Webarchivierung im Auftrag als Dienstleistung des BSZ

Renate Hannemann
Bibliotheksservice-Zentrum Baden-Württemberg (BSZ)

2004-2016 – die Geschichte

- seit 2004 Dienstleistung für Webharvesting mit Eigenentwicklung (HTTrack, Heritrix)
- u.a. 3 Landesbibliotheken, Landesarchiv BW, kommunale und Kreisarchive
- selektives Harvesting (institutionelle Websites, Events)
- Fokus: Vollständigkeit, Spiegelungstiefe, Authentizität
- Archivkopien weitgehend öffentlich zugänglich
- Servicemodell: dezentrales Harvesting durch Archive

2016 – der Umbruch

- Ablösung erforderlich
- Evaluation Fremdsysteme (u.a. WCT, NAS, edoweb, UK Webarchive)
- „harte“ Kriterien:
  - WARC-Standard
  - Heritrix3
  - technische Entwicklung und Fortbestand gesichert
  - Nutzungskomfort
  - (HTTrack-Migration)
- Auswahl: Archive-It (AIT)
- Servicemodell: zentrales Harvesting durch BSZ

2017 – neue Wege

- zentraler Crawlingservice durch BSZ unter Nutzung von AIT
- vollständige Übernahme/Migration aller vorhandenen Spiegelungen
- Dienstleistung „SWBregio“ für rd. 20 Kommunal- und Kreisarchive sowie die Saarländische Universitäts- und Landesbibliothek Saarbrücken
- Auftragscrawling von derzeit lfd. 550 Websites
- Preismodell: elastisch, nach Nutzungsintensität
- Archivkopien öffentlich oder eingeschränkt zugänglich

Aufgaben – im Archiv

- allg. Festlegungen (Sammlung bei AIT, Metadaten)
- Auswahl Webauftritte, Einholung Spiegelungserlaubnisse, Festlegung Intervalle und Zugänglichkeit der Archivkopien
- Beauftragung des BSZ
- Abnahme vorgeprüfter Testcrawls

Aufgaben – im BSZ

- Pflege der Sammlungen, Seeds und Metadaten in AIT
- Analyse der zu spiegelnden Sites, Parametrisierung des Crawlers
- iterative Testcrawls, Analyse, QA, Vor-Abnahme
- Scheduling und Monitoring regelmäßiger Crawls
- Überwachung laufender Websites (Relaunches etc.)
- QA aller produktiven Crawls
- Download WARC ins LSDF / Karlsruher Institut für Technologie (KIT)
- Dokumentation der Spiegelungen

Erfahrungen - dezentrales Crawling

- individuelle Schulungen und Dokumentationen erforderlich
- hohes Kommunikationsaufkommen
- erheblicher Supportaufwand - vermeidbare Problematiken:
  - zeitintensive Fehleranalyse problematischer Crawls aufgrund fehlerhafter Parametrisierung
  - aufwendige Nachjustierung und nachträgliche Ertüchtigung von Crawls
- inhomogene Qualität der Ergebnisse
- personelle Aufwände in den Archiven

Erfahrung - zentrales Crawling

- deutlicher Rückgang Support-Aufwand im BSZ
- Schulungen, Dokumentationen entfallen
- Service aus einer Hand:
  - Kompetenz konzentriert im BSZ
  - einheitlicher Workflow, standardisierte Abläufe
  - einheitliche Methoden der Parametrisierung, große Homogenität
  - durchgehende Qualitätskontrolle aller Crawls
  - gute Qualität der Spiegelungsergebnisse
- reduzierte Aufwände in den Archiven

Fazit

- zentrale Organisation der Dienstleistung hat unsere Erwartungen erfüllt
- personelle Ressourcen im BSZ zielgerichtet und effizient eingesetzt
- Aufwände in den Archiven reduziert
- gesteigerte Qualität der Spiegelungsergebnisse

Vielen Dank für Ihre Aufmerksamkeit!

Fragen gerne an

renate.hannemann@bsz-bw.de