Archiving the Internet: Challenges, Projects, and the Austrian Perspective

Andreas Rauber[1], Andreas Aschenbrenner[1], Alfred Schmidt[2]

[1]Department of Software Technology (ifs)
Vienna University of Technology
http://www.ifs.tuwien.ac.at

[2]Austrian National Library
http://www.onb.ac.at

AOLA

Overview

- Motivation
- Technical challenges
  - Selection
  - Archiving
  - Preservation
  - Access
- Related projects
  - Internet Archive
  - Kulturaw3
  - Nedlib
- AOLA: Austria On-Line Archive

Motivation

Why Should We Archive the Internet?

- Collection of sex & crime
- Masses of useless and/or wrong information
- Incredibly huge
- Only 0.00x % of all information is actually being looked at
- Who is interested in some fellow’s homepage?
- Important information is published in “real” media anyway

Motivation

The Invention of the Press

- Internet often compared to invention of the printing press
- Explosion of printed information
- Quality much lower than manually crafted codices
- Not to be considered important?

- Letters more interesting than books
- Ads, posters, and snippets tell more about a society than “high-quality” information sources
- What if only codices had been preserved?

Motivation

Some Considerations

- Increasing masses of information published electronically
- Volatility of Internet resources
- Social and cultural dimension - modern cultural heritage!
- Need to preserve the Internet
  - information / content
  - look-and-feel
- The early days of the Internet are already lost!

Motivation

Challenges

- Legal challenges
  - copyright issues
  - authenticity

- Technical challenges
  - what to archive
  - how to archive
  - how to keep the archive in good condition
  - how to provide access to the archive

- Financial Challenges
  - who is willing to pay
  - what do we gain (earn?) from it

AOLA

Overview

- Motivation
- Technical challenges
  - Selection
  - Archiving
  - Preservation
  - Access
- Related Projects
  - Internet Archive
  - Kulturaw3
  - Nedlib
- AOLA: Austria On-Line Archive

Technical Challenges

1. Selection

- Building a complete archive is technically impossible
  - enormous amount of data
  - no central catalogue
  - high dynamics and volatility

- Manual Selection:
  select specific sites plus archiving frequency
- Automatic harvesting:
  automatically crawl hyperlinks to download sites

- Which sites to archive: *.at, .com, .cc, Austriaca, ...
- Questions of liability: who is responsible for content?

Technical Challenges

2. Archiving

- Selection of suitable storage media
  - high capacity
  - long durability
  - stable technology
- Migration to new storage media
  - when reaching lifetime of storage medium
  - when storage technology becomes obsolete
  - no “museum” of old devices
  - automatic transfer
- Media of choice currently
  - harddisk arrays
  - tapes

Technical Challenges

3. Preservation

- Digital objects have to be “interpreted”
- Software required for access
- Software needs specific hardware platform
- Ensure, that access to documents is possible in the future
- “Museum” of old hardware impossible to sustain

- 2 approaches
  - Conversion:
    converting to “standard file formats”
  - Emulation:
    emulating obsolete hardware on new systems

Technical Challenges

3.1 Preservation: Conversion

- Files are converted into (few) selected standard file formats (z.B.: text, (series of) image(s), sound, ...)

+ Access via a few file formats -> small set of access software
+ flexible and cheap, especially for immediate access
+ When standard file format becomes obsolete, converters will be around due to critical mass of existing files

- Loss of information at conversion (functionality, looks-and-feel)
- not suitable for all materials (e.g. interactive art)
- constantly maintain all data

Technical Challenges

3.2 Preservation: Emulation

- Storing description of system environment required for executing access software (metadata)
- Emulators for hardware platforms are created as the necessity arises
- Intermediate representation language

+ Theoretically most stable model

+ Conceptually clean solution

- Very expensive (development of specialized emulators)

- Not useful for quick, casual access

- Information required for emulator development might not be known

- Applicability has so far only been demonstrated on some selected examples, several open questions

Technical Challenges

4. Access

- Mostly legal issues
- Technical issues
  - provide access to large data stores within reasonable time frames
  - navigating the archive:
    * by content within a time frame
    * browsing through time (evolution of websites)
  - providing transparent access through emulators or migrated file formats

AOLA

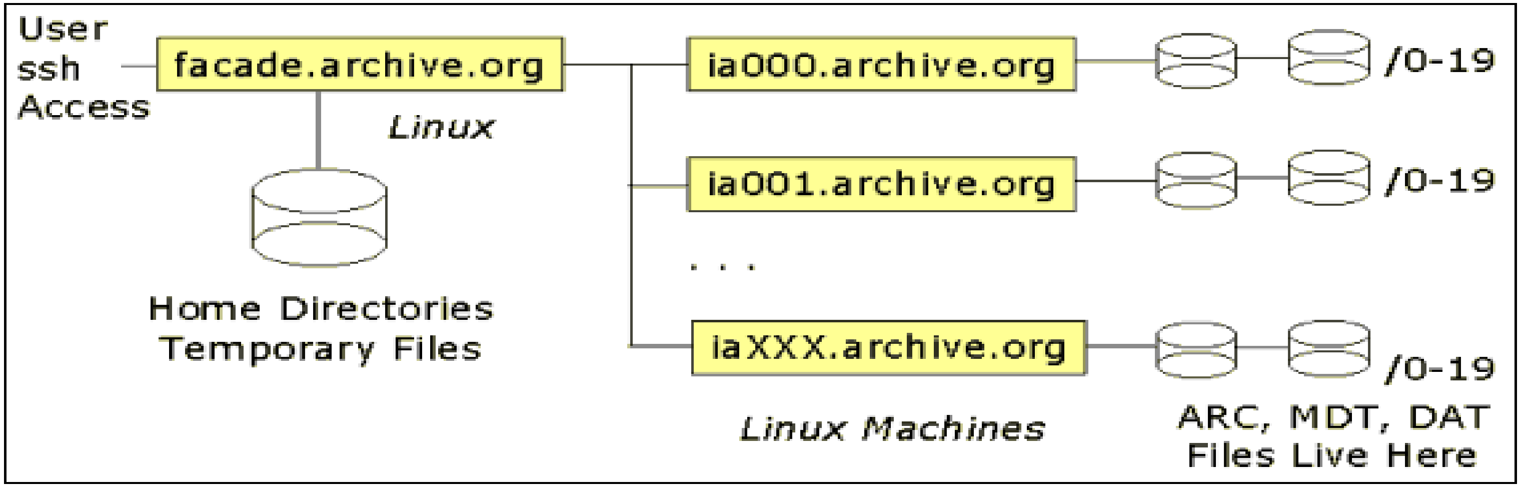
Overview

- Motivation
- Technical issues
  - selection
  - archiving
  - preservation
  - access
- Related projects
  - Internet Archive
  - Kulturaw3
  - Nedlib
- AOLA: Austria On-Line Archive

Related Projects

Internet Archive

- Since 1996, URL: *www.archive.org*
- Set of Linux-Systems with harddisk arrays



- Archives “donated” data collections
- Mostly based on free harvesting (Alexa)
- Initially text-only, now all data types

Related Projects

Internet Archive (2)

- March 2001: approx. 43 TB (March 2000: 14 TB)
- Daily growth: up to 100 GB per day
- Redundancy: distributed across several sites
- Automatic migration onto new storage media
- Collecting existing emulators
- Access limited to research institutions
- “Programming skills” required for using the archive

Related Projects

Kulturaw3

- URL: *http://kulturarw3.kb.se/html/kulturarw3.eng.html*
- Project of the Swedish National Library, since 1996
- Sun Sparc Stations with tape robot archive
- Uses modified indexer (Combine) for harvesting
- Snapshots of the swedish web (.se, .nu, special domains)
- Preservation: originals plus possibly standard file formats
- 5 snapshots so far, last crawl:
  15 Mio. URLs from 58,400 websites, total approx. 280 GB data
- Tendency: dramatically increasing (incomplete 6. crawl: 360 GB !)
- Access tool under development
- Plan to use hierarchical storage media

Related Projects

NEDLIB

- URL: *http://www.kb.nl/coop/nedlib/*
- Networked European Deposit Library
- EU-Project January 1998 till January 2001
- Framework for Archiving on-line media
- Open for all concepts
- Guidelines, technical standards, “best-practice” models
- Nedlib harvester for archiving issues
- Platforms: Linux, SUN, ...
- Further tools under development
- No large-scale experiments so far

Related Projects

Further Projects:

- Nordic Web Archive - *http://nwa.nb.no*
- EVA - *http://www.lib.helsinki.fi/eva/english.html*
- Pandora - *http://www.nla.gov.au/policy/plan/pandora.html*
- CAMiLEON - *http://www.si.umich.edu/CAMILEON/*
- CEDARS - *http://www.leeds.ac.uk/cedars/*
- Prism - *http://prism.cornell.edu/PrismWeb/*
- LOCKSS - *http://lockss.stanford.edu/*
- Arches - *http://www.rlg.org/strat/projarch.html*
- InterPARES - *http://www.interpares.org/*
- Victorian Electronic Records Strategy - *http://www.prov.vic.gov.au/vers/*
- National Library of Canada Electronic Collection - *http://collection.nlc-bnc.ca/e-coll-e/index-e.htm*

AOLA

Overview

- Technical issues
  - selection
  - archiving
  - preservation
  - access

- Related Projects
  - Internet Archive
  - Kulturaw3
  - Nedlib

- AOLA: Austria On-Line Archive

AOLA

Austria On-line Archive

- URL: *http://www.ifs.tuwien.ac.at/~aola/*
- Cooperation between the Austrian National Library and the Department of Software Technology, Vienna Univ. of Technology
- Pilot study: preparations since 1999, 1. phase since March 2001
- Linux-System with 240 GB harddisk plus 6-fold tapedrive
- Open source approach to ensure independent access
- Initially: Nedlib harvester (incl. modifications and expansions)
- Goal: snapshot of the Austrian webspace

AOLA

Austria On-line Archive (2)

- between May 7 and May 16 2001 approx. 10 crawler parallel
- Download during pilot phase: approx. 1GB per day
- at-domain as well as selected subdomains,
  esp. **.cc, *.com, *.edu,* etc.

- Statistics May 7. - 16.:
  - about 666.000 unique URLs harvested
  - 1.210 sites accessed
  - total of 8.3 GB of data stored
  - numerous problems with Nedlib harvester encountered

AOLA

Results of Pilot Phase

- Basically, the setup works!
- Archiving system problems:
  - XFS file system for Linux still unstable (pre-release)
- Nedlib Harvester
  - problems with mal-formatted links in html pages
  - communication problems within system
  - several pages downloaded numerous times
  - still in development phase
- --> crawl needed to be stopped

AOLA

Statistics - Domains (excerpt)

Domain (47)	Size	#Docs	#Hosts
at	4.345.098.283	239.821	8.740
ac.at	454.072.064	19.248	676
co.at	138.067.628	12.557	427
gv.at	75.164.569	4.584	234
or.at	55.349.125	5.576	197
com	331.110.660	18.419	813
edu	737.588	24	9
int	1.183.712	80	1
net	202.837.209	13.108	457
org	45.412.967	1.908	93
cc	402.520.513	13.513	119
de	32.043.054	2.233	250
hu	516.579	70	1
tw	43	1	1

AOLA

Statistics - Extensions (excerpt)

Extension (337)	Size	#Docs
htm	483.073.276	58.554
html	504.579.811	43.815
txt	5.499.481	1.452
wav	38.212.215	107
mp3	216.255.942	169
avi	8.955.594	12
mpg	179.078.751	19
jpeg	2.090.006	133
jpg	549.196.700	35.298
gif	388.089.230	76.498
zip	184.489.636	
gz	3.091.367	9

Extension	Size	#Docs
exe	266.082.863	513
bin	240.768	1
cgi	49.386.025	3.671
java	7.489	1
jsp	18.854.236	684
asp	848.447.298	26.527
php	160.913.685	7.881
xls	4.000.256	28
doc	41.327.637	328
f94	957	1
fangan	10.022	3
woa	4.046	7
346a	43	1

AOLA - Next Steps

- Currently switching to Combine harvester for the next crawl
- Transform pilot study into permanent institution
- Archiving frequent snapshots of the Austrian webspace
- Develop long-term strategies for preservation
- Combination of conversion and emulation approaches
- Setting up technical and personnel infrastructure

AOLA

Conclusions

- **Goal**: Preservation of (modern) cultural heritage
- **Selection**: Combination of manual selection and free harvesting
- **Archiving**: Migration of (hierarchical) storage media
- **Preservation**: Emulation and conversion approaches
- **Access**: Interfaces and legal aspects
- **Urgency**: We have to start **NOW**!

AOLA Project-Homepage:

http://www.ifs.tuwien.ac.at/~aola