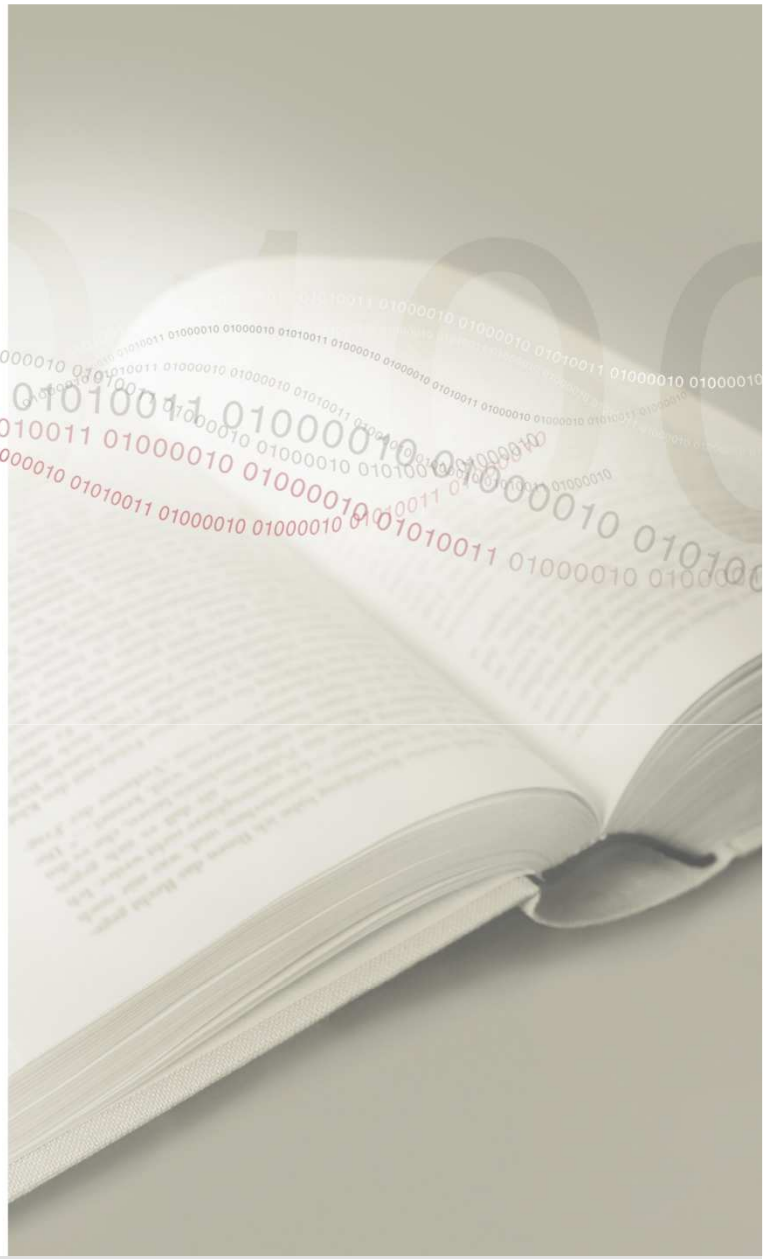
BSB Bayerische StaatsBibliothek
Information in erster Linie

MDZ Münchener DigitalisierungsZentrum Digitale Bibliothek



Heute ernten für die Wissenschaft von morgen

Sammlung und Archivierung von Websites als bibliothekarische Dienstleistung

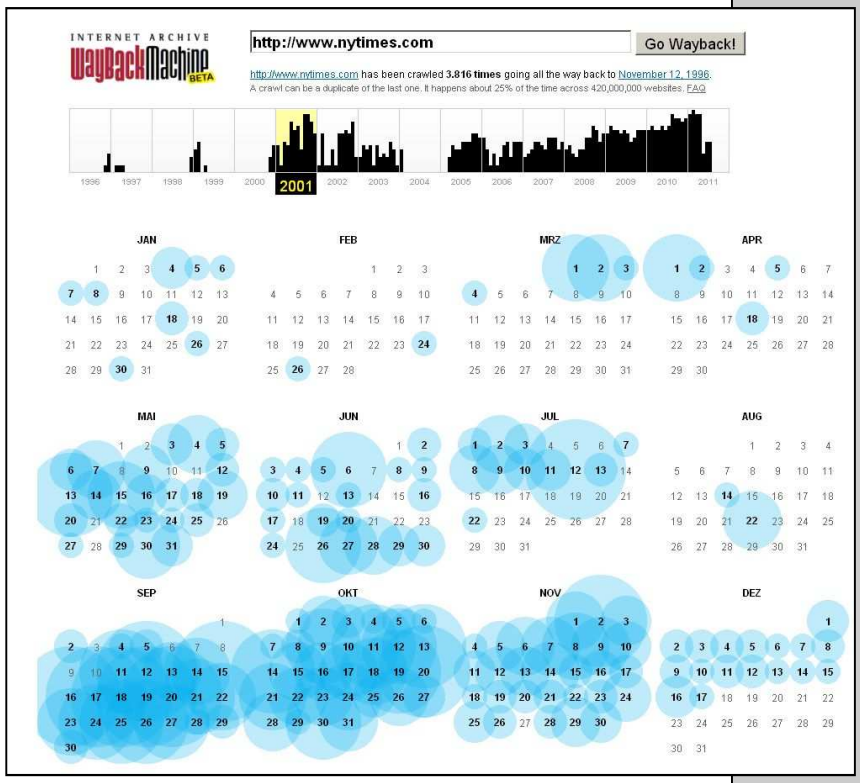
Tobias Beinert, Bayerische Staatsbibliothek

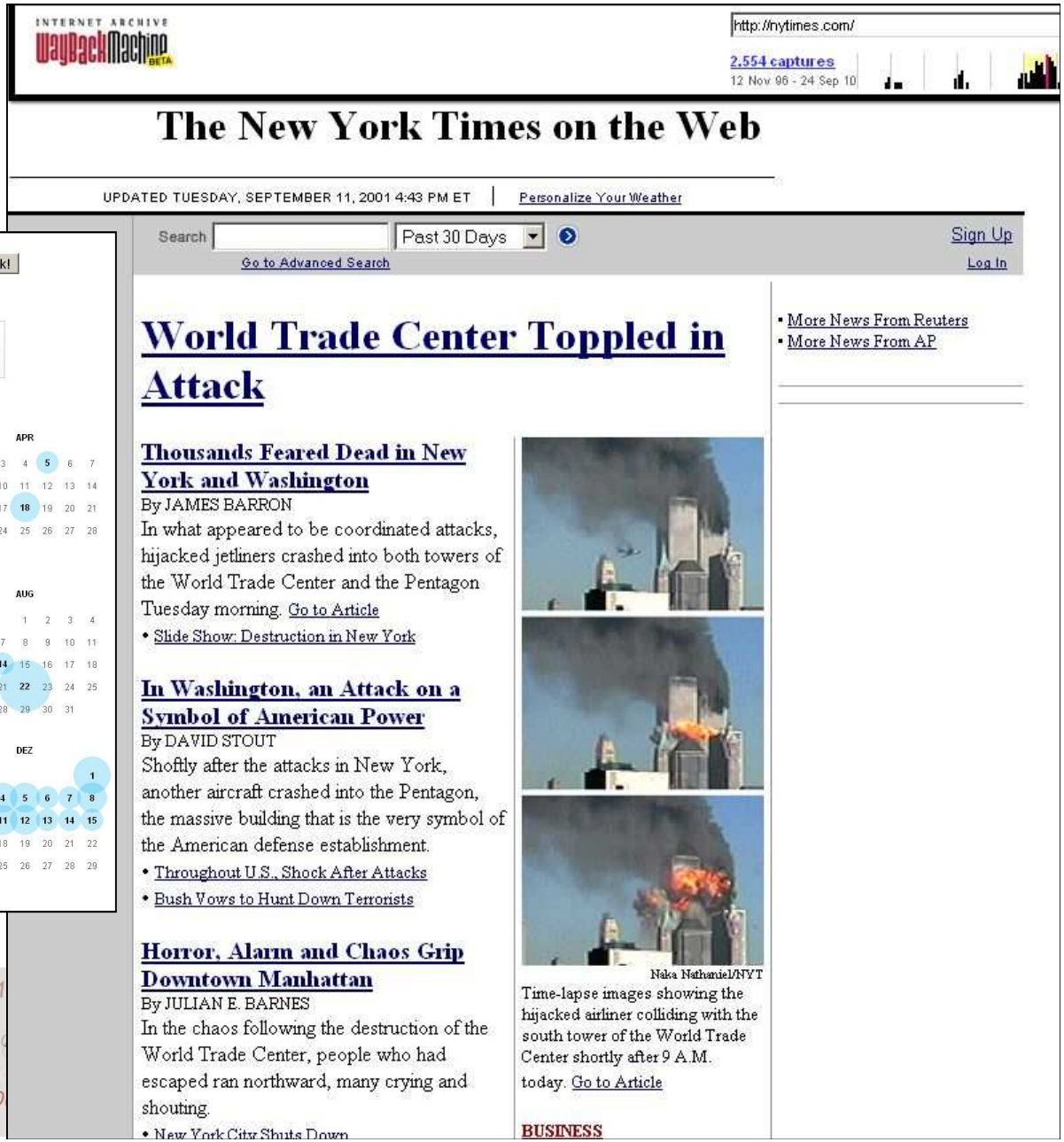
101. Deutscher Bibliothekartag
Hamburg, 24.05.2012

Übersicht

- Ziele und Grundbegriffe der Webarchivierung
- Übersicht: Stand der Webarchivierung in deutschen Bibliotheken
- Webarchivierung an der Bayerischen Staatsbibliothek
- Möglichkeiten und Grenzen in der Praxis
- Entwicklungsperspektiven / Ausblick / Nutzungsszenarien

Website der NY Times vom 11. Sept. 2001





Gründe für die Webarchivierung

- Hohe Flüchtigkeit der Inhalte des Webs als Publikations-, Informations- und Kommunikationsmedium für die Wissenschaft
- Das Internet als Ausdruck der Gegenwartskultur
- Webarchive sind eine wichtige Quelle für Forschungsliteratur und Gegenstand kulturgeschichtlicher Forschung

⇨ Auffindbarkeit der Inhalte und Referenzierbarkeit bzw. Zitierbarkeit des Webs muss für die Wissenschaft dauerhaft ermöglicht werden

⇨ Es können immer nur Ausschnitte bzw. Schnappschüsse gesammelt werden: Ein umfassendes Bild ergibt sich erst in der Gesamtschau der unterschiedlichsten Sammlungen von Websites

Grundbegriffe der Webarchivierung

- Webarchivierung umfasst die Auswahl, Sammlung, Langzeitarchivierung, Bereitstellung von Websites bzw. Webressourcen

- Website = Webpräsenz, die aus mehreren Webpages bzw. Dokumenten bestehen kann, verknüpft durch eine logische Linkstruktur
  ⇨ Webarchivierung meint in der Regel “Website-Archivierung”

- Harvesting / Crawling: Verfolgen der (internen) Links und lokale Speicherung der einzelnen Dateien sowie Überführung in eine Struktur bzw. Format, die den Gesamtkontext einer Website erhält
  - Domain / selektives / eventbezogenes Harvesting
  ⇨ Entstehung von unterschiedlichen Snapshots

Webarchivierung an der DNB

- Gesetzliche Grundlage: DNB-Gesetz
- Erste Tests 2005/2007 und Mitgliedschaft im International Internet Preservation Consortium (IIPC)
- Konzeption und Ausschreibung des Verfahrens mittlerweile abgeschlossen -> derzeit: Umsetzung mit externem Dienstleister
- Auswahlkriterien: Websites von Bundesbehörden, Verbänden, gemeinnützigen Organisationen, Parteien, Kirchen / Ereignisse mit Bedeutung für Gesellschaft, Geschichte, Politik, Wirtschaft oder Kultur Deutschlands / thematische Sammlungslisten, Einzelmeldungen -> modular gestalteter Sammlungsaufbau mit zunächst ca. 1050 Websites
- Öffentliche Zugänglichmachung: Lesesaal-Bereitstellung

Webarchivierung mit edoweb

- Betreiber seit 2003: Landesbibliothekszentrum Rheinland-Pfalz und hbz
- Gesetzliche Grundlage: Verwaltungsvorschrift für Behörden/Dienststellen + Rechteinholung für Angebote Dritter
- Lösung mit HTTrack und DigiTool inkl. Volltextsuche
- Auswahlkriterien: Websites von Landesministerien, Landesbehörden, der Kommunen sowie in Auswahl Websites wichtiger Einrichtungen des Landes und öffentlich zugängliche, privat betriebene Websites mit landeskundlichen Inhalten (ca. 500-700 Websites)
- Zugriff: Frei im Web und Lesesaal-Bereitstellung

Webarchivierungsservice SWBContent

- Anbieter seit 2002: Bibliotheksservicezentrum Baden-Württemberg
- Zahlreiche lokale Instanzen für Bibliotheken und Archive (u.a. BOA – Baden-Württembergisches Online Archiv und SaarDok)
- Technische Lösung jüngst mit Heritrix und Wayback auf den internationalen Stand gebracht
- Auswahlkriterien:
  - BOA - Websites der Landesverwaltung und mit landeskundlichen Inhalten
  - SaarDok: Institutionen des Saarlands und landeskundliche Aspekte
  - je ca. 120 Websites
- Zugriff: Frei im Web

Übersicht: Webarchivierung in dt. Bibliotheken

- Nur an wenigen Institutionen fest etablierte Geschäftsgänge für die Webarchivierung

- Aktivitäten von nestor erfreuen sich eines großen Interesses

- Fleckenteppich mit vielen regionalen und thematischen Löchern
  -> Verantwortlichkeiten der Bibliotheken sind untereinander aber auch in Abgrenzung zu anderen Einrichtungen (z.B. Archiven) bislang nicht abgestimmt

- Anpassung an internationalen Entwicklungen erfolgen derzeit

- Nutzen für die Wissenschaft bislang nur partiell bzw. punktuell gegeben

-> im internationalen Vergleich ist ein Entwicklungsrückstand zu konstatieren!

Webarchivierung an der BSB

- Selektives Harvesting für die Virtuellen Fachbibliotheken

-> Sicherung der Nachhaltigkeit der Internetressourcen-Erschließung

- Harvesting der Websites der Bayerischen Staatsministerien und der Staatlichen Bibliotheken

- Halbjährliche Snapshots, derzeit ca. 150 archivierte Websites

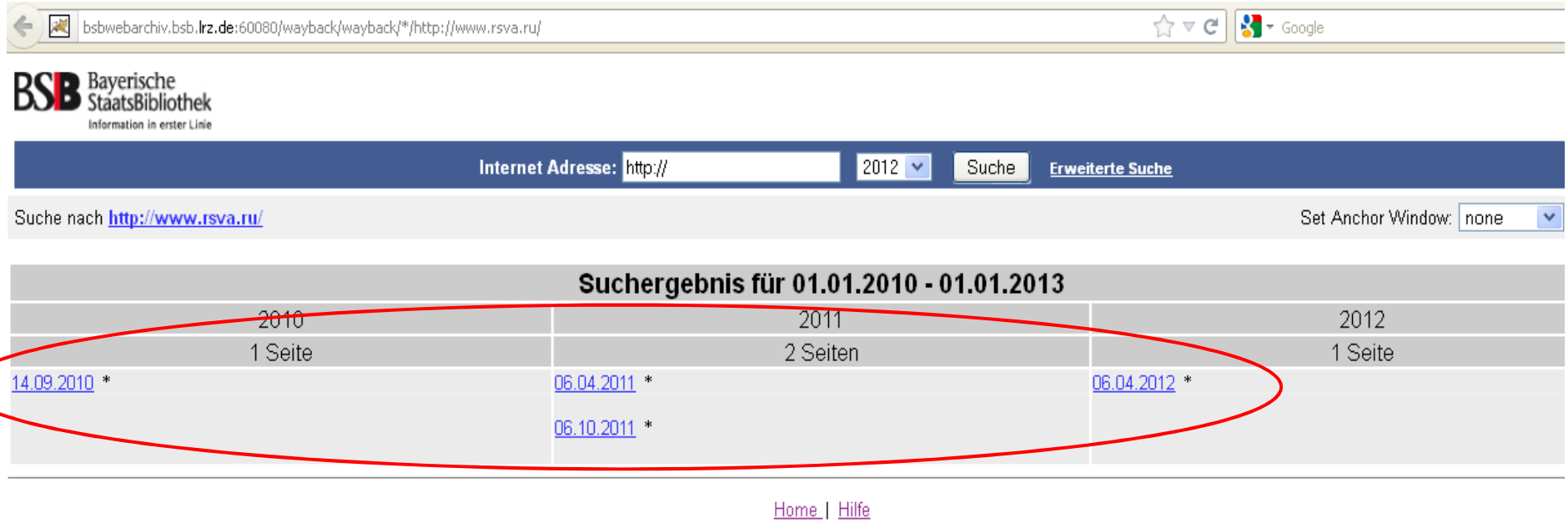
- Genehmigungsverfahren (inkl. Rechteeinholung für Langzeitarchivierung und öffentliche Bereitstellung)

- Manuelle Qualitätskontrolle

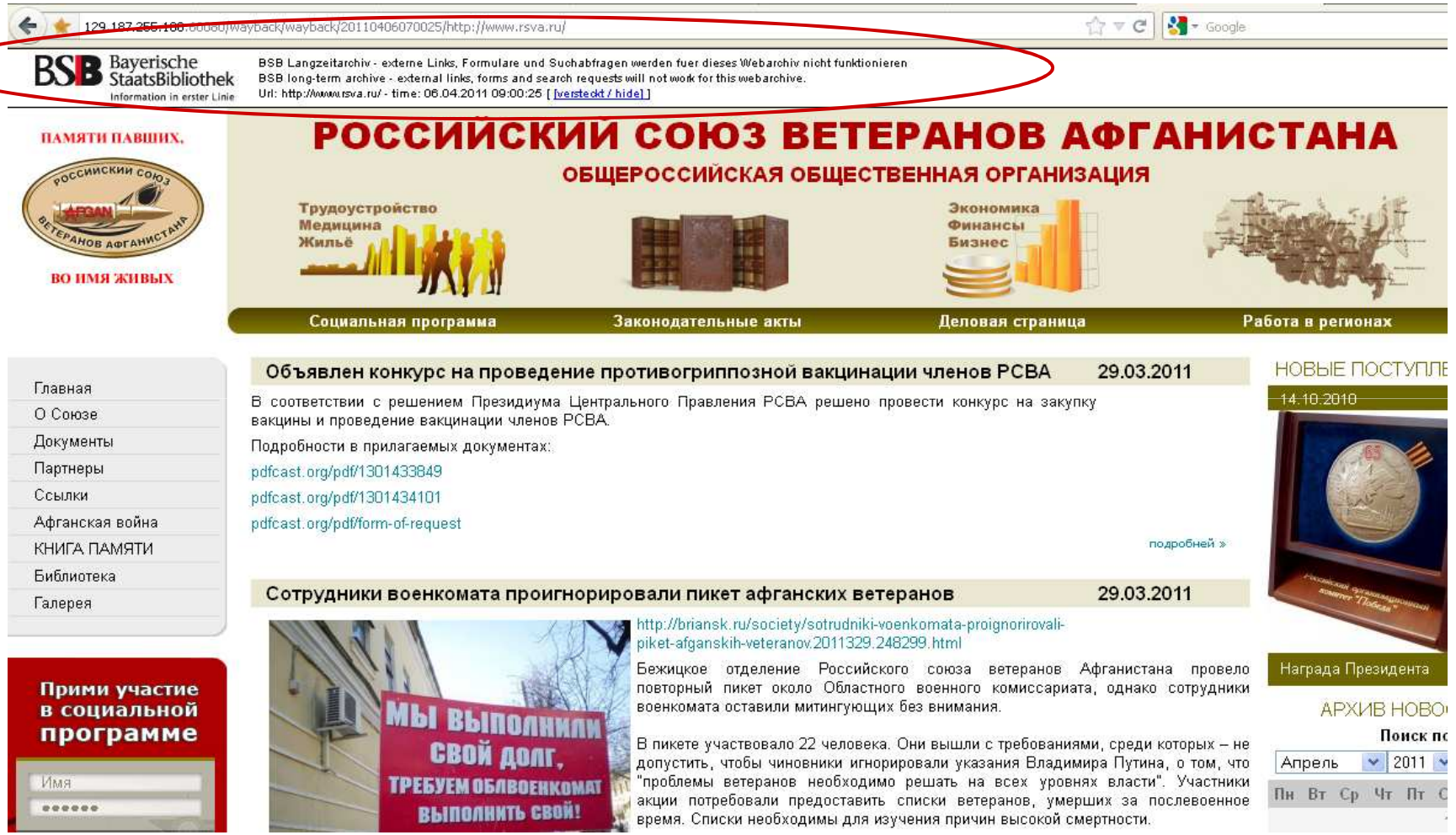
- Zugänglichmachung erfolgt über den Katalog und die Fachportale der ViFas, Darstellung mit der Wayback-Machine

Darstellung im BSB-OPAC

Übersicht in der Wayback Machine



Ansicht des Archivobjekts



Das Web Curator Tool an der BSB

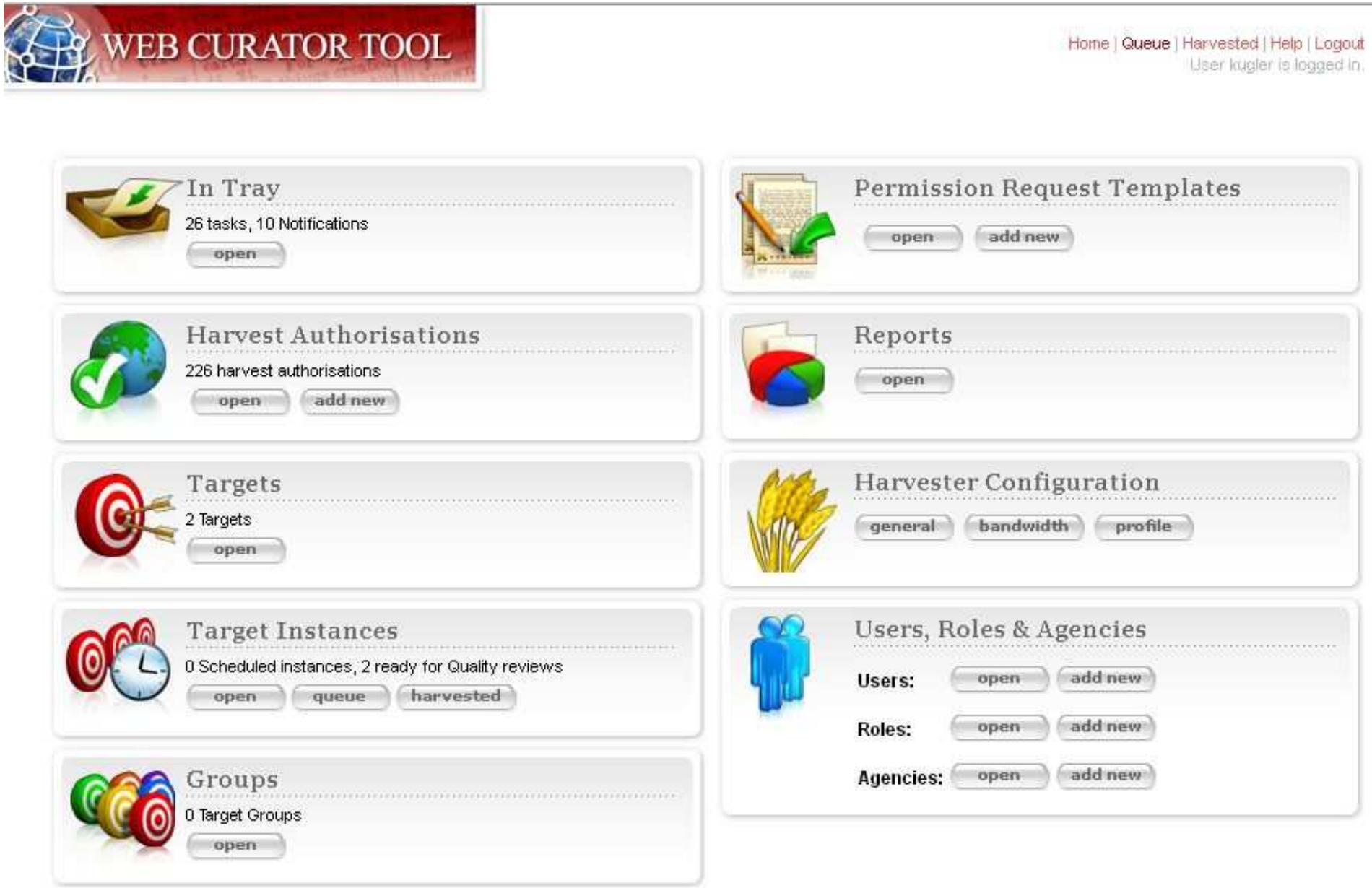
MDZ Münchener DigitalisierungsZentrum Digitale Bibliothek

- Open Source-Tool mit Heritrix-Crawler und Wayback Machine
- Optimiert für selektives Harvesting
- Integrierte Prozesse:
  - Genehmigungsbeantragung und –verwaltung
  - Auswahl und Scheduling von Harvest-Jobs
  - Harvesting
  - Qualitätskontrolle
  - Archivierung im WARC-Format
  - automatische Generierung deskriptiver Metadaten
- Anbindung an Discovery-Systeme der BSB und das Bibliothekarische Archivierungs- und Bereitstellungssystem (BABS)

BSB Bayerische StaatsBibliothek Information in erster Linie

Module des Web Curator Tool

MDZ Münchener DigitalisierungsZentrum Digitale Bibliothek



Herausforderungen der Praxis

- Technische Einschränkungen: Flash, Javascript, Stream Videos, Datenbanken, dynamische Inhalte nur schwer archivierbar

-> Oftmals keine vollständige Archivierung möglich/gewünscht
-> Wesentliche Eigenschaften definieren, gewisse Verluste akzeptieren!

- Fragen des Urheberrechts

- Auswahl des Archivmaterials

- Derzeit noch weitgehend manuelle Prozesse

- Abstimmung der Verantwortlichkeiten / Organisation

Perspektiven für deutsche Bibliotheken

- Bildung von thematischen Kollektionen / größeren Datensets
- Langzeitarchivierung und Langzeitverfügbarkeit
- Bessere Vernetzung von Praxis in Gedächtnisorganisationen und Forschung in der Informatik
- Fortführung und Ausbau von kooperativen Ansätzen
- Urheberrecht mitgestalten

Ausweitung der Nutzungsmöglichkeiten

- Erforschung und Darstellung der Nutzungsmöglichkeiten von Webarchiven
  - Volltextindexierung
  - Auswertung von Textcorpora mit Ngram
  - -> Internet Content als Forschungdaten

- Zielsetzung sollte ein zentraler Nutzerzugang sein

- Verbesserung der Integration in das Live-Web
  - Memento
  - Redirection Services

Fazit

- Flickenteppich mit mehreren Akteuren
- Langsam entstehendes Problembewusstsein
- 2012: Entwicklungsrückstand im internationalen Vergleich
- Hohes Entwicklungs- und Zukunftspotential für Bibliotheken und Archive
- Steigende Komplexität der Daten und deren Vernetzung sowie vielfältige neue Publikationsformen halten die Aufgabe auch in Zukunft spannend
- Es gilt nach wie vor:

Collect now - Ask later why!


Fragen?
beinert@bsb-muenchen.de