Automatisierte Erstellung bibliographischer Informationen zu lokal gespeicherten Internet Dokumenten. Ein Projekt zum Aufbau eines Archivs elektronischer Pressedienste von Parteien und Gewerkschaften.

Ziel und Umfang des Projektes

Im Rahmen des im Frühjahr 2001 begonnenen Projekts werden Pressemitteilungen von Parteien und Gewerkschaften gesammelt, die im World Wide Web angeboten werden. Auf nationaler Ebene werden Pressemitteilungen der im Deutschen Bundestag vertretenen Parteien gesammelt. Im Bereich der deutschen Parteien liegt ein besonderer Schwerpunkt auf Pressemitteilungen der Sozialdemokratischen Partei Deutschlands (SPD). Hier werden Mitteilungen auch auf der Ebene der Bundesländer archiviert. Hinzu kommen die Pressemitteilungen der SPD-Fraktionen in den Landesparlamenten. Mitteilungen deutscher Gewerkschaften werden bis auf Landesbezirks-Ebene gesammelt.

Auf europäischer Ebene werden Pressemitteilungen von Mitgliedsparteien der Sozialistischen Internationale sowie von nationalen Dachgewerkschaften archiviert. Hinzu kommen Pressemitteilungen von internationalen Berufssekretariaten und anderen internationalen Organisationen der Arbeiterbewegung.

Zur Zeit werden im Rahmen des Projekts ca. 100 Pressedienste laufend archiviert. Der Schwerpunkt liegt noch im nationalen Bereich. Das Projekt wird jedoch systematisch auf den europäischen und internationalen Bereich ausgeweitet.

Archivierung von Pressediensten – eine sinnvolle Ergänzung der konventionellen Sammeltätigkeit

Die Entscheidung, systematisch Pressedienste von Parteien und Gewerkschaften zu archivieren basiert auf einer Reihe von Gesichtspunkten.

1. Die systematische Archivierung von Pressediensten ergänzt das von der deutschen Forschungsgemeinschaft geförderte Projekt zur Sammlung von grauer Literatur ausländischer Parteien und Gewerkschaften, das bisher hauptsächlich im Rahmen von Beschaffungsreisen gepflegt wurde. Es trägt der Tatsache Rechnung, dass das Internet als Quelle relevanter Dokumente zunehmend größere Bedeutung gewinnt.
2. Die Bibliothek der Friedrich-Ebert-Stiftung und das Archiv der Sozialen Demokratie sammeln bereits bisher Pressedienste in Papierform. Diese Sammlungen werden durch das neue Projekt ergänzt und weiter geführt.
3. Im Rahmen zweier DFG geförderter Projekte wurden die Pressedienste der SPD bis zum Jahre 1998 retrodigitalisiert. Das hier vorgestellte Projekt stellt u.a. die laufende Archivierung neuerer SPD Pressemitteilungen sicher.
4. Bei Pressediensten handelt es sich um eine Quellengattung, die für die Präsentation im Internet geradezu prädestiniert ist und die folglich von einer Vielzahl von Organisationen auf diesem Wege der Öffentlichkeit zugänglich gemacht werden.
5. Es gibt für den im Sammelprofil des Projekts definierten Bereich teilweise im Internet zugängliche Archive einzelner Organisationen. Es existieren aber keine kumulierten Archive, in denen die Mitteilungen unterschiedlicher Organisationen gesammelt werden.

Durch den nun möglichen Vergleich von Standpunkten und Diskussionsansätzen entsteht ein erheblicher informatorischer Mehrwert für den Nutzer eines solchen Angebots.

6. Bei Pressediensten gibt es keine Urheberrechtsprobleme bei der Redistribution im Internet, da sie explizit zur Verbreitung in der Öffentlichkeit bestimmt sind. Sie enthalten überwiegend kein Bildmaterial, für das die Urheberrechtsproblematik ja noch wesentlich komplexer ist, als für Informationen in Textform.

Konzeptionelle Überlegungen zur Erschließung der archivierten Materials

Der naheliegendste Ansatz zur Erschließung der archivierten Dokumente ist die Indexierung mittels eines Volltextretrieval-Programms. Diese Suchmöglichkeit wird im Rahmen des Projekts auch angeboten, jedoch als Ergänzung zur Suche mit Hilfe von Metadaten oder – bibliothekarisch gesprochen – Titelaufnahmen.

Bestimmte Fragestellungen lassen sich mit Hilfe von Volltextrecherchen nicht beantworten. Insbesondere die Eingrenzung von Suchanfragen auf bestimmte Zeiträume ist mit aktuell verfügbaren Volltextretrieval-Systemen, die den finanziellen Rahmen der Bibliothek nicht sprengen, nicht realisierbar.

So entstand der Ansatz, Metadaten der archivierten Dokumente zu erfassen, um das leistungsfähige Katalogprogramm der Bibliothek nutzen zu können. Wegen der Menge der gesicherten Dokumente verbietet sich jedoch eine Erstellung von Titelaufnahmen per Hand auf Grund der angespannten Personallage der Bibliothek von selbst.

Es wurde ein Verfahren entwickelt, bei dem die archivierten Pressemitteilungen mit Hilfe von in PERL geschriebenen Programmen katalogisiert werden. Dieses Verfahren beruht auf der Tatsache, dass im World Wide Web Pressedienste fast ausschließlich in Form von Übersichtslisten dargestellt werden, einer Präsentationsform, die sich geradezu anbietet. Vielfach handelt es sich um dynamisch erzeugte HTML-Seiten, die auf der Basis von Datenbankabfragen generiert werden.

Diese geordnete Darstellungsform bedingt einen HTML-Quellcode, der ebenfalls entsprechend regelmäßig strukturiert ist. Auf dieser Basis wird es möglich, mit Hilfe von PERL Programmen die für die Erstellung von Titelaufnahmen erforderlichen Informationen aus dem HTML Quellcode der Übersichtsseiten zu extrahieren. Diese einmal extrahierten Informationen können nun - ebenfalls mit Hilfe von PERL Programmen - in ein strukturiertes Format überführt und mit Hilfe einer Datenbank verwaltet werden. Die Bibliothek der Friedrich-Ebert-Stiftung benutzt zu diesem Zweck ihre Bibliothekssoftware ALLEGRO. Es wäre aber auch möglich die erzeugten Datensätze so zu modifizieren, dass sie mit relationalen Datenbanken, wie beispielsweise ORACLE verwaltet werden können. Diese automatisch erzeugten Datensätze mit Metadaten werden über die jeweiligen Datenbank spezifischen Update-Routinen in die Datenbank übernommen.

Von der Theorie zur Praxis – die Umsetzung des Konzepts

Die Programmierarbeiten, die die Basis der Umsetzung bilden, werden in der Bibliothek selbst durchgeführt. Ein bedeutender Nachteil des entwickelten Konzepts ist die Tatsache, dass für jeden archivierten Pressedienst ein eigenes Unterprogramm in PERL erstellt wird, das auf das beim jeweiligen Pressedienst verwendete Layout angepasst werden muss. Bei Layoutänderungen muss dieses Unterprogramm jeweils modifiziert werden.

Auf der Basis dieser Vorarbeiten wird die eigentliche Archivierung von einer studentischen Hilfskraft durchgeführt.

Zunächst werden die jeweiligen Übersichtslisten im WWW aufgesucht und unter einem, für jeden Pressedienst normierten Dateinamen lokal auf der Festplatte abgespeichert. Bei Datenbank gestützter Generierung von Übersichtslisten müssen zunächst die optimalen Aufrufparameter ermittelt werden, die ein möglichst effizientes Arbeiten erlauben. So sind entsprechende WWW Angebote vielfach standardmäßig so konfiguriert, dass sie pro Aufruf Listen mit nur wenigen Pressemitteilungen übermitteln. Eine Veränderung der Aufrufparameter, durch die beispielsweise Listen mit Hunderten von Pressemitteilungen erzeugt werden, erhöht die Effizienz des Verfahrens erheblich.

Nach lokaler Abspeicherung der im WWW verfügbaren Übersichtslisten, wird das PERL Programm aufgerufen und erzeugt die Titelaufnahmen. Die Integration dieser Aufnahmen in die Datenbank erfolgt ebenfalls im Batchverfahren. Teil dieses Vorgangs sind Zeichensatz Konvertierungen, die ein nicht zu unterschätzendes Detailproblem darstellen. Es gibt vielfältige Möglichkeiten, diakritische Zeichen im HTML Quellcode zu kodieren, die jeweils durch entsprechende Skripten abgefangen werden müssen.

Die lokale Archivierung der Pressemitteilungen

Die lokale Archivierung der Pressemitteilungen selbst erfolgt ebenfalls durch eine studentische Hilfskraft. Verwendet werden dabei sogenannte Offline Reader. Dies sind Produkte, die es ermöglichen, ausgehend von einer oder mehreren Startadressen miteinander verknüpfte WWW Seiten automatisiert auf Festplatte zu sichern. Verwendet wird hauptsächlich das Produkt Teleport Pro. Da dieses Programm Schwächen zeigt, wenn der Aufruf verknüpfter Dokumente über Javascript Konstruktionen erfolgt, wird in diesen Fällen das Produkt WinHttrack eingesetzt. WinHttrack ist allerdings nicht so leicht handhabbar, da es sich ursprünglich um ein UNIX Programm handelt.

Grundsätzlich gilt, dass als Startadressen genau die Adressen Verwendung finden, die auch beim Aufruf der lokal gespeicherten Übersichtslisten angegeben werden. Dadurch wird sicher gestellt, dass zu jeder generierten Titelaufnahme auch eine lokal archivierte Pressemitteilung vorhanden ist.

Lokal werden die Pressemitteilungen in einer Verzeichnisstruktur abgelegt, deren Hauptelement jeweils die Adresse des Servers ist, von dem die Pressemitteilung stammt. Eine weitere, in der Regel chronologische Untergliederung dieser Verzeichnisse ist je nach vorhandener Datenmenge möglich. Bilder – in der Regel grafische Navigationselemente und ähnliches – werden für jeden Server in einem separaten Verzeichnis abgelegt.

Lokale Weiterbearbeitung der archivierten Pressemitteilungen

Die lokal archivierten Pressemitteilungen werden durch ein weiteres PERL Programm modifiziert, das die automatisiert erzeugten Titelaufnahmen auswertet. Modifiziert werden nur HTML Dateien, Pressemitteilungen die im PDF- oder WORD Format vorliegen, bleiben unverändert. Es werden folgende Modifikationen vorgenommen:

1. Links auf andere Seiten werden deaktiviert. So enthalten die Pressedienste meist Verknüpfungen zur Homepage der Organisation, die ja lokal nicht archiviert wird. Auch

die Möglichkeit, über einen Link Mail an den Webmaster zu schicken wird so deaktiviert, um diesen vor Mails zu schützen, die sich auf unsere Anwendung beziehen.

2. Die Aufrufe von Bildern, die in das Dokument eingebunden sind, werden an die lokale Archivierungsstruktur angepasst. Bei diesen Bildern handelt es sich – wie erwähnt – in der Regel um grafische Navigationselemente, Logos und ähnliches.
3. Dem Dokument wird ein normierter Kopf in tabellarischer Form voran gestellt, der das Dokument kurz beschreibt.
4. Die Titelaufnahme wird als Dublin Core Element Set im Header der HTML Datei verankert. Dadurch halten wir uns die Möglichkeit offen, diese Informationen zukünftig durch verbesserte Volltext Retrieval Programme auszuwerten.
5. Jedes lokal archivierte Dokument wird mit einer Style Sheet Datei verbunden. Dadurch wird es möglich, das Layout aller archivierten Dokumente in gewissem Umfang über eine einzige Datei zu kontrollieren.

Alle gesicherten Pressemitteilungen sind nachträglich jederzeit modifizierbar. So wird es möglich, die Sammlung jederzeit an neue Entwicklungen, beispielsweise XML anzupassen.

Volltextrecherche als ergänzender Sucheinsteig

Auf die Schwächen einer Volltext Recherche wurde bereits hingewiesen. Trotzdem erscheint es sinnvoll, eine solche Recherchemöglichkeit zumindest als ergänzenden Sucheinstieg anzubieten.

Ein erster konzeptioneller Entwurf sah vor, das gesamte sinntragende Wortmaterial aus den archivierten Pressemitteilungen zu extrahieren und im Bibliothekssystem ALLEGRO mit den erzeugten Titelaufnahmen in einer Datenbank zu verbinden. Leider erwiesen sich die für diesen Ansatz erforderlichen Batchprozeduren im Routinebetrieb als zu zeitaufwendig.

In der aktuellen Version des Projektes ist die Volltextrecherche nun als zusätzlicher ergänzender Sucheinstieg etabliert worden. Die Indexierung der Volltexte erfolgt mit dem Programm htdig, das auch im Occasio Projekt des Internationalen Instituts für Sozialgeschichte eingesetzt wird. Bei htdig handelt es sich um ein Unix basiertes Freeware Programm. In der Anwendung der Bibliothek der Friedrich-Ebert-Stiftung läuft htdig (ebenso wie die im WWW angebotene ALLEGRO Datenbank) unter Linux.

Leider bietet htdig in der aktuell vorliegenden Version keine Möglichkeit, die im Header der archivierten Dokumente als Dublin Core verankerten Informationen auszuwerten. Es erlaubt lediglich eine grobe geografische Klassifizierung des archivierten Materials, die auf einer Auswertung der Länderkürzel in der bereits skizzierten Verzeichnisstruktur basiert, in der das gesicherte Material abgelegt wird. Die im Internet gebräuchlichen Länderkürzel sind ja Teil des Servernamens.

Eine vorläufige Bilanz

Wie bereits erwähnt werden zur Zeit ca. 100 Pressedienste laufend gesichert. Mittlerweile sind rund 137.000 Pressemitteilungen lokal archiviert, mit den beschriebenen Verfahren erschlossen und im Internet zugänglich gemacht worden. Die gemachten Erfahrungen sind mit Einschränkungen zufriedenstellend. Das skizzierte Verfahren funktioniert bei einem großen Teil der archivierten Pressedienste zuverlässig und verbindet traditionelle bibliothekarische Arbeitsweisen mit innovativen Ansätzen. Es haben sich jedoch bei der Konfrontation des

Konzepts mit den zuweilen anarchischen Gegebenheiten im WWW einige Problembereiche heraus kristallisiert:

1. Der Programmieraufwand für das Projekt ist weit höher als ursprünglich angenommen. Dies resultiert daraus, dass viele Organisationen, das Layout der ausgewerteten WWW Seiten weit häufiger ändern, als erwartet. Jede Layoutänderung bedeutet aber zusätzlichen Programmieraufwand im Rahmen des Projektes.
2. WWW Seiten werden vielfach mit JavaScript Elementen überfrachtet. Dies verhindert in Extremfällen, das die mit diesen Seiten verknüpften Dokumente über einen Offline Reader lokal gesichert werden können. Eine manuelle Sicherung ist aber wegen der Fülle der Dateien ebenfalls unmöglich. Solche Angebote können dann aus rein technischen Gründen nicht im Rahmen des Projektes ausgewertet werden, obwohl es inhaltlich sinnvoll wäre.
3. Immer mehr Pressedienste werden nicht mehr im WWW angeboten, sondern an interessierte Personen mittels E-Mail verschickt. Zur Zeit existiert in der Bibliothek der Friedrich-Ebert-Stiftung noch kein griffiges Konzept, wie mit diesen Angeboten umgegangen werden kann. Insbesondere die Extraktion sinnvoller Metadaten aus den E-Mails erweist sich als kaum lösbar, da vielfach im Betreff der Mail keine relevanten Informationen angegeben werden.
4. Die Effizienz des Verfahrens differiert sehr stark. Grundsätzlich steigt die Effizienz des Verfahrens mit der Zahl der Pressemitteilungen, die über eine einzige Startadresse erreicht werden können. Diese Zahl differiert innerhalb der einzelnen Informationsangebote stark und hängt von einer Reihe von Faktoren ab. Neben dem Umfang des Angebots im allgemeinen spielen hier besonders die Organisation der WEB Side und die Flexibilität eventuell eingesetzter Datenbank-Schnittstellen eine Rolle.

Ein Teil dieser Probleme dürfte durch den Einsatz verbesserter Software lösbar sein. Dies gilt insbesondere für Probleme im Zusammenhang mit dem Einsatz von Offline-Readern. Es besteht aber die Möglichkeit, dass die Realitäten im Internet eine konzeptionelle Neuorientierung erforderlich machen. So könnte beispielsweise die systematische Archivierung von Pressediensten in E-Mail-Form eine Hinwendung zur reinen Volltext-Indexierung und somit eine Annäherung an das Occasio-Projekt des IISG nötig machen. Es steht zu hoffen, dass die frei im Internet erhältlichen Volltext-Retrievalsysteme bald das Dublin Core Element Set unterstützen, um auf diesem Weg doch wieder eine Synthese mit dem ursprünglichen Projektansatz zu erreichen.