Das EU-Projekt MEMORIAL - Digitalisierung, Zugang, Erhaltung

The EU Project MEMORIAL – Digitisation, Access, Preservation

Dr. Alexander Geschke
Zentrum für Bestandserhaltung, Leipzig
Mommsenstraße 7, 04329 Leipzig
geschke@zfb.com

Dr. Wolfgang Schade
Gesellschaft zur Förderung angewandter Informatik e. V.
Rudower Chaussee 30, 12489 Berlin
schade@gfai.de

1. Einleitung

Für viele Einrichtungen ist ein FuE-EU-Projekt eine willkommene Gelegenheit anwendungsorientierte Forschungen und Untersuchungen, die man eigentlich schon immer vor hatte, mit spürbarer finanzieller Unterstützung zu realisieren. Dazu kommen die internationalen Kontakte, die oft Eindrücke über die Stellung von Mitbewerbern oder potentiellen Kunden vermitteln. Wenn da nicht der dornige und arbeitsintensive Weg der Antragsstellung, das Risiko der Ablehnung und die administrativ-bürokratisch aufwändige Abwicklung des Projektes wären. Aber das muss offensichtlich wirklich sein, um Missbrauch vorzubeugen. Trotzdem bleibt es eine schwere Entscheidung den Aufwand im Vorfeld zu treiben. Es gibt nur ein probates Mittel, um die Entscheidung leichter zu machen: Man muss vorher genau wissen, was man braucht, erreichen will und möglichst auch noch wie man zu diesem Ziel gelangt. Je genauer die eigene Aufgabenstellung ist, die sich natürlich in die Grundrichtung der EU-Ausschreibung einpassen lassen muss, um so mehr Chancen hat der Antrag, um so erfolgreicher läuft das Projekt.

Das Zentrum für Bestandserhaltung beobachtete seit geraumer Zeit die Entwicklung der Sekundärformen. Als Zentrum wurde schon immer die Mikroverfilmung als eine der sinnvollen Methoden zur Sicherung und zum Schutz des Originals, beispielsweise vor zu intensiver Nutzung mit ihren bekannten Folgen, angeboten. Die große Freude um die hervorragende Haltbarkeit, die günstigen Kosten und technisch einfache Lesbarkeit der Filme haben in einigen Ländern zu großen Verfilmungskampagnen geführt. Andere hatten damit schon vor längerer Zeit als Sicherheitsverfilmung begonnen. Und wie es Kampagnen so an sich haben, wurde teilweise - nehmen wir nur die Vereinigten Staaten - das Kind mit dem Bad ausgeschüttet. Nicht nur, dass oft nach dem nicht nur dort üblichen Motto "quick and dirty" verfahren wurde , was all die oben genannten Vorteile aufhebt, nein, es wurden zum Teil danach sogar die platzraubenden Originale (Zeitungen und Zeitschriften) entsorgt. Die Information war ja gesichert.

Diese Ausführung ist kein hämischer Fingerzeig, sondern eine Erfahrung für die eigene Verfahrensweise und vor allem auch für unseren Umgang mit der Digitalisierung. Damit wir nicht missverstanden werde: Wir sind keine Apologeten die ewig abwarten, nie etwas Neues versuchen. Denn gerade in unserem Metier heißt abwarten oft, dem Zerfall der Originale tatenlos zusehen. Nein, im Gegenteil, wir möchten Ihnen im Folgenden eine Herangehensweise an die Problematik der Digitalisierung schildern, die weder dem Kampagnencharakter noch der ewigen Diskussion um Vor- und Nachteile folgt.

2. Grundaussagen zur Digitalisierung

Digitalisieren hat andere Charakteristika als der Mikrofilm. Die Beständigkeit ist nur gewährleistet, wenn die Daten ständig genutzt werden, d.h. eine Migration auf neue (im Vergleich zur Lupe hochentwickelte) technische Systeme und auf neue Datenspeicher realisiert wird. Dem steht der Vorteil des breitesten und unmittelbaren Zugangs gegenüber. Aber aus internationalen und nationalen Untersuchungen ist auch bekannt, dass die relativ hohen Kosten der reinen Digitalisierung nur ca. 1/3 der Gesamtkosten eines zugriffsfähigen elektronischen Verteilungssystems darstellen. In ein Gesamtprojekt gehen ja auch Transport, Logistik, Bildbearbeitung, Indexierung, Speicherung, Datenbankanpassung, Realisierung des Internetzugriff, Internet-Site-Wartung und Datenpflege ein. Deshalb ist es im Zentrum für Bestandserhaltung wichtig, mit dem Kunden den auf ihn zugeschnittenen Gesamtrahmen zu realisieren. Dies kann in Stufen erfolgen, wie bei der bereits erfolgten Digitalisierung von Bachautographen. In der ersten Stufe wurden die Blätter in abgestimmter Reihenfolge digitalisiert, indexiert und auf CD-ROM an den Kunden übergeben. Die zweite Stufe wird der Realisierung des Internetzugriffs gewidmet sein.

Das Projekt MEMORIAL hat an einem ausgewählten Anwendungsbereich den Gesamtprozess von der Digitalisierung bis zur Verfügbarmachung von personenbezogenen Daten aus Registern u.ä. zum Ziel. Dazu werden grundsätzliche Probleme aus der gesamten technologischen Kette der Digitalisierung von Dokumenten aus Archiven untersucht und einer Lösung zugeführt.

3. MEMORIAL-Projekt-Überblick

Bei dem Projekt stehen personenbezogene Daten im Mittelpunkt, die am Beispiel von Aufzeichnungen aus ehemaligen Konzentrationslagern so in einem technologischen Prozess eingebunden werden sollen, dass die Informationserhaltung und die optimale Auswertung garantiert werden können. Die Partner aus vier Ländern garantieren sowohl die Archivseite mit ihrer Aufgabe der Verfügbarmachung als auch die technische Seite für die Lösung der entsprechenden Probleme. Das Archiv der KZ-Gedenkstätte Stutthof (MST) in Polen hat die meisten Dokumente und die Gedenkstätte Moreshet in Israel ist ebenfalls stark an der erweiterten Zugänglichkeit seiner Dokumente interessiert. Die technischen Fragen der Dokumentqualität und entsprechender Evaluierungssoftware stehen für die TU Gdansk (TUG) im Vordergrund, Probleme der Bildverarbeitung wie Hintergrundbereinigung sind Schwerpunkt für die Universität Liverpool (UniLiv) in Großbritannien, die Erkennung von Schrift und Schnittstelle zu Datenbanken werden durch die Gesellschaft zur Förderung der angewandten Informatik (GFaI) in Berlin realisiert, die Probleme der Web-Präsentation und Nutzerschnittstellen bearbeitet die Firma ID-Knowledge (IDK) in Israel und die Fragen zum optimalen Scannen der Dokumente sowie die Gesamtkoordinierung und Projektleitung erfolgt durch das ZFB in Leipzig. Die Laufzeit des Projektes beträgt zweieinhalb Jahre.

Das Projekt hat die folgenden wissenschaftlichen, technologischen und anwendungsorientierten Ziele:

- Verbesserung der Dokumenten-Digitalisierung und (-Bild-)Verarbeitung durch weiterentwickelte Programme um eine zuverlässigere Erkennung des Inhalts gegenüber dem Hintergrund zu gewährleisten.
- Entwicklung eines elektronischen Dokumentenformates, das für die Speicherung, Suche und den Abruf der Information solcher Dokumente in einer zukünftigen virtuellen Gedenkstätte geeignet ist.
- Untersuchung der rechtlichen, sozialen, ethischen und politischen Bedingungen für den Aufbau und die Nutzung digitaler Archive über den Nationalsozialismus
- Initiieren einer pan-europäischen Infrastruktur digitaler Archive, die Dienstleistungen einer virtuellen Gedenkstätte für individuelle Nutzer und für Forschungseinrichtungen erbringen kann.

Der konzeptionelle Grundgedanke für das MEMORIAL-Projekt ist in Abbildung 1 enthalten.

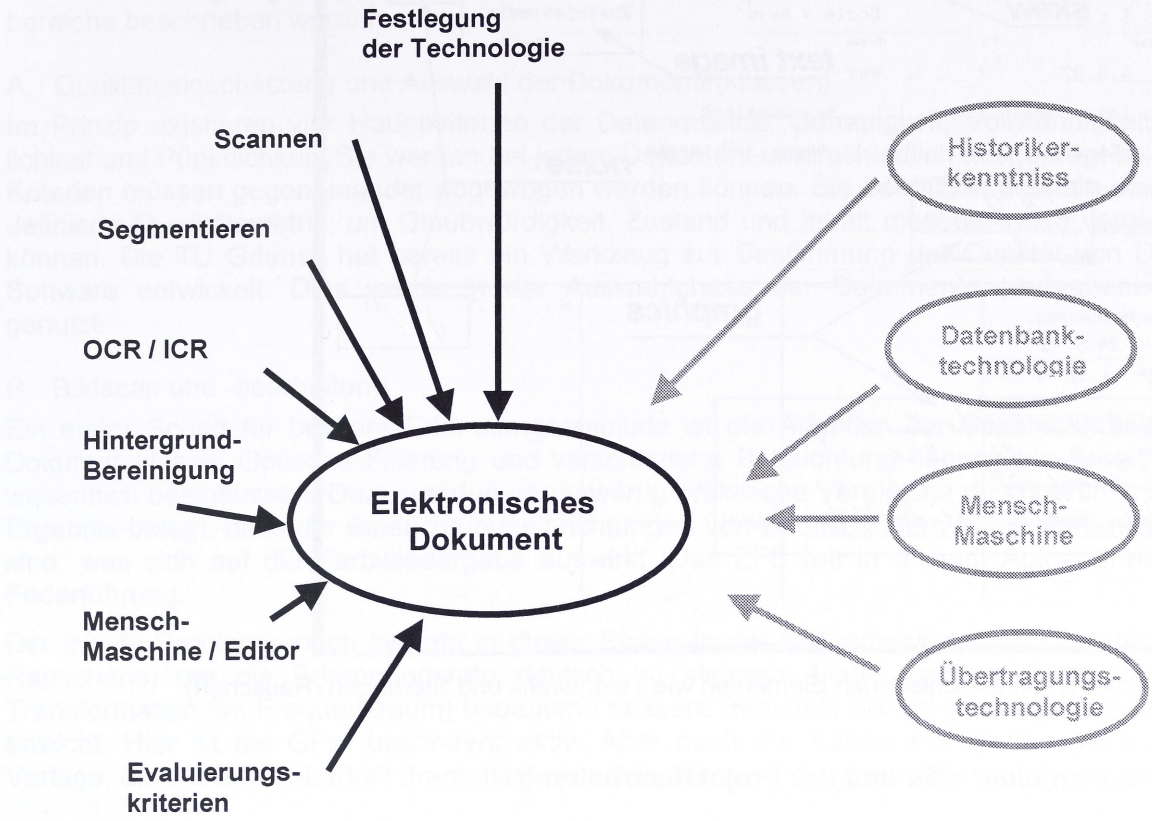

Abb. 1: Konzeptioneller Grundgedanke für das MEMORIAL-Projekt und weiterführende Aufgaben

Das Projekt ist in 7 Arbeitsaufgaben unterteilt. Die Aufgaben sind wie folgt festgelegt:

1. Definition verschiedener Dokumentklassen
2. Integrationskonzept für die Dokumenteingabe
3. Wissensbasierte (Bild-)Vorverarbeitung
4. Digitale Werkbank für Dokumentbehandlung (Software-Werkzeugsammlung) und Infrastruktur für den Informationsaustausch
5. Test & debugging aus Nutzersicht
6. Ergebnispublikation und Produktvermarktung
7. Projektleitung

Während die ersten 5 Aufgaben den technologischen Fortschritt umsetzen, sind die letzten beiden auf die Organisation und praktische Einführung gerichtet. An dem Beispiel einer Karteikarte sind in Abbildung 2 die technologischen Probleme veranschaulicht. Auffällig ist, dass verschiedenste Schriften benutzt werden und Störungen vom Hintergrund hinzukommen: Altdeutsche Druckbuchstaben (Frakturschrift) an geometrisch grob vorbestimmten Orten, Schreibmaschinenschrift mit lateinischen Buchstaben und handschriftliche Eintragungen in Sütterlin, Störungen wie Tintenflecke. Da das Projekt erst im März diesen Jahres begonnen wurde, kann nur von den Ergebnissen der ersten Aufgaben berichtet werden.

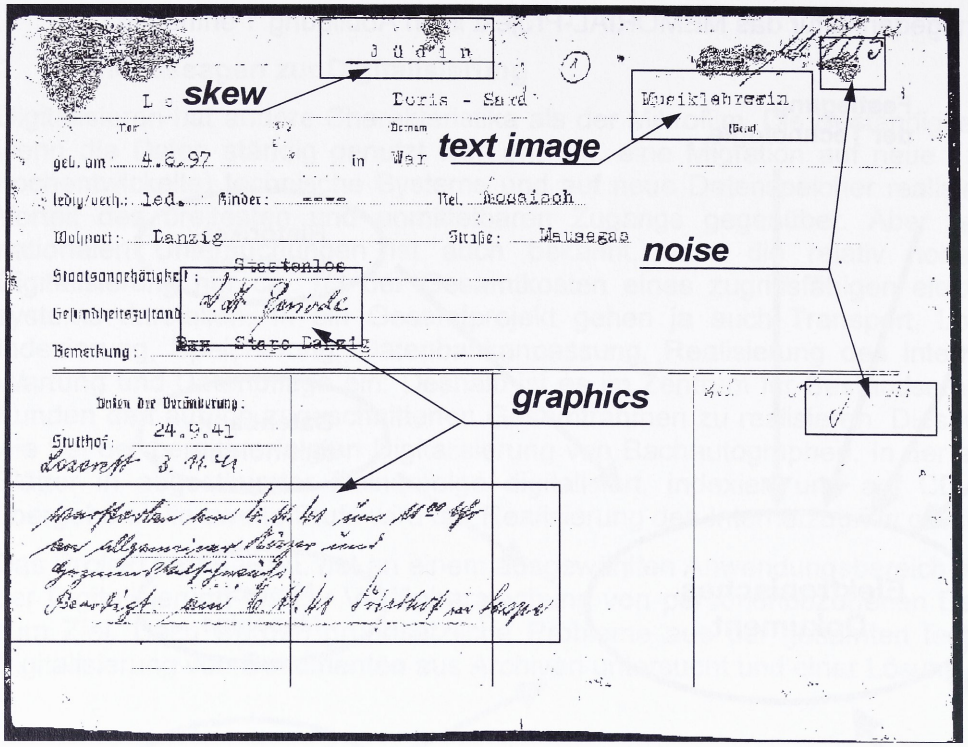

Abb.2: Karteikarte mit verschiedenen Elementen wie Text, Grafik und Störungen (Rauschen)

4. Die Problemfelder während der Projektbearbeitung

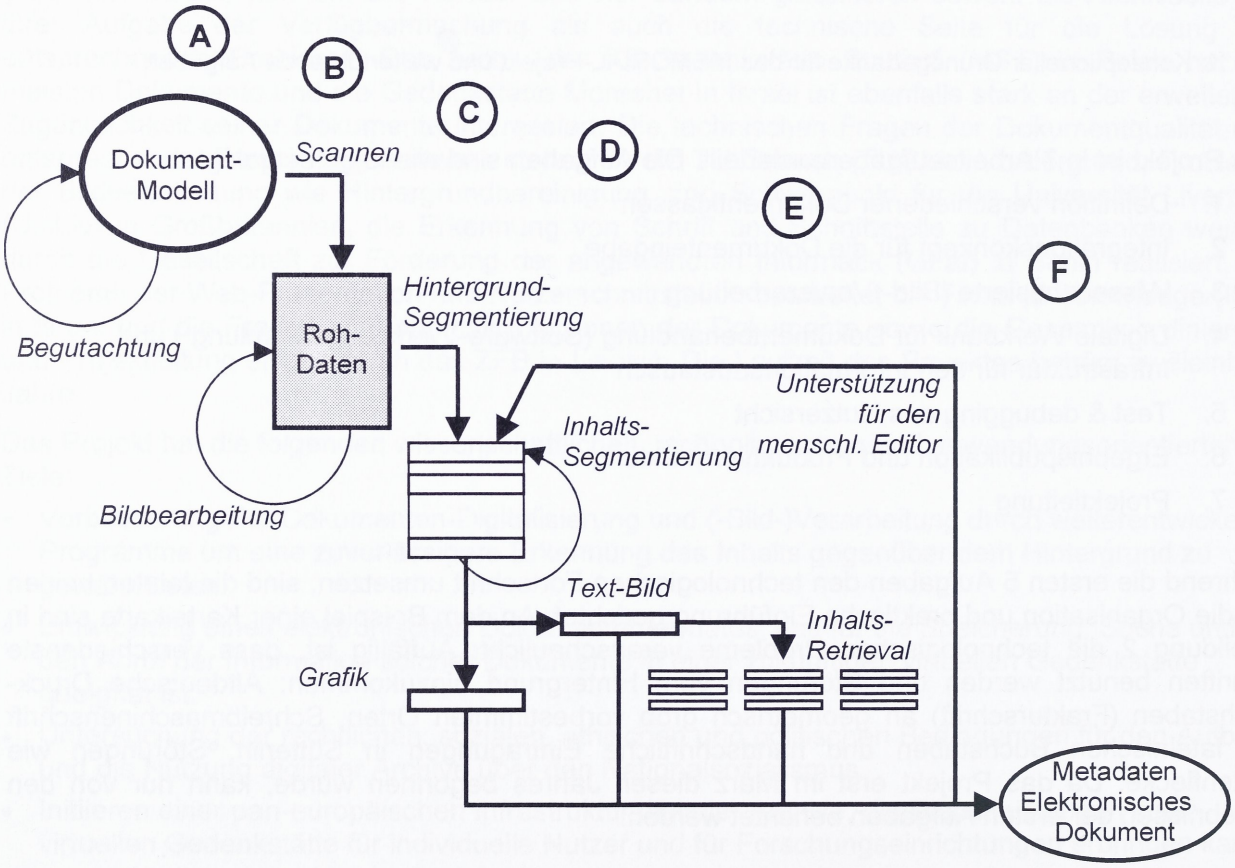

Abb. 3: Innovationsfelder des Projekts

Die in Abbildung 3 gezeigten Innovationsfelder (A bis F) können auch gleichzeitig als Problembereiche beschrieben werden.

A. Qualitätseinschätzung und Auswahl der Dokumente(klassen)

Im Prinzip existieren vier Hauptkriterien der Datenqualität: Genauigkeit, Vollständigkeit, Einheitlichkeit und Pünktlichkeit. Sie werden bei jedem Dokument unterschiedlich ausgeprägt sein und die Kriterien müssen gegeneinander abgewogen werden können. Sie benötigen deshalb eine speziell definierte Qualitätsmetrik, um Glaubwürdigkeit, Zustand und Inhalt messen - also vergleichen zu können. Die TU Gdansk hat bereits ein Werkzeug zur Bestimmung der Qualität von Daten und Software entwickelt. Dies wurde in der Auswahlphase der Dokumente (Aufgabenbereich 1) genutzt.

B. Bildscan und -bearbeitung

Ein erster Schritt für bessere Erkennungsresultate ist die Adaption der Scantechnologie an die Dokumentklasse. Optische Filterung und verschiedene Beleuchtung können die Erkennungsrate wesentlich beeinflussen. Dazu werden gegenwärtig praktische Vergleiche durchgeführt. Ein erstes Ergebnis belegt, dass die Beleuchtungseinrichtungen von Buchscannern bei weitem nicht optimal sind, was sich auf die Farbwiedergabe auswirkt. Das ZFB hat in diesem Aufgabenbereich die Federführung.

Der zweite Problembereich besteht in dieser Phase in der Bildverbesserung (Unterdrückung des Rauschens) um die Erkennungsrate deutlich zu steigern. Dabei werden durch die Wavelet Transformation (im Frequenzraum) bedeutend bessere Resultate als bei gewöhnlichen Filterungen erreicht. Hier ist die GFaI besonders aktiv. Aber auch der Kampf mit leichten Neigungen der Vorlage, die die Erkennbarkeit dramatisch verschlechtern, wird aufgenommen.

C. Dokumenthintergrund und Inhaltssegmentierung

Nach der Rausch- und Neigungskorrektur beginnt die Uni Liverpool mit der Unterscheidung des Textbildes von Grafikelementen und der dafür wichtigen Verringerung des Hintergrundeinflusses. Dies ist besonders wichtig, um die handgeschriebenen Informationen, die einer automatischen Schrifterkennung nicht zugänglich sind, als Grafik sicher zu erkennen und separieren zu können.

D. Inhalts-Abfrage (retrieval) der Seite

Konfrontiert mit verschiedensten Dokumentenarten, die Text und Anmerkungen enthalten in unterschiedlichem layout enthalten, besteht das Problem in der Identifizierung und dem Abruf der nützlichen Textinformationen und jeder einem korrekten Sammelbegriff (z.B. "Name") zuzuordnen. Dies erfolgte bisher nur für streng vorgegebene und relativ einfache Layouts, wie beispielsweise wissenschaftliche Vorträge oder Formulare.

Die Wiederherstellung des Textes soll sich vor allem auf die contextmäßige Analyse unter Nutzung syntaktischer Methoden stützen. Dabei werden auch Modellgrammatiken verwandt, die die Zuordnung verbessern. Auch hierfür liegen erste Erfahrungen bei unseren Partnern von der TUG vor.

E. Unterstützung für den Bediener-Editor

Unter Berücksichtigung der Beschränkungen unserer heutigen Technologien müssen wir die Erhaltung der originalen Digitalbilder (ohne Bearbeitung) mit den komplexen Dokumentbereichen für die menschliche Interpretation oder zukünftige Bearbeitungstechnologien erhalten. Folglich müssen die elektronischen Dokumente interaktive Arbeit zulassen und Raum für Annotationen, markierte Modifikationen und Links zu andern relevanten Informationsquellen ermöglichen. Gültige Standards für elektronische Dokumente unterstützen Annotationen. Der zu entwickelnde Editor muss ebenfalls einen hohen Grad an Flexibilität aufweisen, um jederzeit erweitert und angepasst werden zu können.

F. Metadaten-Kompatibilität

Die Extraktion von Metadaten aus einem Dokument während des in Abb. 3 gezeigten Zyklus stellt eine Gemeinsamkeit mit einem anderen laufenden EU-Projekt METAe dar. Dabei geht es vor allem um die Gewinnung von Metadaten aus dem layout von Büchern. Damit sind jedoch schon Unterschiede zwischen den Projekten vorgegeben. Trotzdem findet ein Informationsaustausch statt und vor allem wird auf dieser Basis eine Kompatibilität unserer Lösung mit der METAe Software angestrebt. In diesem Zusammenhang ist es von Vorteil, dass das METAe Projekt bereits seit 2 Jahren läuft und somit auf fertige Lösungen zurückgegriffen werden kann, um Parallelentwicklungen zu vermeiden.

5. Erste Ergebnisse

OCR - State of the Art

Die meisten kommerziellen OCR-Systeme haben gute Wiedererkennungsraten. Dabei ist jedoch Voraussetzung, dass die Dokumente einem bestimmten Qualitätsniveau entsprechen. Aber gerade dies trifft für Archivdokumente nicht zu. Deshalb müssen Vorverarbeitungsschritte die gescannten Dokumente in eine Form transformieren, die den erfolgreicheren Einsatz von OCR-Systemen ermöglichen.

Die folgenden kommerziellen OCR-Systeme wurden getestet: Textbridge, Eyes and Hands, Abby (Finereader), DOKu-Star. Effektivität und Ergebnisse der Systeme sind vergleichbar. Alle Systeme arbeiten nur mit Binärbildern. Bei neueren Dokumenten/ Ausdrucken ergeben sich kaum Probleme. Nur extrem kleine Schriften (unter 6 Punkte) bereiten Schwierigkeiten. Handgeschriebene Anmerkungen waren überhaupt nur zu erkennen, wenn die Buchstaben durch Abstände voneinander getrennt waren (Druckbuchstaben, Zahlen). Folglich hatte die Anwendung auf Archivdokumente insgesamt unbefriedigende Resultate. Hintergrundrauschen (von schlechter Papierqualität), vorgedruckte Linien (auf Karteikarten) und ausgelaufene, unscharfe Buchstaben von Schreibmaschinentypen führten bei allen OCR-Systemen zu Problemen. Hier ein Beispiel.

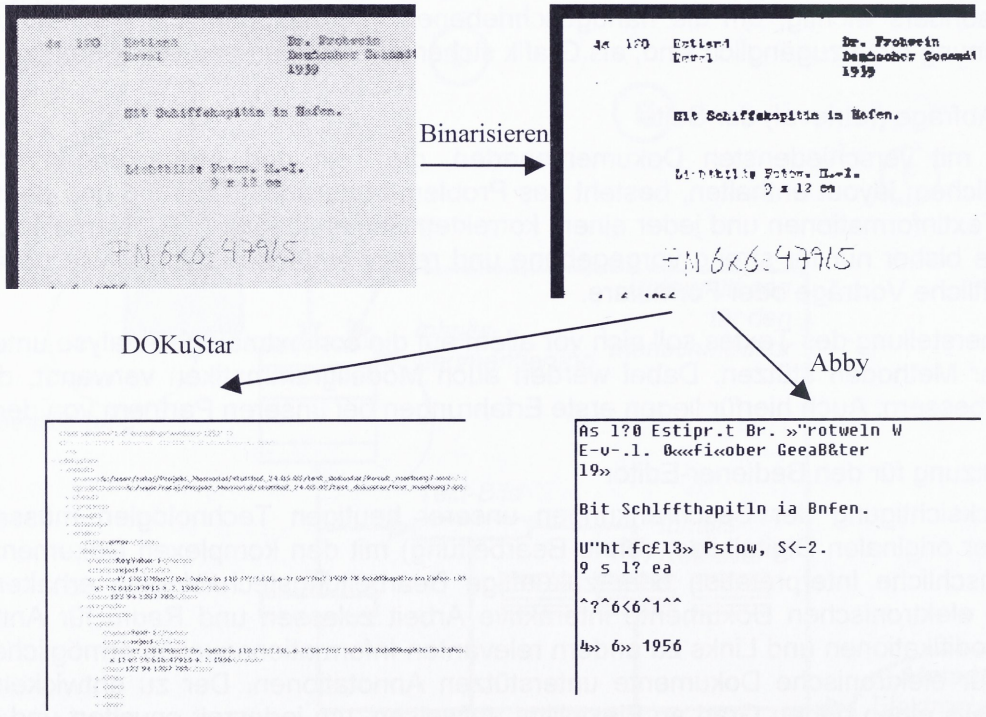

Abb. 4: Auslaufende Buchstaben und handgeschriebene Bemerkungen in DOKuStar und Abby

Die mögliche Definition von Regionen in einem Bild erleichtert die Erkennung zwar im Prinzip, ist aber nicht durchgängig anwendbar, da die Drucke der Karteikarten (Auflagen, Druckerei) nicht identisch sind und somit zu geometrischen Abweichungen führen. Forschung und Entwicklung konzentrieren sich bei OCR-Herstellern auf Bürodokumente, Rechnungen, Barcodes und Lieferpapiere, immer jedoch auf der Basis von Binärbildern.

Bildverbesserung - State of the Art

Wenn man berücksichtigt, dass die OCR-Systeme nur mit Binärbildern arbeiten und auch die Bildverbesserung auf diesem Niveau erfolgt, ist die einzige Möglichkeit hochqualitative digitale Farbbilder herzustellen und diese zu bearbeiten. Dazu ist eine wesentliche Vorstufe die Qualität des Farbscans, um die auf der Binärebene unmögliche Entscheidung zu vermeiden, welch Pixel zum Hintergrund und welches zum Buchstaben gehört. Bei Flachbettscannern ist durch die feste Anordnung der Beleuchtung immer ein gewisser Einfallswinkel des Lichtes von einer Seite gegeben, was zu einer Betonung der Strukturierung des Papiers führt. Bei dem getesteten Zeilenscanner von i2s ist die Beleuchtung beiderseitig des zu scannenden Spaltes angebracht, so dass der Hintergrund sehr flach abgebildet wird. Im Beispiel der Abb. 5 ist dieser Einfluss des Einfallswinkels bei bei den beiden Scannertypen demonstriert.

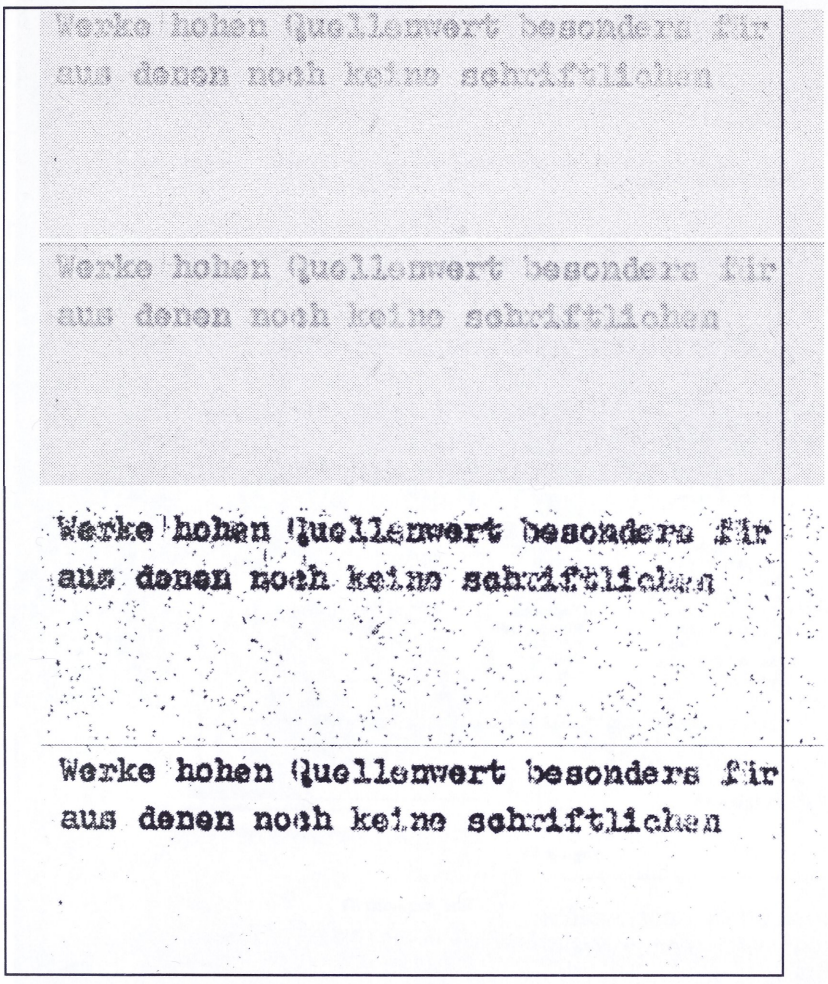

Abb. 5: Von oben nach unten: Scan mit Flachbettscanner (Papierstruktur erkennbar),
Scan mit Buchscanner I2S (flache Wiedergabe des Hintergrundes), manuell optimierte Tonwertkorrektur und Binarisierung des oberen Bildes und darunter des I2S-Bildes.

Die Hauptkorrekturen an den Bildern nach einem qualitativ guten Scannen sind Beseitigung der Winkelabweichung, Erkennung des Layouts, Säuberung des Hintergrundes und Verbesserung der Buchstabenqualität. Der Hintergrund wird für die Erkennung homogener, wenn das Rauschen minimiert (siehe Abb. 5 durch die Beleuchtung) oder durch spezielle Programme beseitigt wird. Ebenso ist die Erkennung und Beseitigung von Linien wesentlich.

Die Erkennbarkeit der Buchstaben kann beispielsweise durch weiterreichende Verarbeitungsschritte wie unscharfe Maske, Diletation und Skelettierung weiter erhöht werden. Dabei ist der Übergang in den Frequenzraum hilfreich.

Formales Modell maschinengeschriebener Dokumente

Diese Formalisierung trägt ebenfalls zur Verbesserung der Erkennbarkeit, aber auch zur Ableitung von Dokumentenauswahlkriterien bei. Abb. 6 zeigt ein Beispiel.

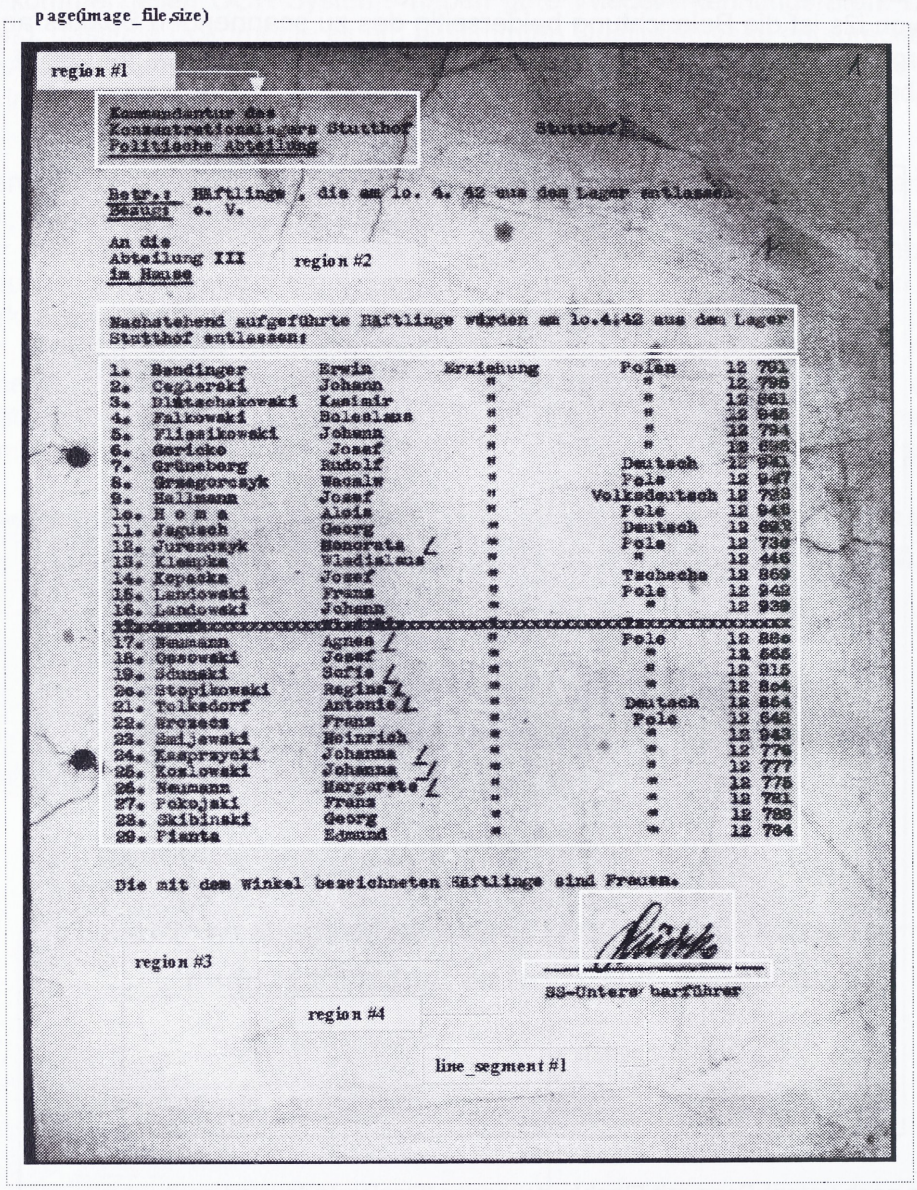

Abb. 6: Komponenten eines Beispieldokuments

Stringorientierter Text in der Region 1 wird in der Region 2 mit normaler Text-/Buchstaben-anordnung fortgesetzt. Danach folgt in Region 3 Text in Tabellenform, in Region 4 eine Unterschrift und darunter ein Liniensegment (No.1).

Nach dieser Prozedur dient auch als Ausgangspunkt für eine Darstellung in XML. Jeder Eckpunkt eines Strukturbaumes identifiziert einen Inhaltsanteil des Dokuments, also eine Region und entspricht einem XML- markierten Element.

Zusätzlich wurden Kriterien zur Bestimmung der Qualität von MEMORIAL-Dokumenten abgeleitet.

Als Reihenfolge lässt sich die Arbeit die erste Arbeitsaufgabe als Kette darstellen, wie sie in Abb. 7 gezeigt ist.

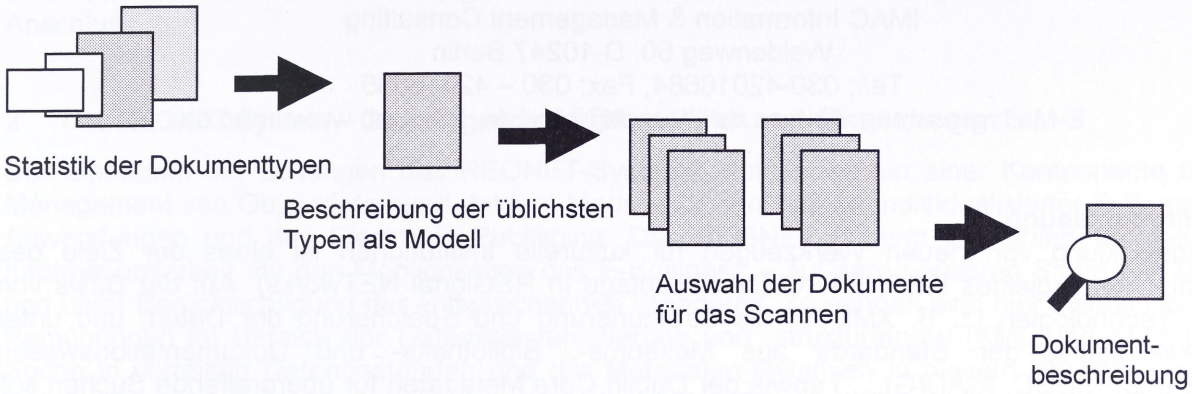

Abb. 7: Arbeitsaufgabe 1: Klassifizierung der Nutzerdokumente

Ein weiteres wichtiges Ergebnis der bisherigen Arbeit war der Vergleich der Zugriffsprozeduren für den Benutzer der Archive in den verschiedenen Staaten. Dabei wurden sowohl die EU-Richtlinien als auch nationale Reglungen in Deutschland, Israel und Polen berücksichtigt.

In der Arbeitsaufgabe 2, die gerade begonnen wurde, wird das Integrationskonzept für die Dokumenteneingabe untersucht und als Technologie etabliert, wie es in Abb. 8 umrissen ist.

Abb. 8: Arbeitsaufgabe 2: Integrationskonzept für die Dokumenteneingabe.

Die Folgeaufgabe 3 konzentriert sich auf die wissensbasierte Bildvorverarbeitung.