Wie gut ist gut genug? Qualität in der Webarchivierung

2. nestor-Expertengespräch zur Archivierung von Websites

von Sabine Schrimpf (Deutsche Nationalbibliothek, Tobias Beinert (Bayerische Staatsbibliothek) und Stefan Wolf (Bibliotheksservice-Zentrum Baden-Württemberg)

Mit dem Workshop „Qualitätskriterien und Möglichkeiten der Qualitätsverbesserung in der Webarchivierung“ am 13. Oktober 2011 an der Bayerischen Staatsbibliothek in München fand die im April in Frankfurt am Main begonnene Gesprächsreihe zur Archivierung von Websites ihre Fortsetzung. Während beim ersten Gespräch der Überblick über aktuelle Initiativen zur Archivierung von Websites und die Möglichkeiten und Grenzen der Webarchivierung im Allgemeinen im Mittelpunkt standen, wurde beim zweiten Treffen das Thema Qualitätsanforderungen an Webarchive in den Fokus genommen.

Einführend stellt Herr Steinke von der Deutschen Nationalbibliothek die Aktivitäten des International Internet Preservation Consortium (IIPC) vor. Da die meisten IIPC-Mitglieder derzeit auf die gleichen Werkzeuge zurückgreifen, zum Beispiel den Web-Crawler Heritrix zum Einsammeln und die Wayback Machine zur Bereitstellung archivierter Websites, können sie kooperativ an Verbesserungen der Werkzeuge arbeiten und nach und nach die Qualität ihrer Ergebnisse steigern. So wurde zum Beispiel das Harvesting von Videos als Schwachstelle erkannt und eine Lösung entwickelt, die zumindest das Einsammeln von Youtube-Videos ermöglicht.
Das IIPC unterstützt die Erarbeitung eines ISO-Reports „Web Archive Quality Issues and Statistics“, der einführenden und normierenden Charakter für die Archivierung von Websites haben soll. Er enthält qualitative und statistische Indikatoren. Als Qualitätsindikatoren für Webarchive werden dort aufgeführt:

1) Existiert eine klare Definition des institutionellen Sammlungsauftrags (Policy)?
2) Entspricht die Sammlung der Policy?
3) Ist der Workflow effizient (Ressourcen)?
4) Wie zugänglich und durchsuchbar ist das Archiv?
5) Ist die Langzeitverfügbarkeit durch geeignete Maßnahmen gesichert?

Der ISO-Report wird voraussichtlich 2012 veröffentlicht werden; bis dahin gibt diese Präsentation des Vorhabens einen guten Überblick.

Am Beispiel von literarischen Weblogs und Netzliteratur veranschaulicht Frau Schmidgall die Qualitätsanforderungen des Deutschen Literaturarchivs in Marbach an das Harvesting einzelner Websites. Die Sammlung wird von der Bibliothek des Literaturarchivs realisiert, die bei der Auswahl und dem Harvesting von Weblogs eng mit den Autoren kooperiert. Die Autoren unterstützen den betreuenden Bibliothekar häufig bei der manuellen Qualitätskontrolle und -verbesserung, so dass es – allerdings manchmal erst nach einigen Anläufen – meist gelingt, ein getreues Abbild der Originalwebsite ins Archiv zu holen. Den intensiven Betreuungsaufwand je Website muss das Literaturarchiv auf sich nehmen, weil es mit den Weblogs und Netzliteratur eine vom Umfang her überschaubare Spezialsammlung betreibt, die aber als zum Kernauftrag gehörend begriffen wird.
Als Herausforderung beschreibt Frau Schmidgall vor allen Dingen die Tatsache, dass jede Spiegelung/Snapshot einen Ausschnitt aus dem eigentlich grenzenlosen Internet bildet und der Umfang des Archivobjekts erst durch den Zeitpunkt und die Auswahl der zu archivierenden Bestandteile bestimmt wird. Jede Spiegelung kann somit als Unikat betrachtet werden. Bei Netzliteratur stellt sich noch mehr als bei Weblogs das Problem, dass hier oftmals der Gesamtkontext mitsamt Software archiviert werden muss, da die Ausführung der Software Bestandteil des literarischen Werks sein kann.

Die Anforderungen des Webarchivs des Südwest-Rundfunk, die Herr Fischer vorstellt, kontrastieren mit denen des Literaturarchivs. Anstelle der Qualität der einzelnen archivierten Websites steht die Qualität des Zugangs zur Sammlung für den SWR im Mittelpunkt. Das SWR-Archiv hat eine dreistellige Millionenzahl von digitalen Ressourcen archiviert, darunter Film- und Audio- und Textbeiträge. Der Bestand wird in erster Linie von Journalisten als Produktionsarchiv genutzt. Die Sammlung wird

derzeit lediglich durch die Wayback Machine erschlossen, die den Anforderungen der Nutzer zunehmend weniger genügt. Der SWR setzt daher Hoffnungen auf eine Volltextindexierung durch SOLR sowie auf neue Visualisierungsmöglichkeiten, die derzeit im EU-Projekt ARCOMEM erarbeitet werden.

Nach den Anforderungen an Webarchive werden die technischen Möglichkeiten und Grenzen am Beispiel der Web-Crawler HTTrack und Heritrix genauer vorgestellt. Herr Wolf vom Bibliothekszentrum Baden-Württemberg beginnt (in Vertretung für Herrn Naumann, Landesarchiv Baden-Württemberg) mit einem Überblick über den Crawler HTTrack. Dieser unterstützt das Abspeichern von Websites inklusive Layout und Navigation. Relativ unproblematisch speichert er Textinhalte, Fotos, PDFs u.ä. ab. Bei audiovisuellen Inhalten, interaktiven Elementen und Datenbankinhalten oder neueren Webtechnologien (z.B. JavaScript) stößt er an seine Grenzen. Die Archivpakete speichert HTTrack bislang in einem internen Format ab, die Möglichkeit zum Umstieg auf das standardisierte ARC/WARC-Format ist allerdings seit kurzem gegeben. Das Portugiesische Nationalarchiv hat diesen Umstieg gerade vollzogen, das BSZ wird mit allen Anwendern der Software SWBcontent den Umstieg von HTTrack auf Heritrix beraten und den Beschlüssen folgend vollziehen. Eine Volltextsuche ist derzeit nicht in SWBcontent integriert, da die Archivfunktion gegenüber der Benutzungsfunktion priorisiert wurde.

Frau Kugler von der Bayerischen Staatsbibliothek gibt einen Überblick über den Crawler Heritrix. Wie HTTrack speichert er Websites inklusive Layout und Navigation ab, allerdings direkt im standardisierten ARC/WARC-Format. Schwierigkeiten hat Heritrix ebenso wie HTTrack mit Flash-Dateien, JavaScript, eingebetteten Video-Streams und Datenbankinhalten. Die Bayerische Staatsbibliothek hat bereits zehn MitarbeiterInnen geschult, die seit einigen Monaten Web-Crawls mit Heritrix über das Web Curator Tool durchführen. Die Erfahrungen zeigen, dass die Entwicklung von Qualitätskriterien unerlässlich ist, um eine möglichst einheitliche Kollektion an Webarchiven erzielen zu können. Aktuell stellt sich z.B. die Frage, fob eine geharvestete Seite verworfen werden soll, wenn zwar alle textuellen Inhalte einer Website eingesammelt wurden, aber Bilder und/oder Layout fehlen. Die Frage wird von den anwesenden Experten nicht eindeutig beantwortet, jedoch der Hinweis gegeben, dass diese Fehlerart generell oftmals minimiert werden kann, wenn die Crawl-Einstellungen so gesetzt werden, dass „robots.txt“ ignoriert wird. In dem Kontext kommt die Frage nach dem rechtlichen Status von „robots.txt“ auf. Da sie nicht eindeutig beantwortet werden kann, soll sie an die nestor-AG Recht weitergegeben werden. Weitgehende Einigkeit herrscht allerdings darüber, dass robots.txt ignoriert werden kann, wenn der Website-Betreiber sein Einverständnis zum Harvesting der Seite gegeben hat.

Abschließend gibt Herr Masanès von der Internet Memory Foundation einen Ausblick auf die internationale Forschung im Bereich Webarchivierung. Er identifiziert vier zentrale Themen, um die Forschungsaktivitäten geclustert werden können:

1) Entwicklung von Webarchiverungs-Dienstleistungen an zentralen Knotenpunkte („Platform Approach“) und Auslagerung der technologisch komplexen Workflows aus den Institutionen, die Webarchivierung betreiben.
2) Intelligentes Crawlen („Application-aware crawling“), das durch die Auswertung von Social Media auch aktuelle thematische Trends erkennt und beachtet.
3) Automatisierte Qualitätskontrolle, z.B. durch die Erzeugung von Images der zu archivierenden Originalwebsite und der archivierten Websites und Vergleich dieser beiden Images mit Hilfe eines Algorithmus.
4) Verbesserung des Zugriffs auf große Webarchive (“Mining Web Archives”)

Insgesamt zeigten die Vorträge und die lebhaften Diskussionen im Anschluss, dass der Begriff der Qualität bei der Archivierung von Websites ganz unterschiedliche Facetten und Aspekte hat, die von unterschiedlichen Organisationen anhand vielfältiger Faktoren bewertet werden. Wichtig erschien den TeilnehmerInnen vor allem, dass zukünftig auch die Qualität des Zugangs verstärkt in die Überlegungen auf institutioneller Seite einfließt und somit die Nutzer von Webarchiven stärker in den Blick nimmt.

In einer kurzen Abschlussdiskussion wird das weitere Vorgehen der Expertenrunde festgelegt. Das Treffen wurde von den Anwesenden als produktiv empfunden und ein weiteres Treffen, wiederum zu einem Schwerpunktthema, wird gewünscht. Um den Erfahrungsaustausch möglichst effektiv zu gestalten, scheint es sinnvoll, den Teilnehmerkreis auf Institutionen beschränkt zu halten, die bereits in der Webarchivierung aktiv sind. Um daneben auch ein größeren Kreis an interessierten Personen zu adressieren und eine breitere Öffentlichkeit für das Thema zu gewinnen haben, soll zusätzlich 2012 auch eine größere Informationsveranstaltung evtl. in Kooperation von nestor und dem AWV-AK 6.2 „Archivieren von Internetpräsenzen“ angeboten werden.