DIGITALE ARCHIVIERUNG VON KABINETTSAKTEN DES ÖSTERREICHISCHEN STAATSARCHIVS.

Ein Zwischenbericht über ein gemeinsames Pilotprojekt mit der „Stiftung Bruno Kreisky-Archiv“

VON MARKUS MAZANEC – THEODOR VENUS – MARIA WIRTH

Einleitung

Im Frühjahr 1998 startete das Österreichische Staatsarchiv in Kooperation mit der Stiftung Bruno Kreisky-Archiv ein vom Jubiläumsfonds der Oesterreichischen Nationalbank gefördertes Projekt zur digitalen Erfassung der im Archiv der Republik aufbewahrten Kabinettsakten von Bundeskanzler a. D. Dr. Fred Sinowatz, die den Zeitraum 1983–1986 umfassen. Bevor im Folgenden die Konzeptualisierung und technische Realisierung dieses Projekts sowie operationelle Erfassung der Akten näher beleuchtet wird, soll in einem kurzen Rückblick auf die bereits im Vorfeld des Projekts gesammelten Erfahrungen hingewiesen werden.

Eine Kooperation zwischen den beiden Institutionen bot sich vom Standpunkt der Generaldirektion des Österreichischen Staatsarchiv aus an, weil es sich dabei sowohl um einen zeitgeschichtlich interessanten, wie auch einen aufgrund des Umfangs überschaubaren Aktenbestand handelt. Die Partnerschaft mit der Stiftung Bruno Kreisky-Archiv ergab sich, weil die Stiftung sich bereits seit Mitte der achtziger Jahre um die Digitalisierung ihrer eigenen Archivbestände, im Wesentlichen die Kabinettsakten des ehemaligen Bundeskanzlers und SPÖ-Parteivorsitzenden Bruno Kreisky's bemüht. Die im Verlauf des letzten Jahrzehnts getätigten Erfahrungen bildeten eine gute Voraussetzung, um dieses bisher ehrgeizigste Projekt gemeinsam mit dem Österreichischen Staatsarchiv in Angriff zu nehmen.

Die ersten Versuche der Stiftung Bruno Kreisky-Archiv zur elektronischen Archivierung und Bereitstellung von Dokumenten wurden bereits Mitte der achtziger Jahre unternommen. Technische Basis sollte ein integriertes System aus Dokumentenarchivierung und ein EDV gesteuertes Suchprogramm sein, wie es von den Firmen Siemens Österreich und Kodak entwickelt worden war. Dieses sollte eine rasche Suche nach wichtigen Dokumenten aus den Kabinettsakten Bruno Kreiskys mit Hilfe eines hierfür eigens entwickelten elektronischen Personen- und Schlagwortregisters ermöglichen. Nach einer Pilotphase wurde das Vorhaben jedoch wegen der schlechten Qualität der Mikrofilme aufgegeben.

Die wichtigsten Teilbestände der Kabinettsakten Bruno Kreiskys wurden im Zuge einer Neuordnung dieses Kernbestandes der Stiftung Bruno Kreisky-Archivs neu und detaillierter als bisher beschrieben. Das Bestandsverzeichnis steht jedem Internet-Benützer auf der Internet-Seite des Archivs zur Verfügung. Eine digitale Erfassung zumindest von Kernbeständen muß aus Gründen knapper Personalkapazitäten bis auf weiteres zurückgestellt werden.

Das bisher umfangreichste Digitalisierungsprojekt im Rahmen der Stiftung stellt die in der ersten Hälfte der neunziger Jahre durchgeführte Erstellung einer Datenbank der dem Archiv übergebenen Originale der Tagebücher des Langzeit-Handelsministers im Kabinett Kreisky Josef Staribacher (1970–1983) dar. Dieses Projekt wurde zwischen 1993–1995 im Rahmen eines EDV-Arbeitsplatzes am Bruno-Kreisky-Forum für internationalen Dialog, in der Armbrustergasse durchgeführt und befindet sich derzeit auch noch dort.

Ziel dieses Projekts war es, die im Gesamtumfang von etwa 20 000 Seiten (Eintragungen von 5–7 Seiten täglich an Werktagen) vorliegenden Staribacher-Tagebücher, die sich über einen Zeitraum von über 13 Jahren erstrecken, samt ausgewählter interessanter Beilagen in Form einer Volltextdatenbank zu erfassen und so der zeitgeschichtlichen, wirtschafts- und sozialgeschichtlichen Forschung zugänglich zu machen.

Das für diesen Zweck eigens entwickelte elektronische Archivierungssystem besteht aus einem Datenbearbeitungsplatz sowohl zur Datenerfassung wie auch Abfrage der digital archivierten Daten, dem Server (durch den die Einbindung in ein Netzwerk gewährleistet wird, einem Scanner zur Datenerfassung (mit OCR-Software) und einem angeschlossenen Drucker. Die erfaßten Daten werden im Zuge der Ablage komprimiert bzw. bei Abfrage dekomprimiert. Ferner dient der OD-Server zur Auslagerung der Archivbestände auf optische Datenträger über ein optisches Laufwerk.

Die Ablage von Text- und Bildinformationen erfolgt im Rahmen des eigens dafür entwickelten Programms „Clarity“, das die Anlage mehrerer diskreter Datenbanken in bis zu maximal 2 Millionen Ordnern ermöglicht, wobei „Clarity“ unterschiedliche (einfache wie auch verknüpfte Suchmodi mittels Verknüpfungsoperatoren) in den zunächst via OCR-Software gescannten Text- und Bildinformationen im Rahmen einer Volltextrecherche gestattet. Der Vorteil der OCR-Erfassung liegt darin, daß die erfaßten Informationen durch die Volltextrecherche tatsächlich ohne Verlust an historischer Information für Abfragen zur Verfügung stehen. Diese Vorgangsweise einer Texterkennung oder textuellen Digitalisierung setzt allerdings eine Mindestqualität der zu erfassenden Dokumentenvorlagen voraus, die bei vielen historischen (auch maschinschriftlichen) Materialien allerdings nicht immer gegeben ist.

Im vorliegenden Projekt, das in der Folge sowohl hinsichtlich der technischen Komponenten, die eigens auf den Zweck hin „maßgeschneidert“ wurden, wie auch der praktischen Realisierung dargestellt wird, wurde daher aus zeitökonomischen Gründen nicht die Kombination von OCR-Erfassung im Kombination mit Volltextrecherche gewählt, sondern die Kombination von Digitalisierung der Dokumente (graphische Erfassung) und Recherchemöglichkeit mit Hilfe eines Index der Vorzug gegeben.

Technische Aspekte (Markus Mazanec)[1]

Marktanalyse

Im Folgenden soll das Projekt aus Sicht der technischen Realisierung beleuchtet werden. Zuerst wurde versucht, ein geeignetes am Markt befindliches System ausfindig zu machen. Der Besuch zahlreicher Produktdemonstrationen, Verkaufs- und Beratungsgespräche führte allerdings zu einer raschen Ernüchterung: Bei dem Großteil der am Markt erhältlichen Software wurde man mit einer ungeeigneten grundsätzlichen Ausrichtung, nämlich Geschäftsanwendungen, d. h. Massenarchivierung von qualitativ ähnlichen (oder gleichartigen) Belegen o.ä. ohne tiefgehende inhaltliche Klassifikation konfrontiert und ebensowenig Erfassungsprobleme aufgrund ständig wechselnder Formate, Farben oder Vorlagenqualität bereiten. Derartige Systeme bieten keine Möglichkeit, eine Datenbasis zu erstellen, welche als Grundlage für wissenschaftliche Auswertungen und Folgeprojekte dienen kann. Viele der untersuchten Lösungen verfügten über mehr oder weniger flexible Möglichkeiten die Datenbankstruktur anzupassen. Diese oft gepriesene Flexibilität wird aber mit verringerter Zugriffsgeschwindigkeit bezahlt; ein Entscheidungskriterium bei der Suche nach geeigneten Systemen stellte aber das Verhältnis der durchschnittlich zu erwartenden Antwortzeit des Systems zu der anvisierten Größe des zu erfassenden Datenbestandes dar. Ein weiterer relevanter Faktor war der Preis – der Großteil der untersuchten Programmpakete war schlichtweg zu teuer.

Man entschied sich schließlich für eine Eigenentwicklung. Es sollte ein System geschaffen werden, das sich für die wissenschaftliche Archivierung und Recherche großer Bestände eignete. Da alle Rechte der Software beim Autor verblieben sind, konnte ein konkurrenzloses Preis-/Leistungsverhältnis erreicht werden.

Datenmodell

Das Datenmodell wurde in intensiver Zusammenarbeit mit MitarbeiterInnen der Stiftung erarbeitet. Es bildet die physikalische Bestandsstruktur ab: Dokumente werden in Mappen zusammengefaßt, Mappen liegen in Boxen, die in Regalen aufbewahrt werden. Als zentrales Element werden die Dokumente gesehen, denen,

---

[1] Geschäftsführer von M-IT Mazanec Informationstechnologie, Gesellschafter bei service.at EDV Dienstleistungen und Mitglied der ACM (Association for Computing Machinery). Für Anregungen und Fragen steht er gerne unter seiner e-mail-Adresse mazanec@acm.org zur Verfügung.

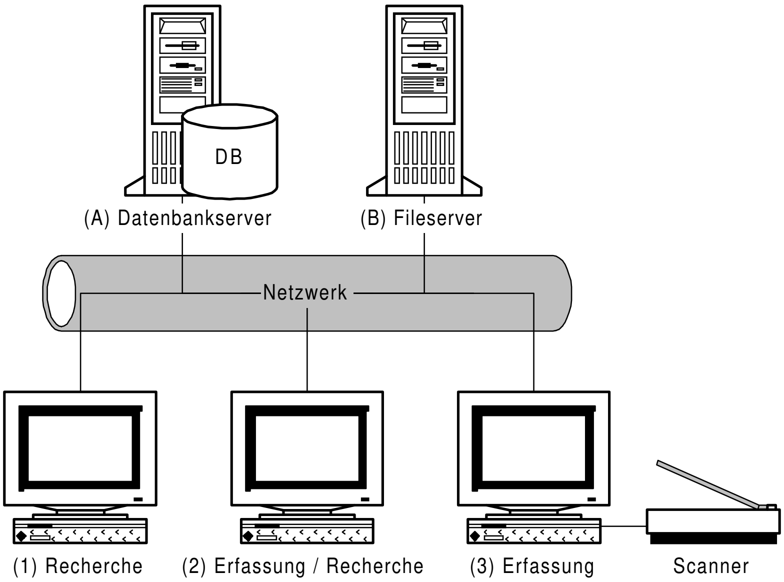
wie Mappen, beliebig viele Schlagworte aus verschiedenen Schlagwortklassen eines benutzerdefinierbaren Schlagwortkataloges zugewiesen werden können. Zusätzlich existiert ein Personenregister zu dem beliebig viele Verbindungen erstellt werden können. Jedes Dokument kann schließlich mit einer unbeschränkten Anzahl von Verknüpfungen zu Dateien (z. B. gescannten Bildern) versehen werden. Mappen wird ein Thema aus einem frei definierbaren Themenbaum zugewiesen (ein Thema kann einem anderen Thema untergeordnet sein). Die Entlehnung bzw. Verwendung von Mappen als auch Dokumenten (z. B. im Rahmen von Archivprojekten) kann mit dem System über das integrierte Bearbeitungsregister aufgezeichnet werden. Das Bearbeitungsregister stellt eine Verbindung zwischen der Benutzerverwaltung des Archivierungssystems *Archiopteryx* mit den Mappen und Dokumenten her. Über die Benutzerverwaltung werden hauptsächlich die Zugriffsrechte gesteuert, es können aber auch persönliche Daten wie Adresse, Firma usw. erfaßt werden.

Architektur

Eine für dieses Projekt wesentliche Eigenschaft des Architekturdesigns ist die Skalierbarkeit, da das System auch für sehr große Datenbestände Ergebnisse in kurzer Zeit liefern muß. Zusätzlich bringt ein umfangreicher Datenbestand die Anforderung einer auf mehrere Arbeitsplätze verteilten, simultanen Dateneingabe und Recherche mit sich. Diese beiden Kriterien waren maßgeblich für die Designentscheidung, ein Client-Server-System zu entwerfen, bei dem die Ablage der digitalisierten Objekte außerhalb der Datenbank erfolgt. In der Datenbank werden zu jedem Dokument Referenzen auf die abgelegten Dateien gespeichert. Es können beliebig viele Ablageorte, sogenannte „Volumina“, definiert werden. Ein Volumen entspricht im wesentlichen einem Verzeichnis auf einem Fileserver. Das Verzeichnis der Volumina wird zentral in der Datenbank verwaltet und kann auf der Arbeitsstation lokal adaptiert werden.

Diese Art der Dateiablage charakterisieren folgende Vorteile:

1. Die Datenbank wird nicht mit großen, schwer administrierbaren Datenobjekten (sogenannten „BLOBs“, binary large objects) vollgefüllt.
2. Das System kann jederzeit auf einfachste Weise um neue Ablageorte (Festplatten, Fileserver) erweitert werden. Durch die zentrale Administration ist dazu kein Eingriff auf den Arbeitsstationen erforderlich.
3. Der verteilte Ansatz entlastet den Datenbankserver und ermöglicht maximale Performance beim Dateizugriff. Durch die Entkopplung entsteht die Möglichkeit Serverbetriebssysteme, deren Stärken und Schwächen bekannt sind, gezielt einzusetzen (So wäre ein Datenbankserver unter Unix in Kombination mit einem Fileserver unter Netware eine denkbare Konfiguration).



Beispielkonfiguration:

PC 1 dient ausschließlich der Recherche, PC 2 wird zusätzlich für die Erfassung von Beschreibungsdaten genutzt und auf dem dritten PC werden Dokumente gescannt und katalogisiert. Der Datenbankserver A hält die Katalogdaten, auf dem Fileserver B liegen alle Faksimiles.

Die Minimalkonfiguration besteht aus einem Server und einer Arbeitsstation. In diesem Fall übernimmt der Datenbankserver auch die Funktion des Fileservers. Findet man damit bei gestiegener Benutzerzahl oder gewachsenem Datenbestand kein Auslangen mehr, können weitere Server hinzugefügt werden und die am Datenbankserver abgelegten Bilddateien verlagert werden.

Digitalisierung

Wie bereits erwähnt, werden Dokumente mit Dateien, wie z. B. Scans der Originalunterlagen, verknüpft und als Sammlung zusammengehöriger Objekte betrachtet [ROB94]. Die digitalisierten Daten werden auf einem Volumen abgelegt und können auf Knopfdruck eingesehen werden. Der Vorgang des Digitalisierens ist einer der kritischen Arbeitsabläufe für die/den ArchivarIn, bei dem der größte arbeitsablaufbedingte Zeitverlust droht. Ein TWAIN-Modul in *Archiopteryx* ermöglicht die nahtlose Integration von Peripheriegeräten, wie z. B. Scanner und Digitalkameras, die diesen Standard unterstützen [TWA98]. Um das Scannen mehrerer Objekte zu vereinfachen, wurde eine Operation vorgesehen, die einen wiederholten Scanvorgang teilautomatisiert. Der Datenimport ist nicht auf die TWAIN-Schnittstelle beschränkt, jede Systemdatei kann in dem Archiv abgelegt werden. Für diverse Bild- und Textformate wurde in das Programm ein

Anzeigemodul integriert, welches das komfortable Blättern in den Verknüpfungen jedes Dokuments erlaubt bzw. Zoom-, Rotier- und Druckfunktionen zur Verfügung stellt. Unbekannte Dateitypen werden identifiziert und, soweit verfügbar, über ein externes Programm dargestellt.

Hardware

Bei der Wahl der Hardware wurde das Hauptaugenmerk auf ein günstiges Preis/Leistungsverhältnis, Einsatz von Standards und Erweiterbarkeit gelegt. Die Entscheidung fiel zugunsten der IBM-kompatiblen PC-Architekur. Wie die dauerhafte physikalische Speicherung der Daten erfolgen soll, ist eine der ungelösten Fragen der digitalen Archivierung. Optische Medien ermöglichen Direktzugriff, sind kostengünstig aber fragil und stellen keine sichere Ablagemöglichkeit dar [TAN98]. Eine Alternative ist die Speicherung auf Festplatten in Kombination mit einer regelmäßigen Sicherung auf Band.

Die Installation im Österreichischen Staatsarchiv besteht aus einem Pentium II 300 MHz mit 256MB RAM und 18,2 GB UltraWide-SCSI Festplattenspeicher als Datenbank- und Fileserver. Es erfolgt eine regelmäßige vollautomatische Sicherung der Daten auf DDS3-Band. Gegen Stromschwankungen und -ausfälle ist der Server durch eine 700VA-USV gesichert. Ein nicht zu vernachlässigender Faktor ist die Temperaturentwicklung, die hochdrehende Serverfestplatten mit sich bringen. Drei Ventilatoren sorgen für ausreichende Luftzirkulation im Gehäuse. Für den selten Fall von Arbeiten an der Serverkonsole ist an den Server ein 15“ Monitor angeschlossen. Als Arbeitsstation dient ein Pentium 200 MHz mit 128MB RAM und 4 GB Festplattenspeicher. Eine CD-ReWriter ermöglicht den Datenexport auf optische Datenträger. Die Arbeitsstation verfügt über einen 21“ Monitor, der für optimale Arbeitsbedingungen beim Einscannen neuer Dokumente und Blättern im Datenbestand sorgt. Beide Computer sind über ein 10-Base-T Netzwerk verbunden.

Software

Die Serversoftware besteht aus der Datenbank-, Betriebssystemsoftware und Software für Dateidienste, soweit diese nicht im Betriebssystem integriert ist.

An die Datenbank wurden folgende Anforderungen gestellt:

I. transaktionsbasiertes Client-Server System
II. leistungsfähig
III. multiuser-tauglich
IV. SQL-92 Unterstützung [DIG92]
V. ODBC-Schnittstelle für genormten Datenzugriff [MIC96]
VI. niedrige Anschaffungs- (insbesonders keine Zusatzkosten für ODBC-Clientsoftware) und Instandhaltungskosten
VII. wartungsarm (für stand-alone Betrieb geeignet)
VIII. unterbrechungsfreier 24 Stunden / 7 Tage Betrieb

Es wurden mehrere Datenbanken unter Windows NT Server und Linux intensiven Vergleichen unterzogen. Man entschied sich schließlich für die Linux-Portierung von Adabas D 6.1 der Software AG, da dieses System alle oben genannten Anforderungen erfüllte und zudem die Möglichkeit der Fernwartung bot.

Die Wahl des Betriebssystems der Arbeitsstation ging an Windows NT Workstation Version 4 von Microsoft. Im Gegensatz zur Serveradministration musste die Bedienung der Arbeitsstation auch von Nicht-IT-ExpertInnen rasch erlernbar sein. Diesem Bestreben kommt der Einsatz verbreiteter und bekannter Technologie entgegen. Die etwas kostengünstigere Alternative Windows 95/98 wurde aus Stabilitäts- und Performancegründen ausgeschieden.

Der 32bit-Client des *Archiopteryx*, der für die/den ArchivarIn das Fenster zum Datenbestand darstellt, wird in dem Teil dieser Gemeinschaftsarbeit, der sich mit dem praktischen Ablauf der Erfassung beschäftigt, detailliert beschrieben.

Das Ablegen gescannter Dokumente und anderer Dateien auf dem Server ermöglicht der Samba-Dämon, durch den der SMB-Dateizugriff auf Linux-Rechner möglich wird [INT88].

Adabas D und Samba wurden unter Linux Kernel Version 2.0.32 installiert. Diese Kombination hat sich als ausgesprochen stabil und zuverlässig erwiesen. Bis zum Zeitpunkt des Verfassens dieses Artikels, war, bei einer Betriebsdauer (uptime) von mehr als einem Jahr, kein einziger Neustart oder administrativer Eingriff seitens einer/s Technikerin/s erforderlich. Die Rotation der Sicherungsmedien beschränkt sich auf das Tauschen der Bänder und kann problemlos auch von BenutzerInnen ohne Spezialkenntnisse vorgenommen werden.

Datenformate

Sobald man sich mit der dauerhaften Archivierung von Daten beschäftigt, stellt sich über kurz oder lang die Frage nach der einzusetzenden Technologie. Schließlich nutzt es niemandem, in zehn Jahren über einen Datenbestand zu verfügen, für den keine geeigneten Zugriffswerkzeuge mehr existieren [ROT95][BIK93]. Leider gibt es keine Garantie, daß die Standards von heute auch die Technologien von morgen sind. Allerdings bringt der Einsatz standardisierter Datenformate in den meisten Fällen die Möglichkeit der Datenmigration mit sich [SMI98]. Bei proprietären Herstellernormen, die ihre ErfinderInnen überleben, hat man diese Option meistens nicht.

Die wichtigste und verbreitetste Datenbanksprache ist zur Zeit SQL-92 (Structured Query Language). Um rasche und problemlose Portierbarkeit auf andere Datenbankplattformen zu ermöglichen, fand nur eine Untermenge von SQL-92 bei diesem Projekt Verwendung. Der *Archiopteryx*-Client kommuniziert über die ebenfalls weit verbreitete ODBC-Schnittstelle (Open Database Connectivity) mit der Datenbank. Zusätzlich ist es möglich die Datenbank unter Umgehung des *Archiopteryx*-Clients direkt anzusprechen.

Verknüpfte Dateien (Bilder, Videos, Tondokumente, Textdateien usw.) werden entweder in ihrem ursprünglichen Format auf ein Volumen transferiert oder auf Wunsch, falls eine Konvertierung möglich ist, in ein anderes Format übertragen. Da Volumina aus Sicht des Dateisystems nichts anderes als Verzeichnisse sind und die Dateien selbst (von einer allfälligen Konvertierung von einem Standardformat in ein anderes abgesehen) nicht verändert werden, besteht auch hier keine Gefahr auf eine Einbahnlösung gesetzt zu haben.

Recherche

Es ist nicht zweckmäßig, von den BenutzerInnen tiefergehende EDV-Kenntnisse für die Bedienung eines Archivierungssystems zu verlangen. Während die Erfassung und simple Recherche (z. B. Abfrage von Listen) problemlos einfach bedienbar gemacht werden können, stellen komplexe Abfragen, die logische Verknüpfungen über mehrere Tabellen enthalten, die/den SystemdesignerIn vor eine größere Herausforderung.

BenutzerInnen von *Archiopteryx* werden keinerlei SQL-Kenntnisse abverlangt. Es wurde ein graphischer Editor integriert, der das Erstellen von Abfragen auf die Eingabe einzelner Kriterien und deren logischer Verknüpfung („und“, „oder“) reduziert. Bei der Kriterieneingabe selbst gibt das System die zulässigen Optionen vor, die mit Maus oder Tastatur ausgewählt werden können. Nachdem die/der BenutzerIn den Auftrag zum Ausführen seiner Abfrage erteilt hat, übersetzt der integrierte Generator die eingegebenen Kriterien in eine SQL-Abfrage. Gleichzeitig findet eine Optimierung statt, die sicherstellt, dass nur die tatsächlich benötigten Tabellen der Datenbank einbezogen werden.

Ausblick

Der ausgezeichnete Erfolg des bisherigen Projektverlaufes hat die Entwickler darin bestärkt, *Archiopteryx* auszubauen und weiterzuentwickeln. Zur Zeit wird an einem Servermodul gearbeitet, das die Integration verschiedener Volltextdatenbanken ermöglichen wird. Mit der Fertigstellung ist noch im vierten Quartal 1999 zu rechnen. Für das erste Halbjahr 2000 ist die Freigabe des Webinterface geplant, mit dessen Hilfe archivierte Daten aus dem World Wide Web kontrolliert abgefragt werden können. Dadurch kann ein breiteres Publikum auf flexiblere Art und Weise angesprochen werden und das Angebot des Archivs attraktiver gestaltet werden [HOR98]. Der Kern des Systems ist bereits auf diese Erweiterungen vorbereitet. Auch bei diesen neuen Modulen wird das Konzept der Benutzerfreundlichkeit und einfachen Bedienbarkeit konsequente Umsetzung finden.

Abschließende Bemerkungen

Im Rahmen dieses Digitalisierungsprojektes konnten einige innovative Ideen entwickelt und realisiert werden. Aufgrund einer klaren Zieldefinition zu Projektbeginn und umfangreicher Markterhebungen war es möglich eine technisch adäquate Lösung zu finden, die für große Datenbestände skaliert, zukunftsorientiert ist und sich durch einfache Bedienung sowie hohe Betriebssicherheit auszeichnet. Aktuelle Informationen zu *Archiopteryx* finden sich im World Wide Web unter http://www.archiopteryx.com/.

Referenzen

[DIG92] Digital Equipment Corporation: Database Language SQL (Second Informal Review Draft) ISO/IEC 9075:1992, July 1992.

[BIK93] Bikson, T. K. – Frinking, E.: Preserving the Present: Toward Viable Electronic Records. Den Haag 1993.

[HOR98] Horsman, Peter: From Users to Visitors: Information and Communication Technology Strategies for Archives. In: Proceedings of the International Conferences of the Round Table on Archives XXXIII CITRA, Paris 1998, pp. 131–134.

[INT88] Intel Corporation – Microsoft Corporation: Microsoft Networks/OpenNET File Sharing Protocol. November 1988.

[MIC96] Microsoft Corporation: ODBC Programmer’s Reference, 1996.

[ROB94] Roberts, D., Defining Electronic Records, Documents and Data. In: Archives and Manuscripts 22/1 (1994), pp. 14–16.

[ROT95] Rothenberg, Jeff: Ensuring the Longevity of Digital Documents. In: Scientific American 272/1 (1995), pp. 24–29.

[SMI98] Smith, Abby: Digital Archiving Models. In: Proceedings of the International Conferences of the Round Table on Archives XXXIII CITRA, Paris 1998, pp. 191–194.

[TAN98] Tangley, L.: Whoops, there goes another CD-ROM. In: US News & World Report: 67–68 (February 16, 1998).

[TWA98] TWAIN Working Group Committee: TWAIN Specification Version 1.8, October 1998.

Rothenberg, Jeff: Avoiding a Viable Technical Foundation for Digital Preservation. A Report to the Council on Library and Informations Resources. Washington DC 1999 [www.rlg.org].

Hedström, Margaret – Montgomery, Sheon: Digital Preservation. Needs and requirements in RGL [Research Libraries Group] Member Institutions. Mountain View CA 1999 [www.rlg.org].

Praktische Erfahrungen bei der Digitalisierung mit *Archiopteryx* (Maria Wirth)

In der Praxis wird das Programm *Archiopteryx* derzeit zur digitalen Erfassung zweier Archivbestände angewendet: zum einen handelt es sich dabei um die Kabinettsunterlagen von Bundeskanzler Dr. Fred Sinowatz, die in rund 114 Archivboxen eine umfassende Korrespondenzsammlung und eine Vielzahl an Materialien zur Politik der Kleinen Koalition enthalten; zum anderen ist der umfangreiche Nachlaß des ehemaligen Justizministers Dr. Christian Broda Gegenstand der Erfassung, welcher neben biographischen Unterlagen insbesondere eine Vielzahl an Dokumenten zur sozialistischen Rechts-, Medien- und Menschenrechtspolitk beinhaltet.

Begonnen wurde im Sommer 1998 zunächst mit der Digitalisierung der Kabinettsunterlagen von Bundeskanzler Dr. Fred Sinowatz, wobei bis Ende August 1999 21 der angeführten 114 Boxen zu folgenden Themen erfaßt wurden: Ankauf von Abfangjägern, Auseinandersetzungen um Hannes Androsch, Atom, Auslandsbesuche und Besuche ausländischer Staatsgäste, Bundespräsidentenwahl 1986, Einladungen, Innenpolitik, Rundfunk und ORF, Weinskandal 1985. Angelegt wurden hierfür 246 Mappen mit 877 Dokumenten, wobei sowohl die Mappen als auch die Dokumente einen Umfang von einigen wenigen bis mehrere hundert Seiten haben können. Verknüpft, d. h. gescannt, wurden dabei die meisten Dokumente in ihrem vollen Umfang; in einigen Fällen jedoch wurden diese aufgrund des großen Umfangs des Bestandes nur auszugsweise erfaßt. So wurden vorwiegend die im Bestand zu vielen Themen vorkommenden Pressespiegel nur teilweise gescannt und auch aus der Vielzahl an Einladungen für Bundeskanzler Sinowatz sowie der umfangreichen Bürgerkorrespondenz zum Kernkraftwerk Zwentendorf lediglich eine Auswahl getroffen. Beim Vorliegen von Sekundärliteratur wie etwa zum Thema Atomenergie wurde hingegen lediglich auf deren Vorhandensein verwiesen bzw. hieraus nur relevante Stellen verknüpft, während Primärunterlagen wie beispielsweise vorbereitende Materialien für Regierungsklausuren sowie Reden, Manuskripte und Notizen vollständig gescannt wurden.

Parallel dazu wurde im Februar 1999 auch mit der digitalen Erfassung des Nachlasses von Dr. Christian Broda begonnen, der – mit Ausnahme weniger, nachträglich an das Österreichische Staatarchiv übergebenen Kartons – von der Handschriftensammlung der Österreichischen Nationalbibliothek aufbewahrt wird. Erfaßt wurden daher bisher nur jene vier Kartons, die nachträglich zu dem durch Dr. Bela Rásky bereits seinerzeit indizierten Nachlaß hinzukamen. Dabei wurden nach der Ordnung dieses Bestandteiles 190 Mappen zu verschiedenen Themenbereichen wie etwa der Asylpolitik, der Justiz, den Menschenrechten, der Strafrechtsreform, dem Parteiprogramm von 1958 und der Wohnungspolitik angelegt. Enthalten sind hierin 876 Dokumente, wobei auch hier die Mappen und Dokumente einen Umfang zwischen einer bis mehrere hundert Seiten haben können. Erforderlich war hier im

Gegensatz zu den Kabinettsunterlagen von Bundeskanzler Dr. Fred Sinowatz jedoch zumeist eine Erfassung in kleinen Objekten, da dieser Nachlaßteil großteils unzusammenhängend ist und insbesondere eine Menge an vereinzelten Zeitungsberichten umfaßt. Verknüpft wurden die einzelnen Dokumente dabei großteils vollständig, wenn auch hier beim Vorhandensein einer Vielzahl von Zeitungsberichten wie etwa zum Fall Habsburg eine Auswahl getroffen wurde. Darüber hinaus wurde beim Vorliegen von persönlichen Unterlagen aus Datenschutzgründen zumeist nur auf deren Existenz verwiesen und die einzelnen Objekte nicht gescannt.

Konkret gestaltet hat sich der Erfassungsprozeß mit *Archiopteryx* als einem sehr benutzerfreundlichen Programm, dessen Anwendung keine besonderen Vorkenntnisse technischer Natur oder ein spezielles Computer-Know-how erfordert, dabei folgendermaßen: entsprechend dem „Regal-Box-Mappe-Dokument"-System, auf dem das Programm in Abhängigkeit der einzelnen Archivobjekte aufbaut, müssen die erfaßten Dokumente zuerst einem Regal zugeordnet werden, in welches dann die Boxen mit den hierin befindlichen Mappen und Dokumenten, abgelegt werden können.

Dabei wurde sowohl für das Sinowatz-Archiv als auch für das Broda-Archiv lediglich ein einziges virtuelles Regal geschaffen, in das dann alle vorhandenen Boxen gestellt wurden. Wirkliche Regalnummern können jederzeit nachträglich vergeben werden. Den Boxen wird hiernach eine kurze Bezeichnung gegeben, unter welcher sie dann in alphabetischer Reihenfolge in der entstehenden Boxenliste, die für eine spätere Recherche relevant ist, aufscheinen.

Detaillierter als die einzelnen Boxen können aufgrund der verschiedenen vorgesehenen Beschreibungsinstrumente die einzelnen Mappen bezeichnet werden. So kann diesen nicht nur ein Titel gegeben werden, unter welchem sie dann – wie auch die Boxen – in einer entsprechenden Liste aufgeführt sind, sondern es ist hier auch möglich ihren Umfang sowie den Zeitraum, über den sich die darin befindlichen Dokumente erstrecken, anzugeben. Vor allem besteht hier die Möglichkeit, einen Kommentar zu verfassen, was sich in der Praxis als besonders vorteilhaft erwiesen hat. Hierdurch kann der Inhalt einer Mappe näher beschrieben werden, beispielsweise, welche Dokumente sich in einer Mappe zu einem bestimmten Thema befinden. Es kann aber auch auf Mappen und Dokumente mit einem ähnlichen Inhalt verwiesen werden oder schlicht angeführt werden, ob in einer Mappe Dokumente mehrfach vorkommen. Diese Kommentiermöglichkeit stellt eine wichtige Hilfestellung zur Unterscheidung von verwandten Mappen für die spätere Recherche dar, wobei es sich jedoch generell empfiehlt, die Mappe bereits möglichst genau zu betiteln.

Des weiteren wird jede Mappe auch einem Thema zugeordnet und beschlagwortet werden, wodurch eine weitere Option für spätere Recherchen geschaffen wird, entsteht doch durch die Erfassung verschiedener Themen, die wiederum in verschiedene Unterthemen gegliedert sein können, im Laufe des Arbeitsprozesses ein Themenbaum, in dem – wie auszuführen ist – mittels eines Datenbank-Explorers gesucht werden kann. Konkret erfaßt wurden dabei, was die Kabinettsunterlagen von Bundeskanzler Dr. Fred Sinowatz betrifft, bis dato 45 solcher Themen wie etwa „Abrüstung", „Priviliegienabbau" und „Verkehrspolitik", die wiederum größeren Themenbereichen wie „Innen- oder Außenpolitik" zugewiesen wurden. Bei der Digitalisierung des Broda-Nachlasses wurden hingegen 56 Themen wie „Ostblock" und „Sozialrecht" aufgelistet, die größeren Themenbereichen wie „Kommunismus" oder „Rechtspolitik" zugehören können.

Was die Beschlagwortung betrifft, so wurde durch das Zusammenführen des Thesaurus der Friedensbibliothek Stadt Schlaining und der im Index zum Nachlaß von Dr. Christian Broda vorhandenen Schlagwortliste ein umfassendes Suchinstrument geschaffen, das bei der näher zu beschreibenden „Kriterien-Suche" verwendet werden kann. Dieser Thesaurus, der jederzeit problemlos erweiter- und korrigierbar ist, umfaßt für das Sinowatz-Archiv derzeit bereits 3 229 Schlagworte, für das Broda-Archiv wurden 3 201 Schlagworte vergeben, aus denen für eine exakte thematische Zuordnung zumeist mehrere Schlagworte ausgewählt wurden. Zugewiesen sind diese Schlagworte zudem auch bestimmten Schlagwortklassen, die insbesondere als Hilfestellung für die Suche im Nachlaß von Christian Broda geschaffen wurden, wo sich häufig Rechtsbezeichnungen – wie etwa „Zwischenzeitengesetz" oder „Divergenzgesetz" finden, die für juristische Laien nur schwer zuzuordnen sind. Die bis heute bestehenden 132 Schlagwortklassen werden dabei ständig erweitert, wobei sowohl bereits bestehende als auch neu geschaffene Schlagworte diesen zugeordnet werden.

Bei der Erfassung von Dokumenten, die wiederum in einer der Recherche dienenden Dokumentenliste verzeichnet sind und bei der sie umfassenden Mappe, die in einem Verknüpfungsfenster aufscheinen, bestehen dann ebenfalls eine Reihe von Beschreibungsinstrumenten. So können auch sie bezeichnet und kommentiert werden, wobei auch hier insbesondere die Kommentiermöglichkeit vielfach verwendet wurde. So kann hier angegeben werden, ob ein Dokument vollständig gescannt wurde oder der Kontext eines Dokuments, etwa der Ort und Anlaß zu dem eine Rede gehalten wurde, näher bezeichnet werden. So wie bei der Mappenbeschreibung kann auch hier der Umfang eines Dokuments angeführt werden, wobei hinsichtlich der zeitlichen Einordnung eines Dokuments nun jeweils optional ein genaues Entstehungsdatum oder ein Zeitraum, auf den sich die Dokumente beziehen, eingegeben werden kann. Ist es in bestimmten Fällen – wie etwa beim Vorliegen einer Aktennotiz - doch nur möglich, ein Datum anzugeben, während bei anderen Objekten – etwa beim Vorhandensein einer Pressedokumentation – eine Zeitspanne angegeben werden muß.

Neu hinzukommt bei der Erfassung eines Dokuments jedoch die Angabe einer Dokumentart: aus einer angelegten Liste möglicher Dokumentarten wie z. B.

„Akten", „Notizen", „Interviews" können die auf den Charakter eines Dokumentes zutreffenden Bezeichnungn gewählt und zugeordnet werden. Aufgelistet wurden dabei für die Kabinettsunterlagen von Bundeskanzler Dr. Sinowatz 41 und für den Nachlaß von Christian Broda 33 unterschiedliche Dokumentbezeichnungen, wobei es sich in der Praxis als vorteilhaft erwiesen hat, diese Dokumentarten schon zu Beginn der Anlegung eines Archivs möglichst konkret zu erfassen. D. h. neu auftauchende Dokumentarten sollten, auch wenn angenommen werden kann, daß diese eher selten vorkommen, bereits bei erstmaligem Aufscheinen angeführt werden, da sonst nach mehrmaligem Vorhandensein nachträgliche Korrekturen nötig sind, die insbesondere dann sehr zeitintensiv sein können, wenn der zu erfassende Bestand noch nicht indiziert ist.

Zusätzlich zu den bereits genannten Beschreibungsinstrumenten einer Mappe besteht bei der Erfassung die Möglichkeit, dem jeweiligen Dokument neben Schlagworten auch Personen, auf die in dem Dokument Bezug genommen wird, aus einem eigenen Personenthesaurus zuzuweisen. Es kann also, wenn ein bestimmtes Dokument etwa eine Person behandelt oder von dieser stammt, diese speziell vermerkt werden, wodurch die zugewiesenen Personen dann ähnlich den Schlagworten in einem Verknüpfungsfenster aufscheinen. Dabei besteht nicht nur die Möglichkeit der Zuordnung des Namens, sondern es können auch Geschlecht und Herkunftsland sowie wahlweise ergänzende personen- oder funktionsbezogene (z. B. Bundeskanzler, Finanzminister) angeführt werden. Die Personenliste ist also wie auch die Schlagwortliste ständig erweiterbar und umfaßt gegenwärtig für die Kabinettsunterlagen von Dr. Sinowatz 139 und für den Nachlaß von Christian Broda 776 Einträge.

Der Scanvorgang, d. h. die digitale Erfassung erfolgt in der Regel erst nach Abschluß der vollständigen Beschreibung eines Dokuments an Hand der eben beschriebenen Möglichkeiten, wobei die Verknüpfung zwischen beiden direkt aus *Archiopteryx* erfolgt, d. h. der Scanvorgang der Datenerfassung unmittelbar angeschlossen ist. Dabei bietet das Programm mit der Wahlmöglichkeit zwischen Einfach- und Mehrfachscannen eine optimale Anpassung an den jeweiligen Umfang eines Dokuments. So ist auch beim Scannen eines noch so umfangreichen Dokuments mittels eines Mehrfachscanvorgangs nur eine einmalige Eingabe eines Samples notwendig, was den Scanvorgang enorm beschleunigt. Zugleich wird aber auch hier – wie beim Einfachscannen – für jede einzelne Seite ein Voransichtsbild erstellt, das – vor der endgültigen Ablage des Dokuments – entsprechend eingerichtet und optimiert (Helligkeit, Schärfe, Vollständigkeit) werden kann. So besteht insbesondere eine Auswahl an bestimmten Bildtypen wie „Schwarz-weiß-Zeichnung", „Schwarz-weiß-Photo", „Farbzeichnung", etc. sowie die Möglichkeit, die Farbstärke einer Seite zu bestimmen und durch eine Umrandung unwesentliche Bildteile wegzulassen, wodurch es möglich ist, die verschiedensten Dokumentarten in bestmöglicher Qualität zu erfassen. Es ist für den Scanvorgang also unerheblich, ob es sich um Handschriften, Druckschriften oder Zeitungsberichte mit Photos handelt, wie sie sowohl Sinowatz-Archiv als auch im Broda-Archiv vorkommen.

Hinsichtlich der Zeitdauer, die ein Scanvorgang in Anspruch nimmt, ist jedoch die Qualität des Originals entscheidend, wobei dieser um so kürzer ist je besser ein Dokument erhalten ist. So hat das Scannen von Originalzeitungsausschnitten sowie Durchschlägen aus den fünfziger Jahren aus dem Nachlaß von Christian Broda zumeist mehr Zeit in Anspruch genommen als die Verknüpfung von Dokumenten aus den Kabinettsunterlagen von Bundeskanzler Dr. Fred Sinowatz. Vereinfacht gilt also, daß das Scannen von jüngeren Dokumenten weniger Zeit in Anspruch nimmt als jenes von älteren, wenn auch diese in guter Qualität wiedergegeben werden können, hierfür aber häufig erst verschiedene Kombinationen von Bildtypen und Farbstärken ausprobiert werden müssen. Der vollständige Scanvorgang einer einzelnen Seite kann dabei 40 Sekunden bis einige Minuten dauern, während in einer Stunde inklusive der Objektbeschreibungen im Durchschnitt zwischen 30 und 60 Seiten – jeweils abhängig von der Anzahl der anzulegenden Mappen und Dokumente und der Qualität der einzelnen Dokumente – gescannt wurden.

Insgesamt erlaubt das Programm somit, sowohl was seine Scanfunktion als auch die umfassende Dokumentenbeschreibung betrifft, eine nahezu optimale Erfassung eines Aktenbestandes, die sich bei der Recherche nach bestimmten Objekten als vorteilhaft erweist. Für die Dokumentenrecherche in einem Bestand bietet das Programm *Archiopteryx* drei große Recherchewerkzeuge mit denen anhand der bekannten Fakten die gewünschten Informationen in einem Bestand aufgefunden werden können.

Dabei handelt es sich zum einen um den bereits angeführten Datenbank-Explorer, der die hierarchische Struktur des Bestandes visuell wiedergibt. D. h. hier kann der aufgebaute Themenbaum mit seinen Haupt- und Unterthemen sowie den dazugehörenden Mappen und Dokumenten betrachtet werden, wobei zunächst nur die Hauptthemen aufscheinen und per Mausklick die unteren Strukturebenen angezeigt werden, die dann durchgeblättert werden können. Es handelt sich hier also vorwiegend um ein Suchinstrument, das eine rasche Orientierung in einem Bestand ermöglicht und selbst Personen, die ein Archiv noch nicht gut kennen, einen Überblick hierüber geben kann.

Zum anderen kann anhand der verschiedenen Objektlisten wie etwa der Boxen-, Mappen- und Dokumentenliste in einem Bestand recherchiert werden, wobei hier auch die Möglichkeit besteht, aus den bereits erwähnten Verknüpfungsfenstern weitere Detailanzeigen zu betrachten. So kann etwa durch ihr Anklicken eine bestimmte Mappe aus der Mappenliste gewählt werden und die im Verknüpfungsfenster angezeigten Dokumente angesehen werden oder aus der Dokumentenliste ein Dokument betrachtet und die näheren Angaben zu einer zugewiesenen Person aufgerufen werden. Diese Suchfunktion richtet sich im Gegensatz zur Recherche mittels Datenbank-Explorer also eher an Personen, die einen Bestand bereits besser kennen, wobei die einzelnen Listen nach einer gewissen Menge an Einträgen jedoch unübersichtlich werden und die Suche hierdurch zeitintensiver wird.

Als drittes Rechercheinstrument bietet sich dann deshalb die sogenannt „Kriterien-Suche" an, die das mächtigste Recherchewerkzeug von *Archiopteryx* darstellt. Dabei ist es hier möglich, die einzelnen Objektlisten nach bestimmten Kriterien oder Vergleichswerten – wie Schlagworten, Personen, Dokumentearten, Entstehungsdaten etc. – zu durchsuchen, wobei hier auch mehrere Abfragen zusammengestellt und auf den Datenbestand angewendet werden können. D. h. es ist nicht nur möglich, alle Mappen oder Dokumente zu einem bestimmten Schlagwort bzw. zu einer bestimmten Person abzurufen oder nach bestimmten Dokumentarten in einem Bestand zu suchen, sondern es besteht auch die Möglichkeit nach Objekten zu suchen, auf die verschiedene Kriterien zutreffen. So kann etwa in einem Suchvorgang nach Mappen gesucht werden, die einem bestimmten Schlagwort – wie etwa „AKH" – zugeordnet wurden und die Dokumente mit einer bestimmten Bezeichnung – wie etwa „Krankenhaus" – enthalten oder sich in einer bestimmten Box – etwa „AKH-Presseberichte" befinden. Insbesondere wenn nur wenige oder ungenaue Angaben zu einem Objekt vorliegen, eignet sich also dieses Suchinstrument zur Recherche, wobei natürlich auch hier beim Vorliegen von Suchergebnissen weitere Detailanzeigen zu bestimmten Objekten abgefragt werden können.

Hat man dann das oder die Objekte, nach denen mittels der angeführten Rechercheinstrumente gesucht wurde, gefunden, ist es rasch möglich die hierzu gescannten Seiten zu betrachten. So können die einzelnen Seiten per Mausklick durchgeblättert und wahlweise in ihrem vollen Umfang oder nur auszugsweise angesehen werden: damit beginnt die eigentliche wissenschaftliche Arbeit.

Projektleitung

Hon.-Prof. Dr. Lorenz Mikoletzky
Generaldirektor

Österreichisches Staatsarchiv
Nottendorfergasse 2
A-1030 Wien
email: gdpost@oesta.gv.at
http://www.oesta.gv.at

Projektausführung

Univ.-Doz. DDr. Oliver Rathkolb,
Dr. Theodor Venus , Mag. Maria Wirth

Stiftung Bruno Kreisky Archiv
Rechte Wienzeile 97
A-1050 Vienna
e-mail: archive@kreisky.vienna.at
http://www.members.vienna.at/kreisky/