Archivierung von Internetseiten : eine Standortbestimmung

Autor(en): **Locher, Hansueli**

Objekttyp: **Article**

Zeitschrift: **Geschichte und Informatik = Histoire et informatique**

Band (Jahr): **13-14 (2002-2003)**

PDF erstellt am: **16.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-118859

Archivierung von Internetseiten – eine Standortbestimmung

Hansueli Locher

Résumé

Les informations disponibles via Internet changent rapidement. La durée d'existence moyenne d'une page Web ne dépasse pas les 100 jours. Une telle fugacité de l'information va à l'encontre des efforts d'archivage.

Cette contribution évoque les expériences faites par «l'Internet Archive» avec sa «Wayback Machine», explique une stratégie intéressante choisie par la Bibliothèque nationale de France et donne une vue d'ensemble des tests que la Bibliothèque nationale Suisse (BNS)a effectués avec «Web Harvester». Elle présente finalement un aperçu des projets de la BNS concernant l'archivage de pages Internet.

Zusammenfassung

Informationen im Internet ändern rasch. Die durchschnittliche Lebensdauer einer Website beträgt knapp 100 Tage. Diese Flüchtigkeit der Information steht den Archivierungsbestrebungen entgegen.

Im Beitrag wird auf die Archivierungsbemühungen von «Internet Archive» mit ihrer «Wayback Machine» eingegangen, eine interessante Strategie der Bibliothèque nationale de France erläutert und ein Einblick in die aktuellen Tests der Schweizerischen Landesbibliothek (SLB) mit einem «Web Harvester» geboten. Zum Schluss erfolgt ein kurzer Ausblick auf die weiteren Pläne der SLB bezüglich der Archivierung von Internetseiten.

1. Einleitung

Lesen Sie schnell, denn nichts ist beständiger als der Wandel im Internet!
(Anita Berres, deutsche Publizistin)

Das Zitat von Anita Berres lässt sich durch Zahlen untermauern: Die mittlere Lebensdauer einer Website beträgt rund 19 Monate und die durchschnittliche Lebensdauer eines HTML-Dokuments gerade mal 100 Tage.[1]

Für die Archivierungsbemühungen hat das grosse Auswirkungen. Wenn wir die im World Wide Web gegenwärtig verfügbaren Dokumente nicht jetzt sammeln, sind sie für immer verloren.

Das ist aber einfacher gesagt als getan. Verschiedene weitere Eigenschaften des World Wide Web stellen sich dieser Absicht entgegen.

- Riesiger Datenumfang
  Für das Jahr 2002 wird der Datenumfang auf dem öffentlich zugänglichen Web ohne datenbankbasierte Webdokumente (Surface Web) auf 167 Terabytes geschätzt.[2] Das entspricht etwa dem Speicherplatz von 120 Millionen 3 1/2-Zoll-Disketten.
- Rasches Wachstum
  Schätzungen aus dem Jahr 2000 gehen von täglich 7,3 Millionen neuen Webdokumenten aus, die Speicherplatz in der Grösse von 0,1 Terabytes (100 Gigabytes) belegen.[3]
- Unübersichtlichkeit
  Vorgaben zur Datenorganisation oder zu inhaltlichen Strukturen fehlen gänzlich. Auf gegenwärtig 172 Millionen im Internet registrierten Servern sind irgendwelche Informationen zu finden, auf die mit verschiedensten Instrumenten zugegriffen werden kann. Das Angebot reicht dabei von Websites über Mail-Server, Datenbanken, FTP-Server bis hin zu Diskussionsgruppen und Chat-Foren.[4]

---

1 Stata, Raymie: Saving the Web (Vortrag an der European Conference on Digital Libraries 2002 in Rom), <http://webapp.bnf.fr/bibnum/ecdl/2002/ia/ia.html>

2 Lyman, Peter; Varian, Hal R. et al.: «How Much Information 2003», <http://www.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf>, 27. Oktober 2003.

3 Murray, Brian H; Moore, Alvin: «Sizing the Internet: A White Paper», Cyveillance, <http://www.cyveillance.com/web/corporate/white_papers.htm>, 10. Juli 2000, S. 2.

4 Quelle: Internet Software Consortium, <http://www.isc.org/ds/>.

Die wesentliche Herausforderung beim Archivieren von Websites liegt darin, die Dokumente zeitgerecht und mit vernünftigem Aufwand im Internet abzuholen und auf einem Archivsystem abzulegen. Das Aufbewahren an und für sich bietet genau die gleichen Probleme wie bei allen übrigen elektronischen Daten auch.

2. Verschiedene Ansätze zur Archivierung von Internetseiten

Verschiedene Stellen, insbesondere aber auch die Nationalbibliotheken, bemühen sich heute um die Archivierung von Websites. Ihre Strategien sind zum Teil sehr unterschiedlich. In den nachfolgenden Unterkapiteln soll auf zwei wichtige Initiativen in diesem Bereich eingegangen werden.

2.1 Internet Archive[5]

Internet Archive ist eine Nonprofit-Organisation, die 1996 mit dem Ziel gegründet wurde, eine «Internet-Bibliothek» zu erstellen, die sowohl Wissenschaftler/innen wie Forscher/innen und Historiker/innen als auch dem breiten Publikum den ständigen Zugriff zu historischen Sammlungen von digitaler Information bieten soll.

Seit der Gründung wurden rund 30 Milliarden Web-Dokumente archiviert und erschlossen. Das Sammeln geschieht in automatisierter Form durch WebCrawler von Alexa.[6] Gesammelt werden dabei im Prinzip alle öffentlich zugänglichen Files im World Wide Web.

Die Datenmenge im Internet Archive beträgt gegenwärtig rund 250 Terabytes. Das entspricht ungefähr 250 Millionen Büchern à je 200 Seiten. Das sind mehr Bücher, als seit der Erfindung der Buchdruckerkunst weltweit produziert worden sind.[7]

Pro Monat nimmt die Datenmenge um rund 10-12 Terabytes zu. Die einzelnen Dateien aus dem Internet werden dabei zu ARC-Files von je 100 Megabytes Grösse zusammengefasst und auf DLT-Tapes abgelegt.

Die Lebensdauer dieser DLT-Tapes beträgt etwa 30 Jahre. Internet Archive plant aber öfter als alle 10 Jahre ihre Archivdaten umzukopieren.[8]

---

5 Siehe Website Internet Archive, <http://www.archive.org>.

6 Informationen zu diesem WebCrawler: <http://pages.alexa.com/company/technology.html>. Die Firma Alexa gehört Amazon.com.

7 «Auffrischung des «nationalen Gedächtnisses». Grundzüge einer schweizerischen Memopolitik». In: *Neue Zürcher Zeitung*, 13. Mai 2003.

8 «About the Internet Archive», <http://www.archive.org/about/about.php>.

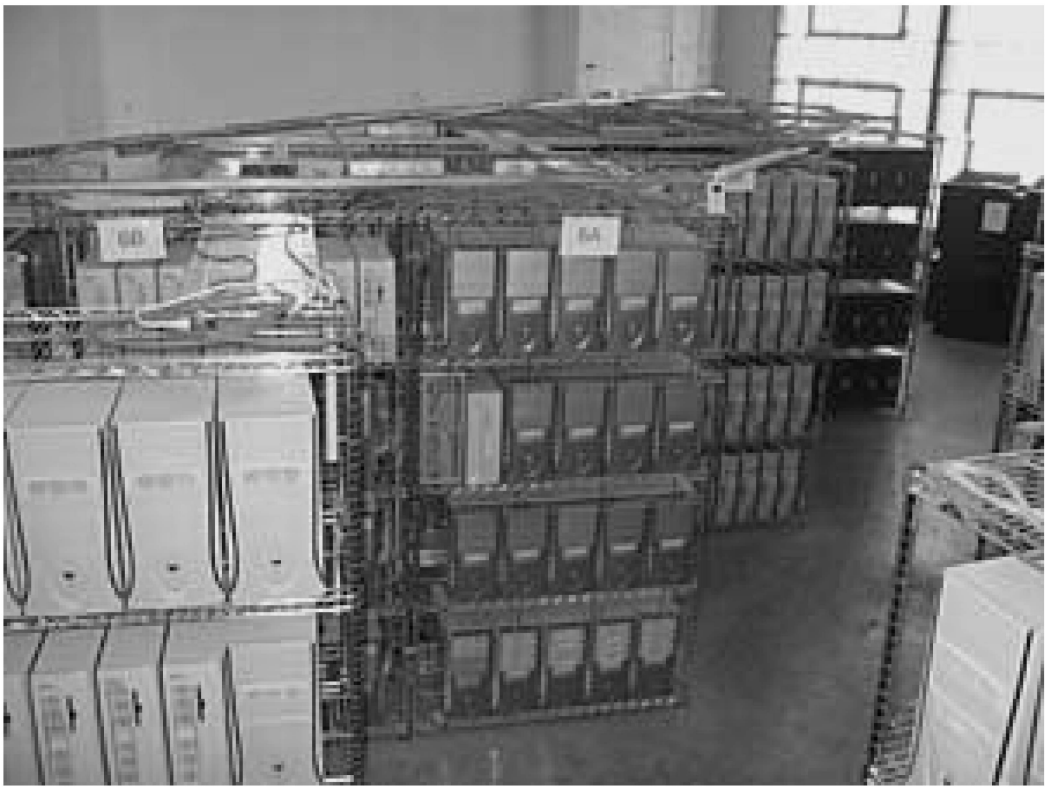

Abb. 1: Teil der Wayback Machine

Mit Hilfe der sogenannten Wayback Machine kann der Benutzer auf die archivierten Webseiten zugreifen, die ihm zu diesem Zweck auf den Harddisks von einigen 100 Servern zur Verfügung gestellt werden. Es genügt dabei die Adresse einer Website oder eines Webdokuments einzutippen, um einen Überblick über die archivierten Versionen zu erhalten, die via Link nun aufgerufen werden können.

2.2 Bibliothèque nationale de France (BnF)

Bei den Nationalbibliotheken stehen zwei Strategien zum Sammeln von Websites im Vordergrund. Die Nordländer führen ein Domain-Harvesting durch, das heisst, sie versuchen alle Websites, deren Domänen-Namen die Kennzeichnung ihres Landes führen, zu sammeln.[9] Andere wie beispielsweise die amerikanische Nationalbibliothek, die Library of Congress, setzen auf ein selektives Harvesting, bei dem ausgewählte Websites oder auch

9 Die nordischen Nationalbibliotheken (Dänemark, Finnland, Island, Norwegen und Schweden) arbeiten im Bereich des Web-Harvesting eng zusammen. Als Forum für Koordination und Erfahrungsaustausch dient das Nordic Web Archive (NWA).

möglichst alle Websites zu einem bestimmten Thema (Präsidentenwahlen 2000 in den USA, 11. September 2001) gesammelt werden.

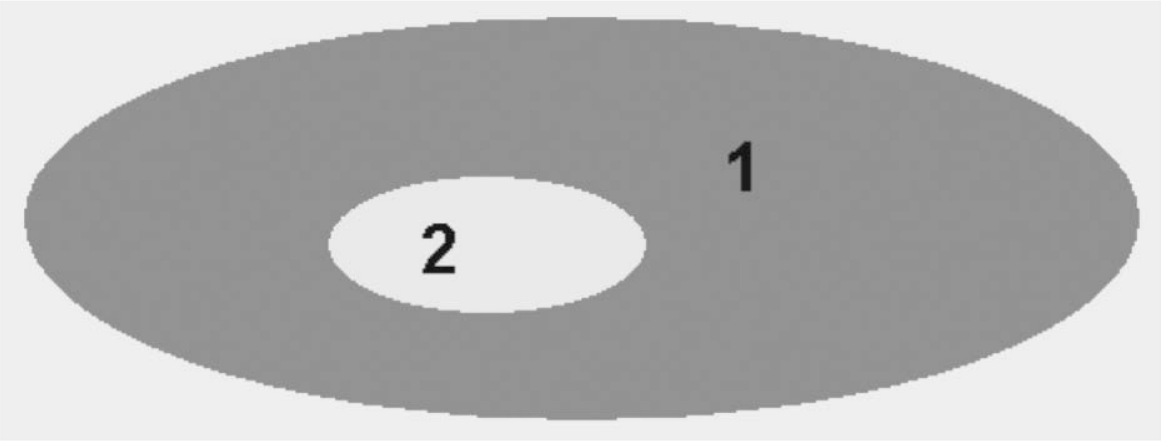

Abb. 2: Selektives Harvesting
1 Gesamte französische Webdomäne (.fr)
2 Ausgewählte Websites

Die Bibliothèque nationale de France (BnF) mischt diese beiden Ansätze. Die gesamte französische Webdomäne wird in einem Harvestig-Durchgang gesammelt. Dieses generelle Harvesting dient dazu, eine Link-Topologie zu erstellen. Damit lässt sich feststellen, auf welche Websites am häufigsten mittels Links referenziert wird. Diese Websites bilden dann zusammen mit weiteren Websites, die durch die Referenzbibliothekarinnen und -bibliothekare bestimmt werden, eine Auswahl. Diese wird genauer analysiert. Je nach Bedarf werden dann für diese Websites, die zur Auswahl gehören, spezielle Massnahmen vorgesehen.

– Bei Sites, die sich rasch verändern (z.B. Online-Zeitungen oder -Zeitschriften), wird der Harvesting-Prozess mit hoher Periodizität durchgeführt.
– Bei dynamischen Websites wird mit der verantwortlichen Stelle vereinbart, dass sie die Inhalte in regelmässigen Abständen auf einem Datenträger der BnF zur Verfügung stellt.

3. Tests der SLB

Die Schweizerische Landesbibliothek (SLB) hat sich im Rahmen des Projekts Networked European Deposit Library (NEDLIB)[10] an der Entwick-

10 Website NEDLIB <http://www.kb.nl/coop/nedlib>.

lung einer Software zum Sammeln von Internet-Dokumenten beteiligt. Dieser NEDLIB-Harvester wurde Ende 2002 mit Hilfe eines externen Dienstleisters auf einem Testsystem in der SLB installiert, um erste Erfahrungen im Umgang mit Websites sammeln zu können.

3.1 Funktionsweise des NEDLIB-Harvesters

Der NEDLIB-Harvester kann auf einer Linux- oder Unix-Plattform eingesetzt werden. Er benötigt eine MySQL-Datenbank. Diese wird sowohl für das Speichern von Informationen zu den archivierten Dokumenten als auch zur Konfiguration der Software verwendet.

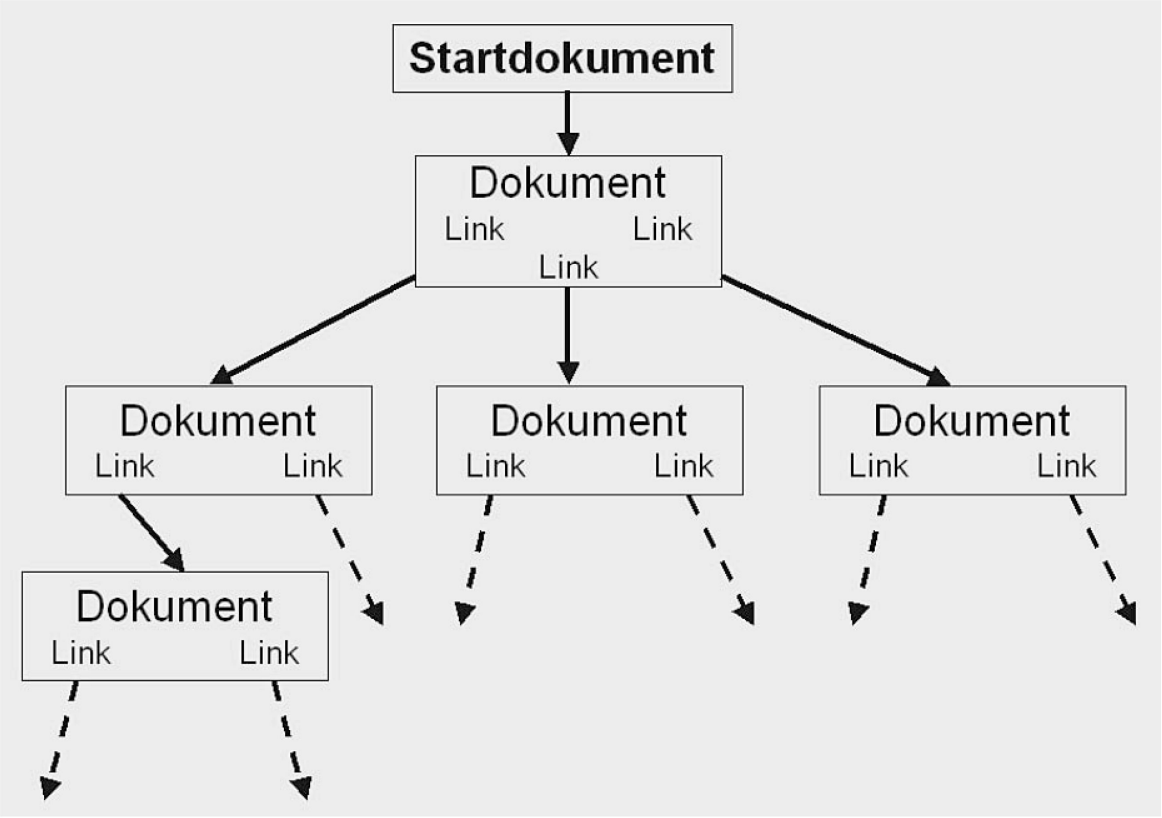

Abb. 3: Harvesting-Prozess

Die Startadresse (Link auf ein Dokument im Internet), von der aus der Sammelprozess durchgeführt werden soll, muss manuell konfiguriert werden. Ausgehend von dieser Adresse beginnt dann das Harvesting, indem den im Startdokument vorhandenen Links gefolgt wird, die zu weiteren Dokumenten führen. In diesen werden erneut die Links ausgewertet.

Dieser Prozess kann mit Hilfe einer vorgängigen Konfiguration auf bestimmte Websites oder Domänen beschränkt werden.

3.2 Die Tests

Der NEDLIB-Harvester wurde in der SLB auf einem normalen Pentium 4 (1,7 GHz Prozessor, 256 MB RAM) getestet, der auch als Arbeitsplatzgerät bei den Mitarbeitenden eingesetzt wird.

In einer ersten Testphase war der Harvester nach rund 40 Minuten und etwa 5000 archivierten Dokumenten überlastet. Es wurden kaum noch Dokumente archiviert. Der Grund dafür lag bei der MySQL-Datenbank, die in der Folge für grosse Datenmengen optimiert wurde.

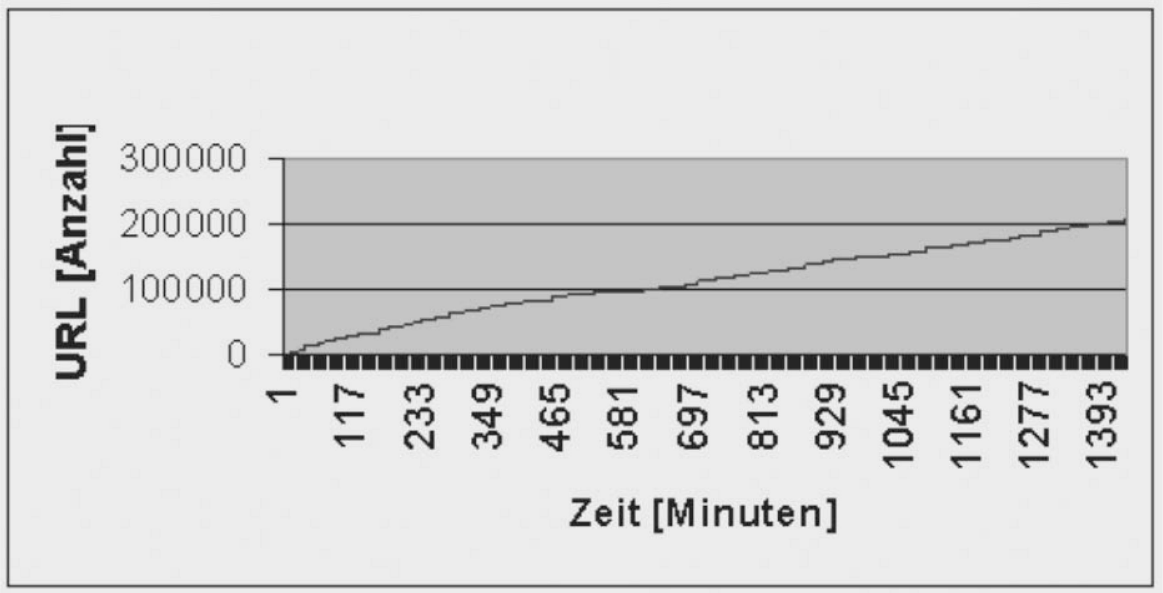

Abb. 4: Leistungsmessung beim Harvesting

Nun war es möglich innerhalb von 24 Stunden 206'000 Dateien mit einem Umfang von gut 8 GB zu sammeln. Die Zuwachskurve verlief linear (vgl. Abbildung). Die Leistungsfähigkeit des Systems blieb konstant hoch.

Als entscheidend für die Geschwindigkeit des Sammelprozesses stellten sich in dieser Testphase verschiedene Faktoren heraus:

– Leistungsfähigkeit der Datenbank
– Datendurchsatz von und zu der Harddisk
– Geschwindigkeit der Harddisk

Der Einsatz mehrerer Harddisks bringt einen weiteren Geschwindigkeitsgewinn.

3.3 Erkenntnisse

Durch den Einsatz eines leistungsfähigen Servers für den Sammel-Prozess und einer Konfiguration, die alle Möglichkeiten zur Optimierung des Pro-

zesses ausschöpft, müsste es möglich sein, rund 25 Gigabytes Daten pro Tag hereinholen zu können. Da die Datenmenge im .ch-Domain im Internet durch die SLB auf gut 2 Terabytes geschätzt wird, heisst das konkret, dass das Sammeln der ganzen Domäne mit dem NEDLIB-Harvester mindestens 80 Tage beanspruchen würde.

Die praktische Arbeit hat gezeigt, dass es dabei eine ganze Reihe von Punkten zu beachten gilt:

- Gewisse Datentypen müssen vom Sammelprozess ausgeschlossen werden können. Die SLB hat beispielsweise kein Interesse an Programm-Files, die in öffentlichen Software-Bibliotheken angeboten werden und mit ihrem Datenumfang den Prozess deutlich verlangsamen.
- Grosse Websites mit zum Teil gegen 100'000 Dokumenten müssen ganz am Anfang des Sammelprozesses bereits besucht werden, weil diese sonst die Dauer des Prozesses wesentlich verlängern können.
- Die gesammelten Datenfiles müssen mit einer eindeutigen, immer gleich bleibenden Identifikation (Persistent Identifier) versehen und verzeichnet werden, damit sie auffindbar bleiben.
- Sobald periodische Sammelprozesse vorgesehen werden, braucht es ein Versionenmanagement.
- Der NEDLIB-Harvester – wie übrigens praktisch alle anderen Harvester auch – kann keine Informationen sammeln, die nur über ein Abfrageformular zu erreichen oder durch Passwörter geschützt sind. Er beschränkt sich hauptsächlich auf statische Websites.

Der Zugriff auf die durch den NEDLIB-Harvester gesammelten Dokumente dürfte kein grosses Problem darstellen, da das Nordic Web Archive mit dem NWA Toolset[11] ein Instrument entwickelt hat, das den Benutzerinnen und Benutzern dafür zur Verfügung gestellt werden könnte.

4. Ausblick

Momentan ist die SLB daran, die Voraussetzungen dafür zu schaffen, die in Zukunft ein Web-Harvesting erlauben sollen. So wird ein Konzept für die Vergabe von Persistent Identifiers erarbeitet, die als Zugriffsadressen auch im Internet verwendet werden können. Die SLB arbeitet dabei mit Der Deutschen Bibliothek zusammen und wird Unified Resource Names (URN) auf der Basis von National Bibliographic Numbers (NBN) vergeben.

11 «About the NWA Toolset», <http://nwa.nb.no/aboutNwaT.php>, 5. September 2002.

Zusammen mit dem Bundesarchiv wird Speicherplatz von vorerst 30 Terabytes beschafft, um unter anderem auch die gesammelten Informationen aus dem Web ablegen zu können.

Voraussichtlich wird die SLB beim Sammeln und Archivieren von Websites zwar gelegentlich ein Domain-Harvesting betreiben, vor allem aber ausgewählte Websites archivieren.

Dies durchaus im Bewusstsein, dass es unmöglich ist, voraus zu sehen, was die Menschen von morgen interessieren wird.

Es gibt trotzdem verschiedene Gründe, die für den selektiven Ansatz sprechen:

- Genau wie im Printbereich sollen auch im Web-Bereich Sammelrichtlinien angewendet und diejenigen Websites aufbewahrt werden, die ihnen entsprechen.
- Die vorhandenen Mittel lassen sich bei einem selektiven Sammeln von Websites zielgerichteter einsetzen, als das mit einem regelmässigen Harvesting der .ch-Domäne möglich wäre. Dieses ist aus technischen Gründen (Zugriff auf Datenbanken, passwortgeschützte Information) ebenfalls nicht vollständig.

Neben der selektiven Sammlung von Websites wird aber auch eine enge Zusammenarbeit mit Verlagen oder anderen Produzenten von elektronischen Publikationen, die online oder auf Datenträgern verfügbar gemacht werden, angestrebt.