Das Netz vergisst nichts? Herausforderungen der Webarchivierung

Jens Crueger
Digital-Historiker.de

„Das Netz vergisst nichts“

(ZEIT Wissen Nr. 5/2011)

„Partyfotos, Gerüchte, Diffarmierungen – das Netz speichert jedes noch so peinliche Detail für die Ewigkeit. Wer unliebsame Daten tilgen will, braucht oft professionelle Hilfe.“

(ZEIT Wissen Nr. 5/2011)

„Das Netz vergisst doch“

(DIE ZEIT Nr. 40/2012)

„Täglich gehen im Internet wichtige Daten verloren – nur leider nicht jene, auf die wir gerne verzichten würden.“

(DIE ZEIT Nr. 40/2012)

Was uns droht…

Dark Ages des Internet

☹ Web *„—elementally— ethereal, ephemeral, unstable, and unreliable“*, *„never-ending present“* (Lepore 2015)

☹ Durchschnittliche Bestandsdauer einer Webseite ~100 Tage

☹ Dynamische Webinhalte (Nachrichtenseiten u.ä.) verändern sich alle ~7min

☹ Leichte Löschbarkeit und Veränderbarkeit macht Webseiten manipulationsanfällig

☹ Frühes WWW kaum noch erhalten

1. Anforderung an gelingende Webarchivierung: Schnelle und konsequente Umsetzung, sonst drohen weitere Dark Ages

„Was der einen Epoche Abfall ist, ist der anderen kostbare Information.“

(Aleida Assmann, Erinnerungsräume, 245)

Geschichtswissenschaft des Digitalen

- *„a researcher in the twenty-second century, it will seem unimaginable that someone studying the twenty-first century would do anything but draw heavily on the online world to tell them about peoples' changing lives."*
- *„an almost untapped source for research."*

- Schroeder, Ralph u. Brügger, Niels: Introduction: The Web as History, in: The Web as History. Using Web Archives to Understand the Past and the Present. Hrsg. von Niels Brügger and Ralph Schroeder. London 2017. S. 1-19. S. 1.

Geschichtswissenschaft des Digitalen

Status quo

- Kaum Forschung mit genuin digitalen („digital born“) Quellen
- Es fehlen Methodiken und Methodologie (Quellenkritik usw.) dafür
- Institutionalisierung der digitalen Geschichtswissenschaft am Anfang
- Geschichtsschreibung der „letzten fünf Minuten“ gilt in der Geschichtswissenschaft seit jeher als heikel (Vergessensprozess nach Kosellek)

2. Anforderung an gelingende Webarchivierung: Nachfrageseite muss Bedarfe und Präferenzen definieren

„Google‘s Mission is to organize the world‘s information and make it universally accessible and useful.“

(Google Mission Statement)

☺ 2001 kauft Google die Deja archives. Daraus wird **Google Groups**, die größte Sammlung archivierter Usenet-Nachrichten, zurückreichend bis 1981

☺ 2004 startet **Google Books**, Ziel: Jedes bekannte Buch zu retrodigitalisieren. Partnerschaften mit Bibliotheken und Entwicklung eines eigenen Buchscanners (1.000 Seiten pro Stunde).

☺ 2006 startet **Google News Archive**, Zeitungsartikel 200 Jahre zurückreichend. Seit 2008 verstärkte Retrodigitalisierung von Zeitungen.

☹ Nach mehreren Redesigns ist **Google Groups** unbrauchbar zu Forschungszwecken. Bspw. Keine Suche nach Datum mehr möglich.

☹ **Google News Archives** gibt es seit 2011 nicht mehr.

☹ **Google Books** droht einzuschlafen; 2012 nur noch ca. 50% der Digitialisierungsrate von 2006; letzter Blogeintrag 2012.

☹ **Google Search** verlor 2011 seine Timeline Ansicht, die eine Suche nach Datum ermöglichten; Änderungen in den Algorithmen bevorzugen neuere vor älteren Webinhalten.

3. Anforderung an gelingende Webarchivierung: Trägerform nachhaltig und verlässlich wählen

Wer archiviert das Web?

- Nationalbibliotheken und Archive (Deutsche Nationalbibliothek, Library of Congress)
- Landes- und Kommunalarchive
- Firmen- und Verbandsarchive
- Grasswurzel-Archive (Special Interest Groups u.a.)

- Internet Archive (WayBack Machine)

„Bizarre Verordnung: Nationalbibliothek will das deutsche Internet kopieren“

(SPON Netzwelt, 23.10.2008)

Deutsche Nationalbibliothek

- Eventharvesting
- selektives Webharvesting (seit 2012 regelmäßig >1.200 Webseiten)
- Domain.de-Crawl (2014; 120 TB)
- Keine gezielte Auswahl von Quellen, keine Selektionskriterien
- Private Nachrichten werden nicht archiviert
- Archivierung als Snapshot mit Metadaten
- „Pragmatische Lernschritte“
- Nutzung nur im Lesesaal der DNB in Frankfurt bzw. Leipzig möglich

WayBack Machine

- Trump-Faktor
- Standorte in Kanada zum Schutz der Daten
- Im Rahmen des Übergangs Obama-Bush wurden 200TB Daten von Regierungswebseiten gesichert
- Eigenes “Trump Archive” wurde initiiert

Freilichtmuseum des World Wide Web?

4. Anforderung an gelingende Webarchivierung: Zugriff möglichst offen und ubiquitär verfügbar

„Future Historians Probably Won‘t Understand Our Internet, and That‘s Okay.“

(Alexis C. Madrigal, 2017)

›So, maybe our times are not so different from previous eras. Lynch himself points out that “the rise of the telephone meant that there were a vast number of person-to-person calls that were never part of the record and that nobody expected to be.” Perhaps Facebook communications should fall into a similar bucket. For a while it seemed exciting and smart to archive everything that happened online because it seemed possible. But now that it might not actually be possible, maybe that’s okay.“
Is it terrible that not everything that happens right now will be remembered forever?” Seaver said. “Yeah, that’s crappy, but it’s historically quite the norm.”‹

Alexis C. Madrigal, 2017

›This paper explores pragmatic approaches that might be employed to document the behavior of large, complex socio-technical systems (often today shorthanded as “algorithms”) that centrally involve some mixture of personalization, opaque rules, and machine learning components. Thinking rooted in traditional archival methodology — focusing on the preservation of physical and digital objects, and perhaps the accompanying preservation of their environments to permit subsequent interpretation or performance of the objects — has been a total failure for many reasons, and we must address this problem.‹

Clifford Lynch, 2017

›Idealized documentation of systems in the Age of Algorithms‹

Clifford Lynch, 2017

›Documenting instead of archiving: Pragmatic approaches‹

(Clifford Lynch, 2017)

- Robotic Witnesses
- Crowd Sourcing
- New Nielson Families
- Session Filming

5. Anforderung an gelingende Webarchivierung: Datenmengen und Algorithmizität gewachsen

›Manual Collection / Submission

Manuelle Sammlung wird einerseits für Websites verwendet, die nicht durch Crawler automatisch erfassbar sind. Dabei handelt es sich meist um Websites, die aus Datenbanken (Content Management Systemen) generiert werden, die nicht durch Linkstrukturen navigierbar sind, sondern z.B. nur ein Abfrage-Interface zur Verfügung stellen (Deep Web, siehe „Sonderformen“ unten). (…)‹

nestor Handbuch

Surface Web

Über gängige Suchmaschinen auffindbare und zugängliche Webressourcen.

Deep Web

Nicht frei zugängliche, sondern passwortgeschützte Inhalte.
Das Deep Web umfasste bereits um die Jahrtausendwende das 400- bis 550fache des seinerzeit frei zugänglichen Surface Web, gegen Ende der ersten Dekade dann gar das tausendfache.

6. Anforderung an gelingende Webarchivierung: Deep Web, Apps, AI, AR und weitere technologische Trends mitdenken

›National Web Archive to preserve top ten websites chosen by Irish public‹

National Library of Ireland, 13.12.2016

https://www.nli.ie/en/list/latest-news.aspx?article=ecebe796-efdf-4086-946d-6c7d73cc9a21

7. Anforderung an gelingende Webarchivierung: Öffentlichkeit sensibilisieren und einbeziehen

Anforderung an gelingende Webarchivierung:

1. Schnelle und konsequente Umsetzung, sonst drohen weitere Dark Ages
2. Nachfrageseite muss Bedarfe und Präferenzen definieren
3. Trägerform nachhaltig und verlässlich wählen
4. Zugriff möglichst offen und ubiquitär verfügbar
5. Datenmengen und Algorithmizität gewachsen
6. Deep Web, Apps, AI, AR und weitere technologische Trends mitdenken
7. Öffentlichkeit sensibilisieren und einbeziehen

Danke für die Aufmerksamkeit!

Jens Crueger

jens@crueger.info
Digital-Historiker.de