Tobias Steinke

Selective Web Archiving at the German National Library

Digital Publications

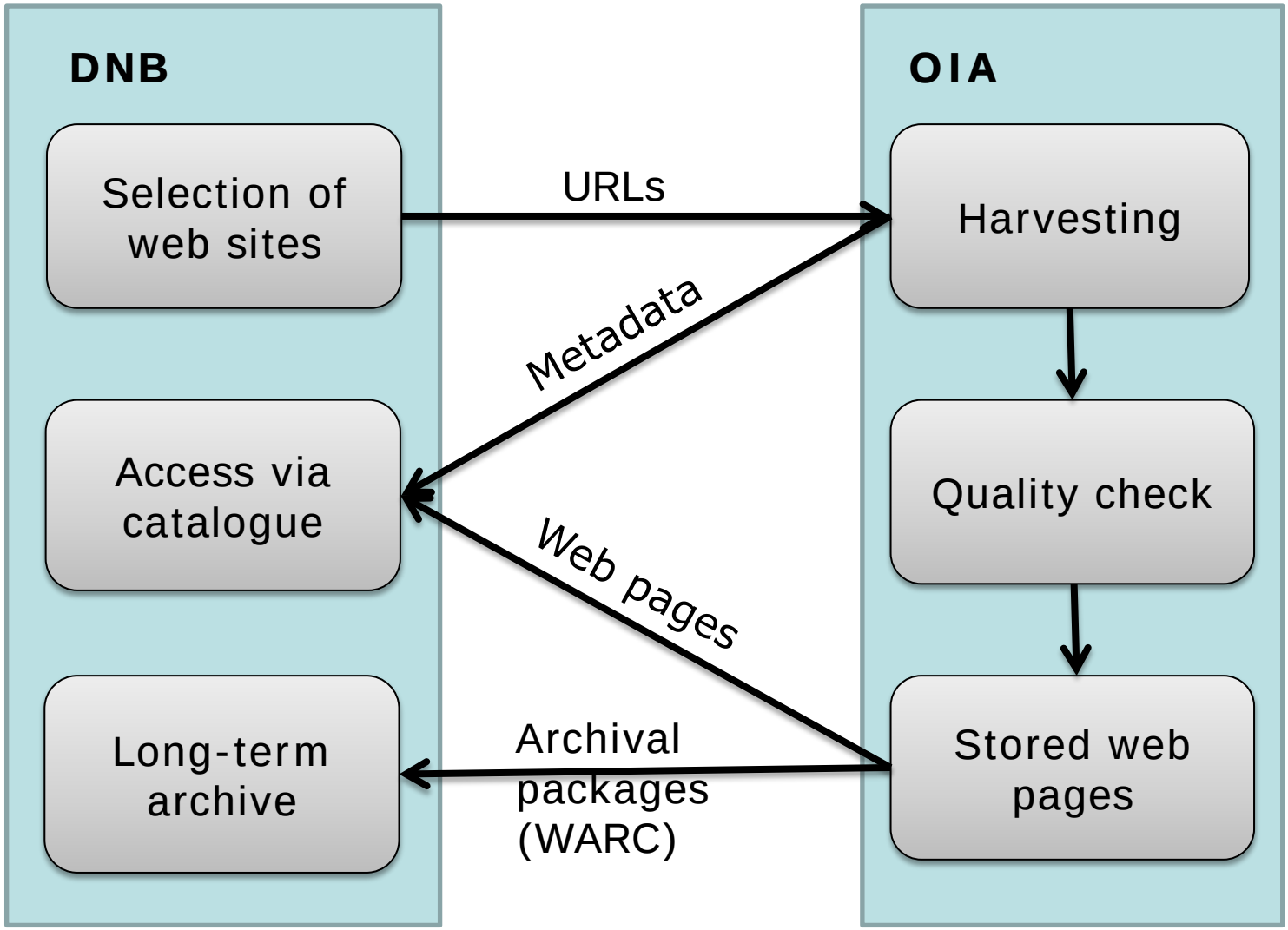
- Legal deposit for digital publications on carriers and net publications (since 2006)
- On carriers (CD-ROM's, floppy disks, DVD-ROM's): Multimedia, educational software, e-books, no games
- E-journals, e-thesis, e-books, digitized books
- Music: CD's, digitized analogue carriers, files
- Web pages
- Access: At least in the reading rooms

Web Harvesting

- Member of the International Internet Preservation Consortium (IIPC)
- Several event harvests (e. g. elections) with the European Archive
- Selective workflow with German company oia since 2012
- Additional broad crawl of .de domain in 2014

Web Harvesting: Workflow

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


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

Web Harvesting: Status

- Topic related web sites (e. g. federal institutions, cultural organizations)
- Default: Sites are crawled twice a year
- Event crawls (e. g. elections, sports events)
- Co-operations planned for selections of web sites
- Currently ca. 1,700 sites with ca. 8,300 crawls

Crawling of news sites

- Challenging: Updated very often, links to articles no longer on start page, pay-walls
- Financial Times Deutschland (www.ftd.de) was closed down in 2013
  - List of links to all articles was provided
  - Complete crawl was archived and is accessible by full text search
- Workshop with German publishers in 2014
  - Articles are not deleted and links don't change
  - Advice to use Google sitemaps
  - Skeptical about giving crawler access behind pay-walls

Crawling of news sites: Status

- Test crawl of SPIEGEL ONLINE (www.spiegel.de)
- Import of XML based Google sitemap as source
- Difficulties with crawler parameters
- In discussion
  - How often?
  - Additional crawls of start page
  - Access to pages and articles behind pay-walls