DEUTSCHE NATIONAL BIBLIOTHEK

Tobias Steinke

Erfahrungsbericht der Deutschen Nationalbibliothek zum ersten Crawl der .de-Domain

DEUTSCHE NATIONAL BIBLIOTHEK

Inhaltsverzeichnis

Sammelauftrag der Deutschen Nationalbibliothek

- 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

Selektives Webharvesting

- Workflow zum regelmäßigen Einsammeln von Momentaufnahmen
- Auswahl von Seiten durch DNB, Erschließung mit Titel und Kategorie
- Sammlung mit eigener Harvester-Software bei Dienstleister 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

DEUTSCHE NATIONAL BIBLIOTHEK

Selektives Webharvesting

- Thematische Kategorien, z. B. Behörden und Institutionen des Bundes, Interessenverbände, Kultureinrichtungen
- Event-Crawls, z. B. Berlinale, Bundestagswahl, Olympia
- Seit 2012: Ca. 1.700 Sites, ca. 8.300 Crawls
- Regelmäßige Crawls zweimal pro Jahr, Events variierend

DE-Crawl: Beauftragung

- Ergänzung zum selektiven Harvesting
- Top-Level-Domain-Crawls seit Jahren üblich in anderen Ländern, jedoch kleinere Größenordnung
- Optionaler Teil der Ausschreibung 2011, keine überzeugenden Angebote
- Erneute separate Ausschreibung 2013: Einmaliger experimenteller Crawl
- Internet Memory Foundation / Research (IMR)

DE-Crawl: Durchführung

- Keine Liste aller registrierten DE-Domains von DENIC
- IMR hat Liste aus früherem Projekt als Ausgangspunkt
- Erster Crawl ohne Speicherung zur Ausweitung der Liste von Start-URLs
- Sammlung mit eigenen Crawlern auf IMR-Servern
- Einschränkungen: Keine Videos, 5.000 Dateien pro Site, 10 MB pro Datei, Gesamtgröße 100 TB
- Berücksichtigung von robots.txt

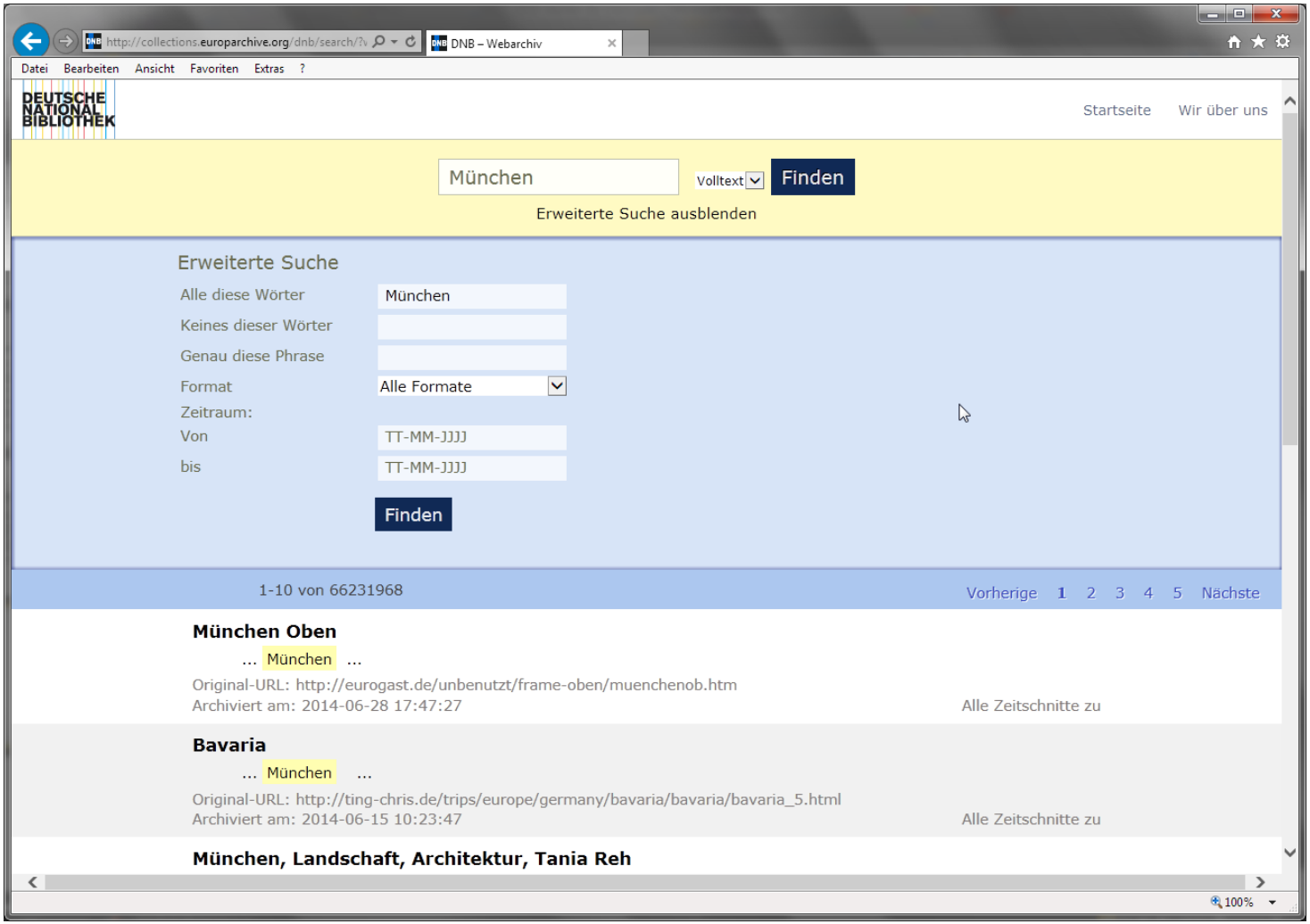
DEUTSCHE NATIONAL BIBLIOTHEK

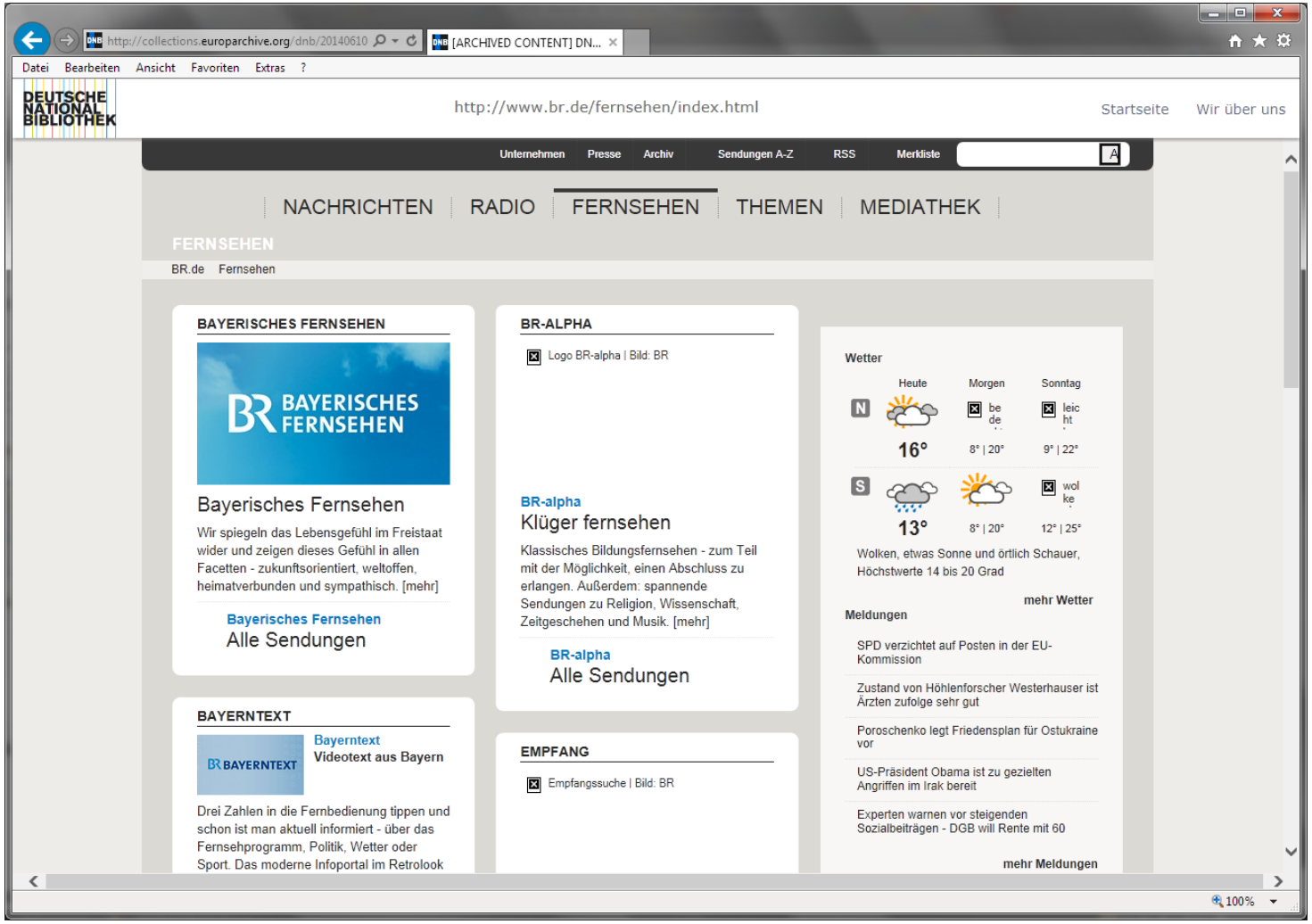
DE-Crawl: Ergebnis

- Crawl im Juni 2014, Dauer 33 Tage
- 2,6 Millionen Start-URLs, 6 Millionen gesammelte Sites, 2,4 Milliarden Dateien
- 120 TB gesammelt (nicht genau 100 TB möglich)
- Bekannte nicht gesammelte Sites: 150 tausend
- Geschätzter Datenumfang ohne Einschränkungen: Mindestens 200 TB
- Abgleich mit Internet Archive: 62% auch dort

DE-Crawl: Bereitstellung

- Eigene Oberfläche von IMR: Per URL und Volltextsuche
- Exklusiver Aufruf aus Lesesälen über Link in Portal und Eintrag im Katalog
- Anpassungen an DNB-Vorgaben und komplette Erneuerung der Oberfläche führte zu Verzögerungen
- Offiziell verfügbar seit 29.02.2016


http://collections.europarchive.org/dnb/search/?v
DNB – Webarchiv
Datei Bearbeiten Ansicht Favoriten Extras ?
DEUTSCHE NATIONAL BIBLIOTHEK
Startseite
Wir über uns
München
Volltext
Finden
Erweiterte Suche ausblenden
Erweiterte Suche
Alle diese Wörter
München
Keines dieser Wörter
Genau diese Phrase
Format
Alle Formate
Zeitraum:
Von
TT-MM-JJJJ
bis
TT-MM-JJJJ
Finden
1-10 von 66231968
Vorherige 1 2 3 4 5 Nächste
München Oben
... München ...
Original-URL: http://eurogast.de/unbenutzt/frame-oben/muenchenob.htm
Archiviert am: 2014-06-28 17:47:27
Alle Zeitschnitte zu
Bavaria
... München ...
Original-URL: http://ting-chris.de/trips/europe/germany/bavaria/bavaria_5.html
Archiviert am: 2014-06-15 10:23:47
Alle Zeitschnitte zu
München, Landschaft, Architektur, Tania Reh
100%


http://collections.europarchive.org/dnb/20140610
[ARCHIVED CONTENT] DN...
Datei Bearbeiten Ansicht Favoriten Extras ?
DEUTSCHE NATIONAL BIBLIOTHEK
http://www.br.de/fernsehen/index.html
Startseite
Wir über uns
Unternehmen
Presse
Archiv
Sendungen A-Z
RSS
Merkliste
NACHRICHTEN
RADIO
FERNSEHEN
THEMEN
MEDIATHEK
FERNSEHEN
BR.de Fernsehen
BAYERISCHES FERNSEHEN
BAYERISCHES FERNSEHEN
Bayerisches Fernsehen
Wir spiegeln das Lebensgefühl im Freistaat wider und zeigen dieses Gefühl in allen Facetten - zukunftsorientiert, weltoffen, heimatverbunden und sympathisch. [mehr]
Bayerisches Fernsehen
Alle Sendungen
BR-ALPHA
Logo BR-alpha | Bild: BR
BR-alpha
Klüger fernsehen
Klassisches Bildungsfernsehen - zum Teil mit der Möglichkeit, einen Abschluss zu erlangen. Außerdem: spannende Sendungen zu Religion, Wissenschaft, Zeitgeschehen und Musik. [mehr]
BR-alpha
Alle Sendungen
BAYERNTEXT
BAYERNTEXT
Bayerntext
Videotext aus Bayern
Drei Zahlen in die Fernbedienung tippen und schon ist man aktuell informiert - über das Fernsehprogramm, Politik, Wetter oder Sport. Das moderne Infoportal im Retrolook
EMPFANG
Empfangssuche | Bild: BR
Wetter
Heute
Morgen
Sonntag
N
16°
8° | 20°
9° | 22°
S
13°
8° | 20°
12° | 25°
Wolken, etwas Sonne und örtlich Schauer, Höchstwerte 14 bis 20 Grad
mehr Wetter
Meldungen
SPD verzichtet auf Posten in der EU-Kommission
Zustand von Höhlenforscher Westerhauser ist Ärzten zufolge sehr gut
Poroschenko legt Friedensplan für Ostukraine vor
US-Präsident Obama ist zu gezielten Angriffen im Irak bereit
Experten warnen vor steigenden Sozialbeiträgen - DGB will Rente mit 60
mehr Meldungen
100%

DE-Crawl: Erkenntnisse

- Sehr gute Quantität trotz einschränkender Vorgaben
- Teilweise enttäuschende Qualität (fehlende CSS)
- Hoher Ressourcenaufwand
- Ständig steigender Umfang erhöht den zu erwartenden Ressourcenaufwand in Zukunft
- Bisher keine Entscheidung und Planung zu weiterem DE-Crawl