

→ Old news in a new shape

The digitisation of historical newspapers and its IT infrastructure

Marian Hellema, Koninklijke Bibliotheek
March 5th 2010

Digitisation at the KB

- mass digitisation
 - 60,000,000 pages in 2010-2013
- mainly textual documents
 - books, newspapers, journals
- historical text
 - 1450-1995

Goals of digitisation

- conservation of original documents
- access to cultural heritage

Digitisation of historical newspapers

- newspapers 1618-1995
- national, regional, local and colonial newspapers
- 8,000,000 pages ($\pm 8\%$ of all printed newspapers)





Steps in digitisation

Preparation

Digitisation

Processing

Search & Retrieval

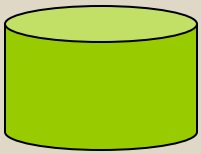
Presentation

Digitisation results

- Images
- Text (OCR)
- Metadata



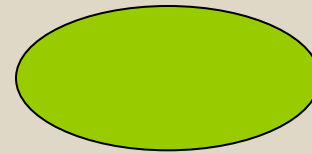
IT infrastructure



storage



metadata



web services

Storage

1. storage for web resources
 - fast access for web services
2. e-Depot
 - archive for long term preservation

Storage web resources

- fast access for web services
- images, text, pdf
- 40 TB for newspaper project
- 100 TB for other digitised documents

e-Depot currently

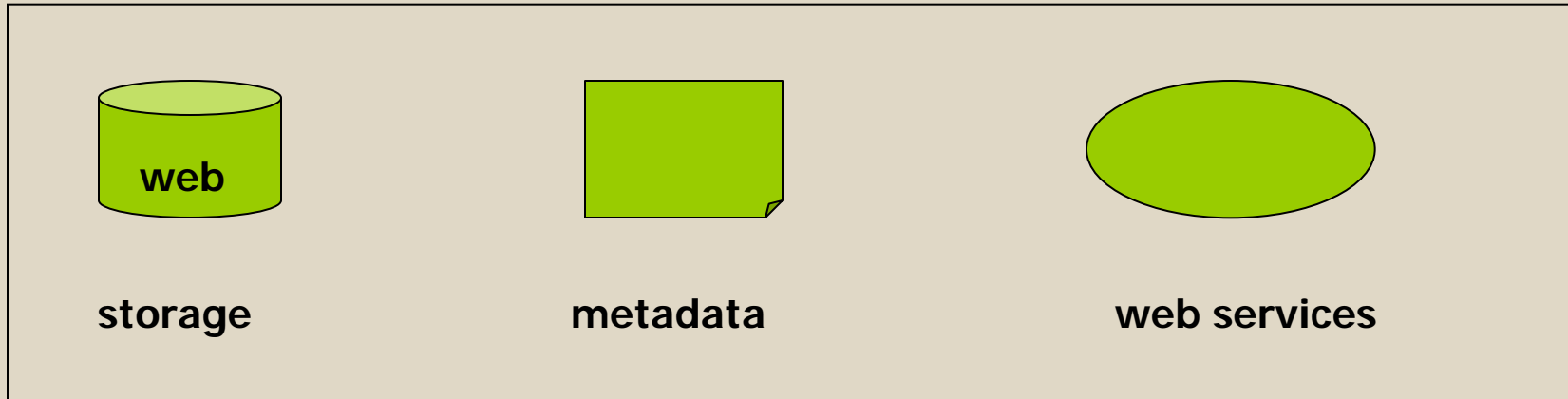
- archive for long term preservation
- OAIS model (Open Archival Information System)
- international scientific publications...
- ... and Dutch digital publications
 - 13,000,000 publications
 - 30 TB storage



e-Depot in future

- in future also master images of digitisation projects
 - master images and metadata
 - 65 TB for newspaper project
 - 700 TB for all digitisation projects

IT infrastructure



Metadata

- Dublin Core
- MPEG21-DIDL
- optionally more formats
- Oracle database
- Verity K2 search engine

Open access

- SRU
- OAI-PMH
- resolver (persistent URLs)

Interoperability: SRU

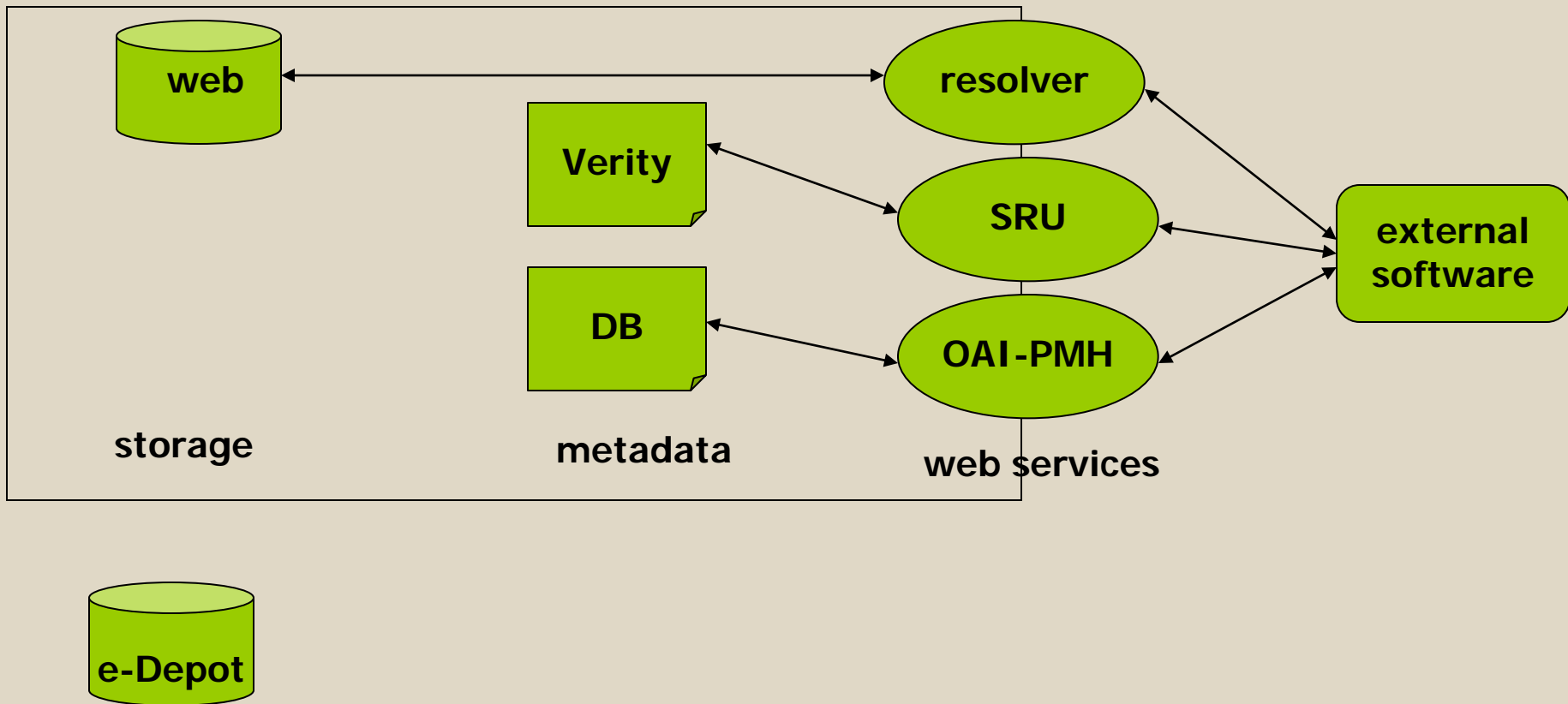
Protocol for Search & Retrieval via URL's

- simplified example SRU request:
www.xyz.nl/?title=Uitvreter&author=Nescio
- web applications can query the search engine via SRU

OAI-PMH

- Open Archives Initiative Protocol for Metadata Harvesting
- central store of metadata
- (web)applications can harvest the metadata

IT infrastructure



Questions ?

marian.hellema@kb.nl

