



# Question Answering

## Classification and Analysis of Semi-Structured Government Data

Matthew King, Mike Branstein and Ryan Coleman

Faculty Advisor: W. Bruce Croft

Project Leader: David Pinto

QA System Software Engineer: David Fisher

Graduate Student: Wei Li

9 August 2001

# What is Question Answering?

- QA is a computer-based activity that involves searching large quantities of text and understanding both questions and textual passages to the degree necessary to recommend a text fragment as an answer to a question.



# The Previous QA System

- Used an InQuery database of plaintext news articles
- Matches types of questions to types of answers. Example:
  - “Who” tries to match to a person.
  - “How much” tries to match to a number.
- Scores each document.
- Returns an answer and the passage containing that answer.



# Our Hypothesis

- We will be able to answer questions better
  - Using a more statistical dataset
  - Sectioning documents
  - Extracting table data

# Steps Involved

- Web Crawl
- Hub/Authority Algorithm (reducing the dataset)
- Finding *METADATA* in the Documents
- Searching the *METADATA*



# Building a Dataset: Goal & Importance

- Wanted to concentrate more on statistical data than on ‘news’ events
- Started from [www.fedstats.gov](http://www.fedstats.gov), a site containing links to other government and non-government statistical sites
- Hopefully would be able to answer factual questions not answerable by other QA systems

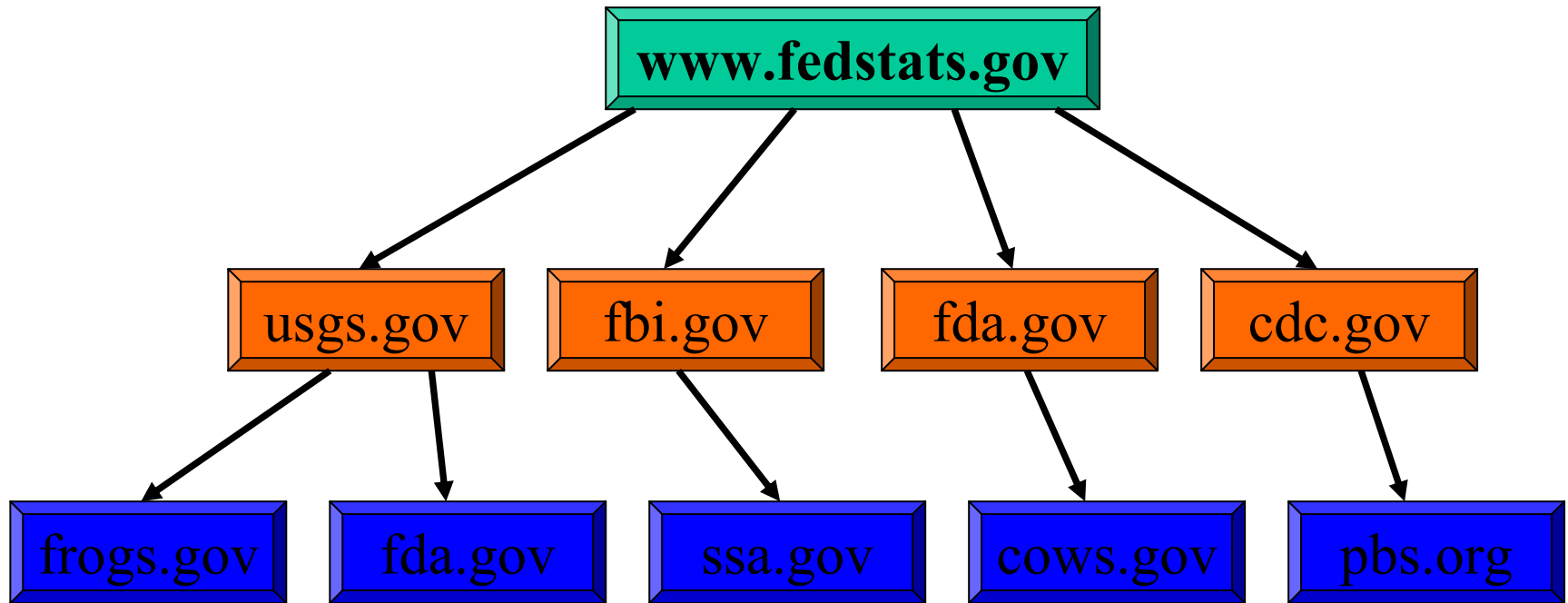


# How we accomplished this

- Focused on retrieving as many documents as possible – extracting links from document as much as possible
  - Extracted links from ‘typical’ link reference tags
  - Extracted links from web forms
  - Resolved http-redirects to retrieve more pages
- Overfetched – i.e. retrieved documents we would later throw out rather than examine them in place
- Preserved HTML structure of documents

# ‘Tiers’ of Web Crawl

- Downloaded web documents limited by domain name
- Executed crawl in ‘tiers’ – each tier consisted of the next layer of domains linked to by the previous layer



# Web Crawl Results

Level	Documents retrieved limited by following domains	Documents retrieved	Total size of documents retrieved
<b>1<sup>st</sup> tier</b>	<p><b>www.fedstats.gov</b></p> <p><b>1 domain</b></p>	<b>64,514</b>	<b>1.8 GB</b>
<b>2<sup>nd</sup> tier</b>	<p><b>Any site linked from www.fedstats.gov</b></p> <p><b>3578 domains</b></p>	<b>143,724</b>	<b>3.1 GB</b>
<b>3<sup>rd</sup> tier</b>	<p><b>Any site linked from 2<sup>nd</sup> tier</b></p> <p><b>16055 domains</b></p> <p><small>*actually incomplete due to time constraints</small></p>	<b>1,408,876</b>	<b>35 GB</b>

# Steps Involved

## √ Web Crawl

- Hub/Authority Algorithm (reducing the dataset)
- Finding *METADATA* in the Documents
- Searching the *METADATA*



# Document Classification

- Hub (Linking) pages
  - high link count
  - small amount of “useful” information
  - “error” pages
- Authoritative pages
  - low link count
  - large amounts of “useful” information



# Example: Hub Page

U.S. Census Bureau

107th Congress - California

---

[State Profile](#)

[U.S. Senate](#)

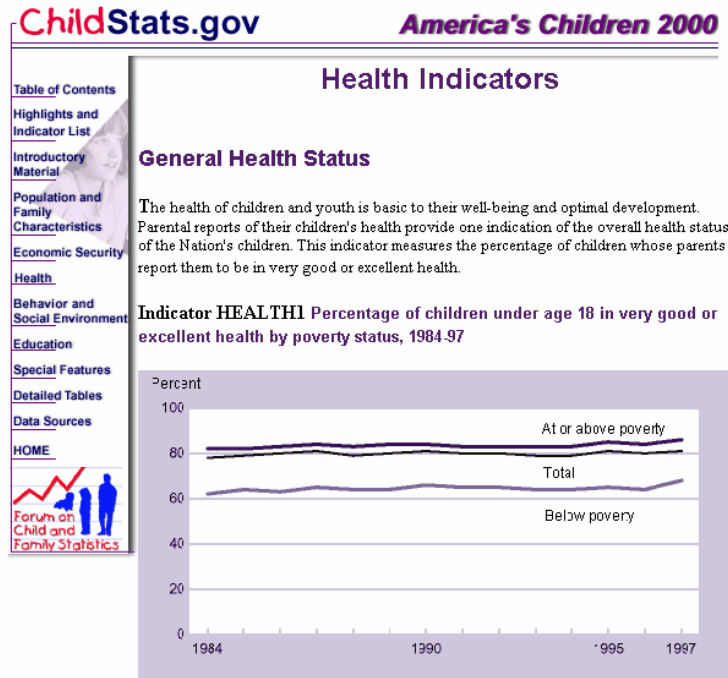
[Senator Barbara Boxer \(D\)](#)

[Senator Dianne Feinstein \(D\)](#)

**Congressional Districts**

- [District 1](#)
- [District 2](#)
- [District 3](#)
- [District 4](#)
- [District 5](#)
- [District 6](#)
- [District 7](#)
- [District 8](#)
- [District 9](#)
- [District 10](#)
- [District 11](#)
- [District 12](#)
- [District 13](#)
- [District 14](#)
- [District 15](#)
- [District 16](#)
- [District 17](#)
- [District 18](#)
- [District 19](#)
- [District 20](#)
- [District 21](#)
- [District 22](#)
- [District 23](#)
- [District 24](#)
- [District 25](#)
- [District 26](#)
- [District 27](#)
- [District 28](#)
- [District 29](#)
- [District 30](#)
- [District 31](#)
- [District 32](#)
- [District 33](#)
- [District 34](#)
- [District 35](#)
- [District 36](#)
- [District 37](#)

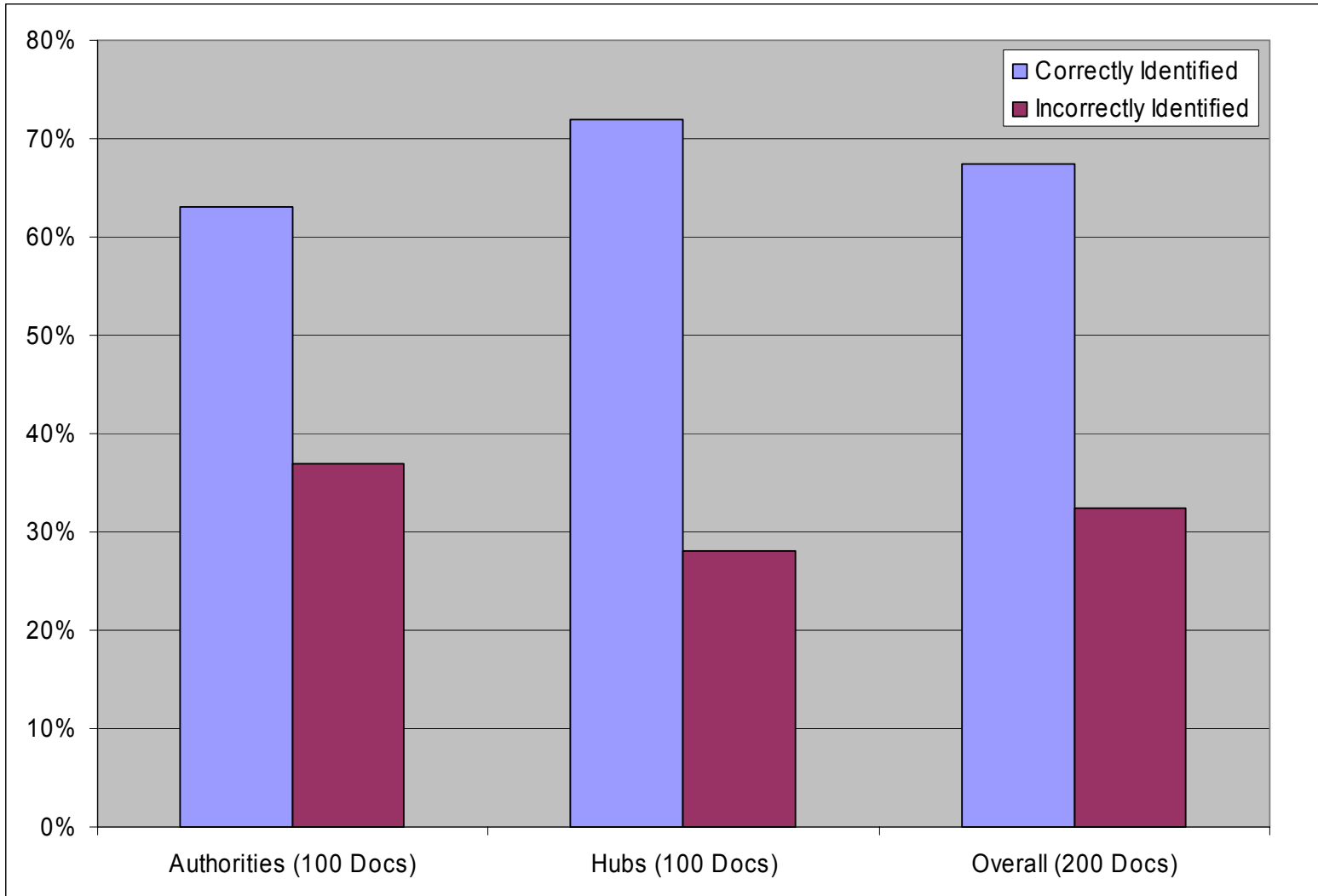
# Example: Authority Page



SOURCE: Centers for Disease Control and Prevention, National Center for Health Statistics, [National Health Interview Survey](#).

- In 1997, about 81 percent of children were reported by their parents to be in very good or excellent health.
- Child health varies by family income. Children living below the poverty line are less likely than children in higher-income families to be in very good or excellent health. In 1997, about 68 percent of children in families below the poverty line were in very good or excellent health, compared with 86 percent of children in families living at or above the poverty line.
- Children under age 5 are about as likely to be in very good or excellent health as children ages 5 to 17.
- The percentage of children in very good or excellent health remained stable between 1984 and 1997. The health gap between children below and those at or above the poverty line also did not change during the time period. Each year, children at or above the poverty line were about 20 percentage points more likely to be in very good or excellent health than children whose families were below poverty.

# Basic Heuristic Results



# Basic Heuristic Results

## Problems

- The Basic Heuristic does poorly
  - static method of identifying the body of documents
  - doesn't attempt to look for “error” documents

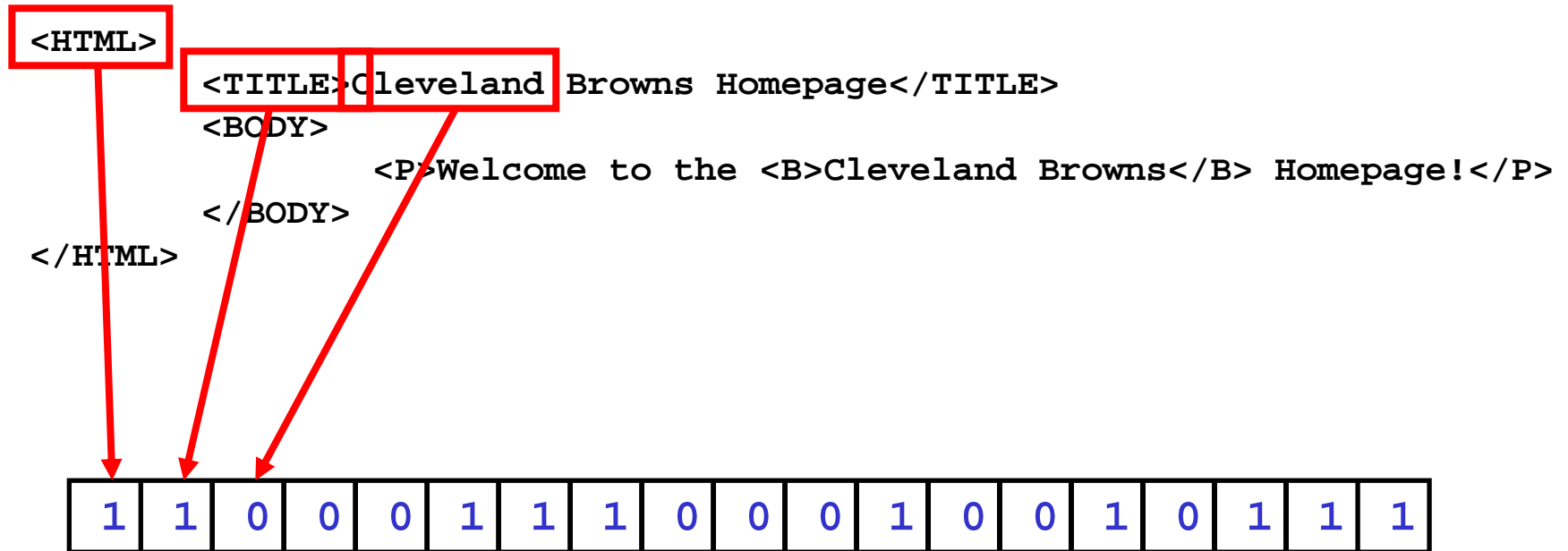
## Solutions

- Need a More Accurate Heuristic
  - dynamically identify the body of documents
  - looks for “error” documents

# Building Vectors from Tokens

- Tokenize each document into two types of tokens
  - text tokens (words)
  - HTML tag tokens
- Assign binary weights to each token
  - text tokens = “0”
  - HTML tag tokens = “1”
- Represent an HTML document as a vector of binary weights

# Vector Construction



# DS Graph Hub Example

Members WebMail Connections BizJournal SmartUpdate Mktplace

**TABLE OF CONTENTS**

[SUMMARY](#)

[BACKGROUND](#)

- [A. Site Description and History](#)
- [B. Site Visit](#)
- [C. Demographics, Land Use, and Natural Resource Use](#)
- [D. Health Outcome Data](#)

[ENVIRONMENTAL CONTAMINATION AND OTHER HAZARDS](#)

- [A. On-Site Contamination](#)
- [B. Off-Site Contamination](#)
- [C. Quality Assurance and Quality Control](#)
- [D. Physical and Other Hazards](#)

[PATHWAYS ANALYSES](#)

- [A. Completed Exposure Pathways](#)
- [B. Potential Exposure Pathways](#)
- [C. Eliminated Pathways](#)

[PUBLIC HEALTH IMPLICATIONS](#)

- [A. Toxicological Evaluation](#)
- [B. Child Health Data Evaluation](#)
- [C. Health Outcome Data Evaluation](#)
- [D. Community Health Concerns Evaluation](#)

[CONCLUSIONS](#)

[RECOMMENDATIONS](#)

[HEALTH ACTIVITIES RECOMMENDATION PANEL](#)

[PUBLIC HEALTH ACTIONS](#)

[PREPARERS OF REPORT](#)

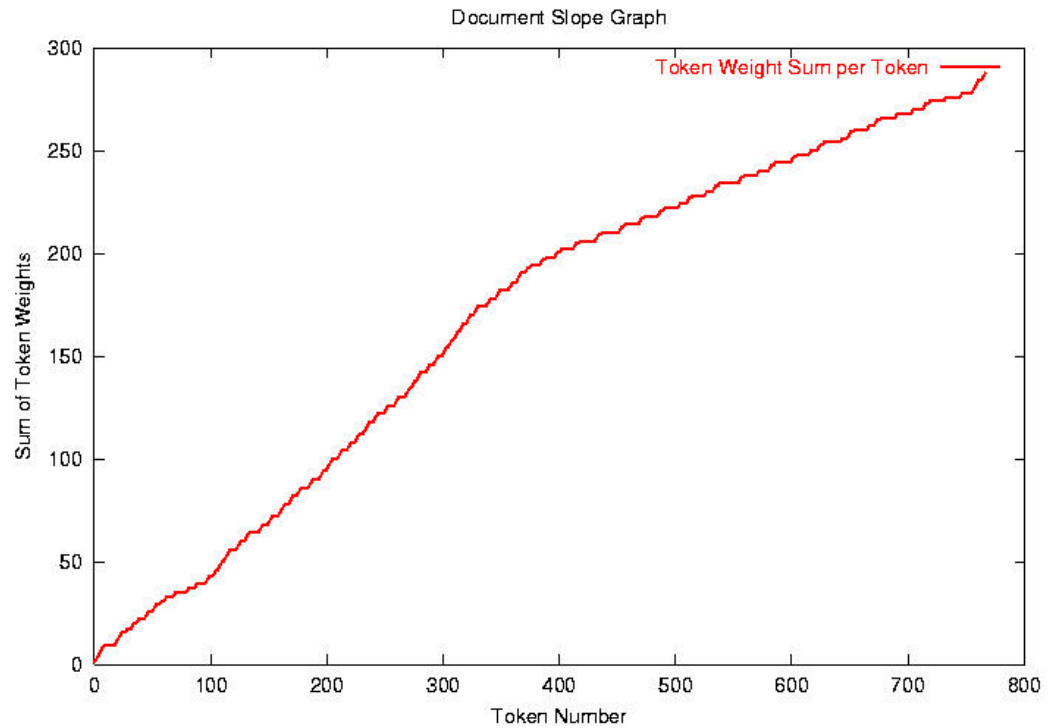
[CERTIFICATION](#)

[REFERENCES](#)

[APPENDICES](#)

- [A. Figures](#)
- [B. Tables](#)
- [C. Louisiana Tumor Registry Cancer Incidence Rates](#)
- [D. Glossary and Acronyms](#)
- [E. Public Availability Outreach Session Summary Report](#)
- [F. Public Comments](#)

100%



# DS Graph Authoritative Example

Office of Solid Waste and Emergency Response

**Brownfields Success Stories:**

**Somerville's Abandoned Mattress Site Springs Back**

---

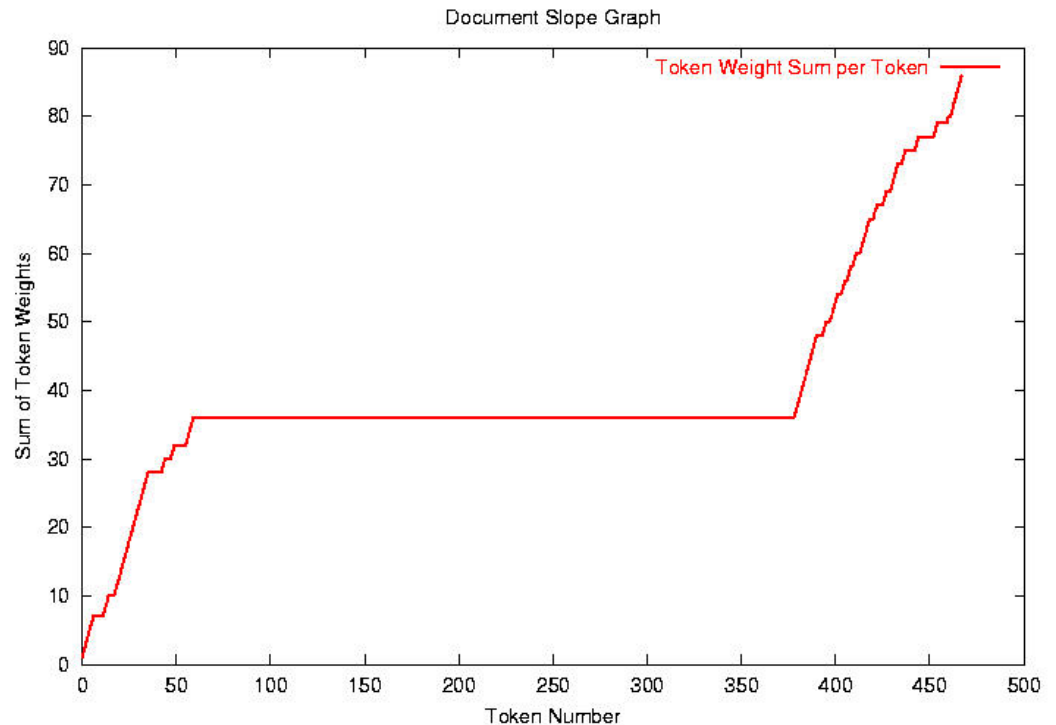
In Somerville, Massachusetts a 1,500-square-foot industrial building at 259 Lowell Street, occupied by the Hostess Bakery Company until the 1970s, and more recently by a series of mattress manufacturers, had sat abandoned and unused since 1995. The empty building fell into disrepair and became a safety concern and an eyesore for residents of the surrounding area. Fears of site contamination limited the site's appeal to developers. This changed in 1996, when Somerville was selected as an EPA Brownfields Pilot and received a \$100,000 grant to jump-start the city's idle properties. The Visiting Nurses Association (VNA) approached the city with an interest in purchasing and redeveloping the Lowell Street site, depending on the Pilot's assessment results. Soil and groundwater assessments revealed lead, petroleum, and barium contamination, with an estimated cleanup cost of \$225,000. To encourage the VNA to move forward with plans to redevelop the property into a 100-unit, assisted living facility and neighborhood health center, Somerville provided \$100,000 in cleanup cost-override coverage. The coverage, which was developed as a component of the Pilot, will protect the VNA should actual site cleanup costs exceed the Pilot's estimate. Community Development Block Grant money from the U.S. Department of Housing and Urban Development (HUD) is being used to finance the program. Other resources for the project include a \$1.25 million loan and a \$250,000 grant from the Federal Home Loan Bank; \$5.9 million in low-income housing tax credits from the Massachusetts Department of Housing and Community Development (DHCD); \$750,000 in low-interest loans from DHCD and the City of Somerville; a \$150,000 loan from the Somerville Affordable Housing Trust fund; and a \$5.4 million Affordable Housing loan from the Massachusetts Housing Partnership. When complete, VNA's \$14 million redevelopment project will bring 45 permanent jobs to the city. The project is now underway, and cleanup of contaminated soil is already complete. For more information on Somerville's Brownfields Pilot, contact Lynne Jennings at (617) 918-1210.

---

[ [Environmental Justice Homepage](#) | [EJFE Home](#) ]  
 [ [EJFE Home](#) | [Search](#) | [Home](#) | [What's New](#) ]

[ [Brownfields Updates](#) | [Join this Site](#) | [Site Statistics](#) ]

URL: [http://www.epa.gov/swevosp/bf/html/doc/sx\\_somvr.htm](http://www.epa.gov/swevosp/bf/html/doc/sx_somvr.htm)  
 Last updated July 19, 1999  
 Please E-mail comments on these web pages to:  
 Jim Shaw, Data Manager at: [jshaw-jones@epamail.epa.gov](mailto:jshaw-jones@epamail.epa.gov)





# Classification from Text Segments

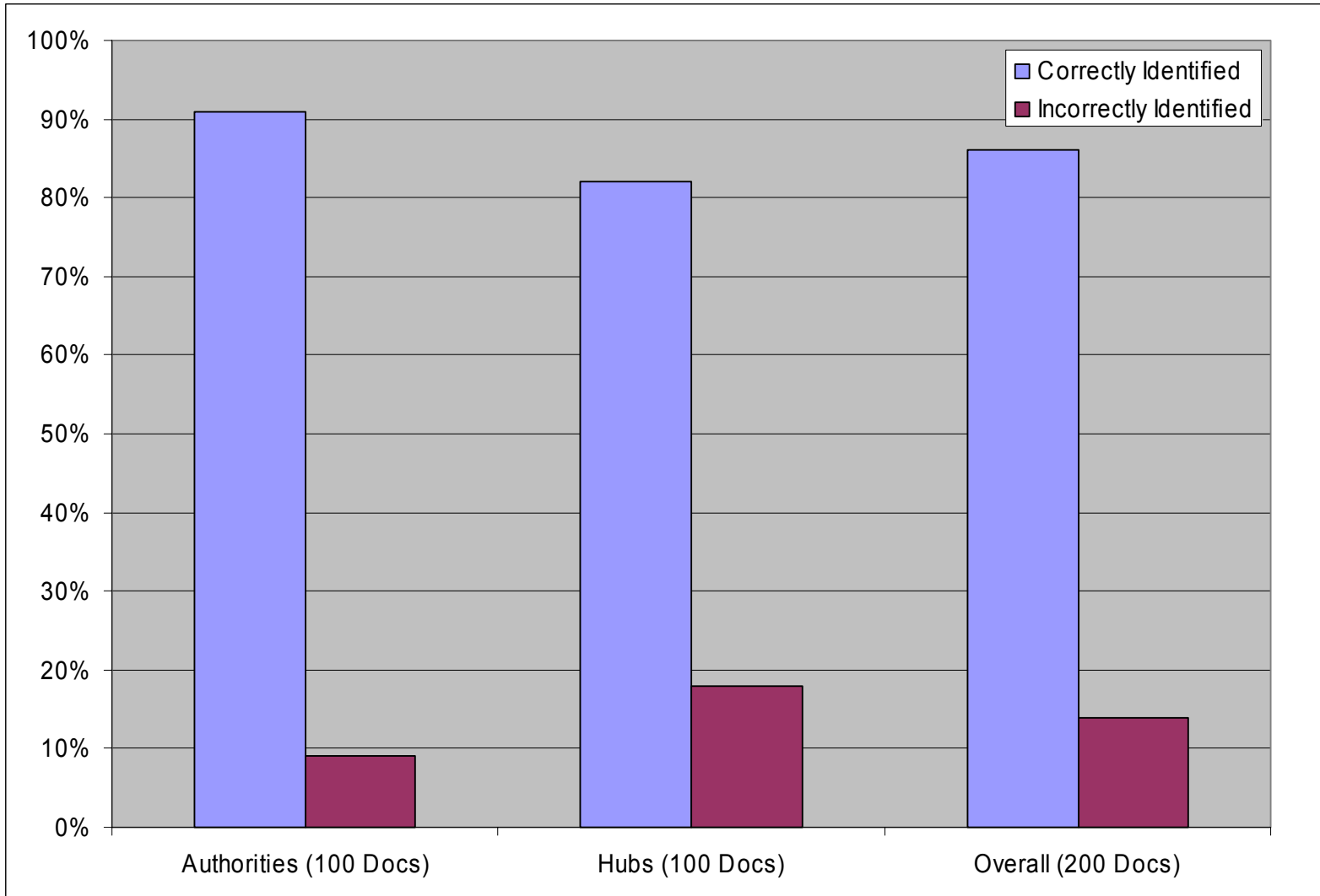
- ***HUB*** Documents
  - sum of text segment lengths  $< 10\%$  of document length
  - average text segment slope  $> 50\%$  of document slope
  - look for “error” documents

**404 File Not Found**

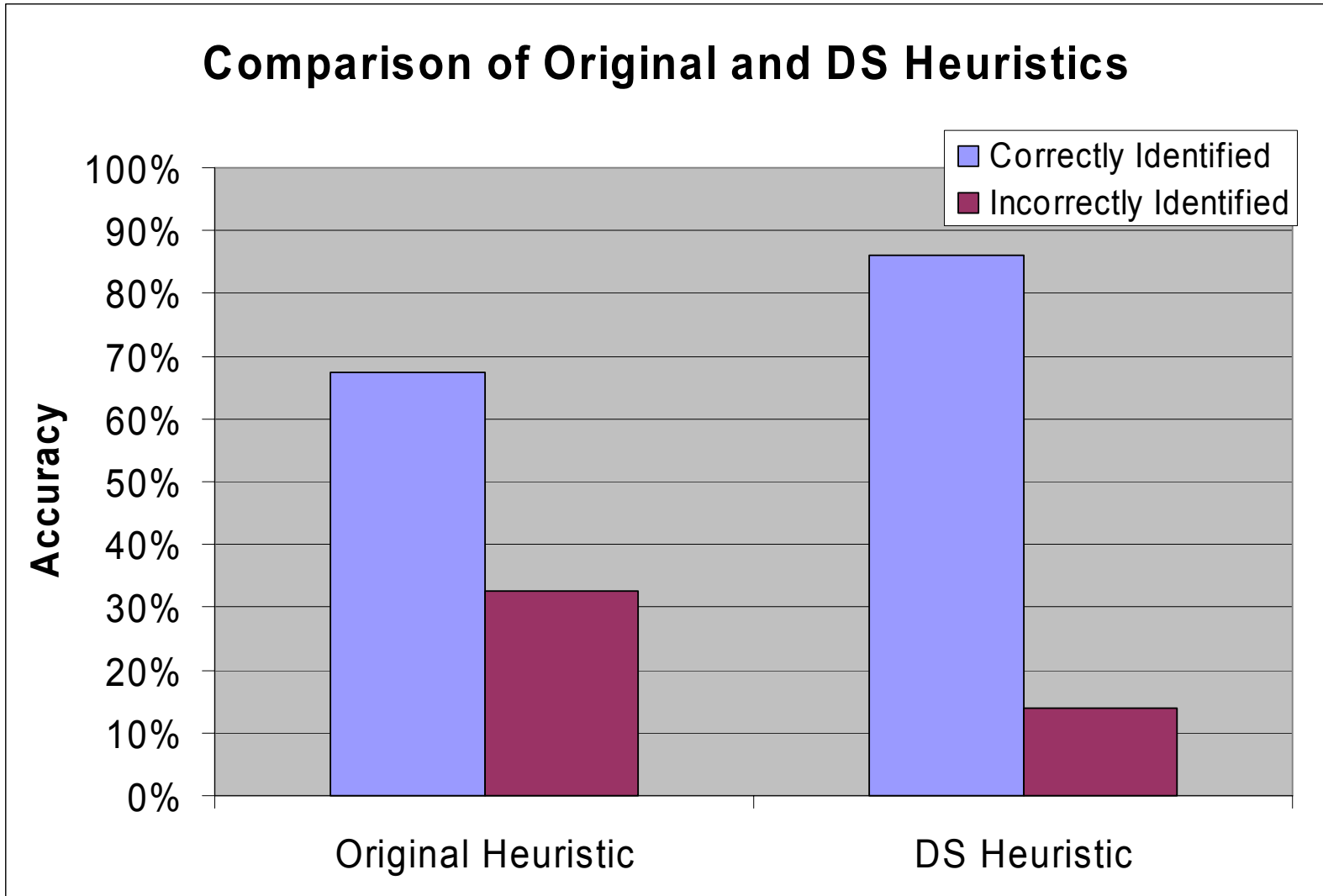
**Page Not Found**

- ***AUTHORITATIVE*** Documents
  - average text segment slope  $< 50\%$  document slope

# DS Heuristic Results



# Comparative Heuristic Results



# Steps Involved

- √ Web Crawl
- √ Hub/Authority Algorithm (reducing the dataset)
  - Finding ***METADATA*** in the Documents
  - Searching the ***METADATA***

# What is *METADATA*?

- *METADATA* describes an object



# Why is *METADATA* Important?

- Adding *METADATA* to a document creates more searchable documents

William J. Clinton » `<NAME>` William J. Clinton `</NAME>`

3 gallons » `<NUMBER TYPE = "VOLUME">` 3 gallons `</NUMBER>`

Amherst, MA » `<PLACE>` Amherst, MA `</PLACE>`

- By adding *METADATA* to documents, we hope to improve the efficiency of our QA searches

# Identifying *METADATA*

- Locate *METADATA* within a document



**CNN.com** **allpolitics.com** with TIME

MAINPAGE  
WORLD  
U.S.  
WEATHER  
BUSINESS  
SPORTS  
**ENTERTAINMENT**  
LAW  
SCI-TECH  
SPACE  
HEALTH  
ENTERTAINMENT  
TRAVEL  
EDUCATION  
CAREER  
LOCAL  
IN-DEPTH

## Bush approval rating shows minor drop, hovers just above 50 percent

July 3, 2001 Posted: 9:06 a.m. EDT (1306 GMT)

By Keating Holland  
CNN Polling Director

WASHINGTON -- Most Americans continue to think that George W. Bush is doing a good job as president, and the public remains unconcerned by Vice President Dick Cheney's heart condition, according to the results of a new CNN/USA Today/Gallup poll released Monday.

But the Democrats score higher than the GOP on the patients' bill of rights, and Bush's energy plan has not won support, even though worries about the country's energy situation have started to wane.

The poll consisted of interviews of 1,015 adult Americans conducted from June 29 through July 1. The question on Cheney's health was posed to 571 respondents on June 30, through July 1.

Some 52 percent of respondents now approve of the way Bush is handling his job as president. That figure -- his lowest to date -- represents a three-point drop since mid-June, and is statistically identical to other polls released last week. But it is important to note that Bush has not been the victim of a sudden drop in his approval, but instead a slow, steady slide spread out over several months.

**LIFE** Collector's Edition  
**OUR CALL TO ARMS:**  
The Attack on Pearl Harbor  
[Click Here](#)

**EDITIONS:**  
[CNN.com Asia](#)  
[CNN.com Europe](#)  
[set default edition](#)

**MULTIMEDIA:**  
[video](#)  
[audio](#)  
[multimedia](#)  
[showcase](#)  
[more services](#)

**E-MAIL:**  
Subscribe to one of our [news e-mail lists](#).



**CNN.com** **allpolitics.com** with TIME

MAINPAGE  
WORLD  
U.S.  
WEATHER  
BUSINESS  
SPORTS  
**ENTERTAINMENT**  
LAW  
SCI-TECH  
SPACE  
HEALTH  
ENTERTAINMENT  
TRAVEL  
EDUCATION  
CAREER  
LOCAL  
IN-DEPTH

## *Bush approval rating shows minor drop, hovers just above 50 percent*

July 3, 2001 Posted: 9:06 a.m. EDT (1306 GMT)

By Keating Holland  
CNN Polling Director

*WASHINGTON -- Most Americans continue to think that George W. Bush is doing a good job as president, and the public remains unconcerned by Vice President Dick Cheney's heart condition, according to the results of a new CNN/USA Today/Gallup poll released Monday.*

But the Democrats score higher than the GOP on the patients' bill of rights, and Bush's energy plan has not won support, even though worries about the country's energy situation have started to wane.

The poll consisted of interviews of 1,015 adult Americans conducted from June 29 through July 1. The question on Cheney's health was posed to 571 respondents on June 30, through July 1.

Some 52 percent of respondents now approve of the way Bush is handling his job as president. That figure -- his lowest to date -- represents a three-point drop since mid-June, and is statistically identical to other polls released last week. But it is important to note that Bush has not been the victim of a sudden drop in his approval, but instead a slow, steady slide spread out over several months.

It is probably incorrect to assume that Bush's current figures are due to anything he has done recently.

**LIFE** Collector's Edition  
**OUR CALL TO ARMS:**  
The Attack on Pearl Harbor  
[Click Here](#)

**EDITIONS:**  
[CNN.com Asia](#)  
[CNN.com Europe](#)  
[set default edition](#)

**MULTIMEDIA:**  
[video](#)  
[audio](#)  
[multimedia](#)  
[showcase](#)  
[more services](#)

**E-MAIL:**  
Subscribe to one of our [news e-mail lists](#).  
Enter your address:

# Example of Finding *METADATA*

[CNN.com Europe](#)  
[CNN.com Asia](#)  
[Spanish](#)  
[Portuguese](#)  
[German](#)  
[Italian](#)  
[Danish](#)  
[Japanese](#)  
[Korean Headlines](#)

TIME INC.  
 SITES:  
 Go To ...

CNN  
 NETWORKS:  
 CNN Networks

[studio tour](#)  
[CNN anchors](#)  
[transcripts](#)

SITE INFO:  
[ABOUT US](#)  
[search](#)

WEB SERVICES:

**CNN/USA Today/Gallup POLL**

Sampling error: +/-3% pts June 29-July 1

**Do you approve or disapprove of the job George W. Bush is doing as president?**

Approve	52%
Disapprove	34%

Sampling error: +/-3% pts

**Bush approval rating**

Now	52%
June	55%
May	56%
April	62%
March	63%
February	57%

**Based on what you have heard or read, do you favor or oppose Congress passing a patient's bill of rights?**

Yes	58%
No	11%
Unsure	31%

**Even if you don't know all of the details, in general, whose approach to a patient's bill of rights would you be more likely to trust -- the Republicans, or the Democrats?**


Democrats	44%
Republicans	34%

**How serious would you say the energy situation is in the United States -- very serious, fairly serious, or not at all serious?**

Very serious	47%
Fairly serious	43%
Not serious	8%

**Comparison to the responses to the same question, when asked in March, and again in May.**

Now	47%
May	58%
March	21%

 **take**  
 the online broker for long-term investors.

[Quick News](#)  
[CNNfy.com](#)  
[CNN.com Europe](#)  
[CNN.com Asia](#)  
[Spanish](#)  
[Portuguese](#)  
[German](#)  
[Italian](#)  
[Danish](#)  
[Japanese](#)  
[Korean Headlines](#)

TIME INC.  
 SITES:  
 Go To ...

CNN  
 NETWORKS:  
 CNN Networks

[studio tour](#)  
[CNN anchors](#)  
[transcripts](#)

SITE INFO:  
[ABOUT US](#)  
[search](#)

WEB SERVICES:

**CNN/USA Today/Gallup POLL**

Sampling error: +/-3% pts June 29-July 1

**Do you approve or disapprove of the job George W. Bush is doing as president?**

Approve	52%
Disapprove	34%

Sampling error: +/-3% pts

**Bush approval rating**

Now	52%
June	55%
May	56%
April	62%
March	63%
February	57%

**Based on what you have heard or read, do you favor or oppose Congress passing a patient's bill of rights?**

Yes	58%
No	11%
Unsure	31%

**Even if you don't know all of the details, in general, whose approach to a patient's bill of rights would you be more likely to trust -- the Republicans, or the Democrats?**


Democrats	44%
Republicans	34%

**How serious would you say the energy situation is in the United States -- very serious, fairly serious, or not at all serious?**

Very serious	47%
Fairly serious	43%
Not serious	8%

**Comparison to the responses to the same question, when asked in March, and again in May.**

Now	47%
May	58%

 **take**  
 the online broker for long-term investors.

# *METADATA*: Focus

- Sectioning documents with *METADATA*
- HTML table markup
- Identifying plain-text tables

# Sectioning with *METADATA*

- Include the document segments between *METADATA* to create sections

**CNN/USA Today/Gallup POLL**

Sampling error: +/-3% pts June 29-July 1

*Do you approve or disapprove of the job George W. Bush is doing as president?*

Approve	52%
Disapprove	34%

Sampling error: +/-3% pts

**Bush approval rating**

Now	52%
June	55%
May	56%
April	62%
March	63%
February	57%

*Based on what you have heard or read, do you favor or oppose Congress passing a patient's bill of rights?*

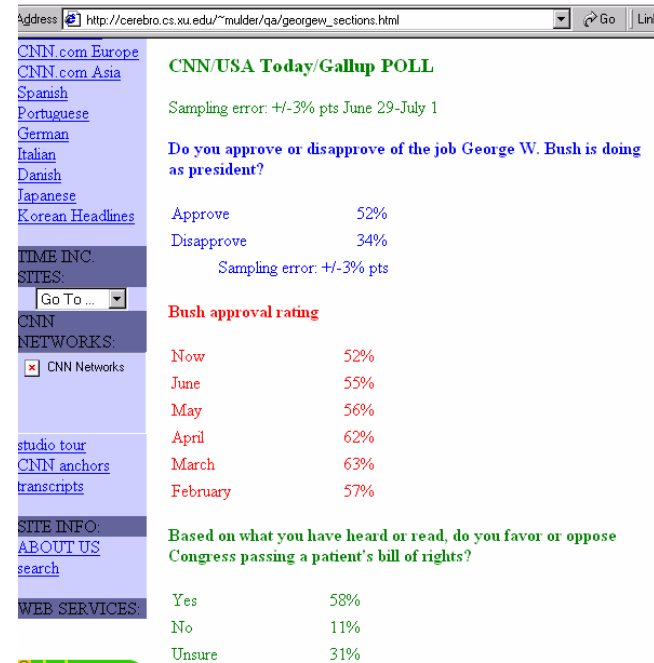
Yes	58%
No	11%
Unsure	31%



*Even if you don't know all of the details, in general, whose approach to a patient's bill of rights would you be more likely to trust -- the Republicans, or the Democrats?*

Democrats	44%
Republicans	34%

*How serious would you say the energy situation is in the United States -- very serious, fairly serious, or not at all serious?*



**CNN/USA Today/Gallup POLL**

Sampling error: +/-3% pts June 29-July 1

*Do you approve or disapprove of the job George W. Bush is doing as president?*

Approve	52%
Disapprove	34%

Sampling error: +/-3% pts

**Bush approval rating**

Now	52%
June	55%
May	56%
April	62%
March	63%
February	57%

*Based on what you have heard or read, do you favor or oppose Congress passing a patient's bill of rights?*

Yes	58%
No	11%
Unsure	31%



*Even if you don't know all of the details, in general, whose approach to a patient's bill of rights would you be more likely to trust -- the Republicans, or the Democrats?*

Democrats	44%
Republicans	34%

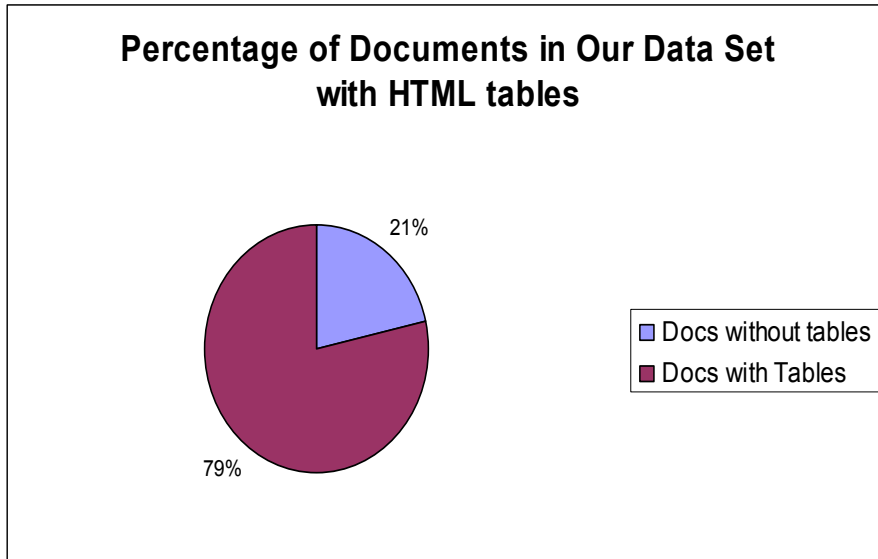
# *METADATA*: Focus

- √ Sectioning documents with *METADATA*
  - HTML Table Markup
  - Identifying plain-text Tables

# Using Table Data

- Two types of tables found in documents on the web:
  - HTML (marked up by the document author)
  - Plain text (no markup given)
- Problem:
  - Answer is often remote from *METADATA*  
(column headings, row headings, title, units)

# HTML Table Markup



- Why do we want to do this?
  - Majority of sites from our data set contain tables.
  - Many numeric answers are contained in tables.



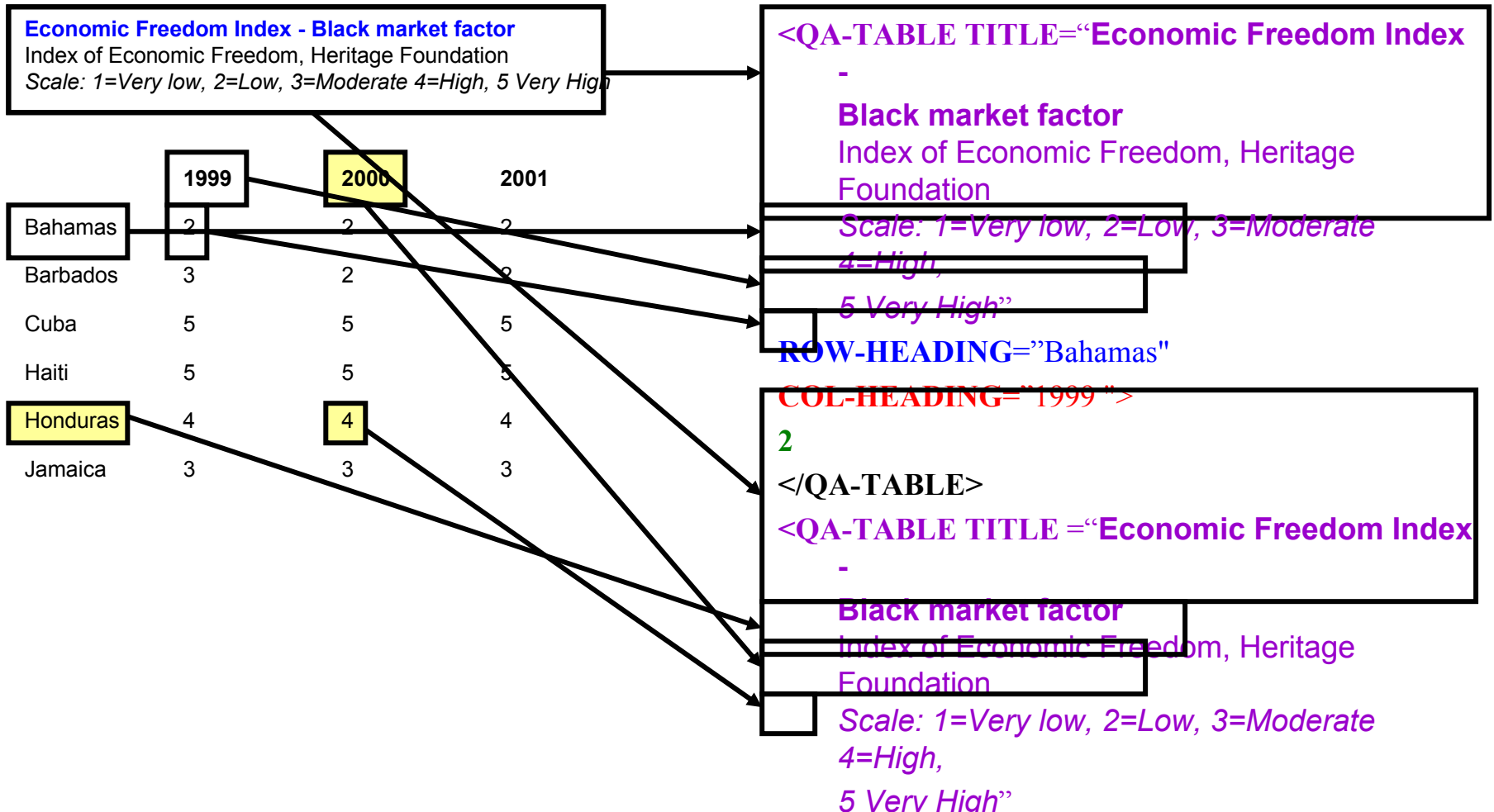
# Our Approach

- Search through each document to see if it contains any `<TABLE>` tags
  - If it does, try to markup individual table cells.
  - Problem: what are the row and column headings for each cell?

# Solution

- Table Generalities:
  - The first row in a table usually contains column headings.
  - The first entry in each row is usually the row heading.
- Using this information, we can add contextual markup to each table cell.

# Examples of *METADATA*



# *METADATA*: Focus

- √ Sectioning documents with *METADATA*
- √ HTML Table Markup
- Identifying plain-text Tables

# Plain Text Tables

- Motivation: A lot of the documents in our data set are ASCII text files which contain tables.
- Problems:
  - Finding tables in text files
  - Table formats and layouts not consistent across a wide collection of documents.

# Examples of plain text tables

## COUNTY NAMES AND CODES

This file lists all counties and equivalent areas in the United States defined as of January 1, 1990, alphabetically by State with related codes. There are four codes shown: the first code is the FIPS MSA/CMSA/NECMA code, the second code is the FIPS PMSA code, the third code is a combination of the FIPS State and county codes, and the fourth code is the geographic summary level (0 = United States total; 1 = state total; 2 = county in CMSA/PMSA; 3 = county in MSA; 4 = county in NECMA; 5 = county not in any metro area). (MSA = metropolitan statistical area; CMSA = consolidated MSA; NECMA = New England county MA; PMSA = primary MSA)

	02000	1	ALASKA
	02013	5	Aleutians East, AK
	02016	5	Aleutians West, AK
0380	02020	3	Anchorage, AK
	02050	5	Bethel, AK
	02060	5	Bristol Bay, AK
	02068	5	Denali, AK
	02070	5	Dillingham, AK
	02090	5	Fairbanks North Star, AK
	02100	5	Haines, AK
	02110	5	Juneau, AK
	02122	5	Kenai Peninsula, AK

## Schedule C - Country List (by name)

Code	Name	ISO
5310	Afghanistan	AF
4810	Albania	AL
7210	Algeria	DZ
9510	American Samoa	AS
4271	Andorra	AD
7620	Angola	AO
2481	Anguilla	AI
2484	Antigua and Barbuda	AG
3570	Argentina	AR
4631	Armenia	AM
2779	Aruba	AW
6021	Australia	AU
4330	Austria	AT
4632	Azerbaijan	AZ
2360	Bahamas	BS
5250	Bahrain	BH
5380	Bangladesh	BD
2720	Barbados	BB
4622	Belarus	BY
4231	Belgium	BE
2080	Belize	BZ
7610	Benin	BJ
2320	Bermuda	BM
5682	Bhutan	BT
3350	Bolivia	BO



# Solution

- We implemented a system for retrieving text tables called TINTIN(Pyreddy and Croft). A similar implementation was written by Margie Connell of the CIIR.
- Our implementation is comprised of 3 steps:
  - An **extractor algorithm**, which finds possible tables in a document.
  - A **captioning algorithm**, which differentiates between table data and data about that table(titles, headings).
  - An **XML table markup program**, which adds XML-compliant ***METADATA*** to each table cell entry.

# Extraction

- **Steps:**

- Eliminate noise in the document. Example:

```
*****  
*****  
*****  
*****  
*****  
*****  
*****
```

- Go through the document and create a **CAG** (Character Alignment Graph). A CAG is formed by checking for whitespace alignment in contiguous lines of text.
- Find gaps by looking for contiguous locations of whitespace in the CAG. These gaps are considered potential column separators and the lines are potential table rows.

# Captioning

- We used five separate heuristics in order to differentiate between table lines and captions(information) about the table lines:
  - Gap Structure
  - Alignment
  - Pattern Regularity
  - Differential Column Count
  - Differential Gap Structure

# XML Markup (*METADATA*)

- Cell data not marked up in the original TINTIN system.
- Similar to HTML table markup at this point, though less precise. Example:

```
<QA-table TITLE="Schedule C - Country List (by name) CODE NAME  
ISO"
```

```
ROW="7620">
```

```
Angola AO
```

```
</QA-table>
```

- Problems:
  - Finding the correct column headings consistently.
  - Finding the title of the table (how many of the non-table lines before and after the table do you count as the title?)

# Steps Involved

- √ Web Crawl
- √ Hub/Authority Algorithm
- √ Finding *METADATA* in the Documents
- Searching the *METADATA*

# Searching Metadata

- Sections
  - Search sections for answers to questions instead of entire documents
- Tables
  - Translate marked up tables into sentences that current system can understand

# State of Current QA System

Answering: Who is the president of the United States? - Microsoft Internet Explorer

Address: <http://canberra.cs.umass.edu/cgi-bin/reu2/QA2.cgi>

## Answering: Who is the president of the United States?

Document Identifier	InQuery Belief	Passage Score	Answer	Answer Passage
<a href="#">00000000028333</a>	4.527	0.038	WILLIAM J. CLINTON	During National Family Caregivers Week, let us honor the many devoted men and women whose efforts do so much to strengthen the bonds of family and community in our Nation. NOW, THEREFORE, I <b>WILLIAM J. CLINTON, President</b> of the <b>United</b> States of America, by virtue of the authority vested in me by the Constitution and laws of the <b>United States</b> , do hereby proclaim November 21 through November 27, 1999, as National Family Caregivers Week.
<a href="#">00000000028333</a>	4.527	-1.000	No answer in passage.	<QA_METADATA>(Copy of <ENAMEX TYPE="ORGANIZATION">White House</ENAMEX> Document on AoA Web Site)</QA_METADATA>(Copy of <ENAMEX TYPE="ORGANIZATION">White House</ENAMEX> Document on AoA Web Site)
<a href="#">00000000028333</a>	4.527	-1.000	No answer in passage.	<QA_METADATA><TIMEX TYPE="DATE">November 19, 1999</TIMEX></QA_METADATA><TIMEX TYPE="DATE">November 19, 1999</TIMEX>
<a href="#">00000000028333</a>	4.527	-1.000	No answer in passage.	<QA_METADATA>Hypertext conversion by <ENAMEX TYPE="PERSON">Saadia Greenberg</ENAMEX> ,</QA_METADATA> Hypertext conversion by <ENAMEX TYPE="PERSON">Saadia Greenberg</ENAMEX> , <a href="http://www.agingstats.gov/pr/">http://www.agingstats.gov/pr/</a> <NUMEX TYPE="NUMBER">1999</NUMEX><ENAMEX TYPE="PERSON">Caregiverweek</ENAMEX> .html <TIMEX TYPE="DATE">Nov 22 99</TIMEX>
<a href="#">00000000028333</a>	4.527	-1.000	No answer in passage.	<QA_METADATA>Go to <ENAMEX TYPE="ORGANIZATION">AoA Home Page</ENAMEX> Go to AoA Press</ENAMEX> Release Menu</QA_METADATA> Go to <ENAMEX TYPE="PERSON">AoA</ENAMEX> Home Page Go to <ENAMEX TYPE="ORGANIZATION">AoA Press</ENAMEX> Release Menu Go to <ENAMEX TYPE="LOCATION">ElderPage</ENAMEX> (AoA Web Site) Go to <ENAMEX TYPE="LOCATION">ElderPage</ENAMEX> (AoA Web Site) Go to Site Index Go to Site Index
<a href="#">00000000028333</a>	4.527	-1.000	No answer in passage.	<QA_METADATA>For further information, contact</QA_METADATA>For further information, contact
<a href="#">00000000028333</a>	4.527	-1.000	No answer in passage.	<QA_METADATA><ENAMEX TYPE="PERSON">Moya Benoit Thompson</ENAMEX></QA_METADATA><ENAMEX TYPE="PERSON">Moya Benoit Thompson</ENAMEX>
<a href="#">00000000028333</a>	4.527	-1.000	No answer in passage.	<QA_METADATA>(<NUMEX TYPE="NUMBER">202</NUMEX>) <NUMEX TYPE="NUMBER">401-4541</NUMEX></QA_METADATA>(<NUMEX TYPE="NUMBER">202</NUMEX>) <NUMEX TYPE="NUMBER">401-4541</NUMEX>
<a href="#">00000000028333</a>	4.527	-1.000	No answer in passage.	<QA_METADATA>(Copy of <ENAMEX TYPE="ORGANIZATION">White House</ENAMEX> Document on AoA Web Site)</QA_METADATA>(Copy of <ENAMEX TYPE="ORGANIZATION">White House</ENAMEX> Document on AoA Web Site)
<a href="#">00000000028333</a>	4.527	-1.000	No answer in passage.	<QA_METADATA><TIMEX TYPE="DATE">November 19, 1999</TIMEX></QA_METADATA><TIMEX TYPE="DATE">November 19, 1999</TIMEX>

Done Internet

# State of Current QA System

Geography MapStats	Choctaw County	Alabama
Land area, 2000 (square miles)	914	50,744
Persons per square mile, 2000	17.4	87.6
Metropolitan Area	None	

Answering: What is the population of Choctaw County, Alabama? - Microsoft Internet Explorer

Address: <http://canberra.cs.umass.edu/cgi-bin/reu2/QA2.cgi>

<p>00000000005867</p>	<p>0.855 -1.000</p>	<p>No answer in passage.</p>	<pre> &lt;QA_METADATA&gt;Geography &lt;ENAMEX TYPE="PERSON"&gt;MapStatsChoctaw CountyAlabama&lt;/ENAMEX&gt;&lt;/QA_METADATA&gt;Geography &lt;ENAMEX TYPE="LOCATION"&gt;MapStats Choctaw County&lt;/ENAMEX&gt; &lt;ENAMEX TYPE="LOCATION"&gt;Alabama&lt;/ENAMEX&gt; &lt;NUMEX TYPE="NUMBER"&gt;914&lt;/NUMEX&gt; &lt;NUMEX TYPE="NUMBER"&gt;914&lt;/NUMEX&gt; , Geography &lt;ENAMEX TYPE="PERSON"&gt;MapStats&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="LOCATION"&gt;Alabama&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="LOCATION"&gt;Land&lt;/ENAMEX&gt; area, &lt;TIMEX TYPE="DATE"&gt;2000&lt;/TIMEX&gt; (square miles) , &lt;NUMEX TYPE="NUMBER"&gt;50,744&lt;/NUMEX&gt; &lt;NUMEX TYPE="NUMBER"&gt;50,744&lt;/NUMEX&gt; , Geography &lt;ENAMEX TYPE="PERSON"&gt;MapStats&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="LOCATION"&gt;Alabama&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="LOCATION"&gt;Land&lt;/ENAMEX&gt; area, &lt;TIMEX TYPE="DATE"&gt;2000&lt;/TIMEX&gt; (square miles) &lt;NUMEX TYPE="MONEY"&gt;17&lt;/NUMEX&gt; &lt;NUMEX TYPE="PERCENT"&gt;4&lt;/NUMEX&gt; &lt;NUMEX TYPE="PERCENT"&gt;17.4&lt;/NUMEX&gt; , Geography &lt;ENAMEX TYPE="PERSON"&gt;MapStats&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="LOCATION"&gt;Alabama&lt;/ENAMEX&gt; , Persons per square mile, &lt;TIMEX TYPE="DATE"&gt;2000&lt;/TIMEX&gt; , &lt;NUMEX TYPE="NUMBER"&gt;87&lt;/NUMEX&gt; &lt;NUMEX TYPE="NUMBER"&gt;6&lt;/NUMEX&gt; &lt;NUMEX TYPE="NUMBER"&gt;87.6&lt;/NUMEX&gt; , Geography &lt;ENAMEX TYPE="PERSON"&gt;MapStats&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="LOCATION"&gt;Alabama&lt;/ENAMEX&gt; , Persons per square mile, &lt;TIMEX TYPE="DATE"&gt;2000&lt;/TIMEX&gt; ,None None , Geography &lt;ENAMEX TYPE="PERSON"&gt;MapStats&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="LOCATION"&gt;Alabama&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="LOCATION"&gt;Metropolitan Area&lt;/ENAMEX&gt; , , Geography &lt;ENAMEX TYPE="PERSON"&gt;MapStats&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="LOCATION"&gt;Alabama&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="LOCATION"&gt;Metropolitan Area&lt;/ENAMEX&gt; ,(a) Includes persons reporting only &lt;NUMEX TYPE="NUMBER"&gt;one&lt;/NUMEX&gt; race. (b) Hispanics may be of any race, so also are included in applicable race categories. Figures are in absolute numbers unless otherwise indicated. FN: Footnote on this item for this area in place of data NA: Not available D: Suppressed to avoid disclosure of confidential information X: Not applicable S: &lt;ENAMEX TYPE="PERSON"&gt;Suppressed&lt;/ENAMEX&gt;; does not meet publication standards Z: &lt;ENAMEX TYPE="ORGANIZATION"&gt;Value&lt;/ENAMEX&gt; greater than &lt;NUMEX TYPE="PERCENT"&gt;zero&lt;/NUMEX&gt; but less than &lt;NUMEX TYPE="NUMBER"&gt;half&lt;/NUMEX&gt; unit of measure shown What do you think of our new MapStats. Send comments to fedstats-mapstats@lists.census.govSource: &lt;ENAMEX TYPE="ORGANIZATION"&gt;Bureau of Economic Analysis&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="ORGANIZATION"&gt;Bureau of Labor Statistics&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="ORGANIZATION"&gt;National Agricultural Statistics Service&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="ORGANIZATION"&gt;National Center for Health Statistics&lt;/ENAMEX&gt; , &lt;ENAMEX TYPE="ORGANIZATION"&gt;U. S. Census Bureau&lt;/ENAMEX&gt;Metadata&lt;/ENAMEX&gt; powered by &lt;ENAMEX TYPE="LOCATION"&gt;DataWeb&lt;/ENAMEX&gt; Last Revised: &lt;TIMEX TYPE="DATE"&gt;Thursday&lt;/TIMEX&gt; , &lt;NUMEX TYPE="NUMBER"&gt;03&lt;/NUMEX&gt; &lt;TIMEX TYPE="DATE"&gt;May&lt;/TIMEX&gt; &lt;TIMEX TYPE="DATE"&gt;2001&lt;/TIMEX&gt; &lt;TIMEX TYPE="TIME"&gt;09:25:20 EDT&lt;/TIMEX&gt;our&lt;/TIMEX&gt; privacy on this site Your privacy on this site Accessibility on this site Accessibility on this site for persons with disabilities for persons with disabilities. , , , &lt;ENAMEX TYPE="ORGANIZATION"&gt;Fedstats - www.fedstats&lt;/ENAMEX&gt;. gov/ About &lt;ENAMEX TYPE="PERSON"&gt;Fedstats&lt;/ENAMEX&gt; Send your feedback to &lt;ENAMEX TYPE="PERSON"&gt;Fedstats&lt;/ENAMEX&gt; .h ttp://www.fedstats.gov/qfstates/&lt;NUMEX TYPE="NUMBER"&gt;01/01023&lt;/NUMEX&gt;.html Unknown &lt;ENAMEX </pre>
-----------------------	---------------------	------------------------------	--



# Web Demo

<http://canberra.cs.umass.edu/~reu2/QA2.html>

[click](#)

# Future Work

- A better hub/authority algorithm
- Sectioning of documents before putting them into an InQuery database
- Mine data from query-type pages
- Use link text as *METADATA*
- Use *METADATA* to augment scoring, perhaps choose passages to score



</Presentation>