

Archivierung von Webseiten und Social Media

Probleme und Lösungsmöglichkeiten

München, 08.04.2024

Johannes Thomé

Programm

I. Begriffsbestimmung

II. Warum Webarchivierung?

III. Spezifika von Webseiten als Archivgut

IV. Praktische Umsetzung

V. Exkurs: Social Media

VI. Praxisbeispiele

VII. Fazit

I. Begriffsbestimmung

- Überlieferung und dauerhafte Aufbewahrung von Internetpublikationen
- Digitale Archivierung von genuin digitalen Inhalten
- Ziel: möglichst langfristiger Erhalt von Inhalt, Form und Funktion
- Unterschied zwischen Webseiten und Social Media
 - Serverzugriff durch eigene Institution
 - Dynamik

II. Warum Webseitenarchivierung?

- Webseiten sind Teil des historischen Erbes
- Wesentliches Medium für Öffentlichkeitsarbeit
- Prägen Bild der eigenen Institution in der Öffentlichkeit
- Stark aggregierte Informationen
- Auch für digitale Unterlagen gilt das Provenienzprinzip

III. Spezifika von Webseiten als Archivgut

- Vielfalt der Formate
- Volatile äußere Form und Darstellung
- Schnelllebigkeit
- Veränderbarkeit
- Dynamische Inhalte unterschiedlichster Herkunft
- Interdependenz von Webseiten

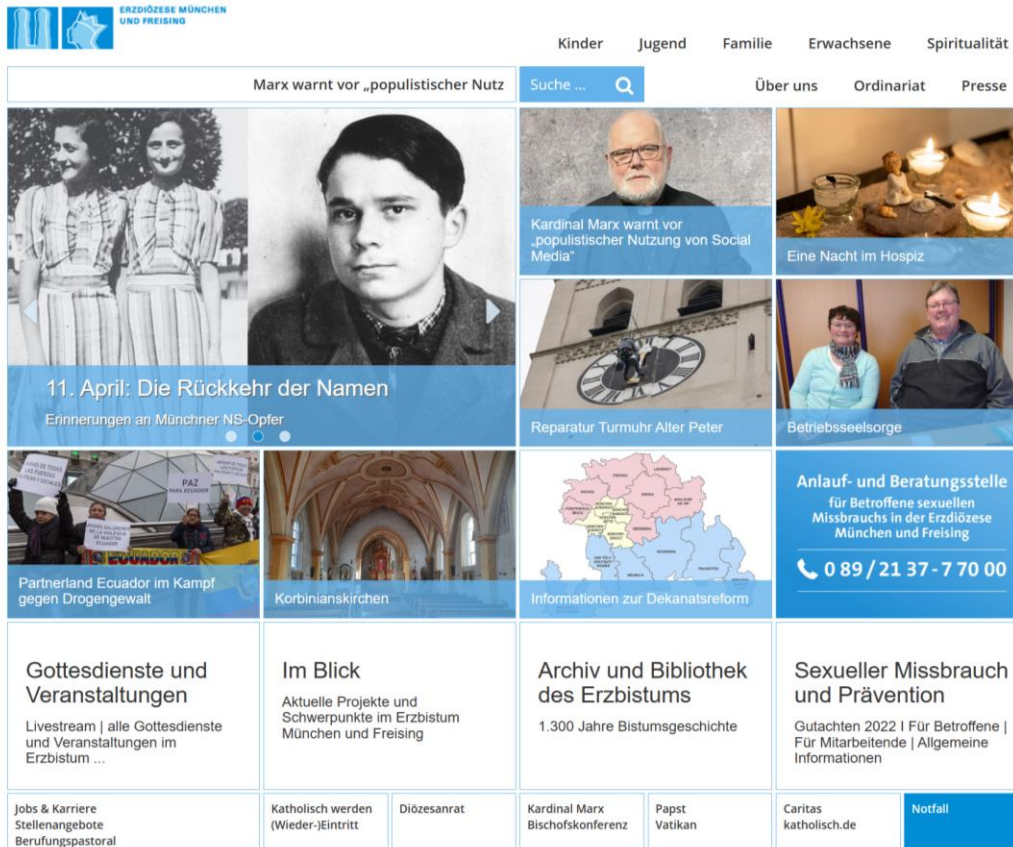
Externe Inhalte, YouTube

The image shows a browser window displaying a website for a virtual tour. The main heading is "VIRTUELLE KULTURTOUR" in large white letters on a dark background. Below it, the text reads: "Entdecke das Stift Wilten mit exklusiven Virtual-Reality-, 3D- und Live-Touren." The background image is a smartphone screen showing a virtual tour interface with a map and various icons. Below the main heading, there is a section titled "Virtuelle Führung" and "Stiftskirche des Prämonstratenser Chörherrenstiftes Wilten".

Below the main content, there is a YouTube video player. The video title is "Next Level | Stift Wilten". The video content shows the text "NEXT LEVEL" in a stylized font with a red play button icon in the center. The video player includes a "Teilen" (Share) button in the top right corner and a "Kann ich dir helfen? Klicke hier und schreibe mir!" (Can I help you? Click here and write to me!) button in the bottom right corner.

<https://stift-wilten.at/shop/virtuelle-kulturtour/>

Interdependenz von Webseiten: Header



ERZBISCHOFSTUHL MÜNCHEN UND FREISING

Kinder Jugend Familie Erwachsene Spiritualität

Marx warnt vor „populistischer Nutz“ Suche ... Über uns Ordinariat Presse

11. April: Die Rückkehr der Namen
Erinnerungen an Münchner NS-Opfer

Kardinal Marx warnt vor „populistischer Nutzung von Social Media“

Eine Nacht im Hospiz

Reparatur Turmuhr Alter Peter

Betriebsseelsorge

Partnerland Ecuador im Kampf gegen Drogengewalt

Korbinianskirchen

Informationen zur Dekanatsreform

Anlauf- und Beratungsstelle für Betroffene sexuellen Missbrauchs in der Erzdiözese München und Freising
0 89 / 21 37 - 7 70 00

Gottesdienste und Veranstaltungen
Livestream | alle Gottesdienste und Veranstaltungen im Erzbistum ...

Im Blick
Aktuelle Projekte und Schwerpunkte im Erzbistum München und Freising

Archiv und Bibliothek des Erzbistums
1.300 Jahre Bistumsgeschichte

Sexueller Missbrauch und Prävention
Gutachten 2022 | Für Betroffene | Für Mitarbeitende | Allgemeine Informationen

Jobs & Karriere
Stellenangebote
Berufungspastoral

Katholisch werden (Wieder-)Eintritt

Diözesanrat

Kardinal Marx
Bischofs-konferenz

Papst
Vatikan

Caritas
katholisch.de

Notfall

© Erzbistum München und Freising 2024

Hinweise für Webseiten-Betreiber zum Kirchlichen Datenschutz | Impressum | Datenschutzerklärung | Kontakt | CMS | Seite drucken | Bildnachweis | Cookie-Einstellungen



<https://www.erzbistum-muenchen.de/>



Deutsche Provinz der
Salesianer Don Boscos
Damit das Leben junger Menschen gelingt.



<https://www.donbosco.de/>

- [Datenschutzerklärung](#)
- [Impressum](#)
- [Sitemap](#)
- [Stellenangebote](#)
- [Interner Bereich](#)
- [Interner Bereich für Mitbrüder](#)



DON BOSCO

Deutsche Provinz der
Salesianer Don Boscos

Damit das Leben junger Menschen gelingt.

- [Home](#)
- [Über uns](#)
 - [Don Bosco](#)
 - [Salesianer Don Boscos](#)
 - [Don Bosco Familie](#)
 - [Salesianer werden](#)
 - [Unsere Arbeit in Deutschland und weltweit](#)
 - [Provinzialat München](#)
 - [Unser Profil](#)
 - [Prävention und Missbrauch](#)
 - [Karriere bei Don Bosco](#)

IV. Praktische Umsetzung

Vor dem Einstieg zu klären:

- Ist Webarchivierung sinnvoll und umsetzbar?
- Überlieferungsziel?
 - Technische vs. Organisatorische Probleme
- Welche und wie viele Seiten sollen archiviert werden?
 - Nur öffentliche Seiten, oder auch Intranet?
- Wie oft soll gespeichert werden?
- Was darf es kosten?

IV. Praktische Umsetzung

Möglichkeiten:

- a) Daten vom Server regelmäßig selbst abspeichern
- b) Selber ein Programm aufsetzen und betreiben
- c) Externe Firma beauftragen

=> b) und c) setzen auf Spezialsoftware und lassen sich (teilweise)
automatisieren

IV. Praktische Umsetzung

Wie funktionieren Webarchivierungstools?

- Crawling und Harvesting
 - „Einsammeln“ von Dokumenten
 - Flächig oder selektiv
- Speichern mitsamt Metadaten
 - WARC-Format
- Zugang über spezielle Oberflächen
- Langzeiterhaltung bleibt als Aufgabe des Archivs!

IV. Praktische Umsetzung

Programme und Anbieter für Webarchivierung

- **Heritrix**
- Web Curator Tool
- Mirror Web
- PABLO
- Conifer

Heritrix

The screenshot shows the Heritrix Engine web interface in a Mozilla Firefox browser window. The browser title is "Heritrix Engine 3.4.0-SNAPSHOT-2020-08-03T23:43:15Z - Mozilla Firefox" and the address bar shows "https://localhost:8000/engine". The page header displays "HERITRIX 3.4.0-SNAPSHOT-2020-08-03T23:43:15Z".

The main content area is titled "Engine" and includes a "Memory" section with a progress bar showing 136 MiB used, 504 MiB current heap, and 1024 MiB max heap. A "run garbage collector" button is present. To the right, an "Exit Java" dialog box is displayed with an "I'm sure" checkbox and an "exit java process" button.

The "Job Directories" section shows a "rescan" button and a "parent directory /home/heritrix/heritrix/jobs" with "1 known job directories". Below this, there are two forms: "create new job" with a "myJob" input field and a "create" button, and "add existing job" with a "/path/to/job" input field and an "add" button.

A job entry "myJob" is listed with the path "/home/heritrix/heritrix/jobs/myJob/crawler-beans.xml" and "0 launches".

https://commons.wikimedia.org/wiki/File:Heritrix_3.4.0_Web_UI.png

IV. Praktische Umsetzung

Programme und Anbieter für Webarchivierung

- Heritrix
- **Web Curator Tool**
- Mirror Web
- PABLO
- Conifer

Web Curator Tool

The screenshot displays the Web Curator Tool dashboard. At the top left is a logo with a globe and the text 'WEB CURATOR TOOL'. At the top right are navigation links: 'Home | Queue | Harvested | Help | Logout' and a status message 'User MPennock is logged in.' The dashboard is organized into two columns of widget boxes. The left column contains: 'In Tray' (2249 tasks, 0 Notifications) with an 'open' button; 'Harvest Authorisations' (6745 harvest authorisations) with 'open' and 'add new' buttons; 'Targets' (0 Targets) with an 'open' button; 'Target Instances' (0 Scheduled instances, 0 ready for Quality reviews) with 'open', 'queue', and 'harvested' buttons; and 'Groups' (0 Target Groups) with an 'open' button. The right column contains: 'Permission Request Templates' with 'open' and 'add new' buttons; 'Reports' with an 'open' button; 'Harvester Configuration' with 'general', 'bandwidth', and 'profile' buttons; and 'Users, Roles, Agencies & Rejection Reasons' with sub-sections for 'Users', 'Roles', 'Agencies', and 'Rejection Reasons', each having 'open' and 'add new' buttons.

https://netpreserveblog.files.wordpress.com/2018/04/web_curator_tool-02.jpg

IV. Praktische Umsetzung

Programme und Anbieter für Webarchivierung

- Heritrix
- Web Curator Tool
- **Mirror Web**
- PABLO
- Conifer

IV. Praktische Umsetzung

Programme und Anbieter für Webarchivierung

- Heritrix
- Web Curator Tool
- Mirror Web
- **PABLO**
- Conifer

Pablo

PABLO

Pablo Digital long-term archiving of websites

URL:

Login: username: password:

Archive Directory:

Include File Types:

Exclude URLs:

Image format: Max depth: Generate presentation: Include external URLs:

Resolution: Page Timeout (in min): Include external Files: On screen logging:

Full Archive: Scale factor (in %):

Delay on root page(sec): Delay on visit page(sec):

04:01:46 INFO : PABLO Version 1.8 -- startext GmbH

Screenshot: Michael Korn, StA St. Augustin / Tobias Schröter, InterKA Lohmar

Pablo

```
</FContent>
</file>
</fileGrp>
<fileGrp ID='FILEGRP_2' USE='page_snapshot'>
  <file ID='FID_00000003' MIMETYPE='image/jpeg' USE='page_snapshot'>
    <FLocat LOCTYPE='URL' xlink:href='index.html.jpeg' />
    <FContent>
      <xmlData>
        <wa:metadata xmlns:wa='http://www.starttext.de/WebArchiver'>
          <wa:link id='LINK_0000000000'>
            <wa:href>https://www.speed-pizza.com/</wa:href>
            <wa:text></wa:text>
            <wa:width unit='pixel'>450</wa:width>
            <wa:height unit='pixel'>92</wa:height>
            <wa:leftOffset unit='pixel'>281</wa:leftOffset>
            <wa:topOffset unit='pixel'>4</wa:topOffset>
          </wa:link>
          <wa:link id='LINK_0000000001'>
            <wa:href>http://www.pizza-onlineshop.com/</wa:href>
            <wa:text>Pizza Shopsystem</wa:text>
            <wa:width unit='pixel'>90</wa:width>
            <wa:height unit='pixel'>15</wa:height>
            <wa:leftOffset unit='pixel'>1207</wa:leftOffset>
            <wa:topOffset unit='pixel'>1</wa:topOffset>
          </wa:link>
          <wa:link id='LINK_0000000002'>
            <wa:href>https://www.speed-pizza.com/</wa:href>
            <wa:text>HOME</wa:text>
            <wa:width unit='pixel'>56</wa:width>
            <wa:height unit='pixel'>21</wa:height>
            <wa:leftOffset unit='pixel'>281</wa:leftOffset>
            <wa:topOffset unit='pixel'>120</wa:topOffset>
          </wa:link>
          <wa:link id='LINK_0000000003'>
            <wa:href>https://www.speed-pizza.com/285/Pizza+bestellen/</wa:href>
            <wa:text>PIZZA</wa:text>
            <wa:width unit='pixel'>53</wa:width>
            <wa:height unit='pixel'>21</wa:height>
            <wa:leftOffset unit='pixel'>341</wa:leftOffset>
            <wa:topOffset unit='pixel'>120</wa:topOffset>
          </wa:link>
        </wa:metadata>
      </xmlData>
    </FContent>
  </file>
</fileGrp>
```

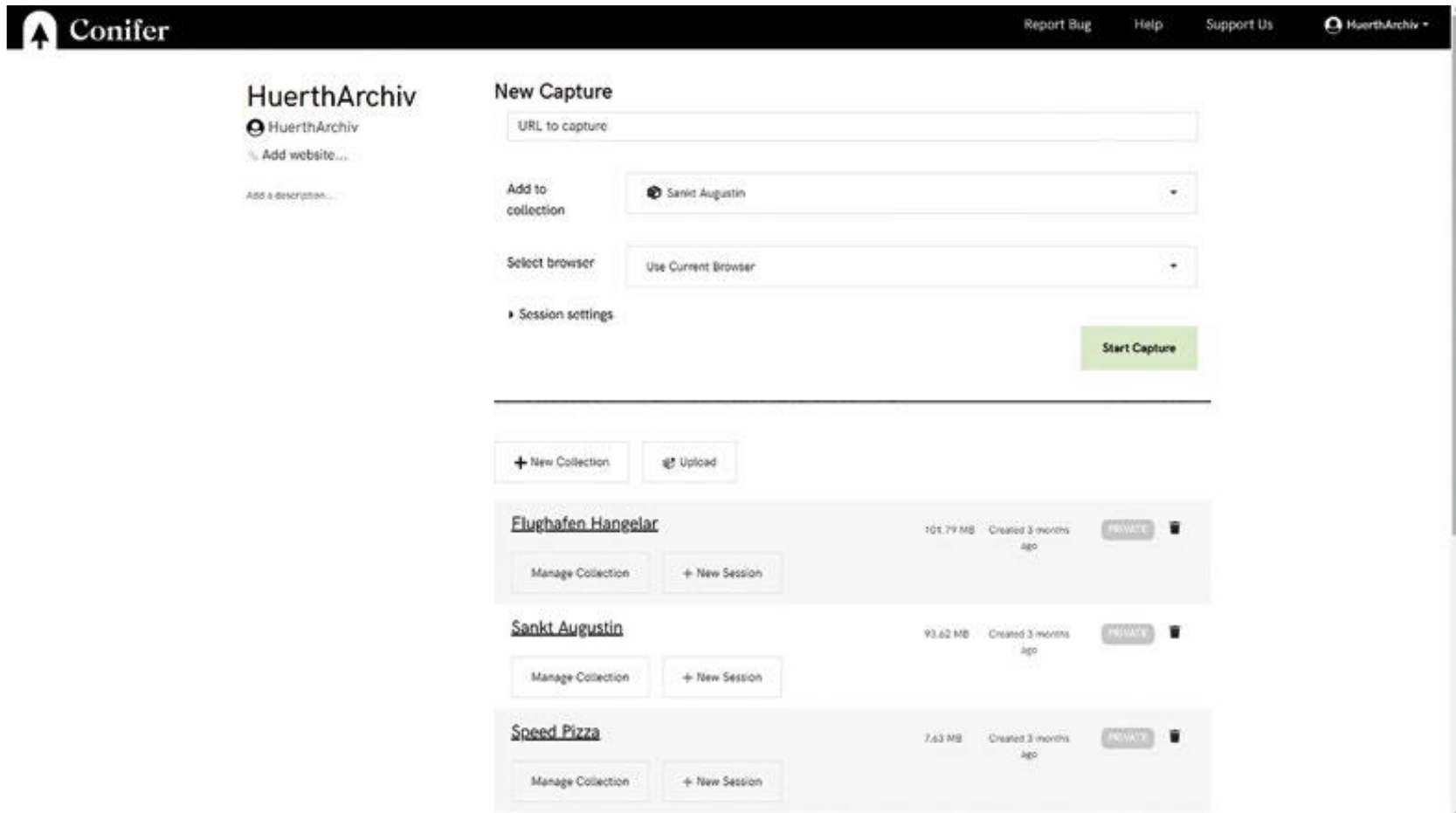
Screenshot: Michael Korn, StA St. Augustin / Tobias Schröter, InterKA Lohmar

IV. Praktische Umsetzung

Programme und Anbieter für Webarchivierung

- Heritrix
- Web Curator Tool
- Mirror Web
- PABLO
- **Conifer**

Conifer



The screenshot shows the Conifer web interface. At the top, there is a navigation bar with the Conifer logo, 'Report Bug', 'Help', 'Support Us', and a user profile for 'HuerthArchiv'. The main content area is divided into two sections. The left section, titled 'HuerthArchiv', contains a profile picture, the name 'HuerthArchiv', and options to 'Add website...' and 'Add a description...'. The right section, titled 'New Capture', contains a form with the following fields: 'URL to capture' (text input), 'Add to collection' (dropdown menu with 'Sankt Augustin' selected), and 'Select browser' (dropdown menu with 'Use Current Browser' selected). Below these fields is a 'Session settings' section and a green 'Start Capture' button. Below the form, there are two buttons: '+ New Collection' and 'Upload'. The bottom section displays a list of three collections:

Collection Name	Size	Created	Privacy	Actions
<u>Flughafen Hanglar</u>	101.79 MB	Created 3 months ago	PRIVATE	Manage Collection, + New Session
<u>Sankt Augustin</u>	93.62 MB	Created 3 months ago	PRIVATE	Manage Collection, + New Session
<u>Speed Pizza</u>	7.63 MB	Created 3 months ago	PRIVATE	Manage Collection, + New Session

Screenshot: Michael Cöln, StA Hürth

V. Exkurs: Social Media

VI. Beispiel DNB-Webseitenarchiv


English Kontakt Glossar Gebärdensprache Leichte Sprache

BAW
Bundesanstalt für Wasserbau

Kompetenz für die Wasserstraßen

Die BAW Infrastruktur Umwelt Mobilität Karriere Service Publikationen

Suche



Mobilität auf den Wasserstraßen fördern

< || >

Neues aus der BAW
13.03.2024
Innovative Bauverfahren entwickeln und erproben

Forschung Xpress
04.04.2024
Entwicklung eines neuen Verfahrens zur Modellierung der Strömungen mi...

Veranstaltungen
22.04.2024
Hydraulik der Wasserbauwerke – Komplexe Aufgaben unter sich...



<https://www.baw.de/de/home/home.html>

DNB-Abbild 2018

Archivierte Netzressource vom 30.08.2018

<https://www.baw.de/%0Ahttps://izw.baw.de>

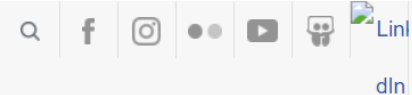
Datensatz im Katalog

ENGLISH KONTAKT PRESSE KARRIERE GLOSSAR  GEBÄRDENSPRACHE  LEICHTE SPRACHE



Kompetenz für die Wasserstraßen

WASSERBAU BAUTECHNIK GEOTECHNIK DIE BAW SERVICE & WISSEN



- THEMEN >
- METHODEN >
- PROJEKTE
- FORSCHUNG & ENTWICKLUNG
- FACHPUBLIKATIONEN

<https://webarchiv.dnb.de/show.aspx?GUID=38c9d77b-20ac-e811-a072-d4ae528b7600&PARENT=46734d33-7bec-470c-80ae-ab3cf20091b4>

DNB-Abbild 2024

Archivierte Netzressource vom 09.01.2024

<https://www.baw.de/%0Ahttps://izw.baw.de>

Datensatz im Katalog

English Kontakt Glossar Gebärdensprache Leichte Sprache



Kompetenz für die Wasserstraßen

Die BAW Infrastruktur Umwelt Mobilität Karriere Service Publikationen

Suche



□ □ □



Karriere

Arbeiten bei der BAW



Qualitätsbewertung

Übersichten von im Verkehrswasserbau
zugelassenen Stoffen, Firmen und
Verfahren



Regelwerke

Merkblätter, Richtlinien, Empfehlungen,
Leitfäden und Arbeitshilfen

<https://webarchiv.dnb.de/show.aspx?GUID=70d867c0-18af-ee11-8f00-0050568c5d1b&PARENT=46734d33-7bec-470c-80ae-ab3cf20091b4>

Verlinkung auf Youtube

Wir empfehlen Google Chrome für die Wiedergabe von YouTube-Videos. [Chrome jetzt installieren.](#)

Schließen

Navigation überspringen

DE



Anmelden

Suchen

Suchen

[BAW - Bundesanstalt für Wasserbau](#)

• **BAW - Bundesanstalt für Wasserbau**

- [Videos](#)
- [Playlists](#)
- [Kanäle](#)
- [Diskussion](#)
- [Kanalinfo](#)



- - [Start](#)
 - [Trends](#)
 - [Verlauf](#)
 - [YouTube Premium abonnieren](#)

VI. Friedrich-Ebert-Stiftung



SOLRWAYBACK

Solidarität|domain:verdi.de

GROUPED SEARCH [?]

IMAGE SEARCH [?] URL SEARCH [?]

SEARCH WITH UPLOADED FILE

GPS IMAGE SEARCH TOOLBOX

FACETS

domain

- verdi.de (1,225,224)

content_type_norm

- html (1,223,712)
- pdf (1,482)
- other (27)
- image (3)

type

- web page (1,223,712)
- document (1,482)
- other (27)
- image (3)

crawl_year

- 2019 (221,996)
- 2020 (850,426)

RESULTS

Showing 0 - 20 of 1,225,224 entries matching query.

[See available export options](#)

Previous 20

Next 20

SOLIDARITÄT - VER.DI

#1
score: 42.269615

Type: html, web page @ verdi.de

Date: 19/07-2021

Url: <https://www.verdi.de/ueber-uns/bundeskongress-2019/solidaritaet/>

Highlighted content:

"Solidarität ist keine WorthülseEuropa ja, aber anders!Die Globalisierung hat ihren"

Images: showing 4 out of 27

[See all images](#)



VII. Fazit

- Aufwand und Kosten-Nutzen-Verhältnis abschätzen
- Seien Sie pragmatisch!
- Lösungen an eigenen Bedarf anpassen
- Vorrangig: Bewusstsein für Webseiten als Teil der eigenen Überlieferung schaffen!

Noch Fragen?

Kontakt:

Johannes.Thome@lvr.de

02234/9854-466