World Library and Information Congress:
71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme: http://www.ifla.org/IV/ifla71/Programme.htm

June 20, 2005

Code Number: 074-G
Meeting: 133 SI - Bibliography

Web-Crawling : Die Praxis der Bibliothèque nationale de France

Christian Lupovici
Head of the National Bibliographic Agency
Bibliothèque nationale de France
Christian.lupovici@bnf.fr

Zusammenfassung:
Die Bibliothèque nationale de France erprobt zurzeit im Rahmen ihres Pflichtexemplarrechts Verfahren des Web-Harvesting und organisiert die Langzeitarchivierung digitaler Dokumente. Um dieses Ziel zu erreichen, sind grundsätzliche Überlegungen zum Wesen des Pflichtexemplarrechts und der bibliografischen Behandlung von Internet-Ressourcen erforderlich. In der Realität ist die enorme Zahl von digitalen Ressourcen ein wichtiger Faktor bei allen Entscheidungen, die die Nationalbibliothek treffen muss.

1. Das Pflichtexemplarrecht der BnF

1.1 Geschichtlicher Hintergrund

Die französische Pflichtexemplargesetzgebung geht zurück bis ins 16. Jahrhundert, als François I. 1537 die Verordnung von Montpellier erließ. Seit dieser Zeit war der Zweck aller späteren Regelungen, jede neue Ausgabeform zu berücksichtigen und sie in die gesetzlichen Vorschriften zu integrieren, um das Pflichtexemplarrecht entsprechend zu erweitern.

1537: Bücher;
1667: Radierungen;
1689: Stiche;
1925: jede Art von Grafik;
1941: Plakate, Noten und Fotografien;
1963: Tonträger jeder Art;
1975: Aufzeichnungen von unbewegten und bewegten Bildern, unabhängig vom technischen Verfahren;
1977: Filme;
1992: Alle Arten von Dokumenten auf physischen Datenträgern, einschließlich Datenbanken und Expertensystemen, soweit sie öffentlich zugänglich sind.

Die Einbeziehung von Online-Ressourcen in die französische Verordnung wird im Jahr 2005 erfolgen, wenn die Diskussion im Parlament abgeschlossen ist.

1.2 Auftrag und Umfang des Pflichtexemplargesetzes

Zweck des Pflichtexemplargesetzes ist die Bewahrung und langfristige Bereitstellung des vollständigen kulturellen Erbes an erschienenen und in zahlreichen Exemplaren veröffentlichten Dokumenten. Das Pflichtexemplargesetz umfasst alle Arten von Veröffentlichungen und ist nicht auf die französische Sprache und auch nicht auf Veröffentlichungen aus Frankreich beschränkt.
Es ist mehr eine Frage des Inhalts und/oder der Vertriebswege (Bücher, Grafiken, Karten, Tonträger, audiovisuelle Medien) als eine Frage des Mediums: Papier, elektronischer Datenträger, Netzpublikation. Im Internet ist der größte Teil der institutionellen grauen Literatur, die öffentlich zugänglich ist, Bestandteil des Pflichtexemplarrechts.
Pflichtexemplarrecht bedeutet auch, dass das Sammeln nicht eine Frage der Erwerbung im Sinne eines Bestandsentwicklungsverfahrens ist. Vielmehr sollen als ein Abbild der Veröffentlichungen, die dem französischen Volk gegenwärtig zugänglich sind, alle diese Materialien gesammelt werden, um sie für die zukünftige historische Forschung zu erhalten. Das bedeutet, dass sowohl verschiedene Ausgaben einer Veröffentlichung wie auch verbotene Schriften aufgenommen werden, so wie es schon immer seit dem 16. Jahrhundert war.
Wie soll das bei Netzpublikationen umgesetzt werden?

2. Web-Crawling

2.1 Internet-Harvesting für das Pflichtexemplarrecht

Die Anwendung der Pflichtexemplar-Philosophie auf Netzpublikationen setzt voraus:
a) Den Ausschluss aller Domain-Namen, die keine Ressourcen zur französischen Kultur enthalten
b) Die Konzentration auf die spezifisch französischen nationalen Domain-Namen (*.fr, .pf, .wf, .pm, .re, .tf, .ad, .yt*)
c) Das Harvesting des „Restes“

In einem ersten Ansatz ist es nicht möglich zu erkennen, was innerhalb des „Restes“ für die französische Kultur relevant ist. Deshalb muss zuerst ein Verfahren gefunden werden, mit dem eine grobe Auswahl auf der Basis einer semantischen Analyse der Seiteninhalte getroffen werden kann. Da Speicherplatz kein entscheidendes Kriterium mehr ist, speichert und archiviert die Bibliothèque nationale de France in der Hoffnung auf eine bessere Lösung eine wesentlich größere Anzahl von Seiten als streng genommen für das Pflichtexemplarrecht erforderlich ist. Auf jeden Fall wird der Teil des Web, der gespeichert und archiviert werden soll, immer sehr umfangreich sein.

2.2 Von Experimenten zur Realität

2.2.1 Erste Versuche

Die Bibliothèque nationale de France sammelt seit 2001[1] französische Webseiten mithilfe von HTTrack-Crawlern. Die ersten Versuche konzentrierten sich auf verschiedene Wahlkampagnen in Frankreich. Diese Versuche dienten zum Testen der folgenden Punkte:
a) Bildung einer Typologie der Webseiten und damit zusammenhängend Bestimmung der richtigen Periodizität;
b) Testen von technischen Problemen (Formate, Tiefe des Crawling) und das Tuning des Crawlers;
c) Festlegung von manuellen und automatischen Mechanismen bei der Auswahl von Webseiten

Diese Phase war wichtig, um automatische Selektionen von Webseiten festzulegen und sich mit der Nutzung eines Tools vertraut zu machen, der Unterstützung bei der Entscheidungsfindung bietet.

[1] http://bibnum.bnf.fr/conservation/aristote2004/roche.pdf

2.2.2 Web-Crawling in der Realität

Was bedeutet es für das französische Pflichtexemplarrecht im Hinblick auf Ressourcen und Umfang des zu Sammelnden?
Im Frühling 2004 wurde mit dem Alexa Web Crawler ein Schnappschuss über das ganze Web hergestellt. Dieser Schnappschuss richtete sich nach den Spezifikationen, in denen festgelegt war, welche Domainnamen verwendet werden (*.fr, .com, .net, .edu, .com, .biz, .info, .int, .pf, .ad, .coop, .name, .aero, .tf, .re, .museum, .pro, .pm, .wf, .yt*) und welche ausgeschlossen werden sollen. Das Ergebnis dieses Crawlings ergab mehr als 2 Billionen URLs und 65 Millionen Hosts und hatte einen Umfang von 27 Terabytes.
Zusätzlich wurde mit dem Heritrix-Crawler am Jahresende 2004 und im Januar 2005 ein gezieltes Crawling auf die *.fr*-Domain durchgeführt. Das Ergebnis dieses späteren Crawlings waren 500.000 amtliche und 4.000 persönliche Seiten mit 118 Millionen URLs und einer halben Million Hosts und hatte einen Umfang von 3 Terabytes.
Bei diesem Datenvolumen ist es nutzlos, sich Gedanken über irgendein Verfahren manueller Katalogisierung oder Indexierung zu machen. Für diese Menge an Daten wären mindestens 500 Mannjahre allein für das Sortieren und Katalogisieren der amtlichen Seiten erforderlich. Das ist zweimal mehr als das Katalogisierungspersonal der BnF für die Nationalbibliografie.
Daher ist es notwendig, verschiedene Techniken zu kombinieren.

2.3 Das Verfahren für die Pflichtexemplare

Die grundlegende Methode ist das automatische Crawling. Das komplette Surface Web und einiges aus dem Deep Web können damit geharvestet werden, abhängig von den technischen Problemen, mit denen der Crawler konfrontiert wird und von seinem Tuning. Aber im Großen und Ganzen kann ein automatisches Crawling hauptsächlich das Surface Web erfassen und damit eine Karte des Web erstellen.
Eine zweite und zusätzliche Methode ist die Konzentration des automatischen Crawlings auf eine einzelne Gruppe von Seiten, z. B. auf die *.fr*-Domain, um dort den Crawler gezielt in der Tiefe suchen zu lassen.
Diese beiden letzten Methoden sind top-down orientiert.
Um das ganze Bild des Web zu erhalten, müssen wir das periodische automatische Harvesting kombinieren mit einer botton-up orientierten Technik, die darin besteht, die Sammlungen mit von Webseiten-Anbietern abgelieferten Deep-Web-Ressourcen zu ergänzen und mit den Links aus dem Surface Web zu verbinden. Solche Datenlieferungen müssen teils manuell, teils automatisch bearbeitet werden. Die Datenlieferung kann Vorteile in Bezug auf die Erhaltungssicherheit bieten, aber es ist ein zeitaufwändiger Prozess, beginnend mit der Analyse der Seite, die durchschnittlich 5 Manntage in Anspruch nimmt.
Das komplette Daten-Set wird in der BnF in einem zentralen Verzeichnis für den internen Gebrauch gespeichert. Das Web-Archiv ist Teil des globalen digitalen Datenspeicherungs- und Langzeitarchivierungsprogramms der BnF, SPAR (*Système de Préservation des ARchives*), das am Jahresende 2004 begonnen hat.

2.4 Fragestellungen und Verfahren zur Auswahl

Grundsätzlich möchte die Bibliothèque nationale de France die relevantesten Webseiten archivieren. Deshalb ist es wichtig, ein Verfahren zu finden, das aus den geharvesteten Webseiten selektieren oder Vorschläge zur Selektion machen kann. Es wurde ein französisches Forschungsprojekt mit dem Namen „Watson“ über linguistische Regelungen durchgeführt, die im Web-Archiv der BnF für Mitarbeiter und Nutzer angewendet werden sollen. Das Projekt dauerte von 2003 bis 2004[2]. Einer der Wege, die erforscht wurden, war die Beschreibung einer Webseite durch linguistische Analyse der Metadaten und durch Phrasen aus dem Volltext. Dieses Verfahren lieferte Abstracts über den Seiteninhalt. Es ist noch nicht gut genug, um ohne menschliche Überprüfung genutzt zu werden, aber es ist gut genug als Hilfe bei der Seitenauswahl.
Die BnF ist dabei, ein Web-Beobachtungsteam zu organisieren. Es besteht hauptsächlich aus Mitarbeitern der verschiedenen Abteilungen, die für die Pflichtablieferung zuständig sind.

[2] José Coch, Julien Masanès**. -** Language engineering techniques for web archiving. In : 4th International Web Archiving Workshop (IWAW04), 16 septembre 2004, Bath, UK - http://www.iwaw.net/04/index.html (visited 30/05/2005).

Aufgabe des Teams ist es, Erfahrungen mit der Umgebung der neuen digitalen Ressourcen zu sammeln und die erforderlichen Verbindungen zu den bisherigen traditionellen Sammlungen herzustellen. Es wurden Zugriffstools und Beobachtungsverfahren eingerichtet, um die Ergebnisse des Crawlings und die Qualität des Harvestings bewerten zu können. Das Team hat Zugriff auf einen Tool, mit dem es möglich ist, verschiedene Überprüfungen sowohl im BnF-Archiv als auch im Web durchzuführen, um bestimmte Webseiten auszuwählen, die vollständig von einem Crawler erfasst werden sollen. In solchen Fällen berücksichtigt der Crawler die Ergebnisse des Beobachtungsteams und erhält die gewünschten URLs als Parameter.
Wie auch immer, es wird niemals möglich sein, eine Gruppe von Webseiten auszuwählen, die exakt dem französischen Pflichtexemplarrecht entsprechen, und daher ist klar, dass das BnF-Web-Archiv immer mehr Seiten als erforderlich enthalten wird. Wenn alle Länder das Gleiche tun, wird es Überschneidungen geben, aber das ist eigentlich gut, da Speicherplatz kein Problem mehr ist und Dubletten in verschiedenen Ländern eine größere Sicherheit bei der Langzeitarchivierung garantieren. Man muss auch bedenken, dass die Schnappschuss-Technik keinerlei Vollständigkeit der Sammlung garantiert, sondern nur ein repräsentatives Beispiel der aktuellen Kultur liefert. Das ist das Wesen des Pflichtexemplarrechts.
Dennoch hat das internationale Konsortium IIPC (International Internet Preservation Consortium) festgestellt, dass Bedarf für einen „smarten Crawler“ besteht, der in der Lage ist, sich nach vorgegebenen Parametern auf bestimmte Seiten zu konzentrieren und in die Tiefe zu gehen, wenn die Relevanz der Seite hoch ist. Die British Library und die Bibliothèque nationale de France haben in diesem Frühjahr einen Call for Proposal für einen solchen „smarten Crawler“ gestartet.

2.5 Neue Umgebung, neues Dokument

Beim Umgang mit dem Web müssen wir die traditionellen Verfahren der Dokumentbearbeitung hinter uns lassen und einen neuen Weg für die Behandlung dieser Ressourcen einschlagen.
Das Web-Archiv hat neue Dimensionen:
Es muss die intern geschaffenen Navigationsmöglichkeiten behalten, da keines der Dokumente eine allein stehende Ressource ist. Das Dokument selbst ist transformiert worden. Im BnF-Modell ist die Dokumenteneinheit die Webseite selbst.

2.6 Linguistische Verfahren zur Analyse der Ressourcen

In Anbetracht der ungeheueren Datenmengen, die in das Speichersystem übernommen werden, brauchen sowohl die Mitarbeiter wie die Leser Tools, mit denen die gewünschten Informationen ermittelt werden können. Manchmal brauchen sie bei der Suche nach präzisen Informationen kleine Teilmengen, manchmal müssen aber auch große Mengen von Informationen verarbeitet werden, um statistische oder linguistische Aufbereitungen durchzuführen.
Das Watson-Projekt richtete sich auf die Auswahl individueller Gruppen von Ressourcen entsprechend der gewünschten Domain, die durch linguistische Analysen ermittelt werden.
Man kann dem Nutzer eine Vorstellung über den Inhalt durch die im Dokument enthaltenen Strukturen, Stichworte, Namen und Orte oder durch ein navigationsfähiges Abstract geben.
In dem Projekt wurden benutzerspezifische Tools für die Ermittlung von Inhaltsangaben, die Kontext-Navigation und die semantische Analyse der Seiteninhalte erstellt. Die Experimente wurden im französischen Wahlarchiv durchgeführt.

3. Neue Katalogisierungsperspektiven

3.1 Die Katalogisierung von Metadaten

Herkömmliche Metadaten werden für die Beschreibung von Dokumenten auf physischen Datenträgern für Retrievalzwecke genutzt. Diese Katalogisierung ist sehr arbeitsaufwändig und vollständig manuell, aber sie gehört zu physischen Objekten und kann bei diesen nicht vermieden werden.
In einer digitalen Umgebung ist die Situation anders:
Digitale Objekte sind (oder sollten es sein) bereit für die Informationsverarbeitung. Sie enthalten Selbstbeschreibungen in Form von technischen and rechtlichen Metadaten, die ihre Verwaltung in einem System möglich machen.

Wenn Text vorhanden ist, kann der Inhalt selbst in diesem Stadium ohne menschlichen Eingriff indexiert und sortiert werden. Damit werden die Struktur und die semantische Inhaltsinformation abgedeckt.
Im Webkontext können unzureichende Metadaten durch den Netzwerk-Effekt ausgeglichen werden. Die digitalen Objekte sind im Webarchiv keine isolierten Dokumente. Sie sind mit anderen Ressourcen durch eindeutige Links, Inhalte und Beschreibungen von Inhalten verbunden, und das schließt Suchhilfen ein, seien es traditionelle oder völlig neue.
Das bedeutet nicht, dass wir keinerlei beschreibende Metadaten oder Stichwörter mehr brauchen. Die Bibliothek ist aber nicht der beste Ort, um solche Informationen zu erstellen. Es ist besser, diese Aufgabe in die Entstehungsphase des Objektes zu verlagern, da der Bedarf für solche Daten weniger groß ist als in einer herkömmlichen Umgebung.

3.2 Weitere Arten von Metadaten

Neue Sachverhalte treten auf wie die Rechteverwaltung und die Sicherung der Authentizität, und dies nicht nur aus rechtlichen Gründen, sondern auch um eine unbegrenzte Wiedergabe der digitalen Ressourcen und ihre sofortige Bereitstellung sicherzustellen. Formate, die in einem Dokument häufig kombiniert vorkommen, sind einer der entscheidenden Faktoren, wenn die Informationen des Servers oder des Dokuments nicht zuverlässig genug sind. Für die gesammelten Dokumente ist eine Formatbewertung erforderlich.
Für eine Nationalbibliothek ist es auch sehr wichtig, Metadaten mit Archivierungsinformationen anzureichern, die auf das Web-Archiv abgestimmt sind.

3.3 Zukünftige Katalogisierung

Die Katalogisierungssituation im Webkontext kann wie folgt zusammengefasst werden: Erstellen von Spezifikationen für die Herstellung von digitalen Dokumenten und Quellenanalyse während der Phase des Einsammelns (OAIS Modell) einschließlich der Crawler-Instruktionen und Zusammenarbeit mit den Webproduzenten.

4. Wie kann eine Nationalbibliografie des Web aussehen?

Das französische Pflichtexemplargesetz (auch das neue) besagt, dass es eine Nationalbibliografie mit den Daten der abgelieferten Dokumente gibt.
Aber wie soll eine Nationalbibliografie im Kontext des Web aussehen?
Da der Zweck einer Nationalbibliografie die Anzeige neuer Publikationen ist, können wir uns vorstellen, auf der BnF-Webseite unter Anwendung eines automatischen Sortierungs- und „Katalogisierungs“-Verfahrens eine Liste der neu hinzugekommenen (nicht der geänderten) Webseiten zum Zeitpunkt der Archivierung zu veröffentlichen (OAIS-Modell). Das Ergebnis wird eine Liste von URLs mit Namensindex und einem Schlagwortindex sein (soweit es möglich ist, diesen mit einer linguistischen Analyse zu erstellen, ähnlich dem Verfahren im Watson-Projekt).

Deutsche Übersetzung: Claudia Werner, Die Deutsche Bibliothek