Österreichische Nationalbibliothek

Web@rchiv Österreich

Webarchivierung an der Österreichischen Nationalbibliothek

Michaela Mayr
Österreichische Nationalbibliothek
michaela.mayr@onb.ac.at
www.onb.ac.at

Österreichische Nationalbibliothek

Umfeld

- Massenmedium, nationales **Kulturgut**
- Sammelauftrag Österreichische Nationalbibliothek: **Mediengesetz** (Novelle seit März 2009 in Kraft)

- **Herausforderungen:**
  - Kurze **Lebenszeit** von Internet-Seiten: durchschnittlich 44-75 Tage (Quelle: Library of Congress)
  - **Deep Web**
  - Neue **Technologien**
  - **Viren** etc.
  - **Langzeitarchivierung**: Migration, Emulation?

Österreichische Nationalbibliothek

Datenmengen global

- „The current size of the world's digital content is equivalent to all the information that could be stored on **75bn Apple iPads**, or the amount that would be generated by **everyone in the world** posting messages on the microblogging site **Twitter** constantly for a **century**…."
- 2007: 161.000 PB
  2009: 8 Mio. PB
  2010: 1,2 ZB
- 1 Zettabyte = 1 Mio. Terabytes oder

1,000,000,000,000,000,000,000 Bytes

Quelle: http://www.telegraph.co.uk/technology/news/7675214/Digital-universe-to-smash-zettabyte-barrier-for-first-time.html, IDC Survey, Mai 2010

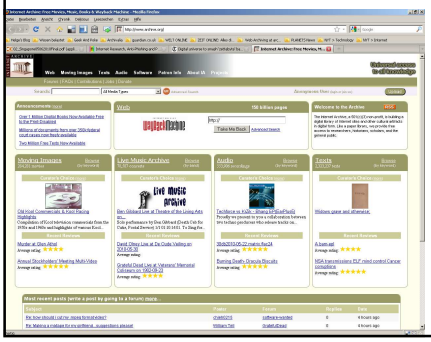
3

Österreichische Nationalbibliothek

Webarchive international (1)

Internet Archive
www.archive.org
USA, seit 1996
Non-Profit Organisation

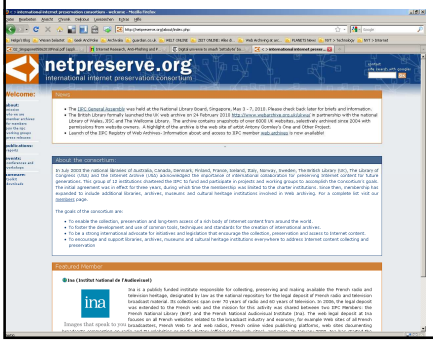
- Derzeit > 4,5 Petabytes Daten
- Zuwachs von 20 Terabytes/Monat
- 150 Milliarden Seiten
- Archiv öffentlich



4

Österreichische Nationalbibliothek

Webarchive international (2)



International Internet Preservation Consortium
netpreserve.org
Gegründet 2003 von 12 Nationalbibliotheken + Internet Archive

- Arbeitsgruppen
- Projekte (Heritrix, WARC)
- ÖNB Mitglied seit 2008
- Kein eigenes Archiv
- Mitglieder Archive
  www.netpreserve.org/about/archiveList.php

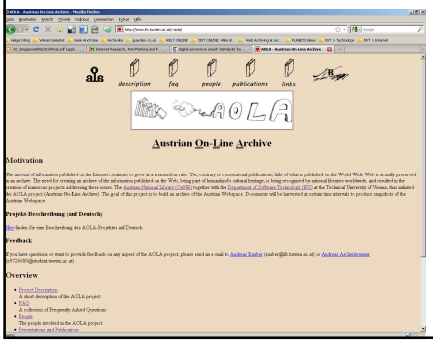
Österreichische Nationalbibliothek

Webarchive international (3)

- Online verfügbar:
  - **European Web Archive**: http://www.europarchive.org/
  - **Pandora**: Start 1996, http://pandora.nla.gov.au/
  - **UK Web Archive**: Start 2005, http://www.webarchive.org.uk/ukwa/
  - **Library of Congress**: Start 2000, http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html
  - **Island**: Start 2004, http://www.landsbokasafn.is/
  - …
- Offline verfügbar:
  - **Netarchive.dk**: Start 2005
  - **Frankreich BnF**: Start 2002
  - ...

Österreichische Nationalbibliothek

Webarchive national



AOLA – Austrian OnLine Archive

www.ifs.tuwien.ac.at/~aola/

2000/2001

- Pilotversuch nationales Webarchiv ÖNB + TU Wien
- Snapshot

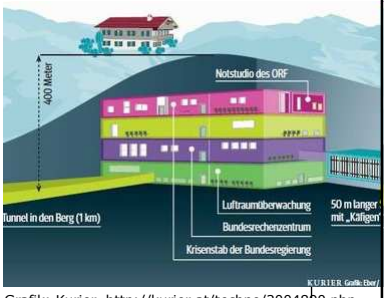
7

Österreichische Nationalbibliothek

Web@rchiv Österreich (1)

- Webarchivierungsprojekt **Start 2008**
- **Mediengesetznovelle** März 2009
- **Team:** 2 VZÄ, Abt. Digitale Bibliothek:
  - Projektmanager
  - Entwickler / Crawl Engineer
  - System Administrator

- **Speicher** und **Back-Up** ausgelagert an Bundesrechenzentrum (+ Kopie ZAS St. Johann)



Grafik: Kurier, http://kurier.at/techno/2004890.php

Österreichische Nationalbibliothek

Web@rchiv Österreich (2)

- **Software** (nur open source)
  - Crawler **Heritrix**
  - Crawl Management mit **NetarchiveSuite** (http://netarchive.dk, Kooperation mit Dänemark, Frankreich)
  - Zugang mit **Wayback Machine**
- **Hardware**
  - 8 Maschinen:
    6 Crawler (mit je 3 Crawlerinstanzen)
    1 für Datentransfer BRZ
    1 DB und Indexierung
  - Betriebssystem Linux

9

Österreichische Nationalbibliothek

Web@rchiv Österreich (3) Zugang

- Nur am Standort der Bibliotheken, **nicht online** (spezielle Terminals)
- Nur Ausdruck, kein Speichern oder Versenden
- Passwortgeschützte Seiten nur Einzeluser

- Berechtigte Bibliotheken
  - Bundeskanzleramt, Parlament
  - Österreichisches Staatsarchiv
  - Universitäts-, Studien- und Landesbibliotheken

10

Österreichische Nationalbibliothek

Sammlungsstrategien (1)

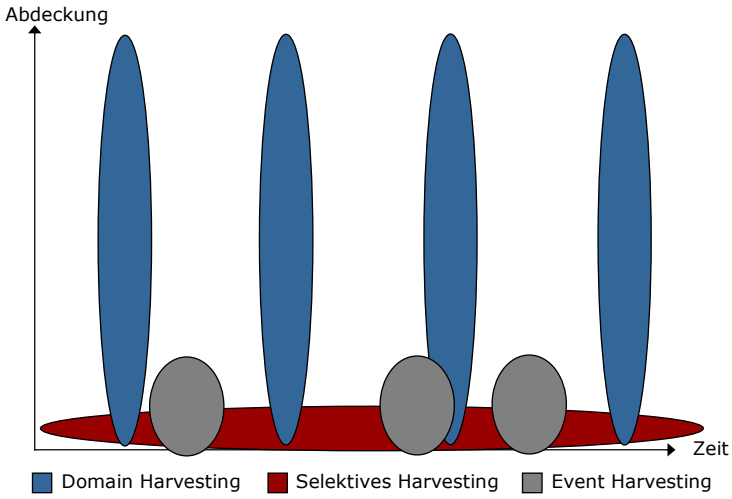
- **Gesamtes Web** (IA)
- **Nationale Domäne** (DK, F, AT)
- **Selektiv/Thematisch** (GB, LOC)

→ International unterschiedliche Ansätze, je nach Gesetzeslage und Ressourcen

11

Österreichische Nationalbibliothek

Sammlungsstrategien (2)



12

Vgl. Bjarne Andersen, http://netarchive.dk/publikationer/DFrevy_english.pdf

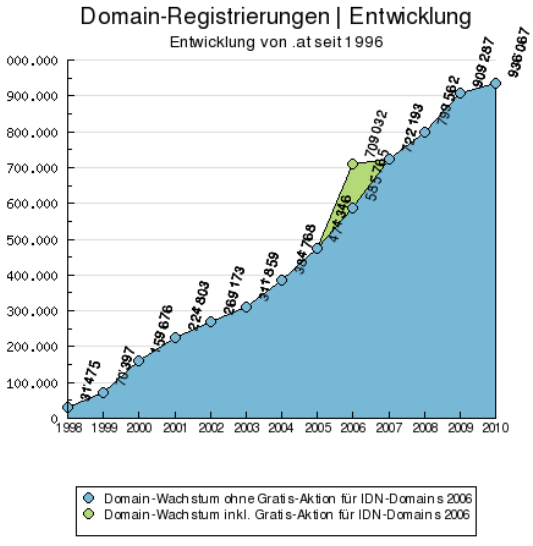
Österreichische Nationalbibliothek

Sammlungsstrategien (3)

- **Domain Harvesting**
  - Gesamte **Top-Level-Domain .at** (Stand Juni 2010: ca. 936.000 Domains, Quelle: nic.at)
  - andere Top-Level-Domains mit **Österreich-Bezug** (keine Definition im Gesetz, manueller Aufwand)
  - Durchführung **alle 2 Jahre**
  - Keine Einhaltung **robots.txt**

Österreichische Nationalbibliothek

Entwicklung .at Domain



Quelle: nic.at

Österreichische Nationalbibliothek

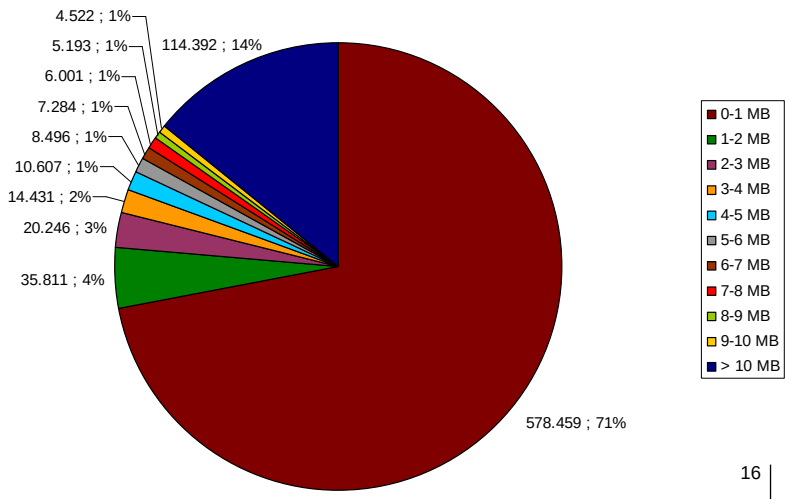
Domain Crawl, 1. Durchlauf

- Beginn: 16.09.2009
- Ende: 10.01.2010
- 895.445 Domains
- Max. 10 MB pro Domain
- Physischer Speicher: 1,57 TB
- Rückmeldungen Seitenbetreiber: 3
- 2. Durchlauf läuft, max. 100 MB pro Domain, Ende in Kürze

15

Österreichische Nationalbibliothek

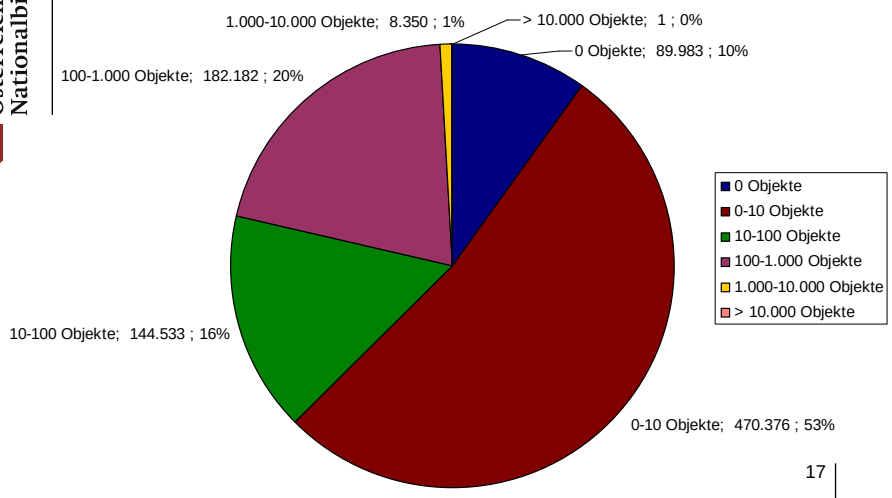
Domain Crawl Stage 1/bis 10MB: Speicherbedarf je Domain



16

Österreichische Nationalbibliothek

Domain Crawl Stage 1/bis 10MB: Objekte je Domain



Österreichische Nationalbibliothek

Sammlungsstrategien (4)

- **Selektives Harvesting**
  - Ausgewählte Seiten, die häufigen Änderungen unterliegen
  - Harvesting in geeigneten Intervallen
  - Inhalte:
    - Medien national und regional,
    - dynamische Seiten aus den Bereichen Gesellschaft, Wirtschaft, Kultur, Verwaltung/Behörden,
    - Wissenschaft/Universitäten sowie
    - experimentelle und/oder einzigartige Webseiten zur Dokumentation von neuen Techniken (z.B. net art).
  - Start 2010

Österreichische Nationalbibliothek

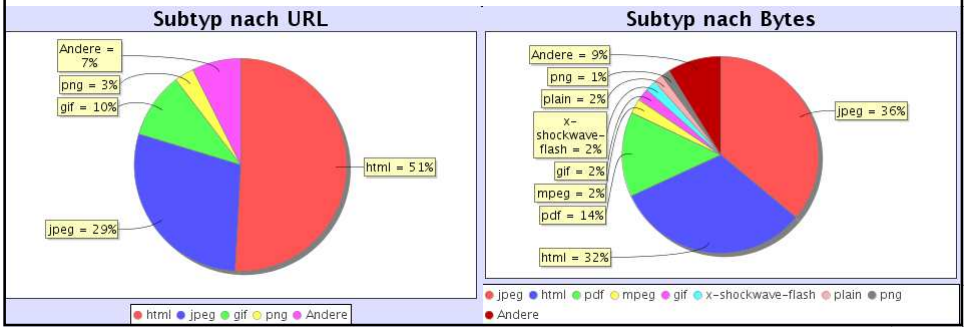
Sammlungsstrategien (5)

- **Event Harvesting**
  - Spezielle Anlässe und Großereignisse (z.B. Wahlen)
  - Webseiten, die nur für den Zeitraum des Ereignisses zur Verfügung stehen
  - Bisherige Event Harvestings:
    - EURO™ 2008
    - Nationalratswahl 2008
    - EU-Wahl 2009
    - Olympische Spiele 2010
    - Bundespräsidentenwahl 2010

Österreichische Nationalbibliothek

Web@rchiv Österreich Statistik

- Aktuell ca. 5 TB Daten (komprimiert und dedupliziert)
- Entspricht ca. 7,5 TB Rohdaten
- 300 GB Metadaten
- 341 Mio. Objekte

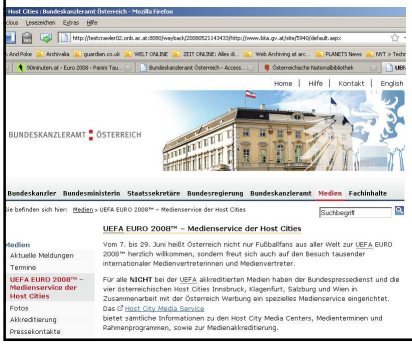


Österreichische Nationalbibliothek

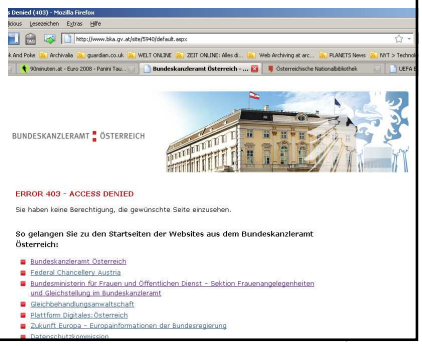
Web@rchiv Österreich Demo

-
-

ARCHIV



LIVE WEB



Österreichische Nationalbibliothek

Vielen Dank für die Aufmerksamkeit!

Weitere Infos:
http://www.onb.ac.at/about/webarchivierung.htm

Webseiten nominieren:
http://www.onb.ac.at/about/seiten_nominieren.htm

Follow us:
http://twitter.com/AT_WebArchive