SAS-Data Warehouse Technologie für die Forschung : eine raum-zeit-thematische Datenbank für Economic Research

Autor(en): **Schüle, Hannes**

Objekttyp: **Article**

Zeitschrift: **Geschichte und Informatik = Histoire et informatique**

Band (Jahr): **9 (1998)**

PDF erstellt am: **17.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-7237

SAS-Data Warehouse Technologie für die Forschung. Eine raum-zeit-thematische Datenbank für Economic Research

Hannes Schüle[1]

Zusammenfassung

Das Ordnen, Ablegen, Verknüpfen und zur Verfügung stellen von raum-zeit-thematischen Informationen für Forschung und Lehre hat die historische Statistik seit langem beschäftigt. Der Beitrag beleuchtet den Weg des «Data Handlings» von Statistikfiles mit systematischer Nomenklatur hin zu einem modernen Data Warehouse (DWH).

Am Beispiel eines solchen DWH für ein Economic Research Team werden die Möglichkeiten der neuesten Technologie des SAS Systems (dem führenden Anbieter von Auswertungssystemen) dargestellt.

Das Arbeiten im Serververbund[2] und die neueren theoretischen Konzepte der Datenarchitektur (multidimensionale DB mit Stern- oder Snowflake-Schema) ermöglichen es auch Forschungsteams, umfassende und komplexe Datenbestände systematisch in einem DWH zu halten, zu pflegen und zur Verfügung zu stellen. Die Metadaten (Daten über Daten) stellen die Datenintegrität sicher und speichern alle Änderungen und Erweiterungen.

Der Autor gelangt zum Schluss, dass die Konzepte aus den Forschungsprojekten der 80er Jahre keineswegs überholt sind, dass viele der damals entwickelten Ideen und Lösungen heute den Weg in die Prospekte der Softwarehäuser gefunden haben. Die interaktiven Oberflächen erleichtern das Arbeiten und das Verstehen der Abläufe für BenutzerInnen gewaltig. Gerade Forschungsprojekte sollten der systematischen Datenhaltung, der Dokumentation und dem Einhalten von definierten Prozessen beim Einfügen neuer Datenbestände genügend Aufmerksamkeit schenken.

---

1 Der Autor hat in verschiedenen historischen Forschungsprojekten raum-zeit-thematische Datenbanken mitaufgebaut. Heute leitet er das Information Delivery Team der EDS (Schweiz) AG. Zu seinen Kunden gehören neben Grossfirmen der Finanzindustrie auch Forschungsteams.
Er möchte an dieser Stelle Jan Burse, Rolf Locher, Christian Schneider und dem SAS Institut für die Unterstützung beim Verfassen dieses Beitrags herzlich danken.

2 Im Serververbund können Daten-, Zugriffs- und Programmlogik optimal getrennt werden, mit der sogenannten «multi tire architecture».

1. Data Warehouse – mehr als ein Modewort?

Das Data Warehouse (DWH), oft wohl richtiger als Information Warehouse bezeichnet, ist seit einigen Jahren ein wichtiges Thema in grossen Konzernen, vorab in jenen der Finanzindustrie – ist doch eine Bank nicht viel mehr als eine umfassende Datenbank in der Kunden, Bestände, Konti und Transaktionen abgelegt sind. Während das *DWH einen umfassenden Prozess* für Aufbau, Update, Haltung und Zugriff der Daten beschreibt, wird für die eigentliche *Datenhaltung ein RDBMS* (Relationales Datenbank-Manangement System) eingesetzt. Auf den DWH-Markt drängen denn auch von der einen Seite Anbieter von Auswertungssystemen (etwa SAS Institut, Information Builders) und von der anderen Seite die Datenbankanbieter (z.B. IBM mit DB2, Oracle). Entsprechend unterschiedlich werden die Prozesse oder die Datenhaltung ins Zentrum gestellt. Mit SAS lassen sich zur Datenhaltung nicht nur SAS-Files, sondern beliebige RDBMS im Hintergrund einsetzen, je nach Plattform und Kundenwünschen.
Die Ausgangslage, die zum Konzept des DWH führt, ist folgende:

- In unterschiedlichsten Bereichen eines Betriebes fallen *operationelle Daten* an, welche dort in Datenbanken oder Tabellen gehalten, gespeichert und verändert werden.
- Für die verschiedensten Zwecke werden in einem Betrieb *Auswertungen aus Datenbeständen unterschiedlichster Provenienz* gemacht. Dazu müssen Daten herumkopiert und angepasst werden. Mit der Erstellung von Auswertungen und v.a. mit der aufwendigen Datenbeschaffung werden viele «IT»-Ressourcen (IT=Informationstechnologie, einst «EDV») gebunden. Es geht viel Zeit verloren von der Vermutung eines möglichen Zusammenhangs bis zum Vorliegen von Auswertungen, welche diese untermauern, präzisieren oder verwerfen.

Zur Rationalisierung und zur umfassenden Verknüpfung der Daten aus allen Bereichen für alle Arten von Auswertungen sollen *alle Daten zentral gehalten und beschrieben* werden. Daraus leitet sich das Konzept eines DWH ab:

- *Trennung von operationellen und Auswertungsdaten.*
- *Metadaten*[3] nicht nur über alle vorhandene Information, sondern auch als elektronische Beschreibung der Prozesse zu ihrer Akquisition und

3 Metadaten können (1) eher technischer Art sein, also Tabellen, in denen Tabellen, Indices und Tabellenbeziehungen beschrieben sind, oder (2) inhaltlicher Art, um Schlüssel aufzulösen, Beziehungen (Oberbegriff von, Summe von etc.) abzubilden, Quellen anzugeben, oder (3) deskriptiver

der Zugriffsmöglichkeiten, also ein umfassendes *Repository*[4] (vgl. Abb. 1).

- Einheitliche, homologisierte, standardisierte zentrale Datenhaltung. Dabei können statt der Datenhaltung auch nur zentral definierte Sichten auf die operationellen Daten zum Zuge kommen. Entscheidend ist der *subjekt- bzw. themenorientierte* Datenzugang.[5]
- Umfassendes *Copy-Management* zum Initialisieren und Updaten des DWH aus operationellen Datenbeständen.
- *Vorverdichtungen*[6] aus der Sicht der Auswertungen oft direkt in «*MDDB*s»[7] (Multidimensionale Datenbank) mit Daten unterschiedlicher *Granularität*.[8]
- Umfassende *Werkzeuge zum Zugriff* auf die DWH-Daten über Management-Informations-Systeme (MIS mit Drill Down[9]), Intranet, SQL für «Data Mining»[10] (Datenanalyse), «Data Marts» (feste, auswertungs-

---

Art: Analysen von Tabellen, wie statistische Werte (Mittel, Min, Max, N) von Variablen oder Beziehungen von Teilmengen beschreiben (Raum-Zeit, Raum-Term, Term-Zeit). In einem DWH werden Metadaten gebraucht, um das Warehouse aufzubauen, gleichzeitig resultieren aber auch (andere) Metadaten aus der Analyse des DWH.

4 Repository: Zentrale Beschreibung von Daten und Applikation sowie von deren Änderungsgeschichte.

5 «Subjekt» meint hier etwa «Kunde» und schliesst alle Kundendaten aus der Werbeabteilung, der Bestellabteilung und der Buchhaltung ein. Dies ganz unabhängig davon, wie die Transaktionen dort erfasst und abgelegt werden. Ein analoges Beispiel aus der Forschung ist etwa das Subjekt «Perinatale Mortalität» das aus unterschiedlichsten Datenbeständen gebildet wurde: aus Toten- und Taufrödeln, aus Daten kommunaler Einwohnerkontrollen, Spitälern und Statistikabteilungen von Gesundheitsämtern.

6 (Vor-)Verdichten: Bereitstellen von Daten auf höheren Hierarchiestufen (verdichten von Gemeindedaten zu Bezirken und Kantonen oder von Produktedaten zu Produktegruppen) unter Anwendung einfacher statistischer Methoden zum Zwecke des raschen Zugriffs bei Drill Down-Analysen. Oft nachgefragte Verdichtungen werden zur Verbesserung der Performance bereits beim Laden der MDDB berechnet und nicht erst bei der Abfrage.

7 In einer MDDB werden Daten nach Analysedimensionen gehalten und meist entsprechend vorverdichtet. In einer MDDB zur Arbeitslosigkeit können Dimensionen wie Raum, Geschlecht, Altersgruppe, Beschäftigungszweig und Herkunft enthalten sein. SAS hat ein eigenes, sehr schnelles Datenformat für die MDDB entwickelt.

8 Granularität: «Körnigkeit», gemeint ist die niedrigste Auflösung von Daten in einem DWH. Möglicherweise ist es nicht notwendig, jeden Einkauf eines jeden Schokoriegels zu speichern, sondern nur die Verdichtung «Süsswaren durch Kunden Einzelhaushalte je Woche, je Filiale». Sowohl der Persönlichkeitsschutz als auch die Kosten des Speicherplatzes führen u.U. zum Verzicht, alle anfallenden operationellen Urdaten dauerhaft abzulegen.

9 Drill Down: Interaktives Hinunterblättern auf tiefere Hierarchiestufen in der Datenbetrachtung. Beispielsweise aus einer Übersicht von Arbeitslosenraten nach Bezirken können per Mausklick nach vordefinierten Dimensionen (Gemeinden, Geschlecht, Altersgruppen, Beschäftigungszweigen) detailliertere Daten angezeigt werden. Dahinter steht eine MDDB, als Werkzeug kann etwa SAS/EIS eingesetzt werden.

10 Data Mining: Datenanalyse zur Bestimmung von Trends oder Mustern; Suche nach dem «Diamanten» in grossen Datenbeständen. Data Mining ist der eigentliche Motor zur Entwicklung von DWH. Wie findet eine Bank genau jene 2000 Kunden heraus, welche am ehesten dazu bereit sind, in der nächsten Woche eine neue Kreditkarte zu erwerben? Vektorisierbare mathematische Modelle haben Hochkonjunktur. Die Resultate von Data Mining können erneut in das DWH einfliessen (zur Erweiterung des Regelwerks in wissensbasierten Systemen).

orientierte Sichten für spezielle NutzerInnen) oder «Information Marts»[11] (vorverarbeitete Datenauswertung).
- Handhabung der *Zugriffsrechte* und Protokollierung der Zugriffe.
- Und schliesslich ein *Werkzeug zur Administration des DWH.*

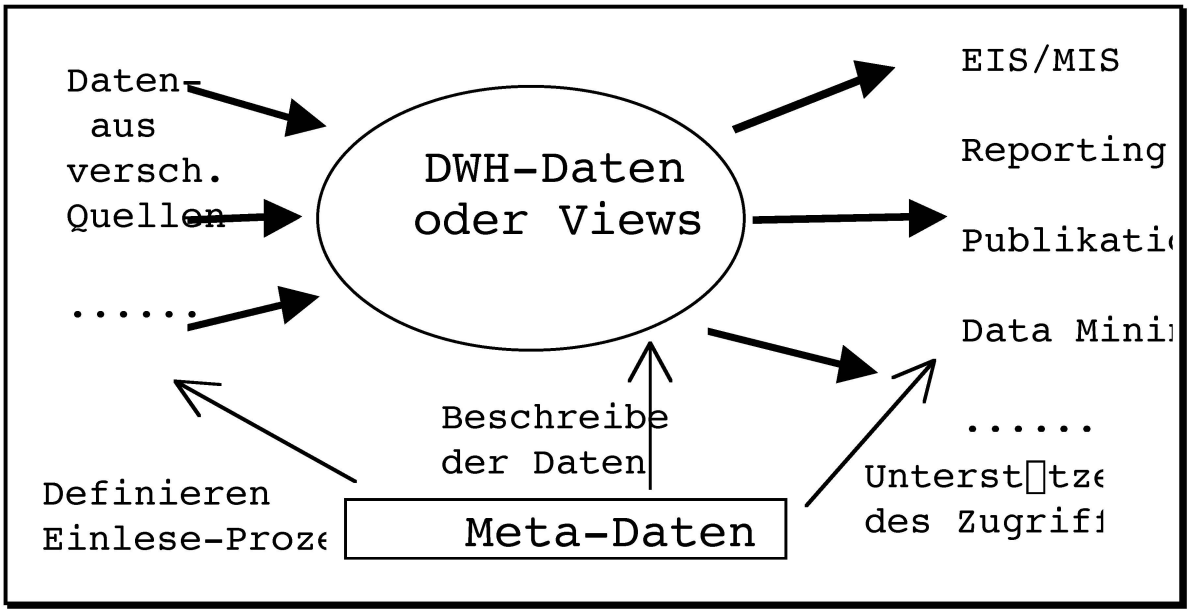


Abb. 1: Grundkonzept eines DWH (vergleiche auch die ganzseitige Abb. 2 mit dem «DWH-Knochen» nach SAS-Institut).

11 Information Mart: Vorverarbeitete Datenauswertung wie standardisierte Reports und Grafiken, welche die BenutzerInnen über POD (Print Output Distribution) oder Internet erreichen.

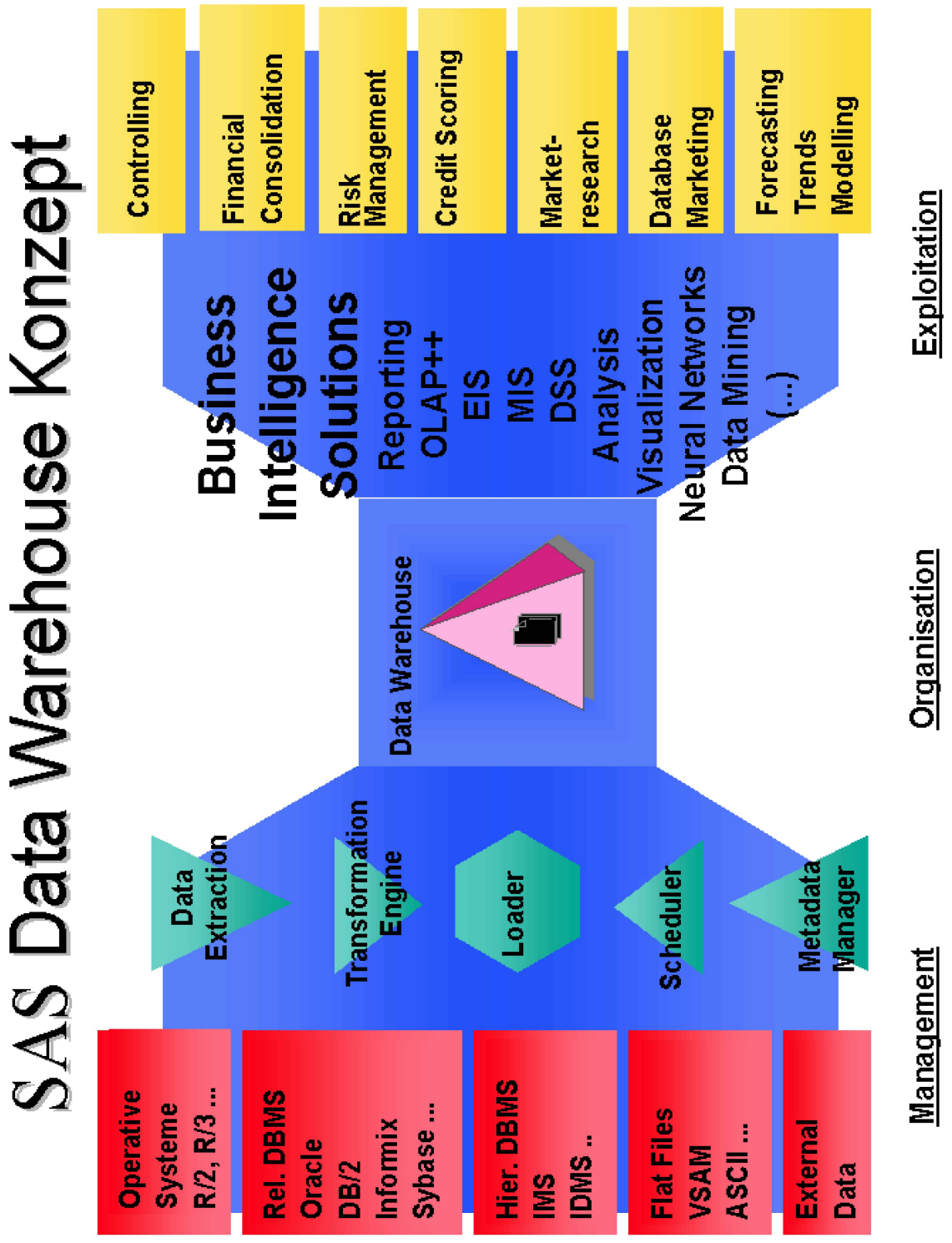


Abb. 2: Grundkonzept eines DWH («DWH-Knochen») nach SAS-Institut (Quelle: Referat C. Caforio, SAS-Institut, Brüttisellen).

Ein DWH steht oder fällt (!) mit der Datenarchitektur und den Administrationswerkzeugen. Wird aber ein durchdachtes DWH sinnvoll in die betrieblichen Prozesse integriert, steht einem «ROI» (return of investment) von bis zu mehreren hundert Prozent per annum nichts im Wege.

Betrieblich scheitert die Idee eines unternehmensweiten DWH vor allem an den fehlenden Sponsoren. Zickert[12] diskutiert, ob ein solches als Profitcenter geführt werden soll und gibt einem Costcenter den Vorzug.

Ein DWH bedingt unweigerlich betriebsorganisatorische Änderungen: Die Aufgaben der Informatik in einem Betrieb verändern sich durch den Einsatz eines DWH. Zwar steht immer noch die IT zwischen den Daten und den EndbenutzerInnen, doch statt Auswertungen baut die IT Sichten und Auswertungssysteme auf die Daten. Die BenutzerInnen erhalten Werkzeuge, um massgeschneiderte Reports selbst zu generieren. Dabei wird nicht weniger, aber ganz andere IT-Unterstützung nachgefragt. Decision Makers können selbständig und bedeutend flexibler Analysen generieren. Es können weit mehr Hypothesen getestet werden als früher.

2. Raum-zeit-thematische Fragestellungen im Geschichts-«Labor»

In den frühen 80er Jahren entstanden mehrere Projekte, welche über die Analyse demographischer Zeitreihen hinausgingen und umfassende raum-zeit-thematische Fragestellungen mit quantitativen Methoden angingen. So wuchs auch – ursprünglich aus dem Erfassen von Toten- und Taufrödeln der Kirchgemeinden – das Konzept «Bernhist»: Eine in ihrer thematischen Breite einmalige Datenbank zur Regionalgeschichte des Kantons Bern.[13] Zur umfassenden geo-historischen Interpretation der Datenmenge wurde ein eigener erkenntnistheoretischer Ansatz entwickelt.[14] Das Projekt führte zu einer Vielzahl von Artikeln, Seminararbeiten, Lizentiaten und Dissertationen und zu einer öffentlichen, relationalen Datenbank. Es fand mit der Publikation einer Monographie[15] nach rund zwölf Jahren seinen vorläufigen Abschluss.[16]

---

12 Zickert, Christina: *Konzeption eines Information Warehouse*. St. Gallen 1992. (Seminararbeit an der Universität St. Gallen), S. 48f.

13 Pfister, Christian und Schüle, Hannes: «BERNHIST: a Laboratory for Regional History». In: *History and Computing II*, hrsg. von Peter Denley, Stefan Fogelvik und Charles Harvey. Manchester 1989, S. 280-286.

14 Pfister, Christian und Schüle, Hannes: «Encompassing Geo-Histoire. Methodological dimensions and historiographical implementations of the «BERNHIST» interdisciplinary information system». In: *Histoire et Informatique V, Actes du Congrès*, hrsg. von Joseph Smets. Montpellier 1992, S. 225-250.

15 Pfister, Christian: *Im Strom der Modernisierung. Bevölkerung, Wirtschaft und Umwelt 1700-1914.* Geschichte des Kantons Bern seit 1798, Band IV, Bern 1995.

16 Zur Darstellung des Projekts «Bernhist» als fächerübergreifendes Historisch-Geographisches Informationssystems vergleiche den Artikel Imfeld et al.; daselbst findet sich auch eine breitere

Am Anfang lagen einige demographische Zeitreihen vor, welche der Interpretation in einem grösseren Kontext bedurften. Das innovative Konzept sah nun nicht in erster Linie die Analyse von Daten, sondern den Aufbau eines «laboratory for regional history» in Form einer «*dynamic database* for the spatial analyses of population, economy and the environment» als «open-ended process» vor. Die Idee eines offenen Prozesses liegt auch dem Data Warehouse zu Grunde: räumlich flexibel und thematisch ausbaubar, in Schritten wachsend, sich im Datenbestand, den Sichten und in den Zugriffswerkzeugen den ändernden Bedürfnissen in kurzen Entwicklungszyklen anpassend.

Mattison zeigt ein Modell, indem sich verschiedene taktische Probleme zu einer strategischen Vision verdichten, statt die Probleme einzeln angehen zu wollen und dabei die Vision aus dem Augen zu verlieren.[17] Das Modell lässt sich leicht auf das Vorgehen in grösseren Forschungsprojekten umsetzen: Statt immer nur einzelne Fragestellungen zu bearbeiten und nur die dazu notwendigen Quellen zu erheben, soll mit dem Aufbau eines grossen, verknüpften Datenbestandes ein längerfristig ergiebigeres Ziel verfolgt werden.

3. Unechte Datenbanken: Strukturierte Statistik-Files

Welche Aufgabenstellungen die Informatik beim Aufbau einer einfachen raum-zeit-thematischen Datenbank in der Form strukturierter Statistikfiles zu bewältigen hat und wie diese bewältigt werden können, zeigt das folgende Beispiel, in welchem Anlehnung an die bereits ältere historische Demographie und an die sozialwissenschaftliche Statistik gesucht worden ist.

In «Bernhist» wurden die Quellen zuerst in Statistikfiles eingelesen, plausibilisiert, standardisiert, homogenisiert und zu nach und nach immer grösseren, themenorientierten Statistikfiles zusammengefügt. Ursprünglich wurde mit SPSSX gearbeitet. Von Stefan Fogelvik, Betreiber des Historischen Datenarchivs in Stockholm[18], haben die Berner SAS als Werkzeug übernommen, um damit systematisch in grossem Umfang Grafiken (Zeit-

---

Auswahlbibliographie. Imfeld, Klaus; Häberli, Peter; Pfister, Christian und Schranz, Niklaus: «BERNHIST – eine Plattform für fächerübergreifendes Forschen und Lehren in Raum und Zeit. Konzept und Potential eines Historisch-Geographischen Informationssystems (H-GIS)». In: *Landesgeschichte und Informatik, Itinera Fasc. 17*, hrsg. von Christoph Döbeli et al. Basel 1996, S. 46-77. Der seit längerem geplante Historisch-Statistische Atlas erscheint 1998.

17 Mattison, Rob: *Data warehousing. Strategies, technologies and techniques*. New York 1996, S. 40f. Vom gleichen Autor in diesem Zusammenhang zu empfehlen: Mattison, Rob: *The object-oriented enterprise. Making Corporate Information Systems Work*. New York 1994.

18 Fogelvik, Stefan: «The Stockholm Historical Database at Work». In: *History and Computing II*, hrsg. von Peter Denley, Stefan Fogelvik und Charles Harvey. Manchester 1989, S. 256-265.

reihen, Altersaufbau, Säulenplots) und Karten zu produzieren. Obwohl die Daten also in Form von Statistikfiles gelagert wurden, konnte in diesen Projekten der Historischen Sozialforschung durchaus von «Datenbank» gesprochen werden[19], waren die Files doch über einheitliche Schlüssel zur Untersuchung einer Vielzahl von Forschungsfragen miteinander verknüpfbar.

Neben der Prüfung der eingelesenen Quellendaten beherrschten sechs, eher technische Aufgaben den Aufbau der Datenbank; Aufgaben die auch zwölf Jahre später für ein DWH von Bedeutung sind:

- Um diachrone Analysen über einen Zeitraum von 250 Jahren zu ermöglichen, erfolgte eine *Standardisierung der Raumstruktur:* Alle Daten wurden auf der Struktur des Kantons Bern in den heutigen Grenzen erfasst oder umgerechnet.[20] Für die urbanisierten Gebiete werden um 1920 die Eingemeindungen nachvollzogen.[21]
- Nur mit gleichwertigen Massen lässt sich rechnen. *Homogenisierung von Masseinheiten:* Zum einen erfolgt die Umrechnungen ins metrische System.[22] Zum andern gibt die Anpassung der *regionaltypischen Masse* an ein einheitliches System grosse Probleme auf.
- In dieser grossen, wachsenden Datenmenge konnte eine *systematische Nomenklatur der Variablen* ein gewisses Mass an Übersicht garantieren. In drei Typen von Statistikfiles gelang es, zehntausende von Variablen einigermassen sprechend[23] zu bezeichnen:
  - In *Raumfiles* mit dem vierstelligen Raumcode als Schlüssel wurden die Variablennamen aus dem Begriffskürzel und dem Jahr gebildet.
  - In *Zeitreihen* (Schlüssel: Jahr) wurden die Variablen analog bezeichnet, doch statt des Jahres stand der Schlüssel des Raumes im Variablennamen: So steht RTAU351 für Rödel-Taufen in der Gemeinde 351.

19 Thaller, Manfred: «Vorüberlegungen für einen internationalen Workshop über die Schaffung, Verbindung und Nutzung grosser interdisziplinärer Quellenbanken in den historischen Wissenschaften». In*: Datenbanken und Datenverwaltungssysteme als Werkzeuge Historischer Forschung*, hrsg. von Manfred Thaller. St. Katharinen 1986, S. 9-30, S. 10f.

20 Lediglich Fussnoten geben noch Hinweise auf die in einer sehr frühen Phase des Dateneinlesens vorgenommenen Umrechnungen und Interpolationen.

21 Für die Gebiete um die Städte Bern und Biel wird für die Jahre 1700 bis 1919 auf die alte und ab 1920 auf die heutige Gemeindestruktur abgestellt.

22 Zudem Umrechnungen von alten Hohlmassen für Getreide auf Gewichte, von der Kategorisierung von Vieh nach Alter auf eine nach Gewicht oder von alten Flächenmassen auf heutige Hektaren.

23 Variablenamen mit acht Zeichen sind wie folgt aufgebaut: Einheit (1 Zeichen, bspw. H für Anbaufläche in Hektaren, D für Ernte oder Schlachtgewicht in Doppelzentnern), Thema (3 Z., bspw. WEI für Weizen, WIE für Wiesland, OCH für Ochsen), Jahr (3 Z., etwa 847 für 1847) und Zusatz (1 Z., etwa G für Anteil am Getreide total in %). Z.B. HWEI847G: Anbaufläche Weizen in % der Getreidefläche total in Jahre 1847 oder GOCH886: Ochsen, umgerecht in Grossvieheinheiten 1886.

  - Für sehr differenzierte Einzelquellen, etwa Volkszählungen, mussten erweiterte Systeme gefunden werden, um beispielsweise ‹männliche Gemeindebürger zw. 30 und 39 Jahren im Jahre 1910› in einem achtstelligen Variablennamen zu bezeichnen.
- Zeitreihen und Raumfiles konnten ineinander *transponiert* werden. Dazu wurde mit einem Zwischenfile gearbeitet, das vier Schlüsselfelder (Jahr, Raumcode, Themakürzel, Einheit) und ein Wertefeld enthielt. Auf dieser Basis ist später auch die öffentliche Datenbank erstellt worden.
- Es galt auch, eine *sinnvolle Grösse* der Datenfiles zu finden, die Speicherplatz, Aufwand für Verknüpfungen, Update, Analysen und Wartung berücksichtigten.
- Einzelne *Anmerkungen* zu Daten, bestimmten Zahlen, Räumen, Begriffen und Quellen wurden in einfachen Textfiles erfasst, mit Steuerzeilen für die Schlüssel in Anlehnung an die Codes für Raum und Term. Beim Ausdruck der Daten werden die zugehörigen Anmerkungen als Fussnoten ausgeben.

Der Ansatz, die unterschiedlichsten Quellendaten zu homogenisierten, themen-(subjekt-)orientierten Files mit Vorverdichtungen in Raum und Thema zusammenzufügen, kann aus heutiger Sicht als *konzeptionelle Stärke* bezeichnet werden.

Zu *bemängeln* hingegen ist zum einen die starke Vermischung von Daten- und Programmlogik: Umrechungsfaktoren, Korrekturfaktoren, Interpolationen, ja sogar eigentliche Quellenkorrekturen finden sich als fixe Statements in den Programmen. Ohne umfassendes Programmstudium lässt sich der Weg von der Viehzählung zur Gesamtsumme der Grossvieheinheiten schwerlich nachvollziehen. Als weitere erhebliche Schwäche der Datenbank muss die ‹festverdrahtete› Standardisierung der Raumstruktur angesehen werden. Auch wenn eine einheitliche Raumstruktur zur Bildung von Zeitreihen und zur diachronen Analyse mehrerer Zeitschnitte sinnvoll ist, müsste der Weg von den Quellendaten zum Themenfile sowohl für den Raum, wie für Masseinheiten tabellengesteuert, nachvollziehbar, automatisch dokumentierbar und leichter änderbar sein.

Bei der Frage, wie weit heute ein Forschungsinstitut eine *zentrale Datenbank* für alle Daten führen soll, gehen die Meinungen auseinander. Während der Autor eher einem allgemeinen, generischen[24] Datenmodell für alle raum-zeit-thematischen Daten zugetan ist, betrachtet Lukas Vogel,

24 «Generisch» heisst, dass das gleiche Datenmodell für eine Vielzahl von themenorientierten Tabellen angewandt werden kann sowie thematisch, räumlich und zeitlich offen angelegt ist.

damals Mitarbeiter der Forschungsstelle für Schweizerische Sozial- und Wirtschaftsgeschichte der Universität Zürich, eine zentrale Datenbank als problematisch, wohl nicht nur aus technischer, sondern auch aus betrieblicher Sicht.[25] Vogel ist aber vehementer Verfechter einer zentralen «Meta-Datenbank» und betrachtet eine gesamtschweizerische Stelle zur Langzeitarchivierung von Forschungsdaten, wie sie die Sozialwissenschaften bereits kennen,[26] als wünschenswert.

4. Beispiel eines Data Warehouses

4.1. Ausgangslage

Die Untersuchung von raum-zeit-thematischen Fragestellungen an grösseren Datenbeständen ist keineswegs HistorikerInnen und SozialwissenschafterInnen vorbehalten. Das Bundesamt für Statistik oder die Konjunkturforschungsstelle der ETHZ gehören zu den grössten Lieferanten solcher Daten, welche dann von Dritten erneut mit weiteren Datenquellen verknüpft und, wie im vorliegenden Fall, etwa nach volkswirtschaftlichen Aspekten untersucht werden. Die dabei an die Datenhaltung gestellten Anforderungen sind jenen von historischen Forschungsprojekten durchaus ähnlich und leiten sich u.a. durch eine sich verändernde Begriffs- und Raumstruktur, durch Lücken, durch unterschiedliche Detaillierungsgrade und durch verschiedene Zuverlässigkeit der Daten ab.

Das Erstellen fundierter, wissenschaftlicher Prognosen und Ratings zu den unterschiedlichsten Wirtschaftsbereichen und Regionen der Schweiz ist die Haupttätigkeit eines Forschungsteams mit rund 15 ÖkonomInnen und StatistikerInnen, unterstützt von einer grösseren Gruppe für Infrastruktur (v.a. Informatik und Publikation). Schlüsselaktivitäten des Teams sind volkswirtschaftliche Analysen und deren Publikation, denen die unterschiedlichsten Datenquellen zu Grunde liegen: zum einen Zeitreihen auf der Basis (Tag,) Monat, Quartal oder Jahr und zum anderen Raumdaten vom Hektarraster über Gemeindedaten bis zu Kantonsdaten. Die Stärken des Teams sind:

- breite thematische Ausrichtung mit umfassenden Schwerpunkten
- klare geographische Fokussierung auf die Schweiz und ihre Regionen
- hervorragend ausgebildetes Team
- umfassende Datenbasis
- kompetente IT-Unterstützung

25 Vogel, Lukas: «Das Projekt FSWbase». In*: Geschichte und Informatik / Histoire et Informatique, Vol 5/6*, hrsg. von Hannes Schüle. Basel 1995, S. 115-118.

26 Sozialwissenschaftliches Datenarchiv SIDOS in Neuchâtel, vgl. den entsprechenden Beitrag von Reto Hadorn in diesem Band.

– gute Kundenbeziehungen in die verschiedensten Sparten der Finanzindustrie.

Die Publikation der Resultate erfolgt in sehr unterschiedlicher Form: in kurzen Memos, als Paper, in Hochglanzprospekten, aber auch in elektronischer Form als Webpages oder als Input für das Regelwerk einer wissensbasierten Anwendung (WBA) zur Beurteilung von Kreditvergaben.

Die Tätigkeit des Teams wird gehemmt durch den grossen Aufwand für die Datenbeschaffung, Datenhaltung und -historisierung sowie die schwierige Austauschbarkeit von Analysen und Teilanalysen zwischen MitarbeiterInnen. Dauernd werden Daten-CD-ROM sowie Excel- und Accessfiles hin und her gereicht. Die Suche nach einer Lösung mit zentraler, einheitlicher Datenhaltung, einem Data Warehouse also, liegt daher nahe.

4.2. Ein Data Warehouse für ein Economic Research Team

Die grundsätzlichen Ziele, die mit dem Aufbau eines Data Warehouses erreicht werden sollen, sind:

– Einheitliche Datenbasis für alle Analysen für alle MitarbeiterInnen
– Raschere Einbindung neuester Daten
– Weniger Reibungsverluste für Datenabstimmung und -beschaffung

Das Ziel für den ersten Release des Data Warehouse Projekt ist es, *möglichst schnell ein bequemes Abfragetool zu entwickeln*. Struktur der Datenbank und Export von Abfrageergebnissen haben Vorrang, während die Schnittstellen zu den Quellen und die Dokumentation nur von zweiter Priorität sind. Ebenfalls als sekundär werden fixe Reports und automatische Updates betrachtet.

Daraus leitet sich zum ersten ab, dass ein *protozyklisches Vorgehen*, bei dem in mehrmonatigen Entwicklungsschritten jeweils wesentliche neue Teile einer angestrebten Gesamtlösung entwickelt werden, die Priorisierung aber für jeden Zyklus (Release) selbst neu definiert wird. Damit kann rasch auf sich ändernde betriebliche Anforderungen und auf Probleme oder Mängel aus den vorangegangenen Zyklen reagiert werden. Zum zweiten ist klar geworden, dass die Entwicklung in enger Zusammenarbeit zwischen den VertreterInnen des Research Teams und deren IT-Support in Angriff genommen werden muss.

Als zentrale Anforderungen an den *ersten Release* wurden gestellt:

– Einfaches, *generisches Datenmodell*, damit eine hohe Flexibilität erreicht werden kann und neue Daten mit SAS/Base-Kenntnissen geladen werden können.

- Ein interaktives *Abfragetool* mit Auswahl von Raum, Themen und Zeitrahmen sowie Periodentyp mit diversen Extras.
- Handhabung einer sich über die Zeit *verändernde Raumstruktur*.
- Export der Abfrageresultate in Excel-Tabellen.
- Schulung mindestens einer/eines MitarbeiterIn des Research Teams als *DWH-AdministratorIn* um selbständig die Datenbank aufbauen und BenutzerInnen schulen zu können. Zudem *Schulung des IT-Supports,* um die SAS-Software und das DWH bei BenutzerInnen installieren und warten zu können.
- Als allgemeine Anforderung muss das *Qualitätsmanagement sicher stellen*, dass die Logik der Datenstruktur und -definition ausserhalb der Programme und Methoden abgebildet wird, damit die *Daten den Lebenszyklus der Applikation überdauern*.

Eine *Evolutionäre Vorgehensweise* bei der Softwareentwicklung wird schon seit länger Zeit diskutiert und angestrebt. Entsprechende Projektstrukturmodelle finden sich in der Literatur genauso wie in Planungsinstrumenten. Die Konzepte des raschen und zyklischen Vorgehens sind jüngeren Datums. Sie heissen etwa «Rapid Application Development»[27], «Rapid Iterative System Engineering»[28], Rapid Prototyping[29] oder eben «protozyklisch». Wobei letzteres besonders den Anspruch hervorstreicht, keine perfekten, dafür um so rascher und effizienter sinnvolle, ausbaufähige Releases zu entwickeln. Im Laufe des Prozesses treten die Konturen der anfänglich eher verschwommenen «Zielwolken» immer deutlicher hervor.

Das Projektteam bestand aus dem Projektleiter, einer SAS-Entwicklerin (vorübergehend) und zwei Junioren, davon ein Informatikstudent mit fundierten Datenbank- und SQL-Kenntnissen. Die Realisierung dauerte knapp vier Monate, der Aufwand betrug etwa vier Personenmonate (wobei die Arbeit der Junioren nicht voll angerechnet wird).

Zur Zeit ist der erste Release ausgeliefert und getestet. Der DWH-Administrator ist voll und ganz mit der Phase des Datenaufbaus beschäftigt, nachdem er sich im Selbststudium SAS/Base beigebracht hat.

Als Hardware steht ein NT-Server mit 9 GB Speicherkapazität zur Verfügung. Eine spätere Migration auf eine Alpha-Maschine (ev. im Zusammenhang mit Zugriff via Intranet) ist vorgesehen.

---

27 Mattison, Rob: *Data warehousing. Strategies, technologies and techniques*. New York 1996, S. 234f.

28 RISE-Guide, EDS 1997 (internal); viele Firmen kennen eigene Vorgehensmodelle zur Entwicklung von Software.

29 *Rapid Warehousing Methodology*, hrsg. von SAS Institut. Cary 1998 (3rd ed.).

4.3. Ein offenes, generisches Datenmodell für ein Forschungs-Data Warehouse

Als Kernstück wurde von Anfang an *eine einzige Datentabelle* mit jeweils nur einem Wert und einem zusammengesetzten Schlüssel aus *Raum*key, *Term*key und *Perioden*key ins Auge gefasst. Darum herum gruppieren sich *polyhierarchische Thesauri* für Raum, (Zeit) und Thema, die je aus der Grundtabelle mit den den Schlüsseln zugeordneten Begriffen und einer Beziehungstabelle (mit Gültigkeitsdauer und Beziehungsart) bestehen. Raumeinheiten und Raumbeziehungen haben eine Gültigkeitsdauer. Die Idee, statt zweien nur einen Thesaurus zu führen, haben wir ausführlich diskutiert und aus pragmatischen Überlegungen verworfen. Zu jedem Begriff gehört noch eine *Quellenangabe* (mit Gültigkeitsdauer).

Die *Abfragen* werden auf der Abfragetabelle mit Timestamp[30] als Schlüssel, Parent (ein «Menu», das ebenfalls in der Abfragetabelle definiert wird), Zeitrahmen und Periodentyp abgelegt. Zu jeder Abfrage gibt es beliebig viele Einträge in den Tabellen «Abfrage-Raum» und «Abfrage-Term». Eine Abfrage-Perioden-Tabelle lässt sich später implementieren. Vorerst ist diese nur eine View auf den gewählten Zeitrahmen und den gewählten Periodentyp in der Abfragetabelle. Frühere Abfragen können jederzeit wieder geholt, bei Bedarf angepasst und neu ausgeführt werden.

Zu jeder Quelle, zu jedem Term und zu jedem Raum sollen beliebig lange *Texte* erfasst und in einer Tabelle abgelegt werden können. SAS stellt dazu einen einfachen Texteditor zur Verfügung. Die Texte können vom/von der AdministratorIn erfasst und von allen BenutzerInnen eingesehen werden. Beim Export der Abfrageergebnisse in Excel werden die Texte zu allen gewählten Begriffen und Räumen am Schluss der Tabelle ausgegeben. Die Textfunktion ist in einem späteren Release erweiterbar.

Das Datenmodell von «Bernhist»[31] stellte sich als nicht genügend heraus. Die Anforderungen *Multidimensionalität* musste *generisch* implementiert werden. Ein vieldimensionales *Stern-Schema*[32], eine Zuordnung von Begriffen zu mehreren Klassen (auch Analysedimensionen) ist notwendig. Auf diesem Bedarf beruht die Klassentabelle (wie: Geschlecht, Berufsgruppe, 5-Jahres-Kohorten) und die Klassencodetabelle (wie: männlich, Bäcker, 25-29-jährige) sowie die Beziehungstabelle Term-Klassencodes.

30 «Zeitstempel»: vom System auf eine Tausendstel- oder Millionstel Sekunde genau generierter, eindeutiger Schlüssel.

31 Imfeld et al. 1996 (vgl. Anm. 16), S. 54.

32 Stern-Schema DB: Datenmodell, welches um einen zentralen Kern mehrere hierarchische Dimensionen als Sichten auf diese Daten anordnet.

Zusätzlich entsteht dann rasch die Erweiterung zum *Snowflake-Schema*[33], also der Bedarf nach Ober- und Unterklassen, die als Klassen gehandhabt werden und deren Beziehung in einer Klassencode-Klassencode-Tabelle abgelegt sind (wie: Berufsgruppe Bäcker als Unterklasse von Branche Lebensmittel).

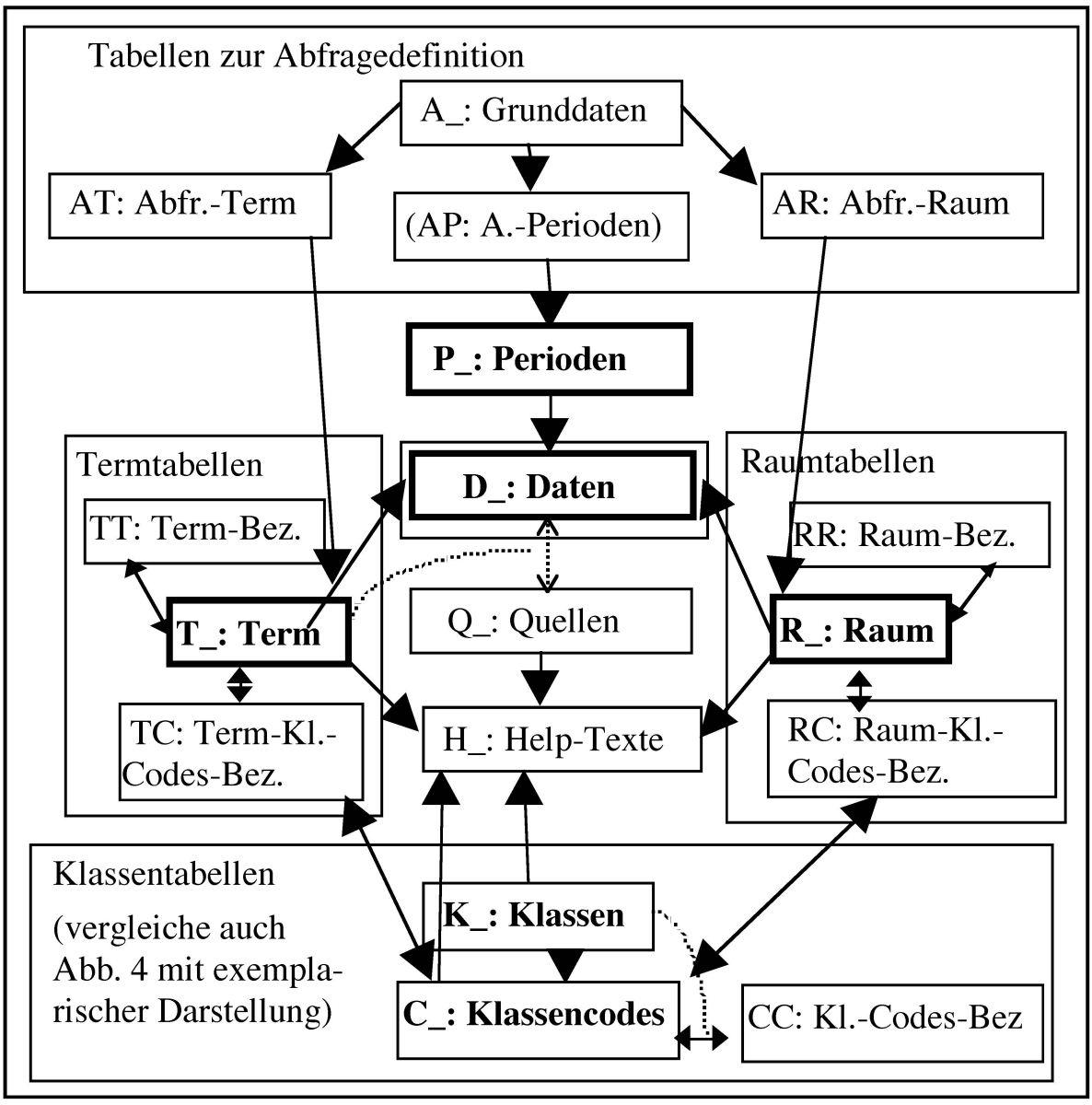
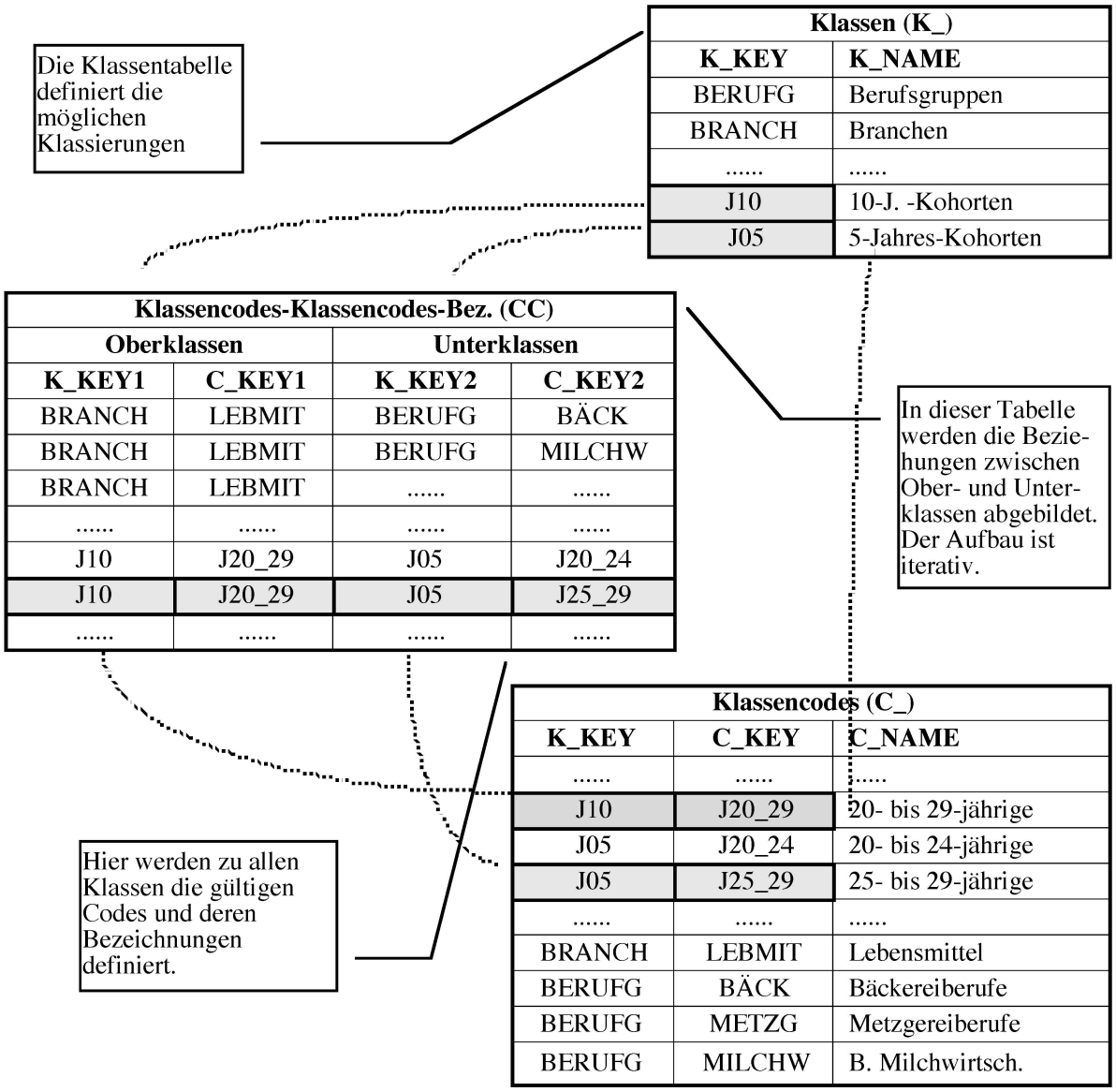


Abb. 3: Generisches Datenmodell des DWH

Die *Klassen bilden nun die zentrale Sicht auf die Daten.* Ein Term hat zwingend mindestens eine Beziehung zu einem Klassencode, ebenso jeder

33 Snowflake-Schema DB: Normalisiertes Stern-Schema Modell, welches auch Dimensionen als Sichten auf Dimensionen enthält.

Raum. Die Idee, nur einen Thesaurus zu führen, ist mit dieser Klassierung wieder entstanden. Raum- und Termthesauri sind noch pragmatische Hilfsmittel, um das Data Warehouse zu handhaben und BenutzerInnen ein einfaches Abfragetool anzubieten. Das logische Snowflake Schema des Datenmodells kann mit diesen wenigen Tabellen abgebildet und zu beliebiger Komplexität ausgebaut werden, ohne ein einziges neues Feld oder gar eine neue Tabelle einfügen zu müssen.



Klassen (K_)

K_KEY	K_NAME
BERUFG	Berufsgruppen
BRANCH	Branchen
.....	.....
J10	10-J. -Kohorten
J05	5-Jahres-Kohorten

Klassencodes-Klassencodes-Bez. (CC)

Oberklassen		Unterklassen	
K_KEY1	C_KEY1	K_KEY2	C_KEY2
BRANCH	LEBMIT	BERUFG	BÄCK
BRANCH	LEBMIT	BERUFG	MILCHW
BRANCH	LEBMIT	.....	.....
.....	.....	.....	.....
J10	J20_29	J05	J20_24
J10	J20_29	J05	J25_29
.....	.....	.....	.....

Klassencodes (C_)

K_KEY	C_KEY	C_NAME
.....	.....	.....
J10	J20_29	20- bis 29-jährige
J05	J20_24	20- bis 24-jährige
J05	J25_29	25- bis 29-jährige
.....	.....	.....
BRANCH	LEBMIT	Lebensmittel
BERUFG	BÄCK	Bäckereiberufe
BERUFG	METZG	Metzgereiberufe
BERUFG	MILCHW	B. Milchwirtsch.

Abb. 4: Zusammenwirken der Tabellen der Klassierung.
Zur Verbesserung der Performance können Views auf alle Ober- resp. Unterklassen aufgelöst und in einer Hilfstabelle abgelegt werden.

4.4. Volatile Raumstruktur

Eine politische Raumstruktur, auf welche sich üblicherweise die meisten obrigkeitlichen Erhebungen beziehen, ist über Zeit nie konstant. Ein Kernthema jeder raum-zeit-thematischen Datenbank muss auch der Umgang mit sich ändernden räumlichen Strukturen sein. Solange, wie in «Bernhist», Daten auf der Raumebene von Einwohnergemeinden in erster Linie zu statistischen Zwecken erhoben und öfters interpoliert oder umgerechnet werden, ist die Vereinheitlichung der Raumstruktur über den ganzen Untersuchungszeitraum ein gangbarer Weg, die Probleme in den Griff zu bekommen, welche durch Fusionen, Aufteilungen und oft auch durch das Verschieben von einzelnen Gemeindeteilen entstehen. Sobald aber räumliche Detaildaten zu präzisen Ratings herangezogen werden, kommt eine allgemeine Standardisierung der Raumstruktur nicht in Betracht.

Die Handhabung von Raumstrukturen im DWH basiert auf der Prämisse, dass der *Raum für alle Daten eines Zeitpunktes gleich* ist. Es kann also nicht sein, dass eine Gemeindefusion in der Arbeitslosenerhebung per 1. September und in der Baustatistik erst per 1. Dezember nachvollzogen wird.

Grundsätzlich wird in der Raumtabelle der Metadaten jeder Raumeinheit eine *Lebensdauer* zugeordnet. In der Raum-Beziehungstabelle werden neben den Beziehungen zwischen unterschiedlichen Raumtypen (etwa Gemeinden zu Bezirken) auch *Beziehungen zwischen gleichen Raumtypen mit einem Zeitrahmen* abgelegt. Ein Beispiel:

R_				**RR**				
Keys	**Name**	**von**	**bis**	**Key1**	**K2**	**Rela**	**von**	**bis**
R1	BRD	1949	1990	R1	R3	Teil	1949	1990
R2	DDR	1949	1990	R2	R3	Teil	1949	1990
R3	Deutschl. BRD&DDR	1949	>					

Abb 5: Einträge in der Raum-Tabelle (R_) und Raum-Beziehungstabelle (RR) für nicht durchgehende Raumeinheiten

Beziehungen aus Fusionen und Aufteilungen sind somit klar abgelegt. Ursprünglich haben wir eine ausgeklügelte, rekursive Methode entworfen, welche für alle Fälle das «kleinste gemeinsame Vielfache» bestimmt (d.h. das kleinstmögliche Konglomerat von Raumeinheiten) und die Summen

automatisch berechnet. Nicht mehr rechenbar ist dies bei Verhältniswerten. Zudem ist eine solche Raumvorgabe für wissenschaftliches Arbeiten gefährlich, oft irreführend, da auch die Art der Datenerhebung mit der Raumänderung ändert. Wir haben uns auf ein Verfahren zur *Raumerweiterung in der Abfrage* entschieden. Für nicht durchgehende Räume im gewählten Zeitraum, werden die *assoziierten Raumeinheiten* mitgewählt. Wird also DDR und der Zeitraum 1960-1995 gewählt, findet die Raumerweiterung die Räume BRD und Deutschland auf Knopfdruck. Auch diese Methode arbeitet rekursiv von unten nach oben und wieder nach unten, bis alle – auch aus mehrfach fusionierten und aufgeteilten Gebilden – zugehörigen Raumkeys gefunden worden sind. Die Raumerweiterung ist im Abfragetool auf der Raumauswahl plaziert.

4.5. Eine interaktive Oberfläche für ein Forschungs-Data Warehouse

Ein möglichst bequemes Abfragetool stand an oberster Stelle der Anforderungen für den ersten Release. Die bisher realisierte Oberfläche besteht denn auch weitgehend aus den acht ‹Tabs' für das Erstellen, Abspeichern, Holen, Ausführen und Exportieren von Abfragen. Entwickelt ist die ganze Oberfläche mit SAS/Frames und insbesondere unter Einsatz der SAS-6.12-Klassen «Tab Layout» (mit einzelnen Tabs) und «Organizational Chart» (Bäume, im folgenden Orgchart genannt) sowie mit SCL-Lists, SQL und DDE (für den Export in Excel).[34] Die gesamte Kommunikation zwischen den Objekten wird mit Messages sichergestellt. In diesem Bereich ist SAS ausserordentlich mächtig. Instanzierungen und generische Subklassen werden erst in einem späteren Release ein Thema sein. Als Beispiel für das Layout der Oberfläche zeigt Abbildung 6 den geöffneten Tab «Term auswählen»: Es soll einen Eindruck des touch-and-feel vermitteln.

Die Oberfläche des Data Warehouse als ganzes wird nächstens im Usability Labor des Ergonomic Dept. der EDS getestet: Potentiellen EndbenutzerInnen wird eine Aufgabe gestellt und das Handbuch neben die Tastatur gelegt. Über Video und durch eine Spiegelglasscheibe verfolgen EntwicklerInnen und Ergonomiefachleute die Erfolge und Misserfolge der Versuchspersonen. Ziel ist es, das User Interface der Software zu testen, nicht die User!

34 Nicht erwähnt sind die eher selbstverständlichen Objekte wie Pushbuttons, Labels, Eingabefelder, Radioboxes und viele mehr.

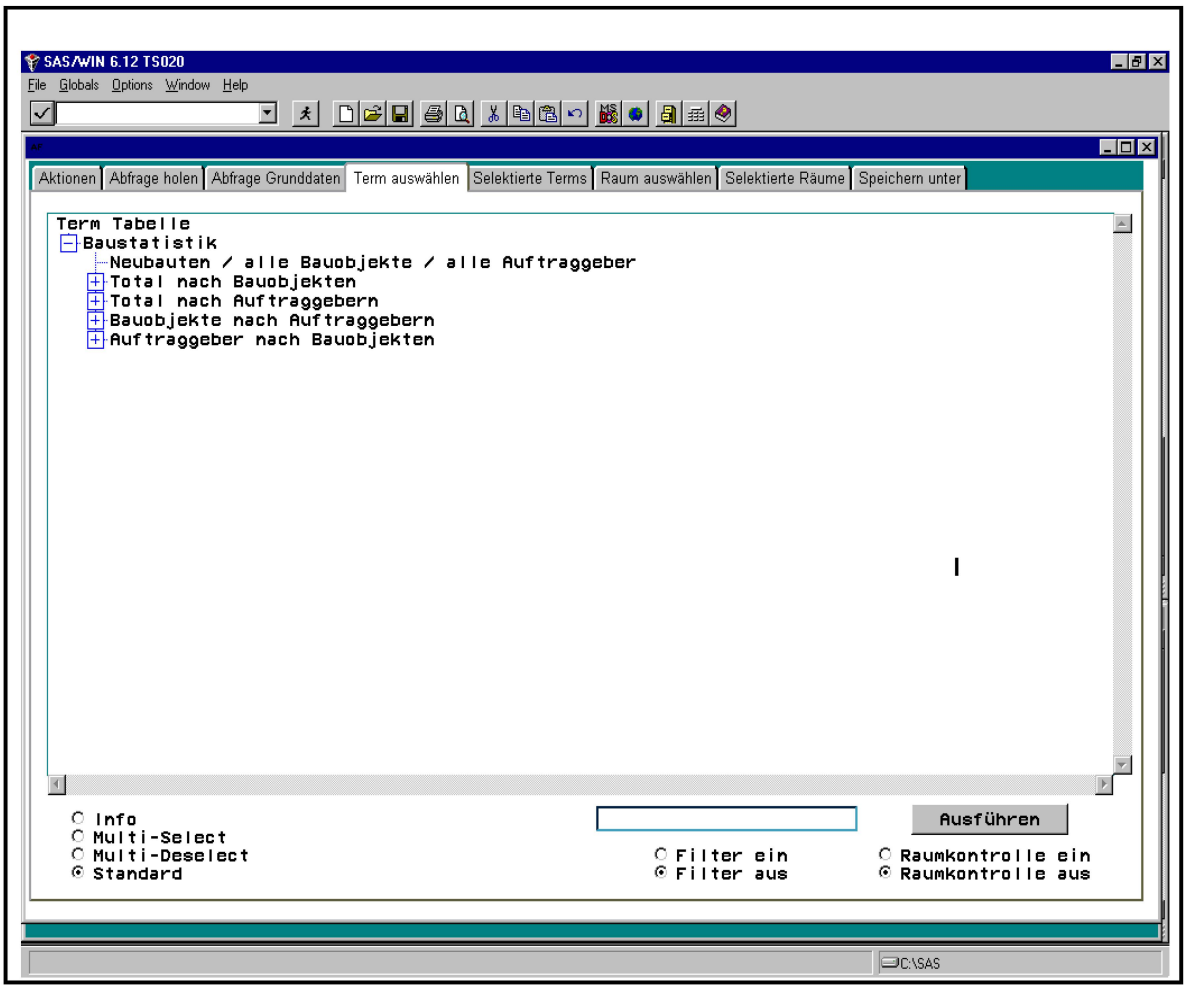


Abb. 6: Snapshoot mit geöffnetem Tab «Term auswählen»:
Tab mit Orgchart, Radioboxes, Eingabefeld und Pushbuttons
- Blättern im Term-Thesaurus: öffnen und schliessen von Unterbäumen
- Erfassen Filterstring, ein- und ausschalten der Filterfunktion
- Ein- und ausschalten der Raumkontrolle (zeigt nur Terme an, die zu den bereits gewählten Räumen Daten haben)
- Ausführen resp. Erneuern der Einschränkung auf Filtertext und der Raumkontrolle
- Auswahl einzelner Terme
- Multi-Select aller Terme direkt unter einem Knoten
- ebenso: Multi*DE*selekt aller Terme direkt unter einem Knoten;
- Farbliche Hervorhebung aller gewählter Terme überall wo diese vorkommen[35]
- Anzeigen des Infofensters zu einem Term oder einem Knoten

4.6. Weiterentwicklung: Methodenbibliothek für iterativen Klassenzugriff

Die schnelle, protozyklische Vorgehensweise birgt die Gefahr in sich, das Gesamtkonzept aus den Augen zu verlieren und Entscheide zu fällen, welche auf spätere Releases erheblichen Einfluss haben, ohne dies recht-

35 Ein Term kann mehreren Oberbegriffen zugeordnet werden: Werden «Arbeitslose BäckerInnen 25-30 J.» unter dem Oberbegriff «AL nach Altersklassen» angewählt, so muss der Term auch unter dem Oberbegriff «AL nach Berufen» als gewählt angezeigt werden (d.h. grey reverse).

zeitig zu erkennen. Um dieser Gefahr vorzubeugen, sollen die Anforderungen oder zumindest die Erwartungen an den nächsten und an den übernächsten Zyklus bereits formuliert oder wenigstens gesammelt werden. Vor dem Start eines Zyklus wird jeweils eine neue, aktuelle Priorisierung vorgenommen. Dieses Vorausschauen entfällt weitgehend, wenn grosse Pausen in der Entwicklung geplant sind, ja gar der nächste Zyklus nur noch der geordnete Abbruch sein wird, d.h. meist die Überführung der Datenlogik in eine neue oder eine andere Applikation.[36] Im Moment ist das beschriebene DWH-Projekt in einer Zwischenphase, in der sogenannte Value Propositions[37] beim Economic Research Team (und dessen Kunden) ausgearbeitet und geprüft werden und wir die Entwicklung verschiedener Instrumente zur Handhabung und optimalen Nutzung der Klassierung ausarbeiten.

Die Nutzung der Klassierung hat vier Aspekte:

- Datenzugriff, d.h. Abfrage nach Klassen und Klassencodes.
- Analyse der Daten nach Klassen. Dies schliesst das Definieren von Klassen als Hypothesen ein.
- Tool zur einfachen Definition von neuen Klassen, um Hypothesen abzubilden.
- Export von Abfrageergebnissen nach Klassen mit wählbarer Zuordnung zu (beliebig) vielen Dimensionen.

Für den *Datenzugriff nach Klassen* kann im Prinzip ein weiterer ‹Tab› angefügt werden, der vergleichbar mit Term- und Raumselektion gestaltet ist. Dabei besteht die Herausforderung im Aufbau eines dynamischen Hilfsinstruments für sinnvolle Klassenwahl. Auswahlen wie «nach 5-Jahres-Kohorten», «nach Bauobjekttyp» und «nach Deliktgruppe» in Verbindung mit dem Hauptbegriff «Einfuhr in Fr.» sollten verhindert werden. Längst nicht in allen Fällen ist Sinn und Unsinn so leicht erkennbar. Einfuhren

---

36 Eine jede Applikation hat einen Lebenszyklus, der mit dem ersten Grobkonzept beginnt und meist mit der geordneten Überführung von Datenbeständen, Datenlogik und Teilen der Programmlogik in eine neue oder eine andere, bestehende Applikation endet. Erfahrungsgemäss gewinnt die Qualität einer Applikation bereits zu «Lebzeiten» durch die konsequente Beachtung der «Bestattungsriten». In Forschungsprojekten sollte das Ende von Anfang an beachtet werden, etwa das wohldokumentierte Übergeben von Daten an ein Datenarchiv.

37 Mattison 1996 (vgl. Anm. 17), S. 115f, stellt das Konzept «value propositions» zur Priorisierung der (Weiter-)Entwicklung von DWH-Projekten vor. Dabei kann es sich um eine Erweiterung der Datenbestände, der DWH-Software, der Analysetools oder um eine weitere Analyse handeln. Er stellt drei Regeln auf:

1. «Each value proposition must be a specific business problem...», keine «laundry list».
2. «Each value proposition must have a single, responsible sponsoring bussiness organisation.»
3. «Each value proposition must define a specific, tangible benefit ... prefered financial, but less tangible as well (i.e. marked share, efficiency)»

Mattison schlägt nun vor, die einzelnen VPs zu priorisieren und zu gruppieren, um daraus weitere Entwicklungsschritte zu definieren. Während sich Zyklen überlappen können, dürfen sich diese niemals überholen.

nach Produkten bei mehreren parallelen Produkteklassierungen handzuhaben, bedürfen eines mächtigen Werkzeuges.

- Wird von einem Hauptbegriff ausgegangen, ist das Anzeigen aller Klassen und Oberklassen durchaus machbar und kann von BenutzerInnen leicht nachvollzogen werden.
- Von gewählten Klassierungen (und auch von gemachten Einschränkungen auf einzelne Klassencodes) sollte nach Hauptbegriffen und deren Klassierungen gesucht werden können. Eine durchaus sinnvolle Fragestellung könnte sein, alle Daten zu finden, welche über mindestens 5 Jahre hinweg mit mindestens vier Dimensionen mit der Zahl der Arbeitslosen verbunden werden können.
- Neben diesem Instrument an der Oberfläche für EndbenutzerInnen, müssen in Hintergrund genügend mächtige Methoden vorhanden sein, um eine Abfrage auch zuverlässig und möglichst schnell auszuführen. Viele Verdichtungen, sicher aber nicht alle, sind in der Datenbank schon gespeichert. Gewisse Zahlen haben Lücken und können somit nicht durchgehend mit anderen verglichen werden.

Für die *Analyse der Daten nach Klassen* (nachdem die Abfrage ausgeführt worden ist) lässt sich recht einfach ein Viewgenerator bauen. Dieser erstellt eine View, welche für die Klassen die Klassenkeys als Kolonnennamen (Variablen) mit den den Termen und Räumen zugeordnete Klassencodes als Werte ausgibt. Vorgängig kann eine Auswahlliste mit den möglichen Klassierung zum Anwählen angezeigt werden. Die daraus generierte View wird dann der Analyseprozedur «gefüttert».

Hypothesen können in Klassen oder Oberklassen abgebildet werden.[38] Dies erlaubt erstens, einen erheblichen Teil der Analysen zu standardisieren, und zweitens, ein Instrument zur Verfügung zu haben, das gemachte Analysen dokumentiert. Dazu brauchen die AnalystInnen ein *Tool zur Definition von Klassen*, wobei auch eine adhoc Klassierung denkbar ist. Damit könnte allerdings auch eine inflationäre Flut von Wünschen an das DWH ausgelöst werden. Insbesondere Regeln für klassen- und subjektübergreifende Filter oder solche für dynamische Klassierungen werden nachgefragt werden.[39]

---

38 Z.B. können die Gemeinden nach Bevölkerungsdaten typisiert werden, etwa durch eine Clusteranalyse. Die verschiedenen Typen werden als Klassen definiert und nach diesen Klassen werden dann beispielsweise Konsumdaten untersucht.

39 Zwei Beispiele von Fragestellungen, welche dynamische Klassierungen, Regelwerke und Filter nutzen könnten: Untersuchung der Arbeitslosigkeit in den Gemeinden nach der differenziert betrachteten Entwicklung der Bautätigkeit der vorangegangenen fünfzehn Jahre oder Strukturfragen wie Zusammenhänge zwischen Pendlerströmen, Bildungsangebot und lokaler Infrastruktur.

Für den dritten Aspekt, den *Export von Abfrageergebnissen nach Klassen*, werden wir die gleiche Auswahlliste wie oben anzeigen. Daraus können dann die erste, zweite, dritte, vierte und so weiter Dimension ausgewählt werden. Nicht gewählte Dimensionen werden verdichtet. Der Export der Resultate kann einfacher in HTML-Format gemacht werden, als direkt (nur) in Excel.

Der Zugriff über die Klassen wurde noch nicht zur Produktionsreife entwickelt, doch die Richtung der Entwicklung ist klar. Einzelne AnalystInnen nutzen die Klassen bereits intensiv.

Andere Ansprüche an einen weiteren Release sind Zugriffseinschränkungen, breitere Streuung ausgewählter Daten, Einsatz von Schedulern und Auslagern von bestimmten Datenbeständen mit niedriger Granularität. Zudem sollen verschiedene Tools zur Administration, Analyse und Publikation von SAS geprüft werden. Dazu gehören SAS/WH-Administrator, SAS/Mining-Cockpit, SAS/Insight, SAS/IntrNet und SAS/Access zum Einsatz eines RDBMS zur Datenhaltung.

Eine Erweiterung des Datenmodells wird zur Zeit diskutiert: Eine zweite *Raumdimension* drängt sich für eine Vielzahl von Subjekten auf. Dazu gehören Ein- und Ausfuhren, Verkehrs- und Güterströme, Migration oder Zu- und Wegpendler.

5. Ein Forschungs-Data Warehouse für HistorikerInnen

Hier will ich der Frage nachgehen, inwieweit sich das dargestellte Data Warehouse für die historische Forschung adaptieren lässt, und klären, worin denn der Fortschritt der Informationstechnologie der letzten 10 Jahre liegt, den sich HistorikerInnen, welche mit quantitativen Methoden arbeiten wollen, zu Nutze machen können.

Für *Datenbestände*, die bei Projekten wie «Bernhist» anfallen, für Daten, wie sie von der Forschungsstelle für Schweizerische Sozial- und Wirtschaftsgeschichte der Universität Zürich gesammelt werden, ja für alle Daten, welche in irgendeiner Weise *bereits Summen* (Einwohner, Flächen, Güter in Tonnen etc.) darstellen, ist das hier skizzierte DWH-Konzept gut geeignet:

- Zur Haltung und Pflege von Daten in der beschriebenen Struktur
- Zur problemlosen Integration neuer Daten, seien dies Fortschreibungen bestehender Zeitreihen (etwa aus Publikationen des BFS), neue Themenbereiche oder räumliche Erweiterungen
- Zur Verwaltung eines Anmerkungsapparates
- Zur Analyse nach beliebigen, erweiterbaren Dimensionen (Klassen)

– Zum einfachen Datenzugriff über eine interaktive Oberfläche
– Und letztlich zur möglichen Überführung der Daten in ein zentrales Archiv für historische Daten.

Voraussetzung für das Gelingen solcher Projekte ist neben dem Datenkonzept die *Administration* des DWH und die *betriebsorganisatorische* Einbindung. Besonders in Forschungsprojekten muss diesen Aspekten wegen der begrenzten Projektdauer und der zeitlich limitierten Anstellung von MitarbeiterInnen hohe Beachtung geschenkt werden. Zudem möchten sich die ProfessorInnen ja mehr mit wissenschaftlichen Fragen als mit der Projektleitung beschäftigen. Doch gerade ein DWH kann sehr viele organisatorische Probleme lösen und die ForscherInnen entsprechend entlasten: Einheitliche Datenbasis für alle Analysen für alle MitarbeiterInnen, raschere Einbindung neuester Daten, weniger Reibungsverluste für Datenabstimmung und -beschaffung.

Für Daten auf der Ebene von Einzelpersonen oder -beobachtungen kann das Datenmodell zwar nicht direkt übernommen werden, doch lässt sich ein angepasstes Modell entwickeln und in eine vergleichbare Oberfläche einbinden. Das Ziel muss immer sein, nichts an ursprünglicher Information zu verlieren, die Überführung der Quellendaten in entsprechende Datenbankeinträge zu definieren und eine systematische Beschreibung (Metadaten) zu pflegen. «Euro-Climhist» kann als Beispiel einer durch Prozesse und Metadaten gesteuerten Forschungs-Datenbank (noch kein DWH) angesehen werden: Owner, Quellen und Timestamp lassen jeden Eintrag auf die ursprüngliche Quelle zurückführen. Andererseits lassen sich themen- und raumorientierte Extrakte zum Analysieren oder zum Ausgeben als Text, Grafik oder Karte bilden.[40] Was «Euro-Climhist» noch fehlt, ist ein modernes Eingabetool für historische Daten (etwa nach Breure[41]).

Der wichtigste Schritt der IT für HistorikerInnen ist wohl die einfache *Möglichkeit zur Gestaltung von ergonomischen Oberflächen*. Dass die Computer leistungsfähiger geworden sind und dadurch auch neue Datenbankmodelle und Analysemethoden einsetzbar sind, darf als zusätzlicher Pluspunkt angesehen werden. Einige Projekte der 80er und frühen 90er

40 Vergleiche Schüle, Hannes: «Coding Climate Proxy Information for the EURO-CLIMHIST database». In: *European climate reconstructed from documentary data: methods and results. Paleoclimate Research. Special Issue ESF Project «European Palaeoclimate and Man» 2*, hrsg. von B. Frenzel; Ch. Pfister; B. Gläser. Stuttgart, New York 1992, S. 211-218. Oder: Schwarz-Zanetti, Gabriela; Schwarz-Zanetti, Werner; Schüle, Hannes: «Berner Datenbank für Klimageschichte. Das historische Wetter Europas aus dem Computer». In: *Angewandte Geographische Informationstechnologie III, Salzburger Geographische Materialien. Heft 16*, hrsg. von Dollinger, F. und Strobl, J. Salzburg 1991, S. 231-241.

41 Breure, Leen: «Interactive Data Entry: Problems, Models, Solutions». In: *History and Computing Vol 7, No. 1*, hrsg. von S. W. Baskerville und R. J. Morris. Edinburgh 1995, S. 30-49.

Jahre in verschiedenen Ländern Europas, welche im weitesten Sinne als Historisch-Geographische Informationssysteme angesehen werden können, haben die zur Verfügung stehenden Technologien und Mittel optimal ausgereizt. Die umgesetzten Konzepte gingen oft an die Grenzen des technologisch gerade noch Machbaren. Es darf nicht vergessen werden, dass die Sponsoren – meist nationale Forschungsförderungsfonds – nicht eine Datenbank, sondern Forschungsresultate erwarten. Die Entstehung von wieder- und weiterverwendbaren Datenbanken ist quasi ein Nebenprodukt der Wissenschaft.

Für Forschungsprojekte erachte ich es als essentiell, *der systematischen Datenhaltung genügend Aufmerksamkeit zu schenken*. Für einzelne, isolierte Projekte sind kleinere, aber gut strukturierte Datenbanken[42] durchaus sinnvoll. Sobald Projekte mehrere MitarbeiterInnen umfassen, länger als ein halbes Jahr dauern oder mehrere Artikel (oder eine Monographie) daraus geschrieben werden, ist der Ansatz eines raum-zeit-thematischen DWH mit klar zugewiesenen Aufgaben («Rollen» wie DWH-AdministratorIn) eine erhebliche Arbeitserleichterung. SAS, zur Analyse von Daten in Universitäten weit verbreitet, bietet auch hervorragende, skalierbare Werkzeuge zur Entwicklung, zum Betrieb und zum Unterhalt eines Forschungs-Data Warehouse, welches mit den sich verändernden Aufgabenstellungen wachsen kann.

42 Greenstein, Daniel I.: *A Historian's Guide to Computing*. Oxford 1994, S. 61-157.