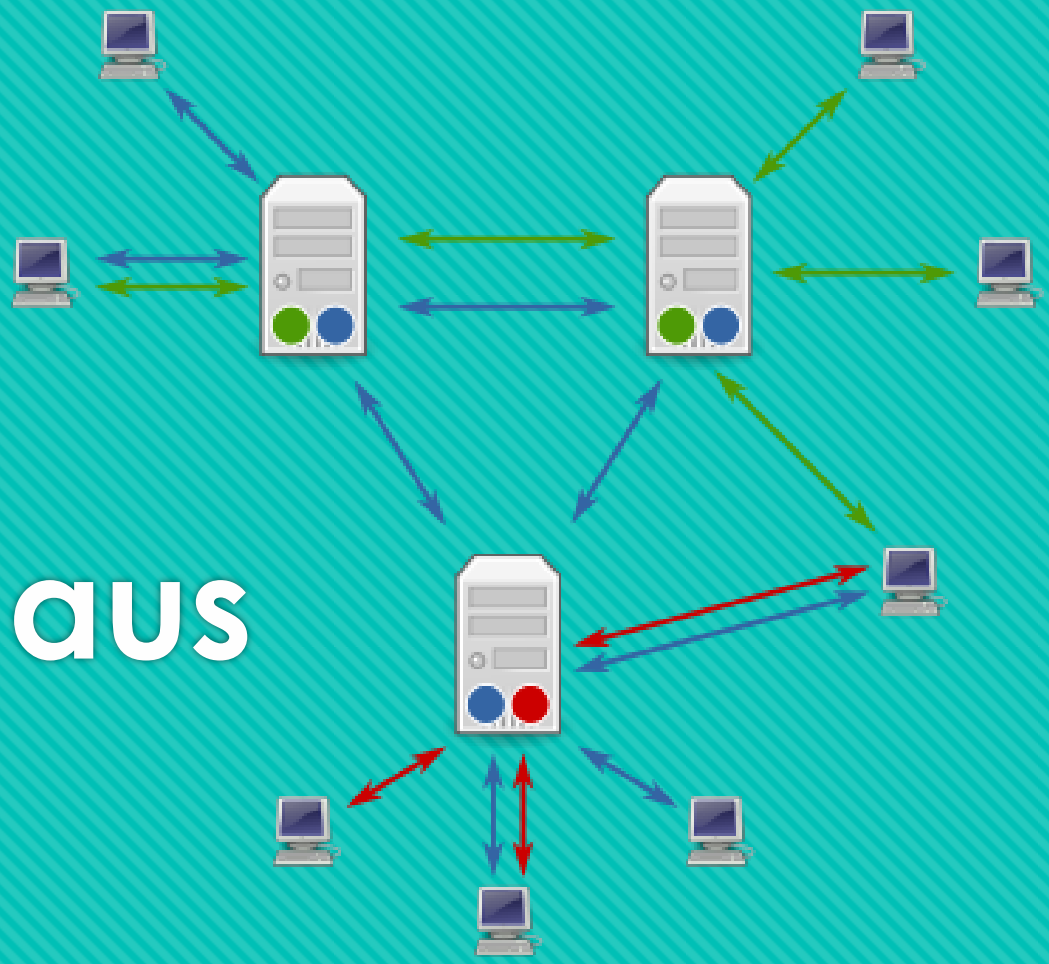
HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF



Oliver Kiechle

Historische Daten aus Webarchiven

Born Digital Sources | Usenet | Datenmanagement

Inhalt

- Born Digital Sources
- Webarchive
- Das Usenet
- Usenet Daten
- Datenmanagement

Born Digital Sources

- Definition:
  Digital entstandene Objekte, die digital verarbeitet, verwaltet und gespeichert werden
- Beispiele:
  Elektronische Dokumente, E-Mails, SMS, Tweets, Digitale Fotos und Videos, Webseiten..
  Aber auch: Programmcode, Logdaten, Trackinginformationen usw.

Born Digital Sources

- Kybernetisch, dynamisch, multimedial
- Original und Kopie
- Content vs. Metadaten
- Datenstandards, z. B. RFC 822

Webarchive

- Definition:
  Webarchive sammeln und speichern Inhalte des *World Wide Web*
- Erweiterte Definition? Zusätzlich:
  - Vorläufer des WWW
  - Deep Web
  - Social Media
- Webcrawler vs. Database Archiving vs. Manual Collection

Webarchive

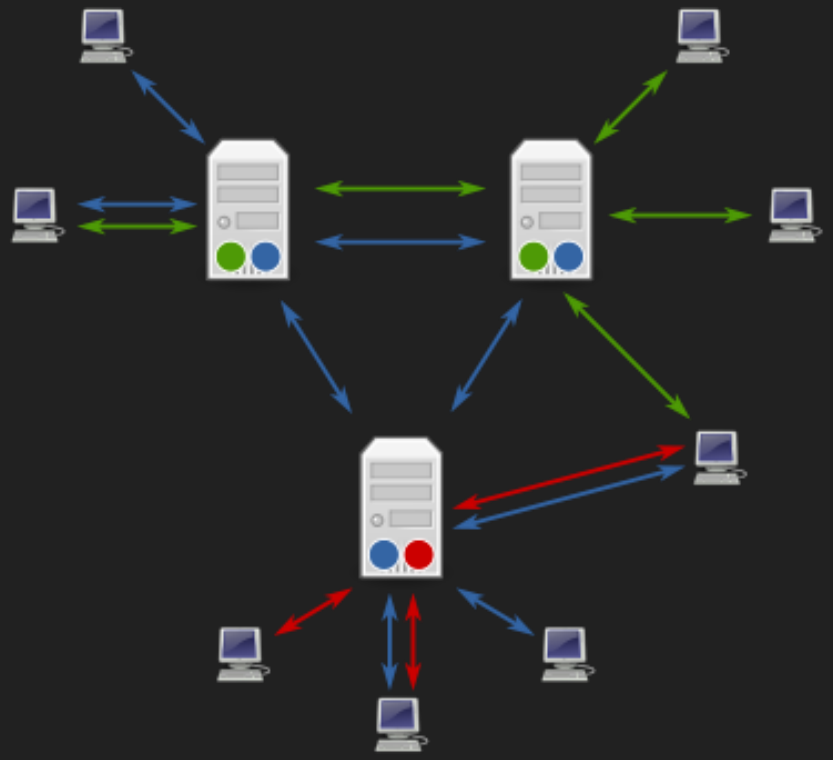
- Internet Archive (1996)
- Nationale Initiativen (z.B. Deutsche Nationalbibliothek)
- International Internet Preservation Coalition (IIPC) seit 2003
  - HERITRIX-Crawler
  - WARC (Web ARChive) Format

Das Usenet

- 1979 Struktur und erste Software ("A-News")
- Sommer 1980 **U**NIX U**se**r **Net**work ("Poor man's ARPANET")
- Themen: Computer, (Natur-)wissenschaft, Unterhaltung, später auch Gesellschaft und Politik
- 1986 Neue Hierarchiestruktur ("Major Seven/Big Eight")
- Eternal September 1993

Das Usenet

- Netzwerk von Newsservern mit textbasierter Kommunikation
- Weitergabe nach dem Flood-Fill-Prinzip; zunächst keine Langzeitspeicherung vorgesehen
- Grundlagen der Onlinekommunikation: Netiquette, FAQ, Spam, Flaming
- Heute vor allem zum Austausch von Binaries (Dateien) verwendet



UTZOO-Tapes

- Newsserver UTZOO (University of Toronto, Zoology Department)
- Seit 1981 Teil des Usenet
- Archiviert von Admin Henry Spencer
- 1991 Beginn der Übertragung auf Festplatten durch David Wiseman (University of Western Ontario) bis 2001

- 141 Magnetbänder
- Über 2 Mio. Nachrichten: Februar 1981 bis Juni 1991
- Nicht alle Newsgroups gespeichert
- (Geringe) Datenverluste durch Aufzeichnungsmedium

- 2001 in Google Groups integriert

Kommerzielle Archivierung

- CD-ROM-Ausgaben, z.B. Sterling Software (1992-1993), Infomagic Usenet (1994), Netnews offline (1995)
- Deja News ab 1995
- 2001 in Google Groups integriert

From: bill@carpet.WLK.COM (Bill Kennedy)
Newsgroups: news.admin
Subject: Re: minimum age for voting on new groups
Message-ID: <80@carpet.WLK.COM>
Date: 4 Jun 88 06:50:08 GMT
References: <441@wpg.UUCP> <346@comdesign.UUCP>
Organization: W.L. Kennedy Jr. and Associates

In article <346@comdesign.UUCP> pst@comdesign.uucp (Paul Traina) writes:
>From article <441@wpg.UUCP>, by russ@wpg.UUCP (Russell Lawrence):
[deleted some stuff]

Speaking only for myself, the three sites I administer [five I feed]

>Net "property owners"? Perhaps we should limit sufferage to TRUE net
>property owners -- those of us that:
>   (a) pay the phone bill   yes [no]
>   (b) actually own the hardware   yes [yes]
>   (c) pay for the maintenance on the hardware   yes [no]
>   (d) pay for the added disk drive that was needed to hold news yes [?]
>
>It would make the voting process much simpler for the vote taker :-), but

I answered your questions but you failed to describe your eligibility to
make the proposal. I feel eligible to criticize since I met your criteria.
Sure, I carry on and generally try to protect my economic turf since it's
mine and mine alone (in truth, lately I have been getting some help from my
neighbors with LD bills), but you don't establish your own position. Would
you qualify within your own conditions? If so, good! Let's proceed.

Rechtliche Aspekte

- Langzeitspeicherung und Suchfunktionen führen bereits Anfang der 1990er Jahre zu Protesten der User
- Autoren und Rechteinhaber können ab 1996 die Löschung von Beiträgen (zunächst bei Deja News, dann bei Google Groups) veranlassen
- Einführung des "X-No-Archive" Headers
- Anonymisierung problematisch
- Google Groups Crawling?
- US-Recht vs. Europäisches Recht

→ Publikation der Forschungsdaten?

Datenmanagement

- Dokumentation aller Bearbeitungs- und Auswertungsschritte:
  - Verschiedene Formate, Metadatenstrukturen, Speicherorte
  - Data Cleaning (z.B. mit Open Refine)
  - Übertragung in einheitliche Datenbankstruktur
  - Analyse (z.B. Topic Modelling) und Visualisierung (z.B. Netzwerkanalyse)
- → Repositorium
- → Publikation?

Vielen Dank für Ihre Aufmerksamkeit!