DEUTSCHE NATIONAL BIBLIOTHEK

Tobias Steinke

Das deutsche Internet archivieren? Zwischen selektivem Ansatz und .de-Domain-Crawl

Sammelauftrag

- Gesetz über die Deutsche Nationalbibliothek von 2006: Sammelauftrag auch für „alle Darstellungen in öffentlichen Netzen“
- Sammlung von Netzpublikationen über Ablieferung: E-Books, elektronische Zeitschriften, E-Paper, Hochschulprüfungsarbeiten, Musikdateien, Hörbücher, Digitalisate
- Erschließung (Katalog), Zugriff in den Lesesälen, Langzeitarchivierung

Das deutsche Internet?

- Gedruckte Publikationen: Deutscher Erscheinungsort im Impressum
- Webseiten: Impressumspflicht nur für kommerzielle Seiten in Deutschland
- .de-Domain nicht zwingend nötig für Webseiten deutscher Herkunft
- Deutsche Sprache und Serverstandort sind keine hinreichenden Merkmale
- Jede Sammlung nur Annäherung, keine Vollständigkeit

Webharvesting an der DNB

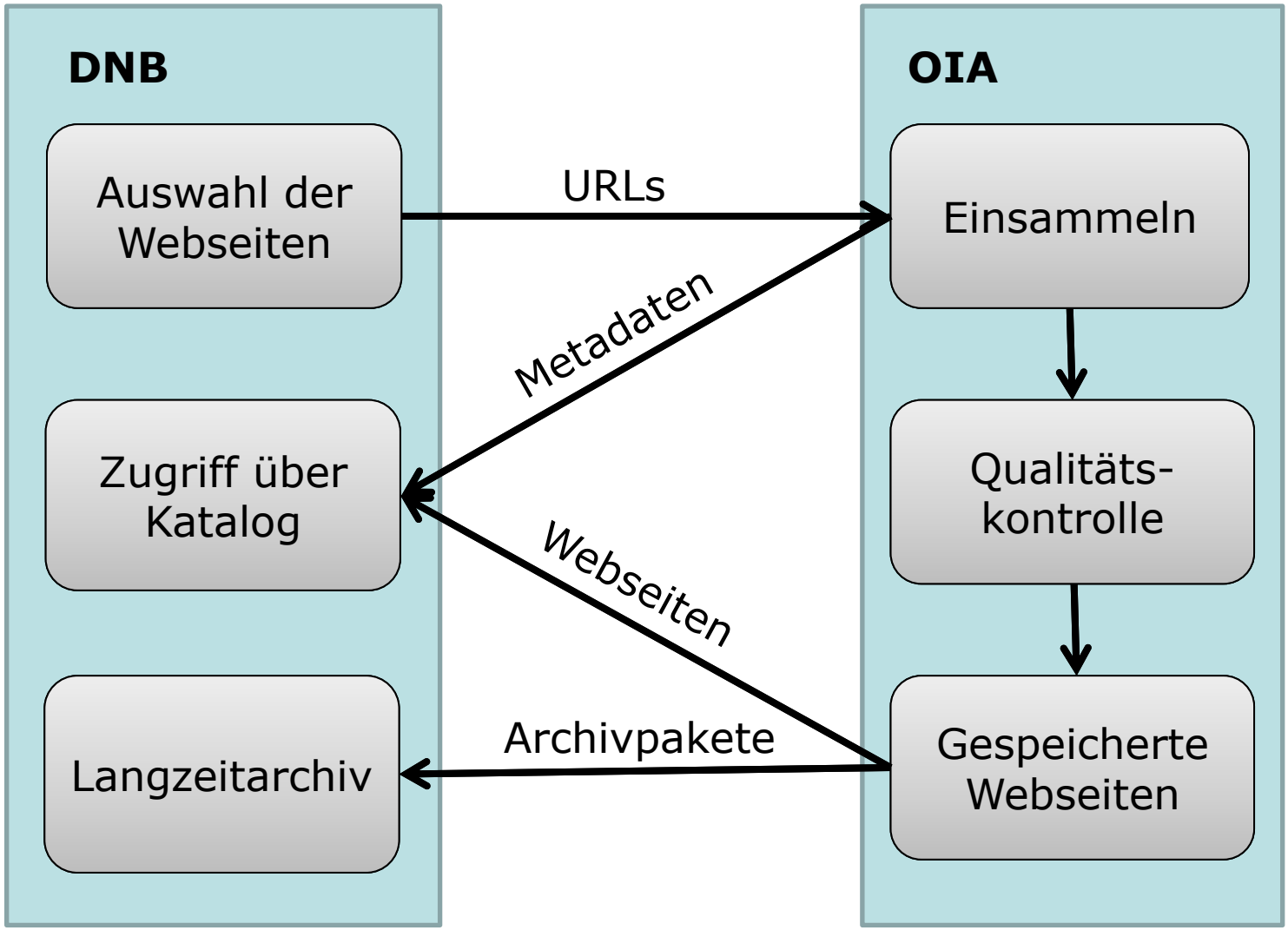
- Event-Crawls zur Bundestagswahl 2004 und EU-Ratspräsidentschaft 2007 mit European Archive
- Mitgliedschaft beim International Internet Preservation Consortium (IIPC) seit 2007
- Seit 2010 Aufbau eines Webarchivierungs-Geschäftsgangs
- 2014: Experimenteller .de-Domain-Crawl

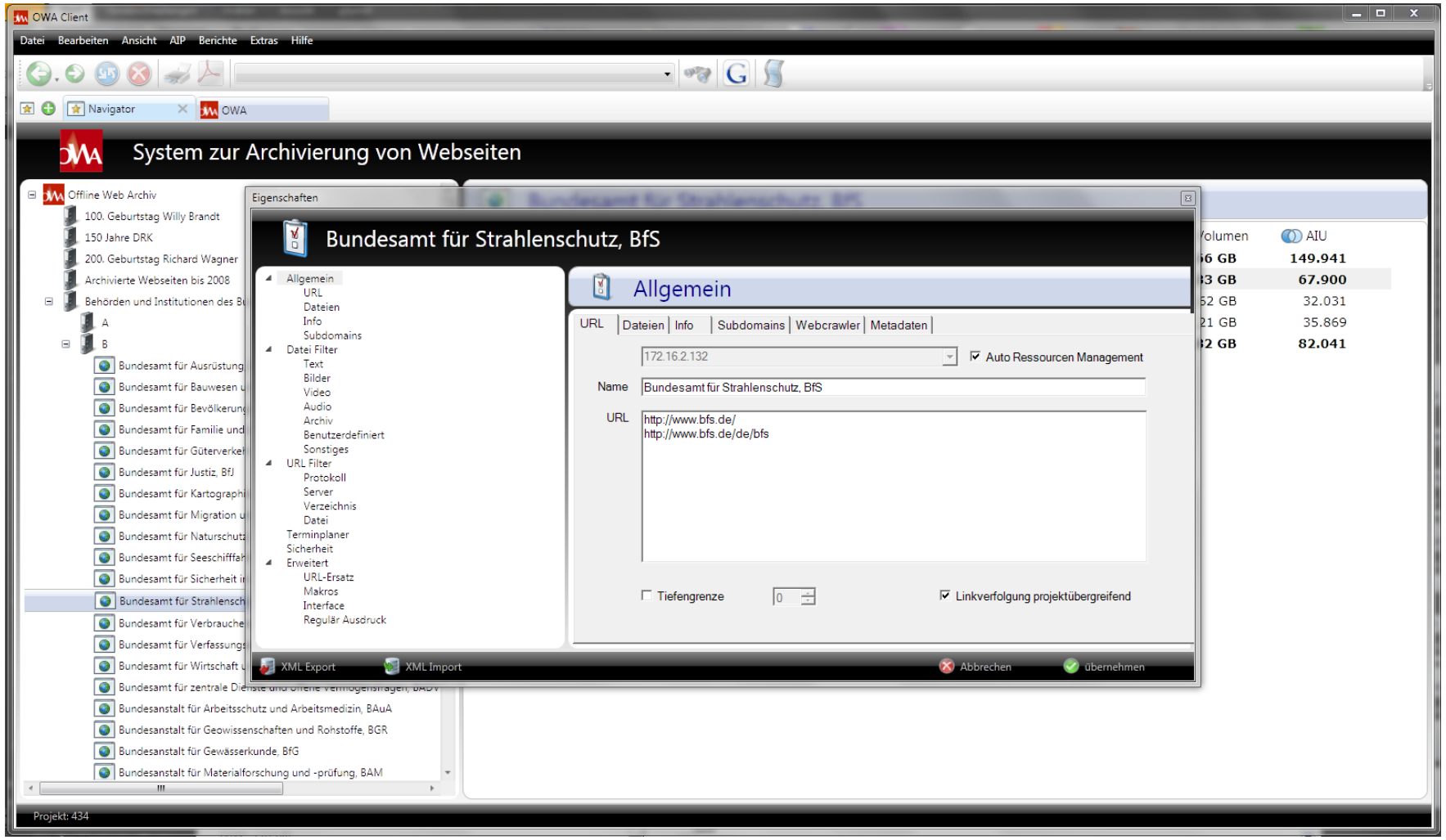
Aufbau eines Webarchivierungs-Geschäftsgangs

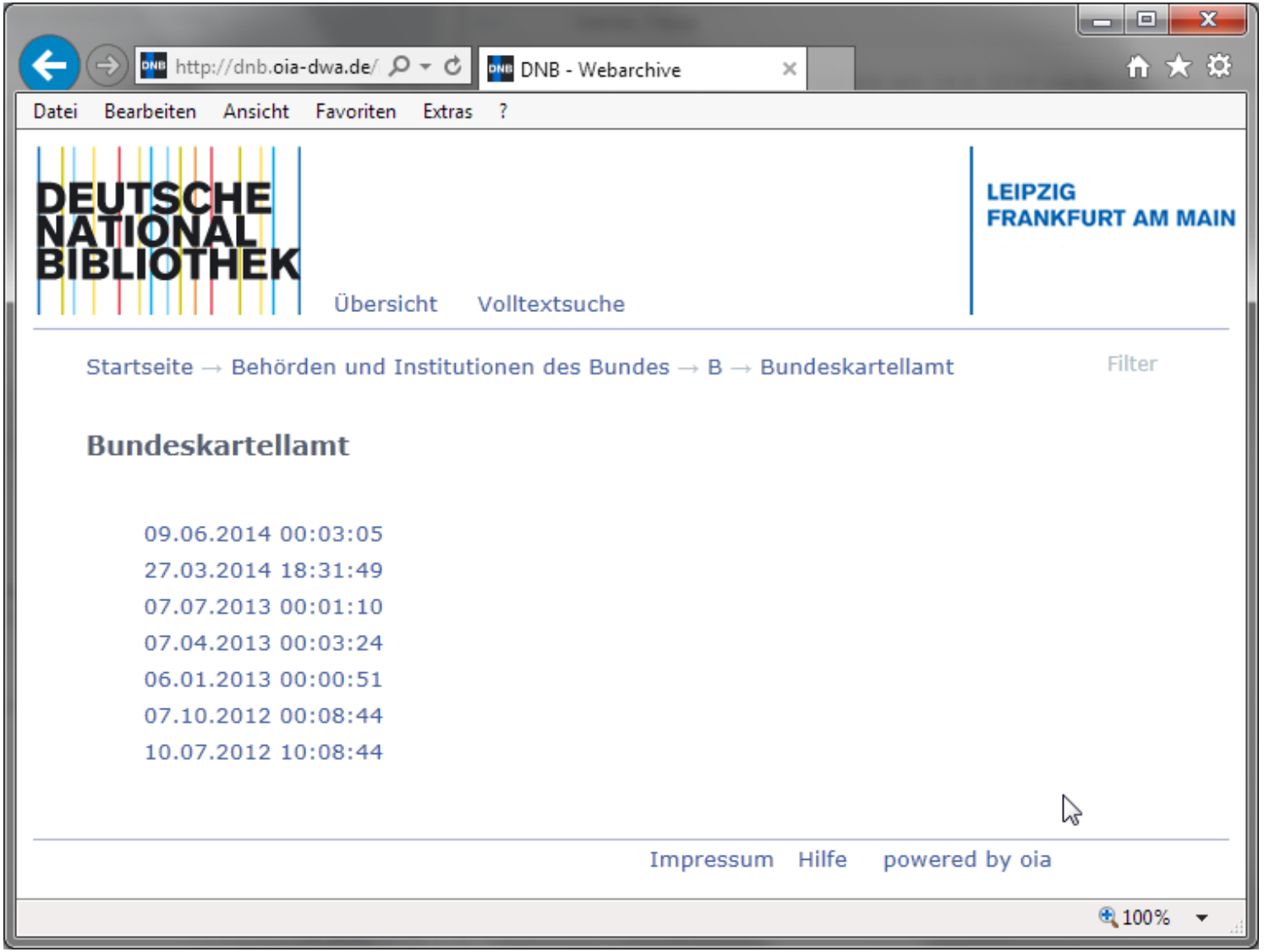
- Internes Projekt zum Vergleich der Optionen für alle Schritte im Geschäftsgang (Auswahl, Sammlung, Erschließung, Bereitstellung, Archivierung)
- Aufwandsabschätzung für Sammlungsentwicklung, Erfahrungen aus vergleichbaren Institutionen
- Ergebnis: Selektives Crawling nach thematischer Auswahl und zu Events; Sammlung, Speicherung, Qualitätssicherung und Bereitstellung über Dienstleister; Langzeitarchivierung bei DNB; .de-Crawl als Ergänzung

Selektiver Workflow

- Europaweite Ausschreibung 2011 für Dienstleister: Vergabe an deutsche Firma oia (www.oia-duesseldorf.de)
- Auswahl von Seiten durch DNB, Erschließung mit Titel und Kategorie
- Sammlung mit eigener Harvester-Software bei oia, manuelle Qualitätskontrolle
- Metadaten pro Site automatisch in DNB-Katalog
- Zugriff im DNB-Lesesaal per Katalog und Volltextsuche
- Archivierung als WARC-Dateien in DNB-Langzeitarchiv


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


OWA Client
Datei Bearbeiten Ansicht AIP Berichte Extras Hilfe
Navigator
OWA
System zur Archivierung von Webseiten
Offline Web Archiv
100. Geburtstag Willy Brandt
150 Jahre DRK
200. Geburtstag Richard Wagner
Archivierte Webseiten bis 2008
Eigenschaften
Bundesamt für Strahlenschutz, BfS
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
Allgemein
URL
Dateien
Info
Subdomains
Webcrawler
Metadaten
172.16.2.132
Auto Ressourcen Management
Name
Bundesamt für Strahlenschutz, BfS
URL
http://www.bfs.de/
http://www.bfs.de/de/bfs
Tiefengrenze
Linkverfolgung projektübergreifend
XML Export
XML Import
Abbrechen
übernehmen
Volumen
AIU
149.941
67.900
32.031
35.869
82.041
Projekt: 434


http://dnb.oia-dwa.de/
DNB - Webarchive
Datei Bearbeiten Ansicht Favoriten Extras ?
DEUTSCHE NATIONAL BIBLIOTHEK
LEIPZIG
FRANKFURT AM MAIN
Übersicht Volltextsuche
Startseite → Behörden und Institutionen des Bundes → B → Bundeskartellamt
Filter
Bundeskartellamt
09.06.2014 00:03:05
27.03.2014 18:31:49
07.07.2013 00:01:10
07.04.2013 00:03:24
06.01.2013 00:00:51
07.10.2012 00:08:44
10.07.2012 10:08:44
Impressum Hilfe powered by oia
100%


http://dnb.oia-dwa.de/show.aspx
DNB - Webarchive
Datei Bearbeiten Ansicht Favoriten Extras ?
DNB
Archivierte Netzressource vom 07.10.2012
www.bundeskartellamt.de
Datensatz im Katalog
Home | Sitemap | RSS | Suche | Kontakt | Impressum | English | Français | Drucken
Bundeskartellamt
Offene Märkte | Fairer Wettbewerb
Suchbegriff eingeben
Aktuelle Meldungen
Über das Bundeskartellamt
Stellenangebote und Praktika
Presse
Rechtsgrundlagen
Kartellverbot
Missbrauchsaufsicht
Fusionskontrolle
Vergaberecht
Stellungnahmen/Publikationen
Veranstaltungen
Internationale Zusammenarbeit
AGB-Verzeichnis (alt)
Links und Adressen
English
Français
Kontakt
Das Bundeskartellamt
Aktuelle Meldungen
05.10.2012 Aktuelle Entscheidungen der Vergabekammern
05.10.2012 Liste der Neuerwerbungen der Bibliothek
05.10.2012 Aktuelle Entscheidung: Klinikum Worms / Hochstift Worms
02.10.2012 Das Bundeskartellamt stellt ein: studentische Aushilfskraft
02.10.2012 Aktuelle Liste der laufenden Zusammenschlussvorhaben (02.10.2012)
02.10.2012 Aktuelle Liste der Hauptprüfverfahren (02.10.2012)
Pressemeldungen
01.10.2012 Unternehmensverflechtungen auf dem Prüfstand – Bundeskartellamt veröffentlicht Abschlussbericht der Sektoruntersuchung Walzasphalt
27.09.2012 Einleitung der Sektoruntersuchung Raffinerien und Mineralölgroßhandel
21.09.2012 OLG Düsseldorf weist Beschwerde von ConocoPhillips gegen
Weitere Meldungen
Hinweise auf Kartellverstöße
Gemeinsames Energie-Monitoring 2012 der Bundesnetzagentur und des Bundeskartellamtes
Häufige Fragen zum Thema Kraftstoffpreise
16. Internationale Kartellkonferenz
Berlin, 20.-22. März 2013
Impressum Hilfe powered by oia
100%

Stand zu selektivem Harvesting

- Thematische Kategorien: Behörden und Institutionen des Bundes, Interessenverbände, Kultureinrichtungen, Parteien, Politiker, Religionsgemeinschaften, Sozialversicherung, Sportverbände
- Event-Crawls: Z. B. 100. Geburtstag Willy Brandt, Berlinale 2013, Bundestagswahl 2013, Grimme Online Award 2013, Hochwasser 2013, Olympia 2014
- Insgesamt derzeit ca. 700 Sites, geplante Steigerung bis Ende 2015 auf ca. 4000 Sites
- Thematische Ausweitung geplant in Kooperationen mit Aggregatoren wie Academic Linkshare und Projekten wie Alexandria

.de-Domain-Crawl

- Top-Level-Domain-Crawl als Ergänzung zum selektiven
- Großer Umfang: ca. 16 Millionen registrierte Domains für .de, Crawl der BnF für .fr mit 33 TB
- Ausschreibung 2013, Vergabe an französische Firma Internet Memory Research (www.internetmemory.org)
- Experiment über Machbarkeit und Umfang: Maximal 100 TB, Dateien mit maximal 10 MB
- Zugriff per Volltextsuche im Lesesaal
- Durchführung erste Hälfte 2014, bei Erfolg alle 2 Jahre

Vielen Dank!

– Deutsche Nationalbibliothek: www.dnb.de