Informatics 225
Computer Science 221

Information Retrieval

Lecture 8

These course materials borrow, with permission, from those of Prof. Cristina Videira Lopes, Prof. Alberto Krone-Martins, Addison Wesley 2008, Chris Manning, Pandu Nayak, Hinrich Schütze, Heike Adel, Sascha Rothe, Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. Powerpoint theme by Prof. André van der Hoek.

Ethics and Law

Information Retrieval

Main point

Just because you can, doesn't mean you should.

Ways of Acquiring Web Data

- Data dumps
- Web APIs
- Targeted downloads

- Web crawling ← last option

Guidelines for bot writers

- An old (1993 !) but still very relevant text
  - http://www.robotstxt.org/guidelines.html

- Summary
  - **Reconsider** – do you really need a bot?
  - **Be Accountable** – if your actions cause problems, be available to take prompt action in response;
  - **Test Locally** – expand the scope gradually before you unleash your crawler on others;
  - **Don't hog resources** – web servers are for people primarily. Walk, don't run.
  - **Stay with it** – "it's vital to know what your robot is doing, and that it remains under control".

Guidelines for bot writers

- **Find out the sites' crawling policies**
  - GitHub (https://help.github.com/en/github/site-policy/github-acceptable-use-policies#5-scraping-and-api-usage-restrictions):

5. Scraping and API Usage Restrictions

Scraping refers to extracting data from our Service via an automated process, such as a bot or webcrawler. It does not refer to the collection of information through our API. Please see Section H of our Terms of Service and Corporate Terms of Service for our API Terms. You may scrape the website for the following reasons:

- *Researchers may scrape public, non-personal information from the Service for research purposes, only if any publications resulting from that research are open access.*
- *Archivists may scrape the Service for public data for archival purposes.*

You may not scrape the Service for spamming purposes, including for the purposes of selling User Personal Information (as defined in the GitHub Privacy Statement), such as to recruiters, headhunters, and job boards.

All use of data gathered through scraping must comply with the GitHub Privacy Statement.

- **Contact the Web admins**

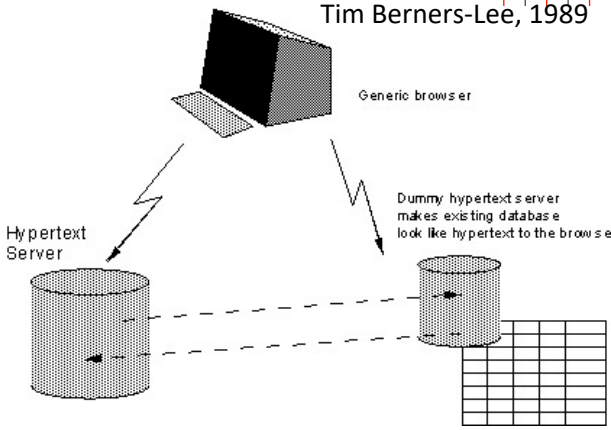
Ethical issues and law

The Web and the Law

- Web: ~30 years old only

The Web and the Law

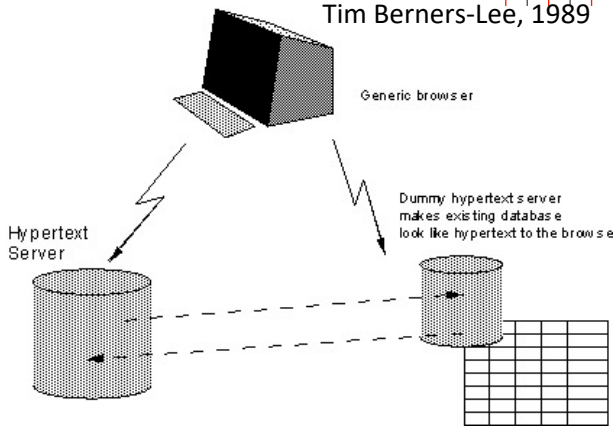
Tim Berners-Lee, 1989



- **Web: ~30 years old only**
- Born at CERN (old name: *Conseil Européen pour la Recherche Nucléaire*), for research in Physics:
  - "This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext system."
- Original proposal:
  https://www.w3.org/History/1989/proposal.html
- The last version (~1993) of the first website is still online:
  http://info.cern.ch/hypertext/WWW/TheProject.html

The Web and the Law

Tim Berners-Lee, 1989

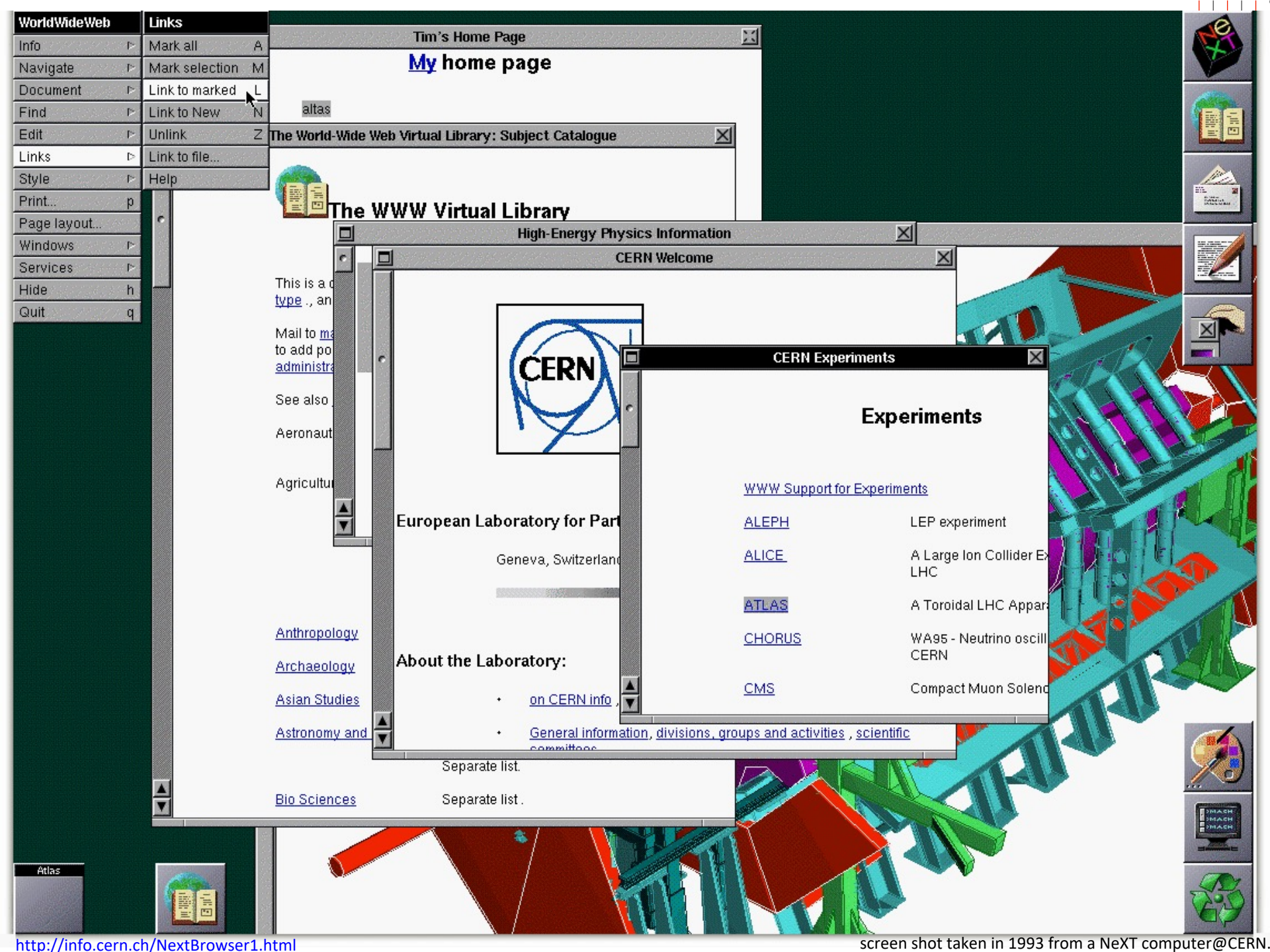


herche Nucléaire),

neral **information about**
sses the **problems of loss of**
nd derives a **solution** based on a

posal.html

is still online:
heProject.html

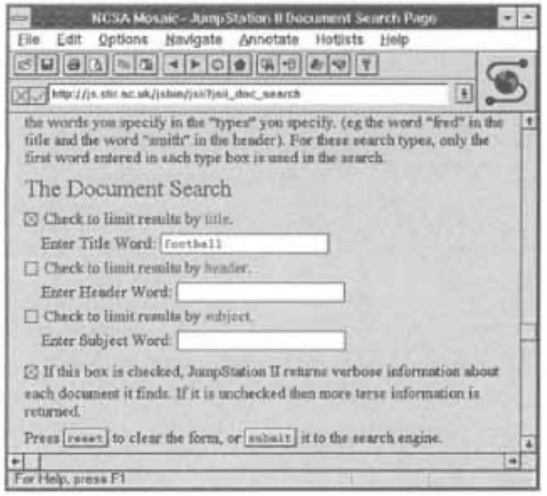


http://info.cern.ch/NextBrowser1.html

screen shot taken in 1993 from a NeXT computer@CERN.

First web search engines

- W3Catalogue (1993) : primitive catalogue
- World Wide Web Wanderer (1993) : *first web crawler to measure the size of the web*
- **JumpSearch (1993) : *crawling, indexing, searching. First "modern" engine.***
- Excite and RankDex (1994) : *ranking system* inspired Google's PageRank
- Webcrawler (1994) : *full text indexing*



The Web and the Law

- **Web: ~30 years old only**
- Wild Wild West (or *Wild Wild Web…*), for the most part

The Web and the Law

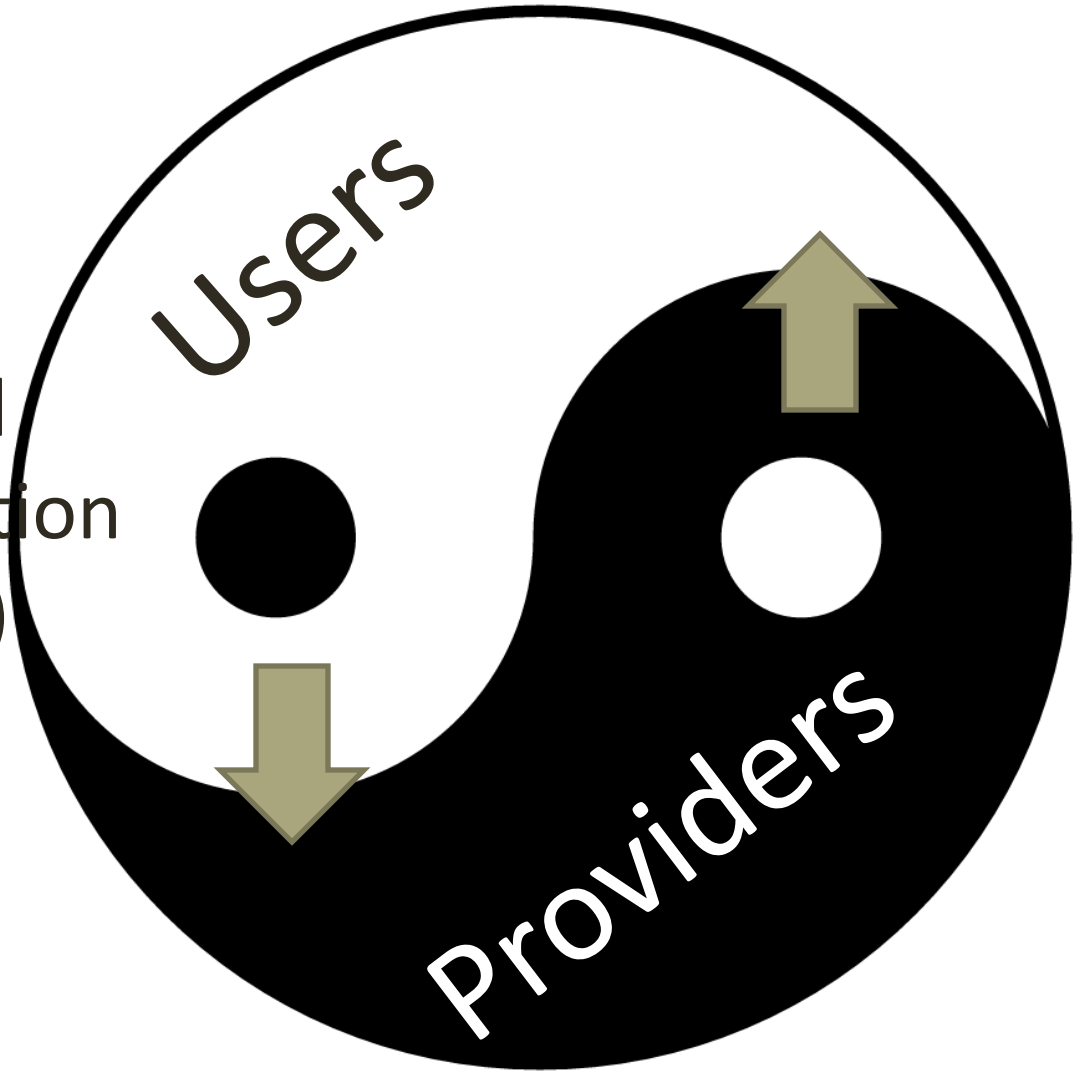
- **Web: ~30 years old only**
- Wild Wild West, for the most part
- Very few laws
- Existing laws either untested or outdated

The Web and the Law

- **Web: ~30 years old only**
- Wild Wild West, for the most part
- Very few laws
- Existing laws either untested or outdated
- **Our own judgments/actions matter**

Who can offend who

Personal Information (Privacy)



Content (Security)

Privacy Statements

- Web sites may have privacy statements:
  - What info they collect, how it's used

- E.g. http://uci.edu/privacy.php

Privacy : Apache logs

- 66.249.66.18 - - [25/Jan/2012:00:18:27 -0800] "GET /xwiki/bin/export/Stats/CurrentYearActivity?format=rtf HTTP/1.1" 404 335 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)”

- 180.76.6.28 - - [25/Jan/2012:00:18:31 -0800] "GET / HTTP/1.1" 200 45 "-" "Mozilla/5.0 (compatible; Baiduspider/2.0; +http://www.baidu.com/search/spider.html)”

- 66.249.66.18 - - [25/Jan/2012:00:19:02 -0800] "GET /calswim/coverage-s1.png HTTP/1.1" 200 961543 "-" "Googlebot-Image/1.0”

- 173.192.98.90 - - [25/Jan/2012:00:33:00 -0800] "GET / HTTP/1.1" 404 290 "-" "Python-urllib/2.4”

- 80.165.186.186 - - [25/Jan/2012:01:21:24 -0800] "GET /events-dataset-api/ HTTP/1.1" 200 3214 "http://www.kdnuggets.com/datasets/index.html" "Mozilla/5.0 (Windows NT 5.1; rv:9.0.1) Gecko/20100101 Firefox/9.0.1”

- 180.76.5.65 - - [25/Jan/2012:02:18:20 -0800] "GET / HTTP/1.1" 200 45 "-" "Mozilla/5.0 (compatible; Baiduspider/2.0;

Main point

Just because you can, doesn't mean you should.

Copyrights

- Products of intellect are ruled by “Copyright Law”
- Web content, source code, etc. are products of intellect
- See : http://www.copyright.gov
  - Next slides, from this site!

What is copyright?

- "Copyright is a form of protection grounded in the U.S. Constitution and granted by law for original works of authorship fixed in a tangible medium of expression. Copyright covers both published and unpublished works."

- *But then, can I use web material?*

https://www.copyright.gov/help/faq/faq-general.html

Fair use

- Various purposes for which the reproduction of a particular work may be considered fair:
  - Criticism
  - Comment
  - News
  - Teaching
  - Scholarship
  - Research

- It's a very grey area (see https://www.copyright.gov/fls/fl102.html)

Fair use

- Four factors to be considered in determining whether or not a particular use is fair:
  - The purpose and character of the use, including whether such use is of commercial nature or is for nonprofit educational purposes
  - The nature of the copyrighted work
  - **The amount and substantiality of the portion used in relation to the copyrighted work as a whole**
  - The effect of the use upon the potential market for, or value of, the copyrighted work

Legal cases

- eBay vs. Bidder's Edge
  - May 24, 2000
  - eBay successfully stopped crawlers from Bidder's Edge
  - "trespass to chattels"
  - Jury (https://law.justia.com/cases/federal/district-courts/FSupp2/100/1058/2478126/):
    - *Bidder's Edge intentionally and without authorization interfered with eBay's possessory interest in the computer system.*
    - *Bidder's Edge's unauthorized use proximately resulted in damage to eBay.*

Legal cases

- American Airlines vs. FareChase
  - March 10, 2003
  - American Airlines stops crawlers from Farechase for scraping AA.com website for "web fares" information (online fare comparison)
  - “trespass to chattels”
  - Jury (https://www.eff.org/document/preliminary-injunction-american-airlines-v-farechase-inc)
    - *Farechase’s actions are intentional and without American’s consent*
    - *Farechase’s conduct interferes with American’s computer network and system (…) such actions adversely affect and harm American and the condition, quality and value of American’s property.*
  - Settled suit later during that year.

Additional information on law

- *Computer Law graduate course:*
  - *https://www.ics.uci.edu/~kay/courses/269/*