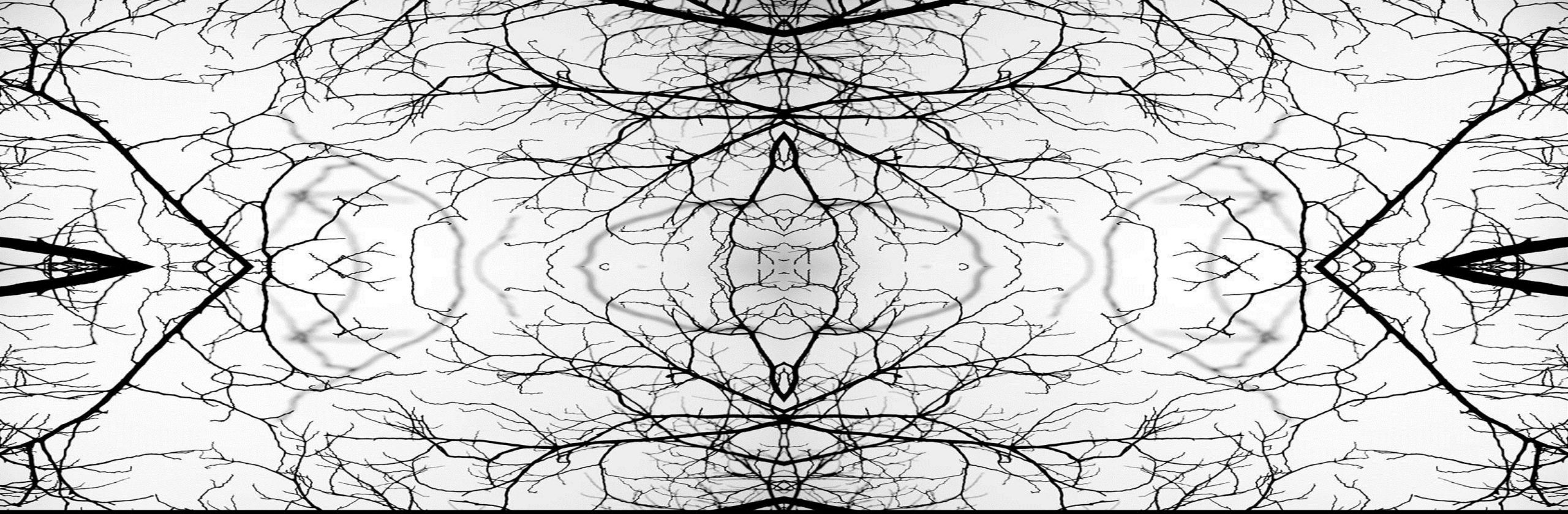

Zur Langzeitarchivierung von Webseiten – ein Lösungsvorschlag

19. AUDS

Wien, 11.03.2015

Alexander Herschung, Alexander.Herschung@startext.de

startext

Die startext

- Seit 1980
- forschungsnaher Softwaredienstleister
- Fokus auf deutschsprachigem Kulturbereich
- Produkte für Museen und Archive
- Projektpartner

Der Schwierigkeiten sind viele…

- Webseiten sind umfangreich und kleinteilig.
  - ➔Was archivieren?
- Webseiten ändern sich oft.
  - ➔ Wann/wie oft archivieren?
- Rechtliche Hürden sind hoch.
  - ➔ gar nicht archivieren?

Was soll bewahrt werden?

- Dargestellte Information?
- Layout und Design?
- Interaktivität?

Was ist eine Webseite? (Dateien)

- Webseiten bestehen aus einer Vielzahl verschiedener Dateien und Dateitypen.
- Viele davon sind kaum als langzeitarchivfähig anzusehen.
- Diese Dateien verlangen eine bestimmte Anordnung und Struktur.

Was ist eine Webseite? (Daten)

- Dargestellter Content liegt in der Regel in einer relationalen Datenbank.
- Viele interaktive Elemente, wie z.B. Suchfunktionen, funktionieren nur bei Vorhandensein dieser Datenbank.

Was ist eine Webseite? (Software)

- Die nutzbare Webseite entsteht erst durch die Interpretation dieser Daten durch einen Browser.
- Browser gibt es mehrere, Browser ändern sich schnell.
- Zugrunde liegende Betriebssysteme verändern sich ebenfalls.
- Ebenso die Hardware.

Was ist eine Webseite? (Nutzererlebnis)

- Eine Webseite ist die Interpretation des zugrundeliegenden Datenkonglomerats durch (Browser-)Software.
- ➔ (vollständige) Webseitenarchivierung = Softwarearchivierung.

Softwarearchivierung

- Archivierung von Software ist mindestens extrem schwierig.
- Ob der Ansatz der Emulation von Hardware, Betriebssystemen und dauerhaftem Weiterbetrieb erforderlicher Softwarekomponenten (z.B. Datenbank) dauerhaft hält, ist zumindest unsicher.

PABLO vereinfacht das Format radikal

- PABLO durchläuft (crawlt) eine gesamte Webseite und erzeugt für jede Einzelseite genau zwei Dateien:
  - Eine Bilddatei, die abbildet, wie die Seite sich im Browser darstellt.
  - Eine METS-XML-Datei, die die Position und das Ziel von Links speichert.
- Eine zusätzliche, weitere METS-XMl-Datei beschreibt die Struktur der Gesamt-Webseite.

PABLO vereinfacht das Format radikal

- Im Ergebnis gibt es nur noch zwei Dateitypen (Bild + METS-XML) mit klar definierten Inhalten.
- So einfach, dass es – auch über Technologiewechsel hinweg – dauerhaft bewahrt werden kann.
- So vollständig, dass daraus eine navigierbare Reproduktion der Website wiederhergestellt werden kann.

PABLO erzeugt eine Präsentationsform der archivierten Webseite

- Aus dem Archivformat (Bild + METS-XML) erzeugt PABLO eine Präsentationsform, die die „Haptik“ der Website reproduziert.
- Diese Präsentationsform wird sich mit der Zeit ändern (müssen), das Archivformat aber nicht.

Eigenschaften, Möglichkeiten und Beschränkungen

- Wählbar:
  - Bildformat.
  - Crawltiefe.
  - Browser (Firefox, IExplorer).
  - Bildschirmauflösung.
- Für Windows (32-/64-bit), MAC, Linux.
- Betriebsbereit ohne Installation.

Eigenschaften, Möglichkeiten und Beschränkungen

- Umsetzung der URL-Bestandteile in eine Ordnerstruktur.
- Funktioniert auch via https.
- Einbettung in Kontext: Archivierung verlinkter (externer) Seiten.
- Sicherung der Texte im METS-XML – Basis für Volltextrecherche.
- Zeitverzögerte Bilderzeugung.

Eigenschaften, Möglichkeiten und Beschränkungen

- Keine Interaktivität.
- Keine animierten Elemente.
- Keine Youtube-Videos o.ä..

- PABLO „fotografiert“ die Website.

Ausbaustufen

- Unterstützung von Login-geschützten Bereichen (in Arbeit).
- Whitelist zur Sicherung direkt verlinkter Dateien (z.B. pdf).
- Umwandlung und Sicherung von Videoinhalten.
- Scripting: wiederholte Snapshots einer Seite.
- Scripting: Suchanfragen und Ergebnisse.
- Interpretation und Verfolgung Java-Script-basierter „Links“.