

# Fedora at The State and University Library, Denmark

Fedora User Group Meeting

Karlsruhe, 2014-09-19

Kåre Fiedler Christiansen

# Digitised newspapers

- 32 million newspaper pages digitised over 3 years
- JPEG2000 files stored in our bit preservation system and referenced from Fedora
- ~2 objects and 7 datastreams per page in Fedora
- Almost 100.000 objects added per day

# Scaling Fedora

- Using fedora 3.7.2
- Scaling issues
  - File system backend unmaintainable (refreshing backup took more than 24 hours)
  - Looking up an object by dc:identifier performance deteriorates with # of objects
  - Mulgara triple store breaks at certain size
  - Rebuilding indexes takes days

# Scaling Fedora

- Tweaked storage backend – millions of XML files not feasible, so bundled in tar files
- Tweaked database index, lookup of a dc:identifier should not require a table scan
- Fixed Mulgara, replacing int with long when memory mapping files

# Scaling Fedora

- With above fixes, Fedora happily keeps up with the load – for now
- Index rebuilds are still a hazzle
- We currently have about 5% of the final size, and scaling is struggling

# Radio and TV in Fedora

- More than ½ million radio programmes
- Almost 1 million tv programmes
- Metadata for programmes in Fedora
- Files with hourly or 24-hourly chunks in bitpreservation system

# Challenges

- Splitting files into programs
- Transcoding programs for dissemination
- See <http://mediestream.dk>

# Get in touch

- Would happily discuss
  - Scaling
    - Speed
    - Size
  - Storing digitized text in Fedora
  - Storing Radio/TV in Fedora



# Get in touch

- [kfc@statsbiblioteket.dk](mailto:kfc@statsbiblioteket.dk)