Kulturarw³³

The Swedish WWW Archiw³³e
Background

- National deposit act, 1661 (cf. 1.5 shelfkm/year)
- Latest revision 1993
  - only electronic documents in fixed form
  - CD-ROM, diskettes
- First Swedish web journal lost (cf. paper 1645)
- New law under way… (stalled)
- Start summer 1996 (”brain storming colloquy”)

frans_lettenstrom@harcourt.com

2002-04-11
Archival Aims

• All www and gopher pages in Sweden
  – images, video clips etc
  – .se, generic TLD’s and .nu
• All articles in electronic journals/newspapers
• All Swedish newsgroups / discussion lists
• Problems: databases behind dynamic pages
Organisation

- Section under National Library of Sweden
- Three technicians (1 PhD, 2 Engineers)
- Engineer student projects
- Collaborates with SveSök
  (www.svesok.kb.se)
Cooperation

• The National Archives and The Archive of Sound and Moving Images
• Nordic cooperation (NWA)
• International cooperation (PreWeb)
• Individual countries (Chile?)
• Central America (1997-1999)
Robot Software

- Modified version of Nordic Web Index’s robot software
- Saving the data in a MIME format
- Temporary storage on DLT tape
Network

- Internal 100Mbit/s switch ethernet
- Connected to Internet via SUNET, 100 Mbit/s
- SUNET academic network 10 Gbit/s
Archive System

- Archive bits and bytes not tapes, cdroms or disks
- HSM
- UNIX server
- Disk system
- Tape robot

- Disk
- (Optical)
- Tape robot
Legal framework

• Legal deposit legislation is stalled by EU directives on privacy and copyright
• ”...a library can produce copies for preservation”? No!
• ”Datainspektionen”, The Data Inspection Board, has halted the archiving for privacy reasons
New Legislation (stalled)

• The Royal Library should collect the “Swedish WWW” at least four times a year.
• If information on WWW intended for public use is protected (by passwords for example) the passwords or access must be given to the Royal Library’s robot.
• The Royal Library can also by law request copies of online databases.
• The WWW archive will be available for “researchers” only.
Statistics (1997 $\rightarrow$ 2001)

- 5.8 $\rightarrow$ 30 M files
- 143 $\rightarrow$ 1334 Gbytes
- 15.7 $\rightarrow$ 127 k websites
- 45% .se
- 43% .com, .edu, .org & .net
- 12% .nu
- suecana extranea
Statistics

• 400 different MIME types found. However, many are the same
• 49% text/html
• 22% image/gif
• 17% image/jpeg
• 4% text/plain
• 8% others (audio 0.7%, video 0.06%)
• html/gif/jpeg/text = 92%
http://www.kb.se/kw3