A Geographically Searchable Local News Aggregator

IPRO 327

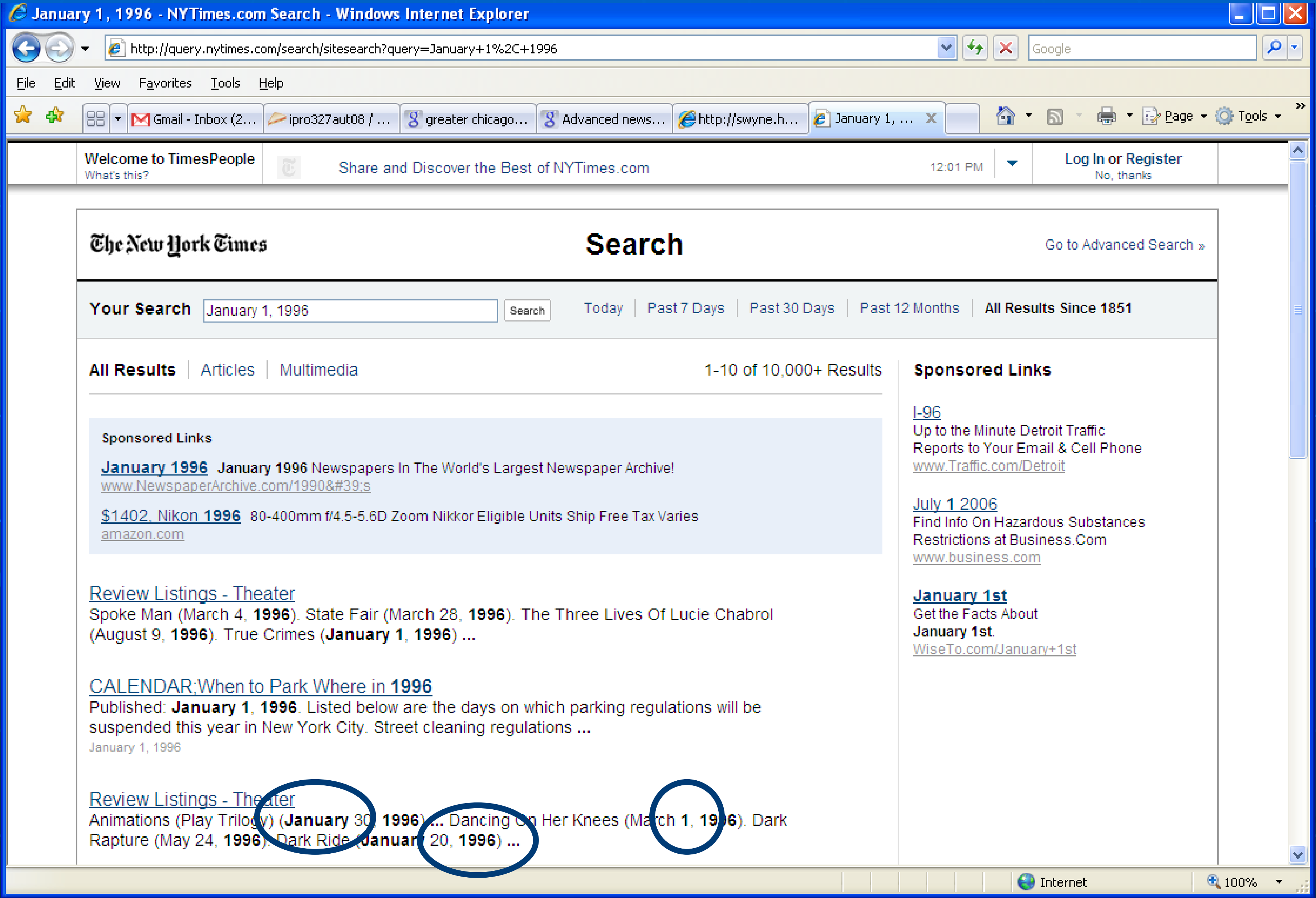
Fall, 2008

Advisors: Frieder, Yee

Goal

- Create a news search engine that understands the contents of articles
  - Allows search by “meaning” versus by “string matching”
  - Example: What happened in Chicago in 1996?

Current Practice – nytimes.com



Welcome to TimesPeople
What's this?

Share and Discover the Best of NYTimes.com

12:01 PM

Log In or Register
No, thanks

The New York Times

Search

Go to Advanced Search »

Your Search: January 1, 1996 [Search]

Today | Past 7 Days | Past 30 Days | Past 12 Months | **All Results Since 1851**

All Results | Articles | Multimedia

1-10 of 10,000+ Results

Sponsored Links

January 1996 **January 1996** Newspapers In The World's Largest Newspaper Archive!
www.NewspaperArchive.com/1990&#39;s

$1402, Nikon **1996** 80-400mm f/4.5-5.6D Zoom Nikkor Eligible Units Ship Free Tax Varies
amazon.com

Review Listings - Theater
Spoke Man (March 4, **1996**). State Fair (March 28, **1996**). The Three Lives Of Lucie Chabrol (August 9, **1996**). True Crimes (**January 1**, **1996**) **...**

CALENDAR;When to Park Where in **1996**
Published: **January 1**, **1996**. Listed below are the days on which parking regulations will be suspended this year in New York City. Street cleaning regulations **...**
January 1, 1996

Review Listings - Theater
Animations (Play Trilogy) (**January** 30, **1996**) **...** Dancing On Her Knees (March **1**, **1996**). Dark Rapture (May 24, **1996**). Dark Ride (**January** 20, **1996**) **...**

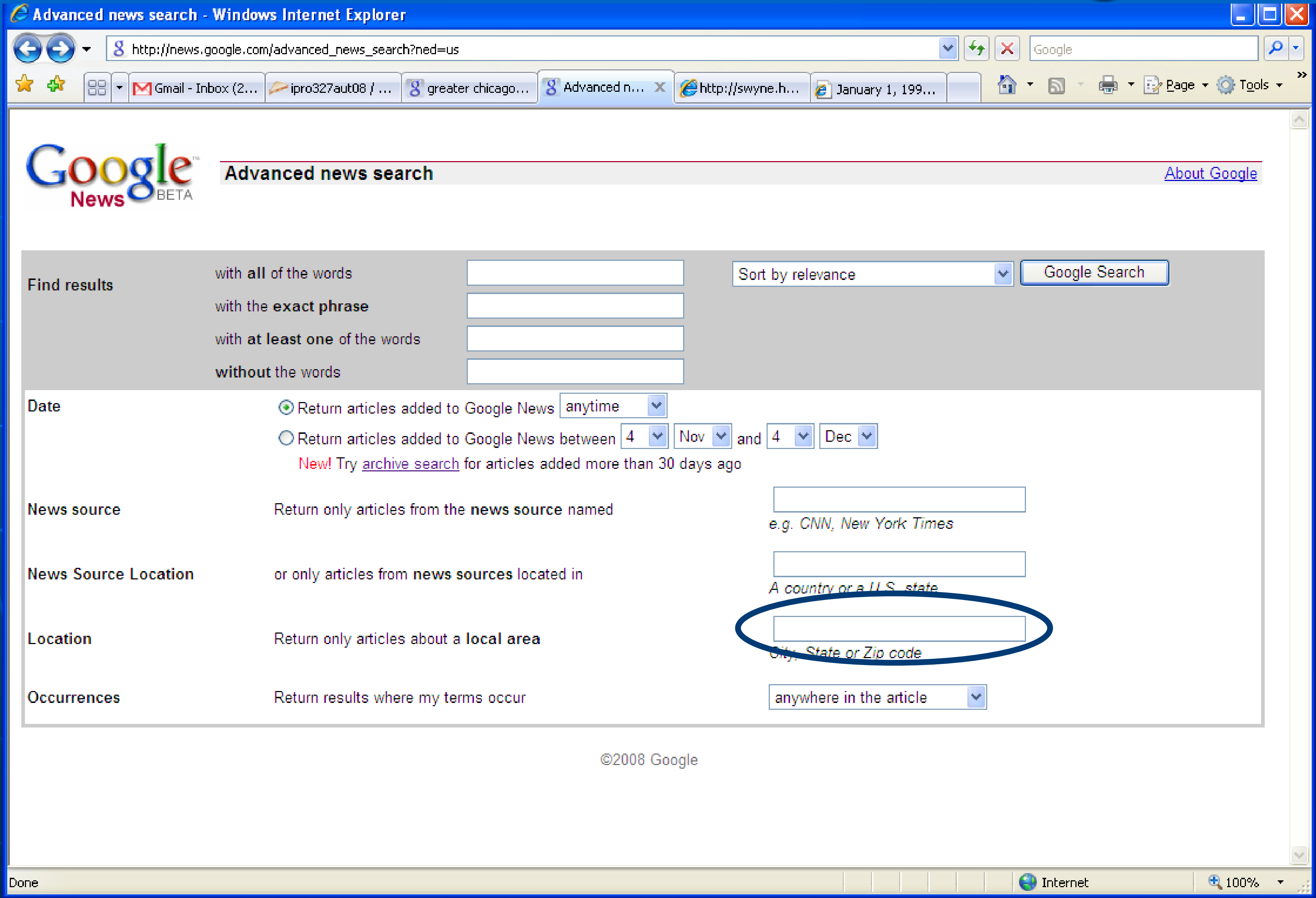
Sponsored Links

I-96
Up to the Minute Detroit Traffic Reports to Your Email & Cell Phone
www.Traffic.com/Detroit

July 1 2006
Find Info On Hazardous Substances Restrictions at Business.Com
www.business.com

January 1st
Get the Facts About **January 1st**.
WiseTo.com/January+1st

Current Practice – Google



Advanced news search About Google

Find results
with **all** of the words
with the **exact phrase**
with **at least one** of the words
without the words

Sort by relevance
Google Search

Date
● Return articles added to Google News anytime
○ Return articles added to Google News between 4 Nov and 4 Dec
New! Try archive search for articles added more than 30 days ago

News source
Return only articles from the **news source** named
e.g. CNN, New York Times

News Source Location
or only articles from **news sources** located in
A country or a U.S. state

Location
Return only articles about a **local area**
City, State or Zip code

Occurrences
Return results where my terms occur
anywhere in the article

©2008 Google

Why News Articles?

- Contains reasonably factual content
- Articles printed regularly
- Articles have relatively standard format

Target Audience

- Researchers (e.g., reporters)
- Intelligence agents
- Real estate agents
- Marketers
- …

Semester Goals

- Design the system
- Build a working prototype
- Lay a foundation for future work

- Focus – search based on time and location

Approach

- Consider functional units of system
- Consider IPRO goals

- Make a team for each functional unit
    - Consider strengths/interests of students

Other Tasks

- Required everyone to do research and present on related topics, such as the semantic web
  - Developed soft skills (reading, writing, presenting)
  - Learned about the problems and benefits of building such a system

Students

- Nick Bathum (CS, 4th yr)
- Evan Estola (CS, 3rd yr)
- Jaeyeon Kihm (CS, 4th yr)
- Jay Mundrawala (ECE, 2nd yr)
- Yacin Nadji (CS, 4th yr)
- Chris Osswald (BME, 3rd yr)
- Pete Schmitz (CS, 4th yr)
- Cameron Zangenehzadeh (PS, 4th yr)

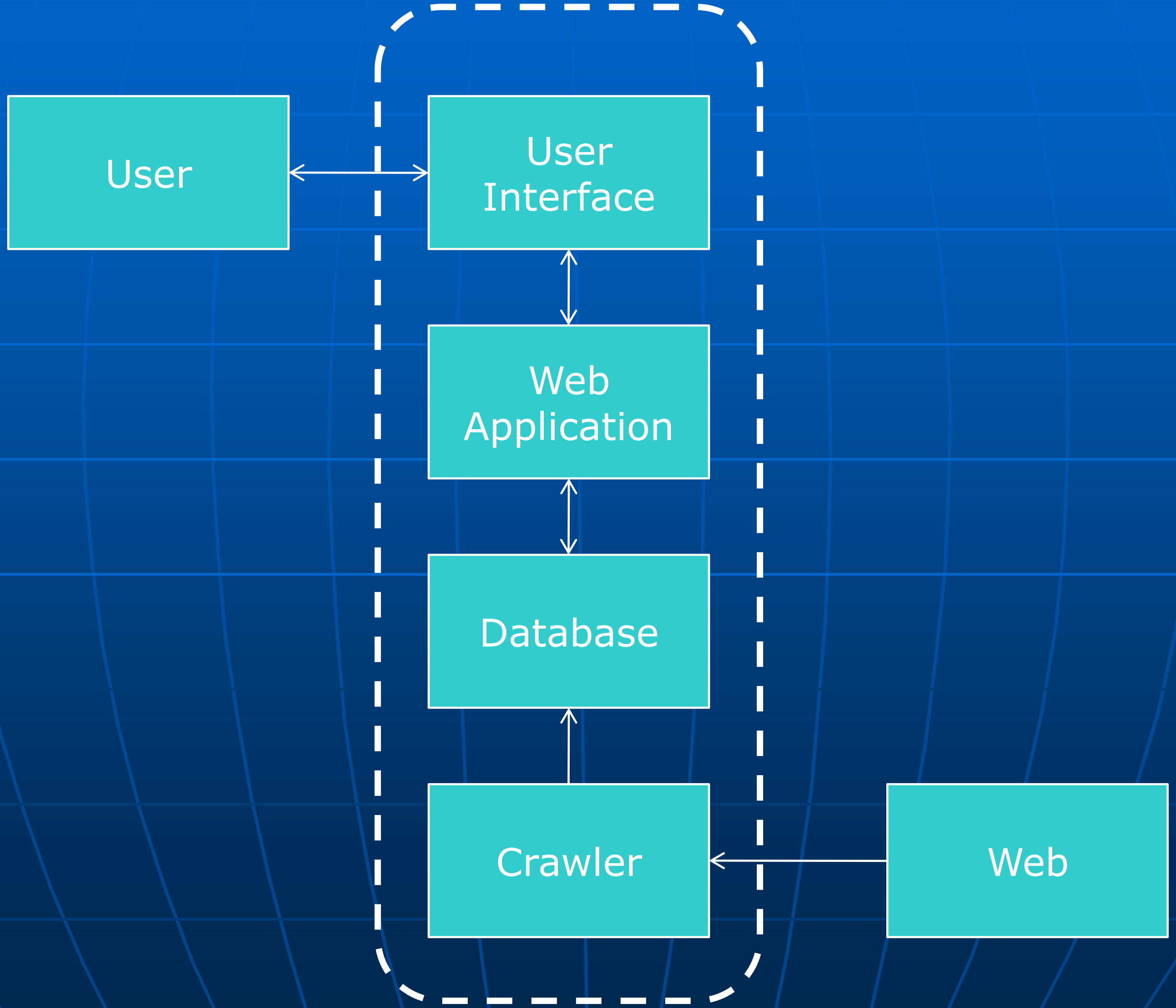
Organization

- Developed Plan, Timeline
- Used a free Wiki service for discussion, file-sharing and status reporting
  - Centralized information
  - Avoid mass emails and information loss

Functional Units

- Design (User Interface)
  - Usability and presentability
- Web Application Architecture
  - Build infrastructure, pull together work of others
- Entity Database
  - Pull “machine readable” information from articles
- Web Crawling
  - Collection of news articles from the Web

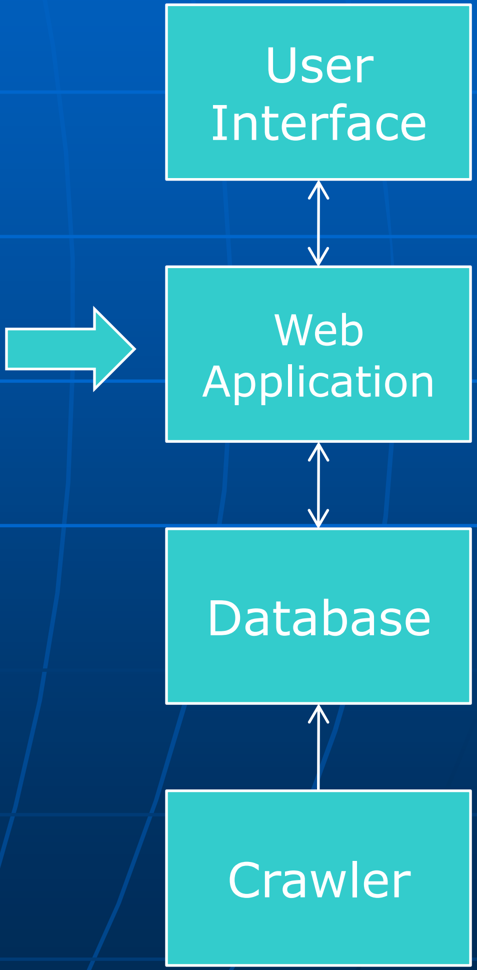
Functional Unit Relationship



Architecture Group

Evan Estola

Yacin Nadji



Architecture Group

- Responsible for building Web infrastructure
  - Interconnect other functional units
- Combine work of other groups

Architecture Approach

- Use leading open source software
  - Sesame semantic database for entities
  - MySQL database for crawled article data
  - Java Server Pages, Java Beans and Tomcat

Architecture Challenges

- Little to no Web development experience
- Working with other groups
  - Dependency on other groups
  - Communication

Architecture Outcome

- Success!
  - Working prototype
  - Dynamically builds interface (from Design team) to display article data (from Crawling team) and entities (from Enitity Extraction team)
  - Learned to learn

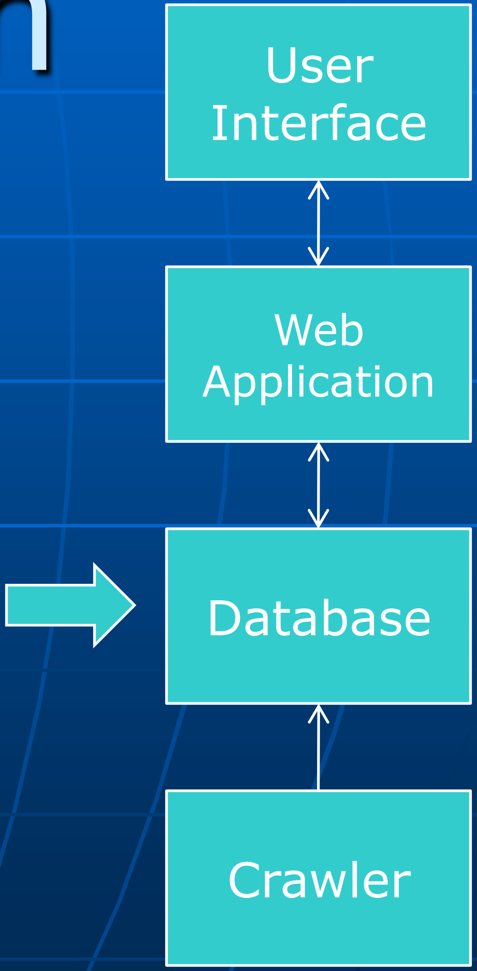
Architecture Future Work

- Optimization (speed)
- Build entity display system

Entity Extraction Team

Jaeyeon Kihm

Jay Mundrawala



Entity Extraction - Goals

- Given news articles, identify
  - Persons
  - Locations
  - Organizations
  - Dates
- Build database for extracted data

Entity Extraction – Entity Recognition

Dr. Wai Gen Yee is a professor in Department of Computer Science at Illinois Institute of Technology in Chicago

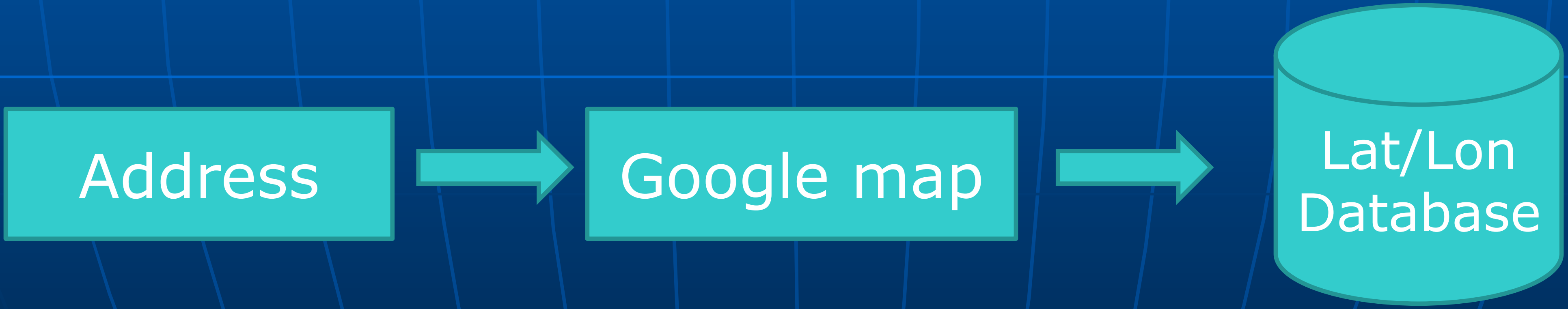
Entity Extraction – Entity Recognition

<Person>Dr. Wai Gen Yee is a professor in <Org>Department of Computer Science at <Org>Illinois Institute of Technology in <Location>Chicago

- Implemented leading open source software (Stanford NER API)
- Person, Location, Organization, Date, Title entities are extracted successfully

Lat/Lon Coordinate Point

- Generate latitude / longitude coordinate points from location entity.



Entity Extraction - Accomplishments

- Accomplishment:
  - Extract entities from articles
  - Build database to store entities and news articles
  - Generate Lat/Lon coordinate points from address

Challenges

- Stanford NER does not recognize addresses
  - Wrote custom module
- Poor database performance
  - Added a second custom database

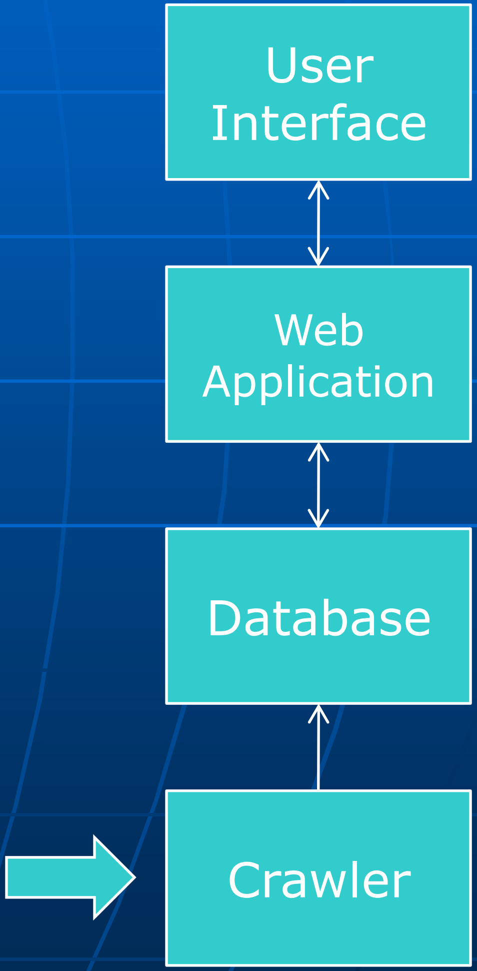
Entity Extraction – Future works

- Future work
  - Unification and disambiguation of entities
    - Ex: “George Bush" vs. “President Bush”
- Ranking of retrieved results
- Support for different types of entities (e.g., sports teams)
- Support for relationships among entities

Web Crawling Team

Nick Bathum

Pete Schmitz



User Interface

Web Application

Database

Crawler

Web Crawling Goal

- Retrieve news articles from the Web
- Submit news articles to Entity Extraction system

Web Crawling Approach

- Select correct software for crawling
  - Nutch selected for its support and scalability
- Determine which news source and how to acquire correct article data

Web Crawling Approach

- Get article links
- Use Nutch software to fetch the articles
- Parse relevant content from articles

Web Crawling Challenges

- Automated extraction of relevant content from Web pages filled with extraneous info
  - Need to identify Web page structure

Web Crawling Outcome

- Functioning article retrieval system
- Retrieve articles from
  - Reuters
  - LA Times

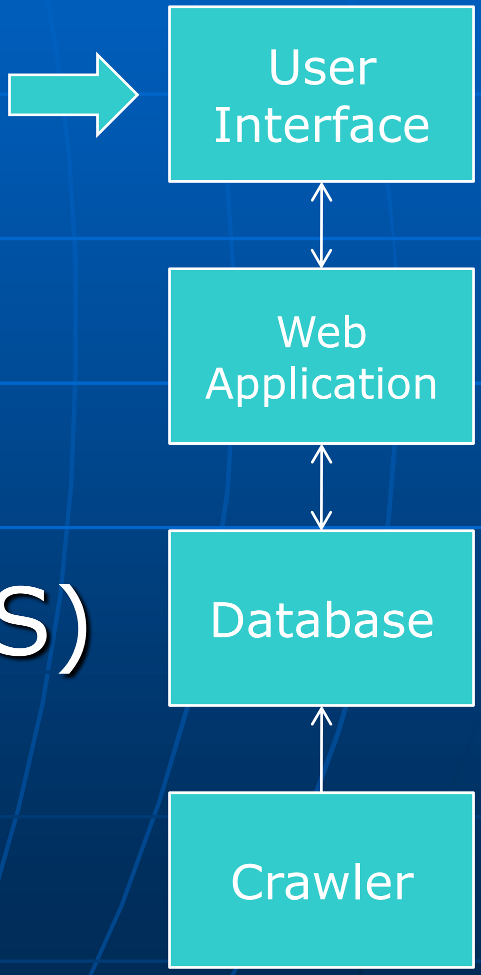
Web Crawling Future Work

- Support for other news sources

Design Team

Chris Osswald (BME)

Cam Zangenehzadeh (PS)



User Interface

Web Application

Database

Crawler

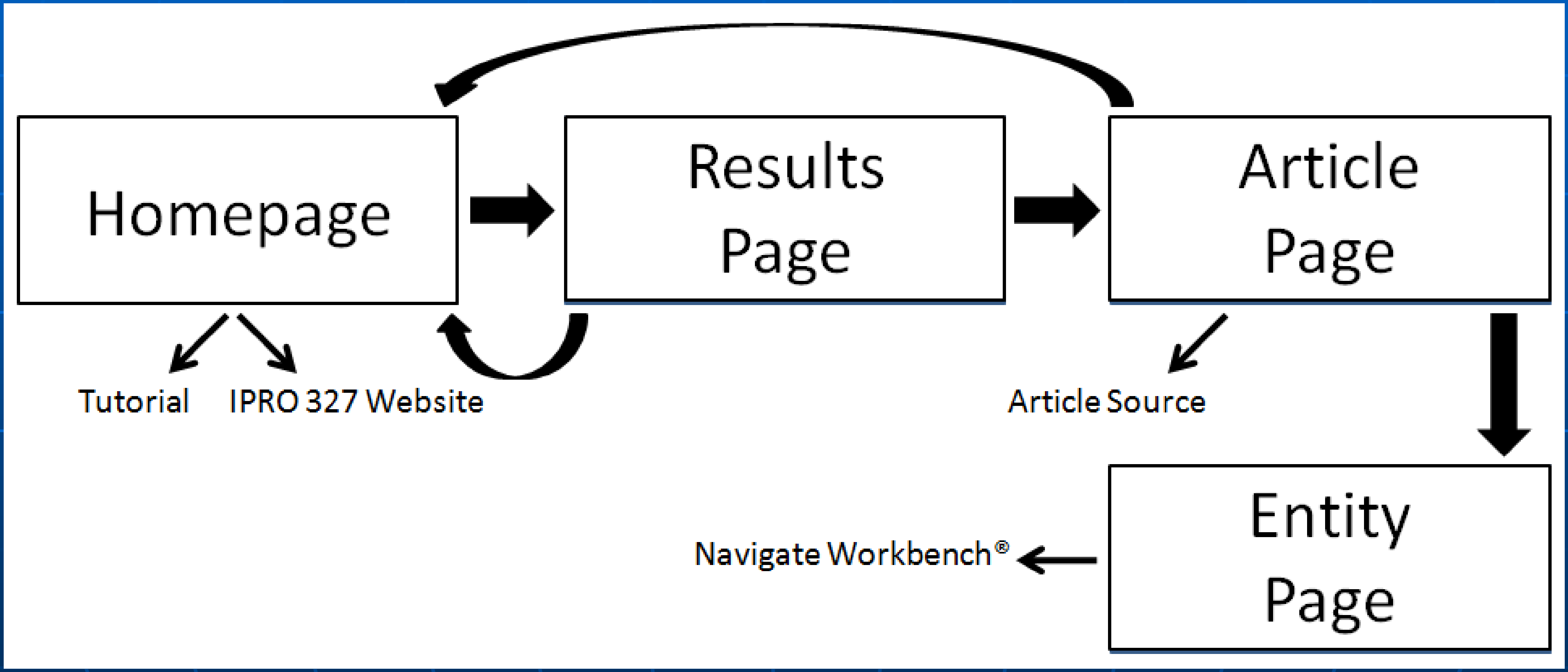
What do we want to do?

- Show off implemented functionality
- Design a simple, user-friendly interface

Approach

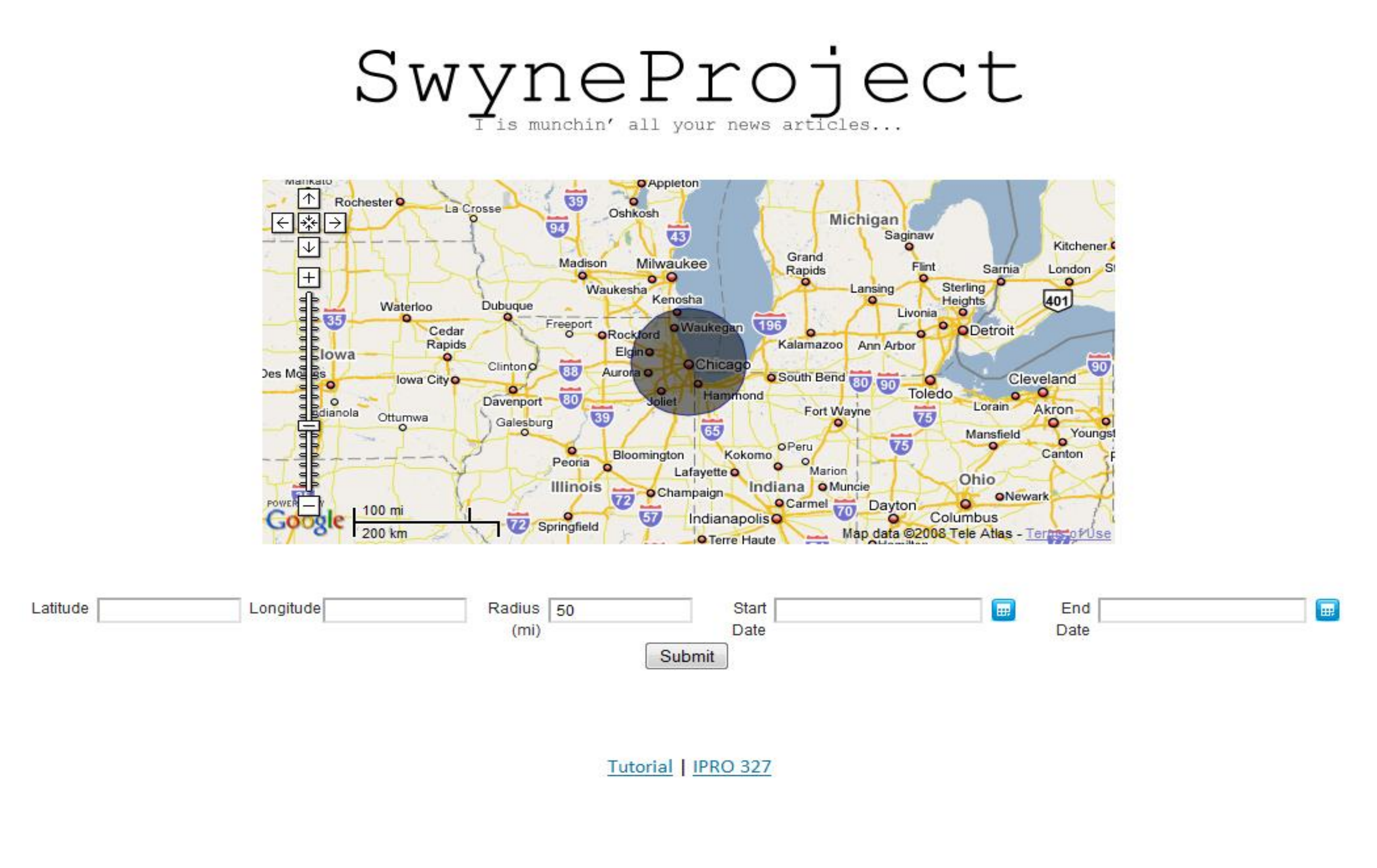
- Looked at other systems (EveryBlock.com, PowerSet.com, etc.)
- Identify core functionality
- Developed appropriate interface
  - Shows core function of system

Flow of Information



The Home Page

- Simple, user-friendly interface for entering query parameters



SwyneProject

I is munchin' all your news articles...

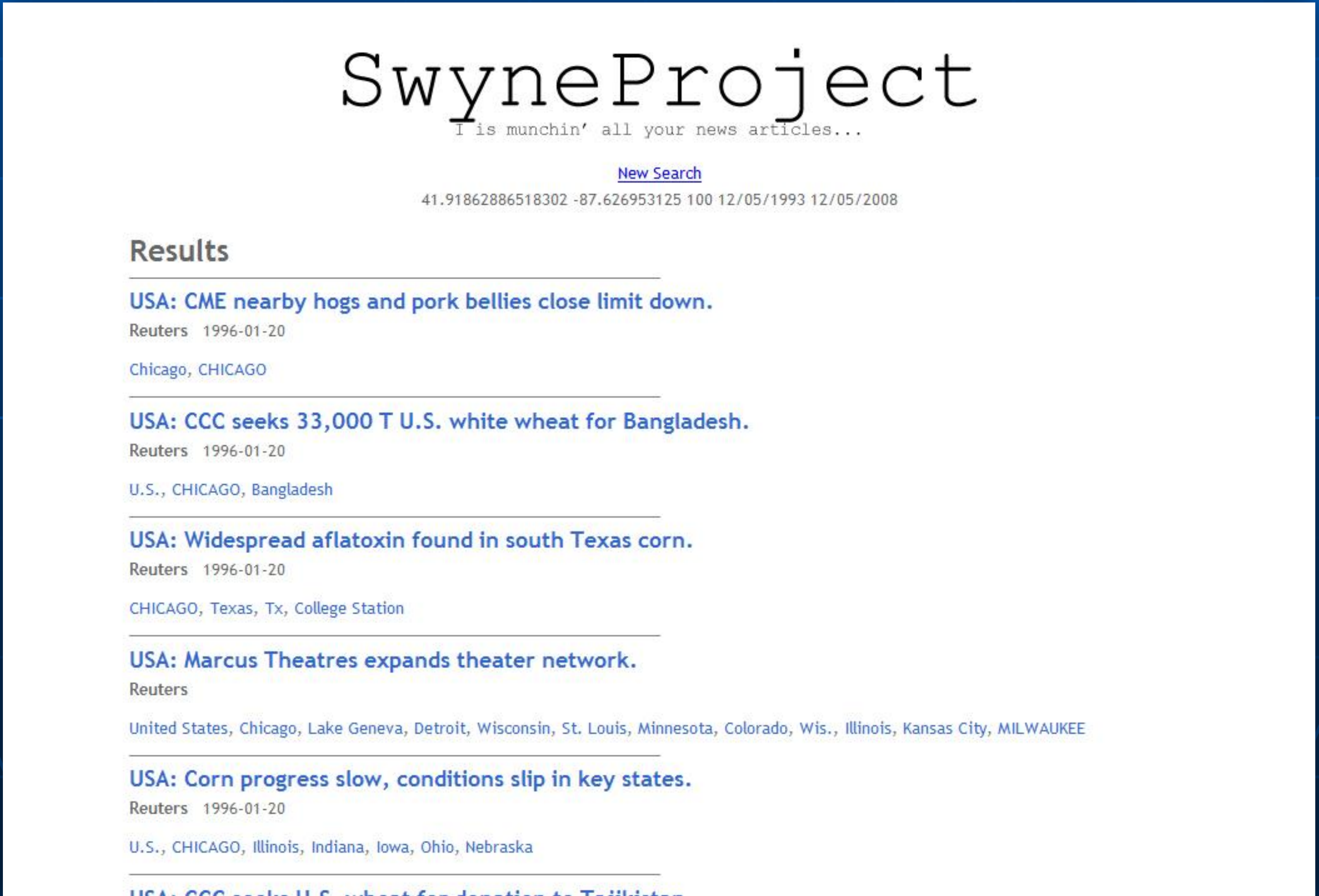
Latitude		Longitude		Radius (mi)	50	Start Date		End Date	

Submit

Tutorial | IPRO 327

The Results Page

- Articles listed in familiar format



SwyneProject

I is munchin' all your news articles...

New Search

41.91862886518302 -87.626953125 100 12/05/1993 12/05/2008

Results

USA: CME nearby hogs and pork bellies close limit down.

Reuters 1996-01-20

Chicago, CHICAGO

USA: CCC seeks 33,000 T U.S. white wheat for Bangladesh.

Reuters 1996-01-20

U.S., CHICAGO, Bangladesh

USA: Widespread aflatoxin found in south Texas corn.

Reuters 1996-01-20

CHICAGO, Texas, Tx, College Station

USA: Marcus Theatres expands theater network.

Reuters

United States, Chicago, Lake Geneva, Detroit, Wisconsin, St. Louis, Minnesota, Colorado, Wis., Illinois, Kansas City, MILWAUKEE

USA: Corn progress slow, conditions slip in key states.

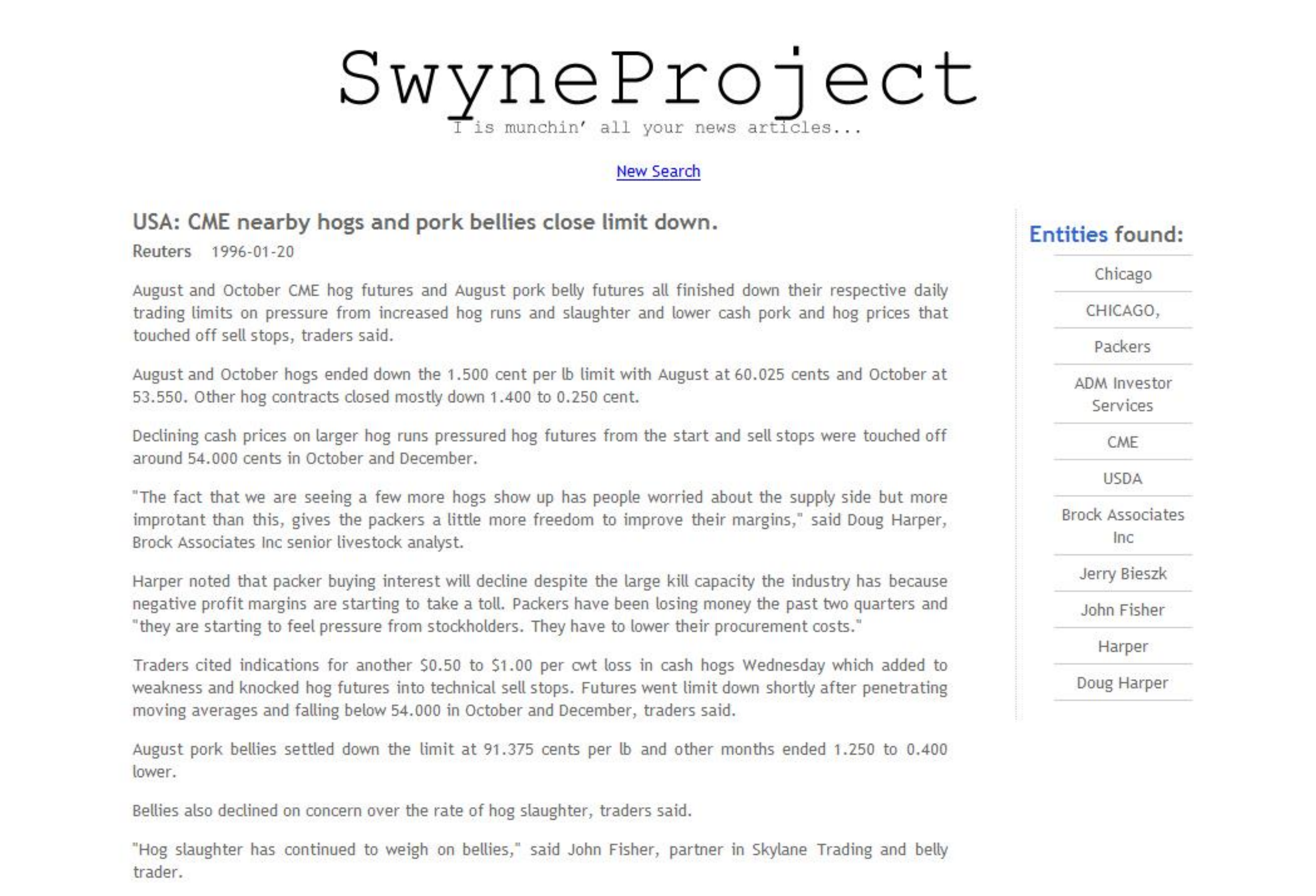
Reuters 1996-01-20

U.S., CHICAGO, Illinois, Indiana, Iowa, Ohio, Nebraska

USA: CCC seeks U.S. wheat for donation to Tajikistan

The Article Page

- Articles shown in familiar format
- Offset with entity list



SwyneProject

I is munchin' all your news articles...

New Search

USA: CME nearby hogs and pork bellies close limit down.

Reuters 1996-01-20

August and October CME hog futures and August pork belly futures all finished down their respective daily trading limits on pressure from increased hog runs and slaughter and lower cash pork and hog prices that touched off sell stops, traders said.

August and October hogs ended down the 1.500 cent per lb limit with August at 60.025 cents and October at 53.550. Other hog contracts closed mostly down 1.400 to 0.250 cent.

Declining cash prices on larger hog runs pressured hog futures from the start and sell stops were touched off around 54.000 cents in October and December.

"The fact that we are seeing a few more hogs show up has people worried about the supply side but more improtant than this, gives the packers a little more freedom to improve their margins," said Doug Harper, Brock Associates Inc senior livestock analyst.

Harper noted that packer buying interest will decline despite the large kill capacity the industry has because negative profit margins are starting to take a toll. Packers have been losing money the past two quarters and "they are starting to feel pressure from stockholders. They have to lower their procurement costs."

Traders cited indications for another $0.50 to $1.00 per cwt loss in cash hogs Wednesday which added to weakness and knocked hog futures into technical sell stops. Futures went limit down shortly after penetrating moving averages and falling below 54.000 in October and December, traders said.

August pork bellies settled down the limit at 91.375 cents per lb and other months ended 1.250 to 0.400 lower.

Bellies also declined on concern over the rate of hog slaughter, traders said.

"Hog slaughter has continued to weigh on bellies," said John Fisher, partner in Skylane Trading and belly trader.

Entities found:

Chicago

CHICAGO,

Packers

ADM Investor Services

CME

USDA

Brock Associates Inc

Jerry Bieszk

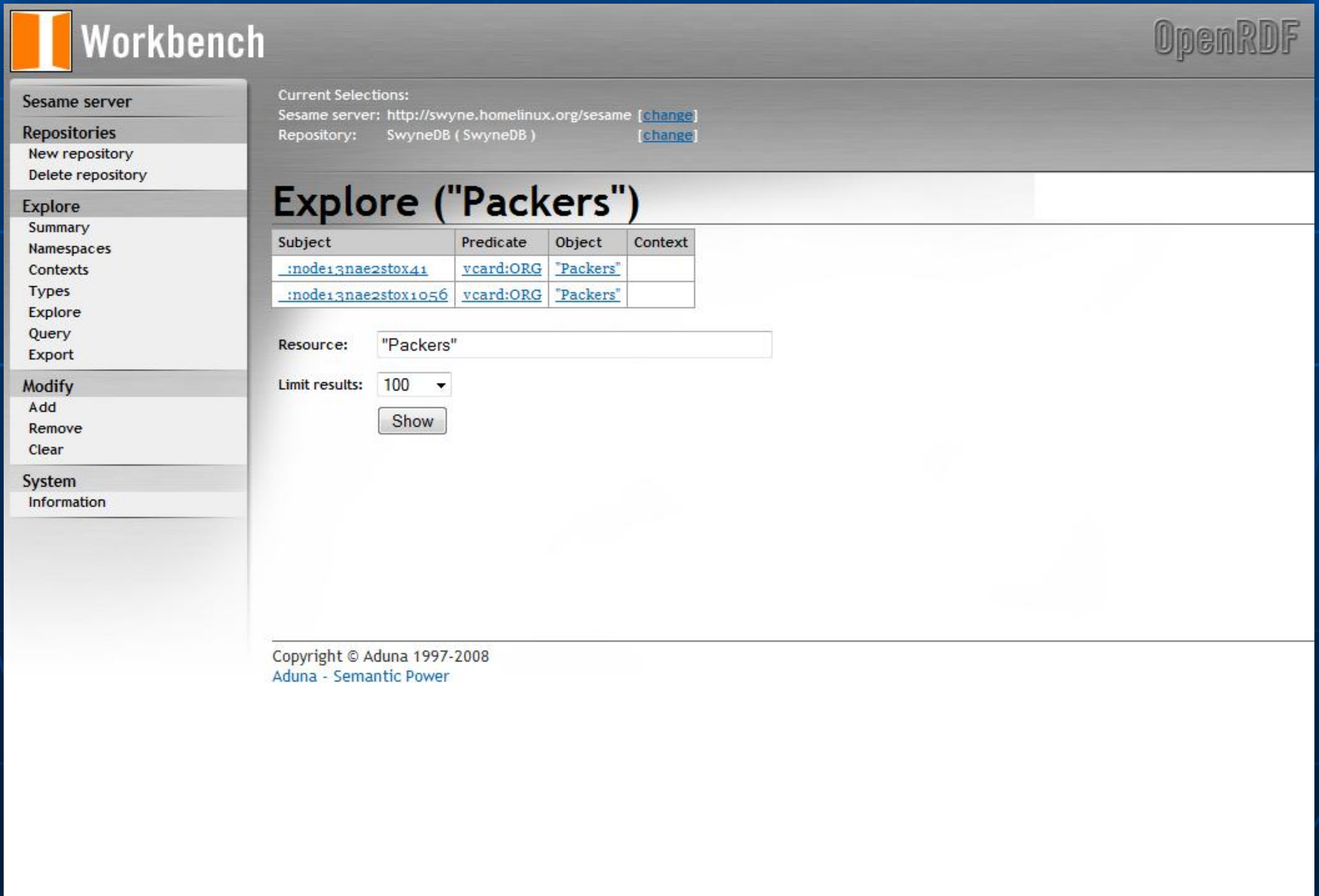
John Fisher

Harper

Doug Harper

Exploring Entities

- Allows user to explore relationships



Current Selections:
Sesame server: http://swyne.homelinux.org/sesame [change]
Repository: SwyneDB (SwyneDB) [change]

Explore ("Packers")

Subject	Predicate	Object	Context
_:node13nae2stox41	vcard:ORG	"Packers"	
_:node13nae2stox1056	vcard:ORG	"Packers"	

Resource: "Packers"

Limit results: 100

Copyright © Aduna 1997-2008
Aduna - Semantic Power

Accomplishments

- Created simple interface
  - Learned Web design
- Access to all implemented functionality
- Created IPRO Web site

Issues Encountered

- Scoping: What is considered "window dressing?"
- Coding of the website itself
- Soft skills
  - Communication
  - Patience
  - Delegating tasks

Future Work

- Support for additional functionality
  - Categorized results
  - Add search support for additional entity types

Ethical Issues

- Professional Issues
  - Software Copyright
  - Content Copyright
  - Terms of Use
- Internal Issues
  - Scheduling Conflicts
  - Interpersonal Conflicts

Group Accomplishments

- Developed working prototype
- Published deliverables on the Web
- Outlined future work

Societal Impact

- Inspire future search engines to use semantic data to provide users with specific answers
  - Improve access to information
  - Published our design documents online

Future Work

- Improve overall speed of system
- Improve accuracy of entity extraction
- Support new the types of queries
- Allow crawling of more types of data sources

Questions & Answers