Elisabeth Niggemann & Reinhard Altenhöner

Processing the National Mandate: Experiences and Ambitions in DNB

Deutsche Nationalbibliothek

German National Library /DNB

A short history

- 3.10.1912 Deutsche Bücherei in Leipzig
- 1946 Deutsche Bibliothek in Frankfurt am Main
- 1970 Deutsches Musikarchiv (German Music Archive) in Berlin becomes a department of Deutsche Bibliothek
- 3.10.1990 merger of Deutsche Bücherei and Deutsche Bibliothek -> one organization at three sites
- 2006 extended mandate, new name
- 2010 The Music Archive moves to Leipzig -> 2 sites
- Federal institution – Federal Government Commissioner for Culture and Media, State Minister Prof. Monika Grütters


DEUTSCHE
BÜCHEREI
KÖRPER UND STIMME
LEIHT DIE SCHRIFT DEN
STUMMEN GEDANKEN
DURCH DER
ATHMENDEN STROM
TRAGT SIES
DAS REDENDE BLATT
FREIE STATT
FÜR FREIES WORT
FREIER FORSCHUNG
SICHRER PORT
REINER WAHRHEIT
SCHUTZ UND HORT

DEUTSCHE NATIONAL BIBLIOTHEK

DEUTSCHE NATIONAL BIBLIOTHEK

DEUTSCHE NATIONAL BIBLIOTHEK

DEUTSCHE NATIONAL BIBLIOTHEK

DEUTSCHE NATIONAL BIBLIOTHEK

Our Legal Mandate

- Collecting and indexing, archiving; providing permanent access.
- Legal deposit for text, music and pictures published in Germany since 1913.
- No matter which physical carrier: microfilms, sound recordings, CD-ROM, DVD, floppy disks …
- 22.06.2006: Online-publications covered by new law: e-theses, e-journals, e-books, newsletters, digital copies as products of digitisation projects, other electronic publications, AND web sites.
- Access in the reading rooms only - if under copyright.

DEUTSCHE NATIONAL BIBLIOTHEK


INPUT 64
Sofies Welt
AMIGA
Computermalschule
Fantasy
INPUT 64
12/85
COMPUTER-PRAXIS PHYSIK
R. J. Brandenburg
Messen und Auswerten mit dem Computer
Leitende Männer und Frauen der Wirtschaft 1992
ATARI ST
DATENSCHUTZ auf dem ATARI ST

Some Figures and Facts

- Holdings: 30 Mio. media units
  - Among them 1.8 Mio. sound recordings
  - More than 1 Mio. digital publications
- Growth per year via legal deposit: about 800,000 physical publications (3,850 per working day)
- 61,520 current periodicals
- Stacks space: 79,000 qm
- Budget: 48 Mio. Euro
- 735 staff = 590 FTE (120 non-permanent staff)

Digital Preservation at DNB

- 2003 kick-off for the competence network for long-term archiving: nestor.
- 2004 start of an operational dp-system.
- Gradual integration into the library's routine workflow for digital publications.
- (Re-)ingest of earlier harvested web sites.
- Limited ingest of newly harvested web sites.
- Evaluation and exploration projects like LUKII → WARC
- International cooperation within IIPC, standardization, …

The Mandate for a „German“ Web

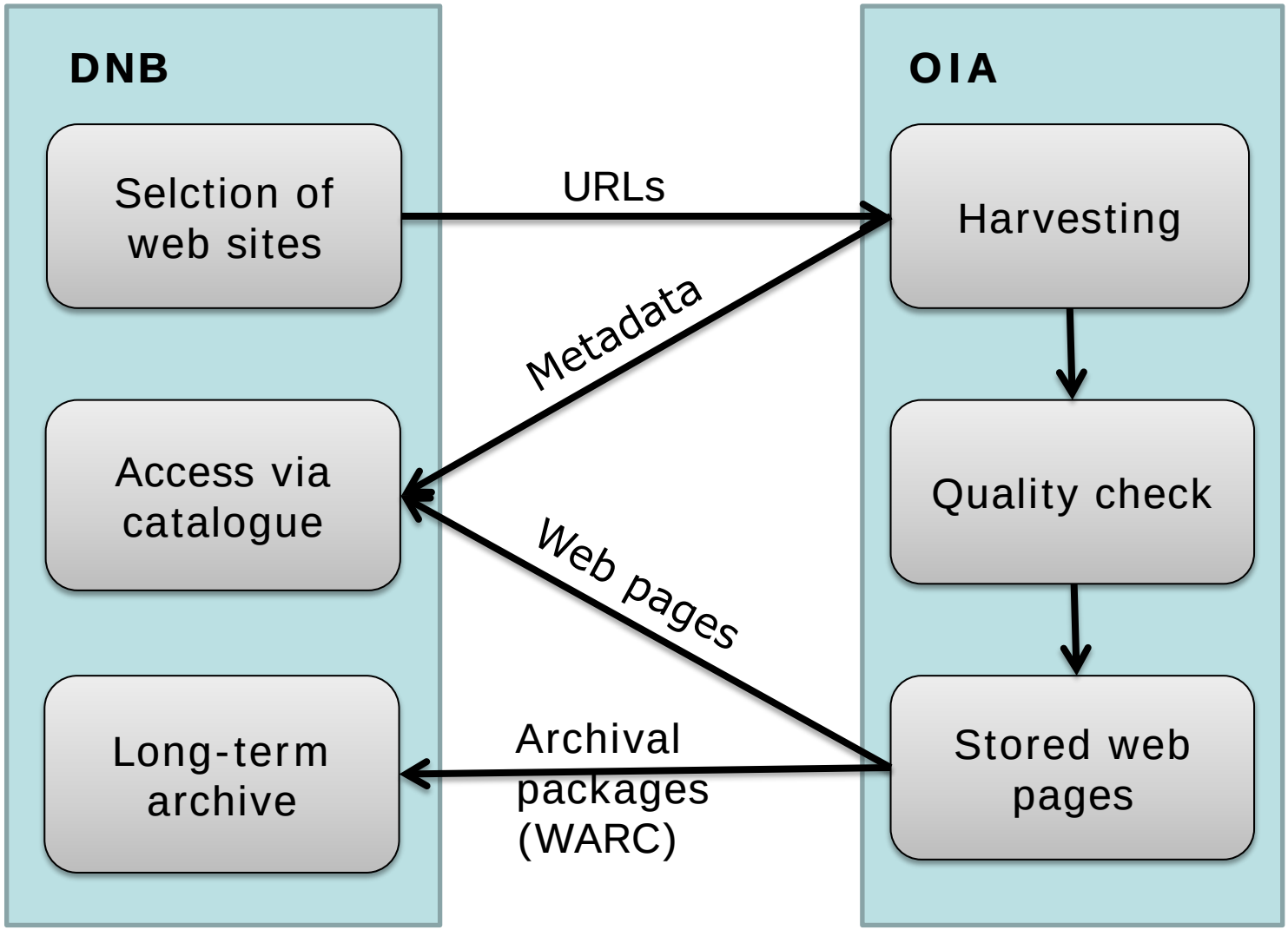
- Cultural and political dimension: The web as an important part of German culture and public life!
- But what is the “German” web?
  - Credit note on a website?
  - German language?
  - Server located in Germany?
  - .de-domain?
- Is “German” web a valid concept for collection purposes anyhow?

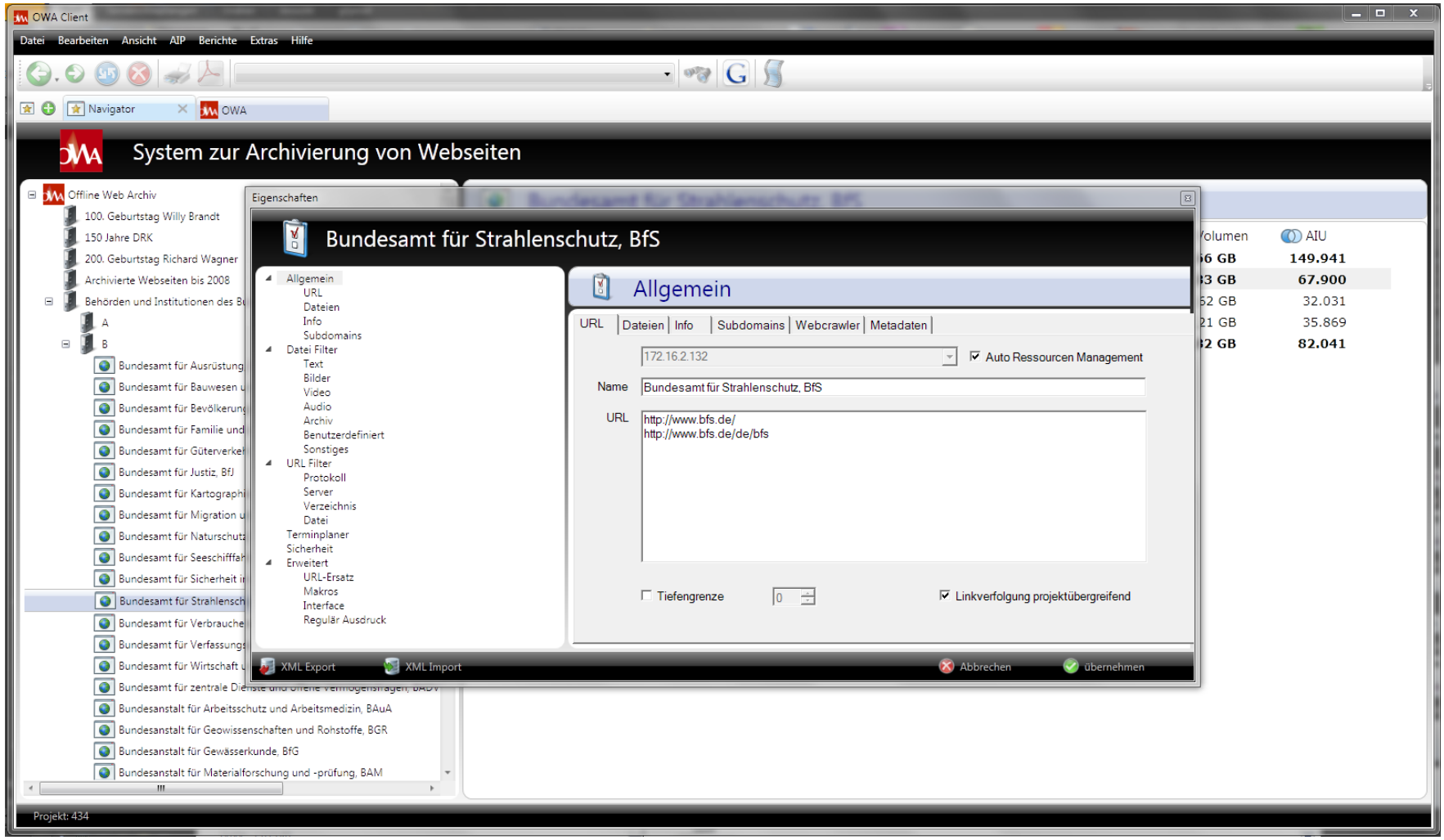
Web Harvesting at DNB

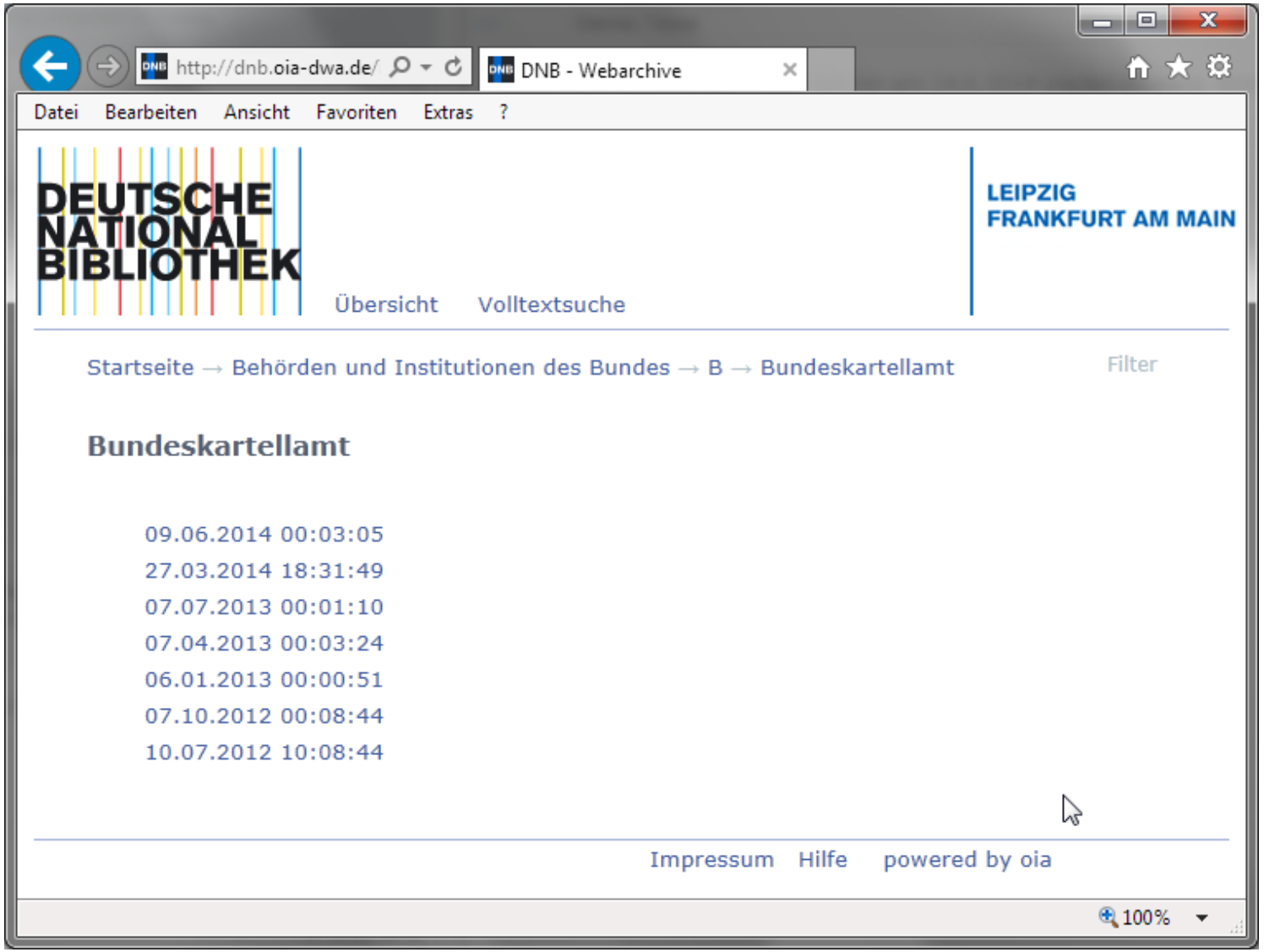
- Event harvesting (e. g. elections) with the European Archive - in close cooperation with the IIPC. Snapshots, samples, approximation as a starting point - not completeness! Challenges: quality, archiving, access.
- 2010: in-house study comparing different approaches, software requirements, costs.
- Strategic decision: Selective approach first (thematic selection and events, quality assurance, dp in place), .de domain harvest later.
- Workflow development together with the German company OIA

Web Harvesting with OIA: Workflow

- DNB uses special tool to select URLs, defining parameters and metadata for the crawl of web sites.
- OIA use their own crawler to harvest web pages, check quality and store data on their own servers.
- Metadata will be automatically integrated into the DNB library catalogue.
- Exclusive access in the reading rooms of DNB via the catalogue and via full text search.
- Interface for long-term preservation in DNB archival system.


DNB
Selction of web sites
Access via catalogue
Long-term archive
URLs
Metadata
Web pages
Archival packages (WARC)
OIA
Harvesting
Quality check
Stored web pages


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%

What Have We Been Collecting?

- Web sites of federal institutions and selected organizations: authorities, institutions, interest groups, cultural institutions, political parties, politicians, religious organizations, social security, sports federations.
- Starting with about 700 sites, increasing gradually to approximately 4,000 sites by the end of 2015.
- Event Crawls (examples): 100th birthday of Willy Brandt, Berlin 2013, federal election 2013, Grimme Online Award 2013, floods in 2013, 2014 Olympics ...

DEUTSCHE NATIONAL BIBLIOTHEK

Plans and Recent Steps

- Thematic expansion planned in cooperation with aggregators such as Academic LinkShare and others.
- Cooperation planned for selecting topic-specific web sites.
- Experimental .de domain crawl with Internet Memory Foundation.
- .de domain: In 2013 European call for tender.

DEUTSCHE NATIONAL BIBLIOTHEK

First Experimental .de-Domain-Crawl

- Partner: Internet Memory Research (www.internetmemory.org)
- Experiment on feasibility and scope: A maximum of 100 TB, with a maximum of 10 MB files
- Wide range: about 16 mio registered domains for .de, (.fr crawl the BnF was 33 TB)
- Results: 130 TB (not complete), 91% .de, 2.6 mio seeds, 2.5 bn resources, some quality concerns
- Data available, full text search still missing, access only in the reading rooms

To Sum Up: 1. Challenges

- Legal restrictions for access and reuse (even for scientific purposes)
- Controversy about legal restrictions even for indexing!
- Privacy
- Quality
- Digital preservation (mass problem)
- Access (is preservation)

2. Needs and Ambitions

– Retrieval:
  - Web Archive as a layered mirror into the past – a relevant source which has to be encoded for science and research
  - Intelligent search: filtering semantics and tailored time-frames, e.g. to examine functions of politicians in a specific time slot
  - Relevance ranking

– Legally:
  - Access control
  - Reuse of data sets

– Setting
  - Common idea, methodology, toolset to take the web as a source (in an historical sense)

DEUTSCHE NATIONAL BIBLIOTHEK

Thank you for your attention!

Questions?