A Summary on the Internet Archive

By Akshat Kukreti
Introduction

• The Internet Archive (IA) contains a collection of historical websites of the world.
• Users can access the archived websites through the Wayback Machine.
• http://www.archive.org
• Founded by Brewster Kahle and Bruce Gilliat in 1996.
Motivation behind IA

• Preservation of Digital Information
• The average of a document on the internet is 75 days [Kahle 96], after that it is lost. “404 Document not found”
• Digital information is easier to store and search in.
• How authentic is a document found on the internet?
• Where to go from the current document?
• A Digital Library
What is currently archived

- Texts
- Audios
- Videos
- Software
Founders and Contributors

- People
  - Brewster Kahle
  - Bruce Gilliat
  - Rick Prelinger

- Institutions
  - Alexa Internet
    - Gathers and analyses data on web content and web usage
    - Uses its crawler to gather data on web content
    - Uses Alexa Toolbar for gathering data on web usage
    - Donates copies of web crawls to IA
Technology

• Capture Technology
  – The IA has developed tools for capturing web content
  – Crawler requirements:
    • Obeys the instructions given in a site’s robots.txt
    • Can run on multiple machines
    • Aggregates the crawl data into large files for easier management
  – Heritrix: A Java-based open source Web Crawler
Technology

• How the crawler works
  – Finds documents (files) based on seed URLs and downloads them to the Archive’s server
  – Looks for reference links and adds them to the list of files to be captured.
  – Relative paths are made absolute before being added to the list
  – The process is repeated for reference links
  – The crawler makes sure that the same page is not retrieved again.
  – Files that link frequently with other files are captured more frequently than files that rarely link to other files.
Technology

• What cannot be captured
  – Databases
  – Password protected files
  – Links found in JavaScript, Flash. May only capture homepage
  – Information restricted by the publisher
  – Successive changes made between crawls.

• IA gives the user a feel of what a page looked like at a given time but not the entire online experience.
Technology

- Storage and Preservation
  - Archive File Format (ARC, .arc extension)
  - Self identifying: No separate index required, easy integration into larger files
  - File Header: URL, size, content type, date and time of retrieval, Name of the Organization that retrieved it
  - Storage:
    - Tape used for the first 3 years
    - Petabox: Stores 1 Petabyte of data
  - Preservation:
    - Mirror sites in Alexandria (Egypt) and Amsterdam (Netherlands)
Technology

• The Wayback Machine
  – Introduced in October 2001, allows the user to find an instance of a web page.
  – User enters URL and is taken to a results page with dates when the capture was made.
  – The user can click on the date to see the version of the page
Technology

• The Wayback Machine contd..
  – The user can browse through domain and time.
  – The Wayback machine rewrites links to refer to archived pages instead of live ones.
  – Example:
    • web.archive.org/web/20050214202400/http://www.google.com
    • Captured in 2005, on Feb 14, at 20:24:00
  – If the user clicks a link, he is redirected to a link that was captured at a time nearest to the home page.
  – Simulates live web environment.
  – Works across multiple domains
Other Services

• Archive-It
  – Launched in 2005
  – Has over 50 members
  – Helps organizations that do not have the infrastructure or the expertise to preserve public web content.
  – Members can manage their collection by paying an annual fee.
  – Specific versions of the Wayback machine are used

• The Bookmobile
  – Gives away paperback books based on material from the archive
References
