Technische Aspekte der Archivierung von On-line Medien

Dr. Andreas Rauber

Institut für Softwaretechnik (ifs)
Technische Universität Wien
Favoritenstr. 9-11/ 188; A - 1040 Wien

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

Archivierung von On-line Medien

Überblick

- Technische Problemstellungen
  - Auswahl
  - Archivierung
  - Preservation
  - Zugriff
- Verwandte Projekte
  - Internet Archive
  - Kulturaw3
  - Nedlib
- Die österreichische Pilotstudie AOLA

Technische Problemstellungen

1. Auswahl

- Vollständige Archivierung ist technisch unmöglich
  - enormes Datenvolumen
  - fehlendes zentrales Verzeichnis
  - hohe Dynamik / Änderungsrate

- Manuelle Selektion:
  Auswahl von bestimmten Seiten plus Archivierungsfrequenz
- Automatisches Harvesting:
  Automatisches Archivieren durch Verfolgung von Hyperlinks

- Was wird archiviert: *.at, .com, .cc, Austriaca, ...
- Wie oft wird archiviert?

Technische Problemstellungen

2. Archivierung

- Auswahl von geeigneten Speichermedien
  - große Kapazität
  - lange Haltbarkeit
  - stabile Technologie
- Migration auf neue Speichermedien
  - bei Erreichen der Lebensdauer eines Datenträgers
  - bei Auslaufen einer Speichertechnologie
  - automatischer Transfer auf neue Speichermedien
  - kein “Museum” alter Geräte

Technische Problemstellungen

3. Preservation

- Digitale Objekte müssen “interpretiert” werden
- Software ermöglicht Zugriff
- Software erfordert Hardwareplattform
- Sicherstellen, dass Zugriff auf “alte” Dokumente möglich bleibt
- “Museum” alter Geräte technisch nicht durchführbar

- 2 Ansätze
  - Migration:
    Konvertierung auf “Standardformate”
  - Emulation:
    Emulation obsoleter Hardware auf neuen Systemen

Technische Problemstellungen

3.1 Preservation: Migration

- Dateien werden in (wenige) Standardformate konvertiert (z.B.: Text, (Serie von) Bilder(n), Sound, ...)

\+ Zugriff über wenige Standardformate möglich

\+ flexibel und kostengünstig, vorallem für unmittelbaren Zugriff

\+ Bei Auslaufen eines Standardformats werden aufgrund “kritischer Masse “ Konvertierungstools zur Verfügung stehen.

\- Informationsverlust bei Migration (Funktionalität)

\- nicht für alle Datenformate realisierbar (z.B. interaktive Kunst)

Technische Problemstellungen

3.2 Preservation: Emulation

- Erfassung der zur Interpretations eines digitalen Objekts benötigten Systemumgebung (Metadaten)
- Bei Bedarf wird ein Emulator für die obsolete Hardware-Plattform entwickelt

\+ theoretisch stabilstes Modell

\+ konzeptuell saubere Lösung

\- sehr kostenintensiv (Emulatorentwicklung)

\- benötigtes Wissen unter Umständen nicht in nötigem Umfang bekannt

\- technische Realisierbarkeit nur an ausgewählten Beispielen vorgeführt, zahlreiche offene Fragen

Technische Problemstellungen

4. Zugriff

- primär rechtliche Fragestellungen:
  - Copyright: wer darf wann wie worauf zugreifen und was damit tun
  - kostenpflichtige Archive
  - Nutzungsrechte
  - Authentizität
- technische Problemstellungen:
  - Zugriff auf große Datenbestände innerhalb vertretbarer Zeitrahmen
  - Navigation durch das Archiv
    * inhaltlich innerhalb eines Zeitfensters
    * durch die Zeit (Evolution von On-line Medien)

Archivierung von On-line Medien

Überblick

- Technische Problemstellungen
  - Auswahl
  - Archivierung
  - Preservation
  - Zugriff

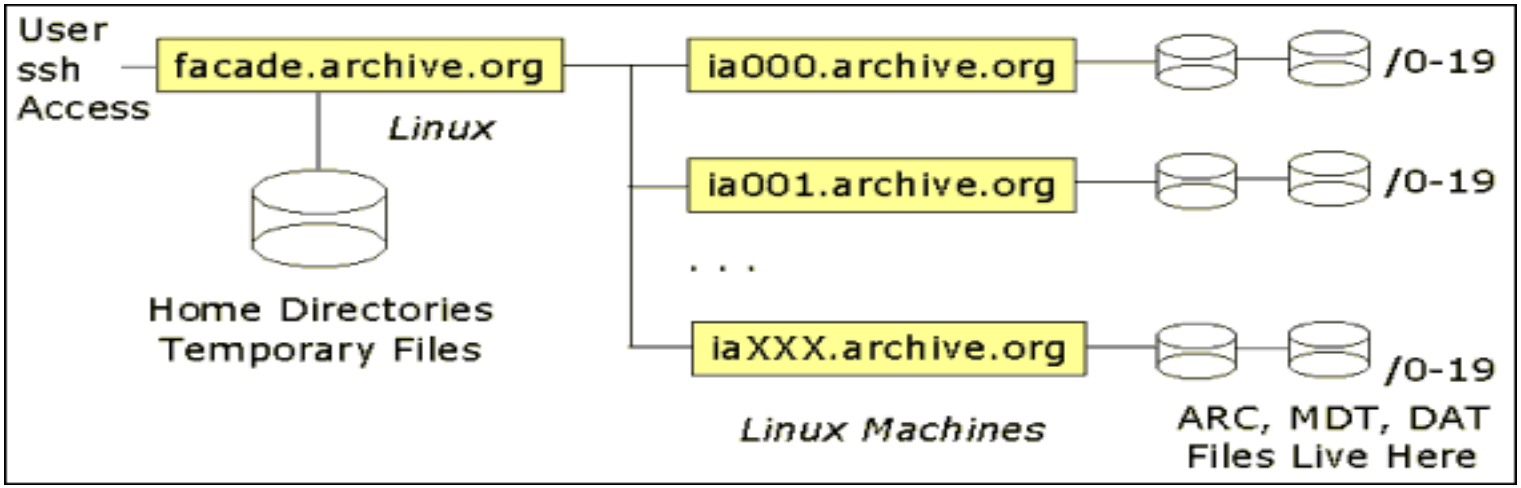
- Verwandte Projekte
  - Internet Archive
  - Kulturaw3
  - Nedlib

- Die österreichische Pilotstudie AOLA

Verwandte Projekte

Internet Archive

- seit 1996, URL: *www.archive.org*
- Reihe von Linux-Systemen mit Festplattenarrays



- Sammelt Datenbestände, die “gespendet” werden
- zum Großteil freies Harvesting (Alexa)
- ursprünglich nur Text, mittlerweile alle Datentypen

Verwandte Projekte

Internet Archive (2)

- März 2001: ca. 43 TB (März 2000: 14 TB)
- tägliches Wachstum: bis zu 100 GB pro Tag
- Redundanz: Verteilung auf mehrere physische Sites
- Automatische Migration auf neue Datenträger
- Sammlung von existierenden Emulatoren
- Zugriff beschränkt auf Forschungsinstitutionen
- “Programmierkenntnisse” erforderlich für Nutzung des Archivs

Verwandte Projekte

Kulturaw3

- URL: *http://kulturarw3.kb.se/html/kulturarw3.eng.html*
- Projekt der Schwedischen Nationalbibliothek, seit 1996
- Sun Sparc Stations mit Tape Roboter Archiv
- verwendet modifizierten Indexer (Combine) für Harvesting
- Snapshots des schwedischen Webs (.se, .nu, spezielle Domains)
- Preservation: Original plus evtl. Standardformate
- 5 Snapshots bisher, letzter Lauf:
  15 Mio. URLs von 58,400 Websites, insges. ca. 280 GB Daten
- Tendenz: stark steigend (unvollst. 6. Lauf: 360 GB !)
- Zugriffstool in Entwicklung
- Hierarchische Speichermedien geplant

Verwandte Projekte

NEDLIB

- URL: *http://www.kb.nl/coop/nedlib/*
- Networked European Deposit Library
- EU-Projekt Jänner 1998 bis Jänner 2001
- Framework zur Archivierung von On-line Medien
- Offen für alle Konzepte
- Guidelines, technische Standards, “best-practice” Modelle
- Nedlib Harvester für Archivierungszwecke
- Plattformen: Linux, SUN, ...
- Weitere Tools noch in der Entwicklungsphase
- Bis Projektende keine größeren Experimente durchgeführt

Verwandte Projekte

Weitere Projekte:

- 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*

Archivierung von On-line Medien

Überblick

- Technische Problemstellungen
  - Auswahl
  - Archivierung
  - Preservation
  - Zugriff
- Verwandte Projekte
  - Internet Archive
  - Kulturaw3
  - Nedlib
- Die österreichische Pilotstudie AOLA

AOLA

Austria On-line Archive

- URL: *http://www.ifs.tuwien.ac.at/~aola/*
- Kooperation zwischen Österr. Nationalbibliothek und Inst. f. Softwaretechnik, TU Wien
- Pilotstudie: Vorbereitung seit 1999, 1. Phase seit März 2001
- Linux-Rechner mit 240 GB Festplatten sowie 6-fach Tapewechsler
- Open Source Lösung
- Nedlib Harvester (incl. Modifikationen und Erweiterungen)
- Ziel: Snapshot des österreichischen Web

AOLA

Austria On-line Archive (2)

- seit 7. Mai 2001 ca. 10 Crawler parallel
- Download in der Testphase: ca 1GB pro Tag
- at-domains sowie ausgewählte Subdomains, insbesondere **.cc, *.com, *.edu,* etc.
- Statistik 7. - 10. Mai: (Exzerpt)

ac.at	*15781*	*294.645.263*	*ch*	*28*	*299.752*
co.at	*10091*	*99.281.054*	*com*	*13926*	*209.888.299*
gv.at	*4417*	*62.645.200*	*cx*	*6*	*108.502*
or.at	*5945*	*55.633.056*	*cz*	*50*	*230.132*
at	*176695*	*2.850.251.545*	*de*	*961*	*10.135.563*
au	*59*	*2.611.691*	*edu*	*29*	*1.207.928*
ca	*2*	*70.566*	*net*	*4013*	*73.352.782*
cc	*6739*	*168.581.373*	*org*	*1747*	*42.288.333*

AOLA - Ausblick

- Überführung der Pilotstudie in permanente Einrichtung
- Wiederholte Archivierung von Snapshots
- Entwicklung einer langfristigen Strategie zur Preservation und Migration
- Kombination von Transformations- und Emulationsansätzen
- Schaffung der technischen und personellen Infrastruktur
- Ausbau der Kooperation mit verwandten Projekten

Archivierung von On-line Medien

Zusammenfassung

- **Ziel**: Erhaltung des (modernen) Kulturerbes
- **Selektion**: Kombination von gezielter Auswahl und freiem Harvesting
- **Archivierung**: Migration von Speichermedien
- **Preservation**: Emulation und Transfer von Datenformaten
- **Zugriff**: Benutzerschnittstellen und rechtliche Aspekte
- **Dringlichkeit**: wir müssen **sofort** damit beginnen!

AOLA Projekt-Homepage:

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