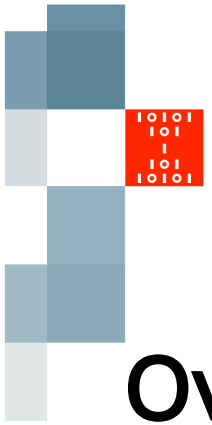


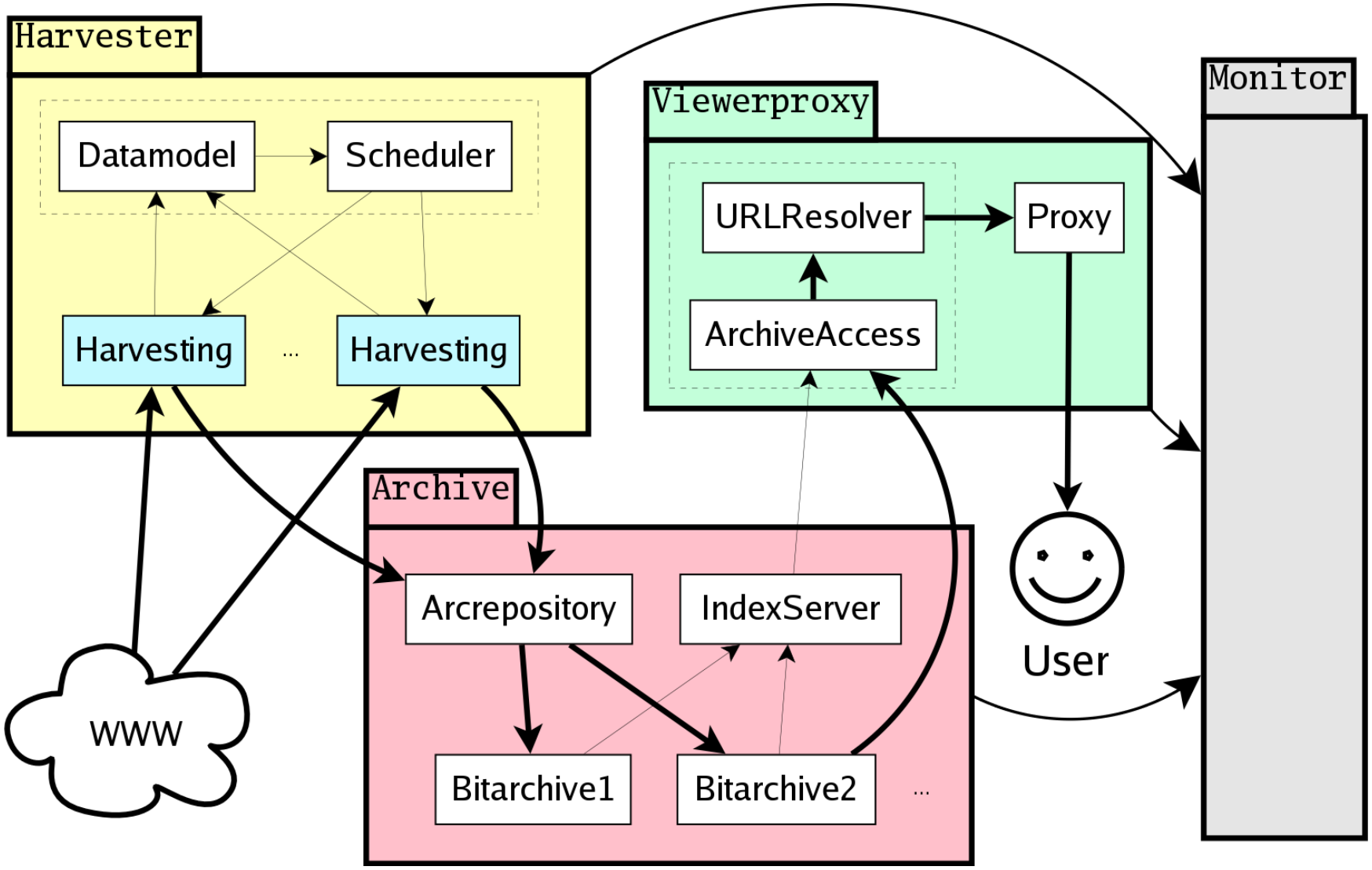
netarchive.dk

Overview of NetarchiveSuite

Lars Clausen
Netarkivet

NetarchiveSuite is...

- a national-scale web archiving system
- based on Heritrix
- highly distributable and modular
- usable by non-technical staff
- written in Java using JMS and Derby DB
- in use in Denmark since July 2005
- now published as Open Source (LGPL)


Harvester
Datamodel
Scheduler
Harvesting
...
Harvesting
Viewerproxy
URLResolver
Proxy
ArchiveAccess
Monitor
Archive
Arcrepository
IndexServer
Bitarchive1
Bitarchive2
...
User
WWW

Harvest types

- Snapshot harvests
  - All known domains
  - One template, one size limit, default configuration
- Selective harvests
  - Selection of domains
  - Specific configurations
- Event harvests
  - Selective harvest with mass ingest

Harvests and domains and seedlists

- Domains have configurations and seedlists
- Configurations set limits, use seedlists
- Harvests specify configurations and parameters
- Harvests are run one or more times
- Harvest runs are split into jobs
- Jobs are run in Heritrix