

Zur Langzeitarchivierung von Webseiten – ein Lösungsvorschlag

32. Österreichischer Bibliothekartag

Wien, 17.09.2015

Alexander Herschung, startext GmbH



Die startext

- Seit 1980 forschungsnaher Softwaredienstleister.
- Fokus auf deutschsprachigem Kulturbereich.
- Produkte und Projekte für kulturgutbewahrende Institutionen.
- Aktuell:
 - OAIS-Repository zur digitalen Langzeitarchivierung: handhabbar, praxisorientiert und bezahlbar.
 - PABLO: Langzeitarchivierung von Websites.

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.

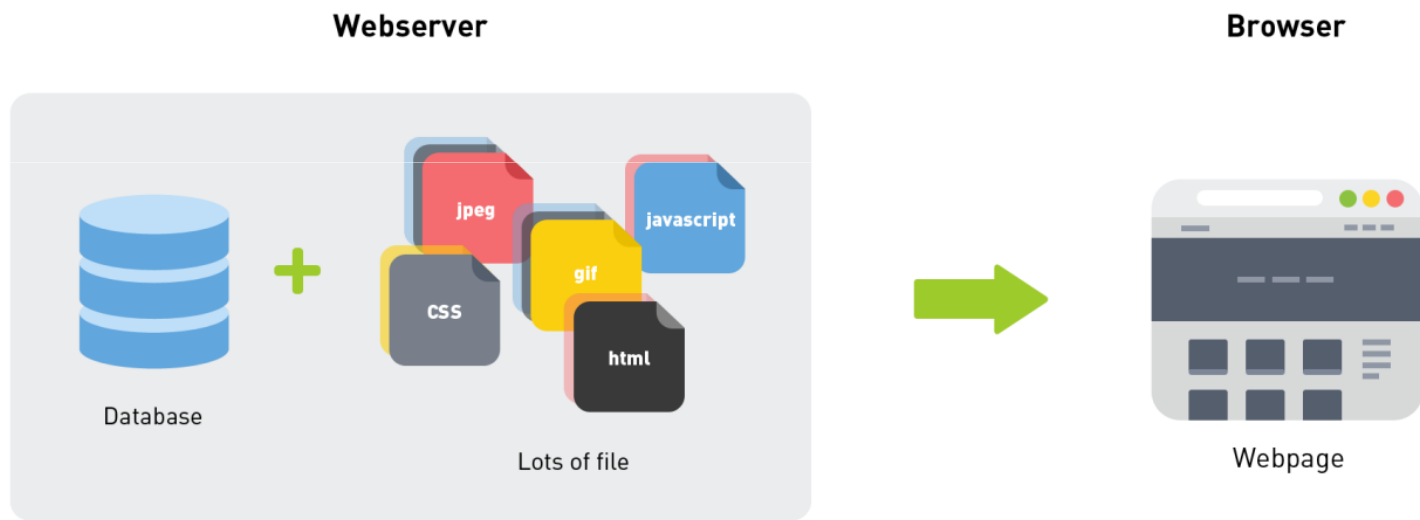
Was ist eine Webseite? (Daten)

- 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 Website?

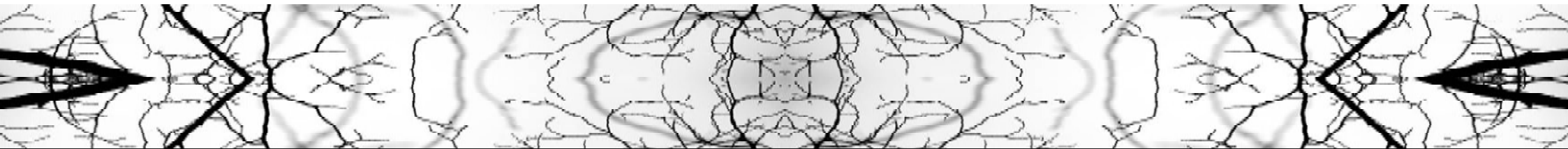


Was ist eine Webseite? (Nutzererlebnis)

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

Softwarearchivierung

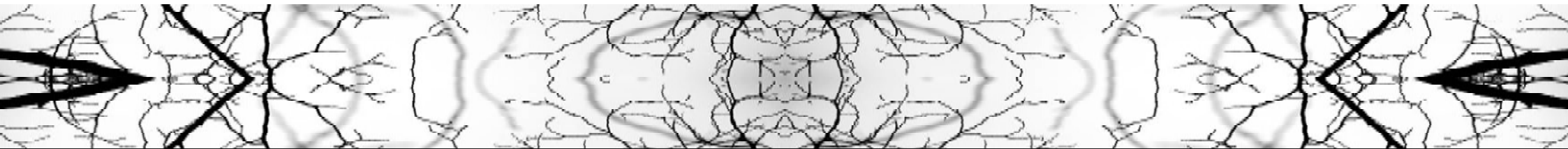
- Archivierung von Software ist mindestens extrem schwierig.
- Ob der Ansatz der Emulation von Hardware, Betriebssystemen und dauerhaftem Weiterbetrieb erforderlicher Softwarekomponenten (z.B. Datenbank) dauerhaft hält, ist zumindest unsicher.



Was soll bewahrt werden?

Was sind signifikante Eigenschaften?

- Interaktivität?
- Dargestellte Information?
- Visuelle Darstellung?
- Surfbarkeit?



Was soll bewahrt werden?

Welche signifikanten Eigenschaften kann PABLO bewahren?

- ~~Interaktivität?~~
- Dargestellte Information!
- Visuelle Darstellung!
- Surfbarkeit!

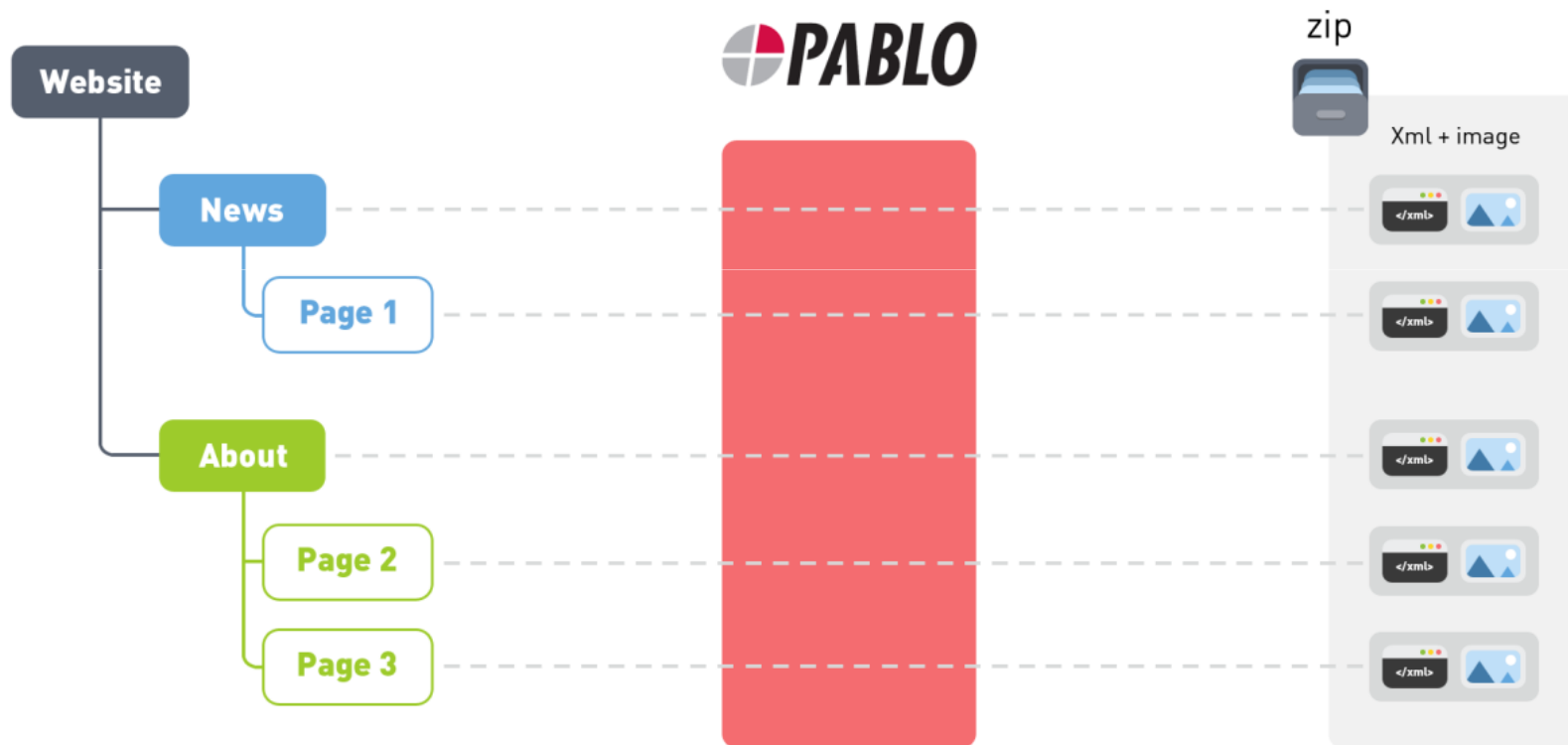
Was macht PABLO?

- 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.

Langzeitarchivierung von Webseiten – PABLO



Langzeitarchivierung von Webseiten – PABLO



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.

Langzeitarchivierung von Webseiten – PABLO

The screenshot shows the PABLO web archiving software interface. The window title is "PABLO" and the main heading is "PABLO Langzeitarchivierung von Webseiten". The interface includes several input fields and controls:

- URL:** An empty text input field.
- Archive Directory:** A text input field containing "C:_Installers\PABLO-1.4 64-bit" and a "Choose" button to the right.
- Include File Types:** A text input field containing "*.pdf".
- Exclude URLs:** A text input field with the instruction "Enter all URLs which should not be parsed. Please enter each URL in a separate line."
- Image format:** A dropdown menu set to "tiff".
- Max depth:** A dropdown menu set to "-1".
- Generate presentation:** A checked checkbox.
- Page Timeout (in min):** A dropdown menu set to "5".
- Resolution:** A dropdown menu set to "Default".
- Include external URLs (directly linked only):** An unchecked checkbox.

At the bottom left of the interface are two buttons: "Start" and "Interrupt". Below these buttons is a log area displaying the message: "13:18:18 INFO : PABLO Version 1.4 -- startext GmbH". At the bottom center of the window, the text "startext GmbH © 2015" is visible.

Eigenschaften, Möglichkeiten und Beschränkungen

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

Eigenschaften, Möglichkeiten und Beschränkungen.

- Sicherung der dargestellten Texte – Basis für Volltextrecherche.
- Whitelist zur Sicherung direkt verlinkter Dateien (z.B. pdf).
- Einbettung in Kontext: Archivierung verlinkter (externer) Seiten.
- Ausschluss von Teilen der Website.
- Funktioniert auch via https.

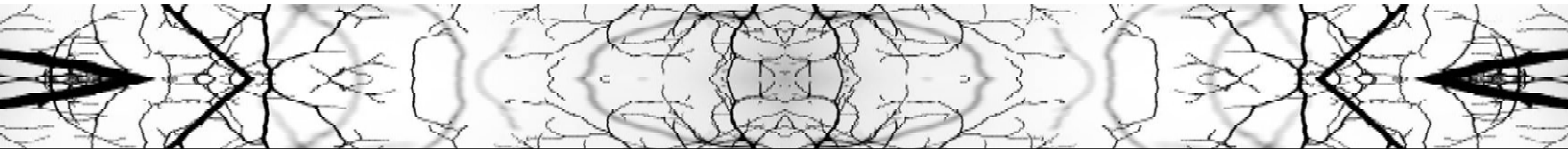
Eigenschaften, Möglichkeiten und Beschränkungen

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

- PABLO „fotografiert“ die Website.

Ausbaustufen

- Unterstützung von Login-geschützten Bereichen (in Arbeit).
- Umwandlung und Sicherung von Videoinhalten.
- Scripting: wiederholte Snapshots einer Seite (für Slider o.ä.).
- Scripting: Suchanfragen und Ergebnisse.
- Interpretation und Verfolgung Java-Script-basierter „Links“.



Kontakt

Alexander.Herschung@startext.de

www.startext.de