Clio - ein datenbankorientiertes System für die historischen Wissenschaften: Fortschreibungsbericht

Thaller, Manfred

CLIO

EIN DATENBANKORIENTIERTES SYSTEM FÜR DIE HISTORISCHEN WISSENSCHAFTEN: FORTSCHREIBUNGSBERICHT

Manfred Thaller (*)

Abstract: *Since 1978 the Max-Planck-Institut für Geschichte is engaged in the development of data base oriented software, tailored specifically to applications in historical research. The system has been designed to be userfriendly and applicable to realistic amounts of data, instead of small sets of test material. It offers: complex input formats, information retrieval, report generation, the preparation of statistical data sets, nominative record linkage, complex merging operations and full-text retrieval capabilities. As a consequence of these developments, the need for a specific data model for historical applications of data processing is postulated.*

Einleitung

Am Max-Planck-Institut für Geschichte in Göttingen laufen seit 1978 Arbeiten zur Entwicklung eines datenbankorientierten Programmsystems, das speziell auf die Bedürfnisse der historischen Forschung abgestellt ist. In einer ersten Phase (1978-1983) wurde eine Version dieses Systems (bekannt als CLIO) fertiggestellt, die an Rechner des Herstellers Sperry (früher UNIVAC) gebunden ist. In den beiden folgenden Jahren wurden umfangreiche Arbeiten zur Implementation zusätzlicher Funktionen in einzelnen Versuchssystemen unternommen (Volltextretrieval, Patternmatching u.a.). Aufbauend auf diesen Vorarbeiten läuft derzeit die Re-Implementation eines erheblich erweiterten Systems (Clio/C) in der Programmiersprache C. Diese Version strebt größtmögliche Portabilität an und wird ab Februar 1987 für eine Reihe von PC's und Mainframes in einer Testversion mit vorläufig freilich kleinerem Funktionsumfang zur Verfügung stehen.

I. Allgemeine Charakteristika

Der Entschluß, über mehrere Jahre hinweg in die Entwicklung von Software zu investieren, hat eine ganze Reihe von Gründen. Am entscheidendsten ist wohl die Auffassung, daß Informationen, die in historischen Quellen enthalten sind, Eigenschaften haben, die die Anwendung herkömmlicher Software mindestens behindern, und in letzter Konsequenz die Erarbeitung eigener fachspezifischer Datenmodelle erfordern.

Neben diesen weiter unten noch erläuterten Überlegungen sind folgende allgemeinen Vorstellungen für die Erarbeitung der Programme maßgeblich: Die Programme sind grundsätzlich objektorientiert, d.h. semantisch aussagekräftige Datennamen. Eine

C) Address ail communications to: Manfred Thaller, Max-Planck-Institut für Geschichte, Hermann-Fôge-Weg 11, D-3400 Gôttingen.

Kenntnis der vom System verwendeten prozeduralen Schritte ist nicht notwendig, ebensowenig wie Erfahrungen mit der Zerlegung von Strings in sinnvolle Einheiten. Generell verwaltet das System - streng unabhängig voneinander - Daten, die als "Texte" im weitesten Sinne aus historischen Quellenmaterial entnommen werden und, in einem System miteinander verbundener Data Dictionaries, Hintergrundwissen über diese Texte. Änderungen im Hintergrundwissen - etwa neuere Erkenntnisse über die Austauschrelationen zwischen verschiedenen Währungen - erfordern keine Modifikation der Daten. Das System ist auf die Produktion von real verwertbaren Ergebnissen abgestellt:die größten in Verwendung befindlichen Datenbanken geben etwa 500.000 bis 750.000 Eingabezeilen wieder. Die bisherigen Versionen sind jedoch in einer Umwelt entstanden, in der Plattenspeicherplatz relativ leicht zugänglich war, daher nicht in erster Linie auf dessen sparsame Verwendung optimiert, so daß das System ziemlich speicherintensiv ist.

II. Leistungskatalog

1. Dateneingabe und Struktur

Daten werden als durch reservierte Sonderzeichen gegliederte Ansammlungen von kürzeren Textpartien eingegeben. In der derzeitigen Version des Systems werden sie als ein Netzwerk interpretiert, wobei hierarchische Zusammenhänge durch die Eingabekonventionen besonders unterstützt werden. Das System legt besonderen Wert auf Flexibilität der Datenstruktur; der Benutzer gibt Kommandos, die abstrakte Aufgaben definieren, und verwendet Quellennähe der Eingabe und Vermeidung von Scheinpräzision. Begrenzungen sind, soweit sie existieren, üblicherweise sehr weit ausgelegt: z.B. kann ein Datenfeld, ohne Verlust an Effizienz, zwischen Null und etwa 80.000 Zeichen variieren. Beliebige Schwankungen in der Feldlänge und/oder die häufige Abwesenheit von "Variablen" oder ganzen "Entitäten" beeinflussen das Systemverhalten nicht. Beide Termini sind allerdings nur mit einigen Modifikationen verwendbar: im Interesse der Flexibilität kennt das System beispielsweise keine skalaren Variablen, sondern nur Vektoren logisch gleichrangiger Werte - die freilich in vielen Fällen nur durch einen Wert besetzt sind. Im Interesse der Quellennähe stehen Hilfsmittel bereit, um die Daten möglichst in der Form des Originals abschreiben zu können: z.B. für eine Reihe von Kalendernotationen, wie etwa die lateinische, die Verwendung von nicht-dezimalen Maßsystemen und die Verarbeitung nicht-standardisierten Namengutes. Um Scheinpräzisionen zu vermeiden, besteht die Möglichkeit, Operatoren wie "circa" sowohl bei der Eingabe der Daten als auch während ihrer Bearbeitung zu verwenden.

2. Information Retrieval

Eine - wie bereits betont nichtprozedurale - Abfragesprache steht zur Verfügung. Sie erlaubt die Auswahl einer Teilmenge der Daten durch entsprechende Operatoren und ihre Weiterverarbeitung als einfachen Ausdruck oder aufbereitet durch eine Familie von Postprozessoren, die z.B. die Erstellung von Registern unterstützen. Dabei besteht grundsätzlich die Möglichkeit, Daten aus beliebigen Teilen der Datenstruktur miteinander zu kombinieren, solange eine eindeutige logische Zuordnung der einzelnen Angaben möglich ist.

3. Vorbereitung statistischer Auswertungen

Ein Zweig des Systems ermöglicht die Aufbereitung von quellennah vorbereiteten Materialien zur statistischen Analyse durch die handelsüblichen Statistiksysteme. Dieser Zweig des Systems - der allerdings auch das Fehlen eigener Statistikfunktionen abdecken muß - erbringt zwei Leistungen: einerseits die Übertragung der verschiedenen Textkategorien (z.B.: Berufe, Kalenderdaten, Preise) in für die herkömmlichen Statistiksysteme geeigneten Kodes, andererseits die Projektion der hierar-

chischen und/oder Netzwerkbeziehungen auf eine rechteckige Datenmatrix, wie sie von der Statistiksoftware nach wie vor gefordert wird. Bei der Kodierung kann der Benutzer sowohl Kodebücher als auch Systeme von Regeln für die Überführung der Texte in numerische Kodes angeben. Die Umstrukturierung der Daten geschieht nach einem von zwei möglichen Modellen ohne Eingreifen des Benutzers aus den Implikationen der in einem data dictionary gespeicherten logischen Struktur des Datenmaterials.

4. Nominative Record Linkage

Das System stellt einen organisatorischen Rahmen für das Nominative Record Linkage, also den systematischen Abgleich zweier Populationen in zwei Quellen auf Grund der vorkommenden Namen, bereit. Zur Überwindung der Schreibungsunterschiede stehen allgemein gefaßte Implementationen von drei Algorithmen bereit; vor ihrer Verwendung wird vom Benutzer erwartet, daß er Angaben über besondere in einem Quellenbestand herrschende Sprach-/Schreibungsverhältnisse in tabellarischer Form vorbereitet (ob z.B. bei der Kombination "1" und "m" voranstehendes "1" häufig wegfällt). Abgesehen von diesem technisch sehr gut lösbaren Teilproblem verfahren die Programme nach folgender grundsätzlicher Logik:

a. Ein Benutzer beginnt mit zwei Datenbanken, die zwei Quellen repräsentieren.
b. Aus beiden Datenbanken werden Dateien abgeleitet, die die identifikationsrelevanten Informationen enthalten.
c. Die Fälle in diesen Dateien (meist Personen) werden auf Grund von vom Benutzer spezifizierten Regeln verglichen.
d. Eine Liste von Vorschlägen, wer mit wem identisch sein können, wird ausgedruckt.
e. Der Benutzer erzeugt eine Datei mit Identifikationsnummern der akzeptablen Vorschläge.
f. Der Rechner entfernt auf Grund dieser Datei aus den im Schritt "b" erzeugten Dateien alle schon verknüpften Fälle.
g. Der Benutzer modifiziert seine Regeln und wiederholt die Schritte "c" bis "g" so oft, bis die Zahl der verbleibenden Fälle so klein wird, daß eine manuelle Identifikation einfacher wird als die Spezifikation zusätzlicher Regeln.
h. Die beiden Ausgangsdatenbanken werden miteinander verbunden, wobei die Dateien mit den akzeptierten Vorschlägen abermals Verwendung finden.

Diese Vorgangsweise hat sich in einigen Fällen sehr gut bewährt und arbeitete meist zufriedenstellend. Drei Problembereiche ergaben sich jedoch, die zu einem teilweisen Neudesign in der derzeitigen Re-Implementation anregen:

- Da das System keine Beschränkungen für die Anzahl auszudruckender Informationen über die zu verknüpfenden Fälle kennt, tendierten viele Benutzer zur Produktion immer umfangreicherer Vorschlaglisten, die nicht nur Papier verschwendeten, sondern vor allem auch ihre Bearbeitung verlangsamten.
- Nahezu alle Projekte fanden etwa 80% aller theoretisch möglichen Identifikationen in sehr kurzer Zeit. Es ist sicher sinnvoll, daß alle Projekte denselben Aufwand nochmals betrieben, um diesen Anteil auf 90% zu erhöhen. Nahezu alle Projekte fielen danach jedoch der Versuchung anheim, sehr lange Zeit "noch einen allerletzten Regelsatz" auszuprobieren um - mit rapide abnehmendem Erfolg - einige wenige Identifikationen zu finden.
- Der gravierendste Mangel war, daß in vielen Fällen Forscher in einem späteren Durchgang bereits auf Informationen zugreifen wollten, die auf Grund früher stattgefundener Identifikationen im Prinzip bereits verfügbar waren, da die Datenbanken jedoch erst am Schluß tatsächlich kombiniert wurden, noch nicht bereit standen.

Diese Erfahrungen führten zu folgenden Modifikationen im Design der Re-Implementation:

- Der Benutzer soll in Zukunft die Wahl haben zwischen der interaktiven Inspektion (mit Hilfe eines gesplitteten Bildschirms) der zu Grunde liegenden Systemdateien mit Hilfe der Standardretrievalbefehle und den bisherigen Vorschlagslisten.
- Der Benutzer soll die Möglichkeit haben, die beiden Datenbanken sofort interaktiv zu vermischen.

5. Verbindung von Dateien

Es besteht die Möglichkeit, beliebige Teilmengen aus einer Datenbank in beliebige Teile des Inhaltes einer anderen zu integrieren. Da dabei versucht wird, eine ganze Reihe von inhaltlichen Überlegungen nach vom Benutzer nur grob vorzugebenden Regeln stillschweigend berücksichtigen zu lassen, ist dies derzeit sicher der unzuverlässigste und komplexeste Systemteil. Die dabei auftretenden Probleme werden am besten durch ein Beispiel illustriert: Wenn zwei Nennungen in zwei Quellen sich sicher auf ein und dieselbe Person beziehen, sich die Geburtsdaten in beiden Nennungen jedoch unterscheiden: Sollen sie als Varianten beibehalten oder zu einem terminus post quem - ante quem Ausdruck verbunden werden? Oder gibt es Kriterien, nach denen eine der beiden Datierungen sicher vorzuziehen ist?

6. Volltextverarbeitung

Die ältere Version des Systems stellt eine Reihe von Generatoren für KWIC-Indices und andere Typen von Wortlisten zur Verfügung. In der letzten Zeit waren Versuche mit zusätzlichen Komponenten, die eine Volltextdatenbank darstellen, jedoch so erfolgreich, daß bei der Re-Implementation die Produktion von Wortlisten nun mehr eine sehr geringe Priorität hat, während bereits die erste Testversion Eigenschaften strukturierter Datenbanksoftware und solche der interaktiven Volltextverarbeitung integrieren wird.

III. Methodische Ziele

Bereits einleitend wurde gesagt, daß eine der wesentlichen Legitimationen für die skizzierten Entwicklungen in der Vorstellung liegt, daß es so etwas wie für die Datenverarbeitung relevante Eigenschaften historischer Quellenmaterials gäbe, die in letzter Konsequenz die Erarbeitung neuer informationswissenschaftlicher Konzepte und Lösungswege nötig machen. Die hier anstehenden Probleme werden vor allem in der grundsätzlichen Kontextsensitivität der Informationen gesehen: "Preußen", als Herkunftsregion in einer Quelle des Jahres 1740, hat eine gänzlich andere Bedeutung als dieselbe Zeichenkette in einer Quelle des Jahres 1820. Ähnlich schwankt die Interpretation der sozialen Bedeutung von Berufsbezeichnungen in Abhängigkeit von der lokalen Sozialstruktur; und auch bei zeitlichen Angaben ist die Frage, wie groß eine Abweichung bei zwei Daten sein darf, um sie noch sinnvollerweise als gleich ansehen zu können, nur durch die Bewertung aller anderen gleichzeitigen Datierungen möglich. Diese Situation spricht nach Ansicht des Verfassers eindeutig gegen das extrem kontextfreie relationale Datenmodell: die Re-Implementation baut daher bewußt auf einem eigens entwickelten Datenmodell auf, das die Daten zunächst als ein semantisch bestimmtes Netzwerk verwaltet, das von einzelnen Systemteilen dann, je nach Bedarf, als ein Netz im herkömmlichen Modellsinn verstanden werden kann, auf das andere Systemteile bei Bedarf aber auch mit einem relationalen Ansatz zugreifen können.