IEEE International Conference on Intelligence and Security Informatics
Public Safety and Security, ISI 2010

23-26 May, Vancouver, BC, Canada

Computational Knowledge & Information Management in Veterinary Epidemiology

Svitlana Volkova and William H. Hsu

Laboratory for Knowledge Discovery in Databases

Department of Computing and Information Sciences

Kansas State University

Sponsors: K-State National Agricultural Biosecurity Center (NABC)
US Department of Defense

Agenda

- Overview
- Animal Disease Monitoring Systems
  - Manually Supported Web-Interfaces
  - Automated Web-Services
- Framework for Epidemiological Analytics
  - Web Crawling & Search
  - Domain-specific Entity Extraction
  - Animal Disease-related Event Recognition
- Summary

Animal Infectious Disease Outbreaks

- influence on the travel and trade
- cause economic crises, political instability
- diseases, zoonotic in type can cause loss of life

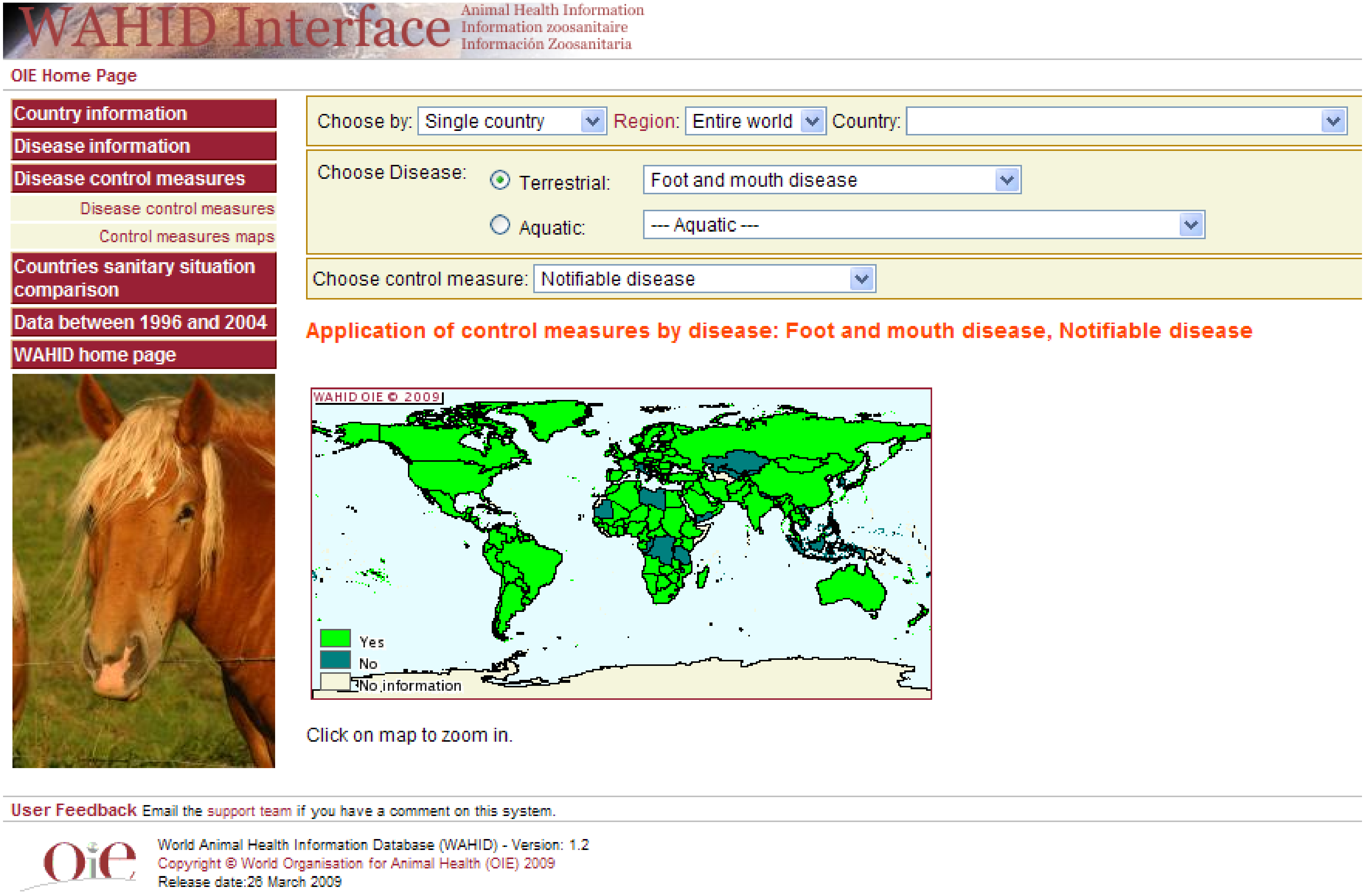
Agenda

- Overview
- Animal Disease Monitoring Systems
  - Manually Supported Web-Interfaces
  - Automated Web-Services
- Framework for Epidemiological Analytics
  - System Functionality
  - Web Crawling
  - Domain-specific Entity Extraction
  - Animal Disease-related Event Recognition
- Summary

Animal Disease Monitoring Systems: Manually Supported Web Interfaces (1)

International:

- World Animal Health Information Database (WAHID) Interface - http://www.oie.int/wahis/public.php?page=home
- WHO Global Atlas of Infectious Diseases - http://diseasemaps.usgs.gov/index.htm
- Emergency Prevention System (EMPRES) for Transboundary Animal and Plant Pests and Diseases - http://www.fao.org/EMPRES/default.html



WAHID Interface

Animal Health Information
Information zoosanitaire
Información Zoosanitaria

OIE Home Page

Choose by: Single country Region: Entire world Country:

Choose Disease: ● Terrestrial: Foot and mouth disease
○ Aquatic: --- Aquatic ---

Choose control measure: Notifiable disease

Application of control measures by disease: Foot and mouth disease, Notifiable disease

Click on map to zoom in.

User Feedback Email the support team if you have a comment on this system.

World Animal Health Information Database (WAHID) - Version: 1.2
Copyright © World Organisation for Animal Health (OIE) 2009
Release date:26 March 2009

Animal Disease Monitoring Systems: Manually Supported Web Interfaces(2)

USA

- Centers for Disease Control and Prevention (CDC) - http://www.cdc.gov
- U.S. Department of Agriculture (USDA) - http://www.usda.gov/wps/portal/usdahome
- U.S. Geological Survey (USGS) and U.S. Geological Survey (USGS) National Wildlife Health Center (NWHC) - http://www.nwhc.usgs.gov
- Iowa State University Center for Food Security and Public Health (CFSPH) - http://www.cfsph.iastate.edu
- BioPortal - http://biocomputingcorp.com/bpsystem.html
- FMD BioPortal - https://fmdbioportal.ucdavis.edu

United Kingdom

- Department for Environment Food and Rural Affairs (DEFRA) - http://www.defra.gov.uk

Agenda

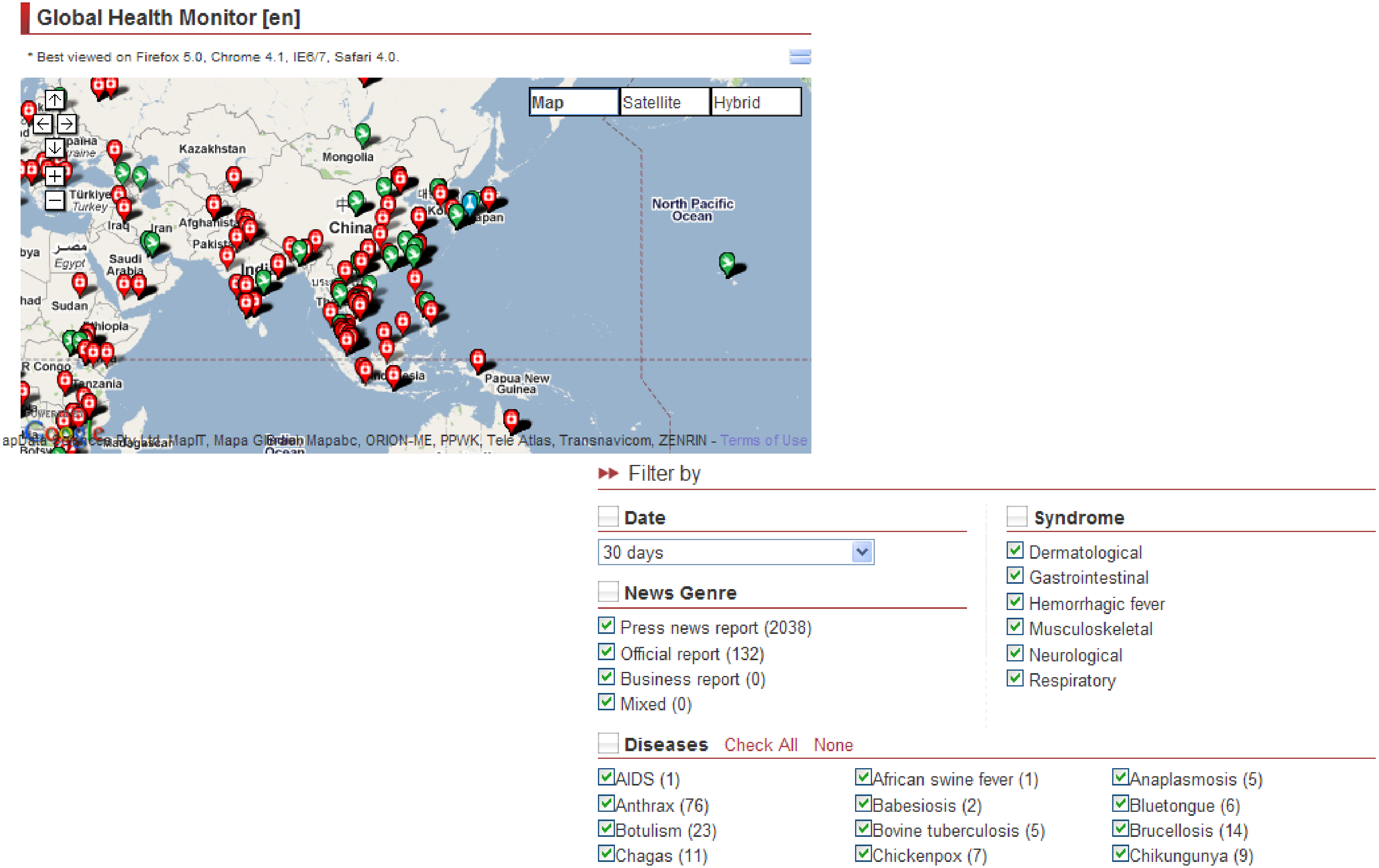
- Overview
- Animal Disease Monitoring Systems
  - Manually Supported Web-Interfaces
  - Automated Web-Services
- Framework for Epidemiological Analytics
  - System Functionality
  - Web Crawling
  - Domain-specific Entity Extraction
  - Animal Disease-related Event Recognition
- Summary

Animal Disease Monitoring Systems: Automated Web Services (1)

BioCaster - http://biocaster.nii.ac.jp/

- follows 1500 RSS feeds hourly
- classifies documents as topically relevant or not
- taxonomy of 4300 named entities (50 disease names, 243 country names, 4025 province/city names, latitudes and longitudes)
- identifies 40 diseases at up to 25-30 locations per day
- multilingual information extraction on to English, French, Spanish, Chinese, Thai, Vietnamese, Japanese
- uses ontology pattern matching approaches to recognize disease-location-verb pairs
- plots events on a Google Map
- does not classify events into categories and does not report past outbreaks
- no timeline visualization

BioCaster - http://biocaster.nii.ac.jp/



Global Health Monitor [en]

* Best viewed on Firefox 5.0, Chrome 4.1, IE6/7, Safari 4.0.

▸▸ Filter by

□ **Date**

30 days

□ **News Genre**

☑ Press news report (2038)
☑ Official report (132)
☑ Business report (0)
☑ Mixed (0)

□ **Syndrome**

☑ Dermatological
☑ Gastrointestinal
☑ Hemorrhagic fever
☑ Musculoskeletal
☑ Neurological
☑ Respiratory

□ **Diseases** Check All None

☑AIDS (1)
☑Anthrax (76)
☑Botulism (23)
☑Chagas (11)

☑African swine fever (1)
☑Babesiosis (2)
☑Bovine tuberculosis (5)
☑Chickenpox (7)

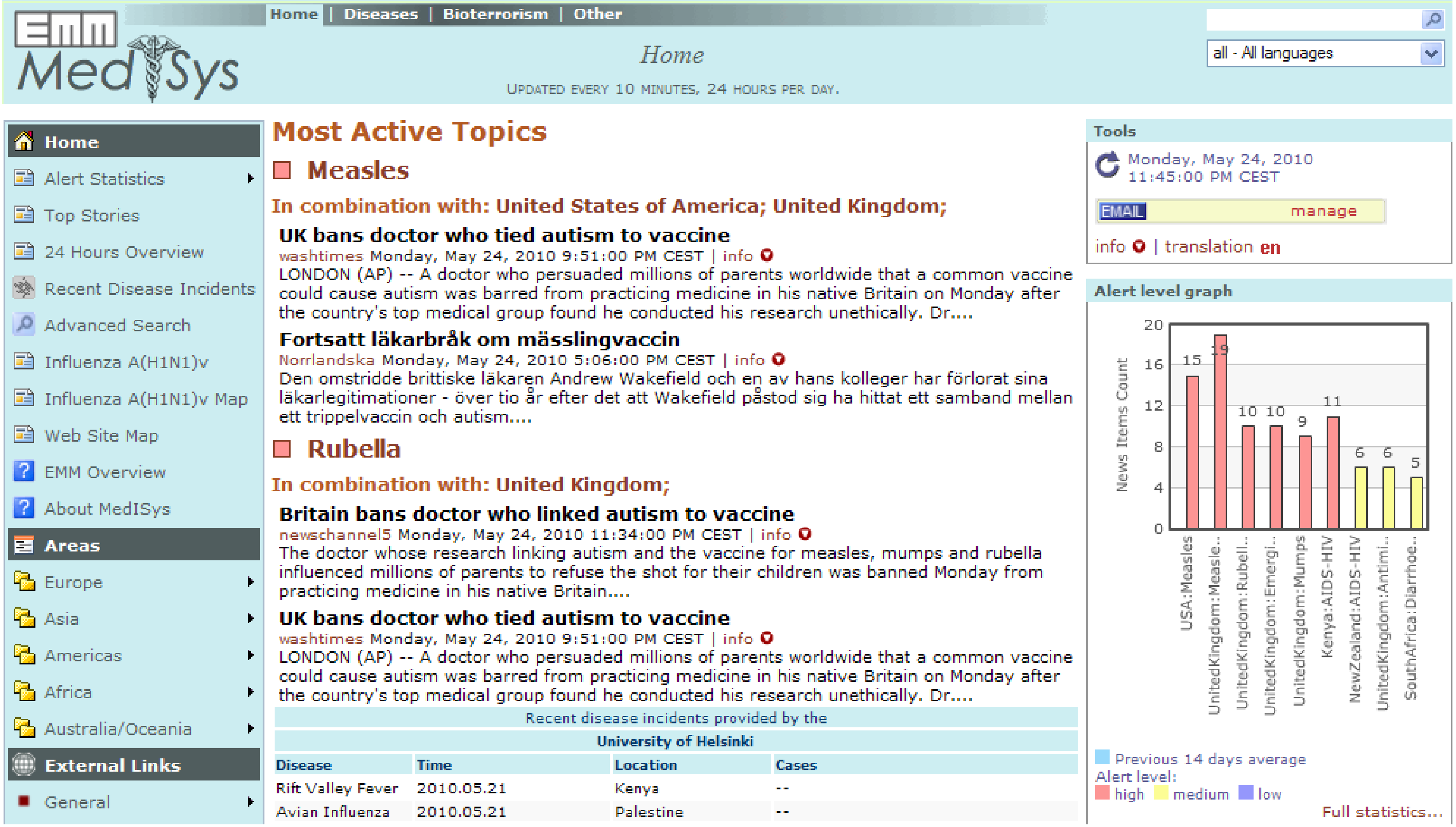
☑Anaplasmosis (5)
☑Bluetongue (6)
☑Brucellosis (14)
☑Chikungunya (9)

Animal Disease Monitoring Systems: Automated Web Services (1)

- **Information retrieval system MedISys -** http://medusa.jrc.it/medisys/homeedition/all/home.html

- **Pattern-based Understanding and Learning System (PULS) -** http://sysdb.cs.helsinki.fi/puls/jrc/all
  - allows automated recognizing of the metadata and structured facts related to the disease outbreaks
  - collects an average 50000 news articles per day from about 1400 news portals and about 150 specialized Public Health sites
  - 43 languages
  - current ontology contains 2400 disease names, 400 organisms, 1500 political entities and over 70000 location names including towns, cities, provinces
  - real-time news clustering and filtering by matching 3000 patterns
  - does not classify events and does not report past outbreaks.

MedISys - http://medusa.jrc.it/medisys/homeedition/all/home.html



EMM MedISys

Home | Diseases | Bioterrorism | Other

all - All languages

Home

UPDATED EVERY 10 MINUTES, 24 HOURS PER DAY.

- **Home**
- Alert Statistics
- Top Stories
- 24 Hours Overview
- Recent Disease Incidents
- Advanced Search
- Influenza A(H1N1)v
- Influenza A(H1N1)v Map
- Web Site Map
- EMM Overview
- About MedISys
- **Areas**
- Europe
- Asia
- Americas
- Africa
- Australia/Oceania
- **External Links**
- General

Most Active Topics

Measles

In combination with: United States of America; United Kingdom;

UK bans doctor who tied autism to vaccine
washtimes Monday, May 24, 2010 9:51:00 PM CEST | info
LONDON (AP) -- A doctor who persuaded millions of parents worldwide that a common vaccine could cause autism was barred from practicing medicine in his native Britain on Monday after the country's top medical group found he conducted his research unethically. Dr....

Fortsatt läkarbråk om mässlingvaccin
Norrlandska Monday, May 24, 2010 5:06:00 PM CEST | info
Den omstridde brittiske läkaren Andrew Wakefield och en av hans kolleger har förlorat sina läkarlegitimationer - över tio år efter det att Wakefield påstod sig ha hittat ett samband mellan ett trippelvaccin och autism....

Rubella

In combination with: United Kingdom;

Britain bans doctor who linked autism to vaccine
newschannel5 Monday, May 24, 2010 11:34:00 PM CEST | info
The doctor whose research linking autism and the vaccine for measles, mumps and rubella influenced millions of parents to refuse the shot for their children was banned Monday from practicing medicine in his native Britain....

UK bans doctor who tied autism to vaccine
washtimes Monday, May 24, 2010 9:51:00 PM CEST | info
LONDON (AP) -- A doctor who persuaded millions of parents worldwide that a common vaccine could cause autism was barred from practicing medicine in his native Britain on Monday after the country's top medical group found he conducted his research unethically. Dr....

Recent disease incidents provided by the
University of Helsinki

Disease	Time	Location	Cases
Rift Valley Fever	2010.05.21	Kenya	--
Avian Influenza	2010.05.21	Palestine	--

Tools

Monday, May 24, 2010
11:45:00 PM CEST

EMAIL manage

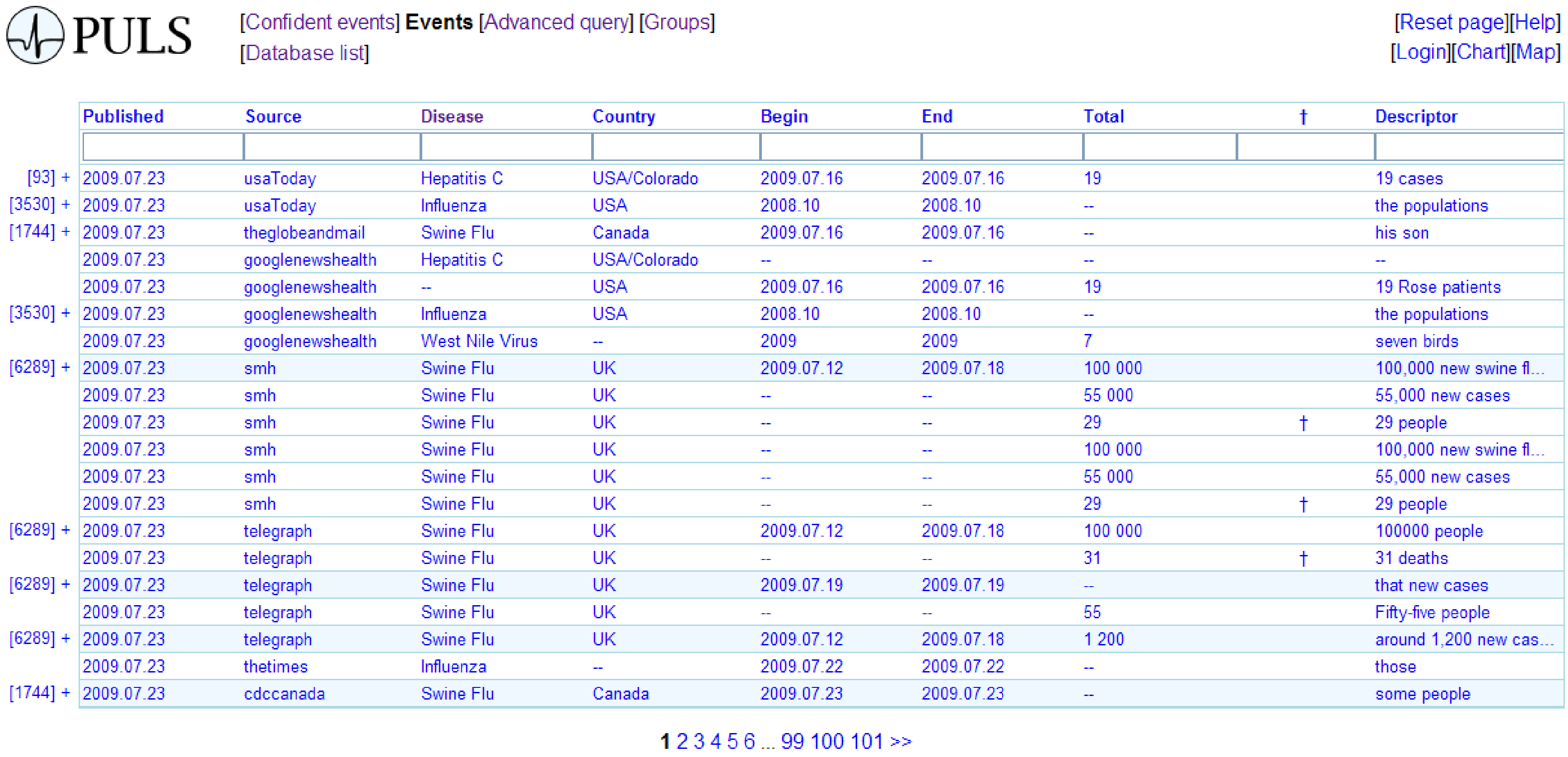
info | translation en

Alert level graph

Full statistics...

*part of the Europe Media Monitor (EMM) product family
http://emm.jrc.it/overview.html

Pattern-based Understanding and Learning System (PULS)



PULS

[Confident events] **Events** [Advanced query] [Groups]
[Database list]

[Reset page][Help]
[Login][Chart][Map]

	Published	Source	Disease	Country	Begin	End	Total	†	Descriptor
[93] +	2009.07.23	usaToday	Hepatitis C	USA/Colorado	2009.07.16	2009.07.16	19		19 cases
[3530] +	2009.07.23	usaToday	Influenza	USA	2008.10	2008.10	--		the populations
[1744] +	2009.07.23	theglobeandmail	Swine Flu	Canada	2009.07.16	2009.07.16	--		his son
	2009.07.23	googlenewshealth	Hepatitis C	USA/Colorado	--	--	--		--
	2009.07.23	googlenewshealth	--	USA	2009.07.16	2009.07.16	19		19 Rose patients
[3530] +	2009.07.23	googlenewshealth	Influenza	USA	2008.10	2008.10	--		the populations
	2009.07.23	googlenewshealth	West Nile Virus	--	2009	2009	7		seven birds
[6289] +	2009.07.23	smh	Swine Flu	UK	2009.07.12	2009.07.18	100 000		100,000 new swine fl...
	2009.07.23	smh	Swine Flu	UK	--	--	55 000		55,000 new cases
	2009.07.23	smh	Swine Flu	UK	--	--	29	†	29 people
	2009.07.23	smh	Swine Flu	UK	--	--	100 000		100,000 new swine fl...
	2009.07.23	smh	Swine Flu	UK	--	--	55 000		55,000 new cases
	2009.07.23	smh	Swine Flu	UK	--	--	29	†	29 people
[6289] +	2009.07.23	telegraph	Swine Flu	UK	2009.07.12	2009.07.18	100 000		100000 people
	2009.07.23	telegraph	Swine Flu	UK	--	--	31	†	31 deaths
[6289] +	2009.07.23	telegraph	Swine Flu	UK	2009.07.19	2009.07.19	--		that new cases
	2009.07.23	telegraph	Swine Flu	UK	--	--	55		Fifty-five people
[6289] +	2009.07.23	telegraph	Swine Flu	UK	2009.07.12	2009.07.18	1 200		around 1,200 new cas...
	2009.07.23	thetimes	Influenza	--	2009.07.22	2009.07.22	--		those
[1744] +	2009.07.23	cdccanada	Swine Flu	Canada	2009.07.23	2009.07.23	--		some people

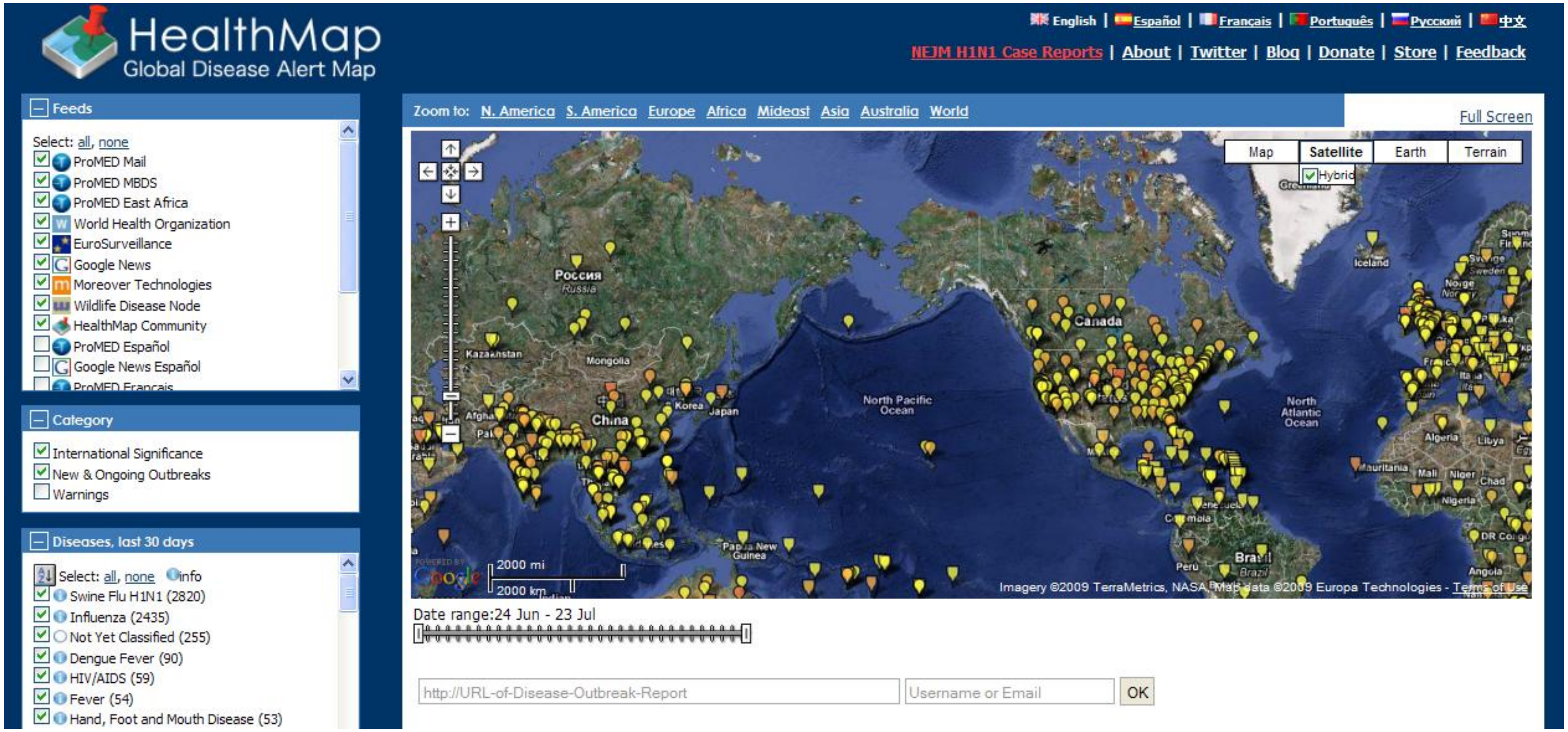
1 2 3 4 5 6 ... 99 100 101 >>

Animal Disease Monitoring Systems: Automated Web Services (1)

- **HealthMap - http://healthmap.org/en**
  - aggregates articles from Google News and ProMED-Mail portal
  - 2300 locations and 1100 disease names
  - identifies between 20-30 outbreaks per day
  - multiple languages English, Russian, Arabic, French, Portuguese, Spanish, Chinese
  - manually supported system

- **EpiSpider- http://www.epispider.org/**
  - combines emerging infectious disease data from:
    - ProMED-Mail - www.promedmail.org
    - The Global Disaster Alert Coordinating System (GDACS) - www.gdacs.org
    - Central Intelligence Agency (CIA) Factbook - https://www.cia.gov/library/publications/the-world-factbook/
    - The United Nations Human Development Report sites - http://hdr.undp.org/en

HealthMap - http://healthmap.org/en



HealthMap
Global Disease Alert Map
English | Español | Français | Português | Русский | 中文
NEJM H1N1 Case Reports | About | Twitter | Blog | Donate | Store | Feedback
Zoom to: N. America S. America Europe Africa Mideast Asia Australia World
Full Screen
Map
Satellite
Earth
Terrain
☑ Hybrid
Россия
Russia
Canada
Kazakhstan
Mongolia
China
Korea
Japan
North Pacific Ocean
North Atlantic Ocean
Iceland
Greenland
Papua New Guinea
Brazil
Peru
Colombia
Algeria
Libya
Mauritania
Mali
Niger
Chad
Nigeria
DR Congo
Angola
2000 mi
2000 km
Imagery ©2009 TerraMetrics, NASA, Map data ©2009 Europa Technologies - Terms of Use
OK

Feeds

Select: all, none

- ☑ ProMED Mail
- ☑ ProMED MBDS
- ☑ ProMED East Africa
- ☑ World Health Organization
- ☑ EuroSurveillance
- ☑ Google News
- ☑ Moreover Technologies
- ☑ Wildlife Disease Node
- ☑ HealthMap Community
- □ ProMED Español
- □ Google News Español
- □ ProMED Francais

Category

- ☑ International Significance
- ☑ New & Ongoing Outbreaks
- □ Warnings

Diseases, last 30 days

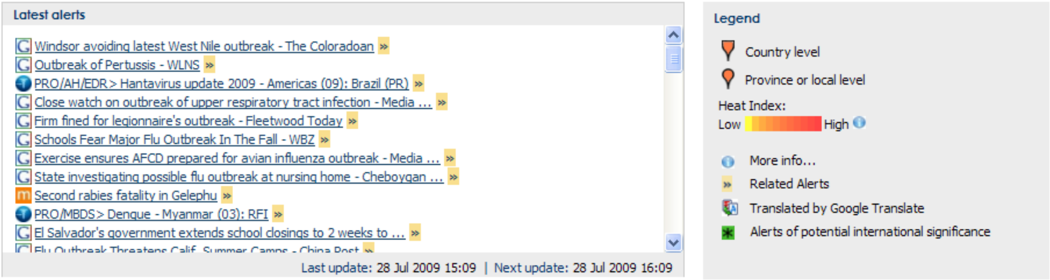
Select: all, none info

- ☑ Swine Flu H1N1 (2820)
- ☑ Influenza (2435)
- ☑ Not Yet Classified (255)
- ☑ Dengue Fever (90)
- ☑ HIV/AIDS (59)
- ☑ Fever (54)
- ☑ Hand, Foot and Mouth Disease (53)

Date range:24 Jun - 23 Jul

http://URL-of-Disease-Outbreak-Report

Username or Email



Latest alerts

- Windsor avoiding latest West Nile outbreak - The Coloradoan »
- Outbreak of Pertussis - WLNS »
- PRO/AH/EDR> Hantavirus update 2009 - Americas (09): Brazil (PR) »
- Close watch on outbreak of upper respiratory tract infection - Media ... »
- Firm fined for legionnaire's outbreak - Fleetwood Today »
- Schools Fear Major Flu Outbreak In The Fall - WBZ »
- Exercise ensures AFCD prepared for avian influenza outbreak - Media ... »
- State investigating possible flu outbreak at nursing home - Cheboygan ... »
- Second rabies fatality in Gelephu »
- PRO/MBDS> Dengue - Myanmar (03): RFI »
- El Salvador's government extends school closings to 2 weeks to ... »
- Flu Outbreak Threatens Calif. Summer Camps - China Post »

Last update: 28 Jul 2009 15:09 | Next update: 28 Jul 2009 16:09

Legend

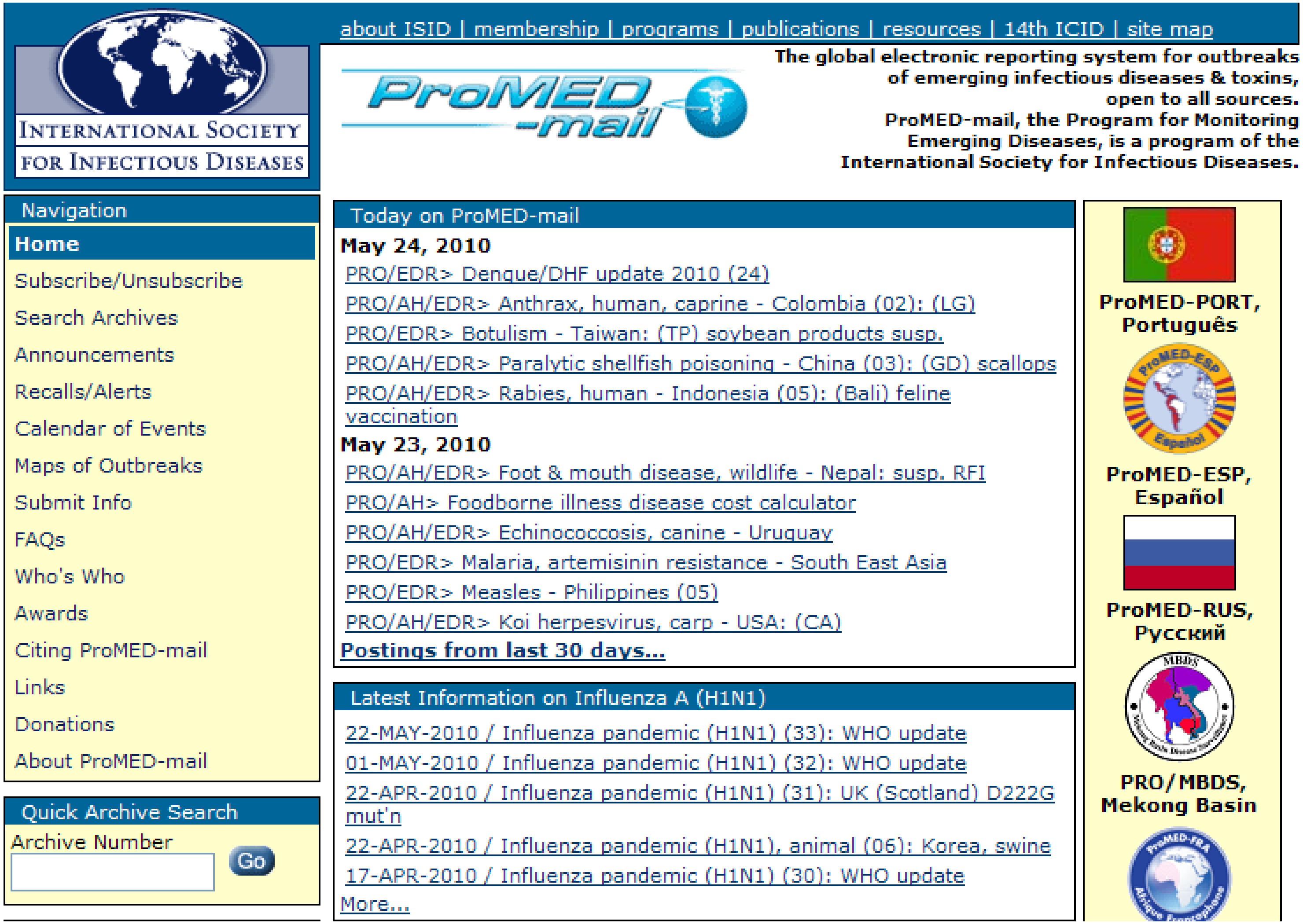
- Country level
- Province or local level

Heat Index:

Low High

- More info...
- » Related Alerts
- Translated by Google Translate
- Alerts of potential international significance

ProMED-Mail - www.promedmail.org



about ISID | membership | programs | publications | resources | 14th ICID | site map

The global electronic reporting system for outbreaks of emerging infectious diseases & toxins, open to all sources. ProMED-mail, the Program for Monitoring Emerging Diseases, is a program of the International Society for Infectious Diseases.

Navigation

- Home
- Subscribe/Unsubscribe
- Search Archives
- Announcements
- Recalls/Alerts
- Calendar of Events
- Maps of Outbreaks
- Submit Info
- FAQs
- Who's Who
- Awards
- Citing ProMED-mail
- Links
- Donations
- About ProMED-mail

Quick Archive Search

Archive Number

Go

Today on ProMED-mail

May 24, 2010

- PRO/EDR> Dengue/DHF update 2010 (24)
- PRO/AH/EDR> Anthrax, human, caprine - Colombia (02): (LG)
- PRO/EDR> Botulism - Taiwan: (TP) soybean products susp.
- PRO/AH/EDR> Paralytic shellfish poisoning - China (03): (GD) scallops
- PRO/AH/EDR> Rabies, human - Indonesia (05): (Bali) feline vaccination

May 23, 2010

- PRO/AH/EDR> Foot & mouth disease, wildlife - Nepal: susp. RFI
- PRO/AH> Foodborne illness disease cost calculator
- PRO/AH/EDR> Echinococcosis, canine - Uruguay
- PRO/EDR> Malaria, artemisinin resistance - South East Asia
- PRO/EDR> Measles - Philippines (05)
- PRO/AH/EDR> Koi herpesvirus, carp - USA: (CA)

Postings from last 30 days...

Latest Information on Influenza A (H1N1)

- 22-MAY-2010 / Influenza pandemic (H1N1) (33): WHO update
- 01-MAY-2010 / Influenza pandemic (H1N1) (32): WHO update
- 22-APR-2010 / Influenza pandemic (H1N1) (31): UK (Scotland) D222G mut'n
- 22-APR-2010 / Influenza pandemic (H1N1), animal (06): Korea, swine
- 17-APR-2010 / Influenza pandemic (H1N1) (30): WHO update

More...

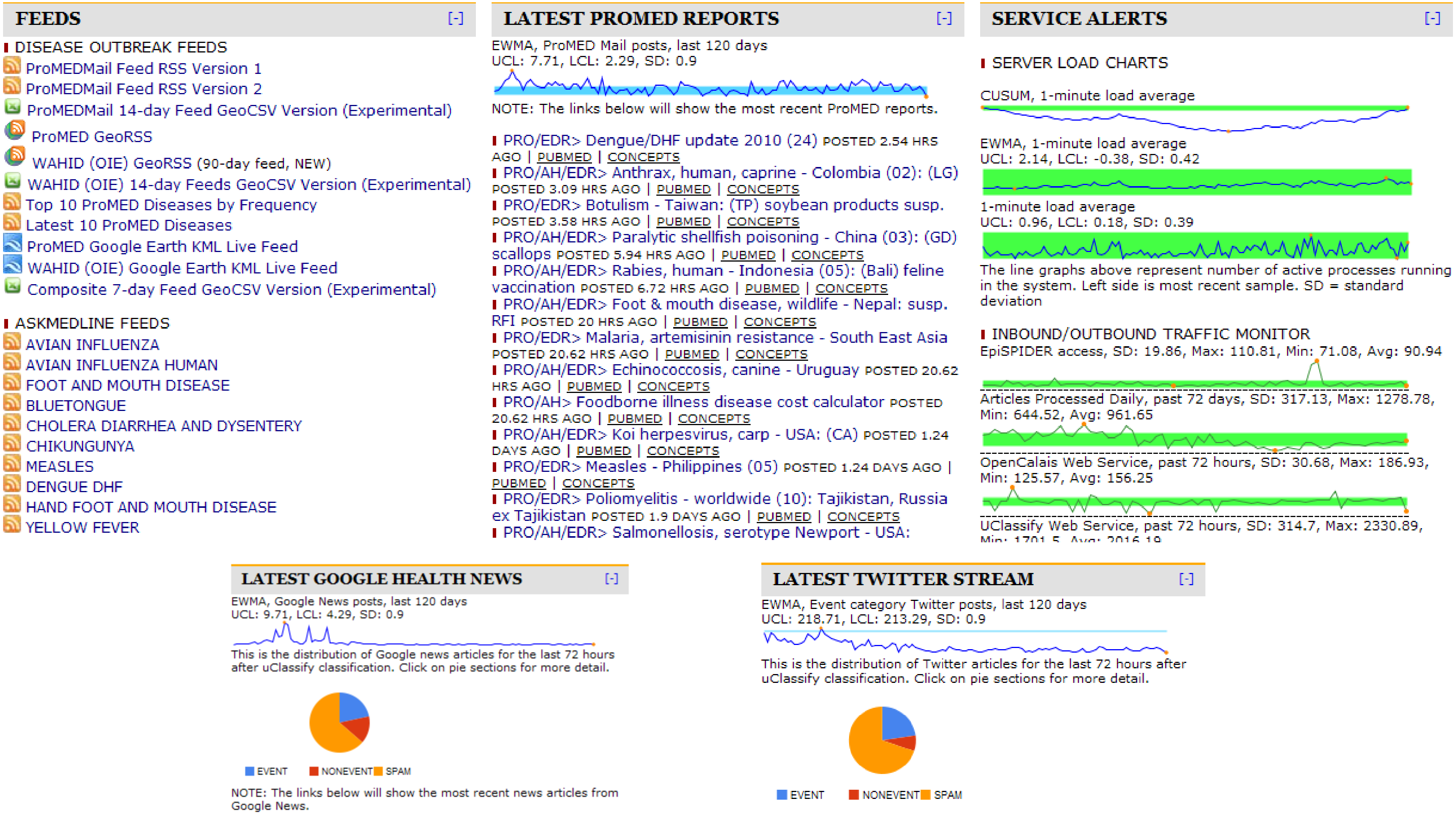
ProMED-PORT, Português

ProMED-ESP, Español

ProMED-RUS, Русский

PRO/MBDS, Mekong Basin

EpiSpider - http://www.epispider.org/



FEEDS [-]

▮ DISEASE OUTBREAK FEEDS

- ProMEDMail Feed RSS Version 1
- ProMEDMail Feed RSS Version 2
- ProMEDMail 14-day Feed GeoCSV Version (Experimental)
- ProMED GeoRSS
- WAHID (OIE) GeoRSS (90-day feed, NEW)
- WAHID (OIE) 14-day Feeds GeoCSV Version (Experimental)
- Top 10 ProMED Diseases by Frequency
- Latest 10 ProMED Diseases
- ProMED Google Earth KML Live Feed
- WAHID (OIE) Google Earth KML Live Feed
- Composite 7-day Feed GeoCSV Version (Experimental)

▮ ASKMEDLINE FEEDS

- AVIAN INFLUENZA
- AVIAN INFLUENZA HUMAN
- FOOT AND MOUTH DISEASE
- BLUETONGUE
- CHOLERA DIARRHEA AND DYSENTERY
- CHIKUNGUNYA
- MEASLES
- DENGUE DHF
- HAND FOOT AND MOUTH DISEASE
- YELLOW FEVER

LATEST PROMED REPORTS [-]

EWMA, ProMED Mail posts, last 120 days
UCL: 7.71, LCL: 2.29, SD: 0.9

NOTE: The links below will show the most recent ProMED reports.

▮ PRO/EDR> Dengue/DHF update 2010 (24) POSTED 2.54 HRS AGO | PUBMED | CONCEPTS
▮ PRO/AH/EDR> Anthrax, human, caprine - Colombia (02): (LG) POSTED 3.09 HRS AGO | PUBMED | CONCEPTS
▮ PRO/EDR> Botulism - Taiwan: (TP) soybean products susp. POSTED 3.58 HRS AGO | PUBMED | CONCEPTS
▮ PRO/AH/EDR> Paralytic shellfish poisoning - China (03): (GD) scallops POSTED 5.94 HRS AGO | PUBMED | CONCEPTS
▮ PRO/AH/EDR> Rabies, human - Indonesia (05): (Bali) feline vaccination POSTED 6.72 HRS AGO | PUBMED | CONCEPTS
▮ PRO/AH/EDR> Foot & mouth disease, wildlife - Nepal: susp. RFI POSTED 20 HRS AGO | PUBMED | CONCEPTS
▮ PRO/EDR> Malaria, artemisinin resistance - South East Asia POSTED 20.62 HRS AGO | PUBMED | CONCEPTS
▮ PRO/AH/EDR> Echinococcosis, canine - Uruguay POSTED 20.62 HRS AGO | PUBMED | CONCEPTS
▮ PRO/AH> Foodborne illness disease cost calculator POSTED 20.62 HRS AGO | PUBMED | CONCEPTS
▮ PRO/AH/EDR> Koi herpesvirus, carp - USA: (CA) POSTED 1.24 DAYS AGO | PUBMED | CONCEPTS
▮ PRO/EDR> Measles - Philippines (05) POSTED 1.24 DAYS AGO | PUBMED | CONCEPTS
▮ PRO/EDR> Poliomyelitis - worldwide (10): Tajikistan, Russia ex Tajikistan POSTED 1.9 DAYS AGO | PUBMED | CONCEPTS
▮ PRO/AH/EDR> Salmonellosis, serotype Newport - USA:

SERVICE ALERTS [-]

▮ SERVER LOAD CHARTS

CUSUM, 1-minute load average

EWMA, 1-minute load average
UCL: 2.14, LCL: -0.38, SD: 0.42

1-minute load average
UCL: 0.96, LCL: 0.18, SD: 0.39

The line graphs above represent number of active processes running in the system. Left side is most recent sample. SD = standard deviation

▮ INBOUND/OUTBOUND TRAFFIC MONITOR

EpiSPIDER access, SD: 19.86, Max: 110.81, Min: 71.08, Avg: 90.94

Articles Processed Daily, past 72 days, SD: 317.13, Max: 1278.78, Min: 644.52, Avg: 961.65

OpenCalais Web Service, past 72 hours, SD: 30.68, Max: 186.93, Min: 125.57, Avg: 156.25

UClassify Web Service, past 72 hours, SD: 314.7, Max: 2330.89, Min: 1701.5, Avg: 2016.19

LATEST GOOGLE HEALTH NEWS [-]

EWMA, Google News posts, last 120 days
UCL: 9.71, LCL: 4.29, SD: 0.9

This is the distribution of Google news articles for the last 72 hours after uClassify classification. Click on pie sections for more detail.

NOTE: The links below will show the most recent news articles from Google News.

LATEST TWITTER STREAM [-]

EWMA, Event category Twitter posts, last 120 days
UCL: 218.71, LCL: 213.29, SD: 0.9

This is the distribution of Twitter articles for the last 72 hours after uClassify classification. Click on pie sections for more detail.

	BioCaster	HealthMap	MedISys + PULS	Our System
Year	2007	2007	2007	2010
Country	Japan	USA	European Union	USA
Mined Sources	1500 News Feeds	Only Google News & ProMED-Mail	1400 news portals + 150 Public Health sites	Personalized predefined set of seeds by domain experts
Productivity	25-30 locations on 40 diseases per/day	20-30 outbreaks per day	50,000 news articles per day	Varies by the size of the crawled collection
Languages	English, French, Spanish, Chinese, Thai, Vietnamese, Japanese	English, French, Spanish, Portuguese, Russian, Chinese, Arabic	43 languages	Future work: wikification for the multilingual IE/IR
Geographical Entities	243 countries & 4,025 sub-countries (province & cities)	2,300 locations	70,000 locations (towns, cities, provinces)	> million locations from NGA GEOnet Names Database
Disease and Other Entities	50 diseases (ontology with synonyms, symptoms)	1100 diseases	2400 animal + human disease names, 400 organisms, 1500 political entities	Automatically constructed ontology with >1000 animal diseases, viruses, serotypes

Animal Disease-related Data Online

Structured Data

- Official reports by different organizations:
  - state and federal laboratories, bioportals;
  - health care providers;
  - governmental agricultural or environmental agencies.

Unstructured Data

- Web-pages

- News
- E-mails (e.g., ProMed-Mail)
- Blogs

- Medical literature (e.g., books)

- Scientific papers (e.g., PubMed)

Research Challenges (1)

- Large amount of information from multiple sources
  - Extract facts/structured information from unstructured text
  - Manage the specificity of the content (e.g. blogosphere, biomedical literature, news, official reports etc.)
  - Necessity of data aggregation from multiple sources

Speculate about event confidence

- Solution: Reliability of the source by majority voting

- Multiple locations, different case status, different victim types lead to multiple data base entries

- Solution: spurious event detection and event disambiguation *e.g.,* source 1: 10 victims vs. source 2: 15 victims

Research Challenges (2)

- Resolve location disambiguation “Rabies in Isle of Wight”

  - What geo-tag in Virginia, USA or UK?
  - Solution: track geo-tag of the original source of information

- Deal with unknown or undiagnosed diseases

  - “The deadly outbreak has so far killed 16 people in Gabon”
  - Solution: Look into the context/recent outbreaks in this location

- Manage specific dates/times occurrences

  - “FMD outbreak was reported last week/today…”
  - Solution: Use set of regular expressions and date/time ontology

Existing Systems vs. Designed System (1)

Existing Systems	Designed System
System Purpose	
disease surveillance	research and epidemiological analytics
Targeted Audience	
public	domain experts and analysts
Processed Data	
news, ProMed - Mail	medical literature, research papers, general web-articles, blog posts, e-mails *etc.*

Existing Systems vs. Designed System (2)

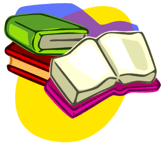
Existing Systems	Designed System
Ontology-based IE (limited by # of diseases, locations etc.)	Automatically expanded ontology for IE
No functionality for past outbreak tracking	Identify events in the historical data, *e.g.* medical literature
No timeline visualization (BioCaster)	More event attributes: disease, date, location, species, confirmation status
Manual Moderation (HealthMap)	Classify events into two categories: suspected or confirmed

Targeting Audience

Research and Public Health communities

1. Managing the specificity of blogosphere

Health Care Providers (e.g. hospitals)



Governmental Agencies (e.g. Center for Disease Control and Prevention)

2. Dealing with biomedical literature

3. News content & official reports processing

Laboratories

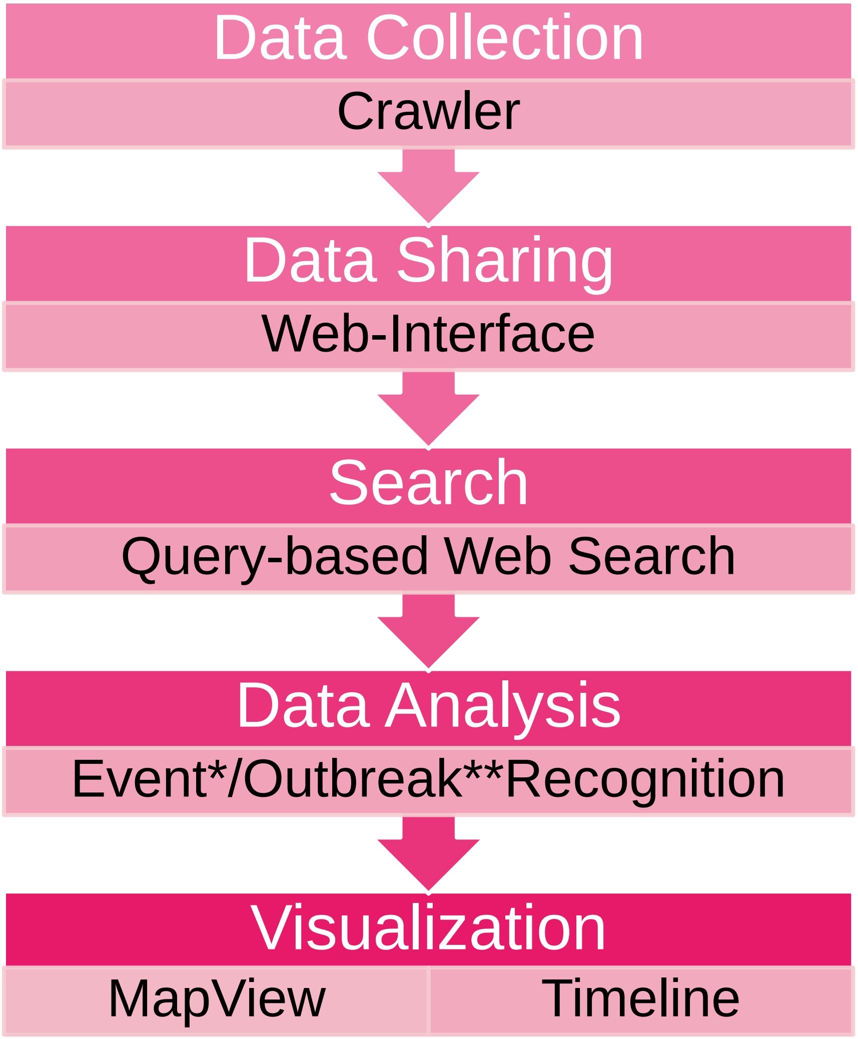
4. Capturing all possible breakdowns in communication channels between levels of animal disease management

State → **National** → **International**

Agenda

- Overview
- Animal Disease Monitoring Systems
  - Manually Supported Web-Interfaces
  - Automated Web-Services
- Framework for Epidemiological Analytics
  - System Functionality
  - Web Crawling
  - Domain-specific Entity Extraction
  - Animal Disease-related Event Recognition
- Summary

Framework for Epidemiological Analytics



Data Collection

Crawler

Data Sharing

Web-Interface

Search

Query-based Web Search

Data Analysis

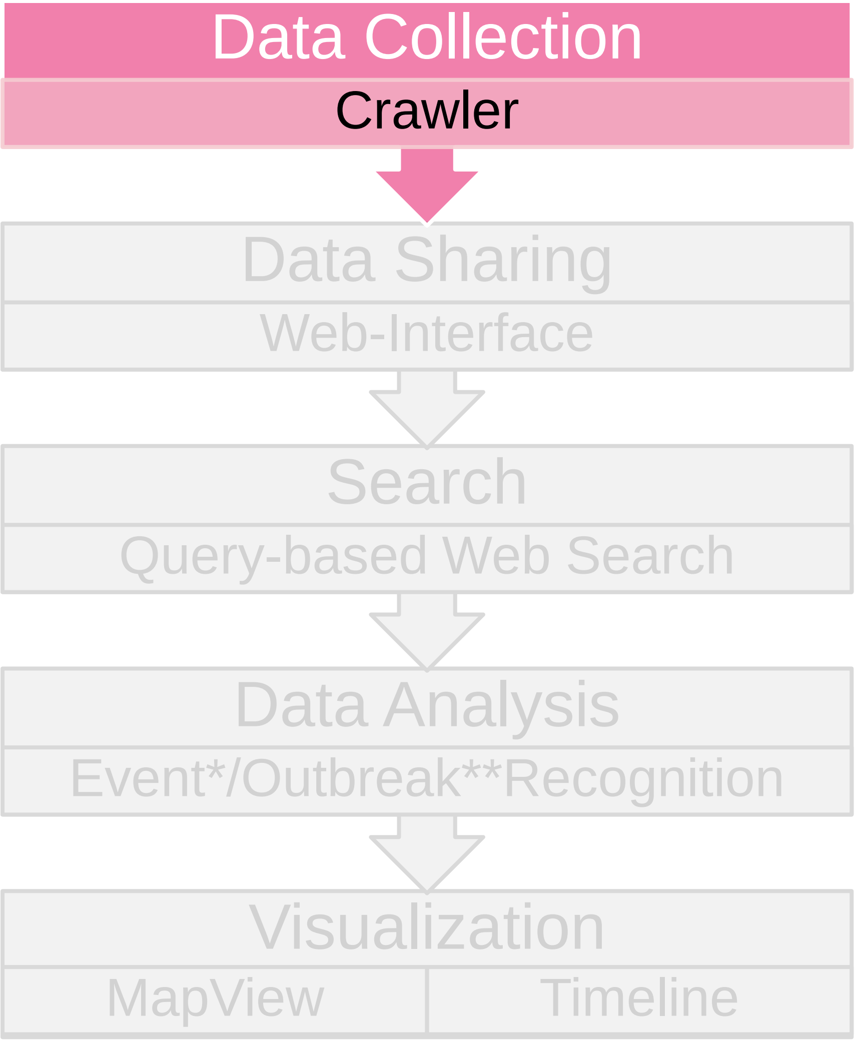
Event*/Outbreak**Recognition

Visualization

MapView

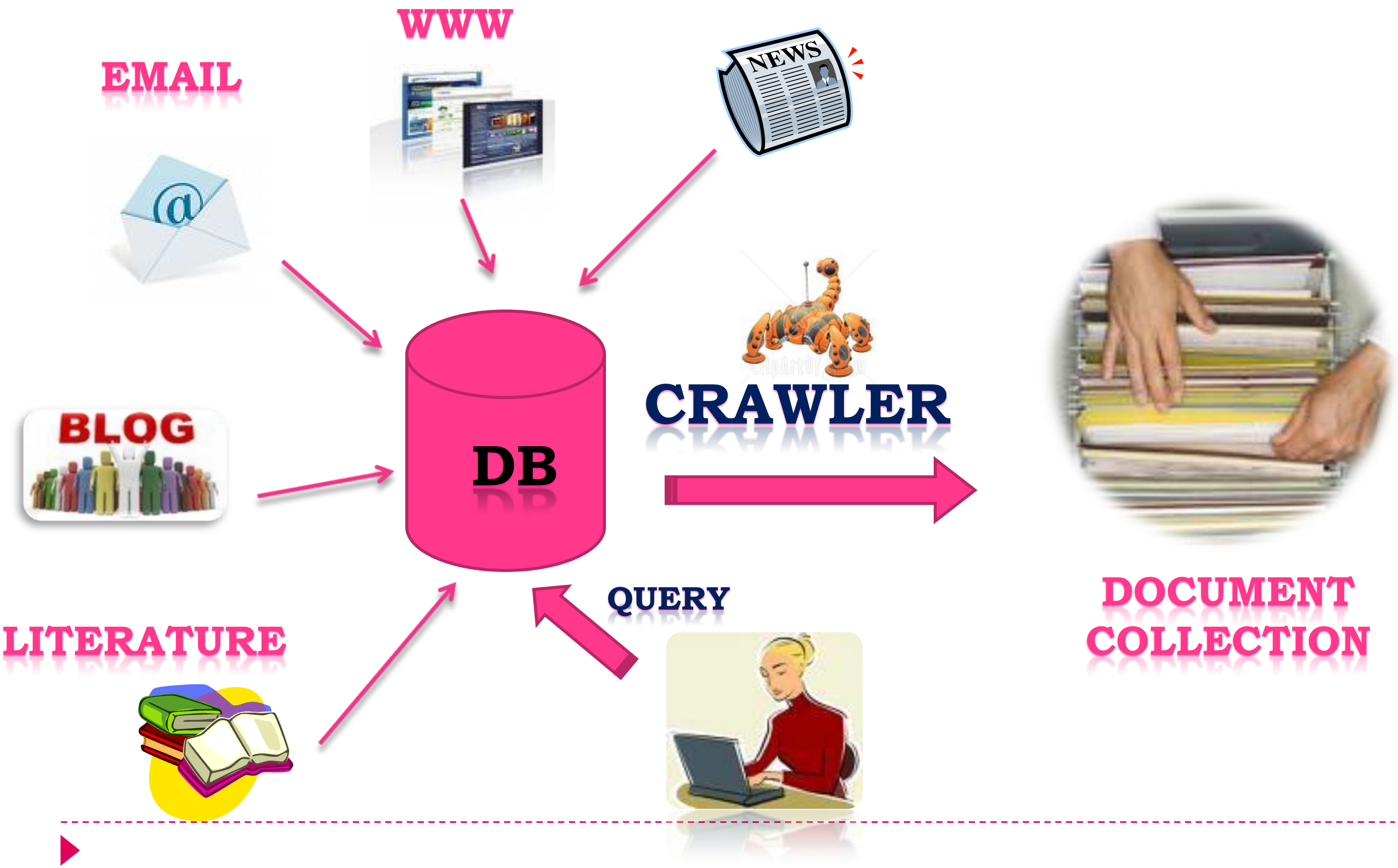
Timeline

1. Data Collection (1)



- Periodically crawl the web using Heritrix crawler - http://crawler.archive.org/
  - set of seeds (ProMED-Mail, DEFRA *etc.*)
  - set of terms (animal disease names from the ontology)
- Text-to-tag ratio-based method for **content extraction** from web pages

1. Data Collection (2)



WWW

EMAIL

BLOG

DB

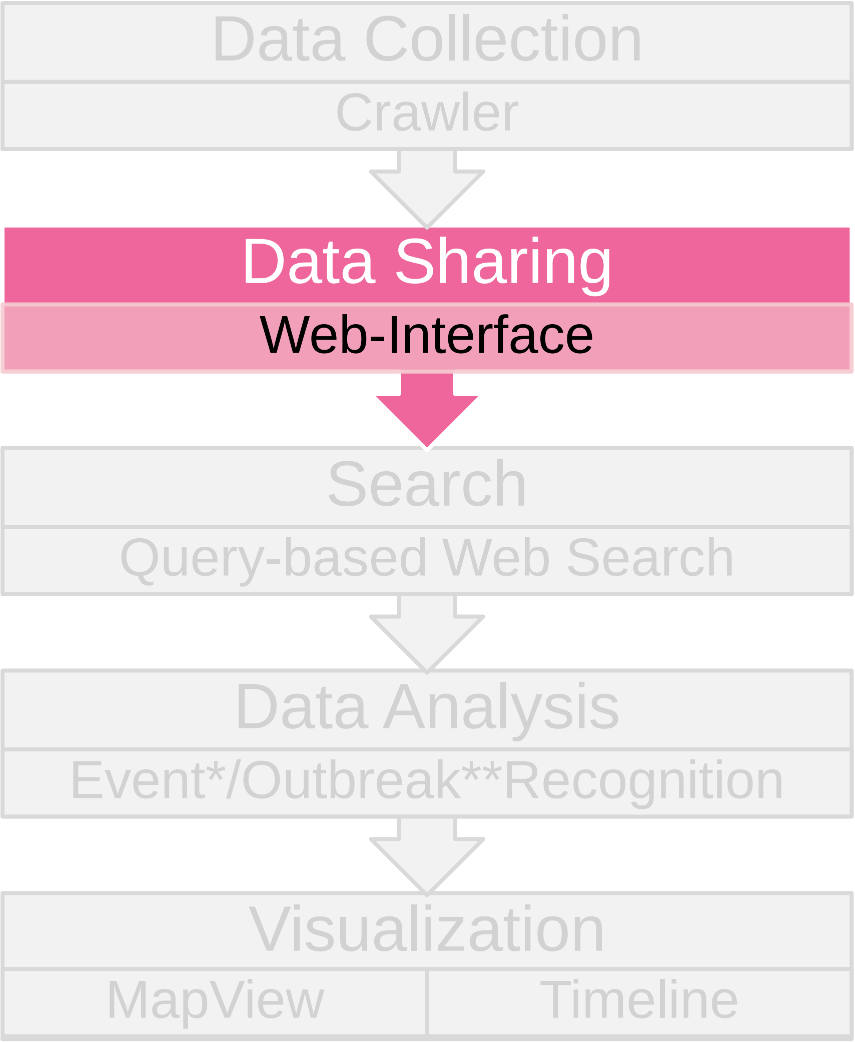
CRAWLER

QUERY

LITERATURE

DOCUMENT COLLECTION

2. Data Sharing



Data Collection

Crawler

Data Sharing

Web-Interface

Search

Query-based Web Search

Data Analysis

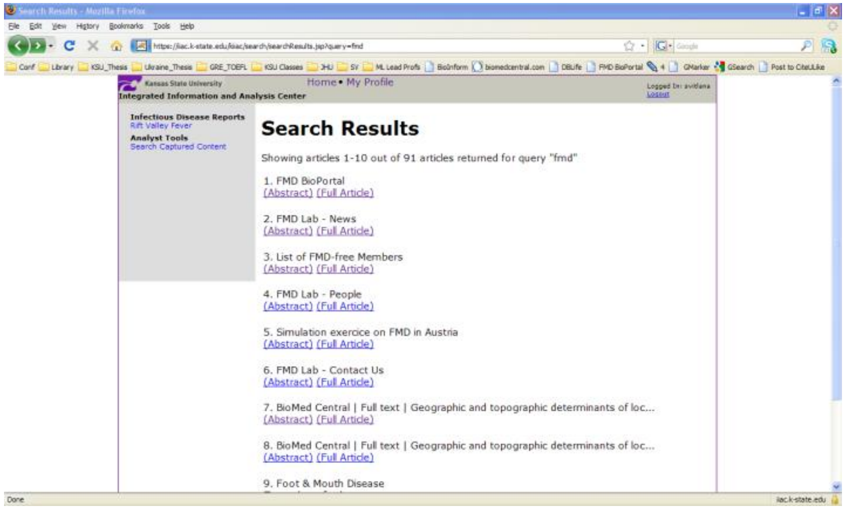
Event*/Outbreak**Recognition

Visualization

MapView

Timeline

- Document relevance classification using Naive Bayes Classifier from Mallet - http://mallet.cs.umass.edu
  - Relevant
  - Non-relevant



Search Results - Mozilla Firefox
File Edit View History Bookmarks Tools Help
https://iiac.k-state.edu/iiac/search/searchResults.jsp?query=fmd
Google
Conf Library KSU_Thesis Ukraine_Thesis GRE_TOEFL KSU Classes JHU SV ML Lead Profs BioInform biomedcentral.com DBLife FMD BioPortal 4 GMarker GSearch Post to CiteULike
Kansas State University
Integrated Information and Analysis Center
Home • My Profile
Logged In: avitlana
Logout
Infectious Disease Reports
Rift Valley Fever
Analyst Tools
Search Captured Content
Done
iiac.k-state.edu

Search Results

Showing articles 1-10 out of 91 articles returned for query "fmd"

1. FMD BioPortal
(Abstract) (Full Article)

2. FMD Lab - News
(Abstract) (Full Article)

3. List of FMD-free Members
(Abstract) (Full Article)

4. FMD Lab - People
(Abstract) (Full Article)

5. Simulation exercice on FMD in Austria
(Abstract) (Full Article)

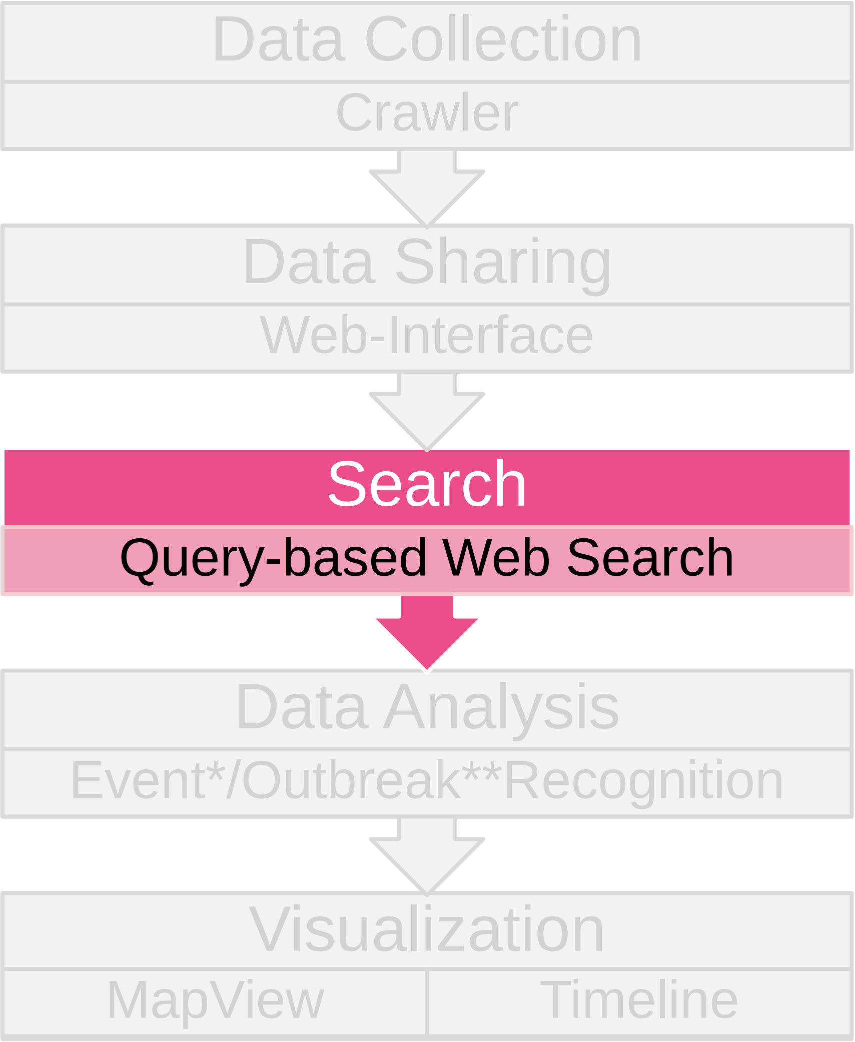
6. FMD Lab - Contact Us
(Abstract) (Full Article)

7. BioMed Central | Full text | Geographic and topographic determinants of loc...
(Abstract) (Full Article)

8. BioMed Central | Full text | Geographic and topographic determinants of loc...
(Abstract) (Full Article)

9. Foot & Mouth Disease

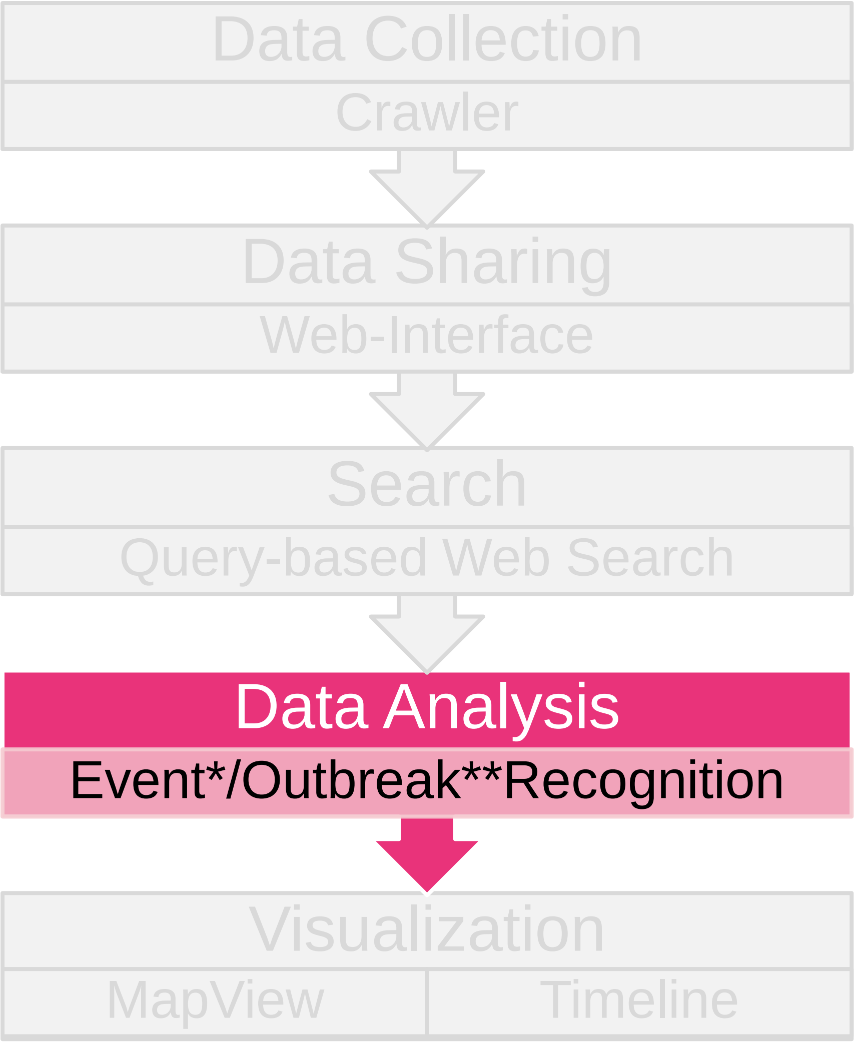
3. Search



- Lucene-based* ranking
- Query-based keyword search
- Search by animal **disease name** and/or **location**

*Lucene - http://lucene.apache.org

4. Data Analysis



Event example:

"On **12 September 2007**, a new **foot-and-mouth disease** outbreak was confirmed in **Egham, Surrey**"

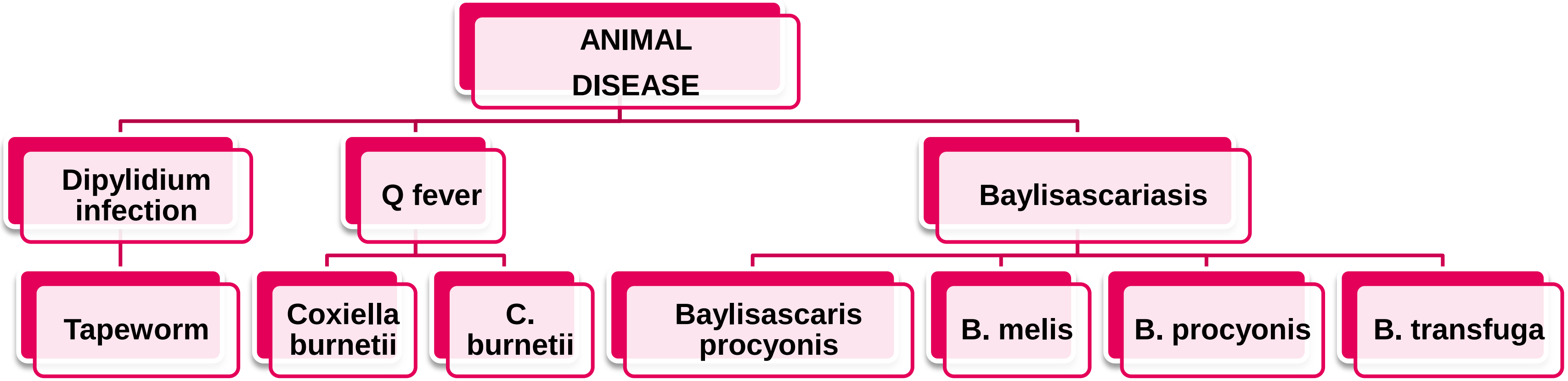
Domain Meta-data

Domain-specific knowledge

- Medical ontology
  - diseases, serotypes, and viruses.

Domain-independent knowledge

- Location hierarchy
  - names of countries, states, cities;
- Time hierarchy
  - canonical dates.



Event Recognition Methodology

- **Step 1.** Entity recognition from raw text.
- **Step 2.** Sentence classification from which entities are extracted as being related to an event or not; if they are related to an event we classify them as confirmed or suspected.
- **Step 3.** Combination of entities within an event sentence into the structured tuples and aggregation of tuples related to the same event into one comprehensive tuple.

Step 1.Entity Recognition

- Locate and classify atomic elements into predefined categories:
  - **Disease names:** “foot and mouth disease”, “rift valley fever”; **viruses:** “picornavirus”; **serotypes:** “Asia-1”;
  - **Species:** “sheep”, “pigs”, “cattle” and “livestock”;
  - **Locations** of events specified at different levels of geo-granularity: “United Kingdom", “eastern provinces of Shandong and Jiangsu, China”;
  - **Dates** in different formats: “last Tuesday”, “two month ago”.

Entity Recognition Tools

- Animal Disease Extractor*
  - relies on a medical ontology, automatically-enriched with synonyms and causative viruses.

- Species Extractor*
  - pattern matching on a stemmed dictionary of animal names from Wikipedia.
- Location Extractor
  - Stanford NER Tool** (uses conditional random fields);
  - NGA GEOnet Names Database (GNS)*** for location disambiguation and retrieving latitude/longitude.
- Date/Time Extractor
  - set of regular expressions.

*KDD KSU DSEx - http://fingolfin.user.cis.ksu.edu:8080/diseaseextractor/
**Stanford NER - http://nlp.stanford.edu/ner/index.shtml
***GNS - http://earth-info.nga.mil/gns/html/

Step 2. Event Sentence Classification

- Constraint: True events should include a disease name together with a status verb from Google Sets* and WordNet** (eliminate event non-related sentences).
  - “Foot and mouth disease **is**$_{[V]}$ a highly pathogenic animal disease”.

- Confirmed status verbs *“happened”* and verb phrases *“strike out”*
  - “On 9 Jun 2009, the farm's owner **reported**$_{[V]}$ symptoms of FMD in more than 30 hogs”.

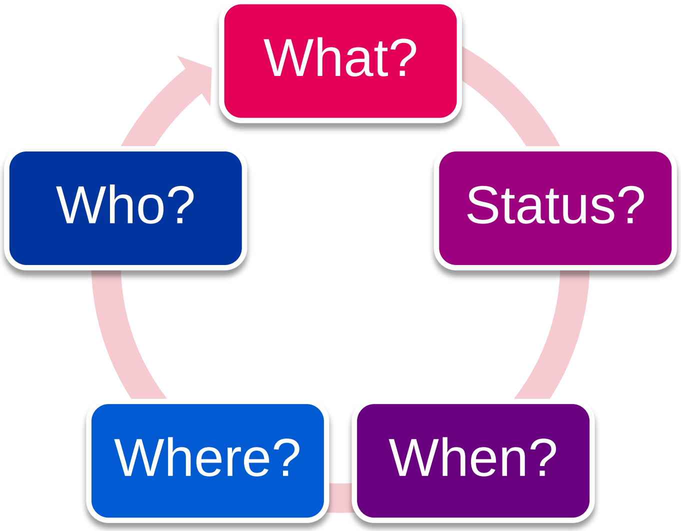
- Suspected status verbs *“catch”* and verb phrases *“be taken in”*
  - “RVF is **suspected**$_{[V]}$ in Saudi Arabia in September 2000”.

*GoogleSets - http://labs.google.com/sets

**WordNet - http://wordnet.princeton.edu/

Step 3. Event Tuple Generation

- Event attributes:
  - disease
  - date
  - location
  - species
  - confirmation status



- Event tuple:
  - Event_i = < disease; date; location; species; status > = <FMD, 9 Jun 2009, Taoyuan, hog, confirmed>

- Event tuple with missing attributes:
  - Event_j = <**FMD**, ?, **?**, ?, **confirmed**>

Event Recognition Workflow

Step 1: Entity Recognition

Foot-and-mouth disease$_{[DIS]}$ on **hog$_{[SP]}$** farm in **Taoyuan$_{[LOC]}$**.

Taiwan's TVBS television station reports that agricultural authorities confirmed **foot-and-mouth disease$_{[DIS]}$** on a **hog$_{[SP]}$** farm in **Taoyuan$_{[LOC]}$**. On **9 Jun 2009$_{[DT]}$**, the farm's owner reported symptoms of **FMD$_{[DIS]}$** in more than 30 **hogs$_{[SP]}$**. Subsequent testing confirmed **FMD$_{[DIS]}$**. Agricultural authorities asked the farmer to strengthen immunization. The outbreak has not affected other farms. Authorities stipulated that the affected **hog$_{[SP]}$** farm may not sell pork for 2 weeks.

Step 2: Sentence Classification

YES 1. **Foot-and-mouth disease$_{[DIS]}$** on **hog$_{[SP]}$** farm in **Taoyuan$_{[LOC]}$**.

YES 2.Taiwan's TVBS television station **reports** that agricultural authorities **confirmed** **foot-and-mouth disease$_{[DIS]}$** on a **hog$_{[SP]}$** farm in **Taoyuan$_{[LOC]}$**.

YES 3. On **9 Jun 2009$_{[DT]}$**, the farm's owner **reported** symptoms of **FMD$_{[DIS]}$** in more than 30 **hogs$_{[SP]}$**.

YES 4. Subsequent testing **confirmed** **FMD$_{[DIS]}$**.

NO 5. Agricultural authorities asked the farmer to strengthen immunization.

NO 6. The outbreak has not affected other farms.

NO 7. Authorities stipulated that the affected **hog$_{[SP]}$** farm may not sell pork for 2 weeks.

Step 3a: Tuple Generation

E_1 = <**Foot-and-mouth disease**, ?, **Taoyuan**, **hog**, ?>

E_2 = <**Foot-and-mouth disease**, ?, **Taoyuan**, **hog**, **confirmed** >

E_3 = <**FMD**, **9 Jun 2009, ?**, **hog**, **reported**>

E_4 = <**FMD**, ?, **?**, ?, **confirmed**>

Step 3b: Tuple Aggregation

E = <disease, date, location, species, status> = <**Foot-and-mouth disease**, **9 Jun 2009**, **Taoyuan**, **hog**, **confirmed** >

Animal Disease Extraction Results (1)

- **Synonymic relationships –** “E1 is a kind of E2”

  E1 = “swine influenza” is a kind of E2 = “swine fever”

- **Hyponymic relationships –** “E1 and E1 are diseases”

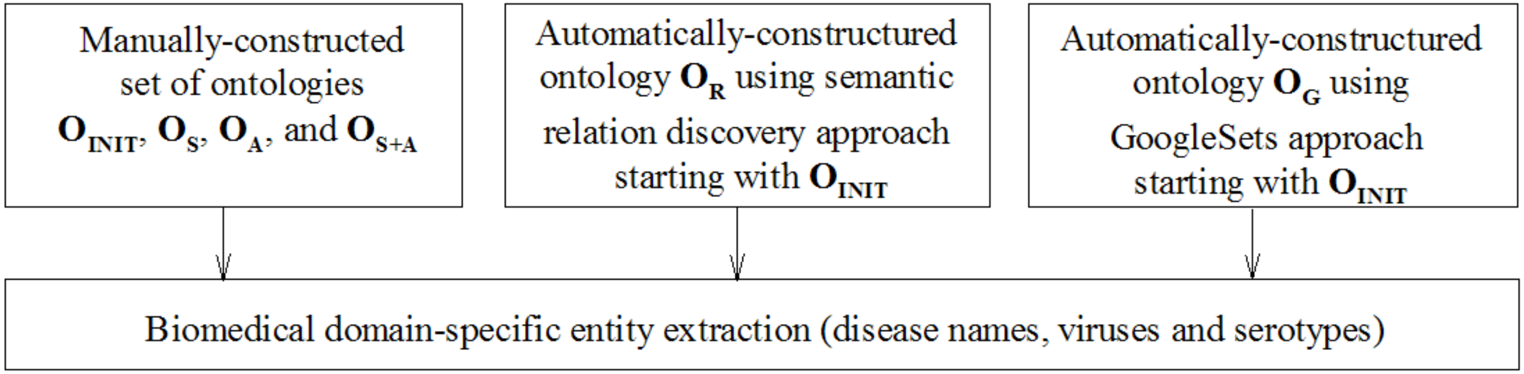
  E1 = “anthrax”, E2 = “yellow fever” are diseases

- **Causal relationships –** “E1 is caused by E2”

  E1 = “Ovine epididymitis” is caused by E2 = “Brucella ovis”

Synonymic	Hyponymic	Causal
• “is a”, “and”	• “such as”	• “is caused by”
• “also known as”	• "for example"	• “causes”
• “is also called ”	• “including”	

Animal Disease Extraction Results (2)

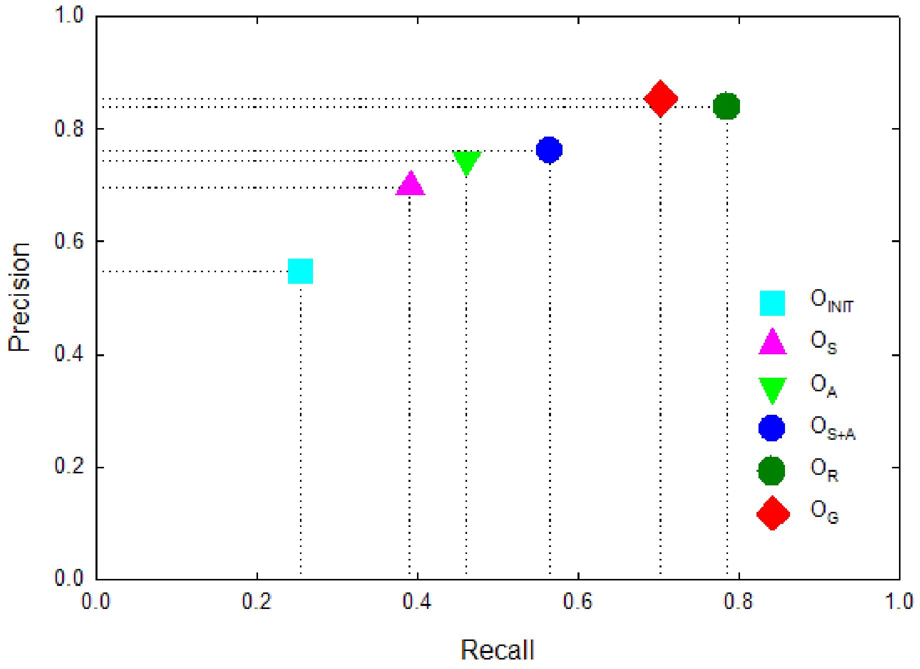


Manually-constructed set of ontologies $\mathbf{O}_{INIT}$, $\mathbf{O}_S$, $\mathbf{O}_A$, and $\mathbf{O}_{S+A}$

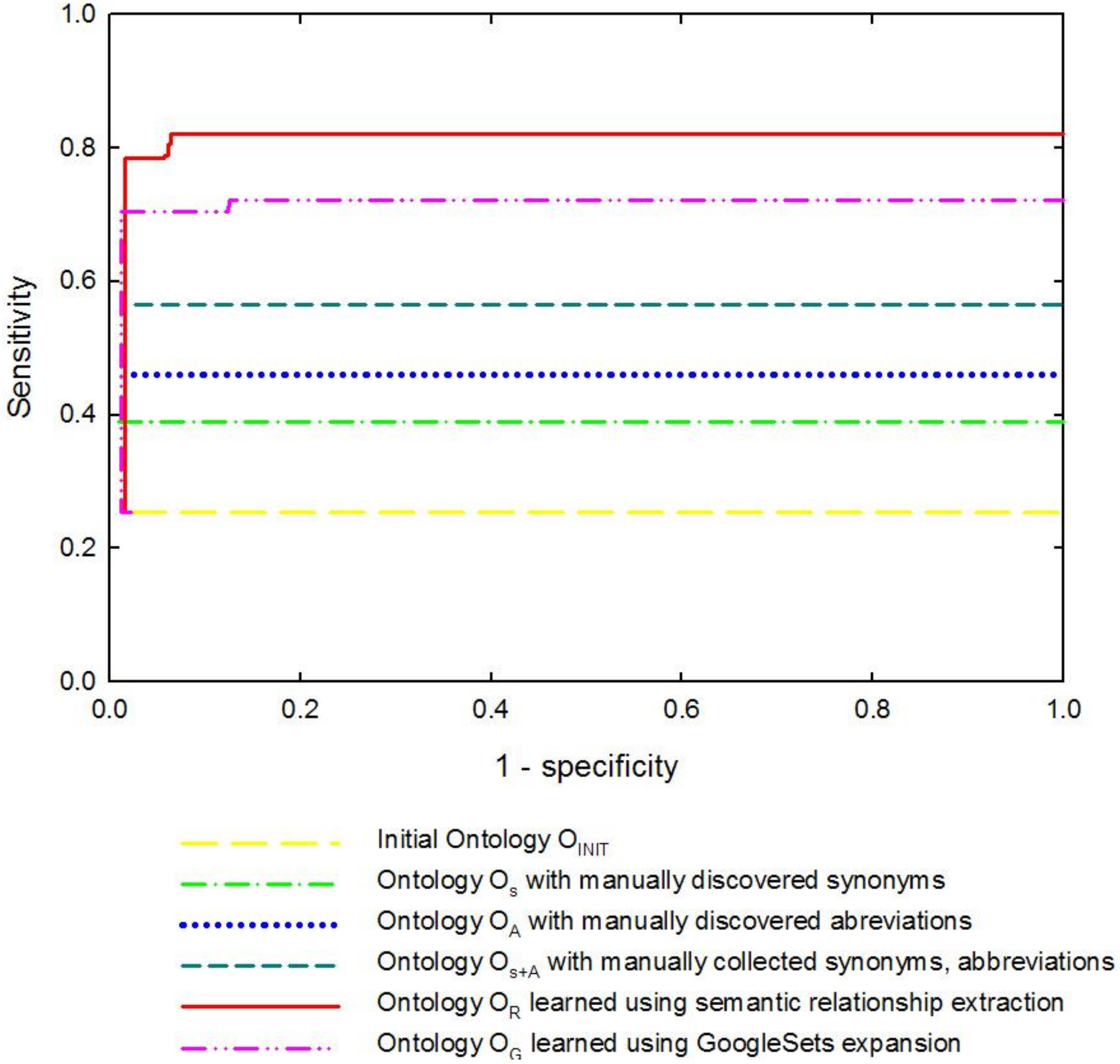
Automatically-constructed ontology $\mathbf{O}_R$ using semantic relation discovery approach starting with $\mathbf{O}_{INIT}$

Automatically-constructed ontology $\mathbf{O}_G$ using GoogleSets approach starting with $\mathbf{O}_{INIT}$

Biomedical domain-specific entity extraction (disease names, viruses and serotypes)



Animal Disease Extraction Results (3)

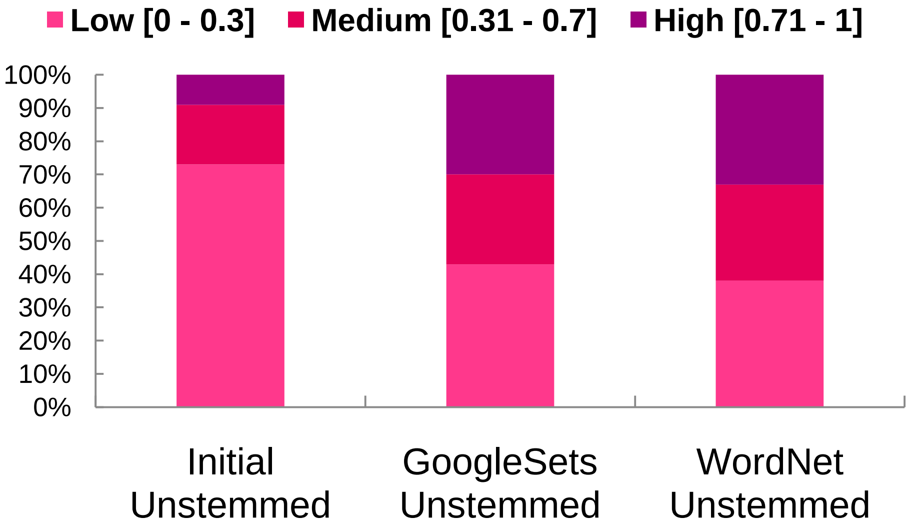


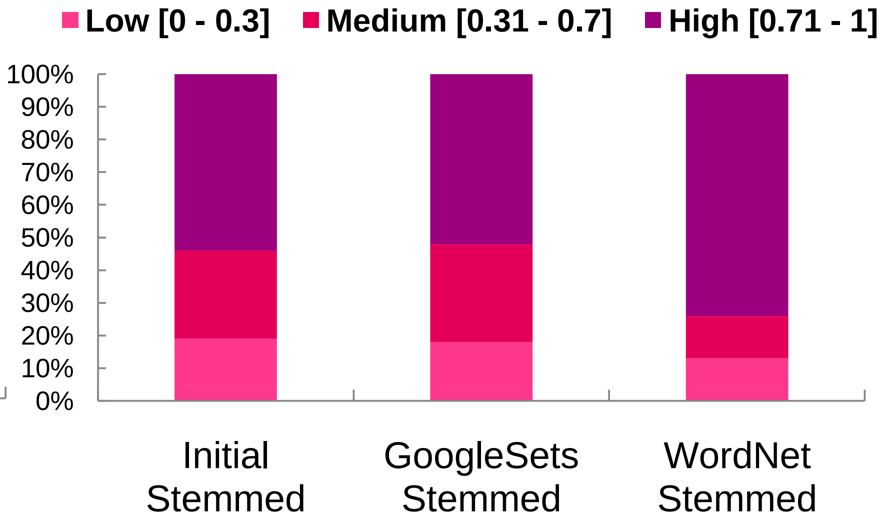
Event Recognition Results

$$Score_i = < w_d disease;\ w_t date;\ w_l location;\ w_s species;\ w_c status\ldots >,$$

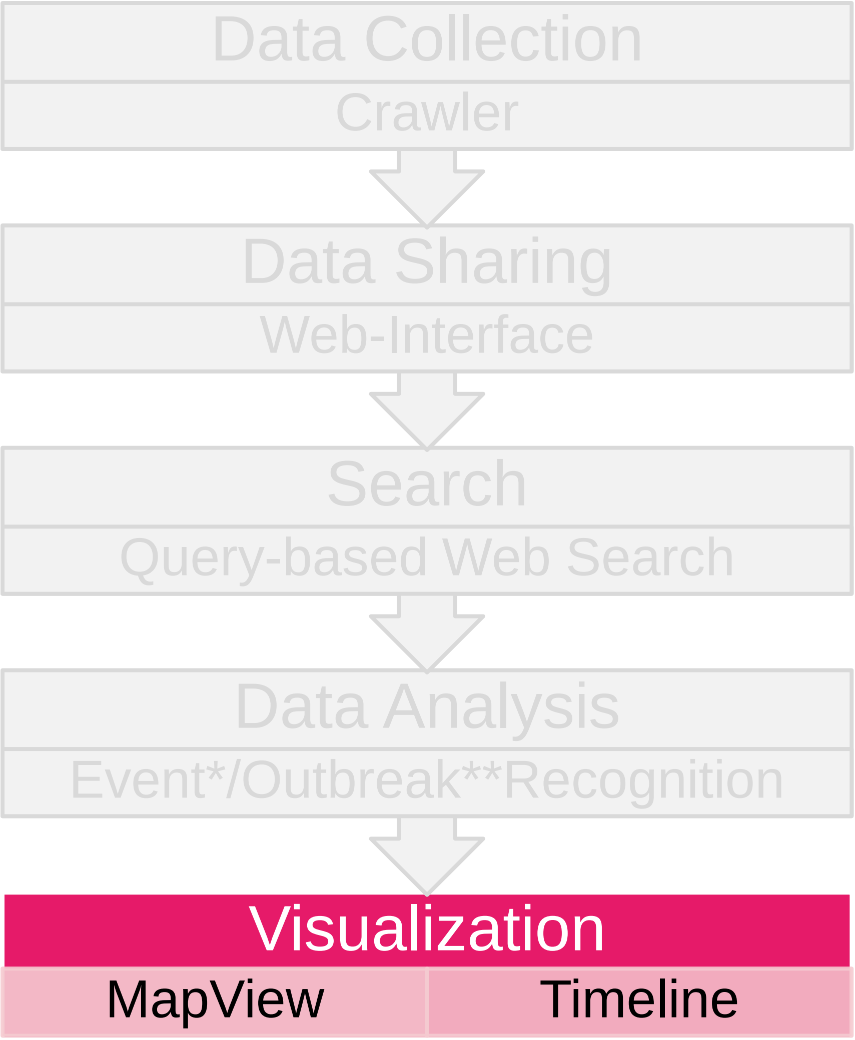
$$subject\ to\ disease + status = 2$$

- Interpret the Pyramid score values - http://www1.cs.columbia.edu/~becky/DUC2006/2006-pyramid-guidelines.html_ducview as an event extraction accuracy
- Apply list of verbs from GoogleSets and WordNet
- We use NS (unstemmed) and S (stemmed) versions of the verb lists

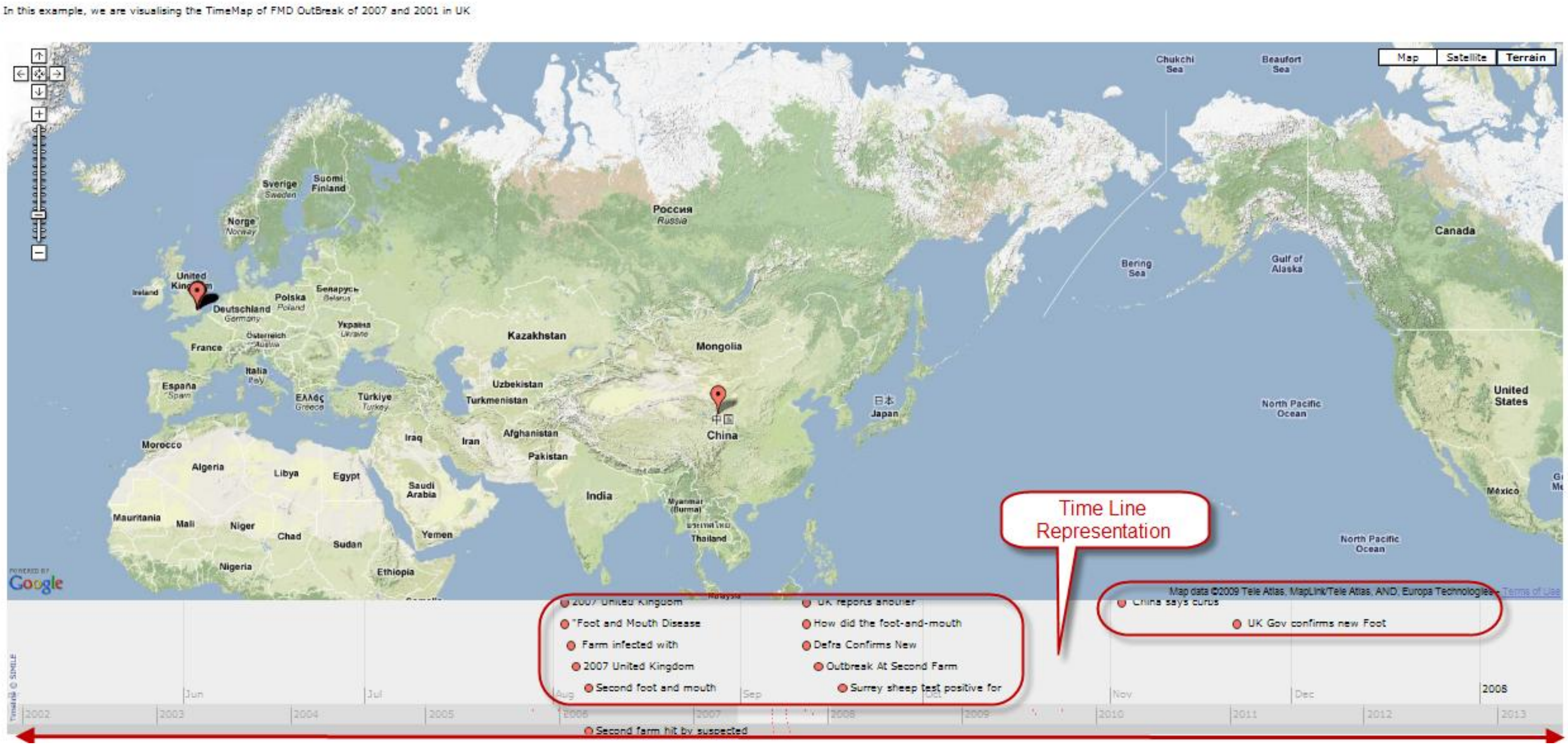




5. Visualization



- Map View
  - GoogleMaps API - http://code.google.com/apis/maps/
- TimeLine View
  - SIMILE API - http://www.simile-widgets.org/timeline/

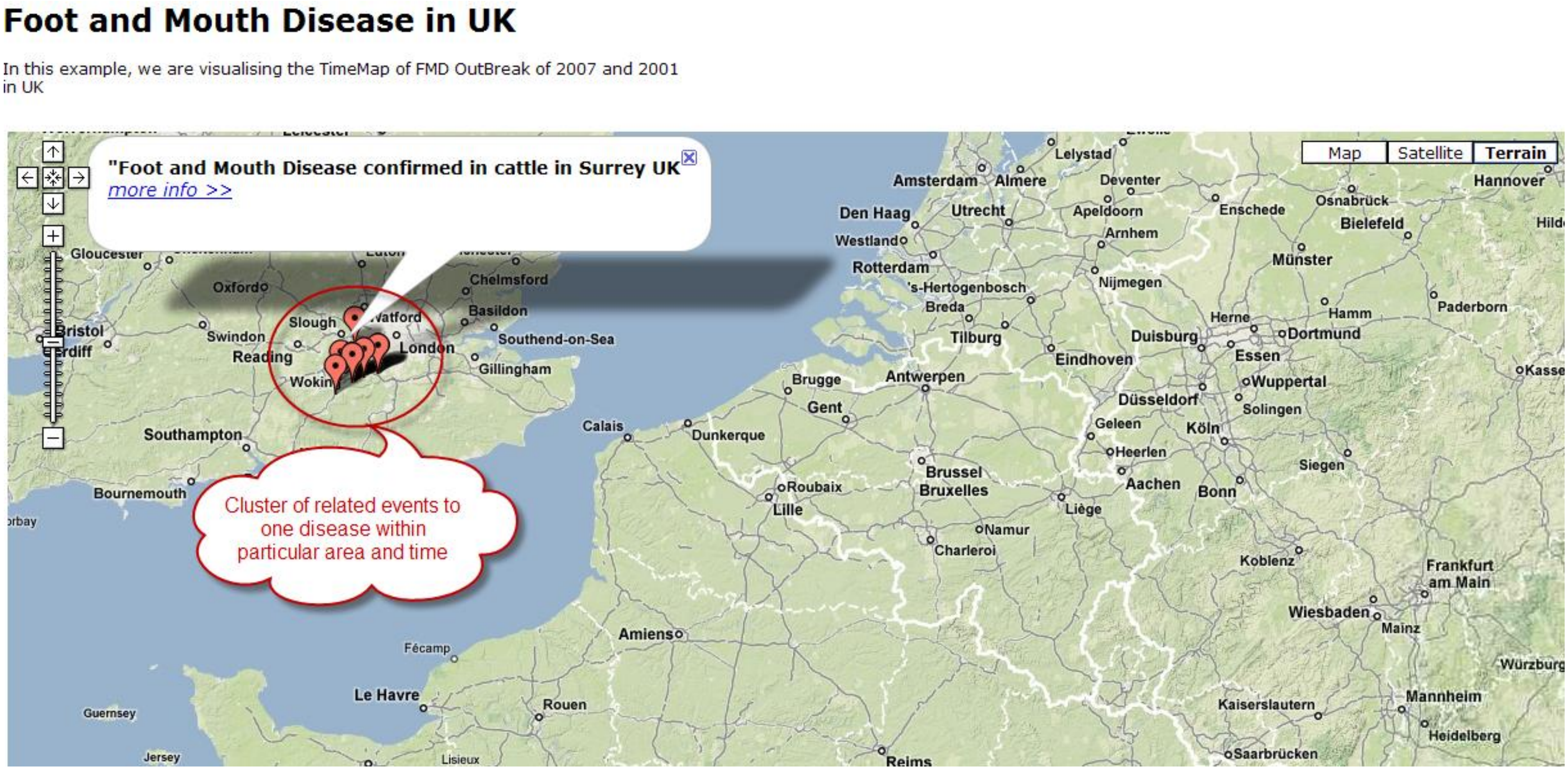


In this example, we are visualising the TimeMap of FMD OutBreak of 2007 and 2001 in UK

Time Line Representation

Event Representation by Date/Time Timeline View

http://fingolfin.user.cis.ksu.edu/timemap.1.4/FMD_2007_UK_Viz/FMD_Viz.htm



Foot and Mouth Disease in UK

In this example, we are visualising the TimeMap of FMD OutBreak of 2007 and 2001 in UK

"Foot and Mouth Disease confirmed in cattle in Surrey UK

more info >>

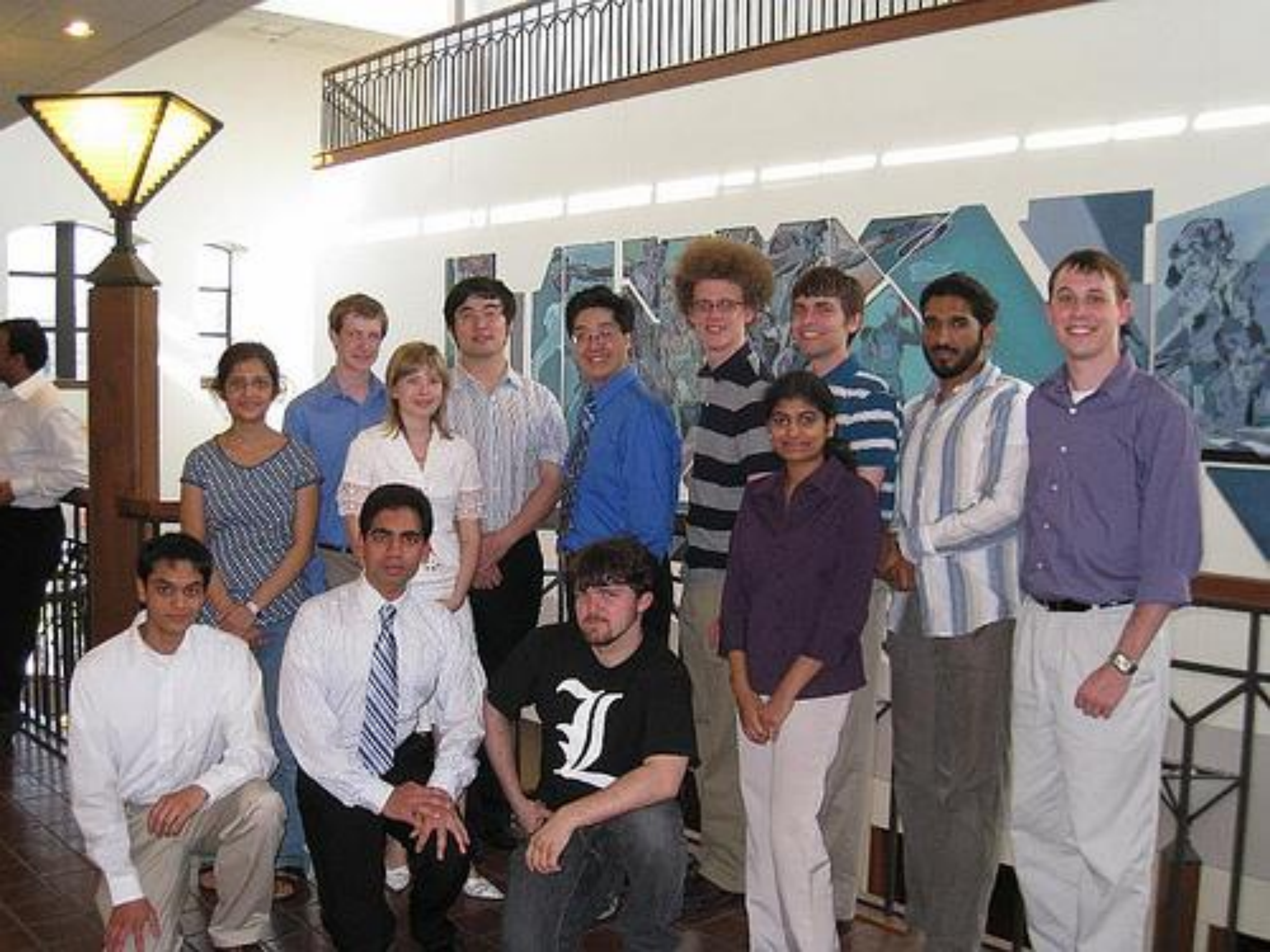
Cluster of related events to one disease within particular area and time

Event Representation by Location Map View

http://fingolfin.user.cis.ksu.edu/timemap.1.4/FMD_2007_UK_Viz/FMD_Viz.htm

Summary

- perform focused crawling of different sources (books, research papers, blogs, governmental sources, etc.)
- use semantic relationship learning approach (including synonymic, hyponymic, causal relationships) for automated-ontology expansion for domain-specific entity extraction (e.g., diseases, viruses)
- recognize geo-entities using CRF approach and disambiguates them using GNServer
- extract animal disease-related events with more descriptive event attributes such as: species, dates, event confirmation status, in contrast to ”disease-location” pairs
- support timeline representation of extracted events in SIMILE in addition to visualized events on GoogleMaps


L

Acknowledgments

- KDD Lab alumni:
  - Tim Weninger (crawler deployment) and Jing Xia (rule-based event extraction)

- KDD Lab assistants:
  - **Information Extraction Team** (John Drouhard, Landon Fowles, Swathi Bujuru)
  - **Spatial Data Mining Team** (Wesam Elshamy, Andrew Berggren)
  - **Topic Detection & Tracking Team** (Surya Kallumadi, Danny Jones, Srinivas Reddy)

- Faculty at the University of Illinois at Urbana-Champaign (2009 Data Sciences Summer Institute)
  - ChengXiang Zhai, Dan Roth, Jiawei Han and Kevin Chang.

References

- S. Volkova, W. Hsu, and D. Caragea, “Named entity recognition and tagging in the domain of epizootics”, In Proc. of Women in Machine Learning Workshop (WiML’09).
- S. Volkova, D. Caragea, W. H. Hsu, and S. Bujuru, “Animal disease event recognition and classification,” In Proc. of The First International Workshop on Web Science and Information Exchange in the Medical Web, 19th World Wide Web Conference WWW-2010.
- S. Volkova, D. Caragea, W. H. Hsu, J. Drouhard, and L. Fowles, “Boosting Biomedical Entity Extraction by using Syntactic Patterns for Semantic Relation Discovery”, ACM Web Intelligence Conference, 2010 (to appear).

Thank you!

Svitlana Volkova, svitlana.volkova@gmail.com

http://people.cis.ksu.edu/~svitlana

William H. Hsu, bhsu@cis.ksu.edu

http://people.cis.ksu.edu/~bhsu