Archiving

SNA Assignment

Due Dec 6th

via dropbox or in my mailbox (outside MGH420)

Why Archive?

Designing The Archive

- Identify Collection & Content to be Archived
- Perform Crawl
- Analyze Objects Crawled
- Develop and Deploy Interface

Ethics

- Suspend to the end... :-)

What to Archive?

- Web page?
- Web site?
- HTML?
- Image?
- Flash?
- What's the difference?

What's a web page made of?

- HTML
- CSS
- Links
- Images
- Javascript
- Some of these can be nested

Methods of Archiving

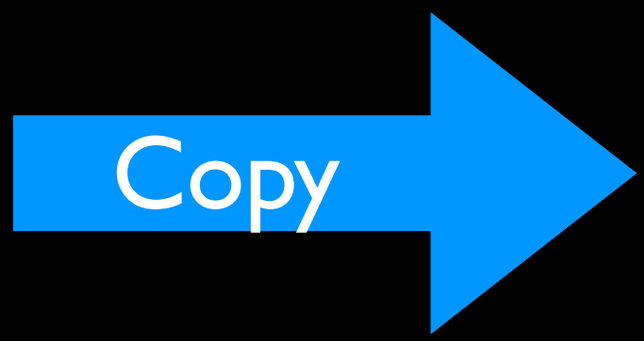
- Crawling (client-side)
  - i.e. like a browser
- Mirroring
  - Obtain a copy of the site or database
- Transactional archiving
  - Copy web server traffic

Crawling

- Request the page using a crawler, tool, or browser
- Process linked/included files
- How is this different than web crawling for indexing the web?

Mirroring

website.com



Backup Server/
Client

Transactional Processing

HTTP Requests

Capture This

website.com

Issues

- Ethics!
  - FB data archive example
  - TOS, robots.txt, etc.
- What should we save? How do we know if it's important?
- Browser technology

Tools

SurfULater

zotero

WebCite

HTTrack

wget

WEB CURATOR TOOL

.webarchive - Safari

LiWA

.mhtml - IE

Tools

- DiscoverText - http://discovertext.com/
- Twapper Keeper - http://twapperkeeper.com
- ContextMiner - http://contextminer.com/
- TubeKit - http://www.tubekit.org/

Assignment

- Archive using different tools (compare)
- Setting the bounds of an archive
- Ethics of archiving question