Tobias Steinke

Impuls 2: Webarchivierung an der DNB

Netzpublikationen

- E-Books, E-Journals, digitale Hochschulschriften
- Ablieferung der Verlage und Hochschulbibliotheken an die DNB
- Schnittstellen: Webformular, Hotfolder, OAI-PHM
- Bibliografische Metadaten als Teil der Ablieferung
- Dateiformate: Vorwiegend PDF und EPUB

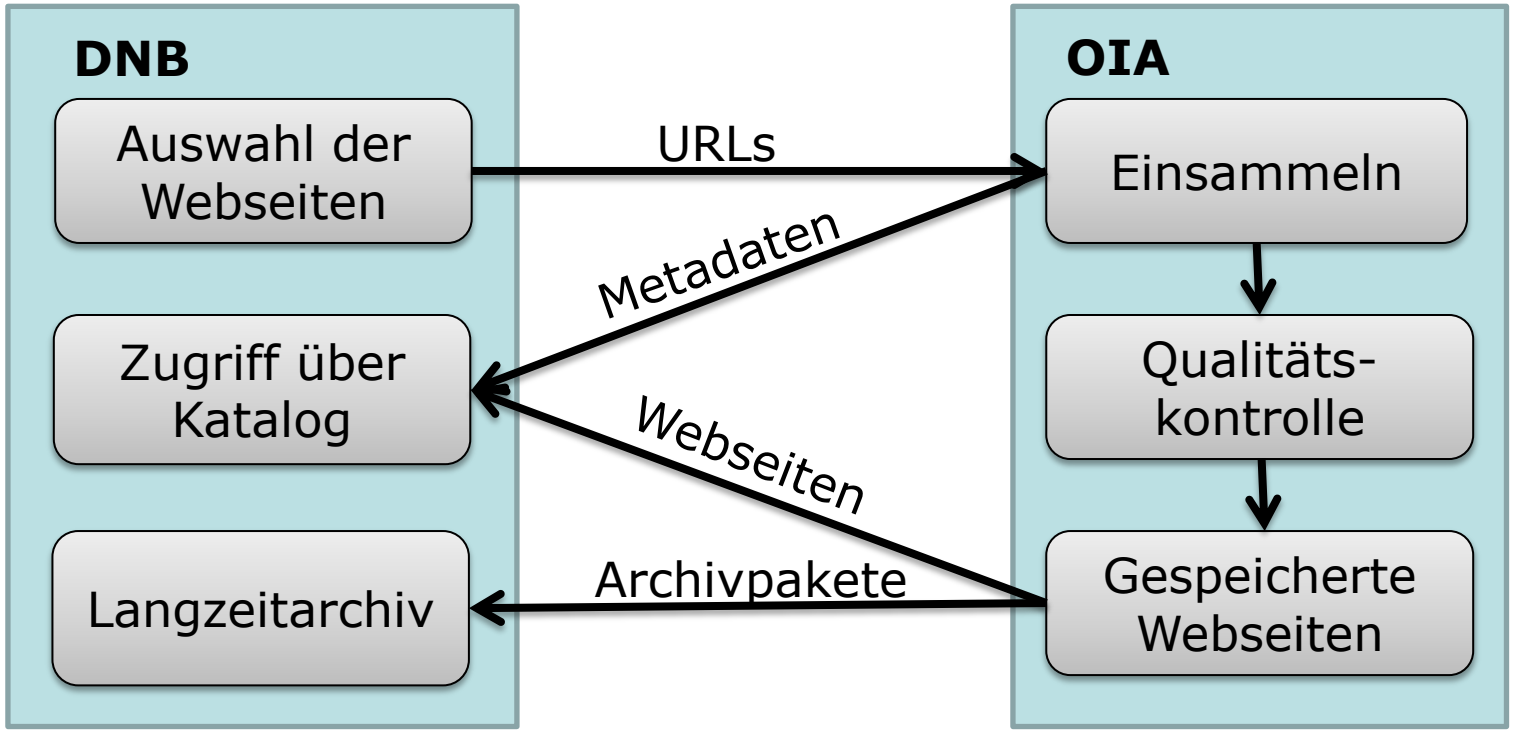
Webseiten

- Wiederholtes Webharvesting zum Einsammeln
- Nur Momentaufnahmen, keine vollständige Sammlung
- Harvesting
  - Speicherung der Seiten mit Daten wie Browser
  - Verfolgen von erkannten Links
  - Interaktion kaum möglich
  - Streaming-Inhalte in der Regel nicht möglich

Selektives Webharvesting

- Seit 2012 Workflow mit Dienstleister oia GmbH
- Auswahl von DNB, Erfassung in Tool OWA von oia
- Sammlung, Qualitätssicherung und Speicherung für Zugriff durch den Dienstleister mit eigener Software auf deren Servern in Düsseldorf
- Zugriff in Lesesälen über den Katalog, eigenes Portal und Volltextsuche

Selektives Webharvesting: Workflow



Langzeitarchivierung

- Speicherung im ISO-Standard WARC
  - Paketformat aller unveränderten Inhalte und Harvesting-Info
- Bereitstellung über geeignete Software (z. B. Wayback)
  - Indexierung der Inhalte (URL, Volltextsuche)
  - Relativierung der Links bei Zugriff
  - Setzt vorhandenen Webbrowser voraus
- Emulation nötig für obsolete Formate (Plugins, Browser)