCHICHTE dient der Dokumentation aller Änderungen am Objekt im Laufe der Zeit.

3.2 Datenmodell für Autoren und Autorenteams

In Abbildung 3 Mitte ist die Konstruktion für die Autoren dargestellt. Die Attribute von Autor bedeuten:

- AUTORENTEAM ist der Fremdschlüssel aus Autorenteams und hält die Verbindung zum Team.
- REIHENFOLGE legt die Reihenfolge der Autoren im Team fest.
- PERSON ist der Fremdschlüssel aus PERSONEN und hält die Verbindung zu einem Objekt im Namenverzeichnis.
- ZUSATZ charakterisiert die Funktion der Person im Autorenteam, z. B. Herausgeber. Die Zusätze sind in dem einfachen Objekt ZUSATZ (Abbildung 3 unten) organisiert.

Das Objekt Autorenteam (Abb. 3 oben) setzt sich aus einem bis beliebig vielen AUTOR-Objekten zusammen und hat folgende Eigenschaften:

- TEAM ist die ausgeschriebene und somit für den Nutzer lesbare Form der 1:n-Verbindung.
- SORTIERUNG entspricht TEAM, jedoch in einer für die korrekte Sortierung notwendigen Form (vgl. 4.5).
- SYNONYM VON zeigt auf ein anderes Autorenteam-Objekt, das als gültig für dieses Team festgelegt wurde. So können u.a. unterschiedlich transliterierte Autorennamen auf einen gültigen zurückgeführt werden.
- TEIL VON zeigt auf einen Vorgänger, von dem das Team abgeleitet wurde.
- BESITZER bezeichnete den Verantwortlichen des Datensatzes.

3.3 Datenmodell für Personen

Am Ende der Kette steht das Objekt PERSON (Abbildung 4) mit folgenden Eigenschaften:

- SORTIERUNG kombiniert NAME und VORNAME bzw. INITIALE in eine für die korrekte Sortierung notwendige Form.
- SYNONYM VON zeigt auf eine andere PERSON, die als gültig für dieses Objekt festgelegt wurde. So können unterschiedliche Schreibungen von Namen auf eine gültige zurückgeführt werden.
- TEIL VON zeigt auf einen Vorgänger, von dem das Objekt abgeleitet wurde.

Weitere Attribute, wie INTERESSENGEBIET (meist taxonomische Gruppen von Insekten) und PORTRÄT deuten an, dass hier eine Vernetzung zu anderen Projekten, wie Bibliographien und Biographien besteht.

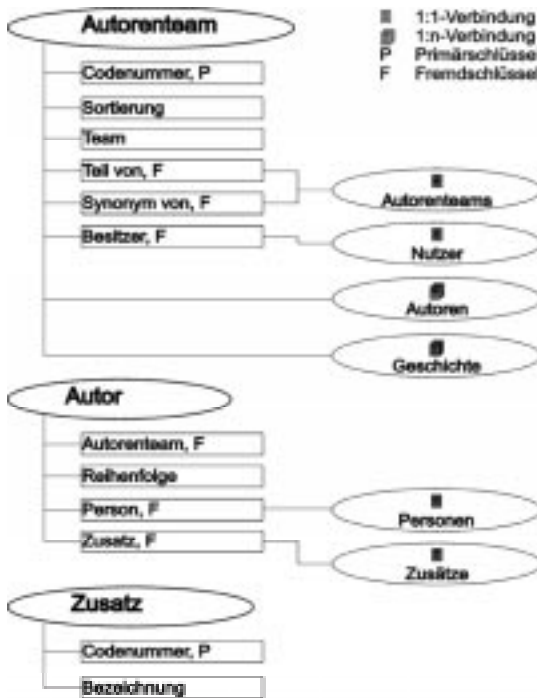


Abb. 3: Datenmodell für Autorenteams und Autoren

3.4 Datenmodell für Quellen

In der Teildatenbank QUELLEN sind Zeitschriften-, Buchtitel und Verlagsnamen abgespeichert, die über das Attribut QUELLE mit der PUBLIKATION verknüpft werden. Eine QUELLE hat folgende Attribute (Abbildung 5):

- SORTIERUNG entspricht NAME jedoch in einer für die korrekte Sortierung notwendigen Form.
- SYNONYM VON zeigt auf eine andere QUELLE, die als gültig für dieses Objekt festgelegt wurde. So können unterschiedliche Schreibungen von QUELLEN auf eine gültige zurückgeführt werden.
- VORGÄNGER zeigt auf eine QUELLE (meist eine Zeitschrift), aus der das Objekt hervorgegangen ist. Zusammen mit ZEITRAUM lassen sich die zahllosen historischen Umbenennungen von Zeitschriften verfolgen.

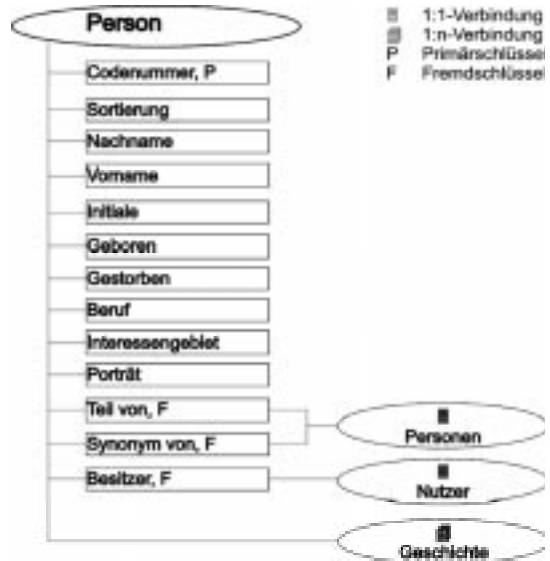


Abb. 4: Datenmodell für Personen

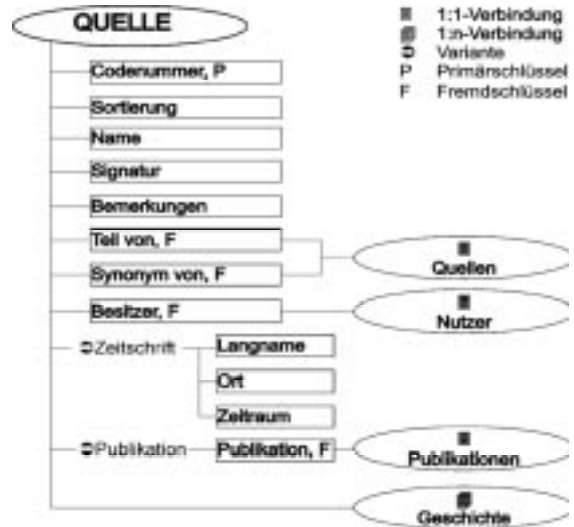


Abb. 5: Datenmodell für Quellen

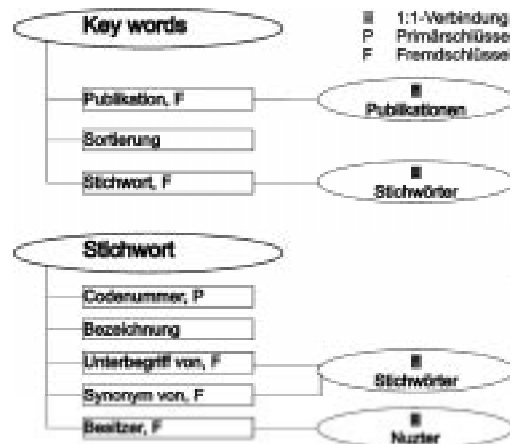


Abb. 6: Datenmodell für Stichwörter und ihre Anbindung

- LANGNAME, ORT und ZEITRAUM variieren die Konstruktion für die Bedürfnisse von Zeitschriftentiteln.
- PUBLIKATION schafft die Verbindung zu den PUBLIKATIONEN, wenn QUELLE der Titel einer anderen Publikation ist. Dieses Attribut muss bei der Implementierung mit Programmcode ergänzt werden, der NAME von QUELLE automatisch aus "In:" und AUTORENTEAM, JAHR, TITEL usw. einer ausgewählten Publikation zusammenstellt. Das z.B. ist der Fall, wenn Kapitel eines Buches als einzelne Publikationen erfasst werden.

3.5 Datenmodell für Stichwörter und ihre Anbindung

Wie aus Abbildung 2 ersichtlich, ist eine Erschließung der Publikationen mit Stichwörtern vorgesehen. Die dazu notwendigen Konstruktionen sind einmal zentral und einmal zusätzlich für jeden Nutzer vorhanden und in Abbildung 6 dargestellt. Das Prinzip ist eine 1:n-Verbindung von Stichwörtern zu Publikationen (Abbildung 6 oben). Die Stichwörter (Abbildung 6 unten) können mittels UNTERBEGRIFF VON beliebig tief hierarchisch verschachtelt werden. Diese einfache Struktur bringt für den Nutzer eine Reihe von Vorteilen:

- einfache, nicht zusammengesetzte Stichwörter ("Fauna", "Deutschland" vs. "Fauna von Deutschland");
- alphabetische Liste für schnelles Wiederfinden bei der Eingabe;
- hierarchische Darstellung für das Wiederfinden im Kontext;
- Eingabe so genau wie möglich! Bei der Eingabe können zunächst allgemeine Stichwörter, wie "Saltatoria" und "Europa" verwendet werden. Später, bei der Literaturauswertung, können sie durch genauere Angaben, wie "Verbreitung", "Decticus", "Brandenburg" ersetzt werden. Eine Eingabe der gesamten Hierarchie eines Begriffs ("Saltatoria-Decticinae-Decticus"), wie in herkömmlichen Systemen ist nicht notwendig;
- Recherche so genau wie notwendig! Spezielle Publikationen wird man mittels genauem, in der Hierarchie hinten stehendem, Stichwort suchen. Ist das Ergebnis unbefriedigend oder sucht man neben speziellen auch allgemeine Veröffentlichungen, kann man übergeordnete Stichwörter verwenden. Die Datenbank kann "Saltatoria" in alle untergeordneten Stichwörter, also auch "Decticus", auflösen und mit diesen anschließend Daten selektieren.

3.6 Datenmodelle für Anmerkungen und Druckformate

Weitere Daten können die Nutzer mittels privaten Anmerkungen ablegen (Abbildung 7 oben). Diese Attribute wurden auf Grund besonderer Wünsche der Nutzer im DEI angelegt. Sie sind vor anderen Nutzern verborgen, repräsentieren eine Teilmenge der gesamten PUBLIKATIONEN und können bei Bedarf auch außerhalb des DEI-Netzes bearbeitet werden.

Abbildung 7 unten zeigt ein Objekt, das nutzerspezifischen Programmcode zur Generierung von Ausgabelisten enthält. Prinzipiell können damit alle gewünschten Zitierformate erzeugt werden.



Abb. 7: Datenmodell für Anmerkungen und Druckformate

4 Implementierung

Nachfolgend werden einige der im Abschnitt 3 herausgearbeiteten Prinzipien nach ihrer Implementierung mittels PARADOX in Form spezieller Felder und Tabellen dargestellt.

4.1 Codenumber

CODENUMMER ist vom Typ Integer (ganze Zahl von -2.147.483.648 bis 2.147.483.647, wobei meist nur der positive Bereich genutzt wird). Jedesmal, wenn ein neuer Datensatz angelegt wird (eine Instanz eines Objektes erzeugt wird), vergibt PARADOX eine neue Nummer, die im Feld CODENUMMER gespeichert wird. Sie ist untrennbar an den Datensatz gekoppelt, d. h. sie kann nicht geändert werden und hört auf zu existieren, wenn der Datensatz gelöscht wird (es entsteht eine Lücke in der Folge der Codenummern). CODENUMMER dient als Primärschlüssel der Identifikation des Datensatzes und der Verknüpfung mit anderen Daten. Sie hat keinerlei semantische Bedeutung.

4.2 Vorgänger

VORGÄNGER ist, wie CODENUMMER, vom Typ Integer und dient als Fremdschlüssel der Referenz auf einen Datensatz in der gleichen Tabelle. Ein Datensatz, dessen Feld VORGÄNGER nicht leer ist, ist eine z. B. durch Teilung entstandene Ableitung von dem Vorfahren, auf den VORGÄNGER zeigt. Ein Vorfahre kann sich in beliebig viele abgeleitete Datensätze aufspalten. Aus Gründen der Lesbarkeit der Programme wird der Feldname seiner jeweiligen Funktion angepasst (Beispiel siehe Tabelle 1).

4.3 Nachfolger

NACHFOLGER ist, wie CODENUMMER, vom Typ Integer und dient als Fremdschlüssel der Referenz auf einen Datensatz in der gleichen Tabelle. Ein Datensatz, dessen Feld NACHFOLGER nicht leer ist, ist z.B. ein Synonym zum Nachfahren, auf den NACHFOLGER verweist. Beliebige viele Datensätze können sich auf einen Nachfahren beziehen (Beispiel siehe Tabelle 2).

4.4 Besitzer

BESITZER ist ein Fremdschlüssel aus der Tabelle NUTZER, die Name und Anschrift jedes Datenbanknutzers enthält (s. Abbildung 8). Das Feld BESITZER wird beim Anlegen des Satzes ausgefüllt. Einerseits wird damit transparent, wer für die Daten verantwortlich ist, andererseits gestattet oder verbietet ein System von Rechten den Zugriff auf Datensätze eines bestimmten Eigentümers. Die Tabelle speichert ebenfalls die E-Mail-Adresse jedes Nutzers. Das Programm erlaubt Nutzern ohne Schreibrecht, eine Nachricht mit Änderungswünschen am aktuellen Datensatz an den Besitzer zu senden.

Die Rechte sind sowohl auf Tabellen- als auch auf Datensatzebene organisiert. Vor dem Öffnen einer Teildatenbank wird überprüft, ob der im Netz angemeldete Mitarbeiter Leserechte (R), Schreib- (W) oder Supervisorrechte (S) für diese Tabelle hat (s. Tabelle 3). Neue Datensätze darf er mit W- oder S-Rechten anlegen. Das Supervisor-Recht ermöglicht einem Nutzer schließlich o.g. Rechte anderen Mitarbeitern zuzuteilen.

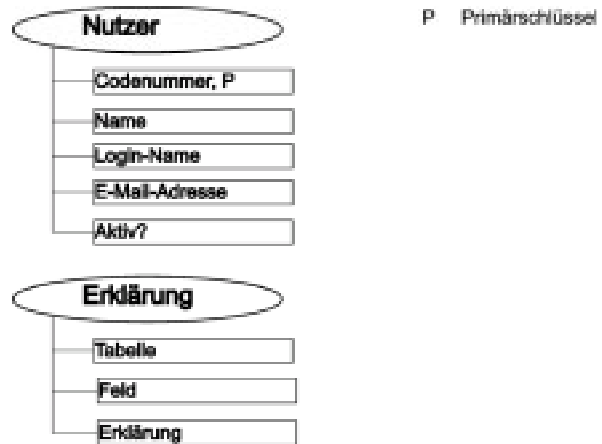


Abb. 8: Datenmodell für Verantwortliche und Hilfesystem

Tab. 1: Beispiel der Teilung eines Datensatzes in der Tabelle PERSONEN

PERSONEN					
CODENUMMER*	NAME	INITIAL	VORNAME	TEIL VON	SYNONYM VON
123	Sanderson	A. R.			
Bis zu einem gewissen Zeitpunkt wurden in der Datenbank 'ZITATE' alle Publikationen von 'Sanderson, A. R.' mit diesem Datensatz verknüpft. Spätestens bei der Erzeugung einer Bibliographie für 'Sanderson, A. R.' wurde sichtbar, dass es mehrere 'Sanderson, A. R.' geben muß. Eine Recherche ergab:					
567	Sanderson	A. R.	Arthur Rufus	123	
568	Sanderson	A. R.	Ann R.	123	
Die beiden neuen Datensätze sind aus der Sicht der ZITATE (und anderer Datenbanken, die PERSONEN benutzen) Teile des Satzes '123'. Das wird durch Nutzung des Feldes TEIL VON dokumentiert. Für eine herkömmliche Ausgabe einer Publikationsliste ist dieser Wissensfortschritt ohne Belang. Für o. g. Bibliographie muß jedoch für jeden abhängigen Datensatz in ZITATE entschieden werden, welche Person der Autor ist.					
123	Sanderson	A. R.			-1
Um den Wissensfortschritt bezüglich des Datensatzes 123 allen weiteren Nutzern sichtbar zu machen, wird das Feld SYNONYM VON des Satzes markiert. In der Anzeige erscheint 'siehe Teile!'. Das Programm stellt eine Funktion bereit, die dem Nutzer eine Liste aller Teile des aktuellen Datensatzes liefert. Der Datensatz ist jedoch weiterhin gültig und kann, wenn der Vorname nicht bekannt ist, für weitere Verknüpfungen benutzt werden.					

Tab. 2: Beispiel der Synonymisierung von Datensätzen in der Tabelle AUTORENTEAMS

AUTORENTEAMS				
CODENUMMER*	AUTORENTEAM	...	TEIL VON	SYNONYM VON
234	Storozhenko, S. Ju.			
bibliothekarische Transkription des Namens				
333	Storozhenko, S. Yu.			234
amerikanische Transkription von oben				
444	Storozhenko, S.			234
Schreibweisen ohne zweites Initial				
444	Strozhenko, S. Yu.			234
Fehlschreibung aus: Strozhenko, S. Yu. (1993): Review of the genus Formosatettix Tinkham (Orthoptera: Tetrigidae) from Japan, Russian Far East and adjacent regions. _ Akitu (134): 1_12 p.				
Mittels des Feldes SYNONYM VON sind alle abweichenden Schreibungen dem Datensatz 234 zugeordnet. In der Anzeige erscheint 'Synonym'. Zur korrekten Zitierung werden alle Schreibungen entsprechend ihres Auftretens in der Literatur benutzt. Soll eine Recherche alle Arbeiten von 'Storozhenko' ermitteln, so muß das als gültig festgelegte Autorenteam verwendet werden. Das Programm ermittelt zunächst alle Synonyme und anschließend die mit diesen Datensätzen verknüpften Zitate.				

Tab. 3: Beispiel für Rechte auf Tabellenebene

TABELLENRECHTE		
DATENBANK*	USER*	RECHT
1	4	R
1	7	W
1	8	S

Modifikationen von Datensätzen werden durch die Einträge in der Tabelle DATENSATZRECHTE gesteuert. Das Beispiel in Tabelle 4 zeigt, dass vom Mitarbeiter 7 in der Datenbank 1 angelegte Datensätze durch die Mitarbeiter 4, 15 und 20 geändert werden können. Alle anderen haben nur Leserechte an diesen Daten und müssen Änderungswünsche über eine Nachricht veranlassen.

Tab. 4: Beispiel für Rechte auf Datensatzebene

DATENSATZRECHTE		
DATENBANK*	EIGENTÜMER*	USER*
1	7	4
1	7	15
1	7	20

4.5 Sortierung

Mindestens ein weiteres Feld enthält die eigentlichen Daten der Tabelle. Sie machen die Bedeutung des Datensatzes aus, bestimmen seine Position bei der Sortierung mittels Sekundärindizes und müssen sich wiederfinden lassen. Sonderzeichenbehaftete Daten erfüllen diese Forderung jedoch nicht. So würde PARADOX den Namen "Åkermann, J." hinter "Zykan, T." einordnen. Beim gegenwärtigen Stand von ca. 46.000 Namen steht er etwa 45.000 Datensätze hinter "Akerman, Conrad", wo man ihn vermuten würde. Ein Nutzer, der diesen Datensatz verwenden wollte, würde ihn nicht finden und eine Dublette eingeben, die er überdies bei der nächsten Anforderung auch nicht finden könnte. Deshalb erhalten Tabellen mit sonderzeichenbehafteten Daten das Feld SORTIERUNG. SORTIERUNG enthält Daten eines oder mehrerer Felder des gleichen Datensatzes in kanonischer, d.h. sonderzeichenloser Form. SORTIERUNG wird beim Anlegen und bei jeder Änderung des Datensatzes ausgefüllt, indem landessprachliche Sonderzeichen auf lateinische Buchstaben zurückgeführt werden (Å wird A, š wird s usw.). Dieses Feld ist die Basis des alphabetischen Sekundärindexes. Es ist strenggenommen ein Verstoß gegen die Normalform der Daten (Sauer, 1998) und die Sortierung ist auch nur hinreichend korrekt. Mit der Überführung der Daten in ein DBMS, welches den Unicode-Zeichensatz unterstützt, kann es jedoch ersatzlos entfernt werden.

Für Felder mit Personen-, Zeitschriften- oder geographischen Name wurde deshalb ein DEI-interner Zeichensatz für WINDOWS entwickelt, der Sonderzeichen nahezu aller auf lateinischen Buchstaben beruhender Sprachen beinhaltet. Er ist in der Tabelle 5 dargestellt. Weitere Felder dieser Tabelle

enthalten Umwandlungsvorschriften in die kanonische Form, in HTML-Code und in Unicode-Zeichen.

4.6 Datensatzgeschichte

Weiterhin wird von allen wesentlichen Tabellen die Geschichte jedes Datensatzes in 1:n verbundenen Tabellen folgender Struktur verwaltet (Abbildung 9). Die Felder CODENUMMER, DATUM und ZEIT bilden den Primärindex in absteigender Folge, so dass das jeweils letzte Ereignis an erster Stelle steht. Bisher sind die Ereignisse "angelegt", "geändert", "synonymisiert", "geteilt", "kommentiert" (veraltet), "gelöscht" (nur intern) und "geprüft" definiert.

ZEICHENSATZ								
INDEX*	8x	9x	Ax	Bx	Cx	Dx	Ex	Fx
0	Å	ç	ž	°	À	á	à	ø
1	â	ž	†	†	Á	Ñ	á	ñ
2	†	†	ž	†	Â	Ò	â	ò
3	†	†	†	ž	Ã	Ó	ã	ó
4	†	†	†	†	Ä	Ô	ä	ô
5	†	†	†	ž	Å	Õ	å	õ
6	†	ž	†	†	Æ	Ö	æ	ö
7	†	Ł	ž	ž	Ç		ç	ž
8	ž	†	†	ž	È	Ø	è	ø
9	†	†	ž	ž	É	Ù	é	ù
A	Š	š	ž	ž	Ê	Ú	ê	ú
B	ž	†	†	ž	Ë	Û	ë	û
C	†	†	†	ž	Ž	Û	ï	ü
D	†	†	ž	ž	Í	Ý	í	ý
E	ž	ž	ž	ž	Î	Þ	î	þ
F	ž	†	†	ž	Ï	ß	ï	ÿ

Tab. 5: DEI-interner Zeichensatz (Ordnungsnummer hexadezimal)



Abb. 9: Datenmodell für die Datensatzgeschichte

4.7 Erklärungen

Auf der Ebene der gesamten Datenbank existiert eine Tabelle ERKLÄRUNGEN, die zu den Feldern aller Tabellen Hilfetexte und insbesondere Beispiele und Ausnahmen enthält. Das Beispiel in Abbildung 10 enthält nicht nur Hinweise, wie das Feld AUFLAGE gefüllt werden soll, sondern auch ein Rezept, wie bei der Erfassung des Index Litteraturae Entomologicae Serie I vorgegangen werden soll. Es entstand durch die Auswertung von Nachfragen der einbezogenen Mitarbeiter und dient somit der Kommunikation zwischen Datenbankverantwortlichen und übrigen Mitarbeitern.



Abb.10: Beispiel einer Erklärungsseite

5 Schlußbemerkungen

Alle vorgestellten Lösungen wurden implementiert, getestet und ständig vervollkommen. Einige Funktionen, wie das Nachschlagen von Feldwerten in Referenztabellen, arbeiten besser, als würde man die Daten per Hand eintragen und werden dementsprechend von den Nutzern akzeptiert. Andere, z. B. das Ableiten von Datensätzen sind schwerer zu vermitteln. Dabei wird deutlich, dass neben den in der vorlie-

genden Arbeit genannten technischen Maßnahmen auch Überzeugungsarbeit geleistet werden muss. Das Ziel, nachhaltig konsistente Daten als Referenz für andere Nutzer zu erzeugen und den Wissensfortschritt transparent zu machen muss höchste Priorität haben. Dateneingabe und -pflege via Internet sollten an solche Nutzer verteilt werden, die über Originaldaten verfügen oder als Experten allgemein akzeptiert sind. Die Nutzung solcher Datenbanken sollte einerseits nicht durch Zugriffsbeschränkungen oder Kosten erschwert und andererseits durch eine faire Quellenangabe unterstützt werden. Erfahrungen bei der Publikation eines Prototyps im Internet liegen bereits vor und sind Gegenstand späterer Veröffentlichungen.

6 Literatur

ANONYM (1994): Anwenderhandbuch. Borland Paradox für Windows. Version 5. Borland GmbH, Monzastraße 4c, D-63225 Langen. 620p.

ANONYM (1994): ObjectPAL-Programmierhandbuch. Borland Paradox für Windows. Version 5. Borland GmbH, Monzastraße 4c, D-63225 Langen. 590 p.

ANONXM (1994): ObjectPAL-Referenzhandbuch. Borland Paradox für Windows. Version 5. Borland GmbH, Monzastraße 4c, D-63225 Langen. 557 p.

ASC (1993): An information model for Biological Collections (Draft). Report of the Biological Collections Data Standards Workshop, August 18-24, 1992. Association of Systematic Collections, Committee on Computerization and Networking, <http://www.bishop.hawaii.org/asc-cnc>.

BIOTA (1997): Biota: The Biodiversity Database Manager. The Biota Data Model (Biota Manual, Appendix A), <http://viceroi.eeb.uconn.edu/BiotaPages/DataModel.html>.

DERKSEN, W.; Scheiding, U. (1963-1975): Index Litteraturae Entomologicae. Serie II: Die Welt-Literatur über die gesamte Entomologie von 1864 bis 1900. Akad. Landwirtschaftswiss. DDR, 5 Bände.

GAEDIKE, R.; Smetana, O. (1978): Ergänzungen und Berichtigungen zu Walter Horn und Sigmund Schenkling: Index Litteraturae Entomologicae, Serie I, die Welt-Literatur über die gesamte Entomologie bis inklusive 1863. Teil I A-K. Beitr. Ent. 28(2): 329-436.

GAEDIKE, R.; Smetana, O. (1984): Ergänzungen und Berichtigungen zu Walter Horn und Sigmund Schenkling: Index Litteraturae Entomologicae, Serie I, die Weltliteratur über die gesamte Entomologie bis inklusive 1863. Teil II: L-Z. - Beitr. Ent. 34(1): 167-291.

GROLL, E. K.; Taeger, A. (1997): Problems of data retrieval from entomological databases. - Abstr.: New Directions in Systematics / ESF Workshop, Crete 1997.

HORN, W.; Kahle, I. (1935): Über entomologische Sammlungen. - Ent. Beih. Berlin-Dahlem Bände 2-4.

HORN, W.; Kahle, I.; Friese, G. und Gaedike, R. (1990): Collectiones entomologicae. Ein Kompendium über den Verbleib entomologischer Sammlungen der Welt bis 1960. Teil I: A bis K. - Akad. Landwirtschaftswiss. DDR 1: 1-220.

HORN, W. und Schenkling, S. (1928-1929): Index Litteraturae Entomologicae Serie I - Berlin-Dahlem, Selbstverlag. 4 Bd. 1426 p.

OTTE, D., Naskrecki, P. (1997): Orthopteran Species File, <http://viceroi.eeb.uconn.edu/Orthoptera>.

SCHMITT, M.; Hübner, H. und Gaedike, R. (1998): Nomina Auctorum - Auflösung von Abkürzungen taxonomischer Autoren-Namen. Nova Suppl. Ent., Berlin 11(1998): 3-189.

SAUER, H. 1998: Relationale Datenbanken. Theorie und Praxis, Addison-Wesley, München.

Probleme und Erfahrungen beim Aufbau nachhaltiger konsistenter Datenbanken (E. K. Groll)

Zusammenfassung

Es werden Probleme der Dateneingabe durch verschiedene Nutzer, der nachhaltigen Konsistenz der Daten und der Abbildung des Wissensfortschritts in Datenbanken diskutiert. Die herausgearbeiteten, verallgemeinerbaren Lösungen, wie Synonymisierung oder Teilung von Datensätzen, Behandlung von Sonderzeichen und Verfolgung der Datensatzgeschichte bilden die Grundlage für Regeln und Datenmodelle, die anhand der von den Mitarbeitern des DEI genutzten Datenbank zur Neuauflage des "Index Litteraturae Entomologicae" vorgestellt werden.

Stichworte: Datenbank, Datenkonsistenz, Entomologie, Bibliographie

Problems and experiences of the construction of permanently consistent databases (E. K. Groll)

Summary

Problems stemming from data input by several users, problems of permanent consistence of data, and problems of representation of the progress of knowledge are discussed. Some generalized solutions as making synonyms of records or differentiation of records, handling of special characters and representation of the history of records are worked out. The deduced rules and data models are shown at the basis of the database of the project "Reprint of the Index Litteraturae Entomologicae Series I" in the Deutsches Entomologisches Institut.

Key words: Database, Consistency of Data, Entomology, Bibliography

Dr. Eckhard K. Groll arbeitet am Zentrum für Agrarlandschafts- und Landnutzungsforschung (ZALF) e.V., Deutsches Entomologisches Institut, Postfach 100238, D16225 Eberswalde, Telefon: 03334-589-816, Fax: 03334-21 23 79, e-Mail: groll@zalf.de