netarchive.dk

STATSBIBLIOTEKET Det Kongelige Bibliotek

The danish national internet archive

A status on the project - by

Bjarne Andersen - manager

netarchive.dk

+45 89462165

bja@netarkivet.dk

Agenda

- New legal deposit law in Denmark
- Netarchive.dk
- Technologies used
- Ease of distribution
- Simple flow animation
- Administrative interface
- Snapshot harvesting
- Future work

Legal deposit law 1

- Revision of the legal deposit law in 1997
  - -> legal deposit included static documents on the internet
- During in 1998-1999 clever people found out that:
  - We were actually perserving the least interesting part
    - Many of the documents in that collection are also available in print
- A lot of work was done between 2000-2004
  - 2 pilot projects run by the two national libraries
    - Testing different software / different strategies for archiving / storing web material
  - A governmental publication on ”preserving the danish digital cultural heritage” (2003)
  - A report to the ministry of culture (2004) outlining
    - Recommondations from the two national libraries on how to solve the ”entire” problem
    - Issues to be covered by a new revision of the legal deposit law

Legal deposit law 2

- A new revision came into force on july 1st 2005
  - Allowing the two national libraries to automatically gather all **danish** websites
  - Danish roughly defined as:
    - Websites on the .dk TLD
    - Websites minded on a danish audience / written in danish
    - Websites about danish poeple (Hans Christian Andersen)
    - More or less any site of interest to Denmark
  - We are by law granted access to all relevant data from the .dk TLD administrator

Legal deposit law 3

- The law covers all **public available** material
  - Material that all danish people *in pricipal* can gain access to
    - Material which requires action before usage (payment, registration….)
    - Pay-sites should hand out username / password upon request (for free)
- Other interesting parts
  - Combined strategy (snapshot, selective and event-harvesting)
  - Robots.txt explicitly mentioned in the remarkings to the law
    - A lot of the very interesting websites have very restrictive robots.txt's (we discovered around 35.000 robots.txt-files)

Legal deposit law 4

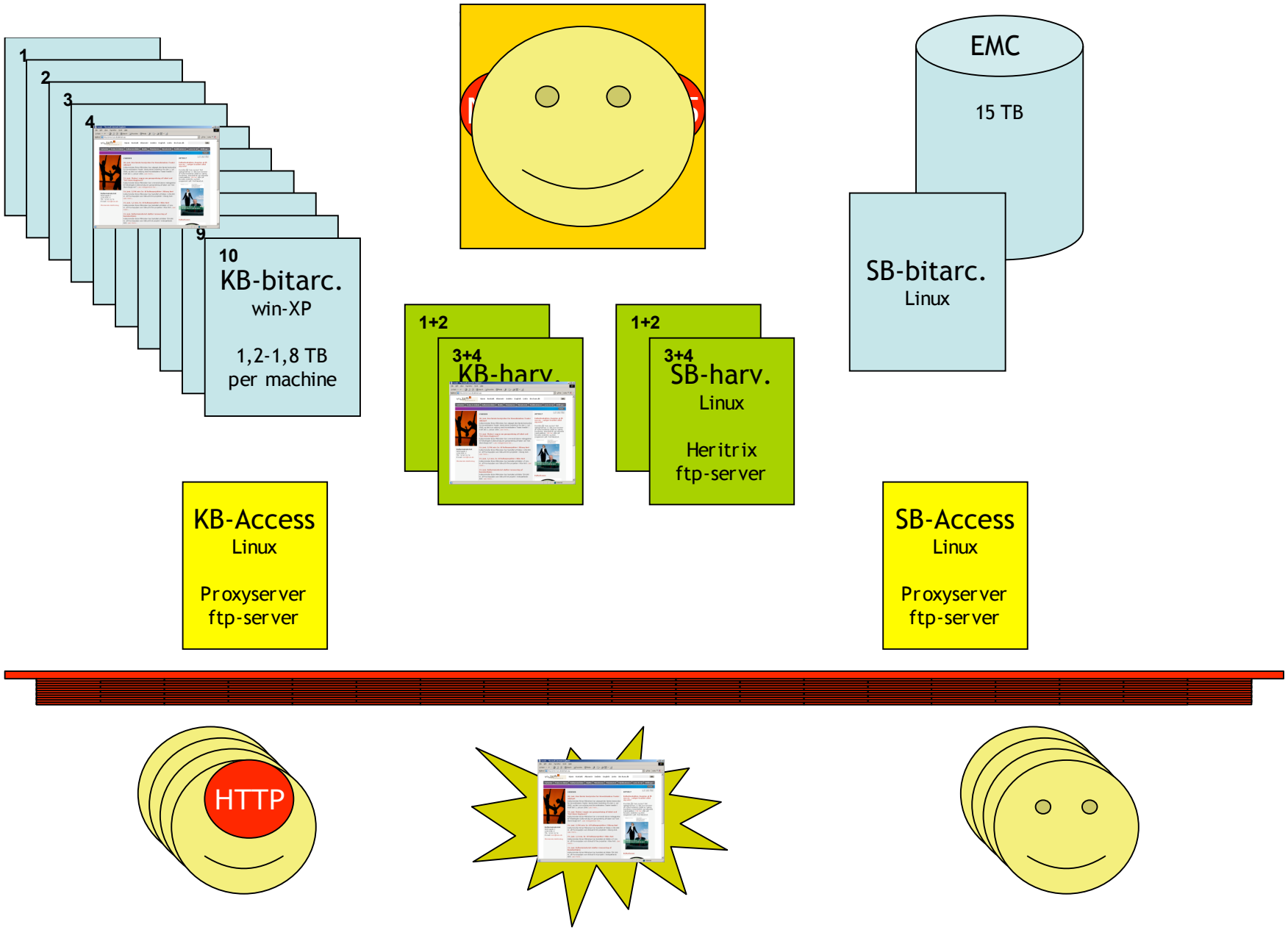
- In the end led to funding of
  - Netarchive.dk
    - Virtual centre in cooperation between
      - The Royal Library, Copenhagen
      - The State & University Library, Aarhus
    - Implementing a complete system
    - Running the system in the future
      - Currently with an anual budget of 400.000 euros

Technologies used

- Pure java (1.5)
  - JMS to distribute
  - Derby to store administrative data
  - Jetty for running our administrative web interface
  - Heritrix for doing the actual crawls
    - Embedded into our own 'Server-application'
- Currently uses ftp for all file transfers
  - Moving to sftp (ssh) in the near future

Ease of distribution

- JMS gives you
  - Asynchronous communication
  - 'Any' node can be plugged / unplugged
  - New nodes can be added to the running system
    - If more harvester resources are needed
    - When more diskspace is required
- Our setup gives **very** easy
  - Installation (xml-file + one click) on 20 machines
    - Using a java deployment application + ssh & ssh-keys
  - Start / Stop of the entire system (one click)


1
2
3
4
9
10
KB-bitarc.
win-XP
1,2-1,8 TB
per machine
1+2
3+4
KB-harv.
1+2
3+4
SB-harv.
Linux
Heritrix
ftp-server
EMC
15 TB
SB-bitarc.
Linux
KB-Access
Linux
Proxyserver
ftp-server
SB-Access
Linux
Proxyserver
ftp-server
HTTP

Administrative interface

- We needed a curator tool
  - Requirement number 1: Operated by librarians
- With the web interface they can:
  - Define harvests (all three types)
    - Based on quite simple settings + a number of different predefined heritrix setups
  - Do quality control
    - Looking at harvest results (reports)
    - Browsing through harvested material
      - Automated pickup of missing URIs (handled by the proxy)
  - Monitoring the entire system
    - Implemented with standard java logging – including a JMS-Handler + SMTP-Handler for serious events

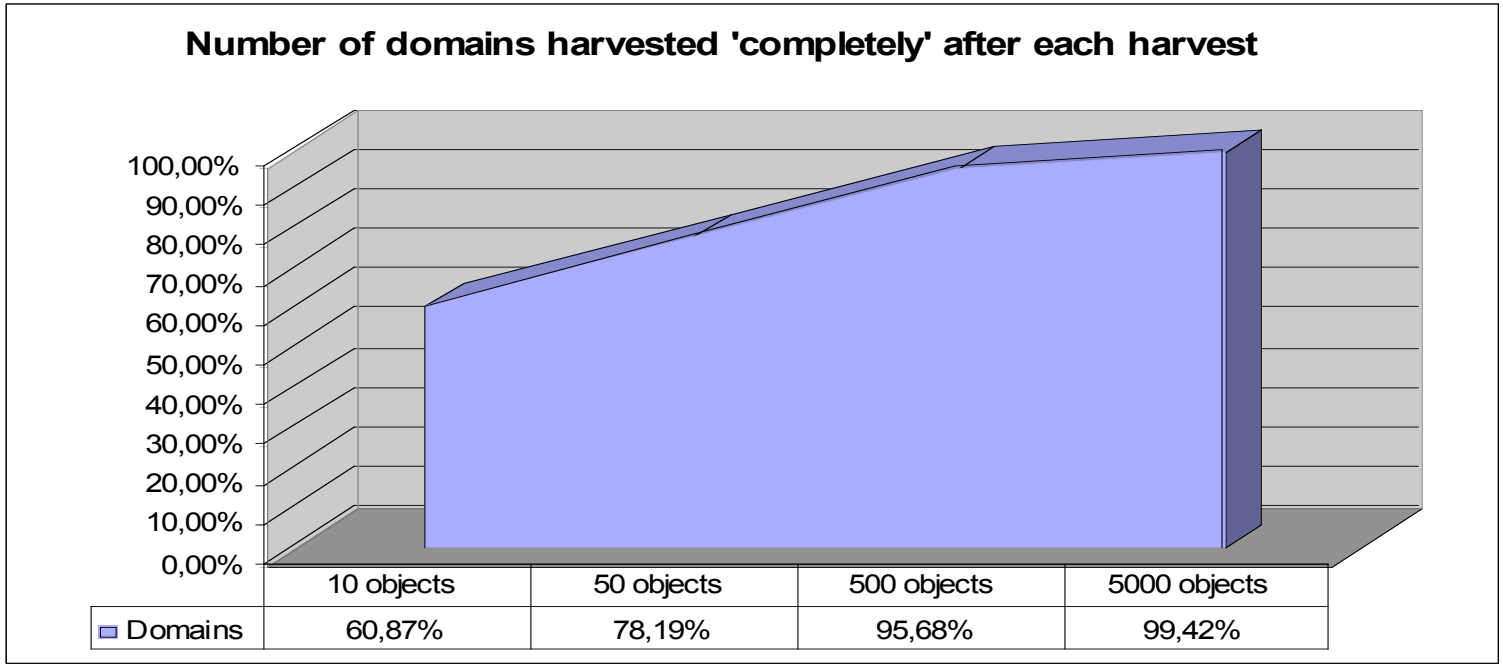
Some words on snapshot harvesting

- The .dk TLD currently holds > 600.000 domains
- Our scheduler splits these into chuncks of domains (between 10 and 10.000)
  - Grouping domains by size
- The size is calculated from privious harvests
- Makes crawling more efficient
  - Ensuring domains of same size gets in the same job
  - Jobs relatively small (< 2M URIs) – no OOMEs
  - Runs for a smaller amount of time (< 24 hrs)

Our first snapshot harvest

- Running i cycles
  - maximum of 10 objects/dom (>600.000)
    - For the doms hitting that limit (230.000) raising the limit to 50 (starting all over again)
    - Same thing -> max 500 (80.000)
    - Same thing -> max 5000 (26.000)
    - Same thing -> max ? (3.500)
  - Desperate need of duplicate handling

Number of domains harvested 'completely'



Future work

- Stabilize a quite complicated distributed technical system
- Integrate access with
  - freetext- search
  - Timeline functionality
- More monitoring of the system
- Using the JMX API to heritrix
- Logical preservation issues
- Automated quality control
  - More sophisticated queries into harvester reports
    - Finding potential crawlertraps
  - Automated browsing throug the archive
    - Via the proxy logging missing URIs

netarchive.dk

Questions