



# The Web Laboratory

## A Research Environment for Studying the Web

Pavel Dmitriev,  
Cornell University

*MSRA, January 5, 2006*

A project of Cornell University and the Internet Archive

<http://www.cs.cornell.edu/wya/weblab/>



# Outline

---

---

- **Background & Motivation**
- **Resources**
- **The Big Picture**
- **Architecture Details**
- **Current Status**

# The Internet Archive

The screenshot shows the Internet Archive website in a Mozilla Firefox browser window. The browser title is "Internet Archive - Mozilla Firefox". The address bar shows "http://www.archive.org/". The website header includes the Internet Archive logo, navigation links for "Web", "Moving Images", "Texts", "Audio", "Software", "Patron Info", and "About IA", and the slogan "Universal access to human knowledge". Below the header is a search bar with "All Media Types" selected and a "GO!" button. The main content area is divided into three columns. The left column has "Announcements" and "This Just In" sections. The middle column features "Archive Collections" with a description of the digital library and a "Browse the Archive" link. The right column contains the "Wayback Machine" logo and a "Take Me Back" button. The footer of the website lists various collections like "Prelinger Archives" and "Open Source Movies".

Internet Archive - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.archive.org/

NSDL RRS

INTERNET ARCHIVE

Web | Moving Images | Texts | Audio | Software | Patron Info | About IA

Universal access to human knowledge

Home Forums | FAQs | Contributions | Terms, Privacy, & Copyright | Contact | Jobs

Search:  All Media Types GO! Anonymous User (login or join us)

Advanced Search

**Announcements** (more)

[Copyright law and Orphans: Suggested solution](#)

[20,000 Live Music Archive Concerts!](#)

[Orphan Works lawsuit appeal filed](#)

**Archive Collections** RSS

The Internet Archive is building a digital library of Internet sites and other cultural artifacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, and the general public.

[Browse the Archive](#)

waybackMachine

http://  Take Me Back

[Advanced Search](#) | [About the Wayback Machine](#)

**This Just In** RSS (more)

[Pugs and Pols - Golden Gloves](#)  
37 minutes ago

[Moving Images](#): [Prelinger Archives](#) | [Open Source Movies](#) | [Feature Films](#) | [Game Videos](#) | [Computer Chronicles](#) | [Net Café](#) | [Election 2004](#) | [Independent News](#) | [Youth Media](#) | [SIGGRAPH](#) | [MSRI Math Lectures](#) | [Open Mind](#) | [Shaping San Francisco](#) |

Done

# The Internet Archive Web Collection

---

---

## The Data

Complete crawls of the Web, every two months since 1996, with some gaps:

- Range of formats and depth of crawl have increased with time
- No sites that are protected by robots.txt or where owners requested not to be archived
- Some missing or lost data

## Sizes

- Current crawls are about 50-60 TByte (compressed)
- Total archive is about 600 TByte (compressed)

# Problem with the Internet Archive Collection

---

---

## Very difficult to use even for Computer Scientists

- System architecture is optimized for storage, not for processing
- No convenient API
- No computing resources

# Goal of the Web Laboratory

---

---

- Provide a mirror for Internet Archive data
- Build a set of tools to support research projects dealing with that data, with special focus on projects from social sciences and humanities

# Motivation: Social Science Research

---

---

## **The Web as a social phenomenon**

Political campaigns

Polarization of opinions

## **The Web as evidence**

The spread of urban legends ("Einstein failed mathematics")

Development of legal concepts across time

# Motivation: Research on Web Structure

---

---

## The Web Graph

Structure and evolution of the Web graph

Hubs & authorities, PageRank, etc.

Social networks

---

*Many of the basic studies have only been done once*

*Few if any large-scale studies across time*

# User Requirements Study

---

---

## Subsets across

- Time
- Web site / Domain name / URL structure
- Topic / Trend / Characteristics

## Search on

- Metadata
- Content

## Link Graphs for

- Full crawls
- Subsets

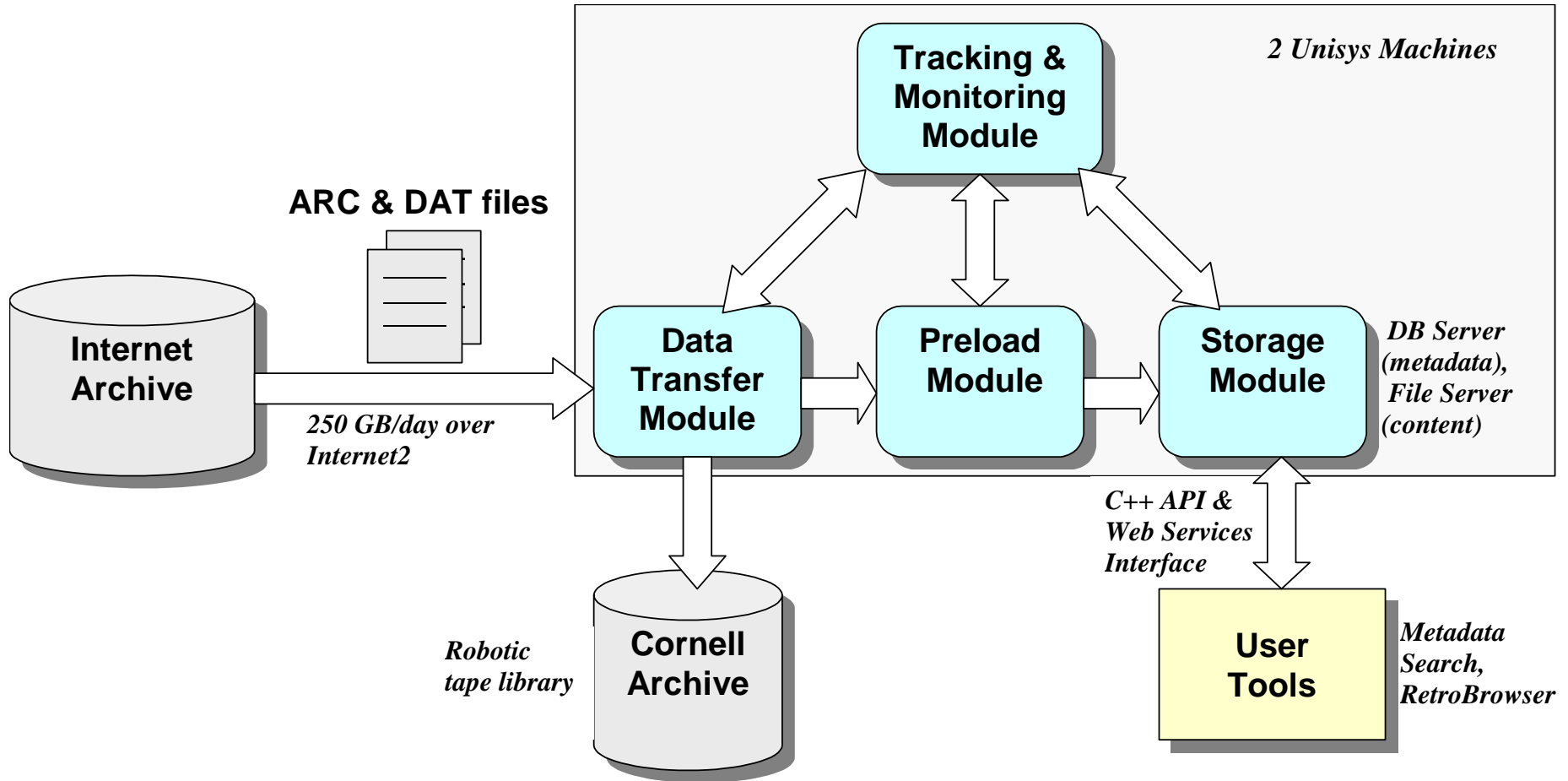
# Resources

---

---

- Two Unisys ES7000 servers with 16 64-bit Itanium 2 processors and 64 GB of RAM, running MS Windows Server 2003
- Two 50 TByte RAID Online Storage Systems (plan to add more in the next 6 month)
- ADIC Scalar 10K robotic tape library for archive
- A Web server

# The Big Picture



# Outline

---

---

- *Background & Motivation*
- *Resources*
- *The Big Picture*
- **Architecture Details**
  - **Preload**
  - **Storage**
  - **User Tools**
- **Current Status**

# Preload

---

---

## Input

- ARC files: store content, ~100 MB (compressed) each
- DAT files: store metadata for ARC files (outlinks, mime-type, crawl time, IP address, etc.), ~15 MB each

## Output

- Content files: one for each ARC file, content duplicates detected using Bloom filter (expect big savings from it)
- DB load files: one for each table, ~ 40GB each)

## Implementation

- ARC and DAT files are processed independently
- Can process ~1 TB/day of DAT files or ~6 TB/day of ARC files

# Storage

---

---

## File Server

- One Unisys machine, stores compressed content in files, 1 file per ARC file

## DB Server

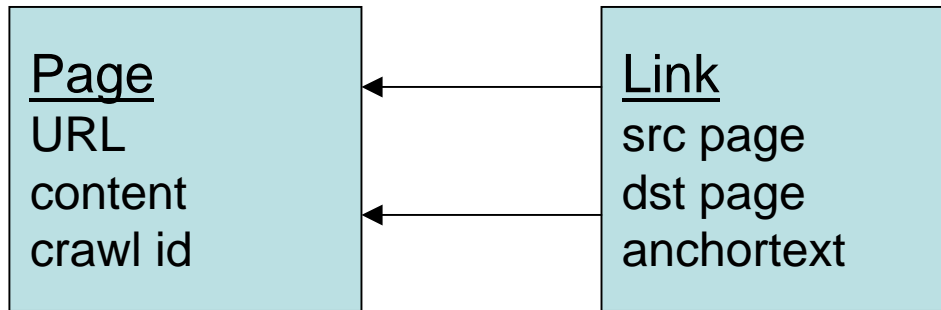
- MS SQL Server 2000 on the second Unisys machine
  - Requirements:
    - Decrease the amount of data stored as much as possible
    - Allow efficient execution of common queries
    - Provide fast loading of data
    - Facilitate incremental backup
    - Minimize transaction and logging expenses
- 
-

# Database Schema

---

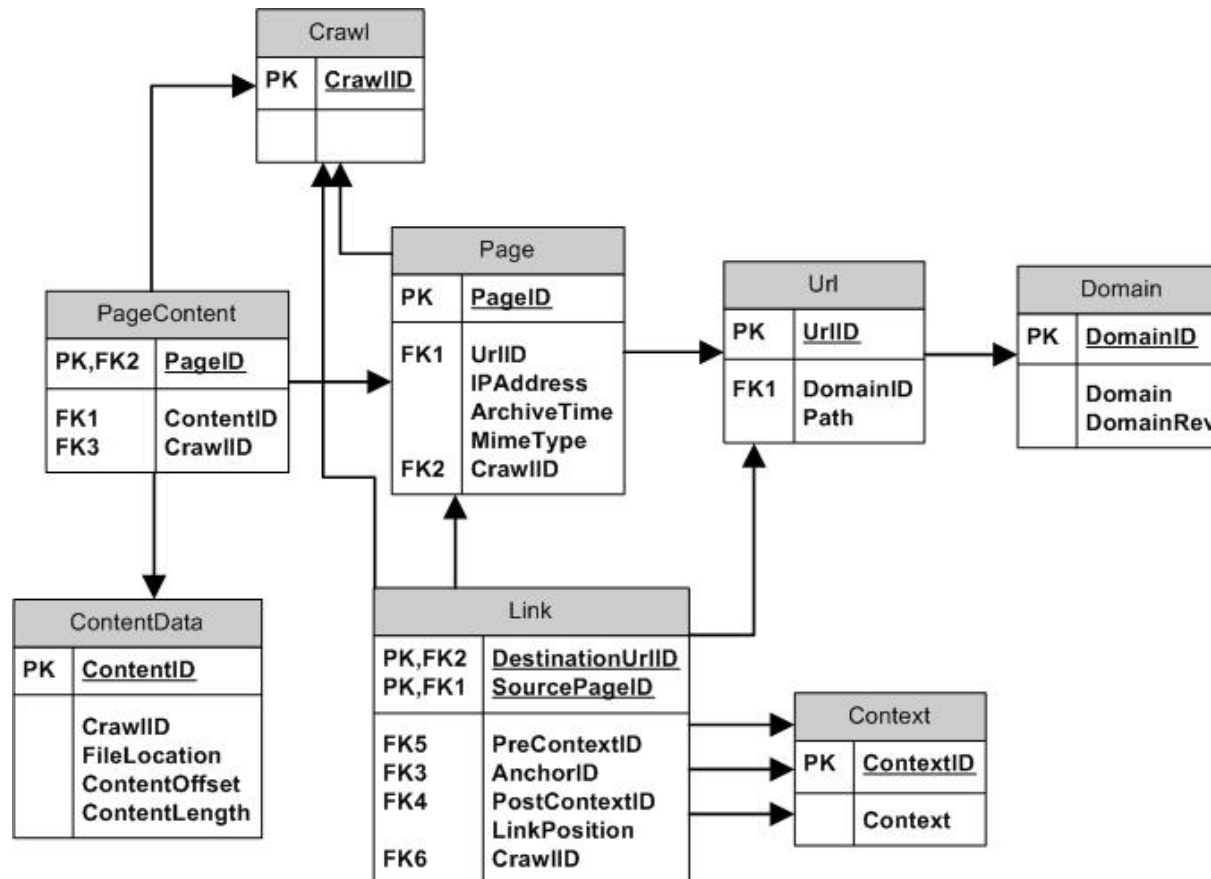
---

## Simple Schema



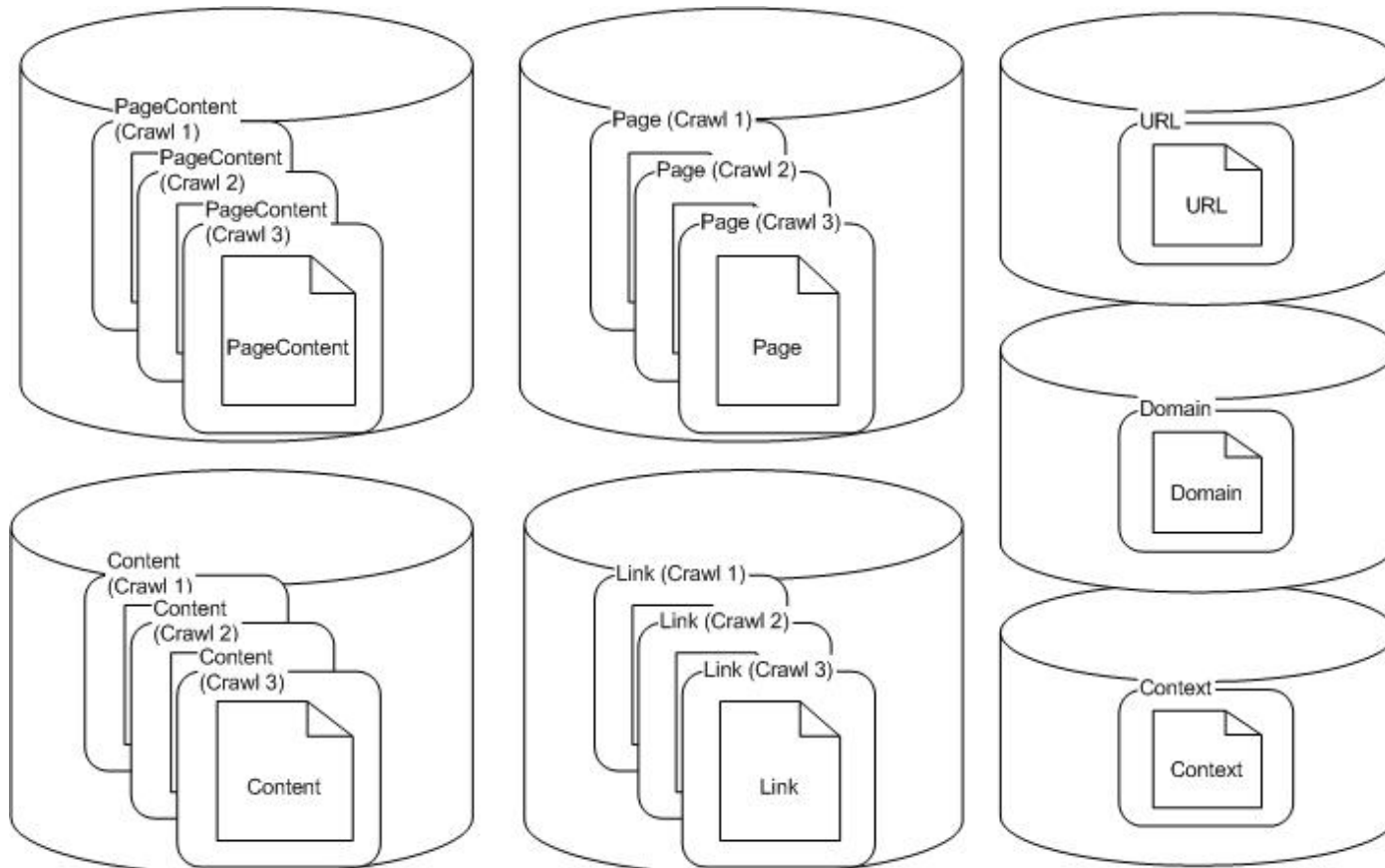
# Database Schema

## Conceptual Schema



# Database Schema

## Physical Schema



# Logging and Recovery

---

---

- Simple recovery model is too expensive
  - Log files grow very large
  - Checkpointing slows down loading process
- Use Bulk Logged recovery model
  - Allows manual control over logging and backups
  - Requires manual recovery in case of a crash

---

*Ongoing experimentation in this area*

# DB Load Performance

---

---

- ~1 TB/day using 4 processors (optimal number)
- Optimal load file size is ~40 GB
- ~ 1:1.25 storage ratio

# User Tools

---

---

## Subset extraction

- Subset is a virtual view generated from a user query
- User receives an identifier that he/she can use to work with the subset

## Metadata Search Service

- Allows retrieving metadata for pages from crawls and/or specified subsets across time

## Retro Browser

- Allows pseudo browsing through past crawls
- Our implementation of the Way Back Machine

# Metadata Search – Query

## Web Research Infrastructure

Enter search terms (domain):

Ex: .gov : All domain names ending with .gov

 OR 

Enter search terms (url):

Ex: \*/~.\*html\$ all urls which have a ~ in them and end with html. This will match www.cs.cornell.edu/~scs49/index.html but w

  
 Find pages that link  this page

Enter date range:

Ex: 1999-2001, 2003 : all pages from the crawls between 1999 to 2001 and those in 2003.

File format:

To select multiple formats, enter values seperated by commas in textbox.

 OR 

Results

Ex: all, 1000, bottom 500. 1500-3000 shows results ranked 1500-3000

Retrieve

Ex: to perform a join of this query with id 345, select join and enter 345

Note: Regular expressions are allowed in all fields. Quick ref:

Retrieve  
Union  
Minus  
Intersection

# Metadata Search – Result

---

---

## Web Research Infrastructure - Results

Your query :

```
domain: .gov
url: .html$
date: 2001
```

resulted in **1,269** hits.

Query id (for future reference): **3971**

Results (1-10):

*Click headers to sort*

<b>Crawl date</b>	<b>URL</b>
12 Jun 2001	<a href="http://www.nasa.gov/index.html">www.nasa.gov/index.html</a> <i>See all versions of this page</i>
01 Apr 2001	<a href="http://www.virginia.gov/index.html">www.virginia.gov/index.html</a> <i>See all versions of this page</i>
15 Aug 2001	<a href="http://www.kentucky.gov/index.html">www.kentucky.gov/index.html</a> <i>See all versions of this page</i>
...	

---

Previous 10

[Next 10](#)

---

[Back to query page](#)

# Current Status & Future Plans

---

---

## Current Status

- 250 GB/day transfer started at the end of November 2005
- All modules support most of the discussed functionality, and tested independently with small amounts of data (up to 100 GB)
- User Tools and APIs work within local network

## Future Plans

- Run the system in production mode early next semester
  - Open the system for external access (basic API, and retro browsing)
  - Use student course projects to test the system with real workload
  - Have 1 complete crawl and metadata from at least 2 more crawls available by the end of the semester
- 
-

# The Cornell Team

---

---

## **Researchers:**

William Arms, Dan Huttenlocher, Jon Kleinberg

## **Cornell Theory Center Staff:**

Dave Lifka, Ruth Mitchell, Lucy Walle

## **Ph.D. Students:**

Selcuk Aya, Pavel Dmitriev, Blazej Kot

## **M.Eng. & Undergraduate Students:**

Samuel Benzaquen, Nick Gerner, Min-Daou Gu, Wei Guo, Parul Jain, Serena Kohli, Lipi Sanghi, Shantanu Shan, Dmitriy Shtokman, Megha Siddavanahalli, Swati Singhal, Chris Sosa, Harsh Tiwari

# Using the Web Laboratory

---

## Using the laboratory

If you would like to use the Web Laboratory for your research, please contact me. We are looking for users and collaborators.

## Question to the audience

What kind of research would you like to use the Web Laboratory for? What kind of services would you like to see?

The order in which we build the services will be decided by the demands of the users.

# Thanks

---

---

This work would not be possible without the forethought and longstanding commitment of the Internet Archive to capture and preserve the content of the Web for future generations.

The Web Laboratory project is funded in part by National Science Foundation grants 0403340, 0127308, and 0537606 with equipment support from Unisys.

The Cornell Theory Center's support for this project is funded in part by Microsoft, Dell and Intel.



# The Web Laboratory

## A Research Environment for Studying the Web

Pavel Dmitriev,  
Cornell University

*MSRA, January 5, 2006*

A project of Cornell University and the Internet Archive

<http://www.cs.cornell.edu/wya/weblab/>

