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
Stefan Strathmann

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

Kontakt

editors@langzeitarchivierung.de
c/o
Niedersächsische Staats- und Universitätsbibliothek Göttingen
Dr. Heike Neuroth
Forschung und Entwicklung
Papendiek 14
37073 Göttingen
Tel. +49 (0) 55 1 39 38 66

15.4 Web-Harvesting zur Langzeiterhaltung von Internet-Dokumenten

Hans Liegmann

(überarbeitete Fassung eines Vortrags auf der 10. Tagung des Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen" - Planungen, Projekte, Perspektiven – Zum Stand der Archivierung elektronischer Unterlagen - Düsseldorf, 14./15. März 2006)

1 Web-Harvesting als Sammelmethode für Internet-Dokumente

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.
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 Anforde-

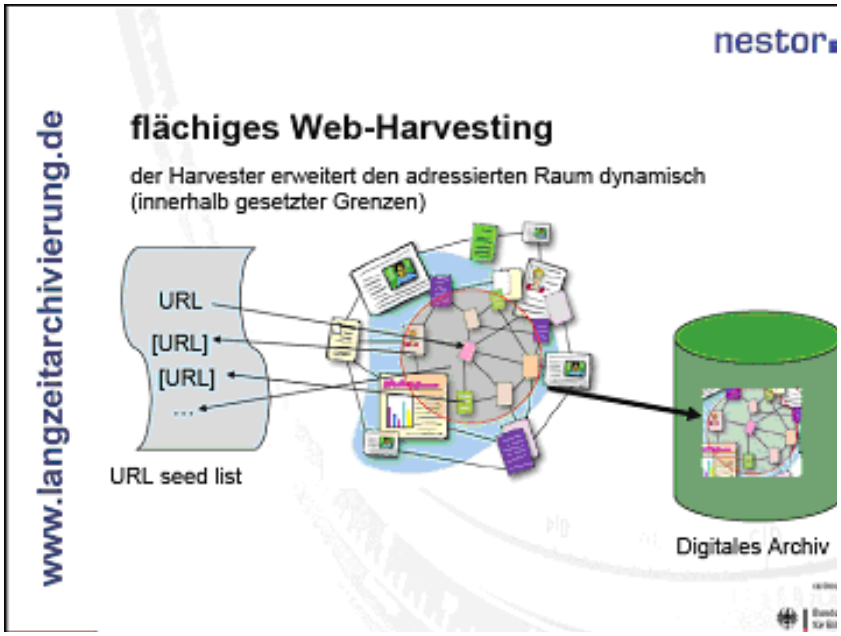

Abbildung 15.4.1

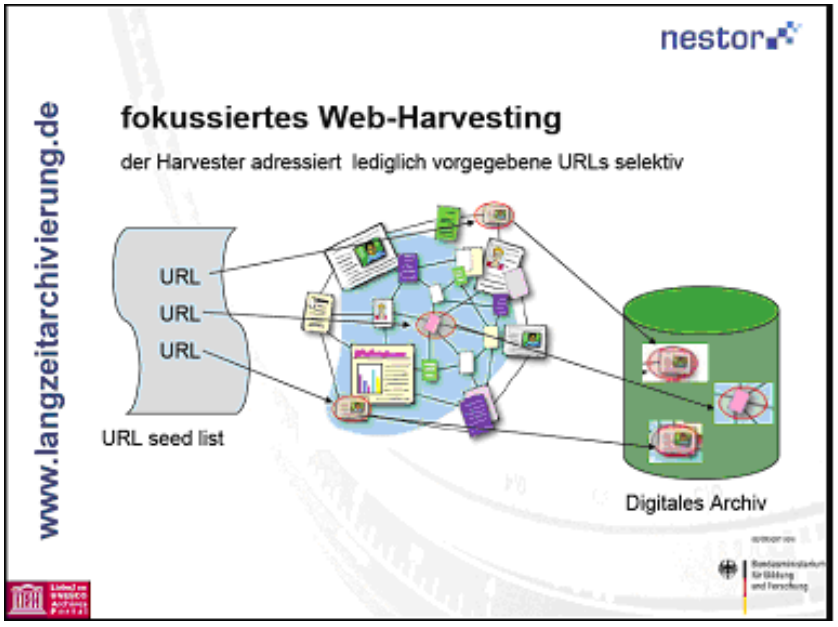

Abbildung 15.4.2

rungen 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 aktuelle Anwendungsbreite von Web-Harvesting-Verfahren ist in folgendem Schaubild dargestellt:

Die eingesetzten 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[27], die Vorgehensweise des Deutschen Parlamentsarchivs[28] mit institutioneller Abdeckung und die kooperativen Aktivitäten einiger deutscher Parteienarchive[29].
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.

27 http://www.sino.uni-heidelberg.de/dachs [DACHS - Digital Archive for Chinese Studios] (Juni 2006)

28 http://www.bundestag.de/bic/archiv/oeffent/ArchivierungNetzressourcenKlein.pdf [Angela Ullmann; Steven Rösler: Archivierung von Netzressourcen des Deutschen Bundestages] (Juni 2006)

29 http://www.fes.de/archiv/spiegelungsprojekt.htm [Politisches Internet-Archiv] (Juni 2006)

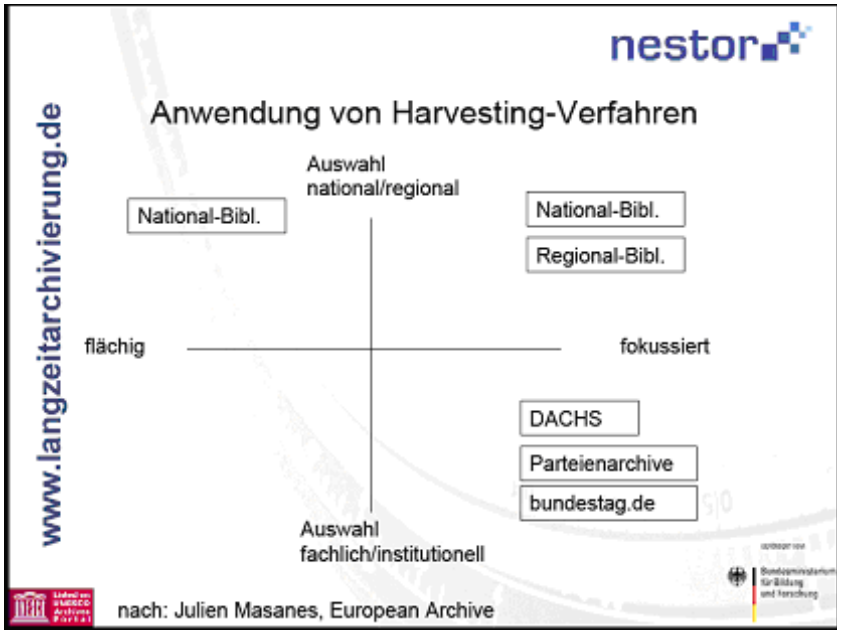

Abbildung 15.4.3

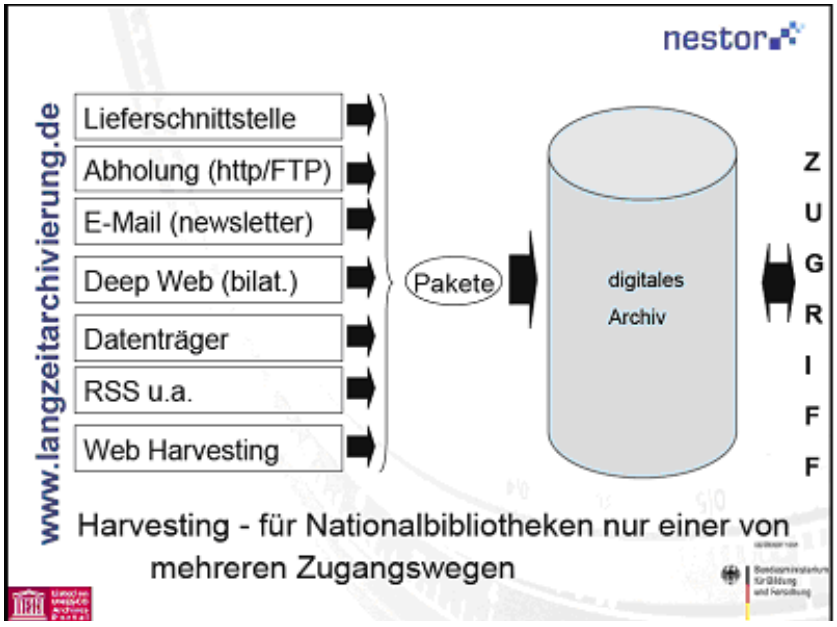

Abbildung 15.4.4

- 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.

2 Nationalbibliotheken und das World Wide Web

Nationalbibliotheken fassen grundsätzlich alle der im World Wide Web erreichbaren Dokumente als Veröffentlichungen auf und beabsichtigen, ihre Sammelaufträge entsprechend zu erweitern, soweit dies noch nicht geschehen ist. Eine Anzahl von Typologien von Online-Publikationen wurde als Arbeitsgrundlage geschaffen, um Prioritäten bei der Aufgabenbewältigung setzen zu können und der Nutzererwartung mit Transparenz in der Aufgabenwahrnehmung begegnen zu können. So ist z.B. eine Klassenbildung, die mit den Begriffen „druckbildähnlich" und „webspezifisch" operiert, in Deutschland entstanden.[30] In allen Nationalbibliotheken hat die Aufnahme von Online-Publikationen zu einer Diskussion von Sammel-, Erschließungs- und Archivierungsverfahren geführt, da konventionelle Geschäftsgänge der Buch- und Zeitschriftenbearbeitung durch neue Zugangsverfahren, die Masse des zu bearbeitenden Materials und neue Methoden zur Nachnutzung von technischen und beschreibenden Metadaten nicht anwendbar waren. Die neue Aufgabe von Gedächtnisorganisationen, die langfristige Verfügbarkeit digitaler Ressourcen zu gewährleisten, hat zu neuen Formen der Kooperation[31] und Verabredungen zur Arbeitsteilung geführt.

30 http://www.zlb.de/aktivitaeten/bd_neu/heftinhalte/heft9-1204/digitalebib1104.pdf [Auswahlkriterien für das Sammeln von Netzpublikationen im Rahmen des elektronischen Pflichtexemplars] (Juni 2006)

31 http://www.langzeitarchivierung.de [nestor - Kompetenznetzwerk Langzeitarchivierung]

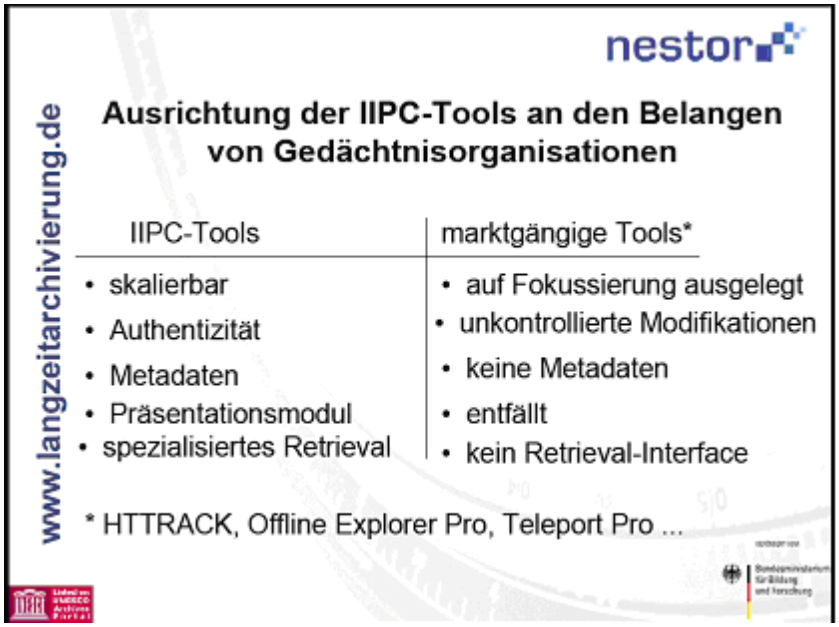

Abbildung 15.4.5

Eine Umfrage der IFLA[32] 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.

Ein „Statement on the Development and Establishment of Voluntary Deposit Schemes for Electronic Publications“[33] der Conference of European National Librarians (CENL) und der Federation of European Publishers (FEP) hat folgende Prinzipien im Umgang zwischen Verlagen und nationalen Archivbibliotheken empfohlen (unabhängig davon, ob sie gesetzlich geregelt werden oder nicht):

(Juni 2006)

32 http://www.ifla.org/ [International Federation of Library Organisations] (Juni 2006)

33 http://www.sne.fr/1_sne/pdf_doc/FINALCENLFEPDraftStatement050822.doc [Statement on the Development and Establishment of Voluntary Deposit Schemes for Electronic Publications] (Juni 2006)

- Ablieferung digitaler Verlagspublikationen an die zuständigen Bibliotheken mit nationaler Archivierungsfunktion
- Geltung des Ursprungsland-Prinzip für die Bestimmung der Depotbibliothek, ggf. ergänzt durch den Stellenwert für das kulturelle Erbe einer europäischen Nation
- Einschluss von Publikationen, die kontinuierlich verändert werden (websites) in die Aufbewahrungspflicht
- nicht im Geltungsbereich der Vereinbarung sind: Unterhaltungsprodukte (z.B. Computerspiele) und identische Inhalte in unterschiedlichen Medienformen (z.B. Online-Zeitschriften zusätzlich zur gedruckten Ausgabe.

Das Statement empfiehlt, technische Maßnahmen zum Schutz des Urheberrechts (z.B. Kopierschutzverfahren) vor der Übergabe an die Archivbibliotheken zu deaktivieren, um die Langzeitverfügbarkeit zu gewährleisten.

3 Nationale Strategien von Nationalbibliotheken

Die norwegische Nationalbibliothek[34] gibt in ihren Planungen für das Jahr 2005 an, viermal im Jahr ein Harvesting des vollständigen nationalen Adressraumes (.no) durchführen zu wollen. Darüber hinaus sollen Online-Tageszeitungen täglich und Online-Zeitschriften in der Häufigkeit ihrer Erscheinungsweise eingesammelt werden. Online-Publikationen mit einer Bedeutung für das norwegische kulturelle Erbe, die in anderen top level domains (z.B. .com, .org, .net) erscheinen, werden in Auswahl archiviert. Datenbanken und Netzpublikationen, die im deep web erscheinen und derzeit nicht durch automatische Harvesting-Verfahren erreichbar sind, bleiben vorerst unberücksichtigt.

Die amerikanische Library of Congress (LoC) hat im Jahr 2000 das MINERVA-Projekt[35] eingerichtet und mit Web Harvesting experimentiert. Dabei hat sich die LoC auf den Aufbau thematischer Sammlungen von Websites konzentriert. In Kooperation mit dem Internet Archive[36] wurden so z.B. folgende Sammlungen eingerichtet: Wahlen zum 107. Kongress, Präsidentschaftswahlen, 11. September 2001. Vorgesehen ist die Sammlung und Archivierung von Websites zu den Olympischen Winterspielen 2002, dem Irak-Krieg und weiteren Wahlen auf nationaler Ebene. Die Aktivitäten der amerikanischen Nationalbibliothek bei der Bildung thematischer Sammlungen stehen im Einklang mit der Vorge-

34 http://www.nb.no/english [The National Library of Norway] (Juni 2006)

35 www.loc.gov/minerva [MINERVA - Mapping the Internet Electronic Resources Virtual Archive] (Juni 2006)

36 http://archive.org [Internet Archive] (Juni 2006)

hensweise bei ihren Digitalisierungsvorhaben zum „American Memory“[37].
Die australische Nationalbibliothek[38] war Vorreiter für die Anwendung innovativer technischer Methoden bei der selektiven Sammlung kulturell bedeutender Websites in Australien. Das dortige digitale Archiv PANDORA[39] wird seit 1996 betrieben. In einem kooperativen Verfahren wird es arbeitsteilig zusammen mit den australischen State Libraries aufgebaut. Eingesetzt wird fokussiertes Harvesting unter Verwendung der Standard-Software HTTRACK[40]. Die zusätzlich durchgeführte intensive Qualitätskontrolle der zu archivierenden Inhalte kostet personelle Ressourcen: bislang konnten durch das mit der Aufgabe betraute Personal (ca. 6 Stellen) insgesamt etwa 12.000 Websites mit 22.000 „Schnappschüssen“ aufgenommen und mit Metadaten versehen werden. Da vorab von jedem einzelnen Urheber das Einverständnis zur Archivierung und öffentlichen Bereitstellung eingeholt wird, ist PANDORA eines der wenigen Web-Archive weltweit, die über das WWW offen zugänglich sind.
Die Nationalbibliotheken von Neuseeland und Großbritannien haben im Rahmen Ihrer selektiven Aktivitäten zur Archivierung wichtiger Websites ihres jeweiligen nationalen Adressraumes ein „Web Curator Tool“[41] entwickelt, das als Freeware allen interessierten Anwendern zur Begutachtung und Verfügung steht.

4 Das International Internet Preservation Consortium (IIPC)

Das IIPC[42] wurde 2003 gegründet. Ihm gehören elf Nationalbibliotheken und das Internet Archive an. Die Gründungsidee des IIPC ist es, Wissen und Informationen aus dem Internet für zukünftige Generationen zu archivieren und verfügbar zu machen. Dies soll durch weltweiten Austausch und Kooperation aller Gedächtnisorganisationen erreicht werden, die sich der neuen Aufgabe stellen.
Die Aktivitäten des IIPC sind vielschichtig. Internationale Kooperation auf einem technischen Gebiet erfordert Standardisierung. So hat das IIPC Mitte 2005 einen Standardisierungsvorschlag (Internet Draft) für das „Web Archive File Format (WARC)“ vorgelegt. Eine Standardisierung des Archivierungsformates vereinfacht die Entwicklung nachnutzbarer technischer Instrumentarien

37 http://memory.loc.gov/ammem/index.html [The Library of Congress - American Memory] (Juni 2006)

38 http://www.nla.gov.au [National Library of Australia] (Juni 2006)

39 http://pandora.nla.gov.au [PANDORA - Australia's Web Archive] (Juni 2006)

40 http://www.httrack.com [HTTrack Website Copier - Offline Browser] (Juni 2006)

41 http://webcurator.sourceforge.net

42 http://www.netpreserve.org [International Internet Preservation Consortium] (Juni 2006)

unter den IIPC-Partnern und erlaubt auch den Austausch von Datenbeständen zur redundanten Speicherung aus Sicherheitsgründen.
Unter dem Projektnamen „HERITRIX“[43] arbeiten die IIPC-Partner an einem Web-Harvester, der allen interessierten Anwendern als Open Source Software frei zur Verfügung steht. HERITRIX tritt mit dem Anspruch an, eine skalierbare und ausbaufähige Software zu entwickeln, die (im Gegensatz zu marktüblichen Produkten) Ergebnisse mit Archiv-Qualität liefert. Standard-Produkte erzeugen normalerweise Veränderungen in den lokalen Kopien von Websites, die den Authentizitätsansprüchen von Gedächtnisorganisationen zuwiderlaufen.
Mit NutchWAX[44] (Nutch & Web Archive Extensions) haben IIPC-Partner eine Suchmaschine für den Einsatz in der Web-Archiv-Umgebung vorbereitet. Damit wird es möglich, die Erwartungen von Web-Archiv-Nutzern im Hinblick auf den Suchkomfort durch die Integration von Standard-Suchmaschinentechnologie zu erfüllen.
WERA[45] (Web Archive Access) ist der Prototyp einer Zugriffskomponente, die als Endnutzer-Schnittstelle den Zugang zum digitalen Archiv erlaubt. Im Gegensatz zu marktüblichen Standard-Tools (z.B. HTTRACK) sind die Ergebnisse des Harvesters HERITRIX als Datenpakete im WARC-Format nicht ohne weiteres von Endnutzern zu betrachten. WERA ergänzt die üblichen Suchfunktionen um die Möglichkeit, einen Zeitpunkt für die Auswahl des gewünschten Schnappschusses im Archiv angeben zu können. Damit ist es möglich, mehrere in zeitlicher Abfolge geharvestete Schnappschüsse zusammen zu verwalten und Endnutzern komfortable Suchmöglichkeiten unter Einbeziehung der Zeitachse zu bieten.
Das IIPC sucht auch nach Lösungen, die oben genannten Defizite automatischer Web-Harvesting-Verfahren auszugleichen. Mit „DeepARC“[46] wurde ein grafischer Editor vorgelegt, der es erlaubt, Strukturen aus relationalen Datenbanken in ein XML-Schema abzubilden. Der Transfer wichtiger Inhalte aus dem deep web kann unter Nutzung dieses Tools durch bilaterale Vereinbarungen zwischen Datenbankbetreibern und Archiven geregelt und unterstützt werden.
Zusammenfassend drückt das folgende Schaubild aus, dass die Tools des IIPC explizit an den Belangen von Gedächtnisorganisationen ausgerichtet sind, die an der Langzeitarchivierung von WWW-Inhalten interessiert sind.

5 Ein Blick nach Deutschland

Eine Anzahl von Aktivitäten in Deutschland hat sich der Aufgabe „Langzeiter-

43 http://crawler.archive.org/ [HERITRIX] (Juni 2006)

44 http://archive-access.sourceforge.net/projects/nutch [NutchWAX] (Juni 2006)

45 http://archive-access.sourceforge.net/projects/wera [WERA] (Juni 2006)

46 http://deeparc.sourceforge.net [DeepARC] (Juni 2006)

haltung von Internetressourcen“ angenommen. Die Internetpräsenz des Projekts „nestor - Kompetenznetzwerk Langzeitarchivierung“[47] listet in der Rubrik „Projekte“ folgende Institutionen und Vorhaben auf, die sich im engeren Sinne mit der Sammlung und Archivierung von WWW-Ressourcen befassen: Parlamentsarchiv des Deutschen Bundestages, Baden-Württembergisches Online-Archiv, Digital Archive for Chinese Studies (Heidelberg), edoweb Rheinland-Pfalz, Archiv der Webseiten politischer Parteien in Deutschland und das Webseitenarchiv des Zentralarchivs zur Erforschung der Geschichte der Juden in Deutschland. Nähere Angaben und weiterführende Hinweise sind auf www.langzeitarchivierung.de zu finden.

Die Deutsche Nationalbibliothek hat in den vergangenen Jahren vor allem auf die individuelle Bearbeitung von Netzpublikationen und das damit erreichbare hohe Qualitätsniveau im Hinblick auf Erschließung und Archivierung gesetzt. Eine interaktive Anmeldeschnittstelle kann seit 2001 zur freiwilligen Übermittlung von Netzpublikationen an den Archivserver info-deposit.d-nb.de[48] genutzt werden. Im Herbst 2005 wurde zum Zeitpunkt der Wahlen zum Deutschen Bundestag in Kooperation mit dem European Archive[49] ein Experiment durchgeführt, um Qualitätsaussagen über die Ergebnisse aus fokussiertem Harvesting zu erhalten.

47 http://www.langzeitarchivierung.de [nestor - Kompetenznetzwerk Langzeitarchivierung] (Juni 2006)

48 http://info-deposit.d-nb.de [Archivserver der Deutschen Nationalbibliothek] (Februar 2007)

49 http://europarchive.org [European Archive] (Februar 2007)