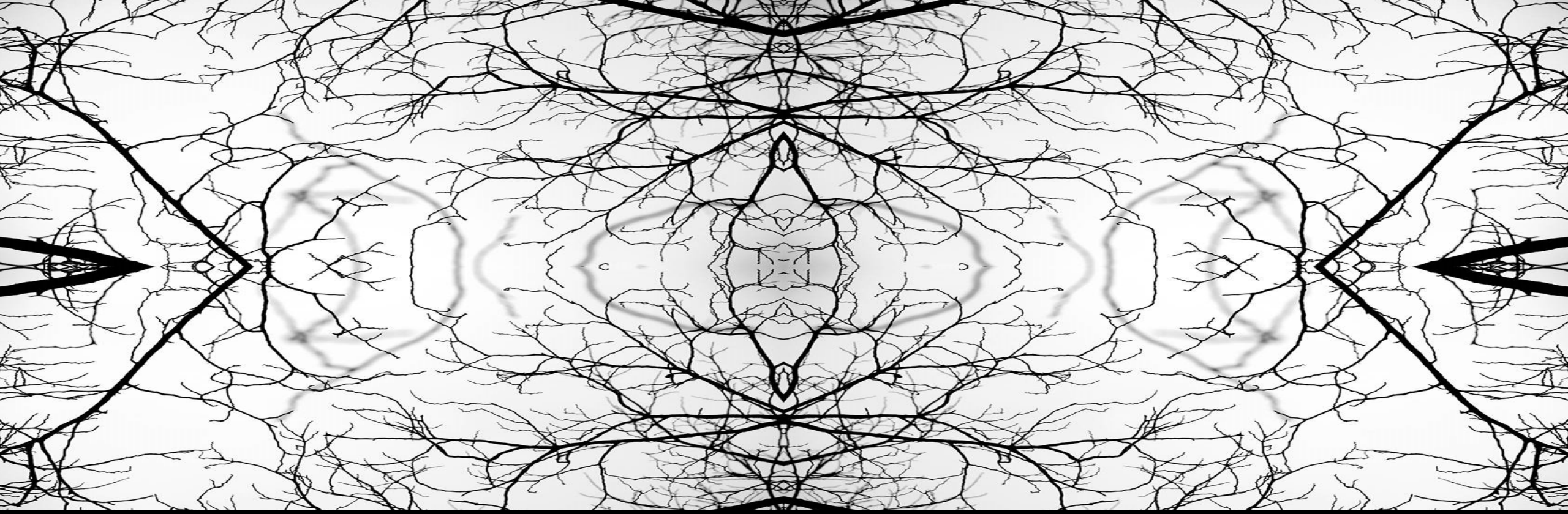
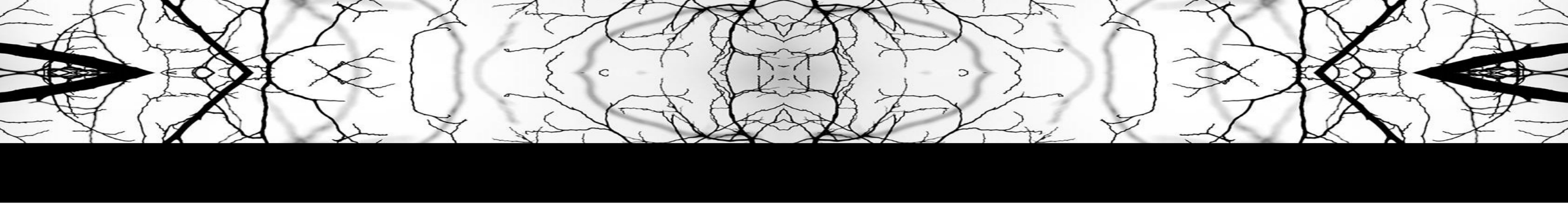


Die Archivierung von Websites. Notwendigkeit und technische Lösung

27.10.2023

Michael Steppes
startext GmbH

startext



Agenda

- Kurzvorstellung startext GmbH und Thema
- Warum Webseitenarchivierung?
- Schwierigkeiten der Webseitenarchivierung
- startext PABLO: technischer Ansatz und Funktionsweise
- startext PABLO: Vertriebsmodell
- Fragen und Antworten

Die startext - wer wir sind

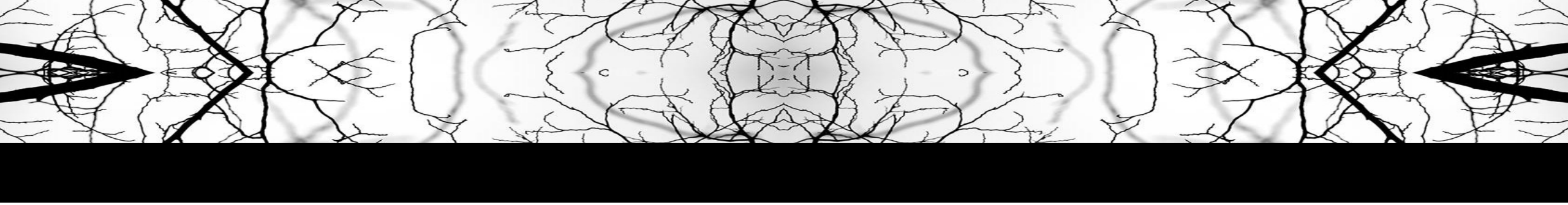
- Inhabergeführtes, unabhängiges Softwareunternehmen
- gegründet 1980
- Fokus: deutschsprachiger Kulturbereich
- HiDA, MidosaXML, VERA, ACTApro
- Firmensitz in Bonn, Niederlassung in Leipzig, Mitarbeiter:innen in Nürnberg und Bremen
- ~30 festangestellte Mitarbeiter:innen

startext

Warum Webseitenarchivierung?

- Homepage der Kommunalverwaltung
- Intranet politischer Parteien
- Webseiten des Lehrkörpers einer Universität
- Firmenhomepage
- …

- → Teil der Überlieferung und damit archivwürdig wie andere digitale Daten auch



Schwierigkeiten der Webseitenarchivierung

- umfangreich
- ändern sich oft
- rechtliche Hürden (zeitlich befristete Nutzungsrechte?)
- Was ist primär:
  - die dargestellte Information?
  - das Layout bzw. Design?
  - die Interaktivität?

Ausgangslage der Webseitenarchivierung

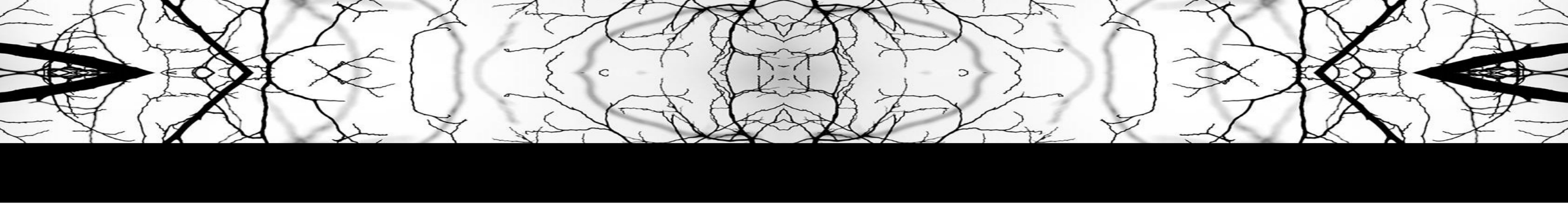
Was ist eine Webseite? Dateien!

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

Ausgangslage der Webseitenarchivierung

Was ist eine Webseite? Daten!

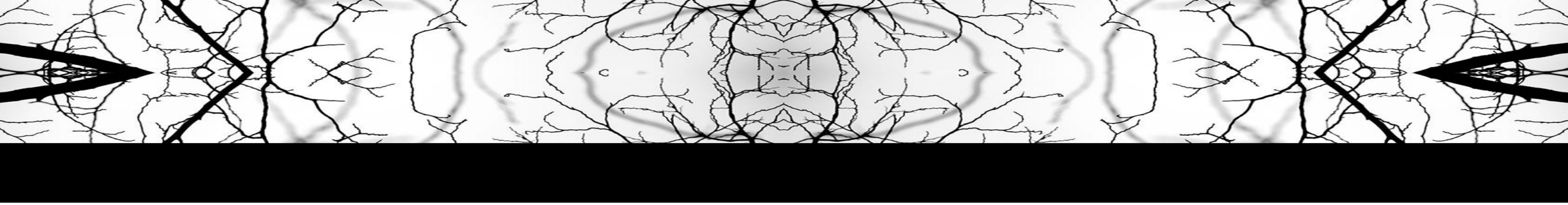
- Dargestellter Content liegt in der Regel in einer relationalen Datenbank.
- Viele interaktive Elemente, wie z.B. Suchfunktionen, funktionieren nur bei Vorhandensein dieser Datenbank.

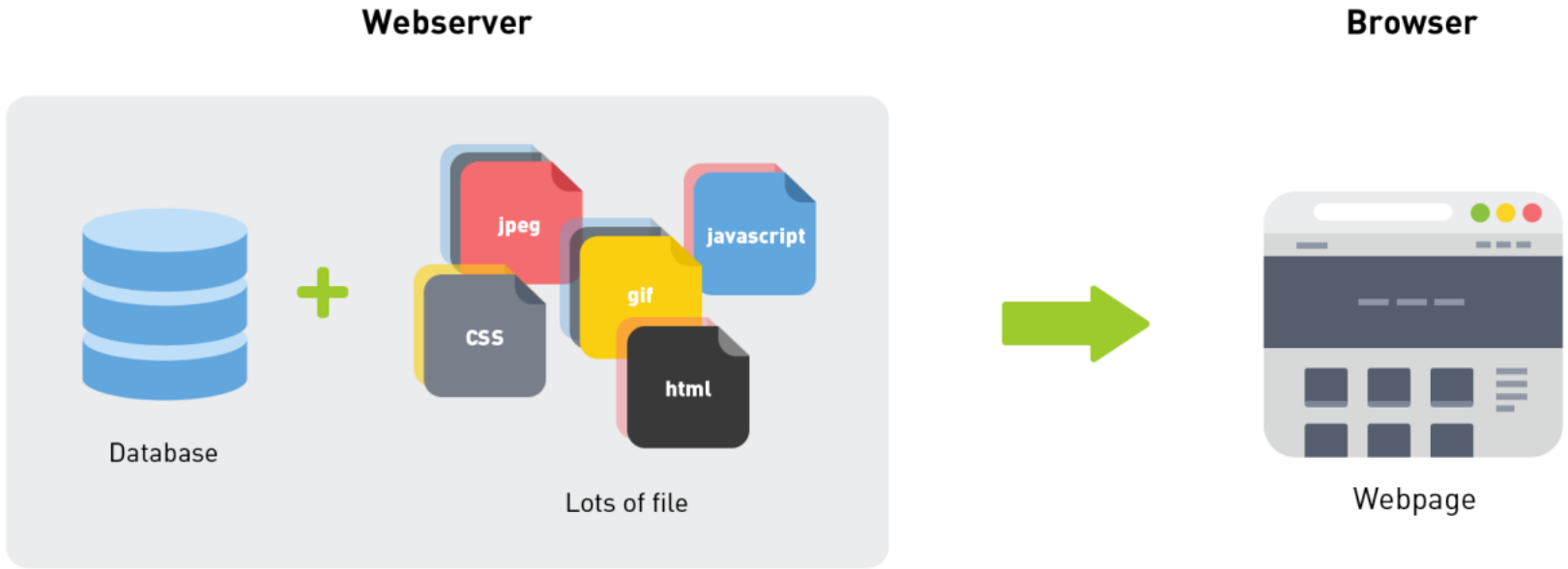


Ausgangslage der Webseitenarchivierung

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.
- Zugrundeliegende Betriebssysteme verändern sich ebenfalls.
- Und auch die Hardware.




Webserver
Browser
jpeg
javascript
gif
css
html
Database
Lots of file
Webpage

startext

Ausgangslage der Webseitenarchivierung

Was ist eine Webseite? Nutzererlebnis!

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

Ausgangslage der Webseitenarchivierung

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.

Ausgangslage der Webseitenarchivierung

Was sind signifikante Eigenschaften?

- dargestellte Information?
- visuelle Darstellung?
- Surfbarkeit?
- Interaktivität?

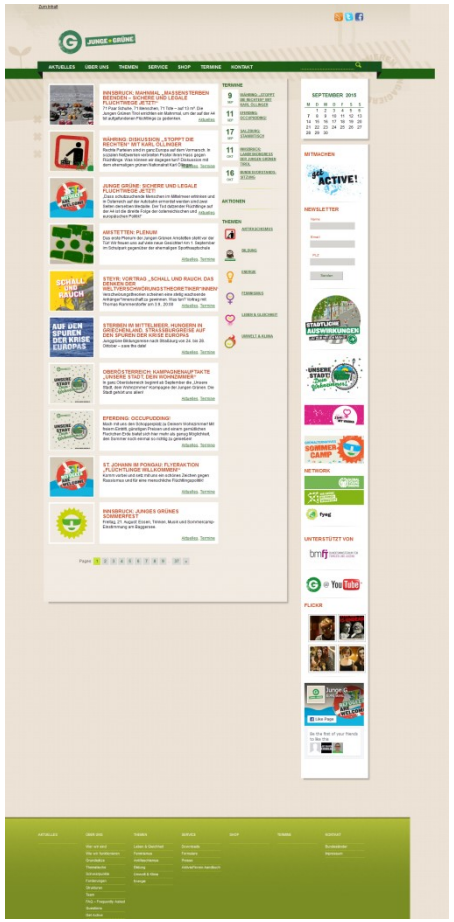
startext PABLO: technischer Ansatz und Funktionsweise

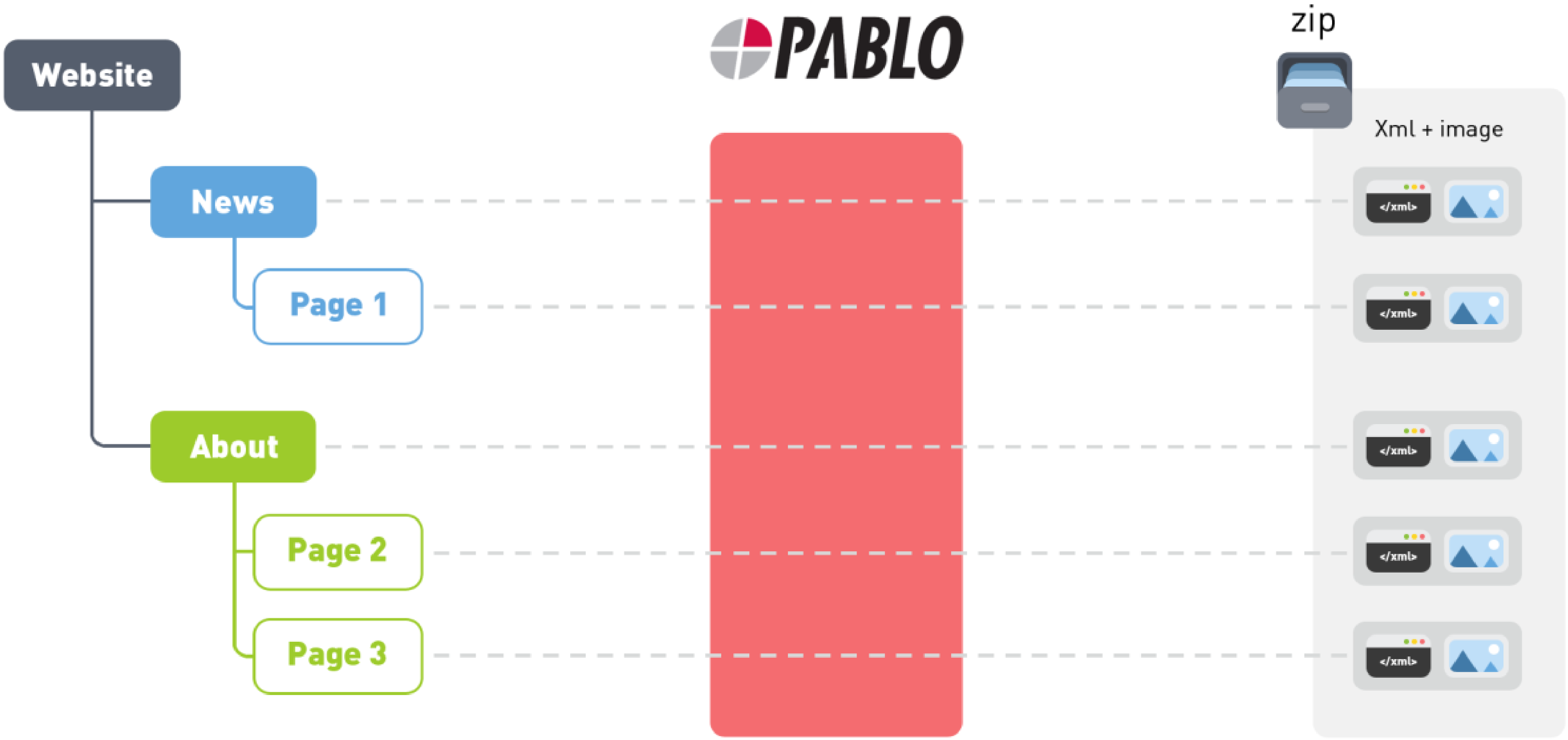
Was macht PABLO?

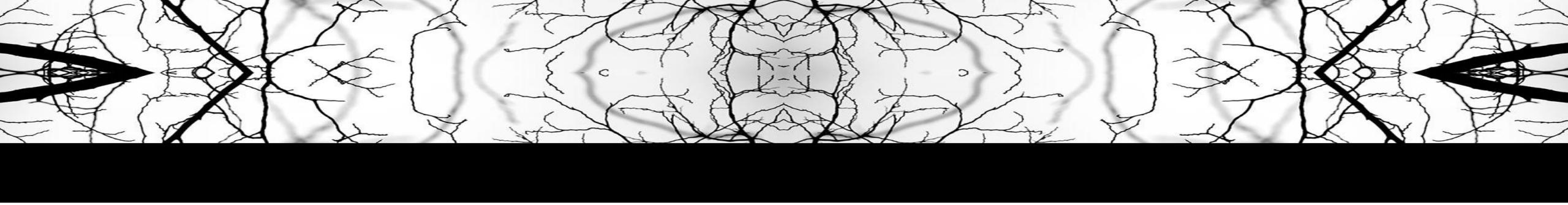
- PABLO vereinfacht das Format radikal.
- PABLO durchläuft (crawlt) eine gesamte Webseite und erzeugt für jede Einzelseite genau zwei Dateien:
  - Eine Bilddatei, die abbildet, wie die Seite sich im Browser darstellt.
  - Eine METS-XML-Datei, die die Position und das Ziel von Links speichert.


Webserver
PABLO
jpeg
javascript
gif
CSS
html
Database
Lots of file
Webpage
</xml>
Xml
Image file

```
    <wa:topOffset unit='pixel'>174</wa:topOffset>
</wa:link>
<wa:link id='LINK_0000000006'>
    <wa:href>http://tirol.junge-gruene.at/ueber-uns/</wa:href>
    <wa:mets>../../../../ueber-uns/index.html.mets.xml</wa:mets>
    <wa:text>ÜBER UNS</wa:text>
    <wa:width unit='pixel'>102</wa:width>
    <wa:height unit='pixel'>30</wa:height>
    <wa:leftOffset unit='pixel'>242</wa:leftOffset>
    <wa:topOffset unit='pixel'>174</wa:topOffset>
</wa:link>
<wa:link id='LINK_0000000007'>
    <wa:href>http://tirol.junge-gruene.at/ueber-uns/get-active/</w
```

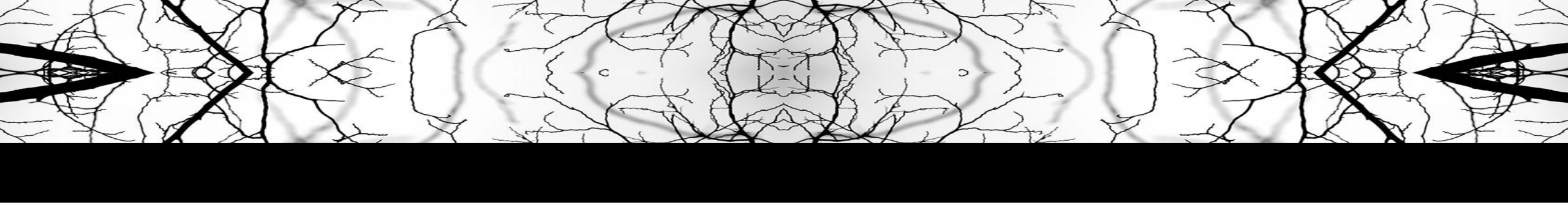



Website
News
Page 1
About
Page 2
Page 3
PABLO
zip
Xml + image
</xml>



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
  - Bildschirmauflösung
- für Windows (32-/64-bit), MacOS, Linux
- betriebsbereit ohne Installation

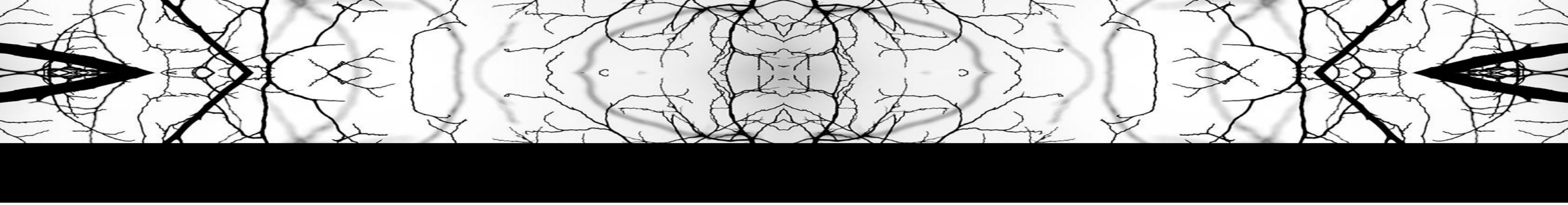
Eigenschaften, Möglichkeiten und Beschränkungen

- 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

Eigenschaften, Möglichkeiten und Beschränkungen

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

- **PABLO „fotografiert“ die Website!**

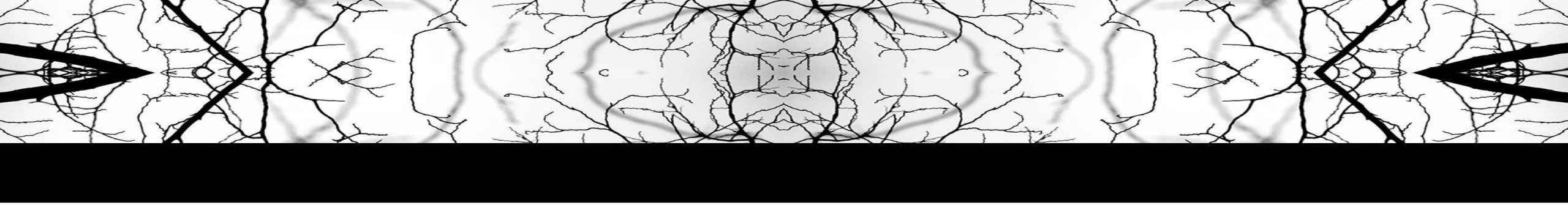


Ausgangslage der Webseitenarchivierung

Welche signifikanten Eigenschaften kann PABLO bewahren?

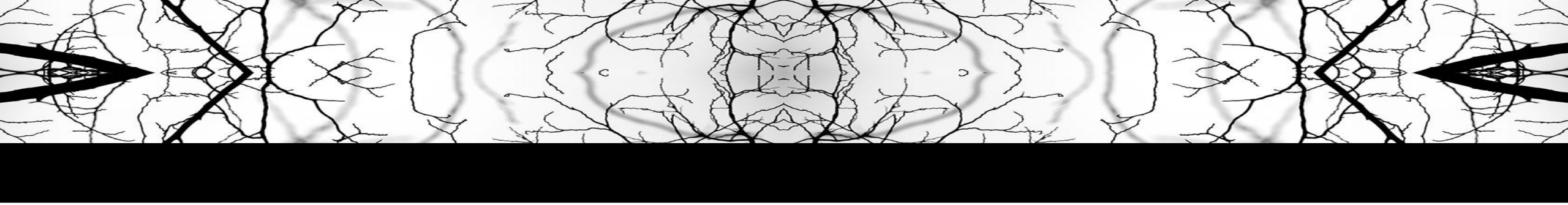
- dargestellte Information!
- visuelle Darstellung!
- Surfbarkeit!
- ~~Interaktivität?~~

startext



Eine PABLO-Archivierung am Beispiel www.startext.de

startext

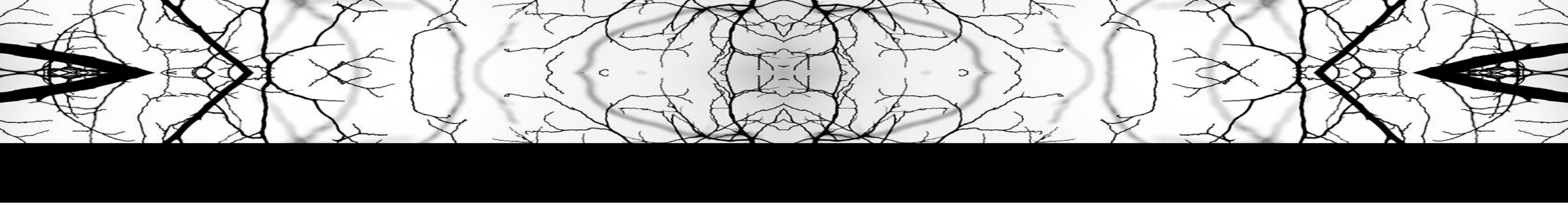


startext PABLO: Vertriebsmodell

- Dienstleistung oder eigene Umgebung
- eine als Virtuelle Maschine (VM) fertig konfigurierte Anwendung: Betriebssystem und Programm zusammen
- Oracle VirtualBox (open source)
- auf Ihre Website zugeschnitten und exakt angepasst
- einmalig: Lizenzpreis plus Einrichtungspauschale
- jährliche Wartungspauschale
- jede weitere Website: erneute reduzierte Einrichtungspauschale
- Gewährleistung bis Tiefe 3

Tips

- Vortrag „Webarchivierung im Spannungsfeld von Nutzung und Langzeitarchivierung. Ein Praxisbericht“: Michael Volpert M.A. M.A., Archiv und Bibliothek des Erzbistums München und Freising (EDV-Tage Theuern, www.edvtage.de – bald als Video verfügbar)
- Webseitenarchivierung im Test: Michael Cöln, Johannes Ehrengruber, Andreas Jüngling, Michael Korn, Jens Löffler, Gregor Patt, Dietmar Pertz, Tobias Schröter, Johannes Thomé (in: „Archivpflege in Westfalen-Lippe“ Heft 97 [2022])



Fragen?

Michael.Steppes@startext.de
www.startext.de

startext