The WebLab Project at Cornell University

Felix Weigel
Postdoctoral Associate
Dept. of Computer Science, Database Group
Cornell University

Could you tell ...

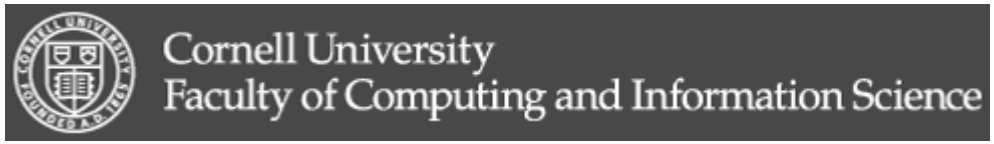
- When did the term “blog” emerge on the web?
- How do rumors spread across communities?
- How to model herd behavior in social networks?
- How do recall campaigns affect a company’s reputation?
- Who is usually cheaper, Amazon or Barnes & Noble?
- How does the web graph evolve over time?
- How to improve PageRank?

Mission of the WebLab Project

- The web is an extraordinarily rich source of information
- Web archives have saved snapshots of the web over many years

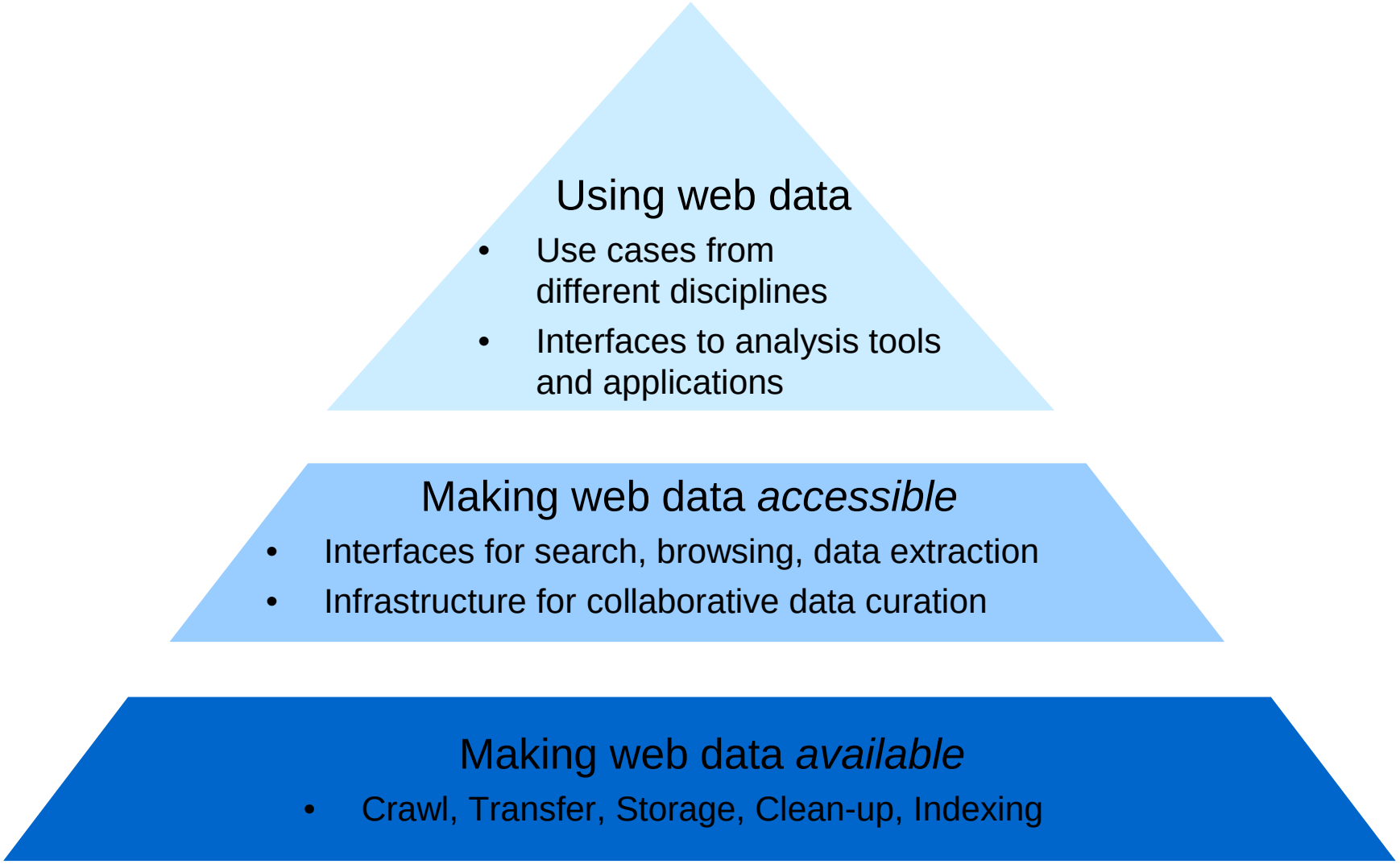
- Goals of the WebLab project:
  - Make this information accessible
    without demanding high technical or computing expertise
  - Provide an infrastructure for archived web data
    to enable all kinds of research about the web

The WebLab Project at Cornell



- Joint project of Cornell University and the Internet Archive
- Started in early 2006
- Participants at Cornell:
  - Computer Science Department
  - Information Science Program
  - Center for Advanced Computing (former Cornell Theory Center)
  - Collaboration with Cornell's Institute for the Social Sciences

Challenges


Using web data
• Use cases from different disciplines
• Interfaces to analysis tools and applications
Making web data *accessible*
• Interfaces for search, browsing, data extraction
• Infrastructure for collaborative data curation
Making web data *available*
• Crawl, Transfer, Storage, Clean-up, Indexing

Availability

- Crawling is done by the Internet Archive
- Compressed raw data is transferred to Cornell
  - Crawled web pages in *ARC Files*
  - Metadata (URL, title, crawl timestamp, ...) in *DAT Files*
- Three types of storage:
  - Tape archive for backups
  - Relational database for metadata
  - ARC files in file system
- Main challenge: *Scalability*
  - Clean-up: e.g., remove duplicate URLs and dangling links
  - Indexing: Hyperlink structure; full text
  - Parallelization: Off-line and on-line tasks on computing clusters

Accessibility

- Research begins with ... Search!
- What's in the archive?

Browsing Historical Web Pages with the Internet Archive's Wayback Machine

INTERNET ARCHIVE
WayBackMachine

Full-text Indexing with NutchWAX

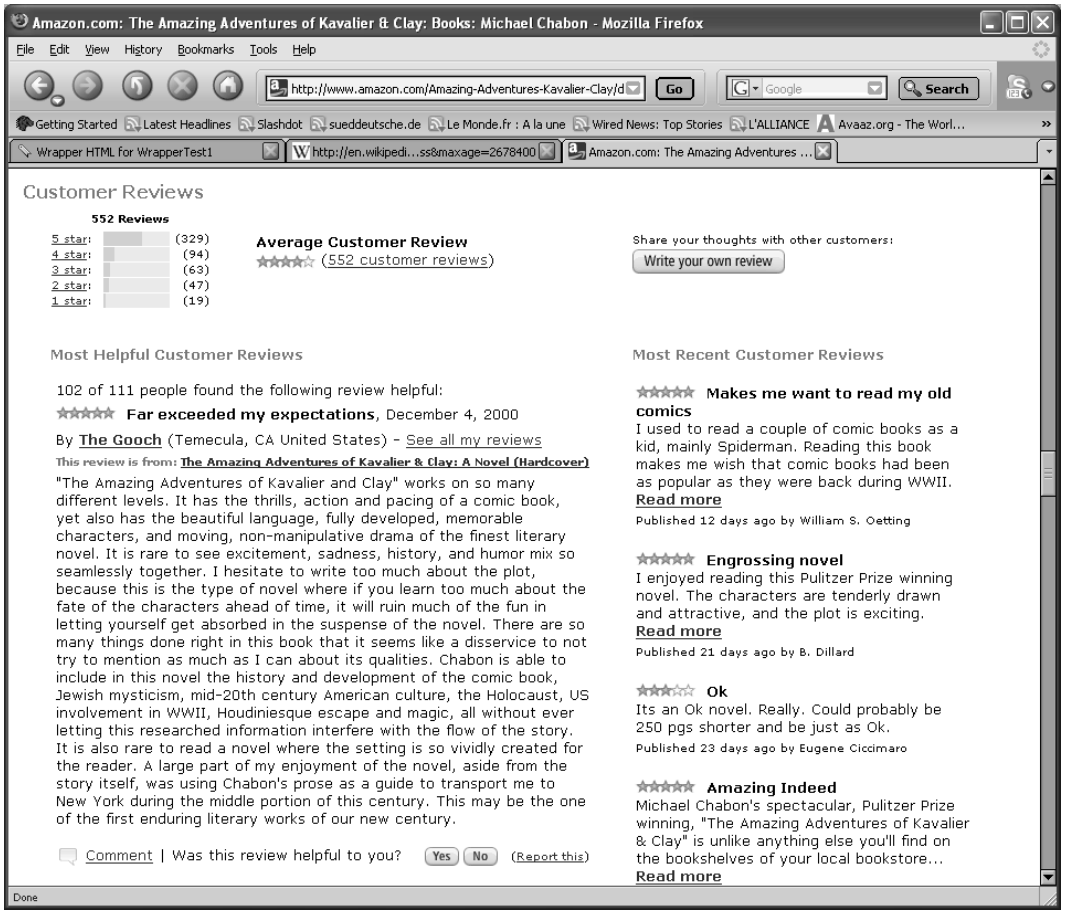
nutchwax

Metadata Search Form

Field	Include In Result	Negate Restriction	Restriction	Example of Query
PageID	☑	☐		LCnODk3L1brt8lIY34I2sA==
UrlID	☐	☐		AAGxclOHjQ7I9w/enYI8Lw==
CrawlID	☐	☐		1 2 3
HostID	☐	☐		1oQXLeTqZevjaDgANLj7qA==
URL:Protocol	☑	☐		http://
URL:Host	☑	☐		"%.edu" OR "%.org"
URL:Port	☑	☐		:80
URL:Path	☑	☐		%/pic/%
URL:Extension	☑	☐		.htm_
URL:QueryString	☑	☐		?no_article=52
ArchiveTime	☑	☐	(Begin) (End)	02/14/2005 08:30:22 NB: MONTH/DAY/YEAR
IPAddress	☐			
MIMEType	☐	☐		text/html
Language	☐	☐		en,us-ascii 0.54363121
Title	☐	☐		%photo%

Web Data Extraction

- Researchers often don't care about web pages, but specific substructures inside the pages
  - Blog postings
  - Online forums
  - Social tags/categories
  - News headlines
  - Tables of content
  - Bibliographies
  - Product details
  - Customer reviews



Web Data Collaboration Server

Data extraction

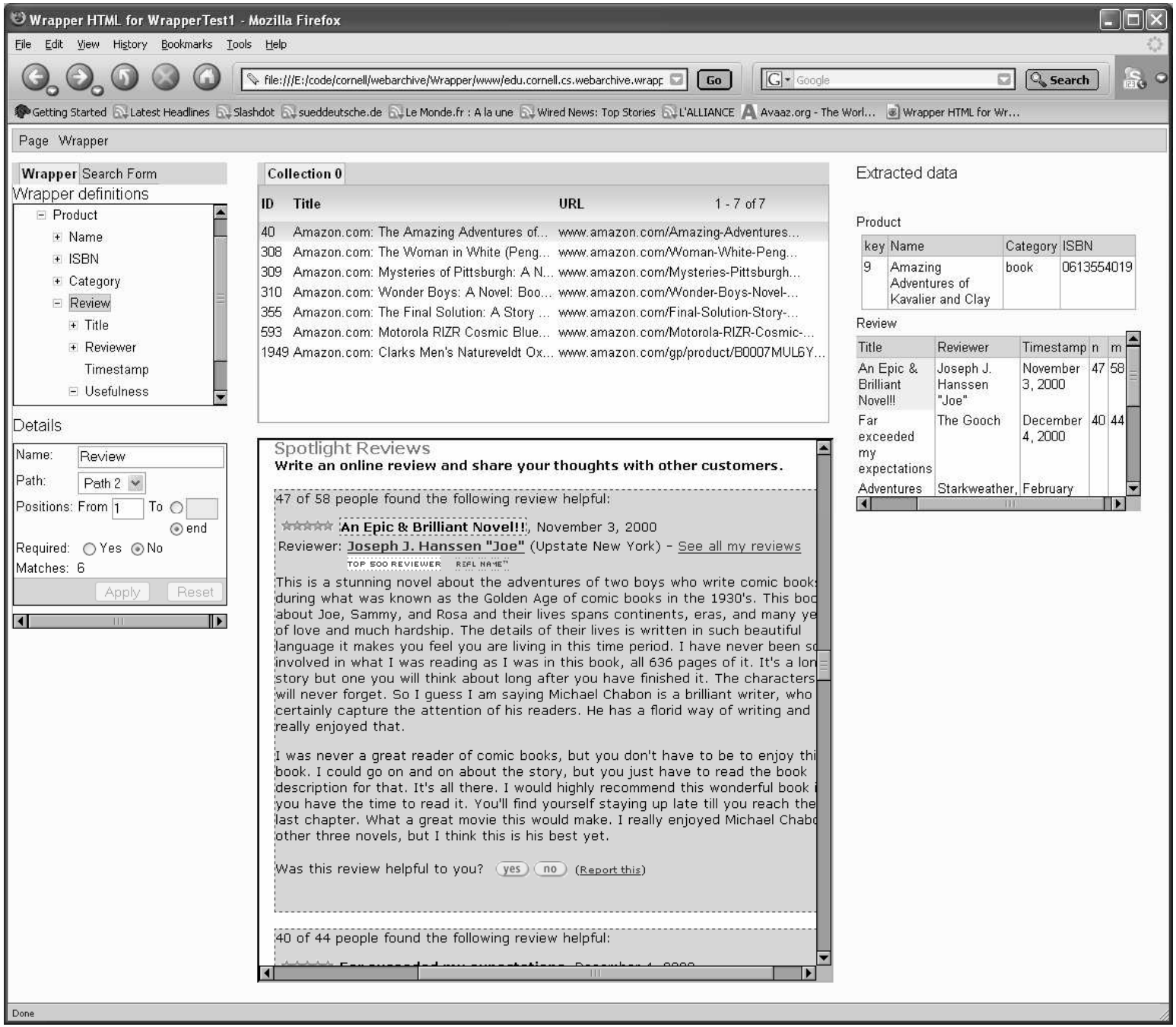
- Writing extraction code is a tedious task
- Create tools to make the data easily accessible in a structured format (e.g., tables in a database)

Data sharing

- Extracting the same data repeatedly is a waste of time and storage space
- Let users share their data and extraction rules

Data curation

- Web data is often incomplete and erroneous


Wrapper HTML for WrapperTest1 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
file:///E:/code/cornell/webarchive/Wrapper/www/edu.cornell.cs.webarchive.wrapp
Go
Google
Search
Getting Started
Latest Headlines
Slashdot
sueddeutsche.de
Le Monde.fr : A la une
Wired News: Top Stories
L'ALLIANCE
Avaaz.org - The Worl...
Wrapper HTML for Wr...
Page Wrapper
Wrapper Search Form
Wrapper definitions
Product
Name
ISBN
Category
Review
Title
Reviewer
Timestamp
Usefulness
Details
Name: Review
Path: Path 2
Positions: From 1 To
end
Required: Yes No
Matches: 6
Apply
Reset
Collection 0
ID Title URL 1 - 7 of 7
40 Amazon.com: The Amazing Adventures of... www.amazon.com/Amazing-Adventures...
308 Amazon.com: The Woman in White (Peng... www.amazon.com/Woman-White-Peng...
309 Amazon.com: Mysteries of Pittsburgh: A N... www.amazon.com/Mysteries-Pittsburgh...
310 Amazon.com: Wonder Boys: A Novel: Boo... www.amazon.com/Wonder-Boys-Novel-...
355 Amazon.com: The Final Solution: A Story ... www.amazon.com/Final-Solution-Story-...
593 Amazon.com: Motorola RIZR Cosmic Blue... www.amazon.com/Motorola-RIZR-Cosmic-...
1949 Amazon.com: Clarks Men's Natureveldt Ox... www.amazon.com/gp/product/B0007MUL8Y...
Spotlight Reviews
Write an online review and share your thoughts with other customers.
47 of 58 people found the following review helpful:
An Epic & Brilliant Novel!!, November 3, 2000
Reviewer: Joseph J. Hanssen "Joe" (Upstate New York) - See all my reviews
TOP 500 REVIEWER REAL NAME
This is a stunning novel about the adventures of two boys who write comic books
during what was known as the Golden Age of comic books in the 1930's. This boo
about Joe, Sammy, and Rosa and their lives spans continents, eras, and many ye
of love and much hardship. The details of their lives is written in such beautiful
language it makes you feel you are living in this time period. I have never been so
involved in what I was reading as I was in this book, all 636 pages of it. It's a lon
story but one you will think about long after you have finished it. The characters
will never forget. So I guess I am saying Michael Chabon is a brilliant writer, who
certainly capture the attention of his readers. He has a florid way of writing and
really enjoyed that.
I was never a great reader of comic books, but you don't have to be to enjoy thi
book. I could go on and on about the story, but you just have to read the book
description for that. It's all there. I would highly recommend this wonderful book i
you have the time to read it. You'll find yourself staying up late till you reach the
last chapter. What a great movie this would make. I really enjoyed Michael Chabo
other three novels, but I think this is his best yet.
Was this review helpful to you? yes no (Report this)
40 of 44 people found the following review helpful:
Extracted data
Product
key Name Category ISBN
9 Amazing Adventures of Kavalier and Clay book 0613554019
Review
Title Reviewer Timestamp n m
An Epic & Brilliant Novel!! Joseph J. Hanssen "Joe" November 3, 2000 47 58
Far exceeded my expectations The Gooch December 4, 2000 40 44
Adventures Starkweather, February
Done

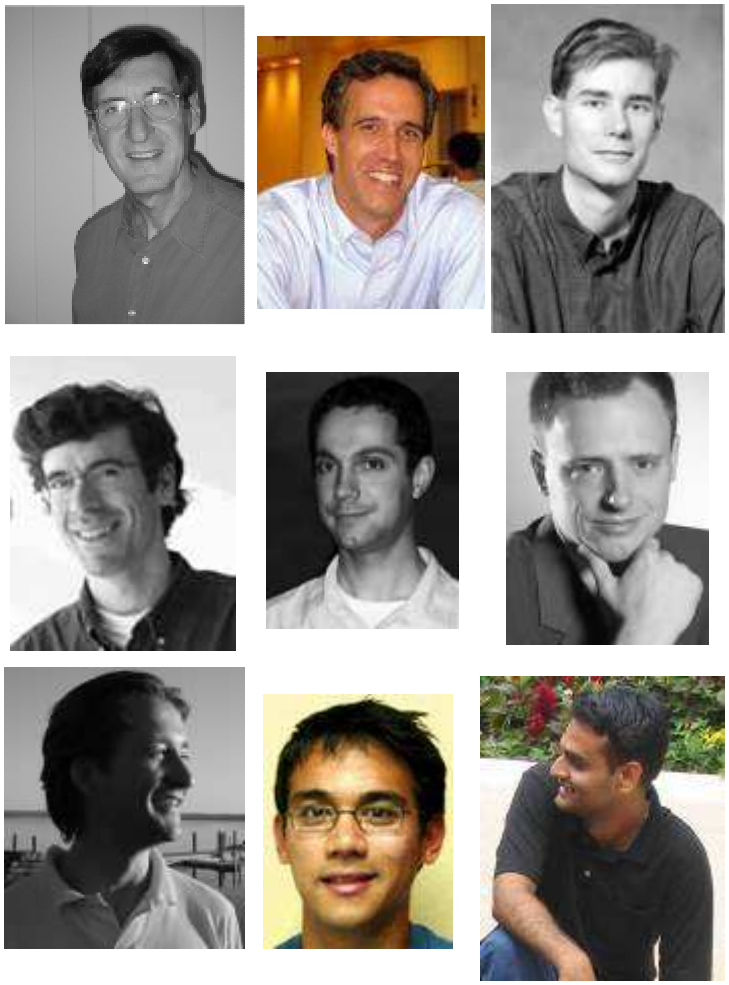
Project Status

- Availability:
  - Infrastructure for (partly) automated crawl download and backup
  - Metadata database of two full crawls
    - 3 billion pages, 20 billion distinct URLs, 130 billion links, 11.5 TB on disk
  - Full-text indexing of two Amazon.com sub-crawls in progress
    - 30 million pages, 164 GB compressed
- Accessibility:
  - Prototype of the Web Data Collaboration Server in progress
  - First deployment to Social Scientists at Cornell later this year
  - No-code web data extraction and sharing will soon be operational
- Usability:
  - Research on exploration of extracted data
  - Novel Methods for “informal” database search in progress

Who's Involved in the WebLab Project

Cornell

- William Arms
- Selcuk Aya
- Manuel Calimlim
- Pavel Dmitriev
- Johannes Gehrke
- Dan Huttenlocher
- Jon Kleinberg
- Christoph Koch
- Blazej Kot
- David Lifka
- Ruth Mitchell
- Biswanath Panda
- Mirek Riedewald
- Lucia Walle
- Felix Weigel



Internet Archive

- John Aizen
- John Berry
- Kris Carpenter
- Tracey Jacquith
- Brewster Kahle
- John Lee
- Gordon Mohr
- Michael Stack