nestor
Kompetenznetzwerk Langzeitarchivierung

nestor Handbuch:

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

15.4 Web-Archivierung

Herausgeber:

Heike Neuroth
Hans Liegmann
Achim Oßwald
Regine Scheffel
Mathias Jehn

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Im Auftrag von:

nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de

Dieser Artikel ist ein Auszug aus dem nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Dieser Artikel ist verfügbar unter der URL:
http://nestor.sub.uni-goettingen.de/handbuch/artikel/text_84.pdf

Die Online Version des Handbuches unter der URL:
http://nestor.sub.uni-goettingen.de/handbuch/

Kontakt:
Niedersächsische Staats- und Universitätsbibliothek Göttingen
Dr. Heike Neuroth
Forschung und Entwicklung
Papendiek 14
37073 Göttingen
neuroth@sub.uni-goettingen.de
Tel. +49 (0) 55 1 39 38 66

15.4 Web-Archivierung

von Hans Liegmann

Der Begriff Web-Archivierung soll hier auf diejenigen Verfahren eingeengt werden, die als "Web-Harvesting" Eingang in die Sammelmethoden von gedächtnisorganisationen gefunden haben.

Unter Web-Harvesting versteht man das automatisierte Einsammeln von Internet-Dokumenten zum Zwecke der Archivierung in einem digitalen Archiv. Zentrales Element des Web-Harvesting ist eine Software-Komponente (crawler). Diese sucht ausgehend von einer Liste vorgegebener Web-Adressen (URL seed list) die erreichbaren Dokumente auf und speichert sie in einer definierten Zielumgebung ab.<?xml:namespace prefix = o ns = "urn:schemas-microsoft-com:office:office" />

Beim selektiven zielgerichteten Web-Harvesting (focused crawl) besteht das Ziel darin, möglichst vollständige und konsistente Archivkopien genau derjenigen Websites zu erhalten, deren Adressen in der vorgegebenen Liste enthalten sind.
Beim flächigen Web-Harvesting (broad crawl) wird eine vorgegebene Adressliste lediglich als Einstieg in ein Sammelverfahren verwendet, das weitergehend ist. Flächiges Web-Harvesting hat definierte formale Regeln als Auswahlgrundlage der zu archivierenden Websites. Eine typische Regel kann lauten, dass zu archivierende Dokumente Bestandteil eines bestimmten Internet-Bereiches (domain, z.B. "de") sein müssen, um als archivierungswürdig angesehen zu werden.
Unabhängig vom Komplexitätsgrad möglicher Regelformulierungen ist die Grundlage des Sammelverfahrens die Verfolgung von Hyperlinks: aus den aufgefundenen Dokumenten werden wiederum die in ihnen enthaltenen Web-Adressen extrahiert und auf Regelkonformität geprüft. Die Liste der aufzusuchenden URLs wird dann ggf. dynamisch erweitert.
Derzeit gibt es verschiedene Produkte auf dem Markt, die zur Durchführung von Web-Harvesting geeignet sind. Das Angebot ist vorrangig auf die Bedürfnisse des selektiven Harvesting ausgerichtet. Dazu gibt es kommerzielle, Freeware- und Open-Source-Angebote. Diese genügen überwiegend den Anforderungen der Langzeitarchivierung nicht, da sie bei der Archivierung der Daten inhaltliche Veränderungen vornehmen.
Flächiges Harvesting unter Berücksichtigung der Authentizität archivierter Objekte wird nur von wenigen Softwareprojekten (z.B. der Crawler HERITRIX des International Internet Preservation Consortium) unterstützt. Bei der Planung produktiver Harvesting-Anwendungen im Massenbetrieb ist zu berücksichtigen, dass kommerzielle Software-Produkte mit garantiertem Leistungsumfang nicht zur Verfügung stehen und ggf. umfangreiche Zusatzinvestitionen notwendig sind, um die gewünschte Funktionalität zu erreichen.
Die eingesetzten Harvesting-Verfahren lassen sich in einer Matrix einordnen, die nach den Kriterien „flächig“ bis „fokussiert“ und „nationale/regionale Auswahl“ bis „fachlich/institutionelle“ Auswahl aufgebaut ist. Die Aktivitäten von Nationalbibliotheken sind zum Teil flächig angelegt (Sammeln nationaler Adressräume) oder auch durch selektives Vorgehen bestimmt (Auswahl der für einen bestimmten Kulturkreis als relevant bewerteten Internetpräsenzen). Im Bereich der fokussierten Harvesting-Ansätze finden sich fachlich orientierte Beispiele wie z.B. das Projekt DACHS [http://www.sino.uni-heidelberg.de/dachs], die Vorgehensweise des Deutschen Parlamentsarchivs [http://www.bundestag.de/bic/archiv/oeffent/ArchivierungNetzressourcenKlein.pdf] mit institutioneller Abdeckung und die kooperativen Aktivitäten einiger deutscher Parteienarchive [http://www.fes.de/archiv/spiegelungsprojekt.htm].
Bei der Darstellung der Methode soll nicht unerwähnt bleiben, dass die technischen Instrumentarien zur Durchführung zurzeit noch mit einigen Defiziten behaftet sind:

- Inhalte des so genannten „deep web“ sind durch Harvester nicht erreichbar. Dies schließt z.B. Informationen ein, die in Datenbanken oder Content Management Systemen gehalten

werden. Harvester sind noch nicht in der Lage, auf Daten zuzugreifen, die erst auf spezifische ad-hoc-Anfragen zusammengestellt werden und nicht durch Verknüpfungen statischer Dokumente repräsentiert sind.

- Inhalte, die erst nach einer Authentisierung zugänglich sind, entziehen sich verständlicherweise dem Harvesting-Prozess.
- dynamische Elemente als Teile von Webseiten (z.B. in Script-Sprachen) können Endlosschleifen (crawler traps) verursachen, in denen sich der Harvester verfängt.
- Hyperlinks in Web-Dokumenten können so gut verborgen sein (deep links), dass der Harvester nicht alle Verknüpfungen verfolgen kann und im Ergebnis inkonsistente Dokumente archiviert werden.

Vor allem bei der Ausführung flächigen Web-Harvestings führen die genannten Schwächen häufig zu Unsicherheiten über die Qualität der erzielten Ergebnisse, da eine Qualitätskontrolle aufgrund der erzeugten Datenmengen nur in Form von Stichproben erfolgen kann. Nationalbibliotheken verfolgen deshalb zunehmend Sammelstrategien, die das Web-Harvesting als eine von mehreren Zugangswegen für Online-Publikationen etablieren.

Der individuelle Transfer von Einzeldokumenten über Einlieferschnittstellen oder teilautomatisierte Zugangsprotokolle sowie bilaterale Vereinbarungen mit Produzenten bilden eine wichtige Ergänzung des „vollautomatischen“ Sammelverfahrens.

Eine Umfrage der IFLA im Jahr 2005 hat ergeben, dass 16 Nationalbibliotheken Web-Harvesting praktizieren, 11 davon flächiges Harvesting in unterschiedlichen Stadien der Produktivität. 21 Nationalbibliotheken setzen parallel oder ausschließlich andere Verfahren zur Sammlung von Online-Publikationen ein. Die Ergebnisse von Web-Harvesting-Verfahren sind aus urheberrechtlichen Gründen fast ausschließlich nur in den Räumen der jeweiligen Nationalbibliothek zugänglich.