APIIS ein neuer Ansatz zur Erstellung konsistenter Datenbanken in der Tierzucht

RALF FISCHER, KÖLLITSCH; EILDERT GROENEVELD, MARIENSEE; UWE BERGFELD, KÖLLITSCH

Abstract
APIIS is a project to design an adaptable platform independent information systems for each kind of animal populations. Main design and requirements are described. Important parts are a generalised database structure, the numbering system to reflect the external identification into database internal sequencies, the definition of data streams and the creation of check rules.

1 Einleitung und Zielstellung

Die Erstellung von Datenbanken zur Abbildung gesammelter tierbezogener Informationen und deren Verwertung ist ein Schwerpunkt für jede Zuchtorganisation. Die anfallenden Daten werden dabei an einer Vielzahl von Standorten erhoben, beispielsweise auf den Betrieben selbst, in den Besamungs- oder Teststationen, bei den Landeskontrollverbänden (Labors), in den Schlachtbetrieben usw. An die Datenbanken, welche diese Informationen abzubilden haben, werden dabei vor allem folgende Anforderungen gestellt:

- Die Datenbank soll in sich konsistent sein,
- sie soll auf einer normalisierten Struktur beruhen,
- sie soll keine redundanten Informationen beinhalten und
- sie muss flexibel erweiterbar sein.

Weitere Schwerpunkte liegen in einer fehlerfreien Abbildung auch unterschiedlicher Nummernsysteme für die Tierkennzeichnung wie sie bei unterschiedlichen Datenquellen auftreten. Und nicht zuletzt sind die Kosten für ein solches System zu berücksichtigen. Hierfür ist die Nutzung einer einheitlichen generalisierten Struktur von Vorteil. Dieser ergibt sich durch die später mögliche gemeinsame Nutzung von Applikationen und der möglichen Verwendung der bereits normalisierten Datenbankgrundstruktur und der erstellten Programme zur Unterstützung für das Laden der Daten.
Ein neuer Ansatz zur Schaffung solcher Datenbanksysteme stellt das APIIS-Projekt dar, das in Grundzügen auf der GIL-Tagung in Bonn (GROENEVELD 1999) vorgestellt wurde.
Ziel war es,

- eine Vorgehensweise zu erarbeiten, welche eine Nutzung von lizenzfreien Komponenten erlaubt (an diese aber nicht gebunden ist),
- ein skalierbares System zu entwickeln sowie ein
- offenes System zu schaffen, welches vom Nutzer leicht adaptiert, modifiziert und erweitert werden kann.

Dabei liegt der Schwerpunkt in der Beschreibung einer Vorgehensweise und der Erstellung von Programmen zu deren Unterstützung. Zielgruppe dabei sind Programmierer welche vor der Aufgabe der Erstellung solcher Datenbanken stehen und diese Schritte mit einem wesentlich geringeren Aufwand an Zeit und Kosten umsetzen können, als dies sonst möglich wäre.
Um die relative Universalität zu gewährleisten beruht das gesamte Design auf Standards wie SQL und auf der Nutzung von frei verfügbarer Software, so dass eine Unabhängigkeit vom Betriebssystem, den eingesetzten Datenbanken wie auch von den Hardwareanforderungen gegeben ist. Ziel ist es, diese Komponenten an den Bedürfnissen der Nutzer entsprechend der Zuchthierarchie anzupassen und somit die Möglichkeit besteht, dieselbe Software sowohl auf dem Zuchtbetrieb als auch in der Herdbuchzentrale einsetzen zu können.
Insgesamt sind für die Umsetzung die folgenden Schritte notwendig:

- Entwicklung einer generischen Datenbankstruktur,

- Formulierung und Integration von Prüfregeln,
- Migration von existierenden Informationen aus bestehenden Informationssystemen,
- Abbildung von Datenströmen entsprechend dem existierenden Informationsfluss.

Der gesamte Prozess gliedert sich dabei wie folgt (modifiziert nach GROENEVELD 2002):

1. Sammeln aller möglichen Datenströme welche Eingang in die Datenbank finden sollen,
2. Feststellen aller Informationselemente,
3. basierend auf einer generalisierten Struktur, verteilen aller Informationen auf normalisierter Basis auf die Datenbanktabellen,
4. Schreiben einer zentralen Modelldatei beruhend auf den Informationen in 3.
5. Beschreiben der Prüfregeln für jedes Element in der Modelldatei,
6. Erstellen der Ladeobjekte für jeden Datenstrom (Nutzung als Stapelverarbeitung oder über GUI),
7. Laden der historischen Daten,
8. Erzeugen des weiteren notwendigen Programmoutputs (Meldebescheinigungen, Züchteranschreiben...),
9. Parallelbetrieb zur Evaluierung und evtl. Anpassen der Prüfregeln und Abschalten der alten Datenbank

2 Grundlagen und Designvoraussetzungen

2.1 Generalisierte Datenbankstruktur

Eine normalisierte und generalisierte Datenbankstruktur ist eine zentrale Voraussetzung zur Umsetzung der gezeigten Schritte. Dabei muss diese einfach an die gegebenen Bedürfnisse angepasst werden können, und sie muss unterschiedliche Tieridentifikationssysteme wie auch Testregime abbilden können. Da nicht eine einzelne Struktur diese Erfordernisse beschreiben kann, ist das Hauptaugenmerk auf die größtmögliche Übereinstimmung gelegt worden. Typische Bereiche für die Abbildung sind Leistungserfassungen im Feld oder auf Station, Besamung und Geburt, sowie der Komplex der Adressverwaltung. Eine Analyse einer Reihe von Zuchtprogrammen (3 Länder Schwein, 2 Länder Fleisch- und Milchrind sowie Schaf und ein System für Kaninchen) hat eine grundlegende gemeinsame Struktur erkennen lassen (GROENEVELD 2002b), die sich in zwei Arten von Tabellen, den immer notwendigen und den an die eigenen Bedingungen anzupassenden Tabellen, wiederspiegelt.

Zu den notwendige Tabellen, welche für alle analysierten Informationssysteme gemeinsam sind, zählen die Tabellen für das Tier selbst, einer Transfertabelle für die Tiere, in welcher sich ändernde Informationen für die Tiere wiederspiegeln, eine Kodetabelle sowie der Block Unit, Adresse und Namen. Weiterhin sind mehrere Transfertabellen erforderlich, welche die externen Kennzeichnungen in datenbankinterne umsetzen. Dies ist für Tiere, für Kode und Personen/Organisationen erforderlich. In diesen Tabellen befinden sich nur offene und somit eindeutige Datenkanäle für externe Kennzeichnungen.

2.2 Nummernsystem

Kern der Abbildung ist die Überführung der externen Kennzeichnungen in eine datenbankinterne Sequenz (GROENEVELD 2002a). Hierbei bilden die externe Kennzeichnung zusammen mit der meldenden Einheit eine eindeutige Identifikation. Für die Tierkennzeichnung wird davon ausgegangen, dass diese innerhalb der Lebenspanne eines Tieres in einem Betrieb eindeutig ist. Das schließt die Nutzung von bereits bestehenden eindeutigen Lebensnummern mit ein. Ziel ist dabei sowohl diese Tierkennzeichnungen eindeutig abbilden zu können wie auch den potentiellen Wechsel der Bedeutung von Kodes, etwa bei Änderung der Schlüssel für auftretende Anomalien oder Erkrankungen, zu ermöglichen.

Die Vorteile dieser Vorgehensweise liegen in der einfach zu handhabenden Umkennzeichnung von externen Identifikationen wie sie beispielsweise in der

Schweineproduktion bei erfolgter Selektion von Mutternummer + Spitze (laufende Ferkelnummer) zu einer eindeutigen Herdbuchnummer routinemäßig erfolgt.
Hierbei ist zu gewährleisten, dass die externen Kennzeichnungen bei Umstellung auf die selben internen Sequenzen verweisen. Weiterhin müssen bei Wiederverwendung von Nummern, beispielsweise in einem Mastbetrieb, die alten Datenkanäle geschlossen werden. Dies kann explizit über eine Meldung erfolgen (z. B. Schlachtung) oder auch wenn einen bestimmten Zeitraum keine weiteren Informationen für diesen Datenkanal erfolgt sind. Besondere Sorgfalt ist hier bei Vatertieren erforderlich, für die noch nach deren Ausscheiden Informationen für ihre Nachkommen auflaufen oder von denen beispielsweise noch tiefgefrorener Samen vorrätig ist. Eine Schließung des Datenkanals ist jedoch nur notwendig, wenn die selbe Kennzeichnung nochmals verwendet werden soll, was bei diesen Tieren kaum vorkommen sollte. Dabei ist es auch möglich nur bestimmte externe Kennzeichnungen für ein Tier zu schließen.
Zur Veranschaulichung der Umkennzeichnung ist ein kleines Beispiel in Tabelle 1 aufgeführt. In den ersten beiden Zeilen wurde eine betriebsinterne Tiernummer nach einer gewissen Zeit wiederverwendet, was dann natürlich die Schließung eines Datenkanals erfordert. Dagegen kann es bei einer Umnummerierung von Mutternummer + Spitze in die endgültige Herdbuchnummer (die unteren beiden Zeilen) auch mehrere offene Datenkanäle geben, wenn weitere Meldungen aus dem Betrieb auf dieser Grundlage erfolgen. Hiermit ist auch die gleichzeitige Nutzung unterschiedlicher externer Kennzeichnungen möglich wie dies beispielsweise bei der Nutzung von separaten Nummern auf den Prüfstationen erforderlich ist.

externe Kennzeichnung	Betrieb	Datenbankinterne Sequenz	Eintritts Datum	Ausscheide Datum	Datenkanal
1294	54	12346	17.06.92	12.12.94	geschlossen
1294	54	21544	24.03.01		offen
4711 / 12	47	54321	10.10.98		offen
320815	47	54321	14.08.99		offen

Tabelle 1 Beispiel für die Umkennzeichnung von externen Tieridentifikationen

2.3 Datenströme

Die Informationen aus der Außenwelt werden mit Hilfe von Datenströmen in die Datenbank aufgenommen .Datenströme sind hier definiert als routinemäßig eintreffende Daten, welche eine logisch und inhaltlich zusammenhängende Informationseinheit darstellen. Diese werden gemeinsam gesammelt und an die Datenbank zu einem definierten Zeitpunkt übertragen. Dies impliziert auch die definierte Zuständigkeit für die Erfassung und Übermittlung wie auch für die evtl. notwendige Fehlerkorrektur der Informationen. Datenströme lassen sich in zeitabhängige und zeitunabhängige Informationsströme unterteilen. Zu ersteren gehören die Informationen, welche dem Lebenszyklus eines Tieres folgen, wie beispielsweise: Geburt, Leistungsprüfung, Selektion und Belegung. Zeitunabhängige Meldungen können unter anderem der Besitzer- oder Standortwechsel sein.
Die Erstellung der aus diesen Datenströmen resultierenden Ladeobjekte erfolgt in zwei Schritten: Sammeln und Identifizieren der Datenherkunft sowie Beschreiben der zugrundeliegenden (redundanzfreien) Informationen und Aufteilen dieser auf die entsprechenden Zieltabellen. Dabei ist dieser Weg unabhängig davon, ob diese Informationen auf elektronischem Wege als Datei eintreffen oder über Masken (GUI) direkt eingegeben werden. Im Rahmen des Projektes wurde zur Erstellung der Ladeobjekte eine leicht zu definierende Struktur erarbeitet, welche die logisch inhaltlichen Abhängigkeiten im Datenstrom beschreibt.

2.4 Prüfregeln

Prüfregeln sind in einer Schicht zwischen Datenbankkern und Anwendungsprogrammen enthalten, wobei mehrere Profile für unterschiedliche Tiergruppen oder Bedingungen möglich sind (GROENEVELD 2002c). Hierbei werden die Daten nicht wie allgemein üblich beim

Aufnehmen in die Datenbank über Filter in jedem Softwaremodul entsprechend deren Funktionalität geprüft, sondern sind Teil der Datenbankstrukturdefinition selbst. Dabei wird auch eine spätere Änderung der Regeln bzw. deren Prüfschärfe sowie deren spätere Prüfung auf Konsistenz mit abgedeckt.
Die Anwendung erfolgt bei jedem schreibenden Zugriff automatisch, auch für die weiteren Abhängigkeiten innerhalb und zwischen den Tabellen, so dass keine gesonderten Plausibilitätsprüfungen auszuführen sind. Diese Prüfregeln sind für jede Datenbankspalte separat zu formulieren.
Die Nutzung verschiedener Prüflevel lässt unterschiedliche Prüfbedingungen sowohl für unterschiedliche Datenströme für die selben Tiere wie auch für unterschiedliche Tiergruppen zu. Beispielsweise ist bei Wurfmeldung das Geschlecht der einzelnen Ferkel noch unbekannt, später ist diese Information jedoch für viele Anwendungen unabdingbar und muss dann auch geprüft werden. Somit ist der Datenbankinhalt immer in einem definierten Zustand und es müssen beispielsweise für eine Zuchtwertschätzung keine weiteren Aufbereitungen erfolgen.

3 Anpassung an die sächsischen Verhältnisse

Für einen historischen Datenbestand des Mitteldeutschen Schweinezuchtverband e. V. würde die Erstellung einer solchen konsistenten Datenbank erprobt. Hierbei wurden neben den eigentlichen Herdbuchdaten auch die Meldungen der Mitgliedsbetriebe über Selektionen und Wurfergebnisse wie auch die in der Prüfstation anfallenden Informationen einbezogen. Insgesamt wurden dabei Informationen von etwas über 200 000 Tieren aufbereitet. Diese stammen aus 6 Datentabellen aus dem Herdbuch sowie einer Quelle aus der Prüfstation.
Als Datenbank kam PostgreSQL unter Linux zum Einsatz. Die gesamte Ladeprozedur für die bereits vorliegenden Daten erfordert dabei auf einem Pentium II mit 256 MB RAM, einschließlich der Prüfung auf Konsistenz sieben Stunden an reiner Rechenzeit.
Anhand dieser Daten zeigte sich, dass in Abhängigkeit von der Schärfe der Prüfregeln, bis zu 25 % der Daten infolge von logischen Inkonsistenzen zwischen den Datenquellen fehlerhaft waren. Diese logischen Verknüpfungen wurden bisher im Herdbuch nicht abgeprüft.
Für die folgenden routinemäßigen Datenströme werden in Abhängigkeit von der Informationsbereitstellung für Stationsdaten insgesamt 13 und für die Informationen aus den Herdbuchbetrieben 6 unabhängige Informationsquellen definiert. Für die sächsischen Verhältnisse sind das, neben den in jeder Organisation anfallenden Informationen, wie die Aufnahme eines neuen Züchters oder ähnlichem, insgesamt 18 spezifische Datenströme.
Insgesamt konnte gezeigt werden, dass die vorgeschlagene Vorgehensweise zur Erstellung durchführbar und auch für beträchtliche Datenmengen praktikabel ist.

4 Literaturverzeichnis

GROENEVELD, E. (1999); Design of a Portable Platform Independent Pig Information System; An Internet based Development Project in: Role and Potential of IT, Intranet and Internet for Advisory Services; Herausgeber G. Schiefer, U. Rickert, R. Helbig; ISBN 3-938227-08-5; S 113-128

GROENEVELD, E. (2002a); Platform independent information systems in animal breeding and research; 7th World Congress on Genetics applied to Livestock Production (WCGALP); 19.-23. August 2002, Montpellier, France

GROENEVELD, E.(2002b); Development of a Adaptable Platform Independent Information system in Animal Agriculture: Framework and Generic Database Structure; in press

GROENEVELD, E. (2002c); Development of a Adaptable Platform Independent Information system in Animal Agriculture: Implementation and Enforcement of Business Rules; in press