Chasing the news

report from 10 years of digital legal deposit in Denmark

Tonny Skovgård Jensen
Director, National Library Division
State and University Library
Denmark
Where is it?
The State and University Library

- 3 main functions:
  - University library
  - National services for public libraries
  - National library

- National Library #2
  - The Royal Library in Copenhagen is the oldest and biggest

- We specialize in media:
  - Newspapers
  - Radio
  - Television
  - The Danish part of the web (In cooperation with the Royal Library)
  - Recorded music
  - Historical audiorecordings
  - DVD / VHS
10 years of digital legal deposit

• Legal deposit act – 10 years anniversary
• Looking up - looking back – looking forward
• What did we learn?
• What did we collect?
• Which challenges do we face?
• Where should we go from here?
Do you remember

2005
The world 2005

• Facebook was just about a year old, and had about 5 million users. Worldwide.
• YouTube was founded this year. Few people could imagine at that time the enormous amount of video that was soon to be distributed online.
• Twitter was founded the year after, in 2006
• You would have to wait two more years to see the first iPhone, one more year to hear of the first android smartphones, and 5 years to see the first Ipad.
• State of the art devices was still PDA’s and also the first smartphones, like the Sony Ericsson P910, but slow connections and slow hardware made them quite limited for practical purposes.
History

- State Media Archive: 1/7/1987
- Legal deposit act Recorded music and film: 1/1/1998
- Legal deposit act Internet, radio, television: 1/7/2005
- Collecting radio and television on tapes
- Radio and television digital
- Web harvesting
- Mediestream

Timeline:
- 1990
- 1995
- 2000
- 2005
- 2010
- 2015
Digital collections
Digital legal deposit

- Publishers and producers do not deliver or upload the content. We have to collect or harvest it.
- News content is generally not handled as a separate collection – it is mixed with other types of content.
Digital collections

Digital repository

LD
Radio & Television
300 TB

LD
Web harvesting
100 TB

PA
Newspapers / PDF

PA
Online music

PA
Online video

PA
Ripping CD’s and DVD’s
50 TB

PA
Newspaper digitization
200 TB

PA
Retro: audio and video digitization

PA
Newspaper

LD
Web harvesting

LD
Radio & Television

LD
Online video

LD
Online music
Harvesting the web

- Broad crawl
- Selective crawl
- Event crawl
- Special crawl

coverage vs. time
Harvesting “everything” ... but ...

- Very complex process
- Huge quantity – harvest / select / curate
- Where is the Danish content?
- A webpage is complex and dynamic
- Technology is developing fast – the harvester need to know how to capture the content
- Streaming
- Databases
- Paywalls
- Complex site structures
- Social media, mobile apps ....
Preservation

• ISO 16363 (Audit and certification of trustworthy digital repositories)
• Bit preservation – not trivial, but also not a complete jungle.
• Ready for conversion
Access

OWN

API

OTHERS

Datamining

Mediestream.dk

Cosound

Cultural heritage cluster

LARM

Digital repository

The Netarchive

SPECIAL

Tonny Skovgård Jensen
tsj@statsbiblioteket.dk
Access to the webarchive

• Everything in one arhive – personal data issues. New EU regulation is on the way.

• Very restrictive - access is only for academic researchers and students writing their thesis
  – Maybe not a problem in 2005, but it is now where a lot of important content is online only

• Combined legal and technical issue

• Access by URL (search is coming soon)
Summary – a new paradigm?

• Looking back – overwhelming amounts of digital content
• Looking forward – even more, and more online only content
• Can we select in a meaningful way?
• Can we get it all?
• Can we still be in control?
• A case …
Discussing if PDF’s can replace microfilming the paper news
Questions?

TSJ@statsbiblioteket.dk
Credits

• Drawings: Jørgen Stamp, from www.digitalbevaring.dk

• Photos from the library: Thomas Søndergaard