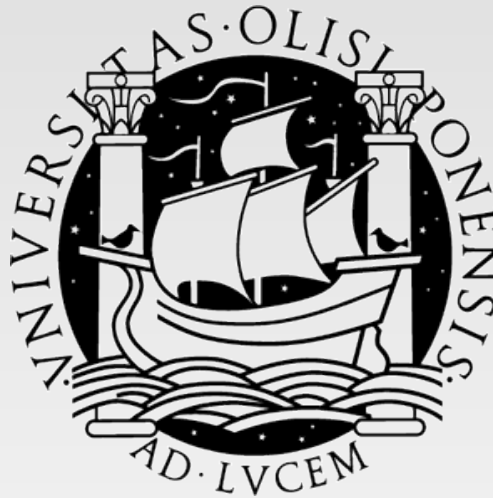


# MEDCollector

## Multisource Epidemic Data Collector

Universidade de Lisboa

Faculdade de Ciências



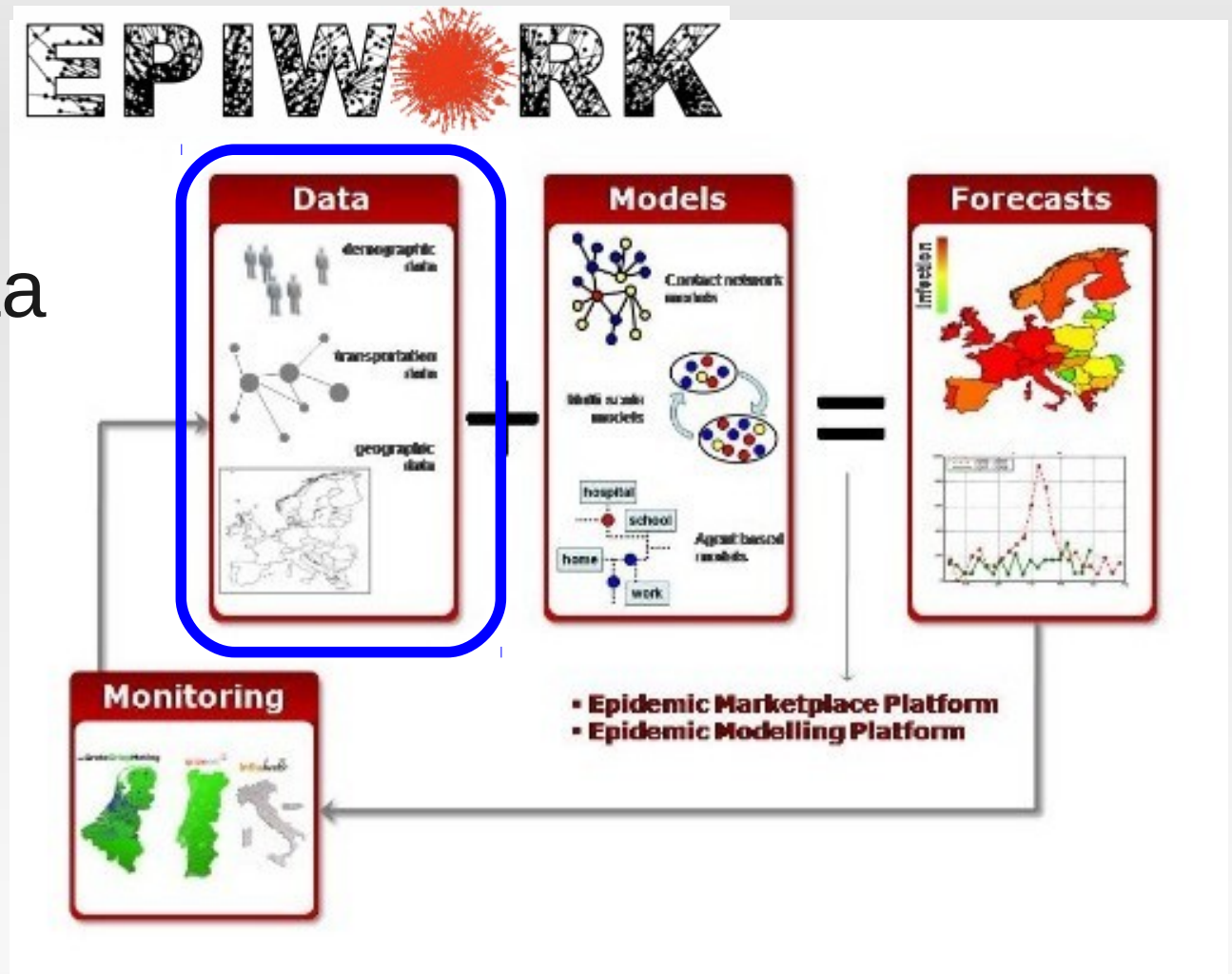
João Zamite, Fabrício A.B. Silva, Francisco Couto,  
Mário J. Silva

LaSIGE

EPIWORK - [epiwork@lasige.di.fc.ul.pt](mailto:epiwork@lasige.di.fc.ul.pt)

# EPIWORK

- What It Is
- Framework
- Getting the Data



# Motivation

- Where is the Data?
- Official Sources
  - ECDC – European Center for Disease Prevention and Control
  - CDC – Center for Disease Control
  - Clinical Data → Restricted Access
- The Web!

# Data on the Web

- Some Official Data
  - Euroflu.org
  - CDC.gov
- Epidemic Data Collection Applications

# GripeNet, FluSurvey, etc...

The image displays a collage of overlapping screenshots from various influenza monitoring websites. The most prominent screenshot is from **gripenet**, which features a news article titled "A estação da gripe chegou ao fim" (The flu season has ended) and a line graph titled "Influenzanet Portugal ILI Incidence". The graph compares the 2008-2009 season (grey line) and the 2009-2010 season (green line) across months from September to May. The 2009-2010 season shows a significantly lower peak in ILI incidence compared to the 2008-2009 season. Other visible screenshots include **Influenzanet**, **deGroteGriepMeting.nl**, **flusurvey**, **UK Flu Surveillance**, and **INFLUWEB**.

Deployed in Portugal, Netherlands, Belgium, Italy, Mexico, Brazil, United Kingdom, Australia and Canada



# Google Flu Trends

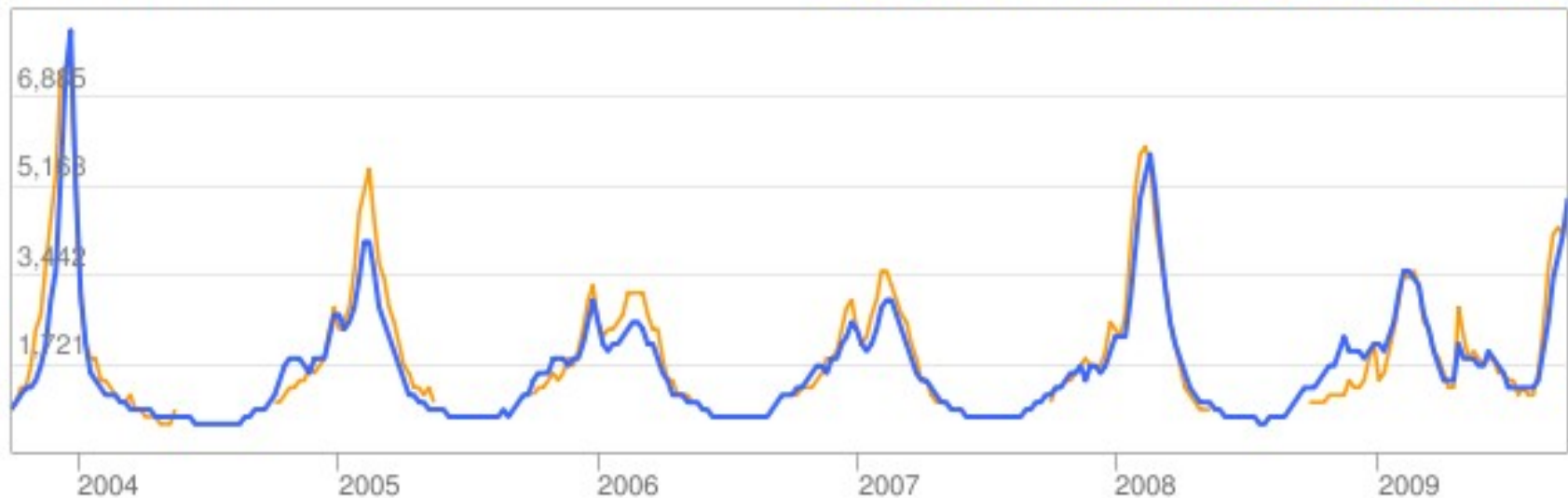
google.org Flu Trends



## United States Flu Activity

Influenza estimate

● Google Flu Trends estimate ● United States data



Estimate predictions → 2 Weeks earlier than CDC's Published Data

# Healthmap



# Social Networks

- Facebook, Twitter, etc



The image shows a screenshot of the Twitter search interface. At the top left is the Twitter logo. To its right is a search bar with the text "Search for a keyword or phrase..." and the word "Flu" entered. A "Search" button is to the right of the search bar. Below the search bar is a horizontal list of trending topics: Dilma, Gamescom, Berbuka, Expendables, Anelka, TRENDING TOPICS, Break Fasting, Laguna Beach, YOG, and Scott. The main content area is divided into two columns. The left column is titled "Top Trending Topics" and lists: #happygday, Rainer, #ausvotes, Inception, Dilma, Gamescom, #rootyq, Berbuka, Expendables, and Anelka. The right column is titled "Realtime results for Flu" and displays three tweets. Each tweet includes a small profile picture, the user's name, the text of the tweet, and the time it was posted.

**twitter**™ Search for a keyword or phrase...

Flu Search

Dilma Gamescom Berbuka Expendables Anelka TRENDING TOPICS Break Fasting Laguna Beach YOG Scott

### Top Trending Topics

- #happygday
- Rainer
- #ausvotes
- Inception
- Dilma
- Gamescom
- #rootyq
- Berbuka
- Expendables
- Anelka

### Realtime results for Flu

 **hokmargaret** Ugh. Down with the **flu**. Becoming best buddies with my couch. Reality TV is almost bearable in a marathon.  
less than 20 seconds ago via API

 **Gorygirl70** woke up to a ouchy throat. Better run to quick care. Its just like me to get the the **flu** before my big celebration. Damn it  
half a minute ago via web

 **DarkEldest** Staying home from school today, I think I have the stomach **flu**.  
half a minute ago via Twitter for iPhone

 **HafizulCarl** Welcome back fever n **flu** n everything in between..miz u guys..crap!!..  
half a minute ago via Twitter for BlackBerry®

# Problem

- Data Availability
  - Some epidemic applications do not share data.
- Heterogeneity
  - Different concepts
    - Diseases, Locations
  - Different formats
    - RSS, Atom, CSV, Plain Text
  - Different protocols
    - REST, SOAP, etc...

# Outline

- Requirements
- Implementation
- Example
- Conclusions
- Future Work

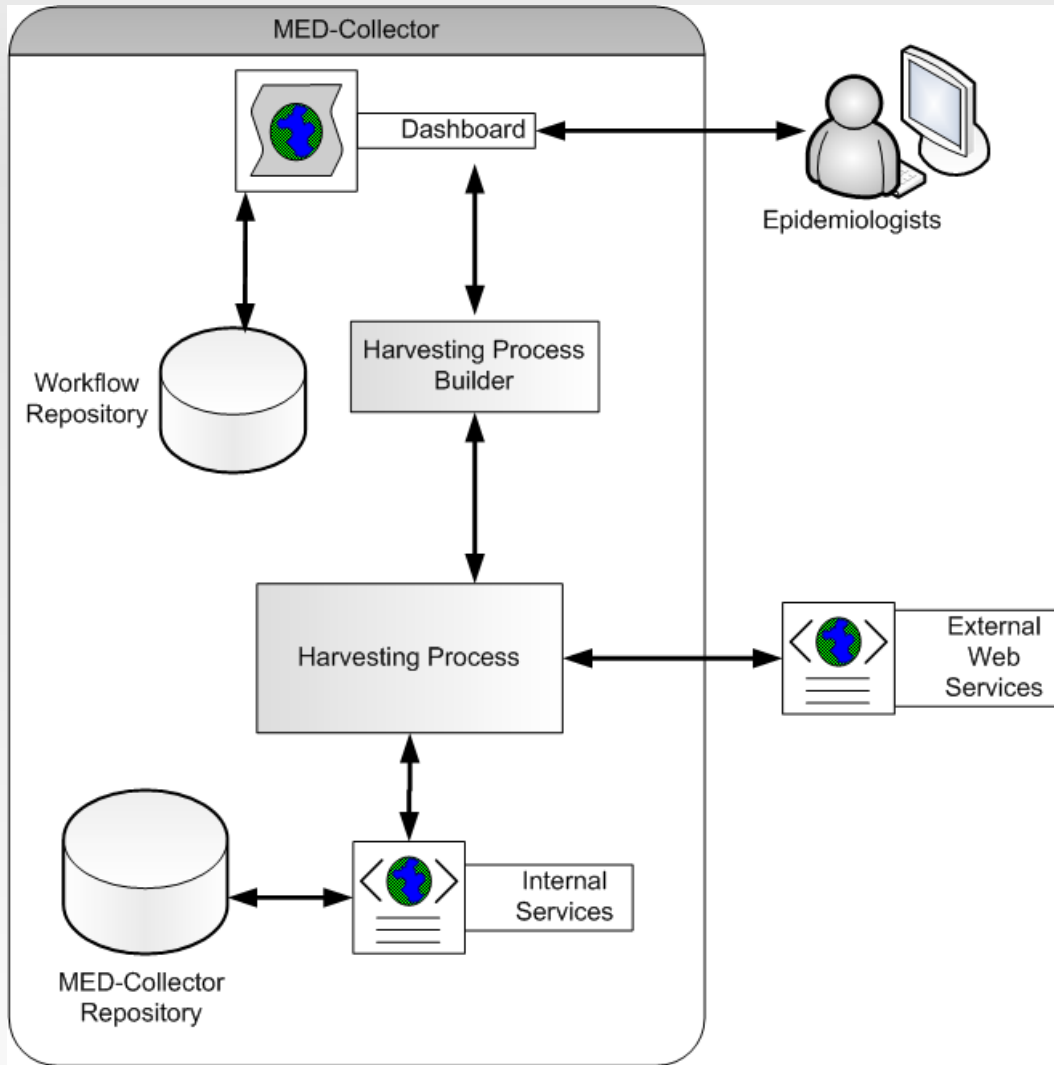
# Requirements

- Active Data Collection
- Passive Data Collection
- Flexible Scheduling
- Local Storage
- Ontologies
- Modularity and Configurability

# Our Approach

- Users Define HOW to collect the data.
- Workflow Design
  - Enables Flexibility and Configurability
  - Adaptable to each source of data

# Implementation



## Web Services

- Basic Functionalities

## BPEL Orchestrations

- Harvesting Processes

## Local Storage

- MEDCollector Repository

## Configurability

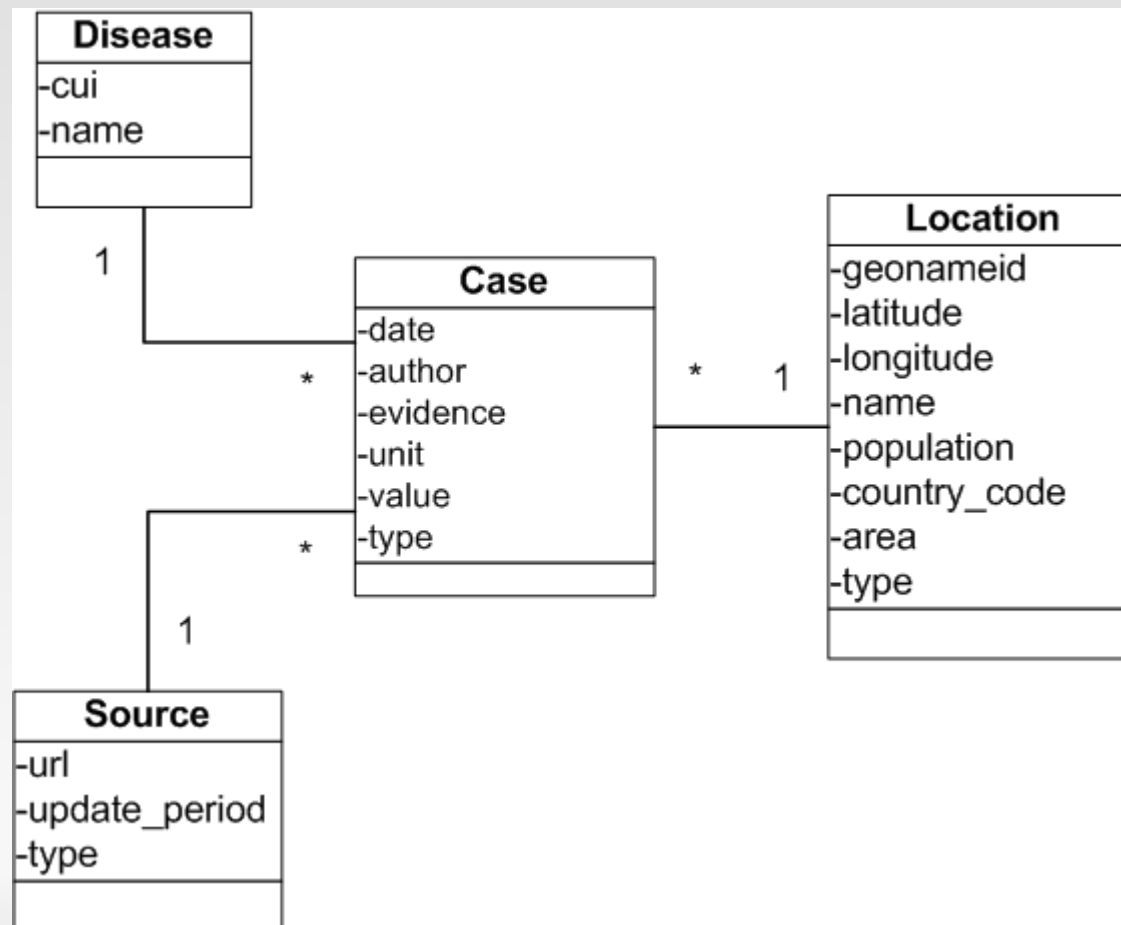
- Dashboard for Workflow Design

# Implementation

- Web Services
  - Query
  - Harvesting
  - XML Operations
    - Transformation, logic operators, etc...
  - Text
    - Translation, Mining...
  - Loading
    - Store the Data
  - Anything Else

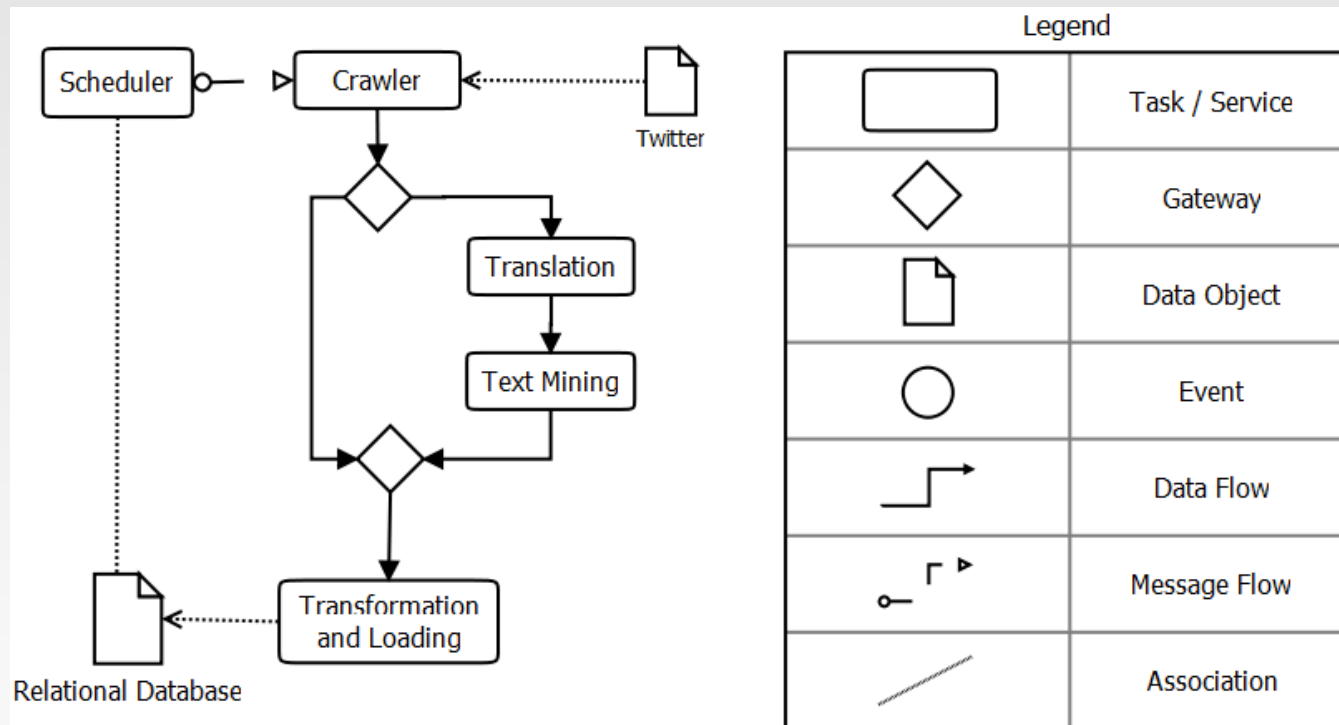
# Implementation

- MEDCollector Repository



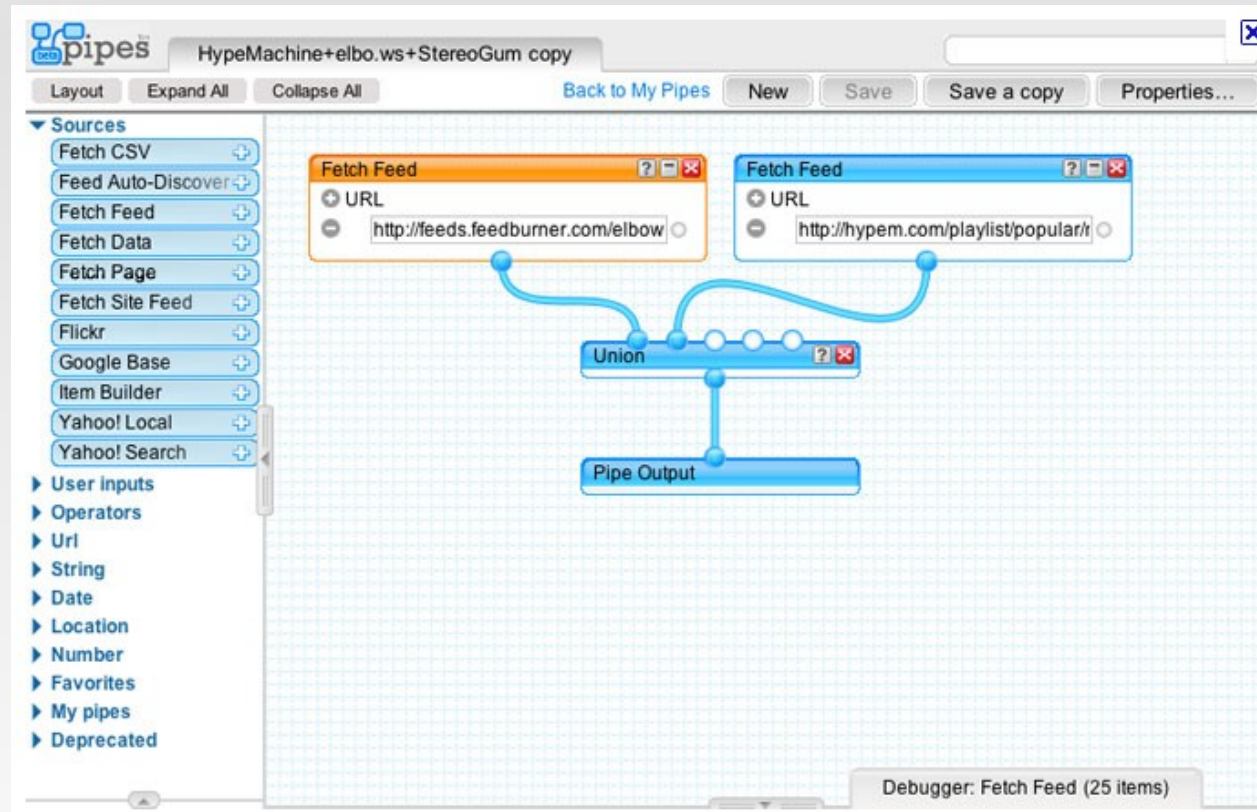
# BPEL Orchestrations

- Harvesting Processes
  - How Web Services are connected



# Some Interesting Approaches

- Yahoo Pipes

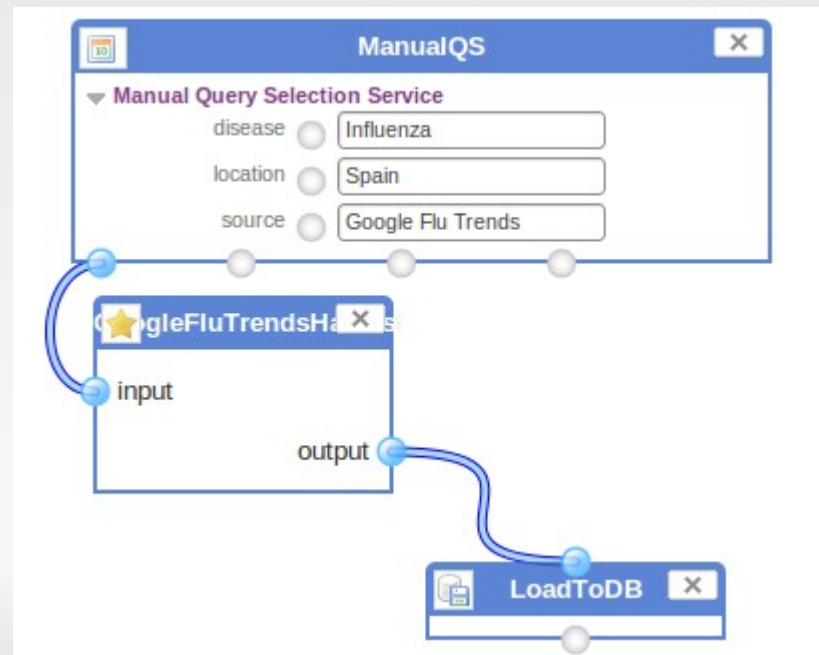


- uDesign

# Implementation

\*<http://javascript.neyric.com/wireit/>

- FrontEnd – WireIt \*
  - Browser Based
  - Similar to Yahoo Pipes
  - Less expressive than BPMN but simpler



# Example

- Collecting Data from CDC RSS Feeds
  - Video.



# Improvements

- View Workflows
  - Select the relevant data
  - Gather workflow design ideas
- More Services
  - More functionalities

# Conclusions

- Flexible Data Collection
- Simple Drag & Drop Interface
- BPEL orchestrations that collect data from the Web
- Active Collection
- Passive Collection
- Focus on Data Collection
  - Users can abstract from the complexities of technical workflow knowledge
  - Simple Data Transformation

# Future Work

- Privacy
  - Disease data could be very sensitive
  - Anonymization vs. Usefulness

# Acknowledgements

- European 7<sup>th</sup> Framework Programme
- EPIWORK Project Partners
- LaSIGE
- University of Lisbon