Internationaler Fachbericht für die Archivierung des Internets

Roswitha Poll

Seit mehr als einem Jahrzehnt haben Bibliotheken damit begonnen, das Internet zu „archivieren". Vor allem Nationalbibliotheken selektieren, sammeln und bewahren Internetressourcen bezogen auf ihre nationale Domain und sehen dies im Zusammenhang mit der traditionellen Pflichtexemplar-Sammlung. Da Umfang und Methodik der Webarchivierung variieren, sind Größe und Qualität der verschiedenen Webarchive nur schwer vergleichbar. Um Vergleichbarkeit besser zu ermöglichen, hat ISO TC 46 SC 8 einen Fachbericht erarbeitet, der Terminologie und Statistik der Webarchivierung normiert und Qualitätsindikatoren für die Evaluierung anbietet. Der Artikel beschreibt Ziele und Inhalt des ISO-Fachberichts.

For more than a decade, libraries have started to „archive the web". National libraries in particular select, collect and store Internet resources related to their national domain, seeing this as a task similar to traditional legal deposit. The collection policies and collecting methods vary, so that it is difficult to compare the quantity and quality of the respective web archives. In order to make comparison possible, ISO TC 46 SC 8 has produced a Technical Report that standardizes the terminology and statistics and offers tested indicators for assessing the quality of web archiving. The paper describes the aims and contents of the ISO Report.

Im Dezember 2013 erschien ein internationaler Fachbericht zur Statistik und Qualitätsbewertung der Webarchivierung:

ISO Technical Report 14873: Information and documentation – Statistics and quality issues for web archiving

Der Fachbericht entstand im Rahmen von ISO TC 46 SC 8 Qualität – Statistik und Leistungsbewertung. Das Komitee entwickelt Normen und Berichte, die sich mit der quantitativen und qualitativen Bewertung von Bibliotheken und anderen Informationseinrichtungen befassen. Die wichtigsten Normen sind:

- ISO 2789: 2013 Internationale Bibliotheksstatistik (5. Ausgabe)
- ISO 11620: 2014 (in Publikation) Leistungsindikatoren für Bibliotheken (2. Ausgabe)
- ISO 16439: 2014 (in Publikation) Methoden und Verfahren zur Bestimmung der Wirkung von Bibliotheken.

Die Normen werden in regelmäßigen Abständen revidiert, um vor allem neue Aufgaben und Methoden der Bibliotheken zu berücksichtigen. Der neue Leistungsbereich der Webarchivierung schien allerdings zu umfangreich und auch noch zu sehr in Entwicklung begriffen, um sofort in die vorhandenen Normen integriert werden zu können. Es bot sich an, zunächst einen Fachbericht (Technical Report) herauszugeben, der Definitionen und Methoden zur Evaluierung der neuen Aufgabe enthält.

Ende 2009 wurde eine ISO-Arbeitsgruppe eingerichtet, in der Experten aus Nationalbibliotheken und einer Forschungsbibliothek zusammenarbeiteten; die Leitung lag bei der französischen Nationalbibliothek. Eine enger Kontakt, zum Teil auch personelle Überschneidung, bestand zu IIPC (International Internet Preservation Consortium).

Webarchivierung

Webarchivierung bedeutet Selektion und Sammlung von Internetressourcen aller Art, die dann in Webarchiven gespeichert und dauerhaft erhalten und zugänglich gemacht werden. Das Einsammeln der Ressourcen wird überwiegend automatisch und in regelmäßigen Abständen durch Harvesting-Software (Crawler) durchgeführt, die nach vorgegebenen Regeln und/oder Listen das Internet durchsuchen.

Das Harvesting geschieht entweder selektiv oder umfassend. Das umfassende Harvesting (bulk crawl) geht nach definierten formalen Regeln vor; z. B. wird eine bestimmte nationale Domain gesammelt (.fr oder .de). Beim selektiven Sammeln sind Adressen von Websites vorgegeben, entweder zu bestimmten Themenbereichen wie Kultur oder Politik eines Landes, oder es werden alle Internetressourcen zu bestimmten Ereignissen gesammelt (event harvesting). Selektives Harvesting wird in der Regel in häufigeren Intervallen durchgeführt als umfassendes, bei dem z.B. halbjährlich eine „Momentaufnahme" erstellt wird.

Die Aufgabe der Webarchivierung steht meist im Zusammenhang mit der traditionellen Sammlung und Bewahrung des kulturellen Erbes, z. B. durch Ablieferung von Pflichtexemplaren. Die Begründung für das Sammeln ist in beiden Fällen die gleiche: die Gefahr, dass Publikationen – sowohl gedruckte wie Websites – für die Nachwelt verloren gehen könnten. Die Gefahr des Verlustes ist dabei für Inhalte des Internets sicherlich noch größer als für die traditionellen Publikationen; das rasche Verschwinden von aktuellen Websites ist bekannt. In vielen Ländern sind für die Webarchivierung bereits Rechtsstrukturen ähnlich dem Pflichtexemplarrecht entstanden. Die entstehenden nationalen Webarchive sollen nicht nur das kulturelle Erbe erhalten, son-

dern auch den dauerhaften Zugang dazu für Forschung und andere Interessen garantieren.

Webarchivierung begann Ende der 1990er-Jahre, vor allem in Nationalbibliotheken, zum Teil auch in nationalen Archiven oder in Institutionen mit speziellen Sammelaufträgen. Von Anfang an war klar, dass diese ungeheure Aufgabe kooperativ angegangen werden sollte. Daher wurde schon 2003 das IIPC gegründet (International Internet Preservation Consortium). Das Konsortium kümmert sich um die rechtlichen und vor allem die technischen Fragen der Webarchivierung.[1] So war IIPC z.B. stark an der Entwicklung des WARC-Formats beteiligt, das als ISO 28500 der Standard zur Verknüpfung, Speicherung und Verarbeitung von Web-Inhalten wurde.[2]

ISO TR 14873

Für die Sammlung und Archivierung von Internetressourcen werden aber nicht nur technische Grundlagen benötigt. Diese neue Aufgabe ist so komplex, riesig und auch kostspielig, dass die Ergebnisse unbedingt quantitativ erfasst und qualitativ evaluiert werden sollten, um Geldgeber und Öffentlichkeit zu informieren und vom Sinn der Aktivitäten zu überzeugen.

ISO/Technical Report 14873 hat den Zweck, Methoden für die Erhebung statistischer Daten und Indikatoren für die Qualitäts-Evaluierung anzubieten. Der Report richtet sich durchaus nicht nur an die Spezialisten der Webarchivierung, sondern auch an das Management der sammelnden Institutionen und an übergeordnete und finanzierende Institutionen. Daher ist mehr an Information über die Praktiken und Probleme der Webarchivierung zu finden, als für die eigentlichen Experten notwendig wäre. Es werden nicht nur die verschiedenen Harvesting-Formen beschrieben, sondern auch die Möglichkeiten der Erschließung: grundlegend durch URL, weitergehend durch Volltextindexierung, automatische Extraktion von Stichwörtern und Metadaten oder sogar – für spezielle Bereiche – Katalogisierung im traditionellen Sinn.

Außerdem geht der Report auf die Probleme der Langzeitarchivierung ein und schließlich auch auf die rechtlichen Grundlagen. Webarchivierung ist meist in der nationalen Gesetzgebung zu Pflichtablieferung und/oder Copyright geregelt. Dabei können bestimmte Webressourcen ausgenommen sein, oder das Sammeln ist nur erlaubt, wenn vorher die Erlaubnis der Rechteinhaber eingeholt wurde. Solches von Genehmigungen anhängiges Sammeln ist allerdings nur als selektives Harvesting möglich. Um dennoch umfassend sammeln zu können, wird manchmal die Genehmigung für die Sammlung und Bereitstellung einfach vorausgesetzt; die jeweilige Ressource wird aber auf Verlangen des Rechteinhabers wieder aus dem Archiv genommen.

Grundlegend für den Report war, dass die Terminologie der Webarchivierung vereinheitlicht und festgelegt wurde. Obgleich die sammelnden Institutionen ähnliche Harvesting-Methoden und oft sogar die gleiche Software anwenden, nutzen sie häufig unterschiedliche Termini. So wird z.B. die Kopie, die eine Harvesting-Software zu einem bestimmten Zeitpunkt von einer Website macht, als „version“, „archive“, „instance“ oder „capture“ bezeichnet. Der Report bringt nun „capture“ als Vorzugsbezeichnung. Der Hauptteil des Reports beschäftigt sich mit der Statistik und den Qualitätsindikatoren für die Webarchivierung.

Statistik der Webarchivierung

ISO 2789[3] nennt folgende Ziele für eine Bibliotheksstatistik:

- Vergleich mit Normen oder Daten ähnlicher Institutionen;
- Verfolgen der Bibliotheksleistung über mehrere Jahre;
- Grundlage für Planung, Entscheidungsfindung und Verbesserungen;
- Information der nationalen oder regionalen Organisationen, die die Bibliothek tragen, finanzieren oder kontrollieren;
- Nachweis des Wertes der Bibliotheksdienste für heutige oder auch zukünftige Nutzer.

Diese Ziele verfolgt auch der Report für die Webarchivierung, wobei hier der Wert für zukünftige Nutzer eine besondere Rolle spielt. Die wichtigsten Statistiken sind die für Größe und Inhalte des Webarchivs und für dessen Nutzung.

Zählung von Größe und Zuwachs des Webarchivs

Statistik für Webarchive fragt natürlich zuerst nach Größe und Zuwachs des Archivs, dann nach den Inhalten. Die Dimension der Webarchive verlangt allerdings andere Zähleinheiten und -methoden als die traditionellen Medien. Zählungen erfolgen schon wegen der Kosteneffizienz unbedingt automatisiert, und die Software, die für das Harvesting, das Indexieren oder die Recherche benutzt wird, kann die Ergebnisse beeinflussen. Problematisch ist, dass ein „Dokument“ im Netz kaum definierbar ist (ein File, eine HTML-Seite mit vielen verschiedenen Files, ein pdf?). Auch die Zahl

1 http://netpreserve.org/ [8. Januar 2014]

2 ISO 28500:2009 Information and documentation – WARC file format

3 ISO 2789: 2013 Information and documentation – International library statistics

der im Webarchiv gespeicherten Websites ist technisch nicht eindeutig ermittelbar, da die Website eher eine intellektuelle Einheit darstellt.
Der Report empfiehlt aufgrund der bisherigen Praxis die folgenden Zähleinheiten für Größe und Zuwachs:

- Targets (für die Archivierung ausgewählte intellektuelle Einheiten, normalerweise Websites). Targets können nur bei selektivem Harvesting gezählt werden.
- captures (erfasste targets). Die Zahl hängt von der Frequenz des Harvestings ab.
- Domains
- URLs
- Bytes. Diese Zählung ist für die Planung der Speicherung wichtig.
- WARC files. Files werden häufig in „Container-Files" gespeichert, weil diese großen Einheiten leichter zu handhaben sind (Speicherung, Datenaustausch, Langzeiterhaltung).

Zählung nach Inhalten des Archivs

Der Inhalt des Webarchivs kann untergliedert werden

- nach geographischer Verteilung (nationale Domains)
- nach Sprachen
- nach Formattypen (z.B. Text, Bild Audio) oder File-Formaten (z.B. html, jpeg)
- nach dem Zeitpunkt der Erfassung.

Untergliederung nach Formattypen ist besonders für die Planung der Erhaltung nützlich. Der Erfassungszeitpunkt zeigt die chronologische Abdeckung des Webarchivs. Ältere Ressourcen könnten inzwischen schon aus dem Netz verschwunden sein, oder ihr Format könnte obsolet sein.

Nutzungsstatistiken

Der Zugang zu den Archiven ist sehr unterschiedlich geregelt, gesetzlichen Bestimmungen oder Regelungen der sammelnden Institution entsprechend. Der Zugang kann auf bestimmte Teile des Archivs beschränkt werden, oder er ist nur innerhalb der Institution möglich. Von diesen Voraussetzungen hängt die Zählung der Nutzungen ab. Bei freiem Zugang können die Nutzungen über Webanalyse ermittelt werden. Kann das Archiv nur innerhalb der Institution, z.B. im Lesesaal einer Nationalbibliothek, genutzt werden, dann können die Benutzer nicht nur gezählt, sondern ggf. auch nach ihren Zielen und ihrem Erfolg befragt werden.
Der Report empfiehlt als grundlegend folgende Statistiken:

- Seitenaufrufe (page views)
- virtuelle Besuche (visits/sessions)
- unterschiedliche, nicht mehrfach gezählte Besucher (unique visitors)

Außer den bereits genannten Statistiken beschreibt der Report auch Daten zur Bestandserhaltung des Archivs und zu den Kosten der Webarchivierung.

Qualitätsindikatoren für die Webarchivierung

Der Report definiert zunächst die Kriterien für die Qualität der Webarchivierung:

- Ziel, Umfang und Inhalt der Sammeltätigkeit müssen klar definiert sein.
- Der angestrebte Umfang und Inhalt soll auch tatsächlich erreicht werden.

- Die Webarchivierung soll kosteneffizient erfolgen.
- Das Archiv soll zugänglich sein und effektive Suchmöglichkeiten anbieten.
- Es soll langzeitig verfügbar bleiben.

Zu diesen Kriterien wurden 15 Indikatoren entwickelt, mit denen sich die Qualität ermitteln lässt. Die Beschreibung der einzelnen Indikatoren folgt weitgehend dem Muster in der ISO-Norm 11620, die Leistungsindikatoren für Bibliotheken anbietet.[4] Ein Beispiel zeigt die Elemente der Beschreibung:

Indikator Nr.	*7*
Name	*Prozentsatz volltextindexierter Ressourcen*
Ziel	*Die Durchsuchbarkeit des Webarchivs zu ermitteln*
Erforderliche Daten	*- Gesamtzahl der Ressourcen in Webarchiv* *- Zahl der durch Volltextindexierung erschlossenen Ressourcen* *Die Berechnung kann auf der Basis von URLs oder Bytes erfolgen.*
Methode	*Der Prozentsatz volltextindexierter Ressourcen ist: **A/B** * **100**w wobei **A** die Zahl der durch Volltextindexierung erschlossenen Ressourcen ist, **B** die Gesamtzahl der Ressourcen in Webarchiv ist.* *Es wird auf eine Dezimalstelle gerundet.*
Kommentar	*Volltext-Recherchen erhöhen die Zugänglichkeit und Nutzbarkeit des Webarchivs erheblich. Die für die Berechnung benutzte Einheit (URL oder Byte) sollte angegeben werden.*

Im Folgenden werden einige der wichtigsten Indikatoren genannt.

Die Qualität des Webarchivs wird daraus ermittelt, ob der vorgegebene Sammelbereich auch tatsächlich umfassend gesammelt wurde und welcher Prozentsatz der Ressourcen im Archiv inzwischen aus dem Netz verschwunden ist. Der zweite Indikator unterstreicht besonders klar die Bedeutung der Webarchivierung.

Die Zugänglichkeit und Nutzbarkeit des Archivs wird belegt durch den Anteil, der Endnutzern direkt zugänglich ist, und den Prozentsatz volltextindexierter Ressourcen (s. oben).

Eine gute Strategie für Bestandserhaltung zeigt sich durch den Prozentsatz der Ressourcen, die wenigstens einmal reproduziert wurden, deren Fileformat identifiziert wurde und die auf Viren gecheckt wurden. Effizientes Management schließlich wird gemessen an den Kosten pro gesammelte URL.

Probleme der Datenerhebung

Die in dem Report beschriebenen Statistiken und Qualitätsindikatoren sind hier nur sehr komprimiert und auch nicht vollständig dargestellt. Vor allem sind die Probleme nicht erwähnt, die mit der Erhebung fast aller Daten verbunden sind. Die Formate der Webressourcen verändern sich rasch, auch die Software und die Sammelroutinen für die Webarchivierung sind in Entwicklung begriffen, und häufig können die für die Qualitätsbewertung sinnvollsten Daten nicht oder nur unvollkommen erhoben werden. Dann wurde ausgewählt, was machbar erschien.

Der Report fasst von Praktikern der Webarchivierung getestete Statistiken und Indikatoren zusammen, die dadurch nun in breiterem Rahmen genutzt werden können. Es wäre sinnvoll, bei der nächsten (3.) Ausgabe der Norm für Leistungsmessung (ISO 11620) Indikatoren in TR 14873, die sich als besonders effektiv erweisen, in diese Norm zu übernehmen.

Literaturverzeichnis

Ball, Alex: Web Archiving (version 1.1), Digital Curation Centre Edinburgh, UK, 2010. http://www.dcc.ac.uk/sites/default/files/documents/reports/sarwa-v1.1.pdf [12. Januar 2014}

Bermes, Emanuelle/ Illien, Gildas: Metrics and Strategies for Web Heritage Management and Preservation, 2009. http://conference.ifla.org/past/ifla75/92-bermes-en.pdf [12. Januar 2014]

Hockx-Yu. Helen/ Crawford, Lewis/ Coram, Roger/ Johnson, Stephen: Capturing and Replaying Streaming Media in a Web Archive – a British Library Case Study. In: Proceedings of iPRES, September 2010. http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/hockxyu-44.pdf [12. Januar 2014]

Jacobsen, Grethe: Web Archiving: Issues and Problems in Collection Building and Access, in: Liber Quarterly 18 (2008), S.366 – 370. http://liber.library.uu.nl/index [12. Januar 2014]

Masanès, Julien (Hrsg.): Web Archiving, Berlin 2006.

Stirling, Peter/ Illien, Gildas: The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future, 2011. http://conference.ifla.org/past/ifla77/193-stirling-en.pdf [12. Januar 2014]

Dr. Roswitha Poll

bis 2014 Vorsitzende von ISO TC 46 SC 8 Quality – statistics and performance evaluation

pollr@uni-muenster.de

4 ISO 11620:2014 (in Publikation), Information and documentation – Library performance indicators