netarchive.dk

The history behind NetarchiveSuite

Bjarne Andersen

daily manager

netarchive.dk

bja@netarkivet.dk

Legal deposit in Denmark I

- 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

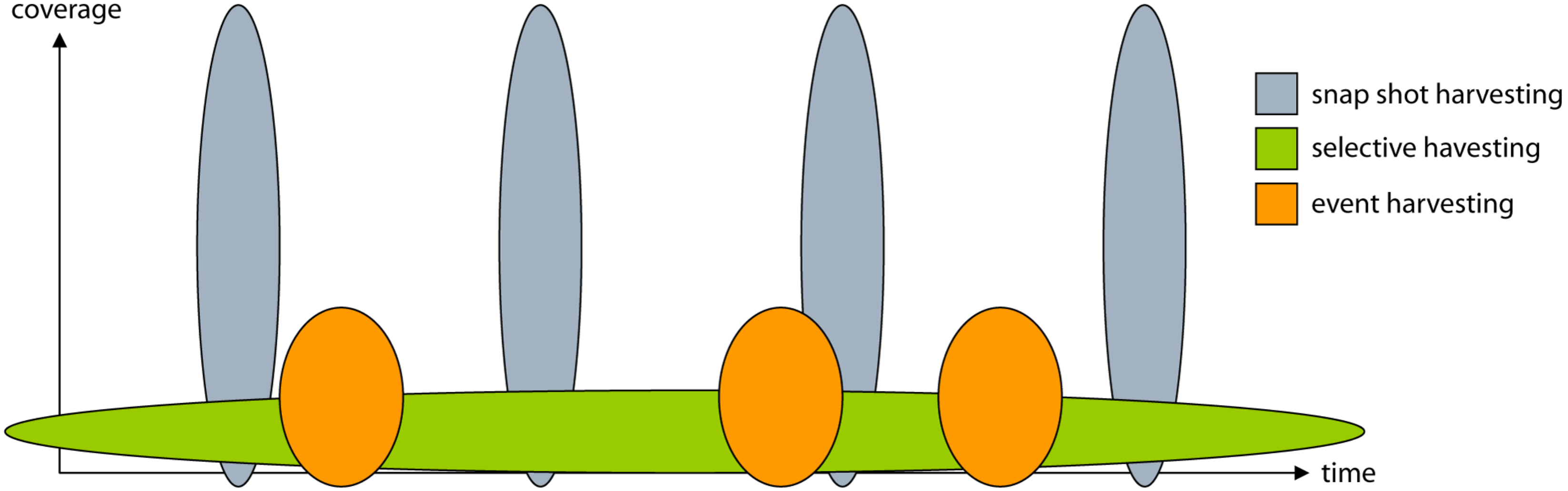
Legal deposit in Denmark II

- 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 in Denmark III

- 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)
- The national libraries are allowed to collect all covered material **without any permissions**
- Combined harvest strategy
  - snapshot, selective and event-harvesting
  - Developed in cooperation with Center for Internet Research, Aarhus University
  - Based on international experience

3 strategies



Important requirements

- Use heritrix as the webcrawler
- Handle all 3 types of harvests
- Distribute between 2 locations
  - Fully automated storage of multiple copies
  - Including active bit-preservation functionality
- Scale to the size of the danish web
  - Currently 5 harvesting machines and 40 storage-nodes (bitarchive machines) holding 45 TB of data
- Easy to maintain and monitor
  - In netarchive.dk: Installable on 50 machines in 5 mins.
  - Monitor all applications on all machines from one central place (JMX / webinterface)
- Must document ”everything” automatically
  - Metadata and crawler-output (logs) are written to metadata ARC-files and uploaded to the archive together with the actual harvested material

Administrative interface

- We needed a curator tool
    - Requirement number 1: Operated by librarians
- With the web interface librarians 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 / statistics
        - Browse through harvested material
            - Automated pickup of missing URIs (handled by the proxyserver application)
- With the web interface you can also
    - Do bit active bit preservation operations
    - Monitor the entire system
    - Handle harvester templates
    - Do mass ingest of domains

NetarchiveSuite

- First version running in Denmark from 1st july 2005 (when the law came into force)
  - Invested about 6 Man-years on development
- Code currently developed/maintained with usage of around 1.6 Man-years per year.
  - Hopefully the open source release will over time grow new cooperations that will increase the effective amount of coding power.
- The initial child diseases should be eliminated – have been running quite stable for 2 years now handling a quite large amout of data