1. Einleitung

SIR steht für Scientific Information Retrieval und ist im Unterschied zu anderen Datenbanksystemen ein Datenmanegementsystem, welches außer der Lösung von Datenverwaltungs- und Retrieval-Problemen großen Wert auf einfache Möglichkeiten der Auswertung selektierter Daten legt. Dies wurde in 2 Ebenen realisiert: zum einen bietet SIR Funktionen an, die das Aggregieren von Daten, das Erstellen von Reports sowie einfache statistische Auswertungen erlauben; zum anderen verfügt SIR über direkte Schnittstellen zu den bekannten Statistik-Paketen SPSS, BMDP und SAS, so daß auch kompliziertere Auswertungen elegant durchzuführen sind.

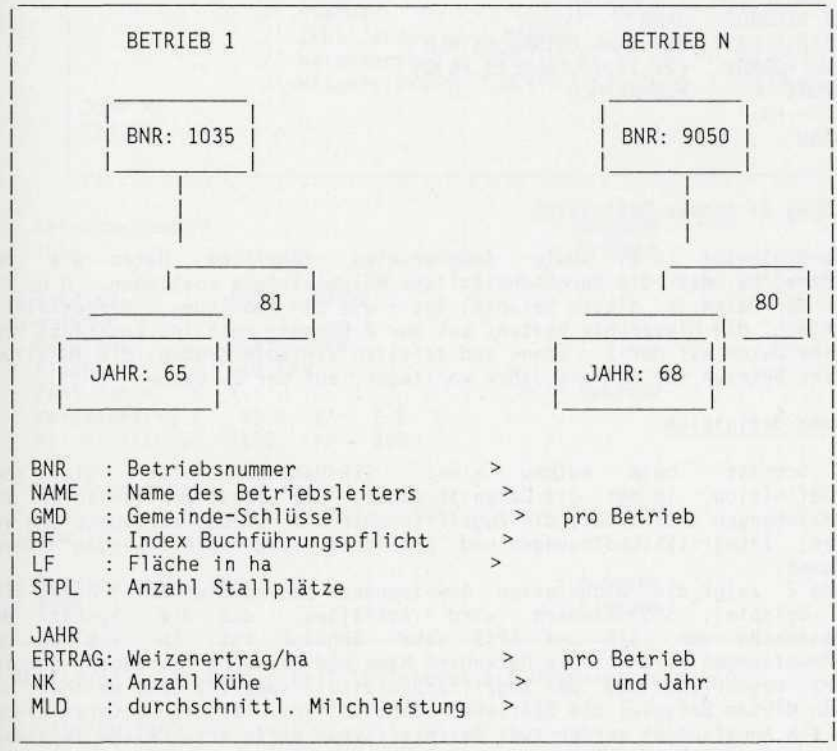


Abbildung 1: Daten-Struktur und Liste der Variablen

2. Ein Anwendungs-Beispiel

Ein einfaches Beispiel soll einen groben Eindruck vermitteln, wie eine SIR-Datenbank aufgebaut und wie mit ihr gearbeitet werden kann. Es soll dabei angenommen werden, daß von landwirtschaftlichen Betrieben einerseits allgemeine Informationen wie die Betriebsnummer, der Name des Betriebsleiters, der

```
FILE NAME       BEISPIEL
PASSWORD        OHM
DOCUMENT        Daten von landwirtschaftl. Betrieben
                Datensatztyp 1 : allgem. Betriebsdaten
                Datensatztyp 2 : jährliche Daten
CASE ID         BNR

RECORD SCHEMA   1,ALLGEMEIN
VARIABLE LIST   BNR,NAME,GMD,BF,FL,STPL
INPUT FORMAT    (2x,I4,A8,I4,I1,2I4)
VALID VALUES    BF (0,1)
VAR RANGES      GMD (100,999)
VAR LABELS      BNR, Betriebsnummer
VALUE LABELS    BF (1)Buchführungspflichtig /
                GMD (100)Odelhausen (200)Kuhdorf

RECORD SCHEMA   2,JÄHRLICH
SORT RECORDS    JAHR
VARIABLE LIST   BNR,JAHR,ERTRAG,NK,MLD
INPUT FORMAT    (2x,I4,I2,F4.0,I3,F6.0)
COMPUTE         MLG=NK*MLD

FINISH
```

Abbildung 2: Schema-Definition

Gemeinde-Schlüssel u.a. sowie andererseits jährliche Daten wie der Weizenertrag/ha oder die durchschnittliche Milchleistung vorliegen. D.h. die Struktur der Daten in diesem Beispiel ist - wie in Abbildung 1 dargestellt - hierarchisch, die Hierarchie besteht aus nur 2 Ebenen: zeitlich konstante bzw. allgemeine Daten auf der 1. Ebene und zeitlich variable Größen, die im allgemeinen pro Betrieb für mehrere Jahre vorliegen, auf der 2. Ebene.

2.1 Schema-Definition

Erster Schritt beim Aufbau einer SIR-Datenbank ist die sog. Schema-Definition, in der die Daten-Struktur, die einzelnen Variablen mit ihren Beziehungen zueinander, die Zugriffsrechte und -pfade, Wertebereiche von Variablen, Integritätsbedingungen und ähnliches spezifiziert werden können bzw. müssen.
Abbildung 2 zeigt die wichtigsten Anweisungen des kompletten Schemas für dieses Beispiel; SPSS-Kennern wird auffallen, daß die Syntax der Kommandosprache von SIR und SPSS sehr ähnlich ist. Im 1.Block der Schema-Anweisungen wird an die Datenbank Name und Passwort vergeben, ihr Inhalt kurz beschrieben und der Zugriffs-Schlüssel spezifiziert, welcher die Fälle (in diesem Beispiel die Betriebe) identifiziert. In den weiteren beiden Blöcken von Anweisungen werden zwei Datensatztypen definiert, welche in diesem Beispiel genau den beiden Ebenen der Datenhierarchie entsprechen. Man kann im Schema der Datensatztypen für einzelne Variablen zulässige Werte bzw. Bereiche angeben, wie dies hier für die Variablen BF und GMD geschehen ist; d.h. als Werte der Variablen BF werden in der Datenbank nur 0 und 1 akzeptiert, die Gemeinde-Nummern müssen im Intervall 100 bis 999 liegen. Um Daten-Konsistenz zu erreichen, können auch komplexere, d.h. mehrere Variablen betreffende Bedingungen angegeben werden; so wird beispielsweise hier verlangt, daß bei positiver Milchleistung die Kuhzahl nicht Null sein darf. Man kann weiterhin Etiketten (Labels) für Variable bzw. Variablenwerte vergeben oder kann die Er-

rechnung neuer Variable - wie der gesamten Milchleistung pro Betrieb (MLG) - veranlassen.

```
     RETRIEVAL
M1:  COMPUTE        BENU=NREAD('Betriebsnummer')
     IF             (BENU EQ 9999) EXIT
     CASE IS        BENU
     MOVE VAR LIST  NAME,GMD,FL
     COMPUTE        GNAME=VALLAB(GMD)
     PROCESS REC    2
     COMPUTE        NJ=CNTR(JAHR)
     COMPUTE        ME=MEANR(ERTRAG); SE=2*STDEVR(ERTRAG)
     COMPUTE        ML=MEANR(MLD); SL=2*STDEVR(MLD)
     WRITE          /' Betrieb        ',BNR(I4),' ',NAME(A10)
                    /' =======        ',
                    /' Gemeinde     : ',GNAME(A30)
                    /' Fläche       : ',FL(I4),' ha'
                    /' Zahl Jahre   : ',NJ(I2)
                    /' Weizenertrag : ',ME(F6.1),'  +/-  ',SE(F3.1)
                    /' Milchleistung: ',ML(F6.0),'  +/-  ',SL(F4.0)/
     JUMP M1
     FINISH
```

```
Betriebsnummer?                              - Ausgabe
1055                                         - Eingabe

Betrieb         1055 , LOISL          >
=======                                >
Gemeinde     :  Odelhausen              >
Fläche       :  110 ha                   >
Zahl Jahre   :  4                         >  - Ausgabe
Weizenertrag :   59.6  +/-  2.6           >
Milchleistung:  6130.  +/-  320.         >
                                        >
Betriebsnummer?                        >

usw.

Betriebsnummer?                              - Ausgabe
9999                                         - Eingabe
```

Abbildung 3: Anwendungsfall 'Retrieval mit Datenaggregation'

2.2 Daten-Eingabe

Nach der Definition des Schemas kann die Datenbank mit Daten gefüllt werden, wobei beim Einlesen Datensätze, welche im Sinne der Vereinbarungen des Schemas fehlerhaft sind, auf eine Ausgabedatei ausgesteuert werden; in diesem Fall sind die entsprechenden Korrekturen an Hand des genauen Fehlerprotokolls durchzuführen und die berichtigten Daten nachzuladen. Selbstverständlich ist das Hinzufügen neuer Daten sowie das Löschen oder Modifizieren bereits gespeicherter Daten jederzeit möglich.

2.3 Retrieval mit Datenaggregation

Die Möglichkeiten der Informationsrückgewinnung (Retrieval) sollen zunächst an Hand eines einfachen interaktiven Programmes, welches nach Eingabe einer Betriebsnummer eine kleine Betriebsstatistik ausgibt, demonstriert werden. Dabei sollen insbesondere die jährlichen Betriebsdaten über alle verfügbaren Jahre aggregiert werden. Das Programm (siehe 1.Teil der Abbildung 3 auf Seite 353) zeigt das Einlesen der Betriebsnummer mittels Aufruf der SIR-Funktion NREAD, den direkten Zugriff auf den entsprechenden Betrieb und die Selektion der Variablen NAME,GMD,FL sowie des zugehörigen Gemeinde-Namens. Durch Aufruf der SIR-Funktionen CNTR, MEANR und STDEVR für die Datensätze des Typs 2 werden Anzahl, Mittelwert und Standardabweichung aller in der Datenbank vorhandener Jahres-Erträge und Milchleistungen des gewünschten Betriebes gebildet. In einen WRITE-Befehl wird die genaue Form der Ausgabe spezifiziert; für Erträge und Milchleistungen werden die 2s-Intervalle ausgegeben. Der 2.Teil der Abbildung 3 auf Seite 353 zeigt einen mit diesem Programm geführten Dialog, wobei die Eingabe des Benutzers und die Ausgabe des Programmes gekennzeichnet wurde. An dieser Stelle muß, um Mißverständnissen vorzubeugen, erwähnt werden, daß in dieser Beispiels-Datenbank keine realen, sondern nur fiktive Daten gespeichert wurden.

2.4 Report-Generator

Der folgende Anwendungsfall soll den Report-Generator von SIR, der die Erstellung von Tabellen ermöglicht, in seiner Funktion zeigen. Dabei sollen pro Gemeinde für ein bestimmtes Zeitintervall die Zahl der Betriebe, die jeweiligen Durchschnittserträge und Milchleistungen ermittelt und in Listenform ausgegeben werden (siehe 2.Teil der Abbildung 4 auf Seite 355). Von dem zur Lösung dieser Aufgabe notwendigen SIR-Programm - bestehend aus einem Retrieval- und Report-Teil - ist in Abbildung 4 auf Seite 355 nur der Report-Teil aufgeführt. Durch Angabe des SORT-Parameters wird veranlaßt, daß die selektierten Daten nach Gemeinden und pro Gemeinde nach Jahren geordnet werden. Als Benutzer des Report-Generators kann man nun angeben, welche Aktionen beim Abarbeiten dieser sortierten Daten an bestimmten Stellen, sog. Breakpoints, vorzunehmen sind. Diese Breakpoints werden in LEVEL- Anweisungen definiert und sind in diesem Beispiel die Sprünge von einer Gemeinden zur nächsten bzw. die Übergänge von einem Jahr zum folgenden. Am Report-Anfang wird der Text 'Gemeinde-Statistiken', bei Änderung des Gemeinde-Schlüssels der Gemeinde-Name ausgegeben. Pro Jahr werden Erträge und Milchleistungen aufsummiert und jeweils vor dem Sprung zum folgenden Jahr deren Mittelwerte ausgegeben.

2.5 Statistische Datenanalyse

Der letzte Anwendungsfall soll die Möglichkeiten der statistischen Datenanalyse mit Hilfe von SIR sowie in Verbindung mit SPSS aufzeigen. Und zwar sollen Betriebsdaten von 2 Gemeinden und 3 Jahren selektiert, in einem Streuungsdiagramm dargestellt und beispielsweise varianzanalytisch ausgewertet werden. Abbildung 5 auf Seite 356 zeigt zunächst den Retrieval-Teil, in dem die Variablen JAHR bis MLD für die Betriebe der Gemeinden 100 und 300 sowie der Jahre 74 bis 81 selektiert werden. SIR bietet für Auswertungen einige Verfahren der deskriptiven Statistik an. Die hier aufgerufene Prozedur PLOT beispielsweise erzeugt das im 2.Teil von Abbildung 5 auf Seite 356 skizzierte Streuungsdiagramm. Zur Durchführung einer Varianzanalyse ist allerdings die Übergabe der Daten an ein Statistik-Paket erforderlich. Da SIR - wie schon erwähnt - Schnittstellen zu SPSS, BMDP und SAS besitzt, kann diese Übergabe an SPSS einfach mit dem Kommando SPSS SAVE FILE bewerkstelligt werden. SPSS kann dann mit dem im letzten Teil der Abbildung 5 auf Seite 356 dargestellten Steuerkarten-Set eine Varianzanalyse durchführen, in der eine Erklärung der Variation von Ertrags- und Milchleistungsdaten durch die Faktoren Jahre, Gemeinden und dem Index für Buchführungspflicht versucht wird.

```
REPORT          FILENAME=SRF/SORT=GMD,JAHR
BEFORE REPORT
WRITE           ' Gemeinde-Statistiken'/' ==================='//
LEVEL           1,GMD
COMPUTE         GNAME=VALLAB(GMD)
WRITE           ' Gemeinde : ',GNAME(A20)/' --------'/
LEVEL           2,JAHR
COMPUTE         JER=0; JML=0; JN=0
WRITE           ' Jahr I Zahl I Ertrag I Milchl.'/
                ' -----I------I--------I--------'
DETAIL
COMPUTE         JER=JER+ERTRAG; JML=JML+MLD; JN=JN+1
AT END
COMPUTE         JER=JER/JN; JML=JML/JN
WRITE           JAHR(I5),' I ',JN(I4),' I ',JER(F6.1),' I ',JML(F6.0)
END REPORT
```

```
Gemeinde-Statistiken
====================


Gemeinde : Odelhausen
--------

Jahr I Zahl I Ertrag I Milchl.
-----I------I--------I--------
  78 I   20 I   56.7 I  5720.
  79 I   25 I   57.1 I  5780.
  80 I   27 I   57.9 I  5810.
  81 I   22 I   58.5 I  5820.

Gemeinde : Kuhdorf
--------

     usw.
```

Abbildung 4: Anwendungsfall 'Retrieval und Report-Generator'

3. Zusammenfassung

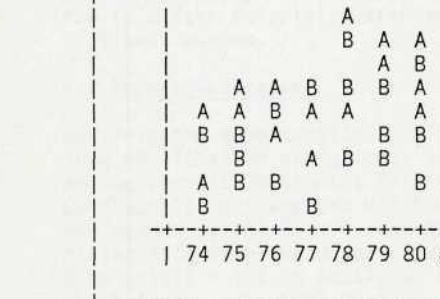
Im Rahmen des oben beschriebenen Anwendungs-Beispiels konnte nur ein Teil der SIR-Leistungen erläutert werden. Daher werden im folgenden die wichtigsten Funktionen von SIR kurz zusammengestellt :

- SIR kann als Datenbankmanagementsystem große, sowie hierarchisch oder netzwerkartig strukturierte Datenmengen verwalten und die anfallenden Retrieval-Probleme lösen.

- SIR bietet über ein vielschichtiges Konzept der Passwort-Vergabe die Möglichkeit, vorhandene Datenschutzprobleme zu bewältigen. Dabei können beispielsweise auch einzelne Variable dem Zugriff bestimmter Benutzer entzogen werden.

```
GET FILE        BEISPIEL                                        .
PASSWORD        OHM                                             .
RETRIEVAL                                                       .
FOR EACH CASE                                                   .
SELECT CASE IF (GMD EQ 100 OR 300)                              .
FOR EACH REC    2, FROM (74) THRU (81)                          .
MOVE VAR LIST   JAHR,GMD,BF,ERTRAG,MLD                          .  SIR -
SUMMARY                                                         .  Programm
                                                                .
PLOT            SCATTERGRAM=ERTRAG WITH JAHR BY GMD             .
SPSS SAVE FILE FILENAME=SSF                                     .
FINISH                                                          .
                                                                .
```

Ertrag



Ergebnis der SIR-Prozedur PLOT

```
RUN NAME        Varianzanalyse                                  .
GET FILE        SSF                                             .
ANOVA           ERTRAG,MLD BY JAHR,GMD,BF                       .  SPSS -
STATISTICS      ALL                                             .  Programm
FINISH                                                          .
```

Abbildung 5: Anwendungsfall 'Retrieval und statistische Datenanalyse'

- Durch Angabe von gültigen Wertemengen, Bereichen und Konsistenz-Bedingungen in der Schema-Definition kann eine Prüfung der Daten bei ihrer Eingabe veranlaßt werden.

- SIR bietet in Verbindung mit SPSS, BMDP und SAS elegante Möglichkeiten der Auswertung selektierter Daten : und zwar teils durch Aufruf spezieller SIR-Prozeduren, teils durch den einfachen Übergang auf eines der erwähnten Statistik-Pakete.

- Zur Aufbereitung selektierter Daten in Tabellenform kann der Report-Generator eingesetzt werden.

- Für das interaktive Arbeiten mit der Datenbank stellt SIR einen leistungsfähigen Editor bereit, der erstens die Erstellung und Korrektur von Programmen und Datensätzen unterstützt sowie zweitens das Starten und Steuern von Dialogprogrammen ermöglicht.

- SIR übernimmt weiterhin die Verwaltung der Retrieval- und Auswertungs-Programme parallel zu den Daten. Dies ist z.B. nützlich, da mit einem Kommando eine Sicherung der kompletten Datenbank mit allen Programmen auf Magnetband veranlaßt werden kann.

- Erwähnenswert ist, daß die Ausführung beliebiger Systemkommandos unter SIR möglich ist. Dies bedeutet beispielsweise, daß man - ohne das System SIR zu verlassen - einen SPSS-Job starten, sich nach dem Zustand gestarteter Jobs erkundigen oder andere System-Funktionen zur Ausführung bringen kann.

Aus der Sicht des Benutzers ist die Ankündigung eines deutschen SIR-Handbuches interessant, in dem - wie zu hoffen ist - die Schwächen des derzeit verfügbaren amerikanischen Manuals hinsichtlich Klarheit und Verständlichkeit weitgehend ausgemerzt sein werden.

Insgesamt jedoch kann SIR als benutzerfreundliches System und brauchbares Instrument zur Lösung von Datenmanagement-Problemen empfohlen werden. SIR ist in erster Linie für Anwender konzipiert, die nach dem Retrieval die selektierten Daten mit mehr oder weniger aufwendigen Methoden weiter auswerten wollen - eine Indikation, welche vor allem im wissenschaftlichen Bereich fast immer vorliegt.

Anhang

SIR wurde in USA entwickelt, wobei Dr.Beutel die Betreuung der europäischen Kunden übernommen hat und auf Anfrage weitere SIR-Informationen zur Verfügung stellt:

SIR Inc.
P.O. Box 1404
Evanston, Illinois 60204, USA

Dr. Peter Beutel
Postfach 101340
6900 Heidelberg

SIR ist derzeit in speziellen Versionen für folgende Rechner verfügbar:
CDC (NOS,NOS/BE), IBM (OS/VS,CMS), PERKIN ELMER (OS/32),
PRIME (450,750 unter PRIMOS), SIEMENS (BS2000,BS3000),
UNIVAC (unter VS9), VAX (11/780 unter VMS).
Weitere Versionen sind geplant bzw. in Vorbereitung.