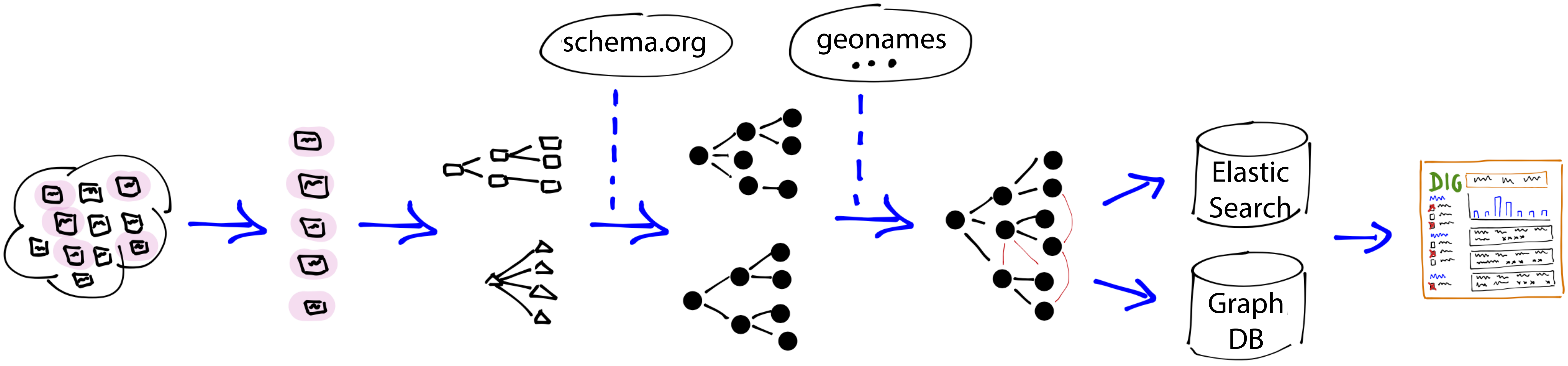
Information integration on the web

Mayank Kejriwal

Let’s go deeper into the architecture

Typical Web information integration architecture



Crawling

Extraction

Data Acquisition

Mapping To Ontology

Entity Linking & Similarity

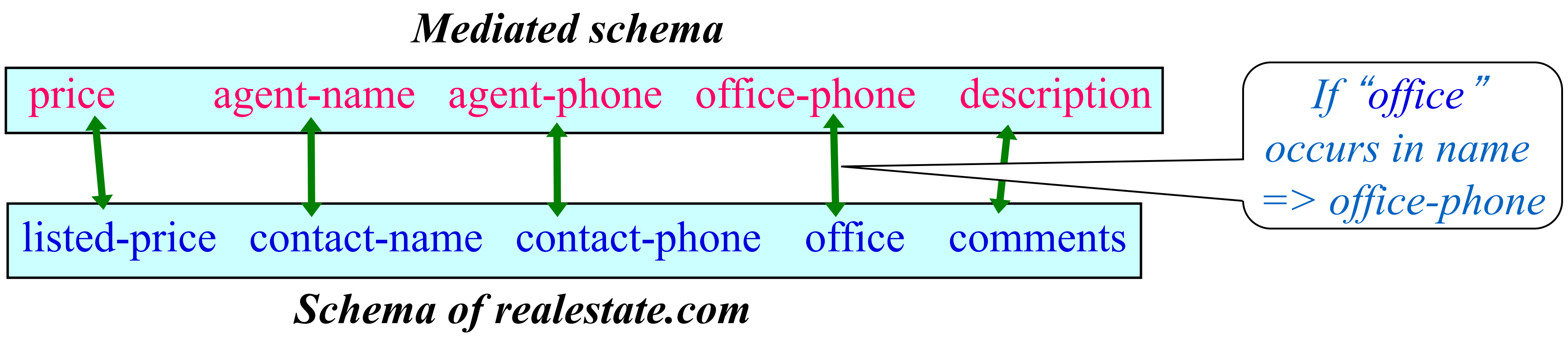
Knowledge Graph Deployment

Query & Visualization

Schema Mapping

- Given two different sources with different schemas, how do we automatically align the information
- Research Topics
  - Automatic schema alignment based on structure and naming
  - Automatic alignment based on the source contents

Schema Mapping



Mediated schema

price agent-name agent-phone office-phone description

listed-price contact-name contact-phone office comments

Schema of realestate.com

If "office" occurs in name => office-phone

realestate.com

listed-price	contact-name	contact-phone	office	comments
$250K	James Smith	(305) 729 0831	(305) 616 1822	Fantastic house
$320K	Mike Doan	(617) 253 1429	(617) 112 2315	Great location
....	....	....	....	....

If "fantastic" & "great" occur frequently in data instances => description

homes.com

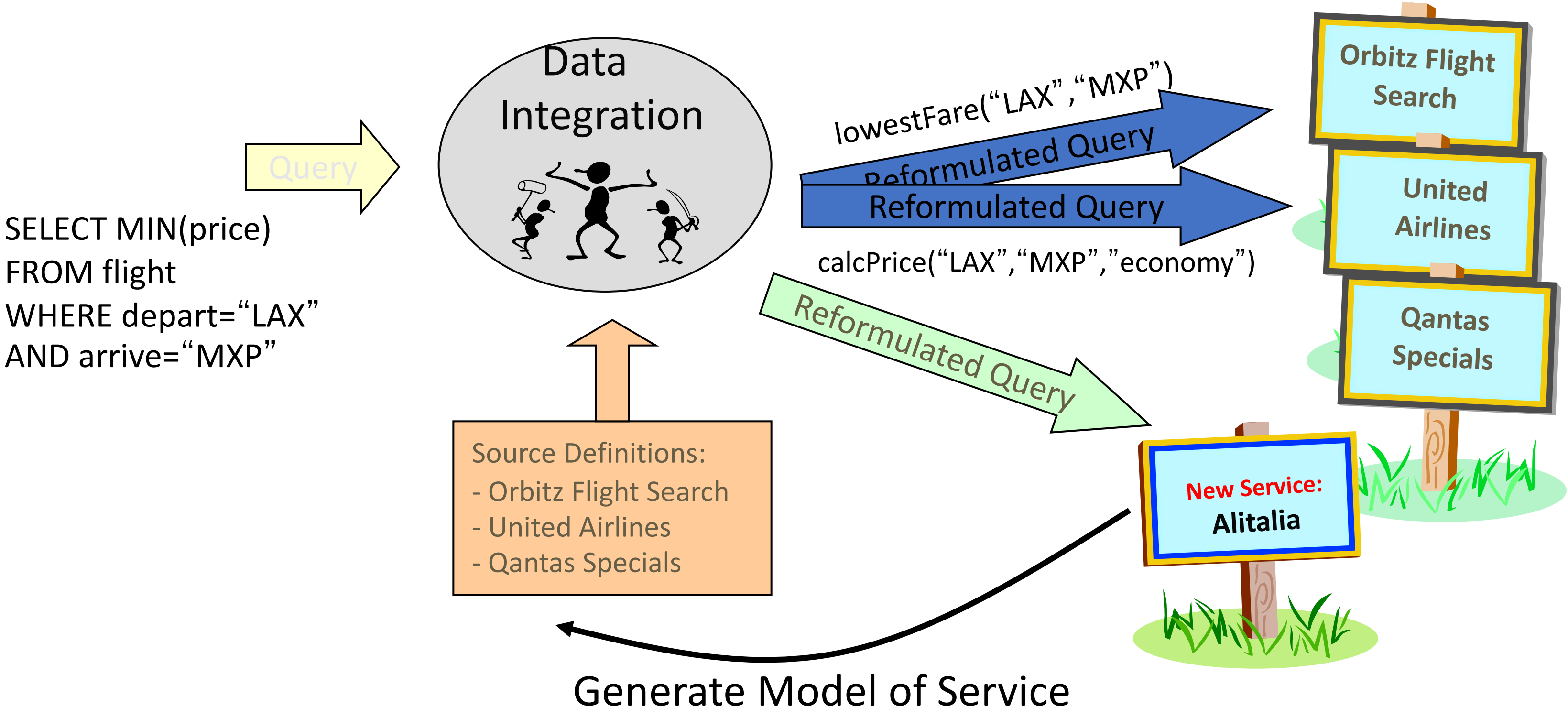
sold-at	contact-agent	extra-info
$350K	(206) 634 9435	Beautiful yard
$230K	(617) 335 4243	Close to Seattle

Source Modeling

- Semantic typing
- Source discovery
- Automatic source modeling
- Interactive source modeling

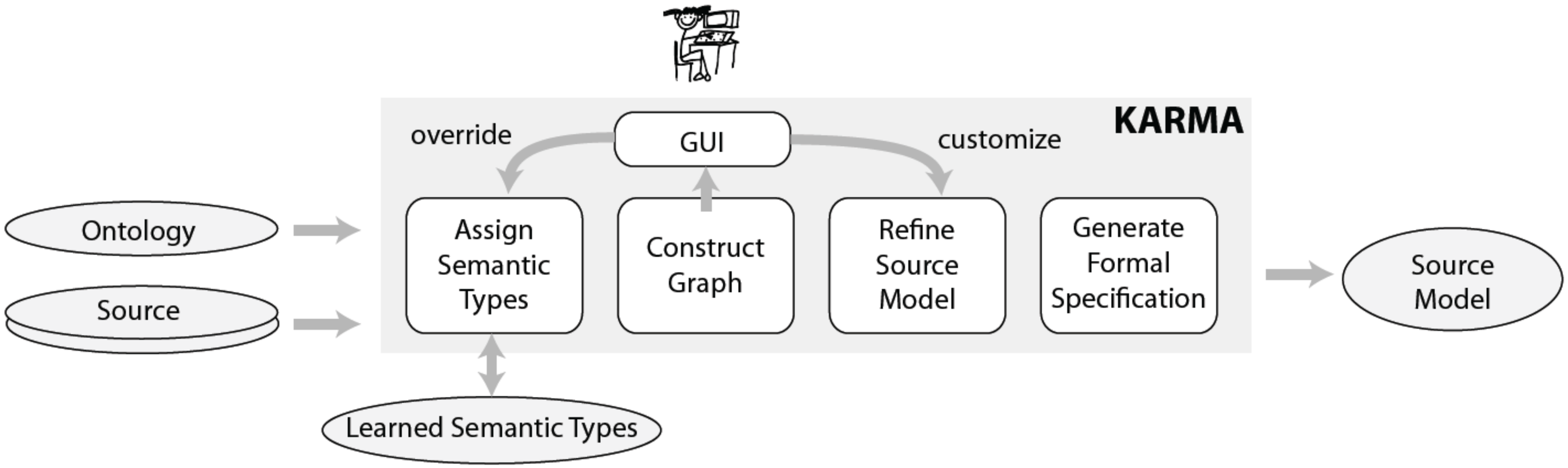
Automatic Source Modeling

- How to learn semantic descriptions of sources and services



```
SELECT MIN(price)
FROM flight
WHERE depart="LAX"
AND arrive="MXP"
```

Semi-Automatic Source Modeling



Relational database to RDF mapping

Anything to RDF mapping

String Similarity:
Why Strings Don’t Match Perfectly?

typos	"Joh” vs “John”
OCR errors	"J0hn” vs “John”
formatting conventions	”03/17” vs “March 17”
abbreviations	”J. S. Sargent” vs “John Singer Sargent”
nick names	”John” vs “Jock”
word order	”Sargent, John S.” vs “John S. Sargent”

String Similarity Problem Definition

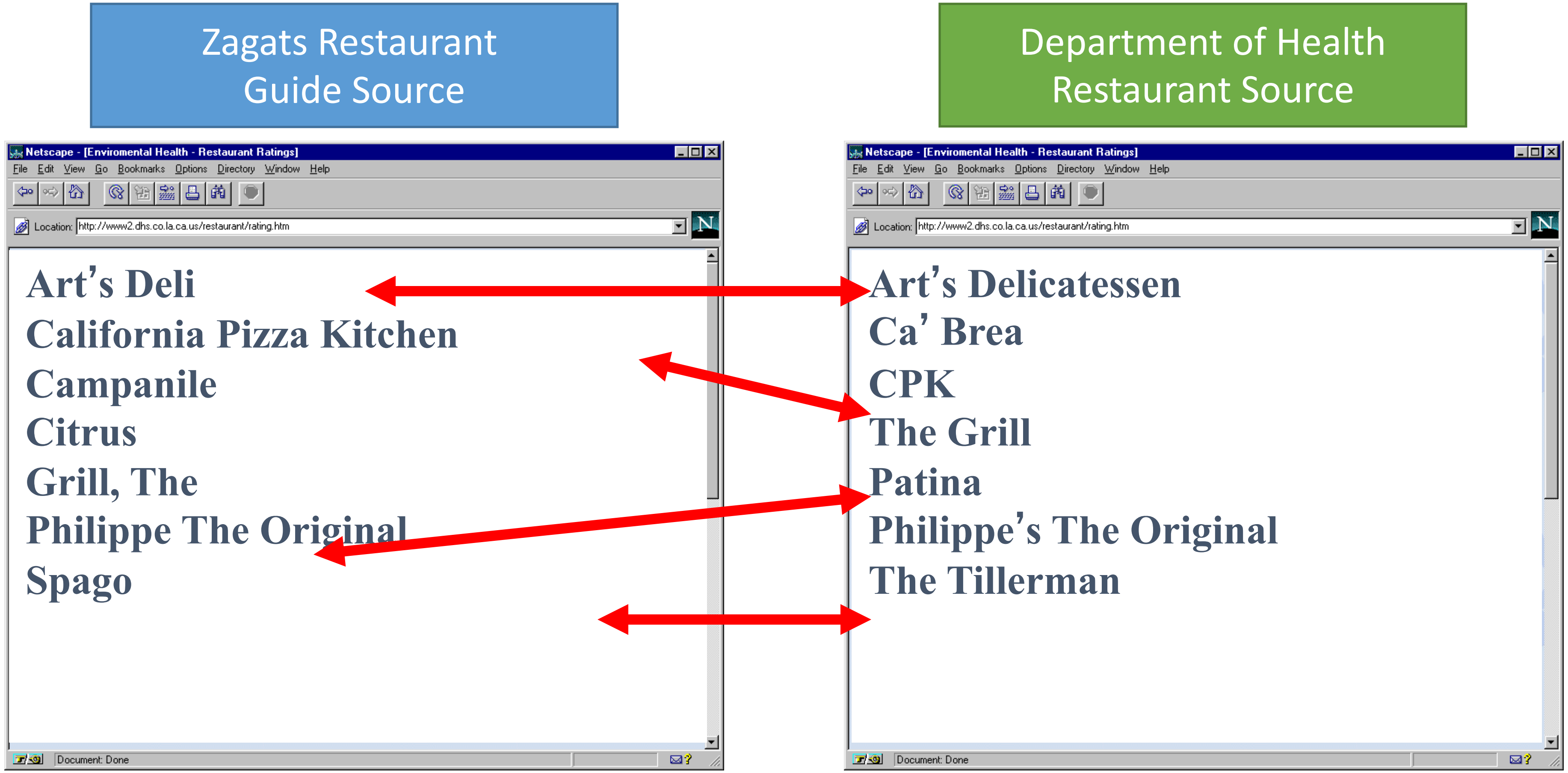
Given X and Y sets of strings

Find pairs (x, y)
such that both x and y
refer to the same real world entity

"John S. Sargent"

"John Singer Sargent"

Record Linkage



Zagats Restaurant Guide Source

Art's Deli
California Pizza Kitchen
Campanile
Citrus
Grill, The
Philippe The Original
Spago

Department of Health Restaurant Source

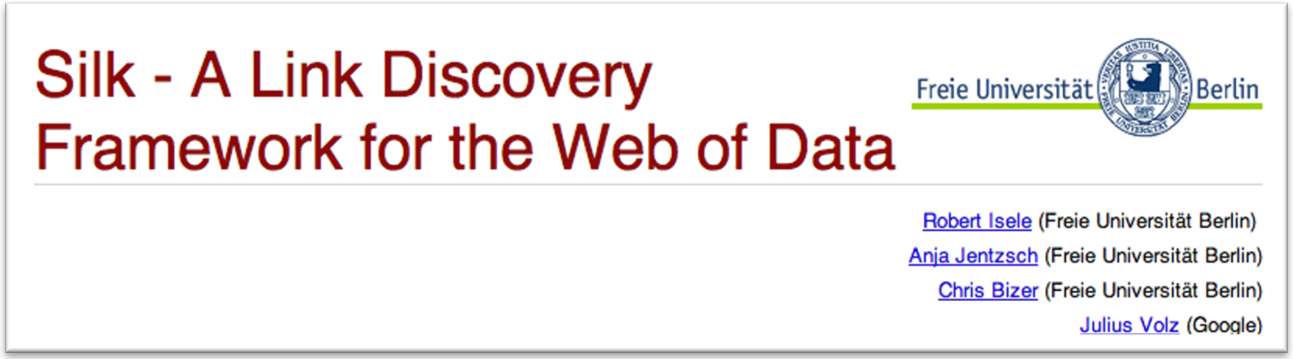
Art's Delicatessen
Ca' Brea
CPK
The Grill
Patina
Philippe's The Original
The Tillerman

Netscape - [Enviromental Health - Restaurant Ratings]
File Edit View Go Bookmarks Options Directory Window Help
Location: http://www2.dhs.co.la.ca.us/restaurant/rating.htm
Document: Done
Netscape - [Enviromental Health - Restaurant Ratings]
File Edit View Go Bookmarks Options Directory Window Help
Location: http://www2.dhs.co.la.ca.us/restaurant/rating.htm
Document: Done

How can the same objects be identified
when they are stored in inconsistent text formats?

Record Linkage

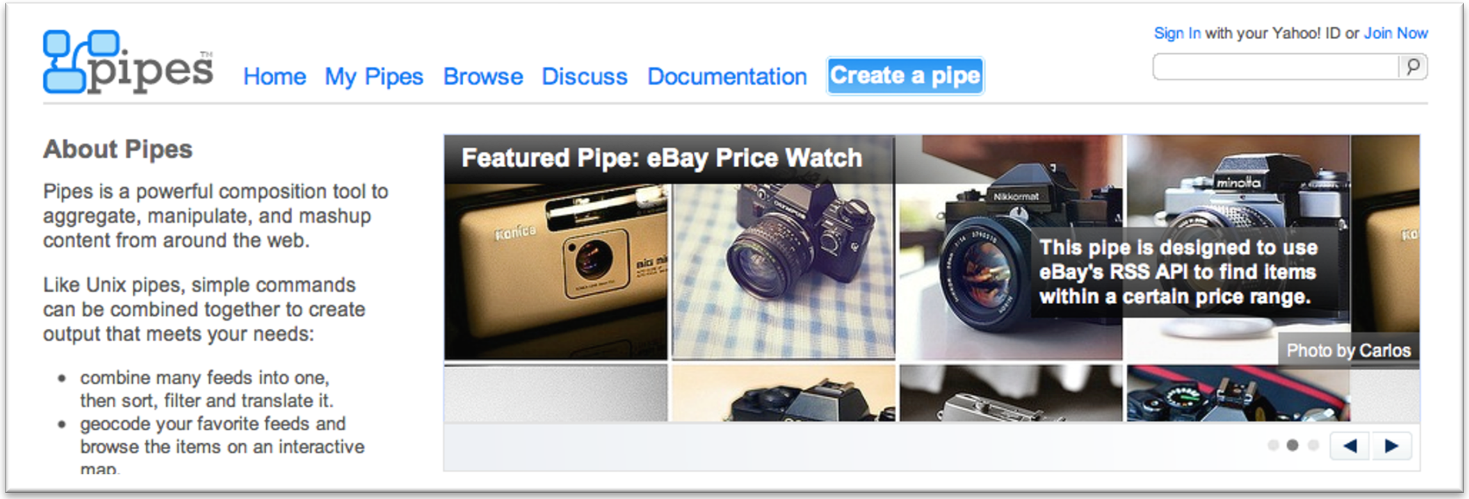
- Align entities across sources
- Research Topics:
  - Blocking
  - Matching individual attributes
  - Matching records
  - Matching entities



Silk - A Link Discovery Framework for the Web of Data

Robert Isele (Freie Universität Berlin)
Anja Jentzsch (Freie Universität Berlin)
Chris Bizer (Freie Universität Berlin)
Julius Volz (Google)

Mashup Construction



About Pipes

Pipes is a powerful composition tool to aggregate, manipulate, and mashup content from around the web.

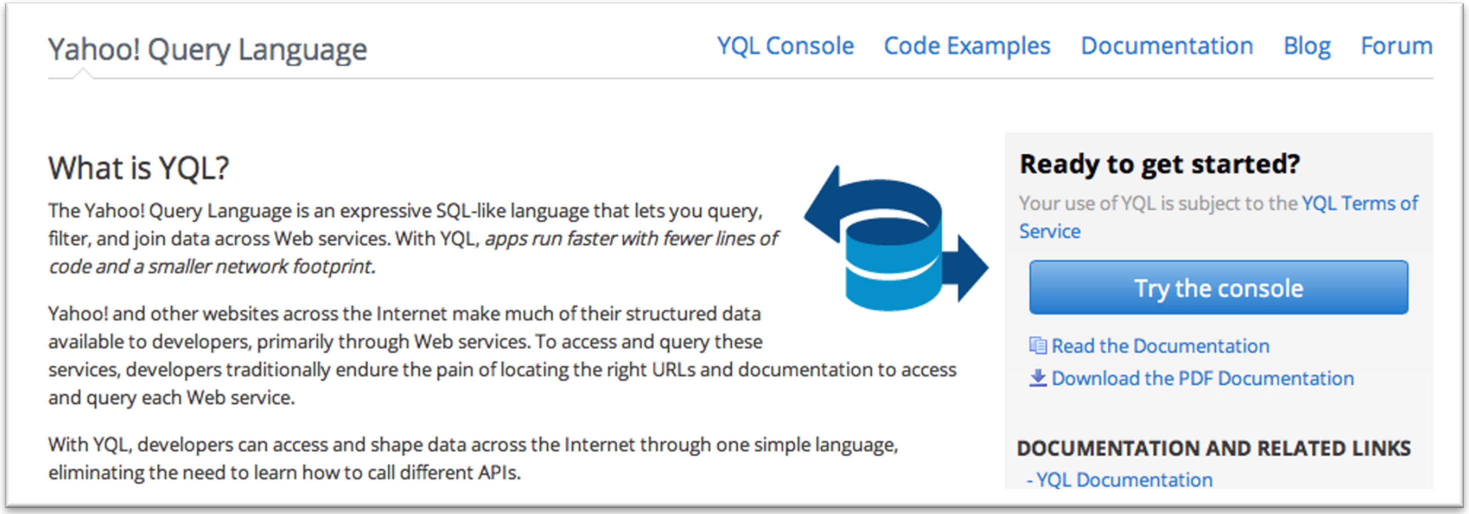
Like Unix pipes, simple commands can be combined together to create output that meets your needs:

- combine many feeds into one, then sort, filter and translate it.
- geocode your favorite feeds and browse the items on an interactive map.

Featured Pipe: eBay Price Watch

This pipe is designed to use eBay's RSS API to find items within a certain price range.

Photo by Carlos



What is YQL?

The Yahoo! Query Language is an expressive SQL-like language that lets you query, filter, and join data across Web services. With YQL, *apps run faster with fewer lines of code and a smaller network footprint.*

Yahoo! and other websites across the Internet make much of their structured data available to developers, primarily through Web services. To access and query these services, developers traditionally endure the pain of locating the right URLs and documentation to access and query each Web service.

With YQL, developers can access and shape data across the Internet through one simple language, eliminating the need to learn how to call different APIs.

Ready to get started?

Your use of YQL is subject to the YQL Terms of Service

DOCUMENTATION AND RELATED LINKS

- YQL Documentation

Ontology-based data access and integration

- Use ontology language as domain model
  - OWL2 profiles
- Answering queries under description logic constraints
  - Unions of conjunctive queries
  - Datalog