# Kostengünstige Digitalisierung eines Zettelkataloges

## Entwurf

Eberhard Pietzsch*

18.06.1998

## 1 Einleitung

Die Universitätsbibliothek Heidelberg bietet als erste deutsche Universitätsbibliothek einen elektronischen Zugang zu ihrem Zettelkatalog. Dazu wurde der Alphabetische Katalog 1936 – 1985 digitalisiert. Seit Juni 1998 steht der *DigiKat* als Recherche- und Nachweisinstrument für Benutzer und Bibliothekare im Internet zur Verfügung[1].

Beim DigiKat handelt es sich um ein *elektronisches Abbild* des Zettelkataloges. Die kostengünstige Realisierung basiert auf der Übertragung der Ordnungsregeln des Papierkataloges auf das elektronische Abbild. Mit dem DigiKat wurde ein Nachweisinstrument geschaffen, in dem i.A. schneller und mit größerer Trefferausbeute als am Papierkatalog recherchiert werden kann.

## 2 Intentionen

Beim Alphabetischen Zettelkatalog 1936 – 1985 (AK) handelt es sich um das wichtigste Recherche- und Nachweisinstrument für die in diesem Zeitraum erschienene Literatur im Besitz der Universitätsbibliothek Heidelberg[2]. Die später erschienenen Titel sind im lokalen Bibliotheks-EDV-System Heidi[3] erfaßt. Während antiquarische Erwerbungen aus den Erscheinungsjahren bis 1985 zunächst noch im Zettelkatalog geführt wurden, erfolgt die Erfassung seit 1997 ausschließlich in Heidi.

Der AK hat einen Umfang von ca. 1.2 Mio Karten und weist etwa 800.000 Titel nach. Seine Ordnung richtet sich nach dem Regelwerk *Preußische Instruktionen (PI)*[1]. Er wird gleichermaßen als Dienst- und Publikumskatalog

*Anschrift des Autors: Universitätsbibliothek Heidelberg, Plöck 107 - 109, D-69117 Heidelberg, E-Mail: Pietzsch@UB.Uni-Heidelberg.de

[1] http://www.ub.uni-heidelberg.de/digikat

[2] Die – größtenteils handschriftlichen – Kataloge bis 1935 liegen weiterhin in Papierform vor.

[3] Die Aufnahme der Titel erfolgt im Verbundsystem beim SWB in Konstanz. Die hier erfaßten Daten werden anschließend nach Heidi überführt.

genutzt.

Unterschiedlichen Schätzungen zufolge würde eine retrospektive Katalogisierung[4] des AK über den SWB in das lokale EDV-System zwischen 50 und 100 Personenjahre erfordern. Selbst bei größten Anstrengungen könnte der Zettelkatalog daher erst nach vielen Bearbeitungsjahren vollständig im EDV-System aufgehen. Gegenwärtig wäre darüber hinaus nicht an eine auch nur annähernd ausreichende Finanzierung eines solchen Vorhabens zu denken. Um Bibliothekaren und Bibliotheksbenutzern bereits heute eine elektronische Recherche in den Daten des Zettelkataloges bieten zu können, haben wir uns zu einer Digitalisierung entschlossen.

Weil für die Digitalisierung keine eigenen Landesmittel bereitstanden, war es um so erfreulicher, daß die Universitäts-Gesellschaft Heidelberg vom Nutzen des Vorhabens überzeugt werden konnte[5]: Sie hat die Sachkosten getragen, während die Softwareentwicklung als Eigenleistung bei der Universitätsbibliothek lag.

Im deutschen Sprachraum gibt es bereits vergleichbare Projekte; die wichtigsten sind hier zusammengefaßt:

**Bayerische Staatsbibliothek** [3]. Die Intention bestand hier im wesentlichen darin, künftig auf herkömmliche retrospektive Katalogisierung ganz zu verzichten. Daher sollten die Textdaten nicht mittels OCR[6] rekonstruiert werden. Stattdessen wurden sämtliche Daten auf den Karten des Papierkataloges nachträglich kategorisiert und manuell in einer Datenbank erfaßt. Letzteres wurde im Firmenauftrag durch Hilfskräfte erledigt.

Die Definition der Kategorien wurde an neuere Regelwerke angelehnt, um eine homogene Überführung des Datenbestandes in das lokale EDV-System mit Recherche- und Ausleihfunktion zu erreichen. Ein wichtiges Projektziel war, die Überführung der nach alten Regeln[7] erstellten Karteninhalte in die neuen Kategorien möglichst verlustfrei und regelkonform zu gestalten. Das Vorgehen weist daher Elemente einer vereinfachten retrospektiven Katalogisierung auf. Die Überführung des Datenbestandes in das lokale EDV-Systen hatte auch eine natürliche Integration in das Ausleihsystem zur Folge. Der digitalisierte Katalog ist unter `http://193.174.99.237/ifk/ifk.html` erreichbar.

[4] Für diese Retro-Katalogisierung wären in erheblichem Ausmaß Fachkräfte erforderlich, denn die Titelaufnahmen würden zugleich in das heute gültige Regelwerk *Regeln für die alphabetische Katalogisierung (RAK)* überführt.

[5] Der Universitäts-Gesellschaft Heidelberg wird an dieser Stelle ausdrücklich für die großzügige Finanzierung der Digitalisierung gedankt. Ohne ihre Mittel wäre diese beträchtliche Angebotsverbesserung der Bibliothek nicht erreichbar gewesen.

[6] Unter *Optical Character Recognition (OCR)* werden automatische Verfahren zur Rekonstruktion von Textinformation aus digitaler Bildinformationen zusammengefaßt.

[7] Der Katalog ist nach der *Münchener Katalogisierungsordnung*, einem Vorläufer der Preußischen Instruktionen, angelegt.

**Österreichische Nationalbibliothek.** Beim Digitalisierungsprojekt an der ÖNB hat man sich im wesentlichen mit der Herstellung digitaler Images der Karten begnügt, auf einen Recherchezugang also zunächst ganz verzichtet. Benutzer bewegen sich allein mittels Navigation zum gewünschten Ziel, und zwar im wesentlichen mit Hilfe einer sogenannten binären Suche. Der Katalog ist unter `http://www.onb.ac.at/online_s/onfr.htm` erreichbar.

**Zentralbibliothek Zürich** [2, 6, 8]. Im Unterschied zu den beiden vorgenannten Bibliotheken wurden hier digitale Images der Karten hergestellt sowie die Texte der Karten maschinell mittels OCR rekonstruiert.

Eine Recherche ist zum einem in den maschinell rekonstruierten Ordnungsbegriffen der Karten möglich, und zum anderen als Freitextsuche im gesamten Kartentext. Dies erforderte eine fehlertolerante Retrievalsoftware, die von einer Spin-Off-Firma der ETH Zürich entwickelt wurde. Der Katalog ist unter `http://www-zb2.unizh.ch` erreichbar. Auch er ist an das Ausleihsystem angebunden.

Das Digitalisierungsprojekt an der Universitätsbibliothek Heidelberg war vom Gedanken geprägt, im Vergleich zu anderen Digitalisierungsprojekten mit einem Bruchteil des Budgets auszukommen. Das Gesamtvorhaben wurde in mehrere Phasen zerlegt, von denen nun die erste abgeschlossen ist:

- Im ersten Teilschritt wurde ein *elektronisches Abbild* des Zettelkataloges, der *DigiKat*, für das WWW geschaffen. Der Begriff elektronisches Abbild bezieht sich dabei vorwiegend auf Benutzungsaspekte: Im wesentlichen bleiben die vom Papierkatalog gewohnten Benutzungsmöglichkeiten erhalten. Es werden aber, gewissermaßen als Zugabe, verschiedene Mehrwertaspekte gewonnen.

  Für die erste Stufe wurden die digitalen Images der Karten hergestellt. Die separat erfaßten Köpfe der Karten (s. Abb 1, 2) lassen sich für die Recherche nutzen.

- Für weitere, später zu realisierende Teilschritte wird – ähnlich wie in Zürich – die fehlertolerante Recherche in den Kartentexten sowie die Anbindung an das lokale Bibliotheks-EDV-System (das etwa 1999 zu erwartende Nachfolgesystem von Heidi) mit simultaner Recherche und Ausleihfunktionalität angestrebt.

Aus Ersparnisgründen wurden nur die Herstellung der digitalen Daten im Firmenauftrag erledigt, ohne jedoch sämtliche Textdaten neu zu erfassen. Die Recherchesoftware ist eine Eigenentwicklung der Universitätsbibliothek.

Von Beginn an waren neben den Bibliotheksbenutzern auch die Bibliothekare eine avisierte Zielgruppe des digitalisierten Kataloges: Sie sollten bereits vom ersten Teilschritt, dem elektronischen Abbild, gegenüber der Papierkatalogbenutzung profitieren. Gerade auch für den Signierdienst wurde ein im

Vergleich zu herkömmlicher Arbeitsweise konkurrenzfähiges Arbeitsmittel angestrebt. Kurze Antwortzeiten standen daher ebenso im Mittelpunkt der Entwicklung wie eine ergonomische Benutzerschnittstelle. Die Vorteile des elektronischen Abbildes gegenüber dem Papierkatalog lassen sich folgendermaßen skizzieren:

**Schnelles Auffinden.** Nach ersten Erfahrungen kommen nur die PI-geübten Bibliothekare des Signierdienstes im Papierkatalog schneller voran als im elektronischen Abbild. Für andere bedeutet letzterer stets einen zeitlichen Gewinn.

**Wegzeiten, örtliche Unabhängigkeit.** Das elektronische Abbild ist von jedem Internet-Arbeitsplatz aus nutzbar, Wegzeiten entfallen.

**Mehrdimensionaler Zugang.** Eine Recherche ist – ebenso wie im Papierkatalog – in den Köpfen der Karten möglich, hier aber zusätzlich über einen permutierten Index.

**Zeitliche Unabhängigkeit.** Der Katalog ist zu jeder Zeit unabhängig von den Öffnungszeiten der Bibliothek verfügbar.

## 3 Der Katalog und sein Regelwerk

Der Alphabetische Zettelkatalog ist nach den *Preußischen Instruktionen* geordnet. Dieses im 19. Jahrhundert entstandene Regelwerk hat zwei Aufgaben: zum einen definiert es, wie die Ordnungsbegriffe aus den Titelangaben der Publikation zu bilden sind. Zum anderen legt es fest, welche Wirkung die Ordnungsbegriffe auf die Reihenfolge der Karten im Katalog haben. Heute sind diese Regeln kaum noch jemandem geläufig. Beispielsweise sind für eine solide Recherche inhaltliche Informationen notwendig.

Ein Beispiel soll verdeutlichen, daß sich die Korrektheit der Abfolge der Ordnungsbegriffe – und damit der zugehörigen Karten – bei einigen Katalogsegmenten erst nach intensivem Studium erschließt; algorithmisch ist sie nicht verifizierbar, was bei fehlenden inhaltlichen Informationen die Auffindbarkeit enorm beeinträchtigen kann. So stehen die Ordnungsworte der nachstehenden

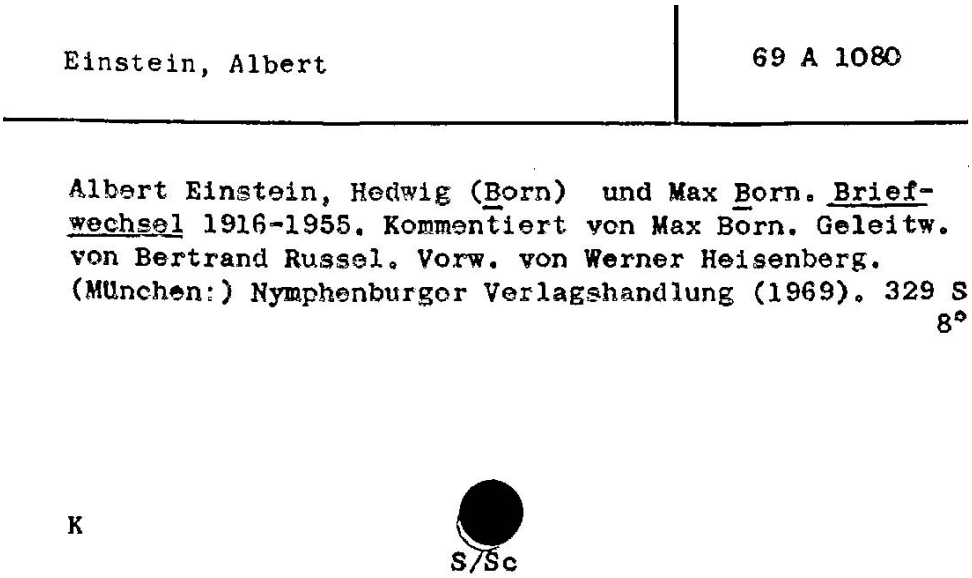


Abbildung 1: Die Segmente einer Karte. Links oben der *Kopf*, in der rechten oberen Ecke die *Signatur* und unterhalb des waagerechten Striches der *Text* oder *Korpus*. Unterstreichungen im Text weisen auf Ordnungselemente hin, die neben dem Kopf die Einordnung der Karte in den Katalog bestimmen. Die physische Qualität der Karten ist recht unterschiedlich: Manche Karten wurden im Laufe der Jahre um Eintragungen ergänzt, wodurch oft ein starkes Kontrastgefälle entstand, andere sind wegen häufiger Benutzung abgegriffen oder gar beschädigt.

Liste gemäß PI in aufsteigend geordneter Folge [1]:

```
Alexander Zweite. - ...                                  (Sachtitel)
Alexander, Grammatiker geb. um 1170                  (kein Beiname)
Alexander de Ales                          (nur der Beiname ordnet)
Alexander Alesius
Alexander Grammaticus                                    (Beiname)
Alexander Sanctus Episcopus Alexandrinus
                        (Beiname ordnet vor Herkunftsbezeichnung)
Alexander Polemius, Julius Valerius                  (Antiker Name)
Alexander 1827                             (einfacher Familienname)
Alexander, Pfarrer 1895
Alexander, Dr. med., Berlin
Alexander, Adolf                          (Familienname mit Vorname)
```

Die Gesetzmäßigkeit dieser Ordnung ist ohne profunde Regelwerkskenntnisse kaum erkennbar. Auch Rechner können ohne zusätzliche inhaltliche Informationen nicht verifizieren, ob die Reihenfolge korrekt ist. Dieses Faktum war bei der Schaffung des elektronischen Abbildes zu berücksichtigen[8].

[8] Die „Gleichwertigkeit“ der Buchstaben I und J bei den PI-Regeln ist algorithmisch hingegen leicht abzuhandeln.

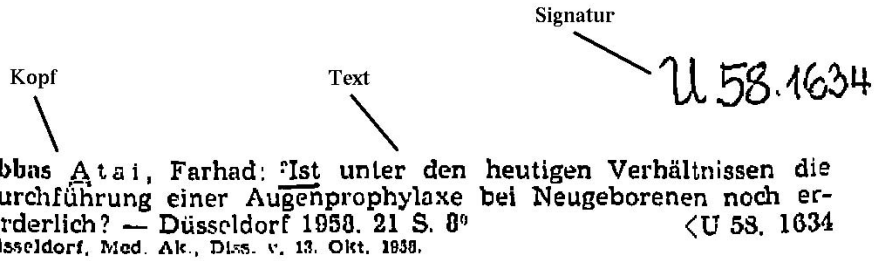


Abbildung 2: Manche Karten sind in der Schriftgröße 10pt oder kleiner bedruckt. Die Segmente sind dann nicht eindeutig erkennbar. Man beachte die handschriftliche Signatureintragung.

Jede Karte des Zettelkataloges besteht aus drei Segmenten: Dem *Kopf*, der *Signatur* und dem *Text* oder *Korpus* (Abb. 1). In der Regel dient ausschließlich der Kopf als Ordnungskriterium der Karten im Katalog. Er kann u.a. Ansetzungen von Autorennamen oder von Sachtiteln enthalten. In einigen Fällen finden sich auch nichtordnende Eintragungen im Kopf, z.B. Geburtsnamen oder Funktionsbezeichnungen von Autoren. Lauten die ordnenden Bestandteile der Köpfe mehrerer aufeinanderfolgender Karten gleich, so enthält der Kartentext weitere Ordnungsworte, die z.B. durch Unterstreichen hervorgehoben sind. Die Gesamtheit der Ordnungsbegriffe einer Karte kann also Teile des Kopfes sowie des Textes enthalten. Leider ist nicht bei allen Karten eine eindeutige visuelle Segmentierung zwischen Kopf, Signatur und Text möglich (vgl. Abb. 2).

Bei Aufbau und Pflege des Kataloges über einen Zeitraum von immerhin etwa 50 Jahren haben sich darüber hinaus lokale Gepflogenheiten herausgebildet, die eine algorithmische Behandlung weiter erschweren:

- Die PI-Regeln sehen vor, in Ordnungsworten keine Umlaute zu verwenden. Tatsächlich sind aber gerade in den letzten Jahren Umlaute und andere Sonderzeichen häufig nicht aufgelöst worden.
- Es gibt Einträge mit Zusätzen, die für den Benutzer als Orientierungshilfe gedacht sind, so z.B. Funktionsbezeichnungen oder Geburtsnamen. Diese zusätzlichen, nicht PI-konformen Elemente sind maschinell nicht als solche zu erkennen. Beispielsweise können Geburtsnamen durchaus vor dem Vornamen stehen, haben aber keine ordnende Wirkung.
- Vornamen und Namenszusätze („von“) sind z.T. abgekürzt, z.T. nicht, auch bei sonst gleichlautenden Köpfen mehrerer aufeinanderfolgender Karten.
- Schließlich soll auch nicht verschwiegen werden, daß so manche Karte falsch eingelegt worden ist.

Diese Skizze zu Aufbau und Status des Papierkataloges läßt bereits erahnen, daß eine elektronische Fassung einige Probleme bereiten wird, wenn man

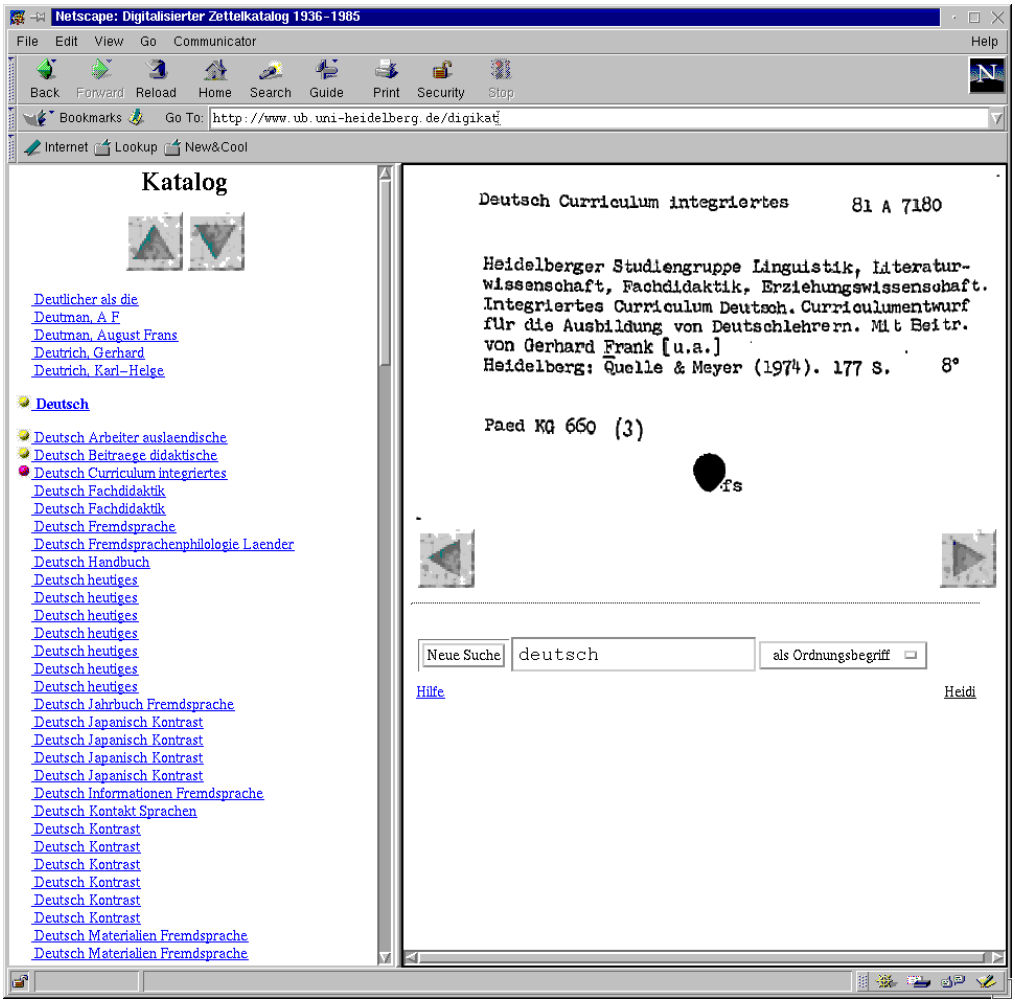


Abbildung 3: Die Benutzerschnittstelle bei der Suche im Katalogindex. Links die Umgebung des Recherchetreffers in der *Liste der Köpfe*: Zuerst die Köpfe von fünf Karten, die dem gewünschten und optisch herausgehobenen Treffer unmittelbar vorausgehen, darauf folgen die Köpfe der nächsten 50 Karten (abhängig vom 1. Ordnungwort können es mehr als 50 sein). Beim abgebildeten Beispiel hat der Benutzer in den als Hyperlinks ausbildeten Listenelementen bereits navigiert und die drei auf den Recherche-Treffer folgenden Karten angeschaut. Rechts das Image der zuletzt angewählten Karte. Im unteren rechten Teil des Schirmbildes ist die Möglichkeit gegeben, die nächste Recherche abzusetzen. Zur Beschleunigung der Navigation wird übrigens die Liste der Köpfe in den Fällen noch um Orientierungshilfen ergänzt, wo viele aufeinanderfolgende Eintragungen gleiche Bedeutung haben („Goethe“, „Steiner“).

auf eine Kategorisierung der Daten und ihre Neuerfassung verzichten will. Eine Kernaufgabe bei der Schaffung des elektronischen Abbildes war die Entwicklung einer Recherchesoftware, die den Benutzer in „fast“ allen Fällen zum gewünschten Ergebnis führt und ihm eine leichte Navigation erlaubt. Die Benutzung des elektronischen Abbildes war von Beginn an als Kombination von Recherche und Navigation vorgesehen. Beides wurde ganz wesentlich aus der gegebenen Abfolge der Karten im Papierkatalog abgeleitet, denn trotz oben beschriebener Eigenheiten des Papierkataloges ist die Reihenfolge „fast aller“ Katalogeinträge algorithmisch verifizierbar.

# 4 Das elektronische Abbild

Das elektronische Abbild des Papierkataloges ermöglicht dem Benutzer die Recherche in den Köpfen der Karten. Die in Abb. 3 präsentierte Benutzerschnittstelle gibt einen Eindruck von der Funktionalität.

Die digitalen Daten für das elektronische Abbild hat eine Partnerfirma[9] kommerziell hergestellt. Folgende Daten wurden geliefert:

**Images.** Von jeder Karte wurde ein digitales Image im Grafikformat TIFF G4 bei einer Auflösung von 240 DPI erstellt.

**OCR-Daten.** Zu jedem Image wurde die gesamte Textinformation mit einer OCR-Software ermittelt. Sie wird im ersten Teilschritt des Vorhabens noch nicht benötigt, steht aber für spätere Entwicklungen schon zur Verfügung.

**Liste der Köpfe.** Unter Zuhilfenahme der OCR-Daten wurden die Köpfe der Karten manuell nachbearbeitet und in einer Liste zusammengefaßt, und zwar in der durch den Papierkatalog gegebenen Reihenfolge. Diese *Liste der Köpfe* dient als Grundlage für die Recherche im elektronischen Abbild (Abb. 3).

Die zur Verfügung stehenden Daten dienten zur Schaffung zweier unterschiedlicher Recherchezugänge:

**Katalogindex.** Getreu dem Vorbild am Papierkatalog ist eine Recherche in den durch die PI-Ordnung gegebenen Köpfen möglich (Abb. 3).

**Permutierter Index.** Er bietet die Recherche in einem permutierten Index der Worte der Köpfe (Abb. 4). Davon profitieren gerade Benutzer, die die regelgemäße Ansetzung und Einordnung der Karte nicht nachvollziehen können und deshalb im Katalogindex nicht fündig werden. Besonders bei der Recherche nach Titelstichwörtern von Sachtiteln (z.B. „deutsch“) zeigt dieser Zugang seine Stärken, wie der Vergleich der Abbildungen 3 und 4 verdeutlicht. Mit Hilfe des permutierten Index läßt sich oft eine höhere Rechercheausbeute erzielen als am Papierkatalog (s. Abb. 4).

[9] Die UB Heidelberg hat mit GM Consult IT GmbH, Stuttgart, zusammengearbeitet.

Die Daten für den permutierten Index entstehen recht einfach aus denjenigen der Köpfe. In den nachfolgenden Ausführungen kann daher auf eine detailiertere Darstellung des permutierten Index verzichtet werden; die Darstellung des Katalogindex reicht aus.

Im vorangehenden Abschnitt wurde festgestellt, daß die Reihenfolge der Köpfe im Papierkatalog nicht immer algorithmisch verifizierbar ist. Die Benutzungsmodalitäten des elektronischen Abbildes sollten diesen Mangel ebenso per Software beheben wie typische Schreibfehler bei der manuellen Nachbearbeitung der Liste der Köpfe: Absolute Fehlerfreiheit der Daten war nicht notwendig, was natürlich dem Budget zugute kam. Folgende Fehler konnten ohne weiteres toleriert werden:

- Die manuelle Nachbearbeitung der Köpfe wurde maschinell indiziert: Nur ein mit höherer Wahrscheinlichkeit vorliegender OCR-Erkennungsfehler führte zur Korrekturaufforderung an den Bediener. Nicht erkannte, aber vorhandene, kleinere Fehler wurden daher nicht korrigiert.

- Die manuellen Korrekturen wurden von Hilfskräften durch Vergleich mit dem jeweiligen Image angefertigt. Mangels bibliothekarischen Fachwissens konnten sie im Image nicht immer zuverläßig zwichen dem Ordnungsbegriff und dahinter plazierter Signatur differenzieren. Darüber hinaus waren herstellungsbedingt Ordnungsbegriffe visuell nicht immer zuverlässig erkennbar (Abb. 2). Schließlich können die Köpfe auch nichtordnende Elemente enthalten, was für Hilfskräfte selten erkennbar ist. All dies hat zur Folge, daß in der manuell korrigierten „Liste der Köpfe“ nicht ausschließlich Ordnungsbegriffe stehen.

Das an der UB Heidelberg entwickelte Indexierungs- und Recherche-/Navigationsverfahren gleicht die vorstehend beschriebenen Merkmale aus. Folgende Überlegungen spielten bei der Softwareentwicklung eine Rolle.

Nach einer Recherche soll dem Benutzer der gefundene Treffer in seiner Umgebung im Katalog angezeigt werden (Abb. 3). Wird der gewünschte Begriff nicht gefunden – etwa weil er im Katalog nicht vorhanden ist – , so gilt der „unmittelbar darauf folgende“ Kopf als Treffer. Anders ausgedrückt: Eine Recherche bewirkt die Ausgabe des „kleinsten“ Treffers in der Folge von Köpfen, der „größer oder gleich“ dem recherchierten Begriff ist. Die Ermittlung dieses Treffers ist jedoch nur mit Hilfe einer *Ordnungsrelation*[10] auf der Menge der Köpfe möglich.

Vergegenwärtigt man sich z.B. die durch das Alphabet der 25 Buchstaben (I=J) gegebene Ordnungsrelation, so wird – wie eingangs bereits herausgearbei-

[10] Dieser mathematische Begriff steht für eine Eigenschaft von Mengen: Sie können (z.B. aufsteigend) angeordnet werden. Einfaches Beispiel ist die Ordnung der Menge der ganzen Zahlen. Sie läßt sich mit Hilfe der Ordnungsrelation „$\leq$“ anordnen. Ein anderes Beispiel ist die „alphabetische“ Ausgabe der Autorennamen oder Titelansetzungen in einem modernen datenbankbasierten Bibliothekskatalog; auch hierfür ist eine Ordnungsrelation verantwortlich.

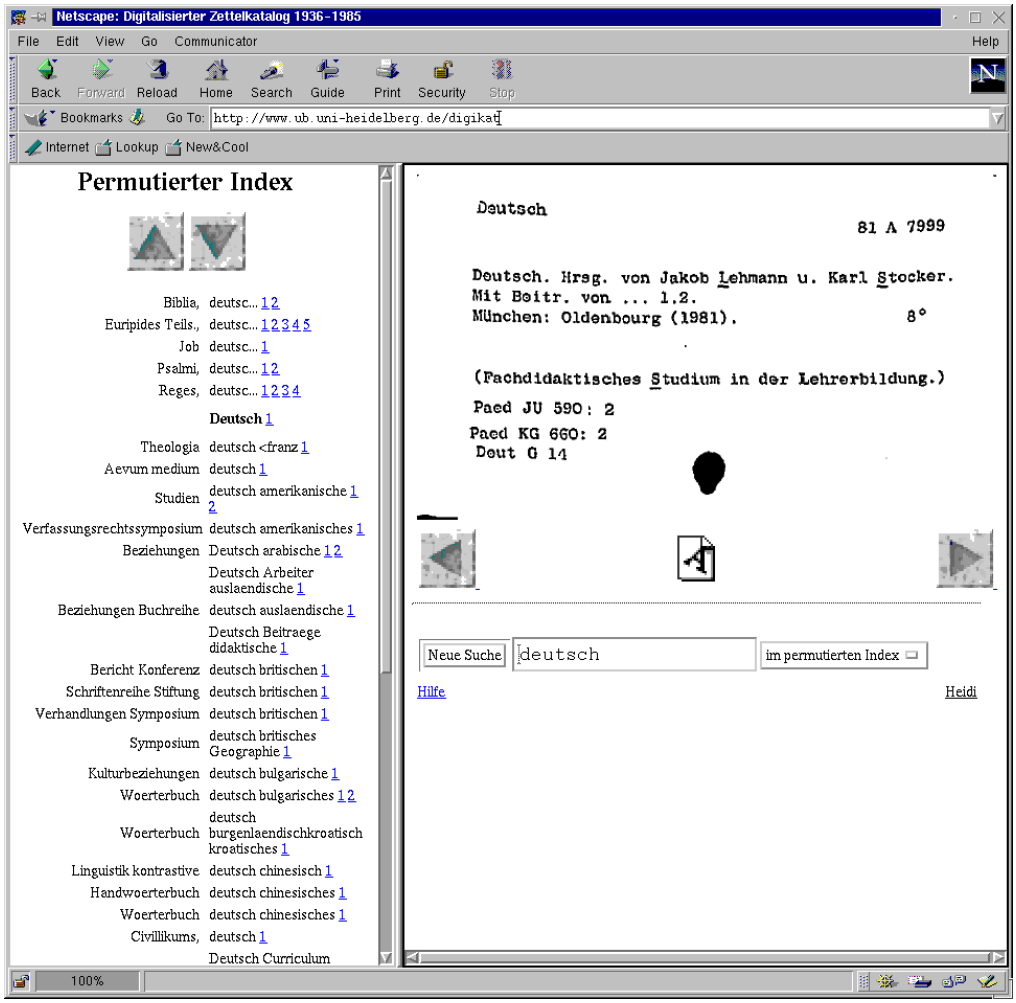


Abbildung 4: Die Benutzerschnittstelle bei der Suche im Permutierten Index. Links ein Ausschnitt aus dem Index in der Umgebung des recherchierten Begriffes „deutsch“; die Zahlen markieren Hyperlinks, die zu den entsprechenden Katalogkarten verweisen. Man vergleiche die Trefferausbeute mit der in Abb. 3.

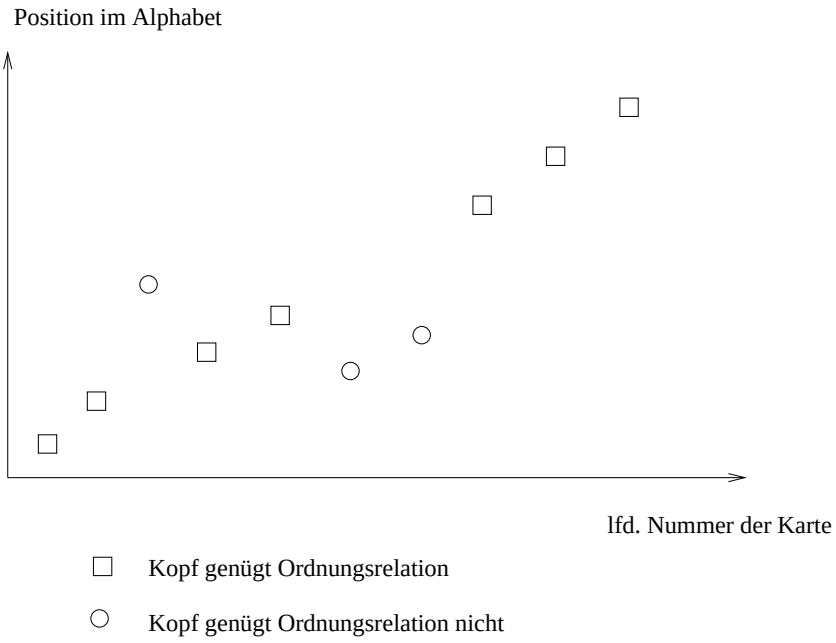


Abbildung 5: Zur alphabetischen Ordnung der Köpfe: Die mit dem Quadrat symbolisierten Köpfe werden bei der Indexierung als recherchierbar deklariert. Die anderen Köpfe sind nicht recherchierbar.

tet – schnell klar, daß die Folge der Köpfe im Katalog dieser Ordnungsrelation nicht sämtlich genügen können. Vielmehr haben die Köpfe eine Abfolge, wie sie in Abb. 5 skizziert ist: Die meisten Karten sind alphabetisch als richtig eingeordnet anzusehen, manche jedoch nicht. Dies gilt auch für andere Ordnungsrelationen als das Alphabet. Überhaupt ist schwerlich eine formale Ordnungsrelation zu finden, der sämtliche digital vorliegenden Köpfe genügen. Wichtige Aufgabe bei der Entwicklung des Recherchezugangs war daher die Definition einer Ordnungsrelation auf der Menge der Köpfe, der eine möglichst große Teilmenge an Karten genügt[11]. Nach Definition einer für den DigiKat sinnvollen Teilordnung sind die Köpfe, die der Ordnungsrelation genügen, recherchierbar. Die anderen werden als nicht recherchierbar deklariert. Eine Begriffssuche führt dann zu genau dem Treffer aus der Menge der recherchierbaren Köpfe, der gemäß der Ordnungsrelation[12] der *kleinste* ist unter allen, die *größer oder gleich* dem Suchbegriff ist. Ein Informationsverlust tritt nicht ein, denn die nicht recherchierbaren Köpfe sind durch Navigation erreichbar.

Für die auf den DigiKat anwendbare Ordnungsrelation sind etwa 94 % aller Köpfe recherchierbar. Dies bedeutet jedoch nicht, daß sämtliche verbleibenden Köpfe (also ca. 6% der Gesamtmenge) nicht retrievalfähig sind. Es ist nämlich zu berücksichtigen, daß mehrere aufeinanderfolgende Köpfe gleich lau-

[11] Mathematisch spricht man hier von einer *Teilordnung*.

[12] Die PI-Regeln sehen für Autorennamen und Sachtitelschriften leicht unterschiedliche Ordnungsregln vor; dies ist bei der gefundenen Ordnungsrelation berücksichtigt.

ten oder gleiche Bedeutung haben können, z.B. Autorennamen in unterschiedlichen Schreibweisen oder mit teilweise abgekürzten Vornamen. In diesen Fällen ist unwesentlich, wenn einige dieser Einträge nicht recherchierbar sind.

Die Auswirkungen auf die Benutzung wurde mit empirischen Tests ermittelt: Es wurde mit einer Stichprobe geprüft, um wieviele Positionen der gefundene Treffer vom gewünschten Treffer abweicht, wenn gerade nach nicht recherchierbaren Köpfen gesucht wird. Das Ergebnis für eine zufällige, 300 Recherchen umfassende Stichprobe ist in nachstehender Tabelle zusammengefaßt.

| Abweichung | 0 | 1 | 2 | 3 | 4 | $\geq 5$ |
|---|---|---|---|---|---|---|
| Anteil | 23% | 44% | 13% | 7% | 5% | 8% |

Tabelle 1: Suche nach nicht recherchierbaren Köpfen (6% der Gesamtmenge). Abweichung in der Position des tatsächlichen Treffers vom erwarteten Treffer.

Die Tabelle ist folgendermaßen zu lesen: In 23% der Fälle gelangt man sofort zum gewünschten bedeutungsgleichen Treffer, in 44% der Fälle weicht der gefundene Treffer um 1 Position von der erwarteten Trefferposition ab, in 13% um 2 Positionen usw. Man erkennt: Nur in wenigen Fällen ist tatsächlich eine nennenswerte Navigation erforderlich, weil ein unpassender Treffer gefunden wurde.

Weil 94% der Köpfe ohnehin recherchierbar sind, ist insgesamt in nur etwa 2% aller Fälle eine nennenswerte Navigation über 2 oder mehr Positionen erforderlich. Dieses Navigieren wird dem Benutzer mit einer intuitiv zu bedienenden Schnittstelle erleichtert. Minimal bessere Rechercheresultate hätten hingegen wegen der dafür erforderlichen Datenqualität beträchtlich höhere Produktionskosten zur Folge gehabt. Dabei sei darauf hingewiesen, daß schon Benutzer des Papierkataloges in Zweifelsfällen auch gut beraten waren, in der Umgebung ihrer Fundstelle noch zu blättern.

# 5 Zur Technik

Die Architektur des Systems ist nach einem mehrschichtigen Client-/Server Modell konzipiert (Abb. 6).

Kern des Server-Systems sind Perl-Scripts, die mittels Apache-Perl [9] so an den httpd-Server angebunden sind, daß zu jedem der laufenden httpd-Prozesse je ein Satz kompilierter Perl-Scripts im Hauptspeicher auf den Aufruf wartet. Die Perl-Scripts sind für einen Teil der Benutzerkommunikation ebenso zuständig wie für den Abruf von Daten aus der Datenbasis. Unmittelbar ausgelesen werden letztere jedoch von Server-Prozessen, die – ebenfalls in Perl mit den Datenbank-Modulen Berkeley-DB [10] – als Daemonen auf Anfragen warten. Bei dieser Konstruktion befinden sich sämtliche Programme als Daemon-Prozesse im Hauptspeicher. Auch die gesamte Datenbasis ist in einem

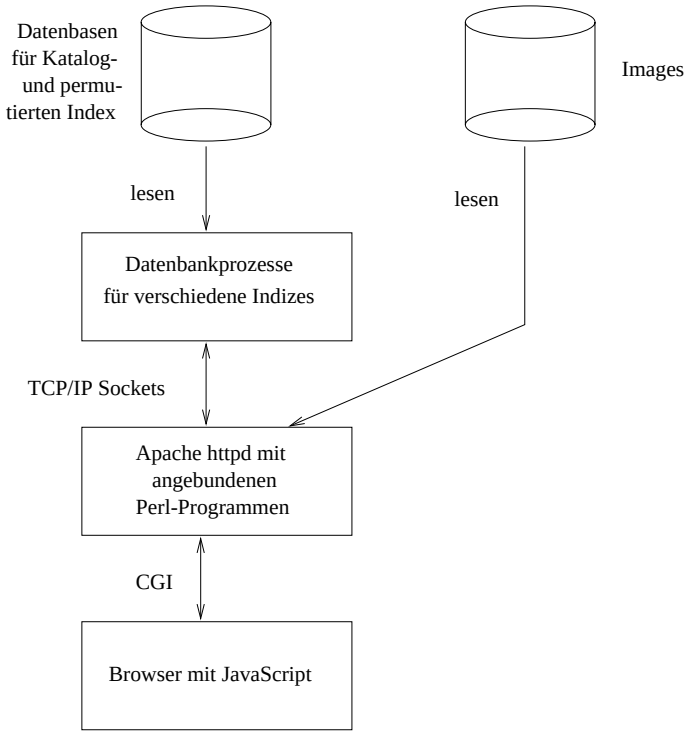


Abbildung 6: Systemaufbau. Bei der Realisierung an der UB Heidelberg laufen sämtliche Module bis auf den Browser auf einem gemeinsamen Host. Je nach Leistungsanforderungen könnten einzelne Module ausgelagert werden.

ausreichend dimensionierten Hauptspeicher angesiedelt. Plattenzugriffe sind im wesentlichen nur für die Imagedateien erforderlich. Außerdem könnten die Datenbankprozesse bei Bedarf auf einen eigenen Server gelegt werden.

Die Serversoftware ist plattformunabhängig und kann auf allen Systemen, auf denen *Perl*, der *Apache* WWW-Server und die Datenbank-Module *Berkeley-DB* ablauffähig sind, eingesetzt werden[13]. Die Hardware des Systems an der UB Heidelberg besteht aus einem Pentium Pro 200 PC mit 256 MB Hauptspeicher unter Linux und einem RAID-System mit einer Nettokapazität von etwa 35 GB. Ein zweiter PC steht als Entwicklungs- und Testsystem zur Verfügung und soll später die Recherche in den Freitexten übernehmen.

Beim Anwender ist ein Internetbrowser erforderlich. Hier sorgen JavaScript-Programme für Ergonomie und Reduzierung der über Netz transferierten Datenmenge[14].

[13] Dies sind sämtliche neueren Unix-Systeme.

[14] Aus Kompatibilitätsgründen können heute erst Netscape-Browser ab Vers. 3, bei denen *JavaScript* und *Cookies* aktiviert sind, die gesamte Palette der Features nutzen. Bei anderen Browsern ist die Benutzung nicht ganz so komfortabel. Textbrowser wie Lynx sind ausgeschlossen.

Die Antwortzeiten bei der Benutzung setzen sich im wesentlichen zusammen aus der Zeit, die der Server für die Bereitstellung der Daten benötigt, der Datenübermittlungszeit und der Zeit für die graphische Aufbereitung beim Benutzersystem. Typische Antwortzeiten des Servers liegen unter 0.1 Sekunden. Neben der Datenübermittlungszeit, die im Universitätscampus vernachlässigbar ist, ist die Geschwindigkeit des Arbeitsplatzrechners beim Benutzer für die Antwortzeit von ausschlaggebender Bedeutung. Bei modernen Arbeitsplatzrechnern liegen die Wartezeiten für den Benutzer in der Regel unter 1 Sekunde.

# 6 Ausblick

Mit der hier beschriebenen Digitalisierung wurde ein kostengünstiges elektronisches Abbild des Alphabetischen Kataloges, der *DigiKat*, geschaffen, das gegenüber dem Zettelkatalog schneller und mit höherer Ausbeute nutzbar ist. Es wurde dargestellt, welche Indexierungs- und Recherchemechanismen dem DigiKat zugrunde liegen.

Für weitere Realisierungsphasen ist die Anbindung an das für 1999 zu erwartende neue lokale EDV-System vorgesehen. Zum einen soll damit die simultane Recherche in sämtlichen elektronischen Katalogen an der UB Heidelberg ermöglicht werden, und zum anderen ist die Integration in das Ausleihsystem beabsichtigt.

Eine fehlertolerante Recherche in den Kartentexten könnte die Benutzung des DigiKat weiter verbessern. Wesentliche Herausforderung dabei ist, kurze Antwortzeiten zu garantieren.

## Literatur


[1] H. Allischewski: *Retrieval nach Preußischen Instruktionen*, Wiesbaden: Reichert, 1982

[2] *Alphabetischer Zentralkatalog der zürcherischen Bibliotheken (AZK) im WWW*, in: ABI-Technik 17 (1997), Nr. 2, pp. 157 f.

[3] C. Fabian, K. Haller: *Der Image-Katalog als alternatives Modell der Konversion*, in: ZfBB 45 (1998), Heft 2

[4] *Instruktionen für die Alphabetischen Kataloge der Preußischen Bibliotheken vom 10. Mai 1899*, 2. Ausg. i.d.F. vom 10. August 1908, unveränderter Nachdruck, Wiesbaden: Otto Harrassowitz, 1966

[5] H. Köstler, P. Schäuble: *Vollautomatische Konversion von Zettelkatalogen*, in: ZfBB: Sonderhefte; 70 (1998), pp 86 ff.

[6] E. Mittendorf, P. Schäuble, P. Sheridan: *Applying Probabilistic Term Weighting to OCR Text in the Case of a Large Alphabetic Library Catalogue*, in: ACM SIGIR Conference on R & D in Information Retrieval (1995) pp. 328-335

[7] E. Pietzsch: *Sichere öffentliche WWW-Zugänge in Bibliotheken*, in Vorbereitung

[8] P. Schäuble: *Kostengünstige Konversion großer Bibliothekskataloge*, in: ABI-Technik 16 (1996), pp. 165 f.

[9] Module `CPAN/modules/by-module/WWW/libwww-perl-5.08.tar.gz` und `CPAN/modules/by-module/CGI/CGI.pm-2.34.tar.gz` abrufbar z.B. unter `ftp://uni-erlangen.de/pub/source/Perl`

[10] `http://www.sleepycat.com/db`