Vorüberlegungen für einen internationalen Workshop über die Schaffung, Verbindung und Nutzung großer interdisziplinärer Quellenbanken in den historischen Wissenschaften [1986]

Thaller, Manfred

Historical Social Research
Historische Sozialforschung

Manfred Thaller:

Vorüberlegungen für einen internationalen Workshop über die Schaffung, Verbindung und Nutzung großer interdisziplinärer Quellenbanken in den historischen Wissenschaften [1986]

doi: 10.12759/hsr.suppl.29.2017.160-177

Published in:

Historical Social Research Supplement 29 (2017)

Cite as:

Manfred Thaller. 2017. Vorüberlegungen für einen internationalen Workshop über die Schaffung, Verbindung und Nutzung großer interdisziplinärer Quellenbanken in den historischen Wissenschaften [1986]. *Historical Social Research* Supplement 29: 160-177. doi: 10.12759/hsr.suppl.29.2017. 160-177.

Historical Social Research
Historische Sozialforschung

Other articles published in this Supplement:

Vorüberlegungen für einen internationalen Workshop über die Schaffung, Verbindung und Nutzung großer interdisziplinärer Quellenbanken in den historischen Wissenschaften [1986]

Manfred Thaller[*]

Abstract: *»Preparatory Thoughts for an International Workshop on the Creation, Connection and Usage of Interdisciplinary Source Banks in the Historical Disciplines«.* The great effort necessary to make historical sources machine readable is a bottle neck for the application of computational methods in history. This makes their reuse for secondary analysis very important; at the same time electronic type setting as well as newly emerging OCR promise to make many more texts available for analysis than so far. As the variety of machine readable data is greater in the historical disciplines than in sociology, the data archives of sociology are only partially useful as model. At least four activities are needed to change the situation: an understanding of the data formats actually used, as well as a drive for standardizing them, organizing models for the preservation of these data for the long run and explicit training of younger researchers for the possibilities of secondary analysis. An agenda for an international workshop to address these problems is derived.

Keywords: Secondary analysis, standardization, long term preservation.

Der Gedanke, durch die Wiederverwendung einmal maschinenlesbar gemachter Daten im Rahmen anderer Projekte Ressourcen zu sparen, ist kaum jünger als die Heranziehung der Datenverarbeitung im geistes- und sozialwissenschaftlichen Bereich überhaupt. Eine große Anzahl von Projekten aller Disziplinen hat immer wieder darauf hingewiesen, daß ihre Bestrebungen auch und insbesonders unter dem Gesichtspunkt zu bewerten seien, daß durch sie exemplarische Daten für weit über die unmittelbaren Ziele hinausgehende Forschungsmöglichkeiten dem jeweiligen Fach zur Verfügung stünden. In den Sozialwissenschaften wurden aus derartigen Überlegungen eigene Einrichtungen der Infrastruktur der jeweiligen Fächer geschaffen, deren Ziel explizit die Archivierung und Aufbereitung von maschinenlesbaren Daten zur sekundären Analyse ist. In den Geisteswissenschaften (worunter wir hier und im Rest dieser Skizze die Literaturwissenschaften und die historischen

[*] Reprint of: Manfred Thaller. 1986. 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,* ed. Manfred Thaller (= HSF 20), 9-30. St. Katharinen: Scripta Mercaturae.

Historical Social Research Supplement 29 (2017), 160-177 | © GESIS
DOI: 10.12759/hsr.suppl.29.2017.160-177

Disziplinen verstehen) wurde diese Institutionalisierung trotz einzelner Ansätze kaum je erreicht, es gibt aber eine ganze Reihe von Einrichtungen, die auch hier die Archivierung und Bereitstellung maschinenlesbarer Daten wenn nicht als den zentralen, so doch als einen wesentlichen Bestandteil ihrer Funktionen ansehen.

Trotz dieser langen Tradition sekundäranalytischer Ansätze wird man die Bilanz tatsächlich durchgeführter sekundäranalytischer Arbeiten aber in wesentlichen Punkten als unbefriedigend empfinden müssen. Einzelne Forscher, die unter beträchtlichen Kosten und hohem persönlichen Einsatz spezialisierte Datenbanken bereitgestellt haben, klagten in aller Öffentlichkeit, daß die von ihnen geleisteten Vorarbeiten nur selten, wenn überhaupt genützt würden. Insbesonders im Bereich der historischen Wissenschaften überwiegt die Zahl zur Weiterverwendung angebotener Daten die dazu herangezogenen derzeit bei weitem.

Daß wir hier von einer in den mit Texten befaßten Disziplinen wesentlich unbefriedigenderen Bilanz sprechen, ist sicher nicht zufällig: einerseits ist selbst ein zusammenhängender Text von 100.000 Zeilen nur für einen unverhältnismäßig kleineren Teil der historischen oder literaturwissenschaftlichen Forschung von Interesse, als es eine in Form numerischer Kodes vorliegende Datei gleicher Größe in den Sozialwissenschaften wäre. Andererseits variieren die Konventionen, nach denen Texte zur maschinellen Analyse aufbereitet werden, wesentlich stärker als die Eingabeformen statistisch orientierter Arbeiten, und hinzu kommt noch, daß die Analyse eines Textes von 100.000 Zeilen für die Ressourcen der meisten akademischen Rechenzentren bereits eine so große Belastung bedeutet, daß es kaum möglich wäre, mehrere andere Korpora gleicher Größe neben dem primär interessierenden nur zur Gewinnung von Vergleichswerten geringerer Signifikanz verarbeiten zu lassen. Demgegenüber bedeutet es bei der Geschwindigkeit derzeitiger Großrechner wenig, neben einer 2000er Stichprobe noch einige andere als Kontrolldateien zu verarbeiten, selbst wenn die dabei gewonnenen Ergebnisse gar nicht in die am Ende veröffentlichten Resultate eingehen sollten.

Im Rahmen der am Max-Planck-Institut für Geschichte laufenden Entwicklungsarbeiten an spezifisch auf die historische Forschung abgestimmten Softwaresystemen besteht der Eindruck, daß diese Situation sich vor allem für die historischen Wissenschaften im nächsten Jahrzehnt wegen einer ganzen Reihe bereits absehbarer technologischer Entwicklungen entscheidend ändern wird.

Grundsätzlich ist davon auszugehen, daß innerhalb weniger Jahre maschinenlesbare Quellen in ganz anderen Größenordnungen zur Verfügung stehen werden, als dies heute der Fall ist, und diese obendrein sehr viel effizienter bearbeitbar sein werden. Diese Entwicklung kann jedoch nur dann voll und fruchtbar genutzt werden, wenn rechtzeitig allgemeine infrastrukturelle Maßnahmen getroffen werden. Unter „maschinenlesbaren Quellen“ verstellen wir alle maschinellen Darstellungsformen einer quellenorientierten Datenverarbeitung. Also sowohl Daten, die signifikante Textpartien zusammen mit einer großen Anzahl kodierter Angaben enthalten – wie sie in den Datenbanken der im weitesten Sinne demographischen Forschung vorliegen –, als auch solche, die fast nur Texte, diese aber in sehr stark strukturierten kleinen Einheiten umfassen – wie sie in den Datenbasen mikroanalytisch ausgerichteter Forschungsprojekte existieren –, und schließlich auch wenig kodierte Repräsentationen fortlaufender maschinenlesbarer Texte, wie sie bei computerunterstützten Editionstechniken entstehen. Unter einer Quellenbank verstehen

wir eine Institution, die große Korpora derartiger Materialien unterschiedlichster Provenienz in integrierter Form so verwaltet, daß sie als Entitäten zwar faßbar bleiben, die gleichzeitige Bearbeitung in völlig verschiedenen Kontexten entstandener Quellen aber möglich wird.

In der derzeitigen Situation der Forschungsfinanzierung wäre es müßig, von der Gewinnung zusätzlicher Mittel für eine quantitative Ausweitung der vorhandenen Infrastruktur der historischen Forschung auszugehen. Es ist daher zu überlegen, wieweit eine verbesserte Zusammenarbeit bestehender infrastruktureller Einrichtungen und die gezielte Abstimmung einschlägiger Bemühungen den sich anbahnenden Entwicklungen gerecht werden kann. Da die beschriebene Situation keineswegs auf ein Land beschränkt ist und die im folgenden angerissenen Probleme zum Teil überhaupt nur im internationalen Rahmen effizient lösbar sind, wären nationale Bemühungen von beschränktem Wert. Wie bereits betont scheinen uns die absehbaren Entwicklungen insbesonders für die datentechnisch arbeitenden Forschungsrichtungen innerhalb der Geschichtswissenschaften von Bedeutung: da die Lage in den Literaturwissenschaften jedoch sehr ähnlich und in deren historischen Zweigen nahezu mit der der historischen Disziplinen identisch ist, ist ihre Heranziehung für alle einschlägigen Überlegungen notwendig. Die längere Erfahrung und die wesentlich bessere einschlägige Infrastruktur der Sozialwissenschaften legt auch deren Einbeziehung nahe.

Als Bestandteil der Infrastruktur der Geschichtswissenschaft in der Bundesrepublik Deutschland schlug das Max-Planck-Institut für Geschichte daher einen internationalen Workshop von Vertretern einschlägiger infrastruktureller Einrichtungen sowie auf die Erzeugung großer Datenbanken ausgerichteter Forschungsprojekte vor. Das Institut erklärte sich gleichzeitig bereit, die Organisation einer derartigen Veranstaltung im Jahre 1985 zu übernehmen.[1]

Ziel dieser Veranstaltung sollte die Erarbeitung klarer Konzepte für mögliche Formen langfristiger und weitreichender Kooperation bei der Ausnützung der durch die Datentechnik gebotenen Möglichkeiten sein, wobei diese Konzepte so konkret zu halten waren, daß ihre Implementation im Rahmen lokal laufender Entwicklungen möglich wird. Wir möchten betonen, daß wir eine derartige Veranstaltung trotz der bewußt sehr weitgespannten Zielvorstellungen als für die unmittelbare Arbeit bedeutungsvoll ansahen. Um die einzurichtenden Arbeitsgespräche von vorneherein straff strukturieren zu können legten wir zunächst ein erstes Grundsatzpapier vor, das unsere Sicht der angesprochenen Themen darlegt und gleichzeitig als erstes Modell für die Strukturierung der vorgeschlagenen Arbeitsgespräche dienen sollte.

Wir gingen von folgenden Themenkreisen aus:

1) Welche Zielvorstellungen für eine optimale datentechnische Infrastruktur der historischen Disziplinen lassen sich aus den derzeit absehbaren technischen Entwicklungen ableiten?
2) Mit welchen organisatorischen Maßnahmen kann eine Annäherung an derartige Idealvorstellungen mit dem geringsten möglichen Aufwand erreicht werden?

[1] Die Kosten wurden zum größten Teil von der Stiftung Volkswagenwerk übernommen.

3) Wieweit muß und kann sich eine Infrastruktur, die dem Fach maschinenlesbares Material in großem Maßstab zur Verfügung stellt, auch als Anregerin seiner Verwendung verstehen? Wieweit besteht dabei innerhalb einer europäisch verstandenen Infrastruktur die Möglichkeit zur Anregung comparatistischer Arbeiten?
4) Welche Möglichkeiten gibt es für ein internationales Ressource Sharing, also: (a) für die Übertragung spezieller Aufgaben der Informationserschließung innerhalb der Gesamtdisziplin an hardware- und/oder softwaremäßig dafür besonders gut ausgerüstete Institutionen? (b) für die Verrechnung derartiger Leistungen zwischen den nationalen Infrastrukturen?

Zielvorstellungen

Der Einsatz der Datenverarbeitung in den Geisteswissenschaften zielte bisher in aller Regel auf die intensive Analyse einer Quelle. Dies unterschied vor allem in der Geschichtswissenschaft derartige Arbeiten vielleicht grundlegender von konventionellen, als der Einsatz der EDV an sich. Die Gründe dafür lagen hauptsächlich in der extrem arbeitsintensiven Vorbereitung, die für die Analyse einzelner Quellen notwendig ist, und in der Beschränkung der verfügbaren Ressourcen an den jeweiligen Rechenzentren, die der Größe selbst mit sehr einfachen Rechentechniken zu bearbeitender Korpora allein schon wegen der entstehenden I/O Zeiten eine Grenze setzten.

In herkömmlichen Forschungsansätzen konzentriert man sich zwar üblicherweise ebenfalls auf ein überschaubares Korpus: gleichzeitig wird aber eine sehr viel größere Anzahl von Quellen für einzelne zusätzliche Argumente herangezogen. Dadurch erscheinen selbst vom heutigen datentechnischen Stand sehr große Datenbanken als relativ eng umgrenzte Quellengruppe, wenn sie mit der Ausdehnung der traditionellen Arbeiten zugrunde gelegten Quellen verglichen werden. Als signifikantes Beispiel dafür mag die Reaktion stehen, die die Bemerkung des Verfassers, integrierte Datenbanken, die den gesamten Bestand der Monumenta Germaniae Historica (nach einer groben Schätzung bereits als unstrukturierte Eingabedaten etwa 0.5 Gigabyte) umfaßten, wären heute mit großem, aber absehbaren Aufwand realisierbar, in einem Kreis überwiegend ohne EDV arbeitender Historiker hervorrief: zweifellos, so wurde bemerkt, handle es sich dabei um eine sehr zentrale Quellengruppe; im Falle des Diskussionsredners müßte eine Datenbank, um alle für seine Arbeit relevanten Quellen aufzunehmen, aber etwa die 300fache Menge an Material enthalten. Als illustratives Beispiel mag dies zeigen, daß man bei Designüberlegungen für eine Infrastruktur, die innerhalb der historischen Disziplinen von zahlreichen Forschungsprojekten mit wirklichem Gewinn genützt werden kann, um Größenordnungen anders denken muß als bisher. Kein Forscher kann freilich in Anspruch nehmen, eine derartige Materialmenge mit der gleichen Intensität bei der Auswertung des Informationsgehaltes ihrer elementaren Bestandteile zu bearbeiten, wie dies etwa bei der statistischen Auswertung dazu geeigneter Quellen geschieht. In aller Regel wird aus derartigen Korpora ein sehr viel kleinerer Teil des Materials zur eigentlichen Bearbeitung herangezogen, wobei auch bei konventioneller Ar-

beitsweise die Phasen der Quellensichtung und der Analyse des dadurch erhaltenen Materials unterscheidbar sind.

Will man die durch derartige Bemerkungen gebotene Herausforderung an eine datentechnische Infrastruktur in den historischen Disziplinen annehmen, ergibt sich als erste Zielvorstellung ein Mechanismus, der eine nach heutigen Vorstellungen extrem große Datenmenge – jedenfalls mehrere Gigabyte – verwaltet, aus der dann von Fall zu Fall mittels geeigneter analytisch „flacher" Zugriffsmechanismen relevante Teile herausgezogen und als „Metaquelle" für die Bearbeitung mittels sehr viel tiefergehender analytischer Routinen zusammengestellt werden.

Jede Umsetzung dieser Zielvorstellung wird sich fünf grundlegenden Problemen gegenüber sehen:

a. dem primären Zugang zu maschinenlesbaren Quellen in großer Zahl,
b. ihrer Sicherstellung in einer Weise, die den Zugriff auf Teile davon im Bedarfsfall mit einem Minimum an menschlicher Intervention jederzeit kurzfristig erlaubt, ohne die permanente Residenz des Gesamtmaterials auf Random Access Speichern zu verlangen,
c. ihrer Sicherstellung in neutraler, also dem Erscheinungsbild der Quelle ohne Zwischenschaltung konzeptueller Vorentscheidungen maximal angenäherter Form,
d. der Schaffung eines Selektionsmechanismus, der in der Lage ist, unterschiedliche Teile des Gesamtmaterials mit unterschiedlicher Eindringtiefe zu bearbeiten, um rechenaufwendiges Suchen in sicherlich irrelevanten Quellen zu vermeiden, und
e. der Notwendigkeit, aus dem Ausgangsmaterial selektierte Teilmengen in einer großen Anzahl extrem unterschiedlicher Darstellungsformen weiterverarbeiten zu können.

Die damit angeschnittenen Probleme sind, wenn man große Quellenmengen betrachtet, sicherlich heute noch nicht endgültig und definitiv lösbar. Unseres Erachtens gibt es jedoch in jedem der angeschnittenen Problemkreise Entwicklungen, die uns jetzt schon zumindest in die Lage versetzen, mittel- bis langfristig gangbare Lösungswege aufzuzeigen.

Unsere Zielvorstellungen für eine integrierte Quellenbank großen Ausmaßes lauten:

Um die Bereitstellung maschinenlesbarer Datensätze für die tägliche Arbeit in der historisch orientierten Forschung wirklich wirksam werden zu lassen, muß von der Verfügbarkeit unverhältnismäßig größerer Datenmengen als bisher ausgegangen werden.

Aufgabe einer als infrastruktureller Einrichtung verstandenen Quellenbank ist zunächst das Sammeln aller einschlägigen Materialien. Solche fallen an als Überbleibsel abgeschlossener Forschungsprojekte im klassischen sekundäranalytischen Sinn, in der Form von maschinenlesbaren Drucksätzen von Quelleneditionen und durch gezielte Umwandlung großer Korpora mittels mechanischer Lesegeräte.[2]

[2] Hier und im folgenden sollen unsere technischen Überlegungen als Anmerkungen folgen, um den programmatischen Charakter der engeren Zielvorstellungen nicht zu stören.

Erstes Ziel einer derartigen Einrichtung ist die Bereitstellung der gesamten Quellenbank in einer abgestuften Speicherungsform, die das Gesamtmaterial in mehrere Ebenen auf die zur Verfügung stehenden Speichermedien verteilt. Data Dictionaries und andere Behelfe, die zur ersten Beurteilung von einzelnen Quellen/Dateien benötigt werden, sind dabei sofort zugänglich; andere benötigte Komponenten des Gesamtbestandes, insbesonders der volle Quellentext, werden bei Bedarf vom System automatisch von geeigneten „unbeschränkten" Hintergrundmedien geladen.[3]

Maschinenlesbares Material, das für die historische Forschung von Interesse ist, fällt heute und in der näheren Zukunft aus drei Quellen an:

1. Bei Projekten, die sich bereits heute der Datenverarbeitung im eingangs skizzierten Bestreben bedienen, ein vergleichsweise kleines Quellenmaterial intensiv zu analysieren.

2. Als Resultat der Druckaufbereitung. Der rapide Vormarsch der Elektronik im Druckwesen liefert bereits heute maschinenlesbare Äquivalente vieler Quellenpublikationen. Da anzunehmen ist, daß sich diese Tendenz bereits in der näheren Zukunft noch erheblich verstärkt, kann man davon ausgehen, daß in absehbarer Zeit jede überhaupt neu edierte Quelle maschinenlesbar vorliegt. Dies gilt noch wesentlich stärker für eine primär statistisch / quantitativ verstandene Forschung; freilich bilden die sich aus dem Datenschutz ergebenden Fragen der Zugänglichkeit primär statistischer Materialien ein besonderes Problem, das aber eher in den organisatorischen als technischen Bereich zu verweisen ist.

3. Schließlich auf Grund neuer Technologien billig zugänglich werdende Materialien. Ohne endgültig entscheiden zu können, ob die Geräte bei ihrem derzeitigen technischen Stand bereits in der Lage sind, ältere, im Bleisatz erstellte Texte in jedem Fall schnell aufzubereiten, kann man doch davon ausgehen, daß die Verfügbarkeit von Kurzweil Data Entry Machines oder ähnlichen Geräten innerhalb kurzer Zeit die Kosten für die Lesbarmachung sehr vieler Quellen radikal senken und auch den Gedanken an die gezielte Aufbereitung großer Korpora, ohne sie sofort mit einer eng verstandenen Analyseabsicht zu verbinden, möglich machen wird. In allen drei Fällen entstehen freilich Daten, die von Projekten, die an Teilen ihres Informationsgehalts interessiert sind, nicht ohne vorbereitende Arbeiten verwendet werden können. Teil der zu skizzierenden Infrastruktur muß also Software sein, die in der Lage ist: (erstens) von einzelnen Projekten erstellte, quellengetreue Daten rasch und mit einem Minimum an menschlichem Eingreifen in eine neutrale, untereinander austauschbare Form umzusetzen, (zweitens) die zur Speisung von Satzmaschinen erzeugten Magnetbänder und andere Datenträger rasch so umzuformen, daß die satzrelevanten Flags entfernt werden und der verbleibende Text ebenfalls auf die bereits erwähnte neutrale Form abgebildet wird, und (drittens) entsprechende Leistungen für mit einem Minimum an Aufbereitung von mechanischen Lesegeräten produzierte maschinenlesbare Texte zu liefern.

[3] Für die Speicherung mit kurzfristigen Zugriffszeiten wird in Eröffnungsreden auf einschlägigen Kongressen gerne global auf die rapide sinkenden Preise sekundärer Speichermedien verwiesen. Wir glauben, daß wir es uns mit einem derartigen Verweis, wenn wir in absehbarer Zukunft mit der Schaffung einer Infrastruktur beginnen wollen, die Datenmengen der angegebenen Größenordnung bewältigen soll, viel zu leicht machen würden. Unseres Erachtens wird in der planbaren Zukunft selbst eine noch so drastische Reduktion der Preise für Speichermedien nicht in der Lage sein, mit der vorhersehbaren Explosion für die Geisteswissenschaften relevanter maschinenlesbarer Daten Schritt zu halten.
Wir glauben, daß sich aber auch hier Lösungen abzeichnen, die sehr wohl mit auf der Reduktion der Speicherpreise beruhen. Wir gehen davon aus, daß die entscheidende Frage ist, wieweit es gelingt, Daten in Form von hierarchisch auf verschiedenste Speichermedien abgelegten diskreten Einheiten, die untereinander durch ein internes Referenzsystem verbunden sind, so abzulegen, daß der Zugriff auf sie keine Kenntnisse des Benutzers über die Ei-

Die Sicherstellung der Quellentreue der verspeicherten Materialien liegt im Ermessen der die einzelnen Quellenbanken unterstützenden Institutionen und bildet Teil des Designs einzelner Entwicklungen: in jedem Fall ist jedoch eine Möglichkeit vorzusehen, die zur Verfügung stehenden Materialien auf ein internationales Austauschformat abzubilden.[4]

---

genheiten des Betriebssystems des Rechners voraussetzt. Im Prinzip gehen wir dabei davon aus, daß heute in sich abgeschlossene Quellen (die Transkription eines bestimmten Bestandes administrativer Aufzeichnungen, ein maschinenlesbar gemachter fortlaufender Text) auch vom Rechner noch als sinnvolle Entitäten erkannt werden. Wir sehen vor, daß auf dem verfügbaren Random Access Speicher nur relativ kleine Teile komplexer Data Dictionaries resident sind, die alle notwendigen Verweise enthalten, um benötigte Entitäten ohne bewußtes Eingreifen des Benutzers von im Prinzip unbegrenzten Datenträgern (wie etwa einer Magnetbandbibliothek) zur Bearbeitung laden lassen zu können.
Wir möchten betonen, daß diese Lösung bereits heute arbeitsfähig ist (in der Tat im Max-Planck-Institut für Geschichte in relativ einfacher Form eben realisiert wird), aber mit der Entwicklung neuer Speichermedien – etwa der Verfügbarkeit großer fix montierter Laserdisks zum teilweisen Ersatz heutiger Plattensysteme oder der Verfügbarkeit wechselbarer Laserdisks zum Ersatz von Bandeinheiten – linear fortgeschrieben werden kann, da sie bloß von der Unterteilung des gesamten verfügbaren Speichermediums in mehrere Klassen mit unterschiedlichen Zugriffscharakteristika ausgeht. Daß Data Dictionaries und File Directories selbst hierarchische Gebilde formen können, ist angesichts der Entwicklung der letzten Jahre auf dem Gebiet der Betriebssysteme ja wohl selbstverständlich.

[4] Die Sicherstellung des Materials in einer neutralen Form, die dem Erscheinungsbild der Quelle maximal angenähert ist, ohne durch konzeptuelle Vorentscheidungen beeinflußt zu sein, ist vielleicht die schwierigste Aufgabe überhaupt.
Nicht so sehr technisch gesehen. Hier liegen zwar ebenfalls einige, keineswegs triviale Probleme vor, schließlich soll ja gewährleistet werden, daß das Material zwar das Erscheinungsbild der Quelle wiedergibt, gleichzeitig aber durch die noch zu diskutierenden Selektionsmechanismen effizient bearbeitet und für Auswertungen leicht umgeformt werden kann. Man wird also kaum umhinkönnen, gleichzeitig mit dem einfachen Quelleninhalt Strukturierungshilfen zu verspeichern, die das Material nicht in konzeptuellem, sehr wohl aber in technischem Sinne vorgliedern und von vorneherein auch eine Reihe von Filtern vorsehen müssen, mit deren Hilfe sich gegebenenfalls der verspeicherte Quelleninhalt auf die relevanten Teile einschränken läßt. Ein Beispiel für letzteres Problem ergibt sich etwa bei der Bearbeitung erzählender historischer Quellen: soll derartiges Material für die sprachwissenschaftlich orientierte Analyse in eine entsprechende "Metaquelle" im oben definierten Sinn umgewandelt werden, wird man Wert darauf legen, vorhandene diakritische Zeichen als Indikatoren für Lautlängen und ähnliches voll zu bewahren. Soll die Auswertung desselben Materials nach sozialhistorischen Gesichtspunkten erfolgen (etwa durch die Selektion aller mit Personennamen verbundenen Textteile und die Verbindung dieser Textteile mit andersgearteten, primär statistisch relevanten Quellen), wird man die nachfolgenden Arbeiten stark vereinfachen, wenn man derartige Auszeichnungen so früh wie möglich entfernt.
Dennoch glauben wir, daß diese technische Seite der Probleme durchaus in den Griff bekommen werden kann; wesentlich gravierender scheint uns die organisatorische Seite. Rein technisch gesehen wäre die verbindliche Vereinbarung eines hochkomplexen internationalen Standards für die langfristige Speicherung großer textorientierter Datenmengen unter besonderer Berücksichtigung der Eigenarten historischen Quellenmaterials die vernünftigste Lösung. In der Praxis müssen wir nach allen Erfahrungen aber davon ausgehen, daß jene Institutionen, die einschlägige Arbeiten schon geleistet haben, nur in den seltensten Fällen bereit sind, auf die Beibehaltung der von ihnen entwickelten Standards zu verzichten; zum

Primäres Ziel einer Quellenbank im hier beschriebenen Sinn ist die Anlage eines möglichst großen Korpus für die historischen/literaturwissenschaftlichen Disziplinen insgesamt, vor allem die Verhinderung der Zerstörung bereits maschinenlesbarer Materialien, nicht so sehr ihre analytische Durchdringung. Um diese zu erleichtern, ist jedoch dafür Sorge zu tragen, daß bestimmte Klassen einmal durchgeführter rechenintensiver Auswertungen innerhalb des Gesamtsystems bekannt bleiben und bei ähnlichen Anfragen einem Benutzer ohne menschliches Eingreifen als „billigere" Einstiege in sein Forschungsproblem angeboten werden. Durch die Kumulation derartiger Resultate und zusätzliche Bemühungen um die Aufnahme globaler Angaben über einzelne Quellen besteht die Möglichkeit, das Gesamtkorpus in mehreren, durch ihre analytische Tiefe und den dadurch bedingten Rechenaufwand unterschiedenen, Stufen nach in ein konkretes Projekt einzubeziehenden Teilen zu durchsuchen.[5]

Teil allein schon deshalb, weil die Finanzierung in vielen Ländern wesentlich von der Vorstellung der Finanzierungsträger bestimmt ist, im jeweils eigenen Land völlig neue Maßstäbe für die internationale Entwicklung zu setzen.
Man wird also auf die technisch optimale Lösung verzichten müssen.
Machbar und wünschenswert ist dagegen die Vereinbarung eines maschinenunabhängigen Austauschformates für in große Quellenbanken zu verspeicherndes Material, etwa im Sinne der Metafiles der GKS Normungsvorschläge. Unter der Voraussetzung, daß die am Datentausch interessierten Institutionen in der Lage sind, einerseits die von ihnen intern in einer Vielzahl von Basisformaten verspeicherten Rohdaten (gegebenenfalls unter Verlust lokaler Besonderheiten) zum Austausch in das Normformat zu übertragen und andererseits in diesem Format empfangene Daten (gegebenenfalls unter Auslassung örtlich als nicht relevant angesehener Angaben) in das jeweils intern gültige zu übertragen, ist der größte Teil der anfallenden Probleme bereits gelöst.

[5] In der voraussehbaren Zukunft ist wohl kaum davon auszugehen, daß die vollständige Durchsuchung mehrerer Gigabyte nach bestimmen Suchbegriffen eine realistische Lösung für ein häufig einzusetzendes Arbeitsinstrument ist. Unseres Erachtens liegen Lösungsmöglichkeiten in einer Verfeinerung der bereits diskutierten Datenorganisation.
Zunächst denkt man natürlich an partielle Inversionen der Teilbestände in der gesamten Quellenbank, die wir uns ja als eine aus zahlreichen kleineren Datenbasen zusammengesetzte übergeordnete Datenbank vorstellen können. Für viele häufig verfolgte Forschungsinteressen wäre dies auch sicher eine Lösung. Allgemein wird man bei einer Quellenbank, die in der Lage sein sollte, gleichmäßig eine große Anzahl von Teildisziplinen zu bedienen, aber kaum in der Lage sein, alle relevanten Aspekte in partielle Inversionen einzugliedern. Zudem bleibt – auch wenn wir uns vorerst nur um Zielvorstellungen bemühen und die Frage ihrer Realisierbarkeit zunächst verschieben – jetzt schon festzuhalten, daß eine derartige Konzeption von einer intensiven Bearbeitung jeder einzelnen Teilquelle durch die die gesamte Quellenbank bereithaltende Institution ausgeht, was mit vorhandenen Ressourcen kaum realisierbar wäre. Unseres Erachtens müßte die Lösung in einer Doppelstrategie gesucht werden:
Einerseits in einer Erweiterung der Funktion der in das Gesamtsystem zu integrierenden Data Dictionaries, die nicht nur innerhalb der einzelnen verwalteten Quellen die elementaren Einheiten (im Sinne von "Variablen") beschreiben, sondern auch so viel wie möglich an Angaben über die Gesamtquelle beinhalten – Entstehungszeit, behandelter Zeitraum, sprachliche Charakteristika, eine kurze Quellenbeschreibung, eine Charakterisierung ihrer Bedeutung für die einzelnen Subdisziplinen etc. etc. Dementsprechend würde der Zugang zu den einzelnen Quellen stufenweise erfolgen – aber mit den erweiterten Möglichkeiten der Da-

Neben diesen grundlegenden Fälligkeiten bietet jede Quellenbank die Möglichkeit, ausgewählte Teile des Gesamtkorpus in eine „Metaquelle" in einer Vielzahl von Formen auszugeben, die der analytischen Weiterverarbeitung durch die jeweils besonders unterstützten Ansätze entgegenkommen.[6]

Soweit zu unseren Zielvorstellungen für die Struktur einer idealen, d.h. die absehbaren technischen Möglichkeiten voll nützenden „großen" Quellenbank. Wir haben bisher bewußt noch nichts über die unserer Ansicht nach anzustrebende Organisation des Zugangs zu einer derartig strukturierten Einrichtung gesagt, da wir glauben, daß hier eine ganze Reihe verschiedener Formen gleichberechtigt nebeneinander vorzusehen sein wird.

Prinzipiell gehen wir davon aus, daß eine Quellenbank, die sich als Beitrag zur Infrastruktur ihres Faches versteht, auf einem Großrechner eines akademischen Rechenzentrums installiert wird. Ohne hier auf die nationalen Unterschiede in der Finanzierung datentechnischer Dienstleistungen für die geisteswissenschaftlichen Disziplinen eingehen zu können oder zu wollen, setzen wir voraus, daß es das Ziel jeder derartigen Einrichtung sein muß, ihre Dienste innerhalb mindestens des nationalen Rahmens auf einer ähnlichen Ebene, wie dies wissenschaftliche Bibliotheken tun, zur Verfügung zu stellen, also im wesentlichen kostenlos.

Dabei ist die Datenhaltung so zu organisieren, daß ein benützender Forscher die oberen Schichten von Selektionsvorgängen (also die Durchsuchung nicht von Quellen, sondern von Quellenbeschreibungen) jederzeit ad hoc vornehmen kann, während jede im System verspeicherte Quelle innerhalb von 6 Stunden zur intensiveren Bearbeitung zur Verfügung steht.

Ein Verbund derartiger Quellenbanken ist vorzusehen, der im nächsten Hauptpunkt noch zu diskutieren sein wird.

---

tenverarbeitung auf allen Ebenen und unter Integration von Suchbehelf und Quellenbestand.

Andererseits ist neben dieses Prinzip einer hierarchischen Datendokumentation noch das der Addidivität der Anstrengungen einzelner Benutzer zu setzen. Wir gehen davon aus, daß in der Regel einmal in den Basisbestand aufgenommene einzelne Quellen kaum mehr verändert werden, daß sie bloß Teile an von einzelnen Benutzern sukzessive aufgebaute Metaquellen abgeben, die dann ihrerseits, als eigentliche Datenbasen der einzelnen Forscher, häufigen und intensiven Veränderungen unterliegen. Dies bedeutet, daß alle einmal für den Basisbestand durchgeführten Berechnungen valide bleiben. Dementsprechend schiene es sinnvoll, zumindest für bestimmte Klassen von Bearbeitungen (wie etwa die Erstellung partieller Inversionen einzelner Quellen innerhalb des Gesamtbestandes) eigene Dictionaries anzulegen. Dies versetzt das Gesamtsystem in die Lage: (a) derartige Berechnungen erst dann durchzuführen, wenn sie wirklich notwendig werden, (b) ihre Ergebnisse danach für alle späteren Benutzer sicher zu stellen und (c), in Verbindung mit einem etwas besser ausgebauten Logging System (also einer „Tagebuchführung" des Gesamtsystems), Benutzer von ihren Interessen wechselseitig in Kenntnis zu setzen.

[6] Die Frage, welche Möglichkeiten zur Erstellung von analyseorientierten Dateien aus der quellenorientierten Gesamtdatenbasis bestehen, wird für die praktische Verwertbarkeit jeder konkreten Quellenbank von ganz zentraler Bedeutung sein – gleichzeitig bildet dies aber ein Problem, das erst für die Designüberlegungen konkret zu realisierender einzelner Quellenbanken von Bedeutung ist, dagegen kaum entscheidend für die Überlegung, ob Quellenbanken als solche realisiert werden können und was die Zielvorstellungen für sie sein müßten.

Die zentral unterhaltenen Quellenbanken stellen interessierten Institutionen Teile des bereitgehaltenen Materials zur Verfügung, setzen dabei aber auch voraus, daß bei diesen Institutionen aufbereitete Daten und als speicherungswürdig klassifizierte Analysen ihnen wieder zugänglich gemacht werden.

Dies bedingt grundsätzlich eine Vernetzung der zentralen Quellenbank mit den Rechnern der Teile des Gesamtkorpus bearbeitenden Institutionen. Wir halten die energische Erinnerung daran für nötig, daß wir hier keine Festtagsreden halten, sondern über Konzepte sprechen, mit deren praktischer Verwirklichung sofort begonnen werden kann. Wir halten Spekulationen über die Zeiten, da man in der Lage sein werde, Quellenbanken, wie wir sie hier beschreiben, via Satellit in allen Teilen der Welt online zu bearbeiten, im derzeitigen Stand der Entwicklung für weitgehend unsinnig.

Wenn wir von einer „Vernetzung von Datenbanken" sprechen, halten wir uns vor Augen, daß:

- insbesonders in den Anfangstagen einer entstehenden Quellenbank die Zahl der Benutzer sehr klein und der Abstand zwischen einzelnen Anfragen sehr groß sein wird. Die Anschaffung von Hardware, die nur gelegentlich im Abstand von Tagen benützt wird, halten wir für nutzlos und gefährlich, da sie nur den prinzipiellen Gegnern der Anwendung formaler Verfahren in den historischen Disziplinen Argumente liefert.
- das Konzept einer Quellenbank – im Unterschied von einer Datenbank – die Übertragung ganzer Korpora von Quellen vorsieht und daher aus Kapazitätsgründen die Verwendung heute und in absehbarer Zeit zur Verfügung stehender kurzfristig anwählbarer Übertragungsleitungen der Postverwaltungen ausschließt.
- bei der Arbeit mit dem hier behandelten Material voluminöse Resultate zu erwarten sind. Eine „rough and ready concordance" z.B. erscheint uns am Bildschirm nicht eben als sinnvoll darstell- und beurteilbar.

Für die Kommunikation zwischen einer zentralen Quellenbank und einzelnen Benutzern, die daraus entnommene Metaquellen mittels lokaler Datenbanken verwalten, schlagen wir folgende Ebenen vor:

- Für die Weitergabe einzelner Korpora und die Rückgabe der Ergebnisse bestimmter rechenaufwendiger Analysen den Versand der in Frage kommenden Daten auf Magnetbändern oder anderen externen Datenträgern[7]. Dabei ist von

[7] Unseres Erachtens liegt das Problem bei der bisherigen Organisation von Archiven maschinenlesbarer Daten nicht darin, daß der Versand von Magnetbändern ein so extrem ineffizienter Weg wäre. Es liegt vielmehr darin, daß vor allem die kleineren derartigen Archive Anfragen sehr selten bekommen und es daher u.U. Monate, wenn nicht Jahre dauert, bis eine angeforderte Quelle auf einem Magnetband lokalisiert und aus einer mittlerweile obsoleten Aufzeichnungsart in eine der jetzigen Anlage verständliche gebracht wird, bevor die Erstellung des andernorts lesbaren Bandes ins Auge gefaßt werden kann. Und es liegt ferner darin, daß in manchen Fällen die bei der Datenerfassung verwendeten Kodiersysteme selbst der archivierenden Institution nicht mehr bekannt sind, ganz zu schweigen von sonstigen Informationen über das Material. Beide Probleme – denen zumindest das Max-Planck-Institut für Geschichte in der Sammeltätigkeit für ein einschlägiges Projekt immer wieder begegnete – können aber durch noch so intensive Bemühungen um die neuesten techni-

einem innerhalb des Netzwerkes der jeweiligen Quellenbank verbindlichen Austauschformat auszugehen, das von entsprechenden sekundären Datenbanksystemen unmittelbar verarbeitet werden kann und eine extensive Datenbeschreibung in standardisierter Form gleich mitlädt. Diese Angaben sind jedoch so zu halten, daß sie, wenn ein derartiges System nicht zur Verfügung steht, gegebenenfalls mit einem Editor einfach vom eigentlichen Text getrennt werden können. Auf dieser Ebene ist dann auch an die Möglichkeit zu denken, nicht nur Quellen, sondern auch Teile der Quellenbeschreibungen auszulagern und in sekundären Datenbanken zugänglich zu machen.

- Eine unmittelbare Vernetzung scheint uns derzeit nicht so sehr zwischen einer einzelnen Quellenbank und von ihr abhängigen Datenbanken von Nutzen, sehr wohl aber als eine Möglichkeit, die arbeitsintensive Verwaltung der Quellenbeschreibungen auf mehrere Institutionen aufzuteilen. Die Realisierung dieses Konzeptes kann etwa bedeuten, daß zwei Quellenbanken einerseits auf die literarischen Quellen einer bestimmten Epoche, andererseits auf die historischen spezialisiert sind. Die eigentlichen maschinenlesbaren Quellen liegen dann in den Hintergrundmagnetbandbibliotheken beider Institutionen; jede Institution verwaltet aber nur die Quellenbeschreibungen und Data Dictionaries ihres Fachgebietes und hat die Möglichkeit, via Netz (eventuell ohne daß der Benutzer dies merkt) die an der anderen Institution verwalteten zu konsultieren, wonach gegebenenfalls die benötigten Quellen aus der eigenen Hintergrundbibliothek geladen werden können.
- Wir halten für eine weitere Möglichkeit der Vernetzung in Verfolgung des eben Gesagten die Zeit für Detailüberlegungen noch nicht für reif, wollen aber doch auf sie verweisen. Grundsätzlich muß es unserer Meinung nach in absehbarer Zeit möglich sein, ein System von Quellenbeschreibungen und Data Dictionaries in erheblichen Teilen in ein Netz voneinander völlig unabhängiger Micros zu verlagern, während die eigentliche Quellenbank auf einer weitgehend als höchst intelligente Memory Bank verstandenen Großrechenanlage gehalten wird.

Derzeitige organisatorische Möglichkeiten

Viel im vorangegangenen Abschnitt Gesagtes mag heute nahezu utopisch erscheinen; wir sind jedoch der Ansicht, daß organisatorische Initiativen auf dem Weg zu diesem Ziele jetzt schon möglich und auch sehr wesentlich sind, wenn vermieden werden soll, daß die fachspezifischen Belange der Geisteswissenschaften bei der Entwicklung der an Universitäten zur Verfügung stehenden informationstechnischen Infrastruktur weiter ins Hintertreffen geraten. Folgende Maßnahmen sind u.E. sofort – und mit einer Ausnahme ohne Ausweitung der bestehenden Infrastruktur – realisierbar und sinnvoll.

a. Die Intensivierung der Sammlung einschlägig erstellter Datenbasen. Es hat zwar bereits zahlreiche Bemühungen um die zentrale Sicherstellung für ei-

schen Errungenschaften nicht, gelöst werden, sondern liegen ganz auf der Ebene der Organisation.

ne Teildisziplin aufbereiteten Quellenmaterials gegeben, die Erfahrungen im Max-Planck-Institut für Geschichte – sowohl beim eigenen Bemühen um Daten, als auch im Gespräch mit Kollegen, die ihre Daten gerne zentral archivieren lassen würden – zeigen aber, daß diese Dienste noch nicht hinreichend bekannt sind.

b. Die Sicherstellung der elektronischen Repräsentationen aller mit Hilfe des Computers gesetzter Werke. Hier ist u.E. zwischen zwei Problemen zu unterscheiden:

   - einerseits hat es Fälle gegeben, wo die Satzunterlagen von Editionen, bei denen eine Zweitauflage unwahrscheinlich war, vom Verlag vernichtet wurden. Dem müßte durch eine rechtzeitige Sammlung derartiger Materialien vorgebeugt werden.
   - andererseits gibt es Verlage, die unter Berufung auf das Copyright Ansprüche auf die elektronische Repräsentation eines Werkes sogar dann erheben, wenn das die Satzmaschinen steuernde Magnetband vom Verfasser an einem akademischen Rechenzentrum selbst erstellt wurde.

   Insbesonders die letzte Situation sollte geklärt werden, bevor sie sich mit unabsehbaren Folgen zum Gewohnheitsrecht verfestigt. Nach Ansicht des Verfassers sollte grundsätzlich festgeschrieben werden, daß die elektronischen Satzunterlagen von Werken mit Editionscharakter einschlägigen infrastrukturellen Einrichtungen zur Verfügung gestellt werden müssen. Einen Nachdruck von einer derartigen Darstellung der Quelle durch Dritte zu verhindern, kann gesetzlich nicht schwer sein, sodaß die legitimen Interessen des Verlagswesens keinesfalls tangiert werden: das ebenso legitime Interesse der Wissenschaft, eine einmal – letztlich auf Kosten der öffentlichen Hand – finanzierte Edition in bestmöglicher Form benutzen zu können, und das heißt eben auch unter Einsatz der elektronischen Datenverarbeitung, darf von allem Anfang an nicht zu einer Umwegsubventionierung mißbraucht werden.

c. Die eine Ausnahme von der Forderung, nicht über zusätzliche infrastrukturelle Mittel zu sprechen, sondern über die bestmögliche Ausnützung bestehender, bildet der Vorschlag, im Rahmen einschlägiger Institutionen die Frage zu besprechen, wie weit es möglich ist, sich gezielt um die langfristige Aufbereitung großer Quellenkorpora unter Einsatz mechanischer Lesegeräte zu bemühen.

   Unserer Ansicht nach haben Quellenbanken, wie wir sie hier besprechen, wesentlich langfristige Bedeutung. Es wäre kindisch – und hat u.E. vergleichbaren Vorschlägen in der Vergangenheit nur geschadet –, so zu tun, als würde die bloße Gründung einer Quellenbank innerhalb der unmittelbar darauf folgenden Jahre zu einer explosionsartigen Vermehrung sie benützender Forschungsprojekte führen. Deshalb sollte der Einsatz eines derartigen Gerätes zur systematischen Aufbereitung großer Korpora, wobei in erster Linie wohl die klassischen Editionsreihen in Frage kämen, nur unter folgenden Rahmenbedingungen angestrebt werden:

   - ein derartiges Gerät sollte grundsätzlich neben der langfristigen Umwandlung eines Korpus für ad hoc anfallende Wünsche der Disziplin zur Verfügung stehen,

- dabei wäre für die langfristige Komponente aber von vorneherein ein bestimmtes Kontingent der Gesamteinsatzzeit zu sichern.
- die Umwandlung sollte nicht nur auf die Speicherung und Bereithaltung der Quellen abgestimmt werden, sondern gleichzeitig auch Grundlage der unmittelbaren Schaffung von Arbeitsbehelfen für die in Frage kommenden Korpora sein.
- im Idealfall sollte von Anfang an ein inhaltliches Projekt, das das aufbereitete Material sofort verwendet, mitlaufen.

d. Die sich konstituierenden Quellenbanken sollten grundsätzlich im freien Austausch miteinander stehen und im Prinzip alle den gleichen Bestand an maschinenlesbaren Quellen bereithalten. Ganz abgesehen vom eher trivialen Vorteil einer dadurch gegebenen absoluten Datensicherheit scheint uns dies einerseits wünschenswert, da dadurch die Möglichkeit zum Leistungsvergleich konkurrierender Designs gegeben ist, andererseits weil dadurch ein erster wichtiger Grundstock für eine überregionale Vernetzung im bereits besprochenen Sinn gegeben wäre. Als Voraussetzung für die Realisierung dieses Prinzips ist davon auszugehen, daß der Datentausch zwischen derartigen Institutionen grundsätzlich kostenfrei erfolgt.

e. Vor allem der letzte Satz, aber auch generell die Punkte a bis d erfordern die Regelung der Fragen des Datentausches durch allgemein als verbindlich anerkannte Richtlinien. Die beiden leitenden Gedanken müssen sein, daß (erstens) im Bereich der Forschung einmal erbrachte Leistungen der gesamten Disziplin ohne weitere Kosten erhalten bleiben müssen, daß aber (zweitens) andere daraus keinen monetären Gewinn ziehen dürfen. Es ist uns klar, daß dies im wesentlichen bereits jetzt der herrschenden Praxis entspricht. Andererseits ist aber immer wieder zu beobachten, daß bei der Anfrage nach Daten, die in allen einschlägigen Publikationsmedien als „auf Anforderung verfügbar“ bezeichnet werden, einerseits recht zeitraubende Prozeduren notwendig werden, um die Genehmigung der ihre Aufbereitung ursprünglich finanzierenden Körperschaft einzuholen, andererseits in einzelnen Fällen durchaus auch versucht wird, einmal aufgelaufene Projektkosten später durch einen „Verkauf“ der Daten zumindest zum kleinen Teil wiederzuerlangen. Wir halten dies auch im Interesse aller datentechnische Methoden propagierenden Forscher in den historischen Wissenschaften für gefährlich, weil die einschlägigen Methodologien sich erst dann wirklich verbreiten werden, wenn es auch und gerade für studentische – d.h. also extrem ressourcenarme – Projekte möglich wird, sich ihrer zu bedienen.

Deshalb schlägt das Max-Planck-Institut für Geschichte zur Regelung der angesprochenen Probleme im Zusammenhang mit den Punkten (a) bis (d) eine Übereinkunft der entsprechenden infrastrukturellen Einrichtungen des europäischen Raumes an Hand folgender Punkte vor:

1. Maschinenlesbare Darstellungen historischer Quellen, die sich im Besitz einer Quellenbank befinden, stehen der Disziplin grundsätzlich kostenlos zur Verfügung.

2. Jede Weitergabe von maschinenlesbaren Quellen ist an eine schriftliche Einverständniserklärung gebunden:

- sie ausschließlich zu wissenschaftlichen Zwecken zu verwenden,

- keine Publikationen daraus zu produzieren, die Partien der maschinenlesbaren Repräsentation erneut in maschinelle Satzanweisungen umsetzen (um z.B. Anthologien unter Umgehung des Copyrights zu erstellen), soweit die betroffenen Teile über den Umfang üblicher Zitate hinausgehen.

3. Die ihre Bestände teilenden Quellenbanken können auch jene Quellen weitergeben, die sie selbst von einer anderen im Rahmen des allgemeinen Austausches empfangen haben.

4. Einzelnen Forschern, die sich der Dienste einer Quellenbank bedienen, erwächst die Verpflichtung, ihrerseits maschinenlesbar gemachte Quellen nach Abschluß der Auswertungen dieser Quellenbank zu übergeben.

5. Die einschlägigen infrastrukturellen Institutionen werden an die nationalen Finanzierungsträger der Forschung mit dem Vorschlag herantreten, in die Förderung aller datentechnisch unterstützten Projekte, bei denen maschinenlesbare Quellen entstehen, die explizite Verpflichtung der Projekte aufzunehmen, die entstandenen Daten kostenfrei nach Ende der Laufzeit des Projektes einer Quellenbank zur Verfügung zu stellen.

f. Die konkrete Implementation einer Quellenbank im beschriebenen Sinn kann kaum Gegenstand einer über die jeweilige Institution hinausgehenden Vereinbarung sein. Von Interesse und einiger Bedeutung für die Verhinderung von Mehrfachentwicklungen wäre jedoch ein detaillierter Austausch über die Entwicklungsabsichten einschlägiger infrastruktureller Einrichtungen. Darüber hinausgehende konkrete Absprachen zur Arbeitsteilung wären wünschenswert. Dabei muß es primär um die Frage gehen, wieweit Designprinzipien für austauschbare Module in diesem Zusammenhang anzustreben sind und gemeinsame Prinzipien für die Softwareentwicklungen formuliert werden können.

g. Große, unmittelbar nützliche und für die weitere Austauschbarkeit von Daten sehr zentrale Fortschritte können durch die Formulierung verbindlicher Regeln für die Dokumentation maschinenlesbarer Quellen, also jene Angaben gemacht werden, die in integrierte Quellenbanken als Bestandteile von Data Dictionaries auf einem hohen Niveau eingefügt werden müßten.

In diesem Zusammenhang ist zu überlegen, wieweit angestrebt werden sollte, im Rahmen einer systematischen Lesbarmachung großer Korpora, wie unter (c) erörtert, vorrangig die klassischen Quellenrepertorien aufzubereiten. Diese stünden der Disziplin damit in Hinkunft als durch Datenbanksysteme verwaltbare Einheiten zur Verfügung, es wäre also sehr viel einfacher, Änderungen und Ergänzungen einzuarbeiten

h. Mittelfristig von gleicher Bedeutung ist die Definition der bereits erwähnten Metafiles für einen Austausch von maschinenlesbaren Quellen zwischen einzelnen Quellenbanken, deren Inhalt dann durch geeignete Schnittstellen ohne menschliche Intervention in das lokale Datenbanksystem geladen werden kann. An besonderen Problemkreisen wären zu erwähnen:

- die einheitliche Beschreibung der für die Darstellung komplexer Zeichensätze verwendeten Konventionen.
- die einheitliche Beschreibung der für die Darstellung struktureller Zusammenhänge zwischen einzelnen Quellenteilen verwendeten Konventionen; für die häufigsten Formen (etwa der hierarchischen Über-

/Unterordnung) wäre in der Definition der Metafiles eine verbindliche Notation vorzusehen.
- die einheitliche Beschreibung der Konventionen, die für die Darstellung formaler, nicht notwendigerweise struktureller Eigenschaften verwendet wurden. (Zu denken wäre etwa an die Verwendung kursiven Drucks für Ergänzungen in einer edierten Quelle, der in manchen Datenbanksystemen als niedrigere Sichtbarkeit der betroffenen Quellenteile zu implementieren ist, in statistisch ausgerichteten durch einen niedrigeren Wert einer Gewichtungsvariablen ausgedrückt wird und bei rein satztechnischen Systemen durch eine Übertragung in die entsprechenden lokal gültigen Druckkonventionen bearbeitet wird.)

i. Einen wesentlichen Punkt bildet schließlich der Meinungsaustausch über die Richtungen, die in den einzelnen Organisationen für die Entwicklung von Software vorgesehen sind, um spezifische Auswertungen vorzunehmen, und die Frage, wieweit auch hier durch entsprechenden modularen Aufbau eine echte Arbeitsteilung durch die Kombinierbarkeit isoliert entwickelter Systeme gewährleistet werden kann.

j. Wichtig wäre schließlich auch ein Meinungsaustausch über verbindliche Vorstellungen einschlägiger Institutionen über die Leistungen, die akademische Rechenzentren in Zukunft erbringen sollten, um den voraussehbaren Interessen der Geisteswissenschaften zu entsprechen, insbesonders auch der Austausch der Vorstellungen, die jeweils über das künftige Verhältnis von Mikros und Mainframes gerade bei daten-/quellenbankorientierten/ Projekten bestehen

Möglichkeiten der Anregung zur Benutzung

Wie wir bereits einleitend bemerkt haben, war eines der großen Probleme zentraler Datenarchivierung im Interesse der Bereitstellung von maschinenlesbaren Materialien für die Wiederverwertung, daß dem Angebot, dessen grundsätzliche Richtigkeit kaum in Frage gestellt wird, nur in relativ seltenen Fällen durch die Nachfrage entsprochen wurde: Einige der dafür verantwortlichen Ursachen haben wir bereits aufgezählt. Über das Gesagte hinaus scheint jedoch für eine erfolgversprechende Installation wirklich verwendeter Quellenbanken eine eingehendere Analyse der Problemlage angezeigt. Vor allem sollte man vier Gruppen potentieller Benutzer einer solchen Einrichtung unterscheiden:

- Forscher, die sich bereits bisher in ihrer jeweiligen Teildisziplin mit Erfolg datentechnischer Verfahren bedient haben.
- Forscher, die den Einsatz der Datentechnik zwar in Teilbereichen erwägen, durch die hohen Kosten der Vorbereitung der benötigten Quellen für die maschinelle Analyse aber davon abgeschreckt worden sind.
- Forscher, die an den einschlägigen Methoden und Techniken an sich zwar wenig oder nicht interessiert sind, sehr wohl aber an selektiven, auf einen bestimmten Quellenbereich bezogenen Dienstleistungen interessiert wären, wie sie durch den Einsatz der Mittel der Datentechnik möglich sind.

- Studenten, die zwar nicht in der Lage sind, aufwendige Projekte selbst zu initiieren, aber die methodischen Entwicklungen mit Interesse verfolgen und die Anwendung des in Veröffentlichungen beschriebenen Instrumentariums gerne in Form kleinerer, oft dissertationsbezogener Studien erproben möchten.

Um ein einschlägiges Angebot für alle vier Gruppen nutzbar zu machen, ist für jede dieser Zielgruppen eine eigene Vorgehensweise erforderlich.

Am einfachsten gestaltet sich das Problem sicherlich jenen Forschern gegenüber, die bereits bisher die Datentechnik eingesetzt haben. Das Angebot einer Quellenbank richtet sich ja auch am unmittelbarsten an diesen Kreis, dem dadurch die Möglichkeit geboten wird, intensiven Studien an Hand eines beschränkten, selbst aufbereiteten Quellenmaterials eine zusätzliche Dimension durch die flankierende Heranziehung weiterer Materialien zu verleihen. Auch hier ist allerdings die Information sicher noch zu verbessern, insbesonders durch die Verfügbarmachung detaillierter Beschreibungen des Inhaltes und der besonderen durch ihre jeweilige Kodierung gebotenen Hilfen. Praktisch wäre diese Aufgabe im Rahmen der bereits verschiedentlich als notwendig bezeichneten Quellenbeschreibung zu lösen.

Der zweiten Gruppe, also Forschern, die bisher an der Anwendung einschlägiger Verfahren durch die Kosten der Datenaufbereitung gehindert wurden, kann eine Quellenbank durch zwei Strategien entgegenkommen: wenn es gelingt, durch zentral zur Verfügung stehende Lesegeräte systematische Bemühungen um die Lesbarmachung größerer Korpora zu erreichen, wird deren Effizienz in entscheidendem Maße davon abhängen, wieweit gerade die eben angesprochene Gruppe hinreichend von den dadurch gebotenen Möglichkeiten informiert werden kann. Unabhängig davon muß erreicht werden, daß einschlägige Institutionen, die bereits bisher methodische/technische Beratung geleistet haben, in die Lage versetzt werden, gegebenenfalls darauf hinzuweisen, daß Teile der angesprochenen Fragestellungen mit bereits verfügbarem Quellenmaterial bearbeitet werden können. Natürlich kann es nicht darum gehen, die Forschung auf jene Quellen zu spezialisieren, die – von der Warte des mit dem Angebot Konfrontierten mehr oder minder zufällig – gerade bearbeitbar vorliegen. Wir glauben aber, daß in allen jenen Fällen, wo der Wert des Einsatzes einschlägiger Verfahren zweifelhaft ist, die Erprobung der geplanten Techniken an Hand nicht allzu unterschiedlichen Materials eine wertvolle Entscheidungshilfe geben könnte. Wovon letztlich auch sehr viele Projekte profitieren könnten, die derzeit den Einsatz der Datenverarbeitung sozusagen „blind“ planen, ohne sich der Möglichkeit zu Pretests vor Beginn der aufwendigen Datenerfassung bewußt zu sein.

Als Dienstleistung für selbst nicht an formalen Analysen arbeitenden Kollegen ist zu prüfen, wieweit die an einer Institution vorgesehenen spezifischen Auswertungen geeignet sind, routinemäßig in vereinfachter Form auf wesentliche Teile des zur Verfügung stehenden Gesamtmaterials angewendet zu werden. Möglich wird dies freilich nur sein, wenn man sich von der Vorstellung frei macht, in jedem Fall perfekte, auf eine Lebensdauer von Jahrzehnten berechnete Behelfe zu liefern und statt dessen beispielsweise die Möglichkeit vorsieht, rein mechanisch Hilfsmittel eingestandenermaßen niedriger Qualität einmal zu erstellen, auf Microfiche zu speichern und nur auf Anforderung zugänglich zu machen.

Vielleicht der wichtigste Punkt ist der letzte: die Schaffung von Möglichkeiten für Studenten, sich bereits zu einem frühen Zeitpunkt mit den neu entstehenden

Techniken und Methoden vertraut zu machen, ohne daß dies in jedem Fall mit größeren Hintergrundinvestitionen verbunden werden muß. Zu diesem Zweck wäre eine enge Kooperation mit den Universitäten anzustreben. Konkret sollte einerseits die Möglichkeit ins Auge gefaßt werden, interessierten Universitätslehrern geeignete maschinenlesbare Quellen zur exemplarischen Analyse im Rahmen einer Lehrveranstaltung zur Verfügung zu stellen, andererseits sollte auch immer wieder mit Nachdruck betont werden, welche Möglichkeiten sich für komparatistische Ansätze durch die Verfügbarkeit großer maschinenlesbarer Korpora ergeben.

Der letzte Punkt scheint uns so wichtig, daß wir ihn auch gesondert nochmals hervorheben möchten: eine der wesentlich neuen Möglichkeiten, die durch die Verfügbarkeit maschinell verarbeitbaren Materials gegeben sind, liegt jedenfalls in der Verbesserung der Ausgangsbedingungen komparatistisch verstandener Ansätze. U.E. sollten Quellenbanken gezielt versuchen, interessierte Forscher von sich aus darauf hinzuweisen, daß für sie relevante Materialien vorliegen, deren Bearbeitung in der Zusammenschau Ergebnisse erbringen kann, die über die Summe der von getrennt voneinander arbeitenden Einzelstudien erbrachten deutlich hinausgehen.

International Ressource Sharing

Wir haben bereits betont, daß die hier vorgeschlagenen Ansätze Teil einer Infrastruktur bilden, die, um effektiv zu sein, über nationale Grenzen hinweg wirken muß.

Unmittelbare Konsequenz dessen war bereits die Forderung, daß alles maschinenlesbar gemachte Material grundsätzlich der ganzen wissenschaftlichen Gemeinschaft ohne weitere Kosten zur Verfügung stehen sollte. Mittelbar bedeutet ein derartiger Grundsatz natürlich, daß in einem Land getätigte Investitionen – etwa ein gezielt eingesetztes Lesegerät – grundsätzlich auch Projekten zugute kommen, die jenseits der Landesgrenzen beheimatet sind.

Wir glauben, daß dies nicht ein bloßer Zufall sein sollte, der sich aus der Natur von Quellenbanken ergibt, sondern daß die Realisierung und Bewältigung dieser Situation eine wesentliche Vorbedingung für Fortschritte bei der Nutzbarmachung der neuen Informationstechnologien für das Fach schlechthin bedeutet. Die Anwendung neuer Informationstechnologien im Bereich der Geschichtswissenschaften wird sicher stets zu einem hohen Grad von der Übernahme in anderen Bereichen entwickelter Problemlösungen getragen werden; andererseits scheint es uns zweifelsfrei, daß aus den allgemeinen Voraussetzungen historischer Forschung letztlich eine ganze Reihe von Anforderungen an eine Methodologie der Informationserschließung historischen Quellenmaterials abgeleitet werden kann und muß, die von in anderen Bereichen geleisteten Entwicklungen nie abgedeckt werden können. Jetzt schon läßt sich absehen, daß die dabei zu bewältigenden Schwierigkeiten über das Potential ad hoc geschaffener Einzellösungen weit hinausgehen.

Allerdings ist angesichts der heutigen Kosten professioneller Programmentwicklung durch kommerzielle Häuser und der vergleichsweise geringen Zahl von Historikern im allgemeinen – ganz zu schweigen von den wenigen, die jetzt schon die Datenverarbeitung in anspruchsvoller Form heranziehen – kaum anzunehmen, daß

Auftragsentwicklungen größeren Maßstabes möglich sind. Zur Bereinigung dieser Situation sind u.E. zwei Strategien möglich:

- einschlägige Entwicklungen in infrastrukturellen Einrichtungen und längerfristig arbeitenden Forschungsprojekten sollten von vorneherein soweit abgestimmt werden, daß sowohl die verwendeten Datenformate austauschbar als auch die einzelnen Programmsysteme so modular und damit übertragbar wie möglich gehalten werden.
- es muß die Möglichkeit geschaffen werden, zur Erprobung neuer Ansätze (in klarerweise beschränktem Ausmaß) Ressourcen der Institutionen, die an entsprechenden Entwicklungen arbeiten, auch durch auswärtige Projekte mit zu benutzen. Dies kann nicht heißen, daß Forschungsaufgaben einer mit einem unzureichenden Rechenzentrum ausgestatteten Universität stillschweigend ins besser dotierte Ausland verlagert werden. Es soll und muß unseres Erachtens aber möglich sein, methodisch interessante Pilotstudien zur Erprobung neuer Methoden und Techniken an allen in der methodischen Entwicklung stehenden Institutionen auch dann durchzuführen, wenn dadurch Projekte eines anderen, auch dem jeweiligen Ausland angehörigen Finanzträgers in substantieller Weise gefördert werden, ohne in jedem einzelnen Fall zeitraubende Überlegungen über die Übertragung eingesetzter Fördermittel über die Grenzen hinweg anzustellen.