

Elisabeth Mödden

# Das Webarchiv der DNB – Kurzer Praxisbericht

# Selektives Webarchiv

- Exemplarische Vielfalt
- Kollektionen
  - Thematisch, z. B. „Behörden und Institutionen des Bundes“, „Themen der Alltagskultur“
  - Ereignis, z. B. „Flüchtlingskrise in Deutschland ab 2015“, „100 Jahre Erster Weltkrieg“, „Bundestagswahl 2017“ etc.
- Thematische Kollektionen: I.d.R. zwei Crawls pro Jahr
- Ereignis: Variierende Anzahl von Crawls
- Derzeit insgesamt über 6.000 Websites mit ca. 28.000 Zeitschnitten
- Erschließung pro Website: Kollektion, Titel, URL, Datum der Archivierung (=Zeitschnitt)

# Bundesebene

Kuratierte Sammlung

mit Sammlungsschwerpunkt Bundesebene:

Die zu sammelnden Websites sollen „landeskundliche Relevanz“ haben:

- sich auf Deutschland beziehen oder
- Themen von historischer, sozialer, politischer, kultureller, religiöser, wissenschaftlicher oder wirtschaftlicher Bedeutung für Deutschland enthalten oder
- von deutschen oder mit Deutschland verbundenen Autoren geschaffen worden sein.

# Landesebene

Kuratierte Sammlung

mit Sammlungsschwerpunkt Landesebene:

Websites mit regionaler Relevanz werden im Rahmen der Kooperation mit regionalen Pflichtexemplarbibliotheken gesammelt:

- seit 2018 Webarchiv Thüringen,
- Start 2020 Webarchiv Hamburg,
- Start 2020 Testphase Webarchiv NRW
- ...

# Kollektionen

- Auktionshäuser, Galerien, Künstler
- Behörden und Institutionen des Bundes
- Biologie
- Frankfurt am Main
- Parteien, parteinahe Organisationen und Politiker
- Kultureinrichtungen
- Musik
- Medizin
- Recht
- ...
- Blogs
- **Ereignisse**
- Diverse / Einzelmeldungen

# Thematische Kollektionen

- Team Webarchiv
- Vorschläge Fachreferent\*innen/ Kolleg\*innen
- Linklisten, wie Academic Linkshare
- Workshops mit Wissenschaftlern
- Externe Vorschläge
- Durch den Einsatz automatischer Verfahren (Test mit Firma Mindup)
- Nach quantitativen Kriterien wie „Reichweitenstärke“ bzw. „Top-Listen“
- Collaborative collections mit [IIPC](#)

# Ereignisse / Event-Harvesting

Gesammelt werden:

- Websites zu einem bestimmten tagesaktuellen Ereignis
- Event-Websites, die sich nur mit dem Ereignis beschäftigen
- Unterseiten zum Event auf den Websites von Institutionen wie z.B. Museen, Stiftungen, Vereinen, Städten sowie Themenseiten auf Nachrichtenwebsites
- klar erkennbarer Deutschland-Bezug

# Ereignisse / Event-Harvesting

Vorhersehbare/geplante Ereignisse:

Themenjahre, Geburtstage / Todestage von Personen des öffentlichen Lebens, Jubiläen von Institutionen, Jahrestage historischer Ereignisse, Sportwettbewerbe, kulturelle Veranstaltungen wie Ausstellungen, Wettbewerbe, Festivals etc., Wahlen etc.

Unvorhergesehene Ereignisse:

Pandemien, Naturkatastrophen, Terroranschläge, Tod von Personen des öffentlichen Lebens etc.



## Ereignisse / Event-Harvesting

- Im Rahmen internationaler Kooperationen (z.B. mit dem IIPC) nimmt die DNB auch an gemeinschaftlichen Event-Harvestings teil, um die deutsche Sichtweise auf bestimmte Ereignisse in die Sammlung einzubringen
- Sammelfrequenz wird für jede Webseite individuell eingestellt (Je nach Änderungshäufigkeit bzw. inhaltlicher Relevanz kann z.B. eine einmalige oder auch tägliche Sammlung stattfinden)

# Webarchiv: Beispielseite

Archivierte Netzressource vom 02.05.2013 [www.auswaertiges-amt.de](http://www.auswaertiges-amt.de) Datensatz im Katalog

**"Stabilitätsanker" Mosambik**  
Anerkennung für die politische und wirtschaftliche Entwicklung - bei seinem Besuch in Maputo sprach sich Außenminister Westerwelle dafür aus, die Zusammenarbeit weiter zu intensivieren.

**Ghana: Schlüsselnd in Westafrika**

**NATO-Rat in Brüssel**

**Unsere Schwerpunkte**  
Abrüstung Europa  
Menschenrechte

**Bundesaußenminister Guido Westerwelle**  
Bundesaußenminister

**Syrien: Suche nach Wegen aus der Gewalt**  
Das Ringen um ein Ende

Es ist leider kein Flash-Plugin installiert. Für die Darstellung wird Flash benötigt.

**Die Mediathek**  
Fotos, Videos und andere Medien stehen zum kostenlosen Abruf in der

Impressum Hilfe powered by oia

# Zeitschnitte Auswärtiges Amt

02.05.2013

14.10.2015

02.02.2017

25.07.2019

Zeitschnitte Auswärtiges Amt

# Ereignis Klimawandel

eine kleine Auswahl der geharvesteten Websites zum Thema

## Ereignis → Klimawandel



Misereor, Thema Klimawandel  
Zeitschnitt: 10.06.2019



Blog: Klimalounge  
Zeitschnitt:  
07.06.2019

Deutsches  
Klimaportal  
Zeitschnitt:  
22.11.2018



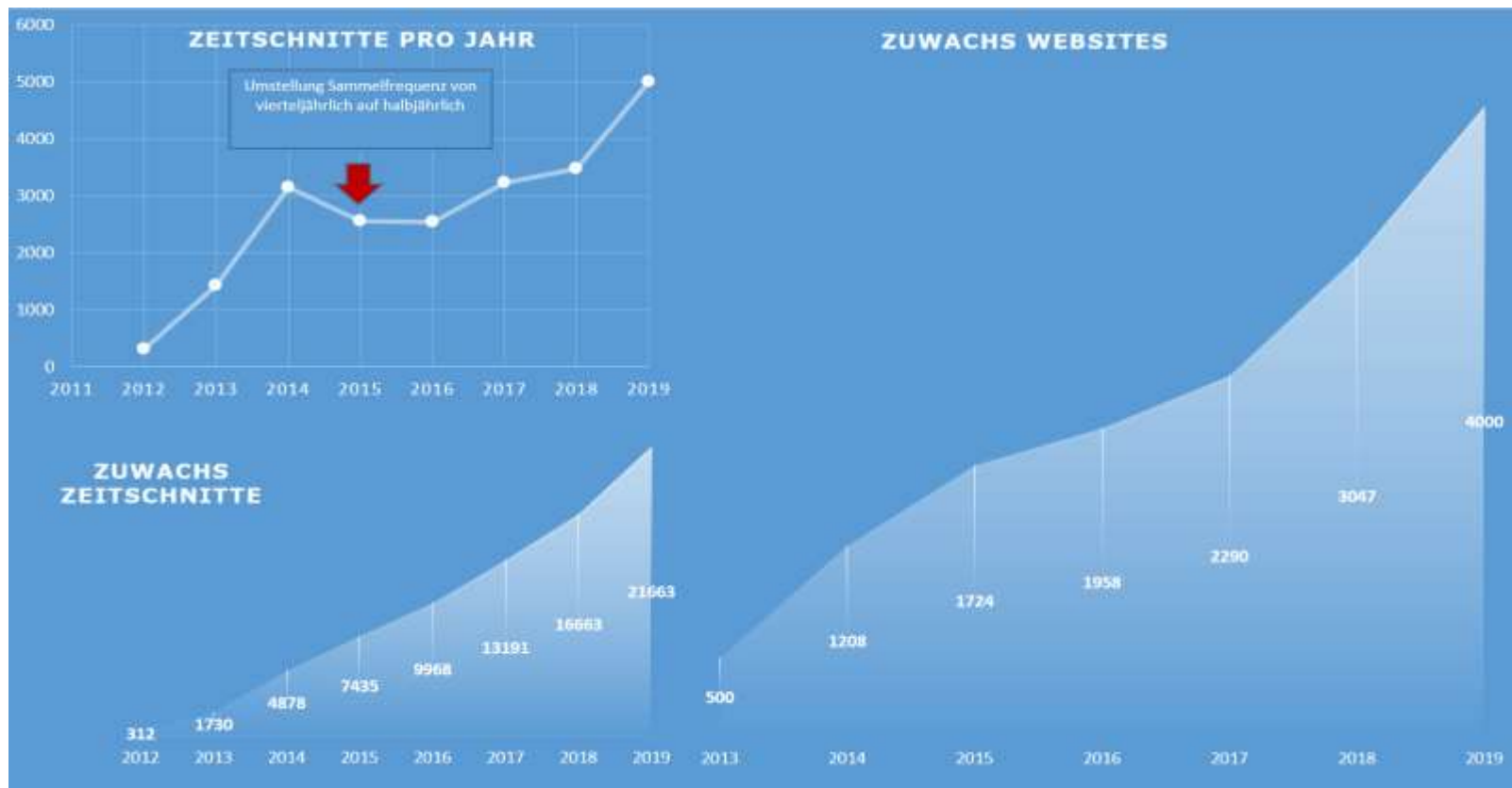
Die Klima-Lüge Zeitschnitt: 24.07.2019



Weltklimarat IPCC, Deutsche Koordinierungsstelle  
Zeitschnitt 16.08.2019



# Selektives Webarchiv

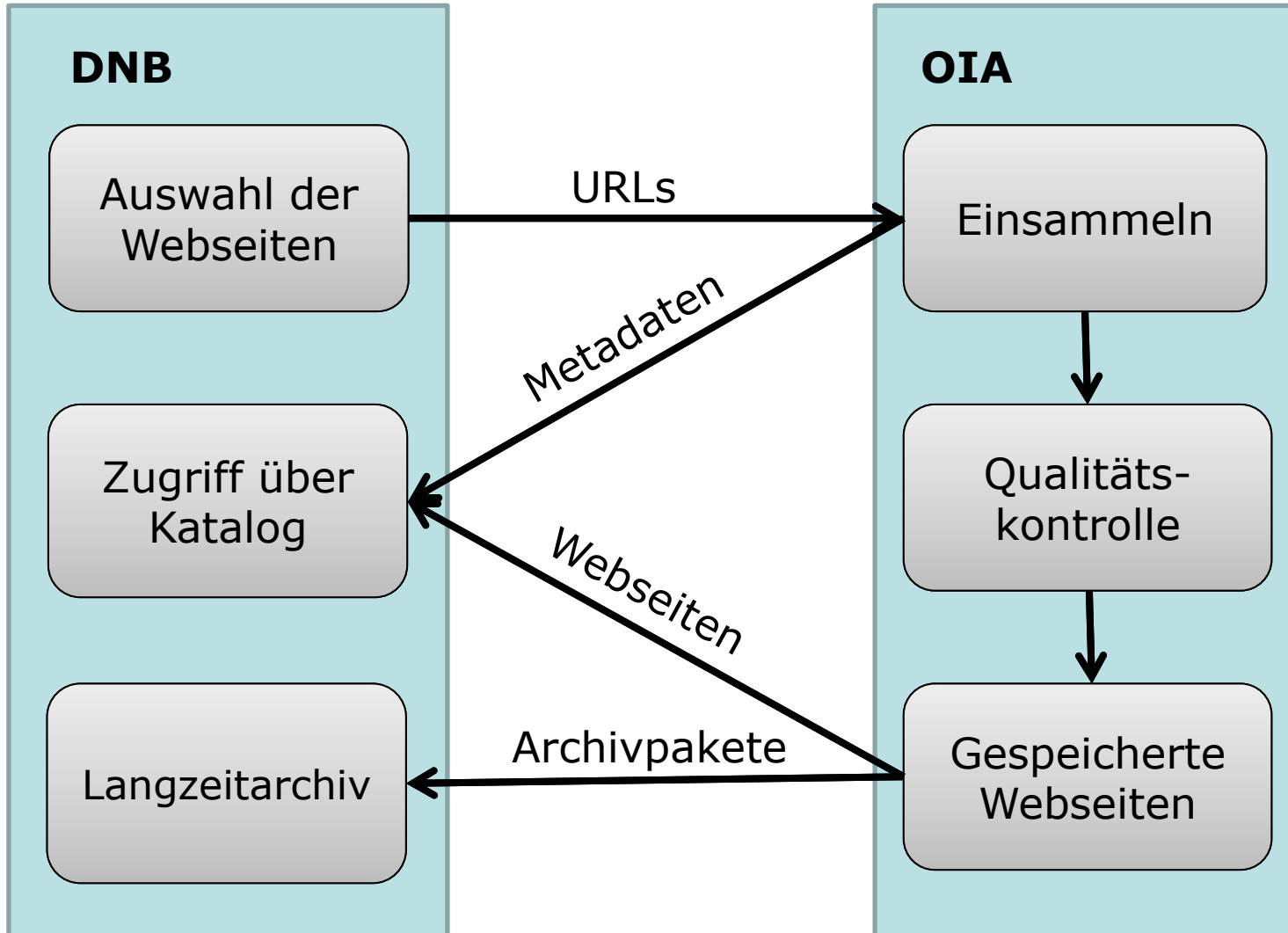


# Zusammenarbeit mit Dienstleister oia



- Seit 2012 Workflow mit Dienstleister oia GmbH
- Auswahl von DNB, Erfassung in Tool OWA von oia
- Sammlung, Qualitätssicherung und Speicherung für Zugriff durch den Dienstleister mit eigener Software auf deren Servern in Düsseldorf
- Zugriff in Lesesälen über den Katalog, eigenes Portal und Volltextsuche
- Derzeit: Infrastrukturübernahme zur DNB Frankfurt

# Workflow



# Erfassungstool: OWA-Client

**System zur Archivierung von Webseiten**

**Offline Web Archiv**

- Archivierte Webseiten bis 2008
- Behörden und Institutionen des Bundes
- Blogs
- Buchwissenschaft
- Digitale Langzeitarchivierung
- Diverse
- Ereignisse
- Forschungseinrichtungen
- Geschichte
- Interessenverbände
- Kultureinrichtungen
- Musik
- Nachrichtenwebsites
- Parteien, parteinahe Organisationen und Politiker
- Religionsgemeinschaften, Kirchen und religiöse Verbände
- Sozialversicherung – Träger, Verbände, Interessenvertretungen
- Sportverbände
- VifaTest
- Webarchiv Thüringen
- Wirtschaftswissenschaften
- Wissenschaftliche Fachgesellschaften**

**Offline Web Archiv**

	$\Sigma$ Anzahl	$\Sigma$ Volumen	$\Sigma$ AIU
<b>Gruppen</b>	<b>69</b>		
<b>Projekte</b>	<b>2660</b>		
<b>Spiegelungen</b>	<b>14.350</b>	<b>15,707 TB</b>	<b>142.455.971</b>



# Erfassung in OWA

The screenshot displays the OWA (Offline Web Archiv) system interface. The main window title is "System zur Archivierung von Webseiten". On the left, a sidebar lists various categories under "Offline Web Archiv", with "Forschungseinrichtungen" selected. A dialog box titled "Projekte" is open in the foreground, containing the following fields:

- Name:** Name der Website
- URL:** URL der Website
- Dropdown 1:** DNB - halbjährlich
- Dropdown 2:** OfflineWebArchive
- Dropdown 3:** 172.16.2.131
- Bemerkung:** (Empty text area)

At the bottom of the dialog box, there are two buttons: "Abbrechen" (with a red X icon) and "übernehmen" (with a green checkmark icon).

# OWA - Sammlungsfrequenz

Eigenschaften

 Akademie für Leseförderung Niedersachsen

- ↳ Allgemein
  - URL
  - Dateien
  - Info
  - Subdomains
- ↳ Datei Filter
  - Text
  - Bilder
  - Video
  - Audio
  - Archiv
  - Benutzerdefiniert
  - Sonstiges
- ↳ URL Filter
  - Protokoll
  - Server
  - Verzeichnis
  - Datei
  - Terminplaner**
  - Sicherheit
- ↳ Erweitert
  - URL-Ersatz
  - Makros
  - Interface
  - Regulär Ausdruck

**Terminplaner**

Scheduler

Halbjährlich 00:00

Kalender

Freitag , 30. März 201

min  max

Spiegelung beenden, spätestens nach [Minuten]

## Datenmodell Webharvesting

- Katalogeintrag für jede archivierte Website
- Drei Ebenen der Erschließung:
  1. Kollektion, innerhalb derer die Website gesammelt wird (thematische Sammlung oder Event-Kollektion)
  2. Website (= Gesamtheit aller Webseiten eines Internet-Angebots, z. B. <http://www.bundeskanzlerin.de>)
  3. Zeitschnitte, d.h. die einzelnen Harvesting-Vorgänge (z. B. Snapshot vom 16.10.2017)

# Katalogansicht (PICA 3)

## Kollektion:

0500 Odf  
 0600 kl  
 1101 cr  
 2150 d6088cff-fbe5-e111-ba3c-d4ae528b7600  
 2240 DNB:1048103935  
 4000 Behörden und Institutionen des Bundes [[Elektronische Ressource]]  
 4060 Online-Ressource

## Website:

0500 Obf  
 0600 ro;ws  
 1100 2013  
 1101 cr  
 2105 14,004  
 2150 abb304f9-59da-46a0-be26-d009fccae6ed  
 2240 DNB:1048133729  
 4000 Auswärtiges Amt, AA [[Elektronische Ressource]]  
 4060 Online-Ressource  
 4083 =A \$AIC=abb304f9-59da-46a0-be26-d009fccae6ed  
 4085 \*HTTP\*=u <http://www.auswaertiges-amt.de/>  
 4190 !1048103935!Behörden und Institutionen des Bundes  
 4233 \*a\*

## Zeitschnitt:

0500 Olfo  
 0501 Text\$bbt  
 0502 Computermedien\$bc  
 0503 Online-Ressource\$bcr  
 0550 PICA+\$bOAI\$cOIA  
 0551 X\$bzm  
 0600 zw  
 1100 2017  
 1101 cr  
 1131 !959344357!Website [Ts1]  
 2150 38ba416b-c533-e711-bc07-d4ae528b7600  
 2240 DNB:1133544223  
 4060 Online-Ressource  
 4070 /d08/m05/b2017  
 4083 =A \$GUID=38ba416b-c533-e711-bc07-d4ae528b7600  
 4241 Zu:!1048133729!--Obf--: Auswärtiges Amt, AA

## Portalansicht: Kollektion

	
<b>Link zu diesem Datensatz</b>	<a href="http://d-nb.info/1048103935">http://d-nb.info/1048103935</a>
<b>Art des Inhalts</b>	Website-Kategorie
<b>Titel</b>	Behörden und Institutionen des Bundes
<b>Umfang/Format</b>	Online-Ressource
<b>Zugehörige Websites</b>	<p>121 Websites</p> <ol style="list-style-type: none"> <li>1. <i>Staatsministerin für Kultur und Medien [Elektronische Ressource]</i></li> <li>2. <i>Bundeskanzleramt [Elektronische Ressource]</i></li> <li>3. ...</li> </ol>

## Portalansicht: Website

	
<b>Link zu diesem Datensatz</b>	<a href="http://d-nb.info/1048133729">http://d-nb.info/1048133729</a>
<b>Art des Inhalts</b>	Archivierte Website
<b>Titel</b>	Auswärtiges Amt, AA
<b>Zeitliche Einordnung</b>	Erscheinungsdatum: 2013-
<b>Umfang/Format</b>	Online-Ressource
<b>URL</b>	<a href="http://www.auswaertiges-amt.de/">http://www.auswaertiges-amt.de/</a>
<b>Beziehungen</b>	Behörden und Institutionen des Bundes
<b>Anmerkungen</b>	Langzeitarchivierung gewährleistet
<b>Zugehörige Zeitschnitte</b>	<p>13 Zeitschnitte</p> <ol style="list-style-type: none"> <li>1. 08.05.2017 Zu: Auswärtiges Amt, AA</li> <li>2. 02.02.2017 Zu: Auswärtiges Amt, AA</li> <li>3. ...</li> </ol>
<b>Online-Zugriff</b>	Webarchiv öffnen (nur im Lesesaal möglich)

## Portalansicht: Zeitschnitt

Treffer 1 von 13

	
<b>Link zu diesem Datensatz</b>	<a href="http://d-nb.info/1133544223">http://d-nb.info/1133544223</a>
<b>Art des Inhalts</b>	Website Zeitschnitt
<b>Umfang/Format</b>	Online-Ressource
<b>Beziehungen</b>	Zu: Auswärtiges Amt, AA (08.05.2017)
<b>Online-Zugriff</b>	<a href="#">Webarchiv öffnen (nur im Lesesaal möglich)</a>

# Zusammenarbeit mit Regionalbibliotheken seit 2018

- Grundlage: Urheberrechtsgesetz (Änderung mit Wirkung zum 01.03.2018)
- Gemeinsame Auswahl von zu sammelnden Seiten: Import von Listen im Erfassungstool von oia möglich
- Bereitstellung des selektiven Webarchivs in den Lesesälen der regionalen Pflichtexemplarbibliothek: Freischaltung von IP-Adressräumen für den Zugriff möglich



# Kooperation



- Angebot, bei der Auswahl der Websites zusammenzuarbeiten und Zugriff auf das gesamte Webarchiv
- Ziel ist es Erfahrungen zu sammeln
- Vertrag mit dem Dienstleister oia ermöglicht es, die Menge der jährlichen Spiegelungen (Crawls) stufenweise zu erweitern
- Start jeweils mit überschaubaren Mengenkontingenten (ca. 250 bis 500 Websites)
- späteren Ausbau der Kontingente nach Bedarf und im Rahmen der zur Verfügung stehenden Möglichkeiten (3 Monate Vorlauf für die zu erwartenden Steigerungsraten)
- Start ab 2021

# Übermittlung von Websites durch Regionalbibliotheken

Die Websites für das Webarchiv werden in einer Excel Tabelle zusammengestellt und an die DNB übermittelt:

Spalte 1	Spalte 2	Spalte 3	Spalte 4	Spalte 5	Spalte 6	Spalte 7
Kollektion 1	Kollektion2	Titel / Projektname	Start-URL	Beteiligte Körperschaften	Beteiligte Personen	Schlagwörter
NRW		Bundesstadt Bonn	<a href="https://www.bonn.de/">https://www.bonn.de/</a>			
NRW	Musik	Beethoven Orchester Bonn	<a href="https://www.beethoven-orchester.de/">https://www.beethoven-orchester.de/</a>			

Die blau markierten Spalten müssen ausgefüllt werden, die weißen Spalten können optional belegt werden.



## DE-Domain-Crawl

- Ergänzendes Angebot zum selektiven Harvesting
- Einmaliger Crawl 2014 durch die Internet Memory Foundation
- Zugriff für die Öffentlichkeit derzeit nicht möglich
- DE-Portal beim Internet Archive
  - Zugriffsseite zu allen DE-Seiten im IA-Datenbestand
  - Stichwort- und URL-Suche
  - Portalzugriff aus den Lesesälen

# Internet Archive DE-Portal


Startseite > .de-Webarchiv-Suche

## .DE-WEBARCHIV-SUCHE


Recherche nach .de-Webseiten und darauf verlinkte Dateiformate:

Stichwort ▾



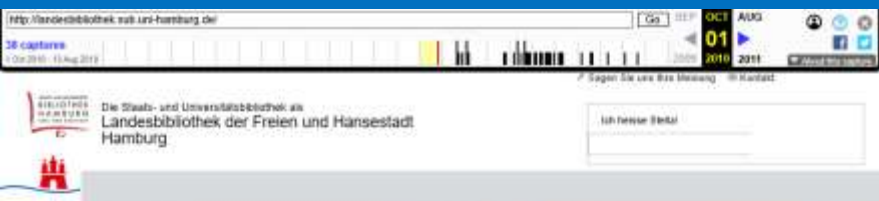
 **Webseiten**

 Bilder

 Audio

 Video

 PDF



DEUTSCHE  
NATIONAL  
BIBLIOTHEK

24.01.2019

01.10.2010

22.01.2012