# Web semantization via dynamic semantics

Peter Vojtáš, MFF UK Praha

# Abstract

Our goal is to extend the semantic web foundations to enable describing the **semantization process**.

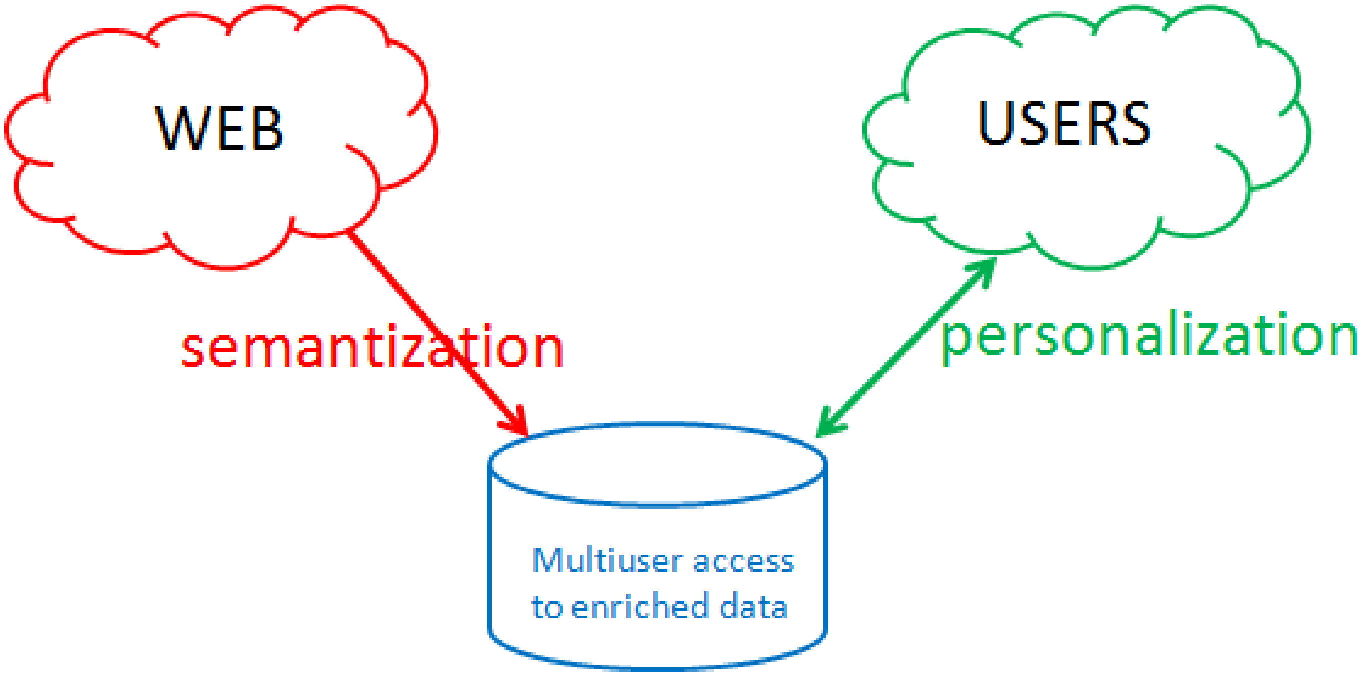
Considering RDF triples, one can ask **where these triples are from**: have they been written by human publishers, extracted (e.g., from structured parts of WikiPedia) by rules edited by humans, or by (inductive) programs trained to extract, e.g., subjects (named entities), properties or property values?

A typical example is the automated extraction of item properties on a retail web. We refer to several diploma/PhD theses containing practical semantization **experiments**.

To describe the reliability of the obtained RDF data we propose a **"half-a-way" extension of dynamic logic**: programs (extractors) remain propositional, Kripke states are web pages, and there is a lot of reification describing the training and testing data and the metrics of learning.

# SemPre research group KSI MFF UK

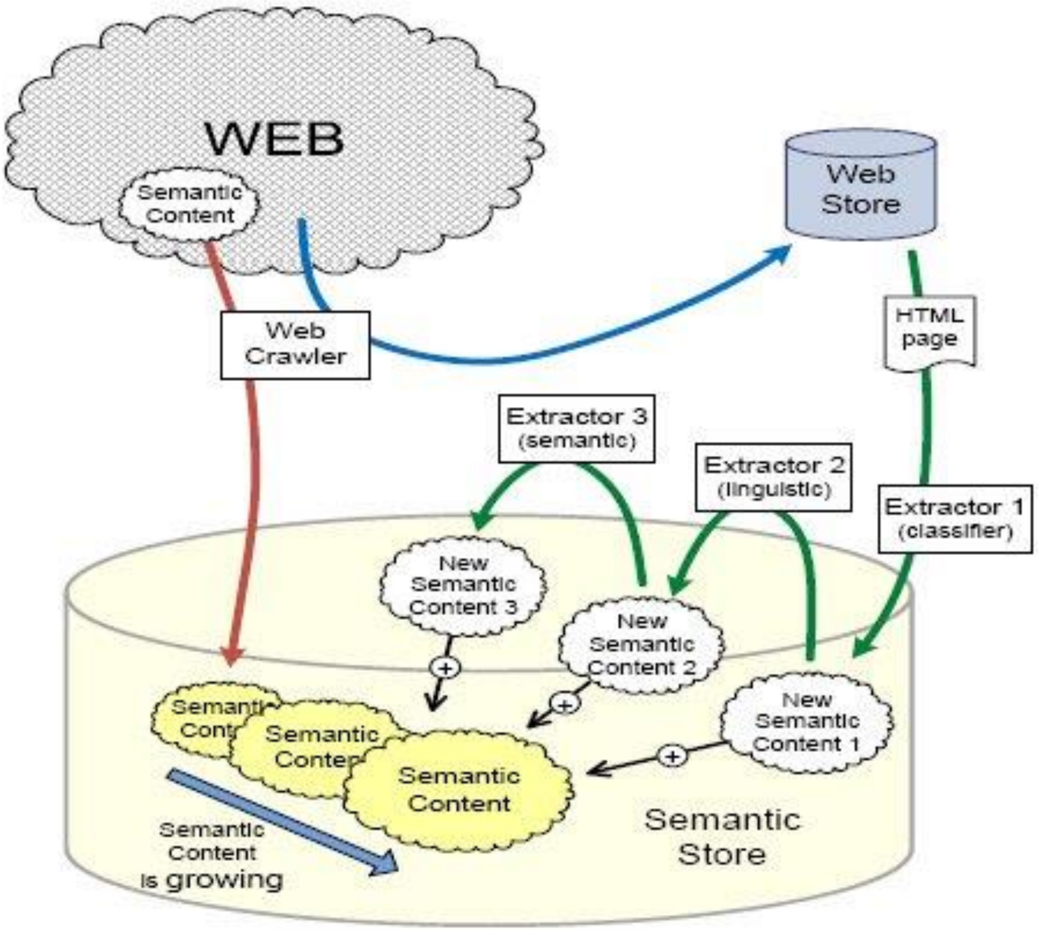
most of materials available from http://www.ksi.mff.cuni.cz/~vojtas/



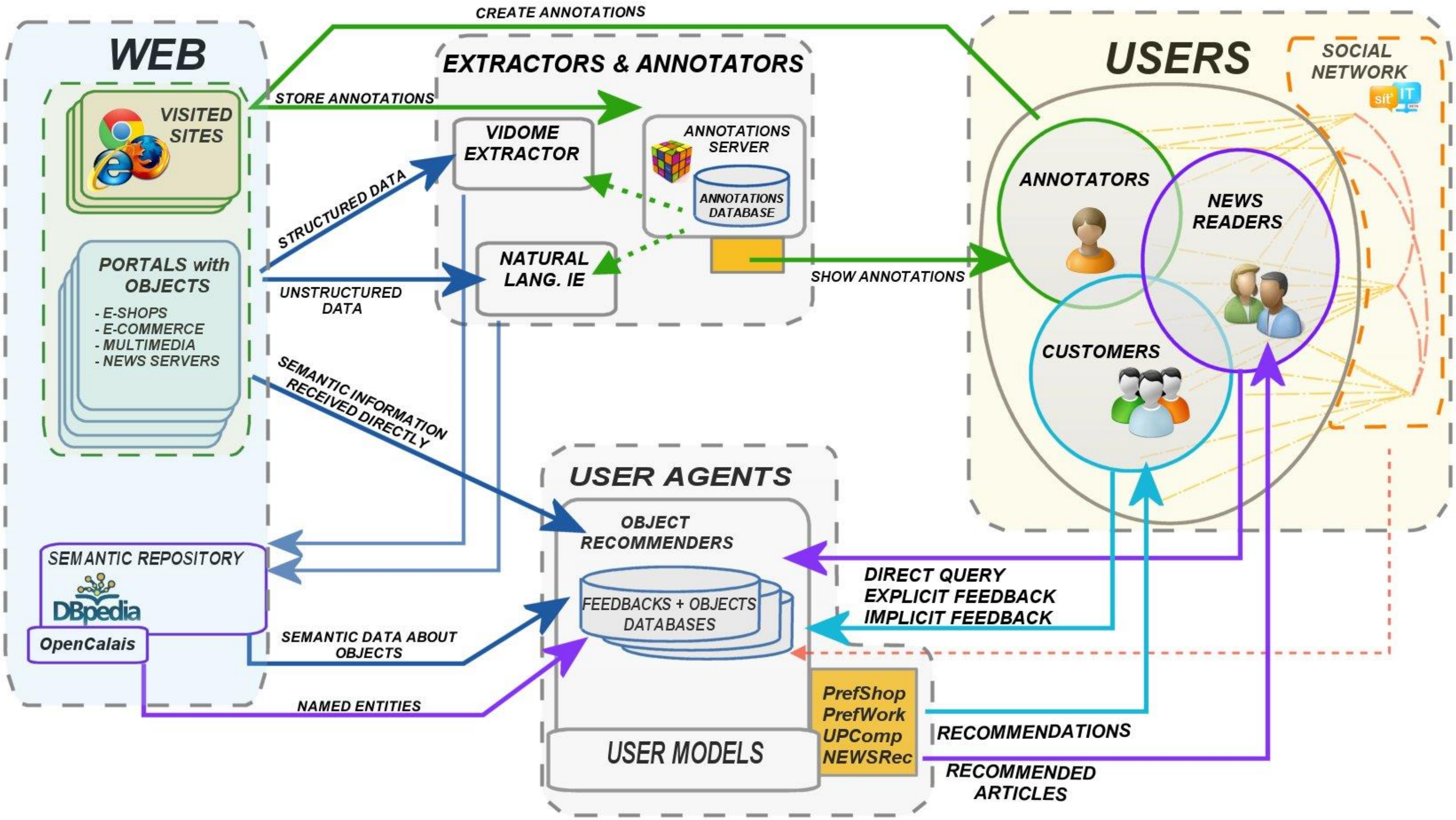
# Web Semantization – Our First Approach

- Generic web crawler, crawl whole czech web
- Various semantic extractors

- No user aspect
- No intended purpose of the data
- Who creates ontologies?



# Web Semantization – Our Vision

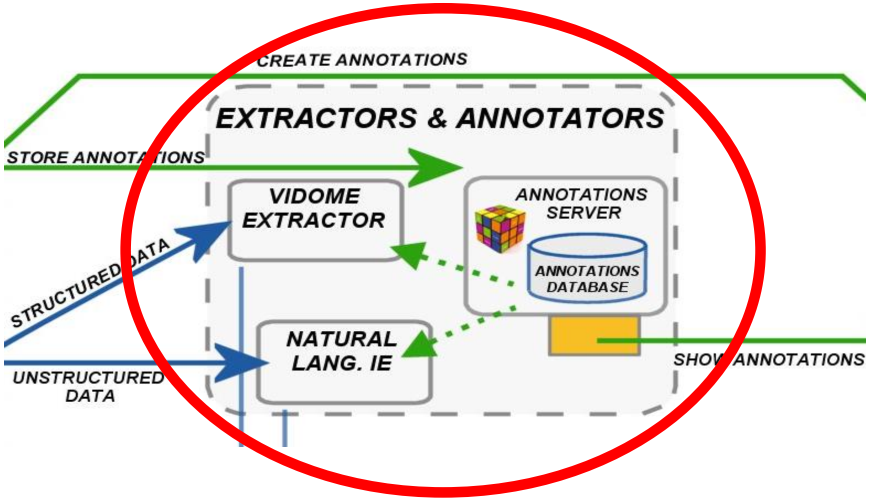


# Web Semantization – Our Vision

- Employ users
  - As a source of some semantical data
  - As consumers of added value
- Semantic data should have reason / application model why to be collected
- Several tools processing parts of our model developed, initial integration steps

# Semantic Data Extracting

- Tools to gather unstructured and semi-structured data:
  - Information Extraction from Natural Language
  - Information Extraction based on structural similarity
  - Domain dependant annotations

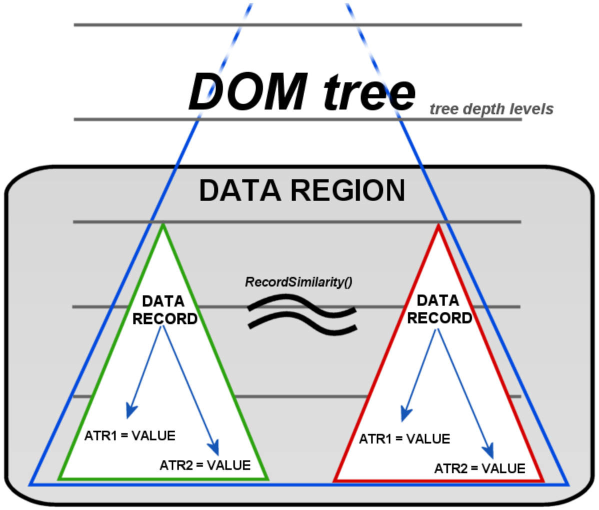


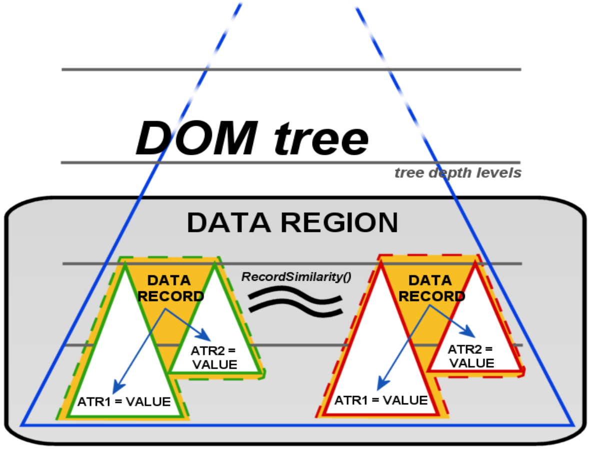
# Semi-Structured Data Extracting

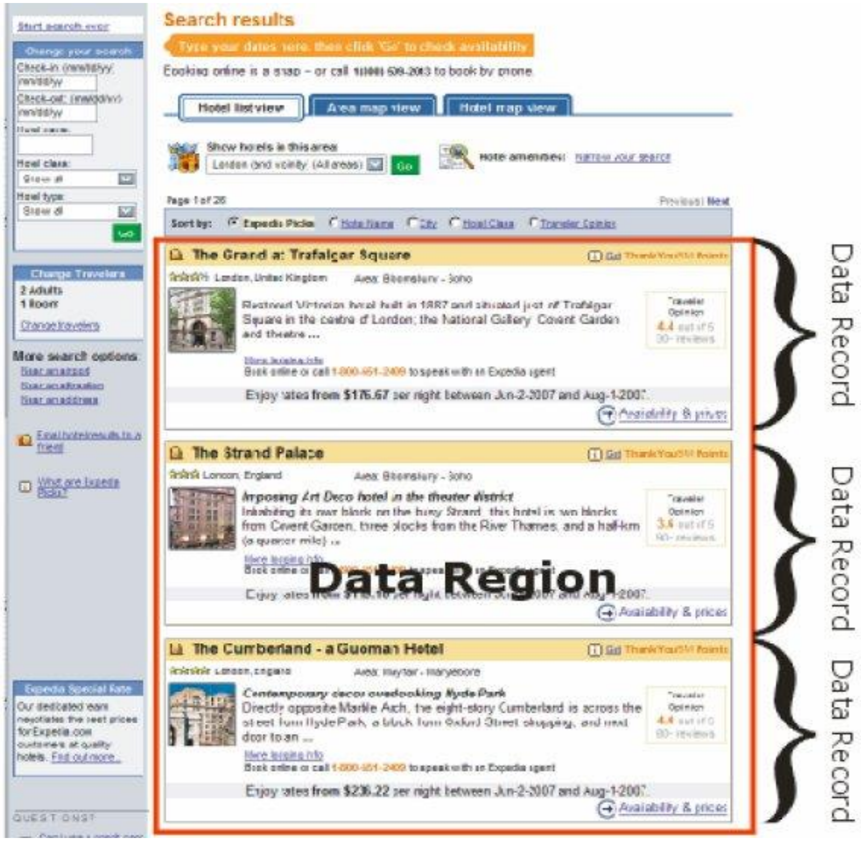
- To identify objects on e.g. e-commerce category page:
  - Contains more objects (records) of the same type
  - The records have similar structure (DOM)
  - The attributes can be identified via ontology + RegExp
  - Record similarity via Levenshtein distance

- Several problems occured e.g. records are not trees, but forrests

# Semi-Structured Data Extracting – Example - Maruscak







Search results

Type your dates here, then click 'Go' to check availability

Booking online is a snap – or call 1(800) 539-2013 to book by phone

Our dedicated team negotiates the best prices for Expedia.com customers at quality hotels. Find out more...

Page 1 of 26

Sort by: Expedia Picks, Hotel Name, City, Hotel Class, Traveler Opinion

**The Grand at Trafalgar Square**

London, United Kingdom Area: Bloomsbury - Soho

Restored Victorian hotel built in 1887 and situated just off Trafalgar Square in the centre of London; the National Gallery, Covent Garden and theatre ...

4.4 out of 5

More lodging info

Book online or call 1-800-451-2409 to speak with an Expedia agent

Enjoy rates from $175.67 per night between Jun-2-2007 and Aug-1-2007.

**The Strand Palace**

London, England Area: Bloomsbury - Soho

*Imposing Art Deco hotel in the theater district*

Inhabiting its own block on the busy Strand, this hotel is two blocks from Covent Garden, three blocks from the River Thames, and a half-km (a quarter mile) ...

3.6 out of 5

More lodging info

Enjoy rates from [illegible] per night between Jun-2-2007 and Aug-1-2007.

**The Cumberland - a Guoman Hotel**

London, England Area: Mayfair - Marylebone

*Contemporary decor overlooking Hyde Park*

Directly opposite Marble Arch, the eight-story Cumberland is across the street from Hyde Park, a block from Oxford Street shopping, and next door to an ...

4.4 out of 5

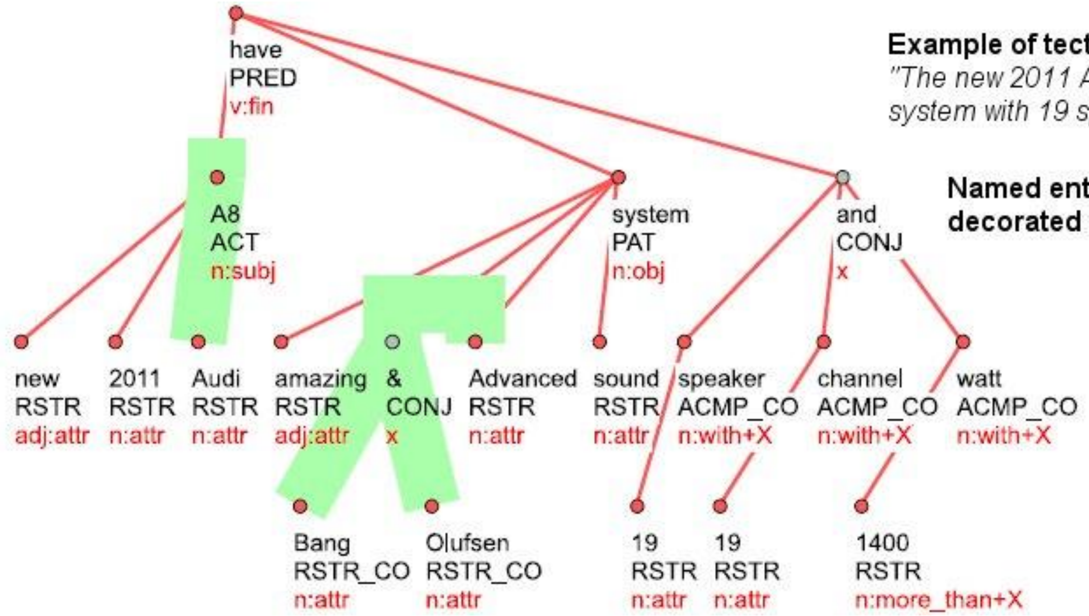
More lodging info

Book online or call 1-800-451-2409 to speak with an Expedia agent

Enjoy rates from $236.22 per night between Jun-2-2007 and Aug-1-2007.

# Unstructured Information Extracting - Dedek

- From natural language – e.g. news articles
  - Several linguistic tools (tokenizer, morphological analyzers…)
  - Named entity recognition, data aggregation, new attributes etc.

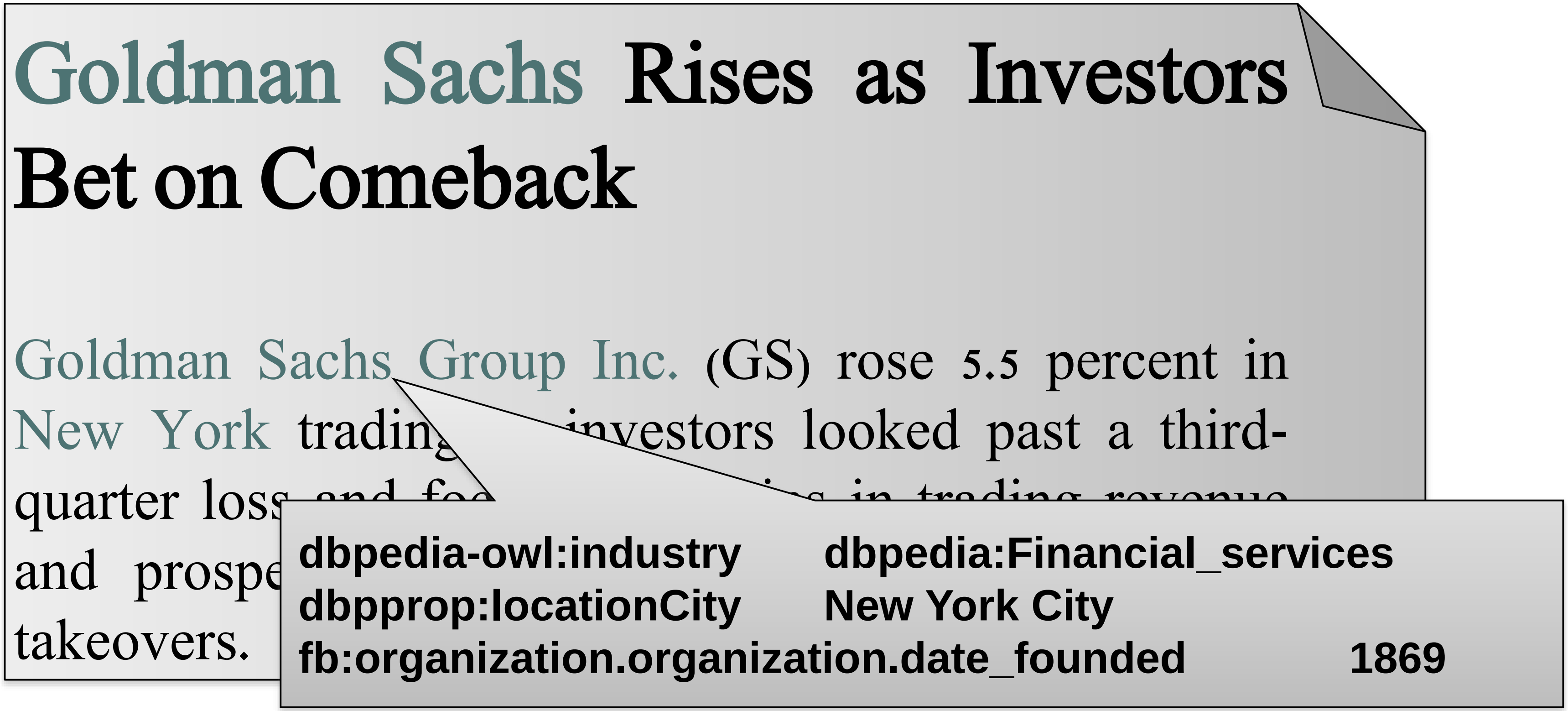


**Example of tectogramatical linguistic tree. Source sentence:**
*"The new 2011 Audi A8 has an Bang & Olufsen Advanced sound system with 19 speakers, 19 channels and more than 1400 wats."*

**Named entities Audi A8 and Bang & Olufsen are decorated by green color over corresponding nodes**

# Unstructured Information Extracting – News Recommending - Lasek

- **Outsourcing named entities**



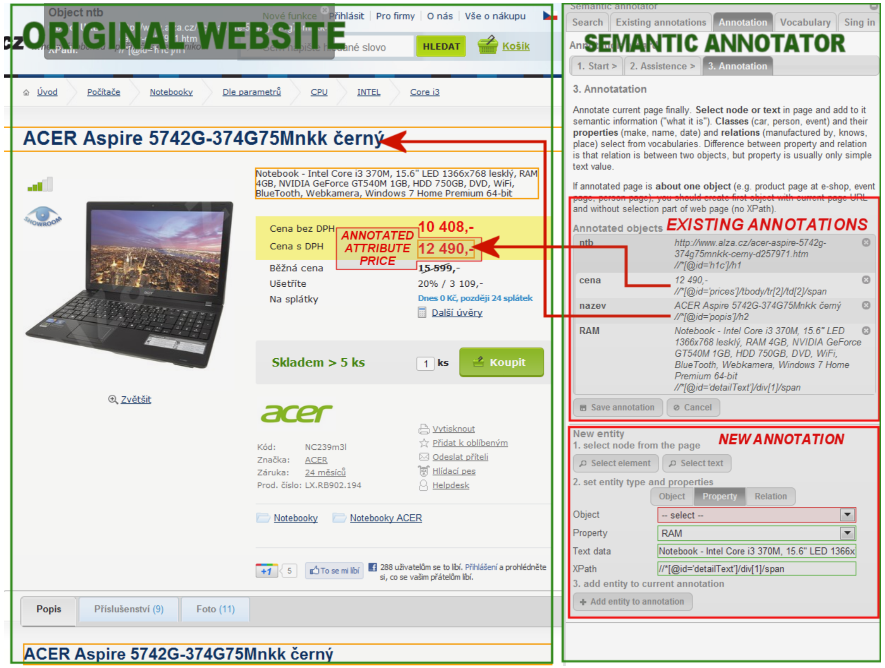
Goldman Sachs Rises as Investors Bet on Comeback

Goldman Sachs Group Inc. (GS) rose 5.5 percent in New York tradin investors looked past a third-quarter loss and fee ies in trading revenue and prospe takeovers.

| | |
|---|---|
| dbpedia-owl:industry | dbpedia:Financial_services |
| dbpprop:locationCity | New York City |
| fb:organization.organization.date_founded | 1869 |

# Domain Specific User Annotations - Fiser

- Annotations based on ontology specified by user
- Collaborative benefit from other users annotations
- Work in progress on machine annotating of similar pages



ACER Aspire 5742G-374G75Mnkk černý

Notebook - Intel Core i3 370M, 15.6" LED 1366x768 lesklý, RAM 4GB, NVIDIA GeForce GT540M 1GB, HDD 750GB, DVD, WiFi, BlueTooth, Webkamera, Windows 7 Home Premium 64-bit

| | |
|---|---|
| Cena bez DPH | 10 408,- |
| Cena s DPH | 12 490,- |
| Běžná cena | ~~15 599~~,- |
| Ušetříte | 20% / 3 109,- |
| Na splátky | Dnes 0 Kč, později 24 splátek |

Další úvěry

1 ks

Kód: NC239m3l
Značka: ACER
Záruka: 24 měsíců
Prod. číslo: LX.RB902.194

5

288 uživatelům se to líbí. Přihlášení a prohlédněte si, co se vašim přátelům líbí.

ACER Aspire 5742G-374G75Mnkk černý

3. Annotatation

Annotate current page finally. **Select node or text** in page and add to it semantic information ("what it is"). **Classes** (car, person, event) and their **properties** (make, name, date) and **relations** (manufactured by, knows, place) select from vocabularies. Difference between property and relation is that relation is between two objects, but property is usually only simple text value.

If annotated page is **about one object** (e.g. product page at e-shop, event page, person page), you should create first object with current page URL and without selection part of web page (no XPath).

Annotated objects

| | |
|---|---|
| ntb | *http://www.alza.cz/acer-aspire-5742g-374g75mnkk-cerny-d257971.htm*<br>*//*[@id='h1c']/h1* |
| cena | *12 490,-*<br>*//*[@id='prices']/tbody/tr[2]/td[2]/span* |
| nazev | *ACER Aspire 5742G-374G75Mnkk černý*<br>*//*[@id='popis']/h2* |
| RAM | *Notebook - Intel Core i3 370M, 15.6" LED 1366x768 lesklý, RAM 4GB, NVIDIA GeForce GT540M 1GB, HDD 750GB, DVD, WiFi, BlueTooth, Webkamera, Windows 7 Home Premium 64-bit*<br>*//*[@id='detailText']/div[1]/span* |

New entity

1. select node from the page

2. set entity type and properties

| | |
|---|---|
| Object | -- select -- |
| Property | RAM |
| Text data | Notebook - Intel Core i3 370M, 15.6" LED 1366x |
| XPath | //*[@id='detailText']/div[1]/span |

3. add entity to current annotation

# Model of dynamic web semantization

- Basic problems and vision of automation of web content processing
  - So far …
  - Challenge of integration data and algorithms models for semantization
  - Recall RDF model
  - Recall PDyL model
  - Let's try to integrate
  - A proposal
  - Conclusions

# Kripke, BoW, NER,metrics, process

For $\mathfrak{K}$ Kripke structure

$\Pi_0 = \{\alpha, \beta\}$

$\Phi_0 = \{p, q\}$

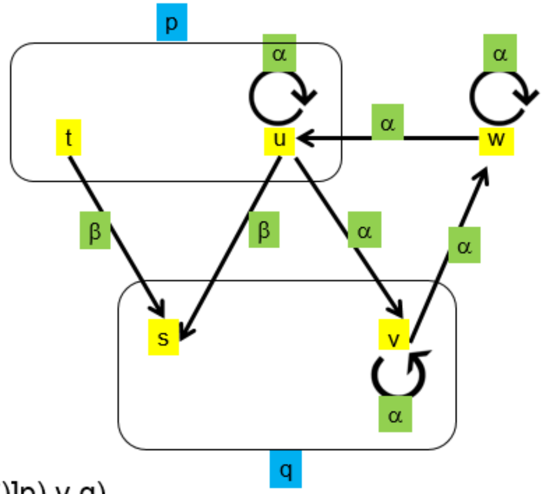
$K = \{u.v.w.s.t\}$

$m_{\mathfrak{K}}(p) = \{u.t\}$

$m_{\mathfrak{K}}(q) = \{s.v\}$

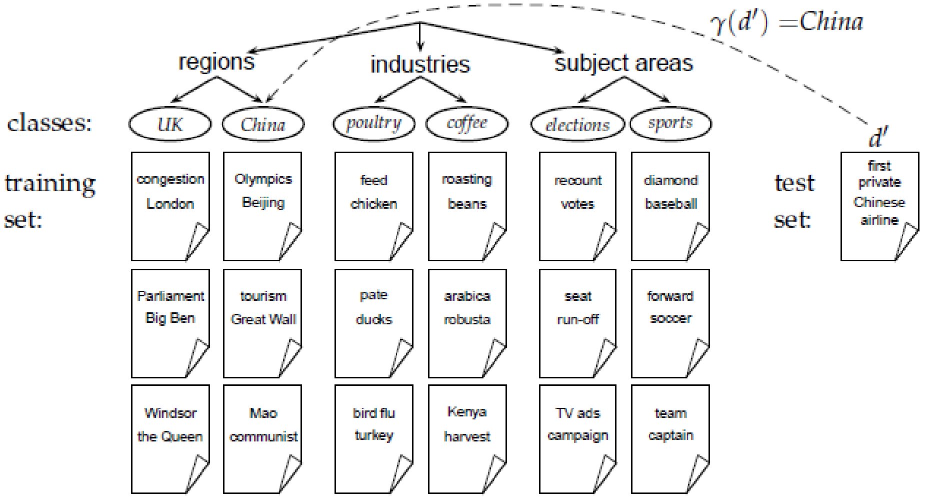
$m_{\mathfrak{K}}(\alpha) = \{(u.u),(u.v)(v.v), (v.w), (w.w), (w.u)\}$

$m_{\mathfrak{K}}(\beta) = \{(u.s), (t.s)\}$

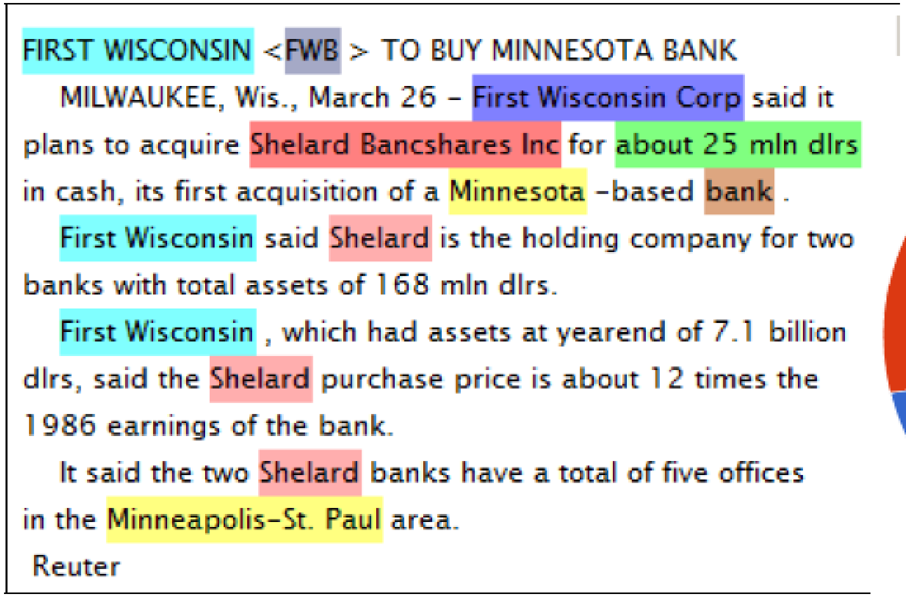


Specify $m_{\mathfrak{K}}(\alpha;(\beta^*))\cup(\beta^*))$,

$m_{\mathfrak{K}}(([(\alpha;(\beta^*))\cup(\beta^*)]p) \vee q)$



► **Figure 13.1** Classes, training set, and test set in text classification .



FIRST WISCONSIN <FWB > TO BUY MINNESOTA BANK

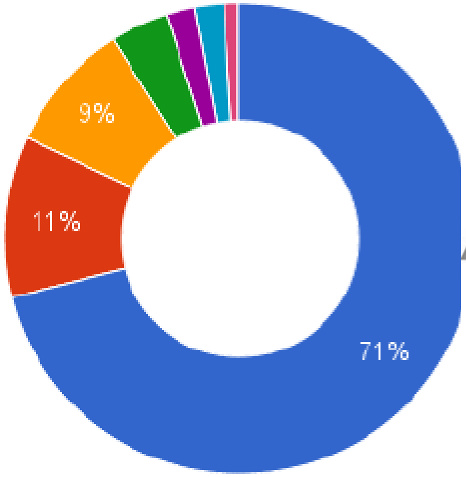
MILWAUKEE, Wis., March 26 - First Wisconsin Corp said it plans to acquire Shelard Bancshares Inc for about 25 mln dlrs in cash, its first acquisition of a Minnesota -based bank .

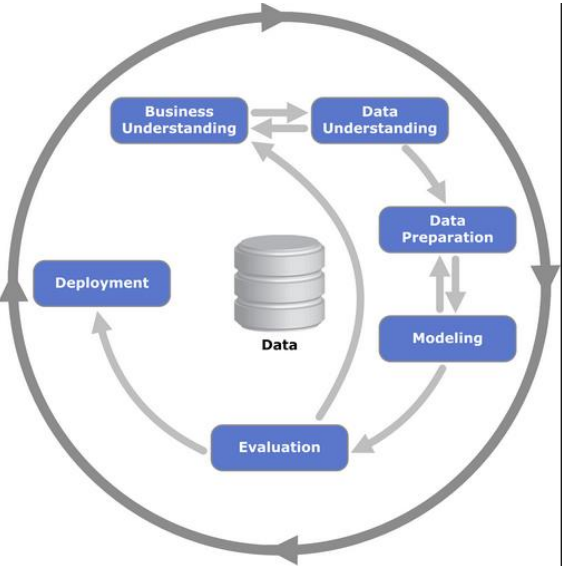
First Wisconsin said Shelard is the holding company for two banks with total assets of 168 mln dlrs.

First Wisconsin , which had assets at yearend of 7.1 billion dlrs, said the Shelard purchase price is about 12 times the 1986 earnings of the bank.

It said the two Shelard banks have a total of five offices in the Minneapolis-St. Paul area.

Reuter





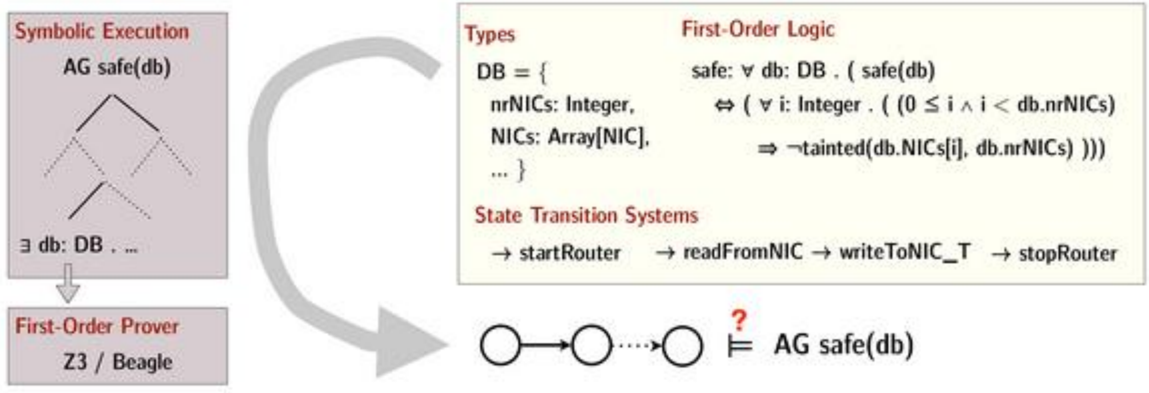
# Challenge of integration of data and algorithm models for semantization

FO (predicate) DyL

x := t only atomic

program

Higher level coding

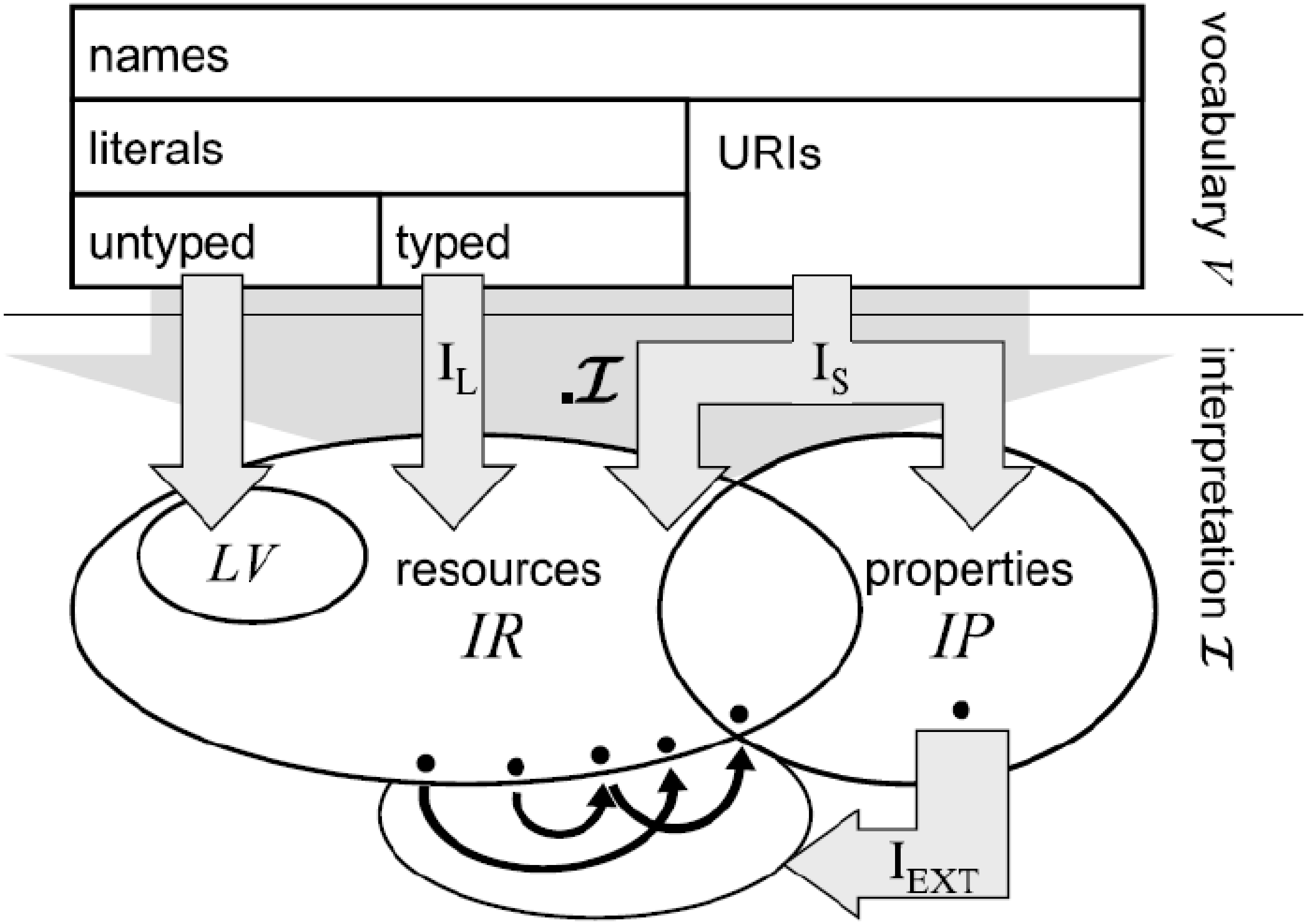


used for dynamic systems verification, model checking

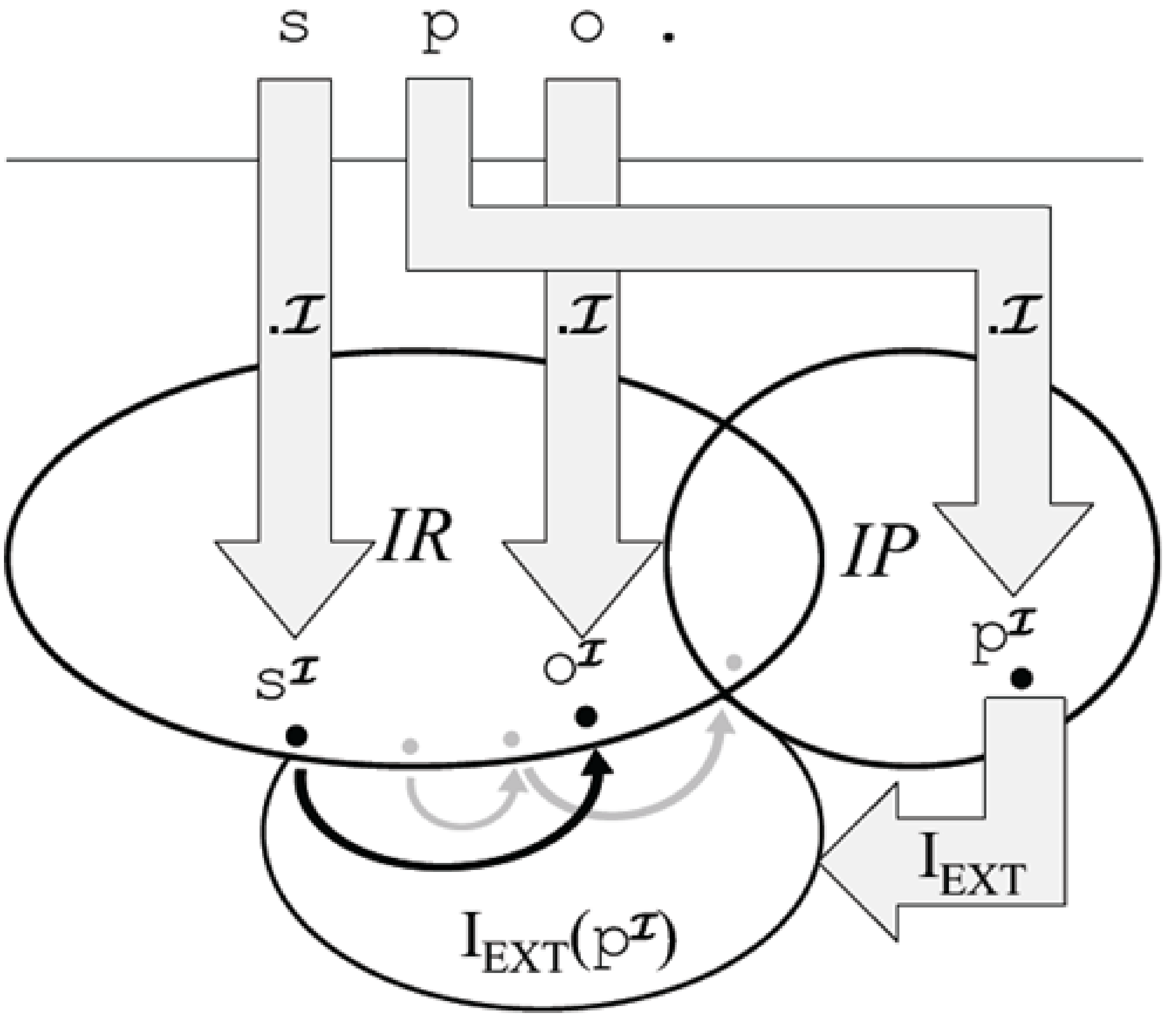
We need to keep:

- Relational data (binarized in RDF), DOM, BoW, XML…
- Propositional algorithms (constructed by induction) rather than code the P/R quality on data counts

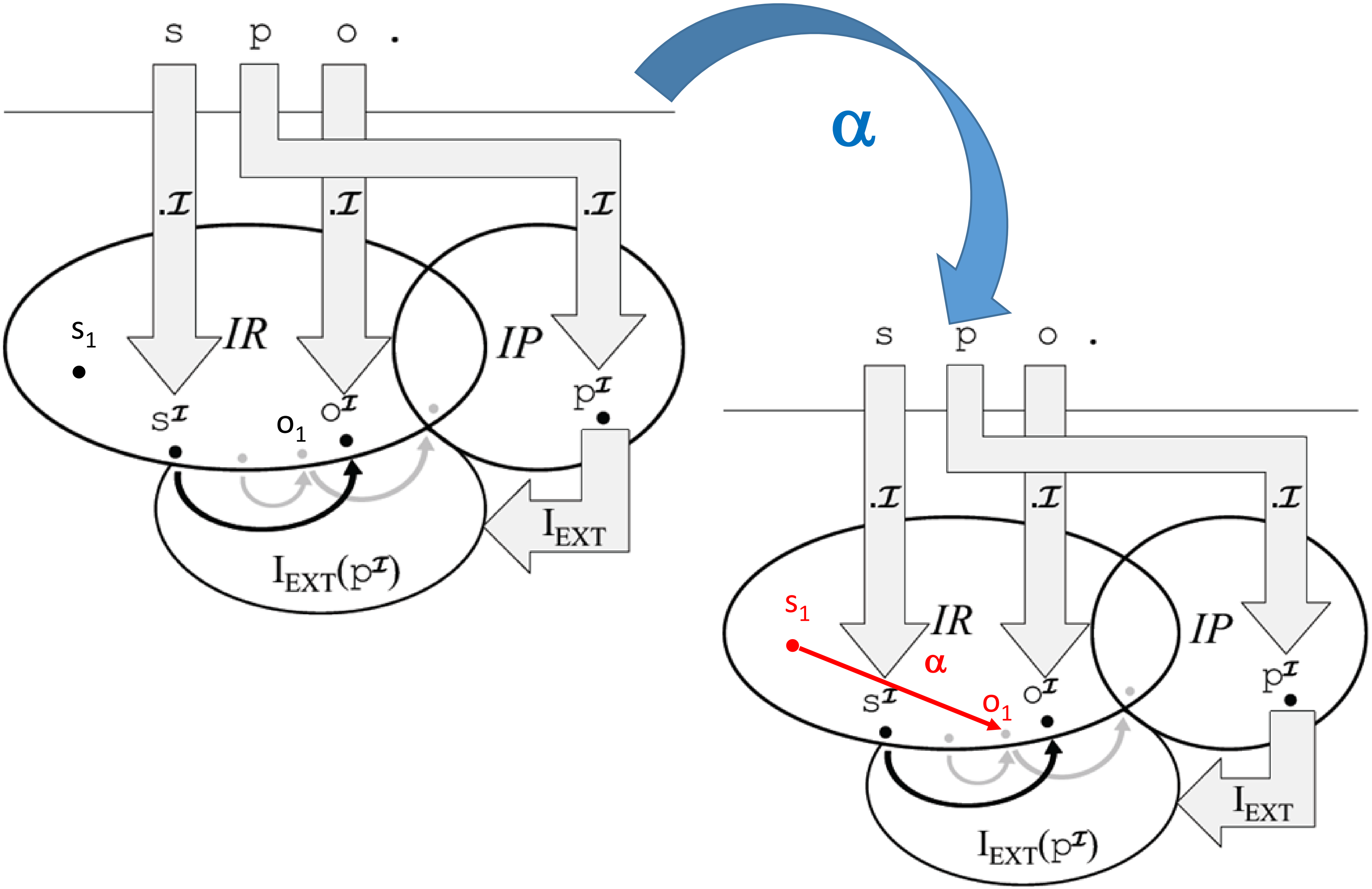
# Simple RDF structure



# Simple RDF structure a triple is "true"



# Simple RDF structure – dynamically



# Named-entity recognition

Most research on NER systems has been structured as taking an unannotated block of text, such as this one:

Jim bought 300 shares of Acme Corp. in 2006.

And producing an annotated block of text that highlights the names of entities:

[Jim]$_{\text{Person}}$ bought 300 shares of [Acme Corp.]$_{\text{Organization}}$ in [2006]$_{\text{Time}}$.

Full named-entity recognition is often broken down, conceptually and possibly also in implementations, as two distinct problems:
- detection of names, and
- classification of the names by the **type** of entity they refer to (e.g. person, organization, location and other).

# Named-entity/resource recognition

Most research on NER systems has been structured as taking an unannotated block of text, such as this one:

[John Fitzgerald Kennedy]https://en.wikipedia.org/wiki/John_F._Kennedy (May 29, 1917 – November 22, 1963), commonly referred to by his initials [JFK]https://en.wikipedia.org/wiki/John_F._Kennedy, was an American politician who served as the 35th [President of the United States] https://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States from January 1961 until his assassination in November 1963.

three distinct problems:

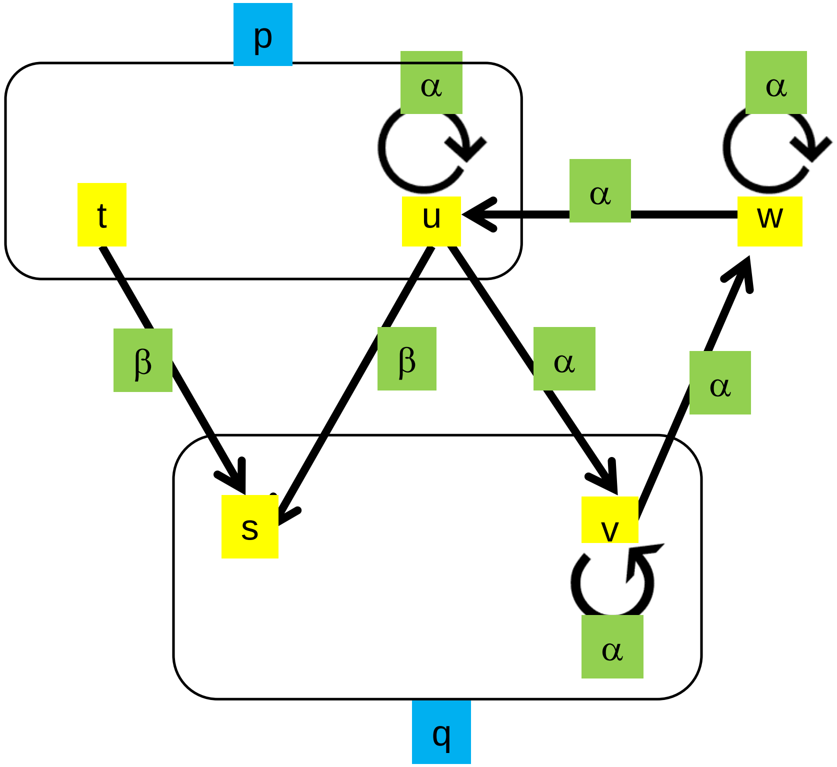
- detection of names, and
- classification of the names by the type of entity they refer to (e.g. person, organization, location and other)
- Detection/recognition/creation of URI.

# Example

We have $\mathfrak{K}$ Kripke structure of
$\Pi_0 = \{\alpha, \beta\}$
$\Phi_0 = \{p, q\}$

As depicted:

K = {u,v,w,s,t}
$m_{\mathfrak{K}}(p) = \{u,t\}$
$m_{\mathfrak{K}}(q) = \{s,v\}$
$m_{\mathfrak{K}}(\alpha) = \{(u,u),(u,v)(v,v),$
$(v,w), (w,w), (w,u)\}$
$m_{\mathfrak{K}}(\beta) = \{(u,s), (t,s)\}$



# Recall: PDyL – Syntax ([HKT, Chapter5, page 164-5])

- PDyL has expressions of two sorts:
  - propositions or formulas: $\Phi_0$ atomic p, q, r, … and $\Phi$ more complex $\varphi, \psi, \ldots$
  - Programs: $\Pi_0$ atomic a, b, c, … and $\Pi$ more complex $\alpha, \beta, \gamma, \ldots$
  - If $\varphi, \psi \in \Phi$, then $\varphi \to \psi \in \Phi$ and $0 \in \Phi$
  - If $\alpha, \beta \in \Pi$, then $\alpha;\beta \in \Pi$, $\alpha \cup \beta \in \Pi$, $\alpha^* \in \Pi$
  - If $\alpha \in \Pi$ and $\varphi \in \Phi$, then $[\alpha]\varphi \in \Phi$
  - If $\varphi \in \Phi$, then $\varphi? \in \Pi$
  - $<\alpha>\varphi \equiv \neg[\alpha] \neg \varphi$
  - skip $\equiv$ 1? And fail $\equiv$ 0?
  - if $\varphi$ then $\alpha$ else $\beta$ $\equiv \varphi?;\alpha \cup \neg \varphi?; \beta$
  - while $\varphi$ do $\alpha$ $\equiv (\varphi?;\alpha)^*;\neg\varphi?$ (repeat $\alpha$ until $\varphi \equiv \alpha$; while $\neg \varphi$ do $\alpha$ $\equiv \alpha;(\neg \varphi?;\alpha)^*; \varphi?$
  - $\{\varphi\}\ \alpha\ \{\psi\} \equiv \varphi \to [\alpha]\ \psi$ (in-conditions, out-conditions)

# Recall: Semantics

Kripke frame is a pair $\mathfrak{K} = (K, m_{\mathfrak{K}})$ , where K is a set of elements u, v, w,… called states and $m_{\mathfrak{K}}$ is meaning function (on atomic extended to whole

$m_{\mathfrak{K}}(p) \subseteq K, \quad \longrightarrow \quad p \in \Phi_0 \qquad m_{\mathfrak{K}}(\varphi) \subseteq K, \; \varphi \in \Phi$

$m_{\mathfrak{K}}(a) \subseteq K \times K, \quad a \in \Pi_0 \quad \longrightarrow \quad m_{\mathfrak{K}}(\alpha) \subseteq K \times K, \quad \alpha \in \Pi$

- $m_{\mathfrak{K}}(\varphi \rightarrow \psi) = (K \setminus m_{\mathfrak{K}}(\varphi)) \cup m_{\mathfrak{K}}(\psi) \quad m_{\mathfrak{K}}(0) = \emptyset \subseteq K$
- $m_{\mathfrak{K}}([\alpha]\varphi) = \{u \in K: (\exists w \in K)\ (\ (u,w) \in m_{\mathfrak{K}}(\alpha) \ \rightarrow\ w \in m_{\mathfrak{K}}(\varphi)\ )\ \}$
- $m_{\mathfrak{K}}(\alpha;\beta) = \{(u,v) \in K^2: (\exists w \in K)\ (\ (u,w) \in m_{\mathfrak{K}}(\alpha)$ and $(w,v) \in m_{\mathfrak{K}}(\beta)\ )\ \}$
- $m_{\mathfrak{K}}(\alpha \cup \beta) = m_{\mathfrak{K}}(\alpha) \cup m_{\mathfrak{K}}(\beta)$
- $m_{\mathfrak{K}}(\alpha^*) = m_{\mathfrak{K}}(\alpha)^* = \cup\{m_{\mathfrak{K}}(\alpha)^n: n \geq 0\} \qquad m_{\mathfrak{K}}(1?) = m_{\mathfrak{K}}(\text{skip})$ = identity relation
- $m_{\mathfrak{K}}(\varphi?) = \{(u,u): u \in m_{\mathfrak{K}}(\varphi)\} \qquad m_{\mathfrak{K}}(0?) = \emptyset \subseteq KxK$

# New proposal W-PDyL: Integration of Web data and PDyL

- Algorithmic part remains propositional, though typed
- Data part needs structure (RDF, FOL, Relational DB, XML, DOM, big data, texts (BoW, sliding window, PoS, morphology, dependency, …)),
- Integrating domain calculus and propositional programs
- W-PDyL has expressions of two sorts (and each sort is/can be typed):
  - Statements about web data: atomic e.g. $\Phi_0^{RDF}$, $\Phi_0^{FOL}$, $\Phi_0^{RDB}$, $\Phi_0^{XML}$, $\Phi_0^{DOM}$, $\Phi_0^{BoW}$, $\Phi_0^{PoS}$, $\Phi_0^{DepTree}$, … and $\Phi$ more complex $\varphi^{RDF}$, $\psi^{FOL}$, … with corresponding data model and metamodel
  - Programs: atomic $\Pi_0^{\sigma}$ for subject extraction, $\Pi_0^{\pi}$ for property extraction, $\Pi_0^{\omega}$ for object value extraction in case of html, xhtml, xml data; $\Pi_0^{ner}$ for named entity extraction in case of text data, … and $\Pi$ more complex $\alpha^{\sigma\pi\omega}$, $\beta^{\sigma\pi\omega}$, $\gamma^{\sigma\pi\omega}$, …
  - Statements are typically accompanied by information about program creation (data mining tool, training data, metric (e.g. precision, recall), …)

# New proposal: Semantics of W-PDyL

W-Kripke frame is a tuple pair $\mathfrak{K} = (K, m_{\mathfrak{K}})$ , where K is a set of elements u, v, w,… called states (possible worlds, web states) and $m_{\mathfrak{K}}$ is meaning function (on atomic statements and programs extended to whole).

Now we have two possibilities, either states

$K = K^{RDF} \cup K^{FOL} \cup K^{RDB} \cup K^{XML} \cup K^{DOM} \cup K^{BoW} \cup K^{PoS} \cup \ldots$

Or each state is a union of corresponding states

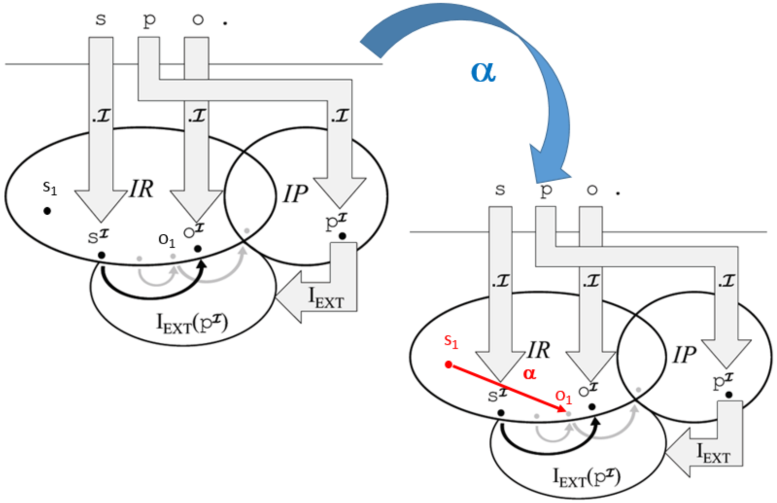
$s = s^{RDF} \cup s^{FOL} \cup s^{RDB} \cup s^{XML} \cup s^{DOM} \cup s^{BoW} \cup s^{PoS} \cup \ldots$

where K is a set of elements u, v, w,… called states and $m_{\mathfrak{K}}$ is meaning function (on atomic extended to whole), e.g.

$m_{\mathfrak{K}}(p) \subseteq K^{RDF}$, if $p \in \Phi_0^{RDF}$,

$m_{\mathfrak{K}}(a) \subseteq K^{html} \times K^{html}$, if $a \in \Pi_0^{\sigma}$

to mention at least one example (for various sorts and types this is generalized



# W-PDyL – how can it help

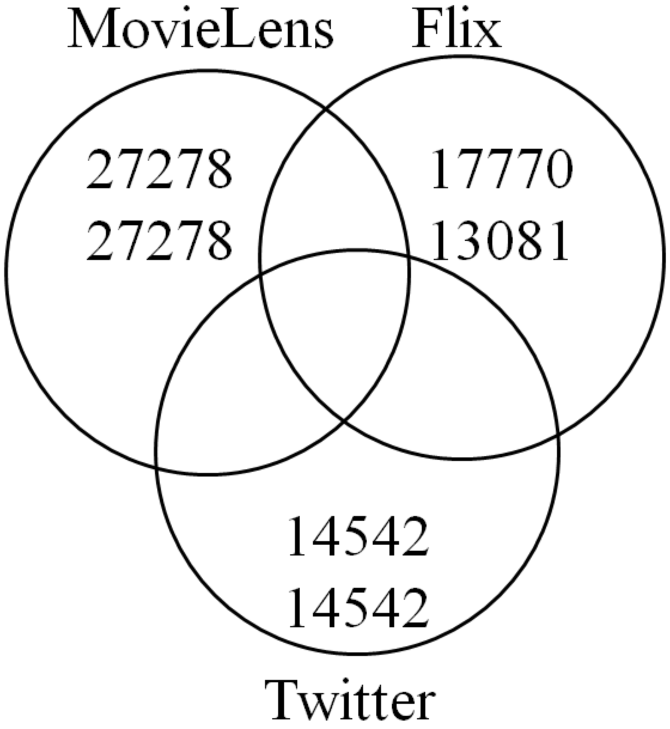
Assume we have a Bag of Words data matrix

| | | Token$_1$ | … | Token$_m$ | Class | Who | computed | |
|---|---|---|---|---|---|---|---|---|
| Train set | Doc1 | 1 | | | 0 | human | | |
| | Doc2 | | | 7 | 1 | human | | |
| | … | | | | … | human | | |
| Test set | Doc$_{n+1}$ | 0 | | | 0 | $\alpha$ | 0 | TN |
| | Doc$_{n+2}$ | | | 1 | 1 | $\alpha$ | 0 | FN |
| | … | | | | … | $\alpha$ | … | |
| | Doc$_{n+m-1}$ | | | | 0 | $\alpha$ | 1 | FP |
| | Doc$_{n+m}$ | | | | 1 | $\alpha$ | 1 | TP |

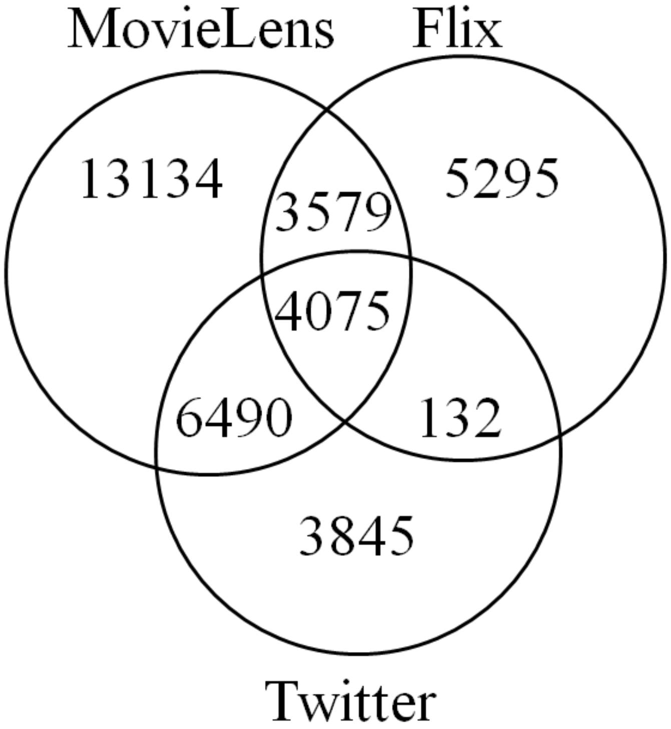
A new document d arrives, can I use $\alpha$ to classify it? One can calculate similarity of d to this collection. The higher the similarity is, the higher will be my confidence in $\alpha$'s output.

# Movie data

**Integrated to IMDB**



**Intersections**



# Proposal

- Our goal is to extend the semantic web foundations to enable describing creation, dynamics and similarities on data.
- To describe the reliability of the obtained RDF data we propose a "half-a-way" extension of dynamic logic.
- Programs (extractors) remain propositional,
- Kripke states correspond to web pages, and there is a lot of reification describing the training and testing data and the metrics of learning. We call this here Dynamic Logic RDF (DLRDF).

# Dynamic Logic RDF (DLRDF)

- The language of DLRDF has expressions of two sorts: propositions or formulas $\varphi$, $\psi$, … and programs $\alpha$, $\beta$, … Atomic programs are denoted a, b, c, … and the set of all atomic programs is denoted $\Pi_0$ . Atomic propositions are denoted p, q, r, … and the set of all atomic propositions is denoted $\Phi_0$. The set of all programs is denoted $\Pi$, and the set of all propositions is denoted $\Phi$.
- Traditional tasks of DyL
- Our task here is different