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:

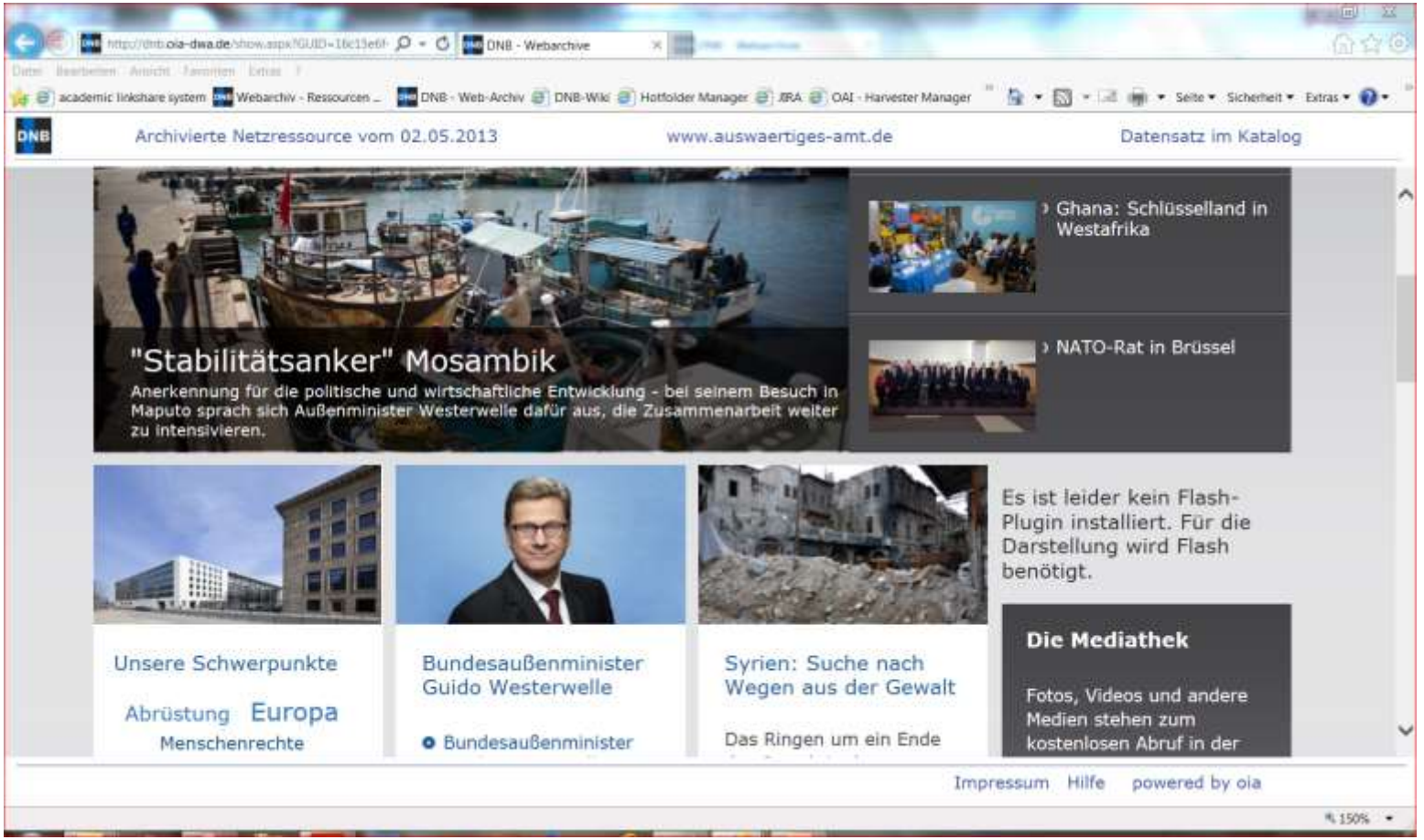
Pandemien, Naturkatastrophen, Terroranschläge, Tod von Personen des öffentlichen Lebens etc.

Ereignisse / Event-Harvesting

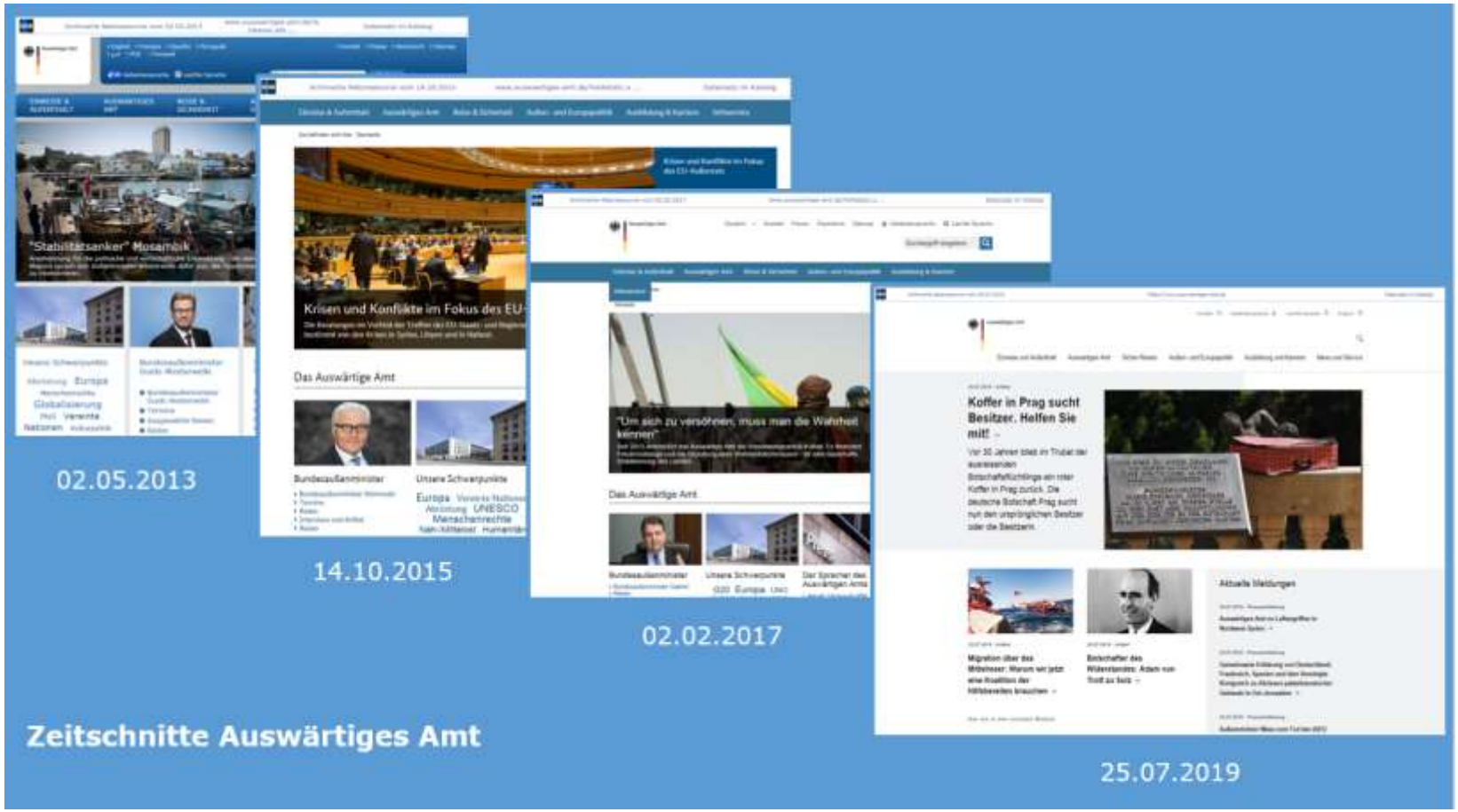
- Im Rahmen internationaler Kooperationen (z.B. mit dem IIPC) nimmt die DNB auch an gemeinschaftlichen Event-Harvestings teil, um die deutsche Sichtweise auf bestimmte Ereignisse in die Sammlung einzubringen

- 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

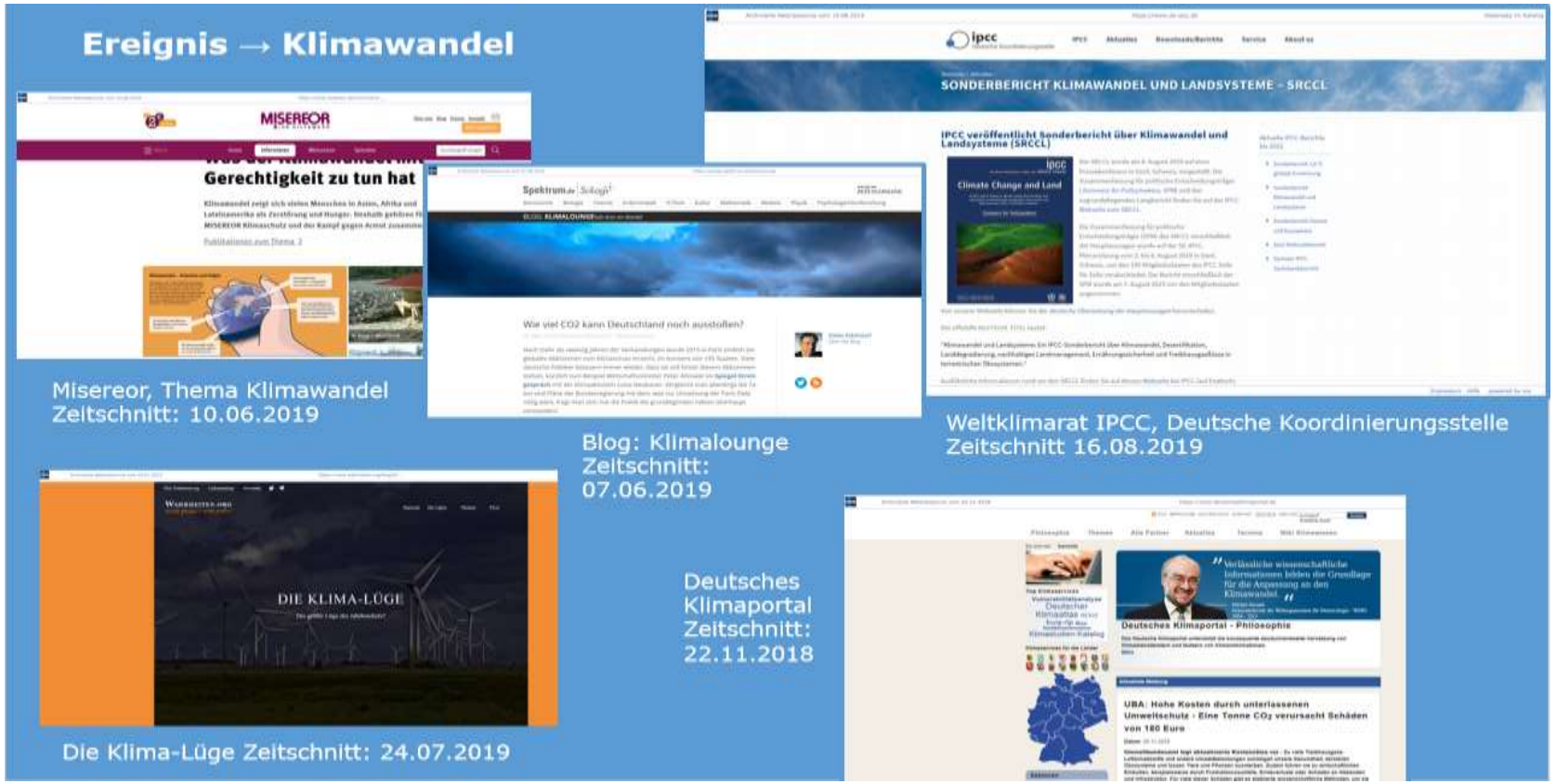


Zeitschnitte Auswärtiges Amt



Ereignis Klimawandel

eine kleine Auswahl der geharvesteten Websites zum Thema

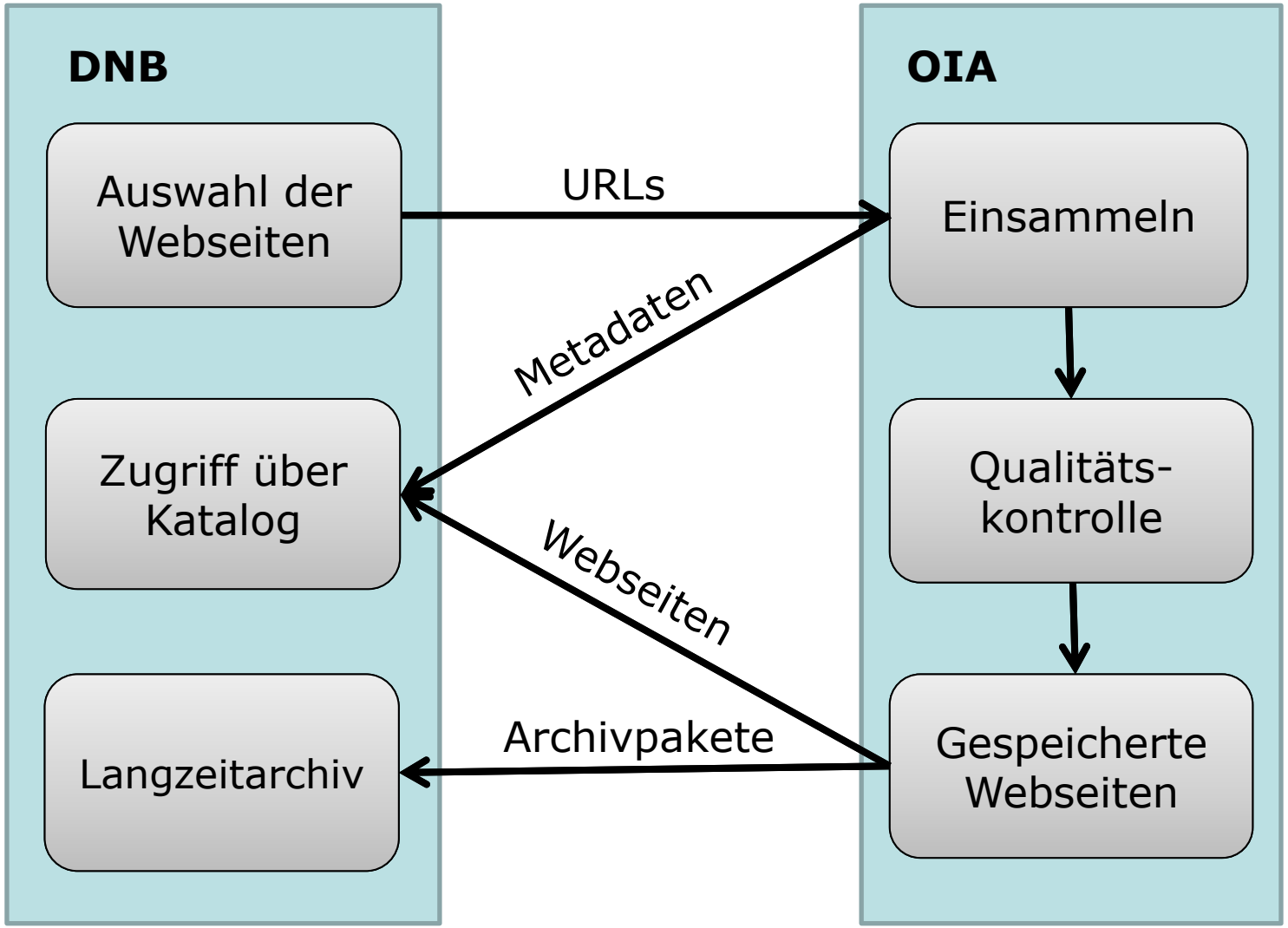


Selektives Webarchiv

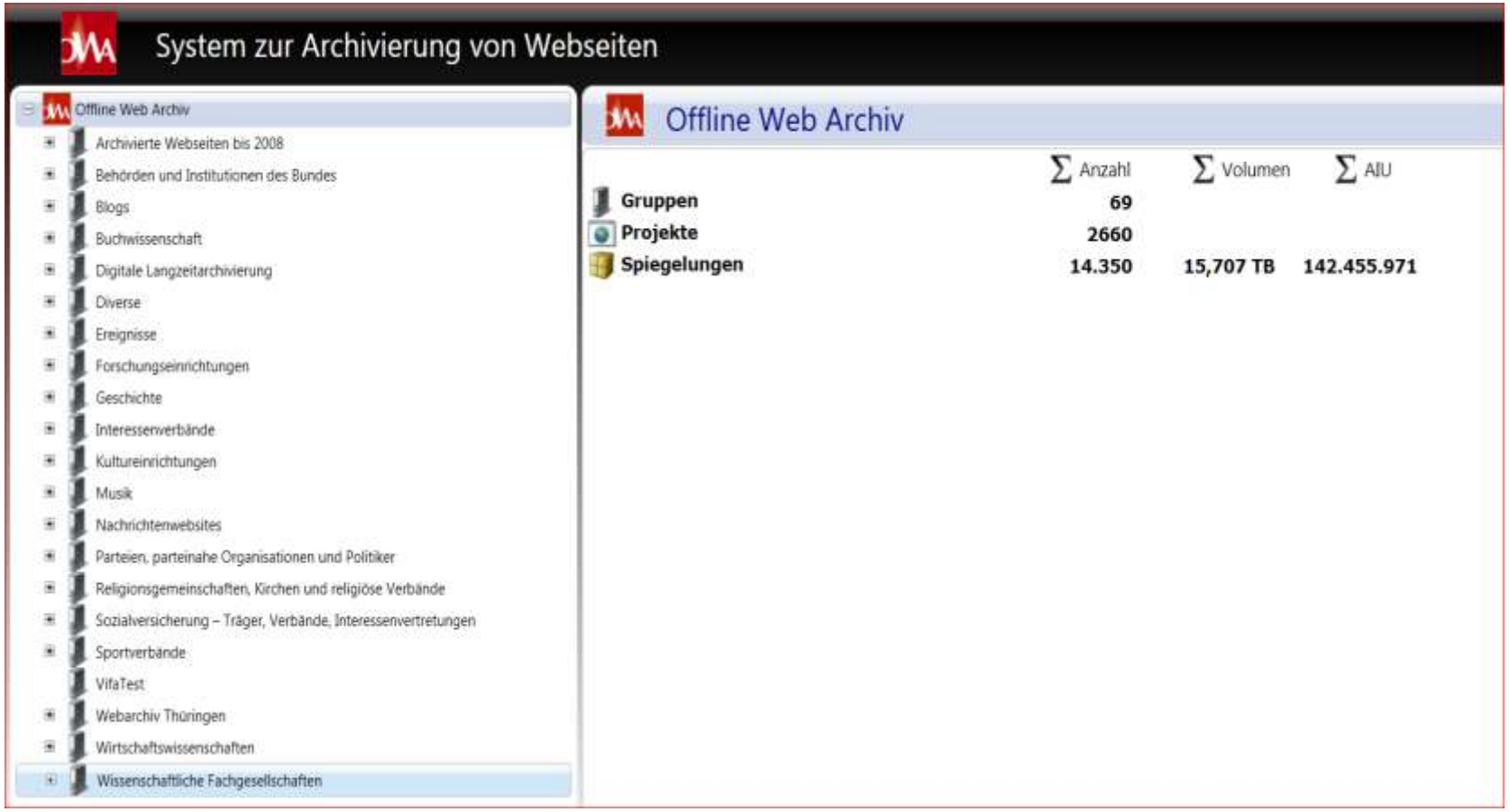
Zusammenarbeit mit Dienstleister oia

- 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

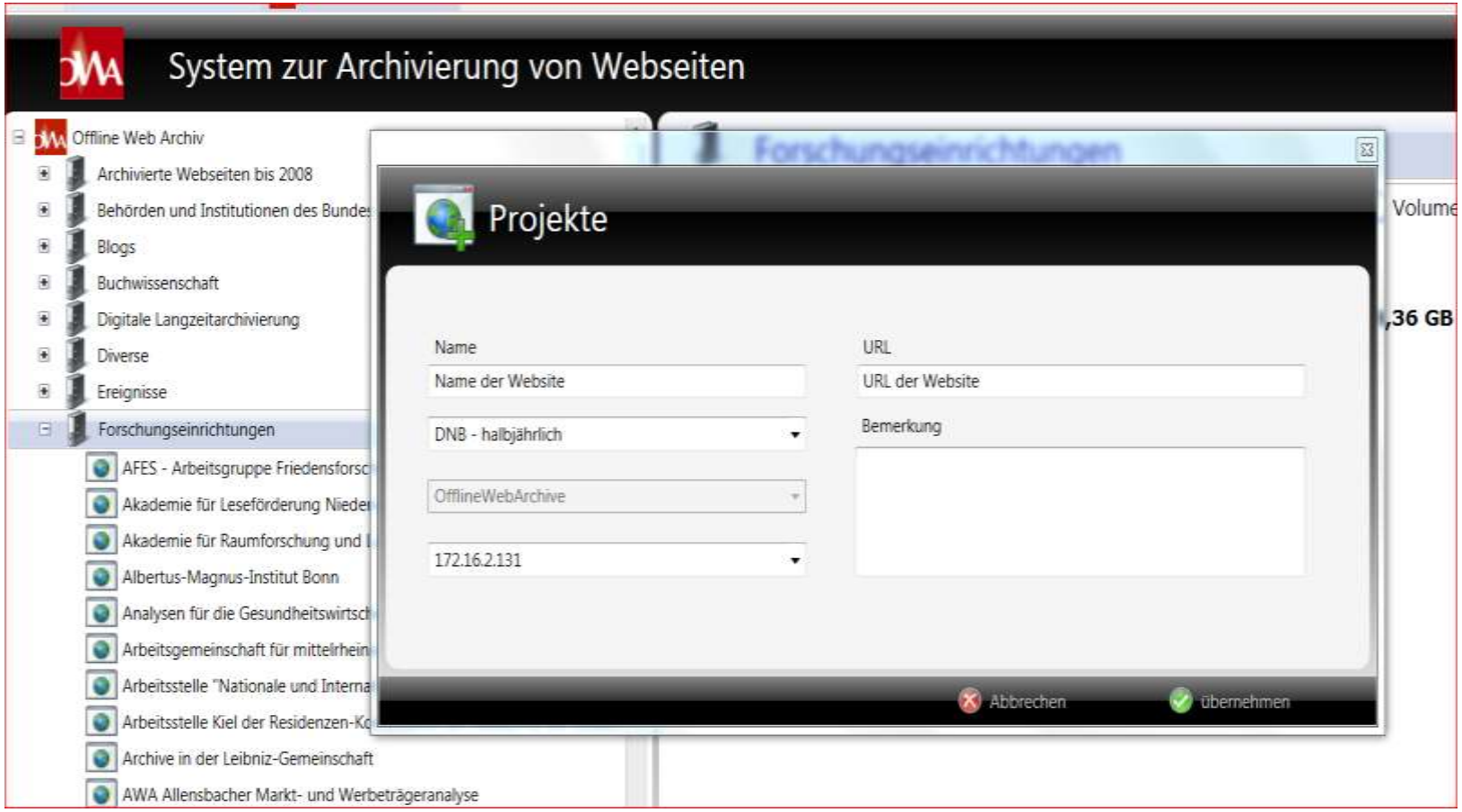
Workflow


DNB
Auswahl der Webseiten
Zugriff über Katalog
Langzeitarchiv
OIA
Einsammeln
Qualitäts-kontrolle
Gespeicherte Webseiten
URLs
Metadaten
Webseiten
Archivpakete

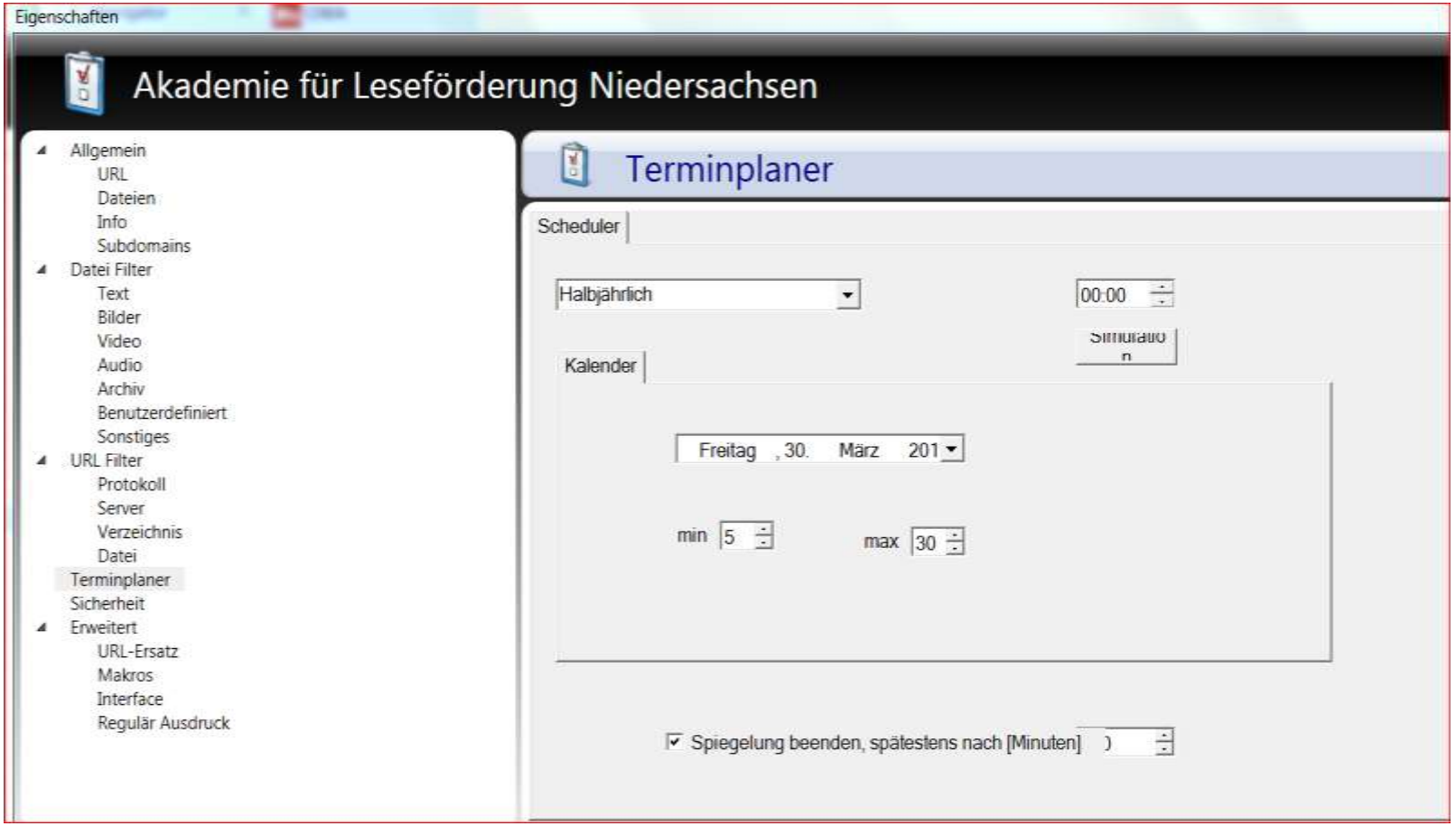
Erfassungstool: OWA-Client



Erfassung in OWA



OWA - Sammlungsfrequenz



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,O04
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$btxt
0502 Computermedien$bc
0503 Online-Ressource$bcr
0550 PICA+$bOAI$cOIA
0551 X$bm
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

Link zu diesem Datensatz	http://d-nb.info/1048103935
Art des Inhalts	Website-Kategorie
Titel	Behörden und Institutionen des Bundes
Umfang/Format	Online-Ressource
Zugehörige Websites	121 Websites 1. *Staatsministerin für Kultur und Medien [Elektronische Ressource]* 2. *Bundeskanzleramt [Elektronische Ressource]* 3. ...

Portalansicht: Website

Link zu diesem Datensatz	http://d-nb.info/1048133729
Art des Inhalts	Archivierte Website
Titel	Auswärtiges Amt, AA
Zeitliche Einordnung	Erscheinungsdatum: 2013-
Umfang/Format	Online-Ressource
URL	http://www.auswaertiges-amt.de/
Beziehungen	Behörden und Institutionen des Bundes
Anmerkungen	Langzeitarchivierung gewährleistet
Zugehörige Zeitschnitte	13 Zeitschnitte 1. *08.05.2017* *Zu: Auswärtiges Amt, AA* 2. *02.02.2017* *Zu: Auswärtiges Amt, AA* 3. ...

Online-Zugriff	Webarchiv öffnen (nur im Lesesaal möglich)

Portalansicht: Zeitschnitt

Treffer 1 von 13

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

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:

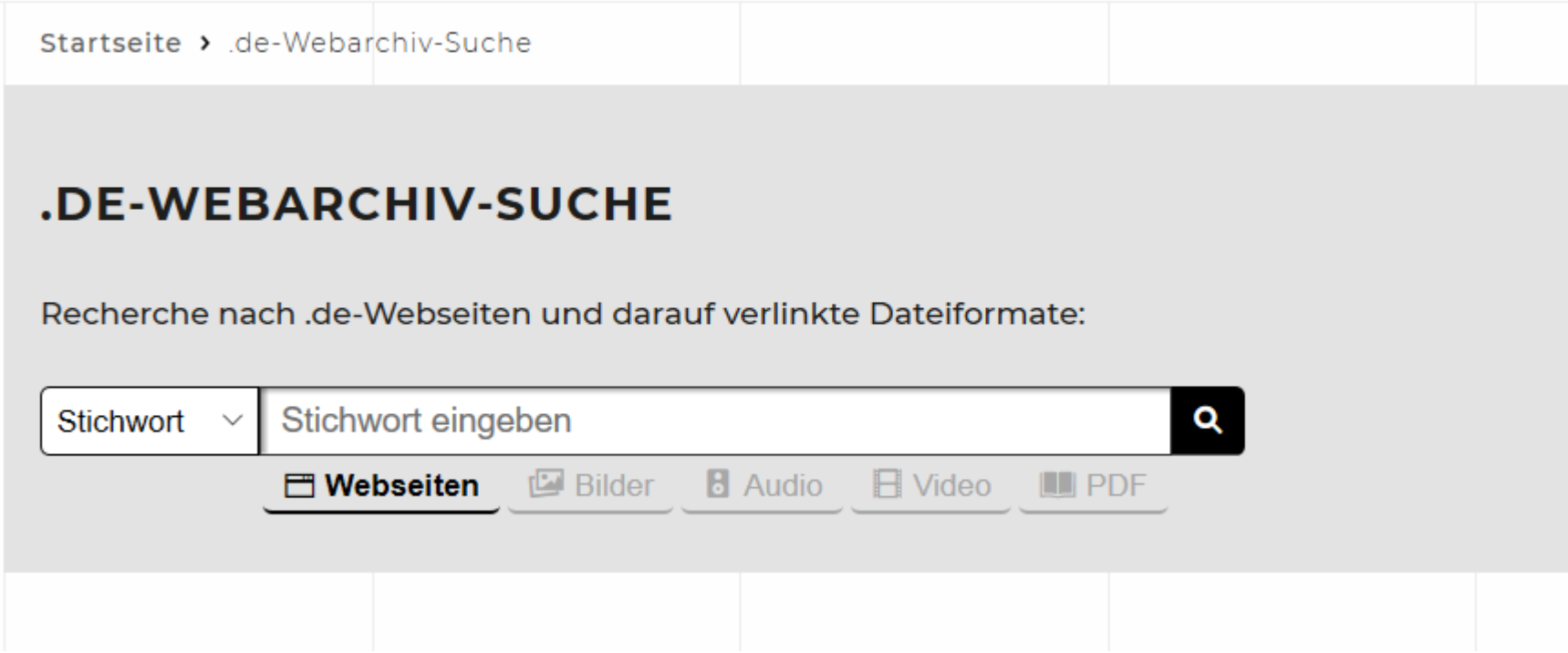
Spalte 1	Spalte 2	Spalte 3	Spalte 4	Spalte 5	Spalte 6	Spalte 7
Kollektion 1	Kollektion2	Titel / Projektname	Start-URL	Beteiligte Körperschaften	Beteiligte Personen	Schlagwörter
NRW		Bundesstadt Bonn	https://www.bonn.de/			
NRW	Musik	Beethoven Orchester Bonn	https://www.beethoven-orchester.de/			

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

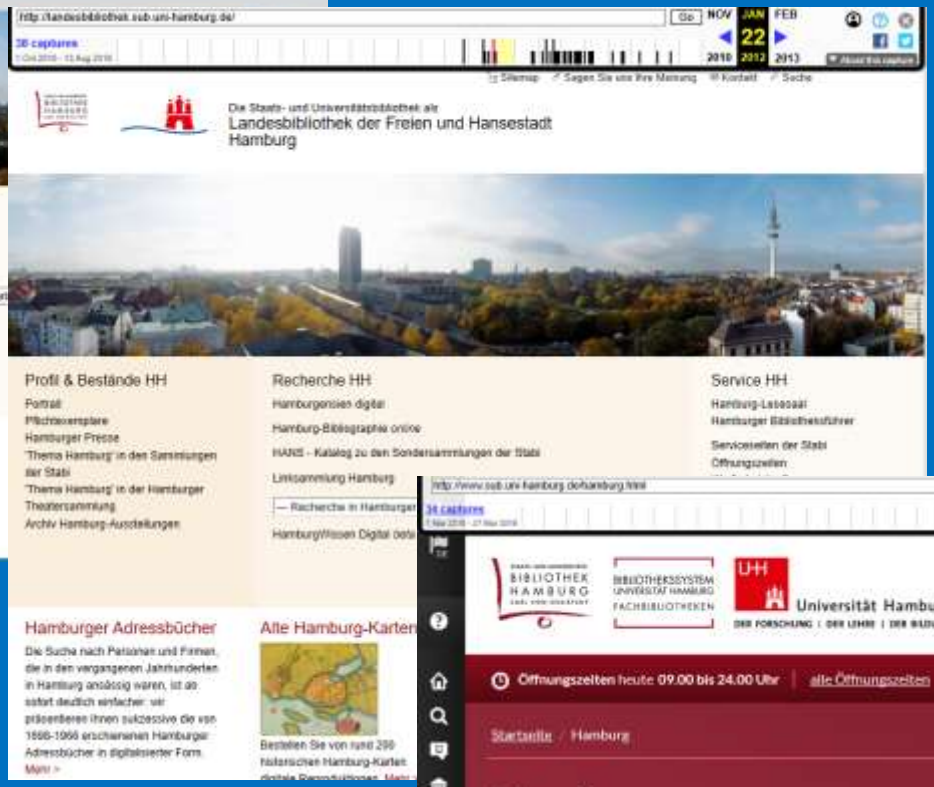
DE-Domain-Crawl

- 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

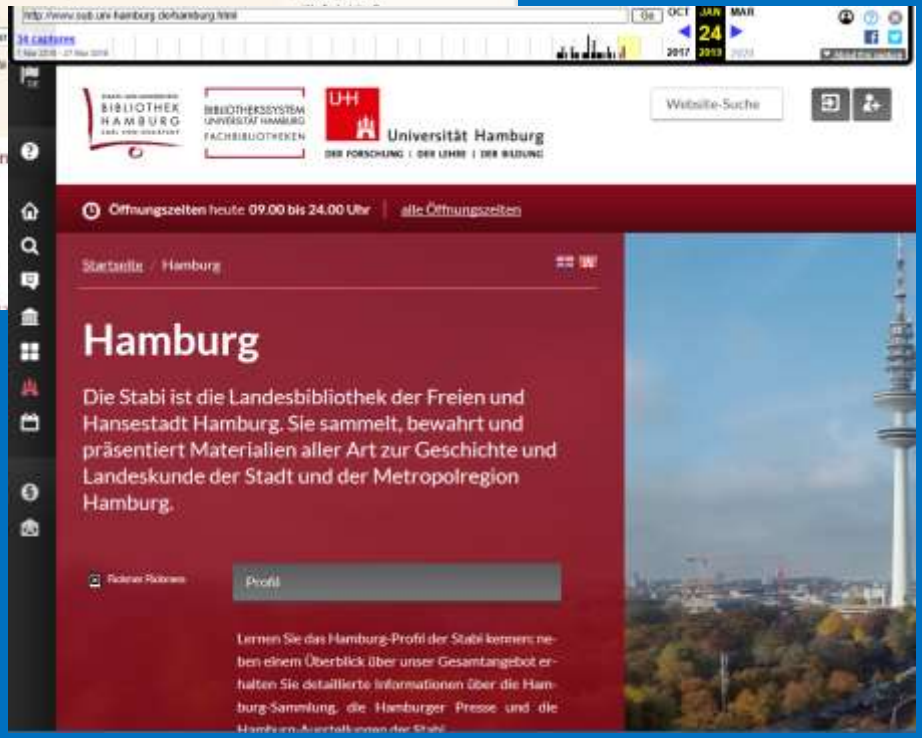
Internet Archive DE-Portal



01.10.2010



22.01.2012



24.01.2019