Semantic Matching: KG Integration and Construction

Part III: Instance Matching

Instance Matching (also called Entity Co-resolution)

Source Dataset

About: Berlino
An Entity of Type : municipality,
Data Space : dbpedia.org

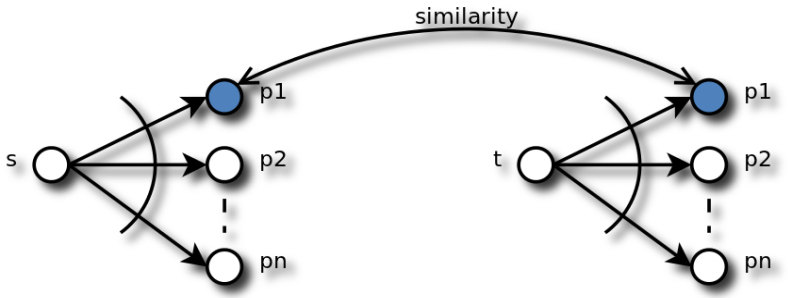
DBpedia

Target Dataset

Berlin
Location, City/Town/Village, Listed Site, Filming location, Adm
bidding city, Travel destination, Olympic host city, Sports Tean
Helynevek, Literature Subject, Place with neighborhoods, Orga
Architectural structure owner, Ontology Instance, Government
Administrative Area, German city, Newspaper circulation area

Berlin (/bɜrˈlɪn/; German pronunciatio
states of Germany. With a population
second most populous city proper an
Located in northeastern Germany, it i
has 4.4 million residents from over 19

Link Discovery



$$\sigma(s,t) = \sigma_1(s.p_1, t.p_1) \cdot w_1 + \ldots + \sigma_n(s.p_n, t.p_n) \cdot w_n$$

Given:

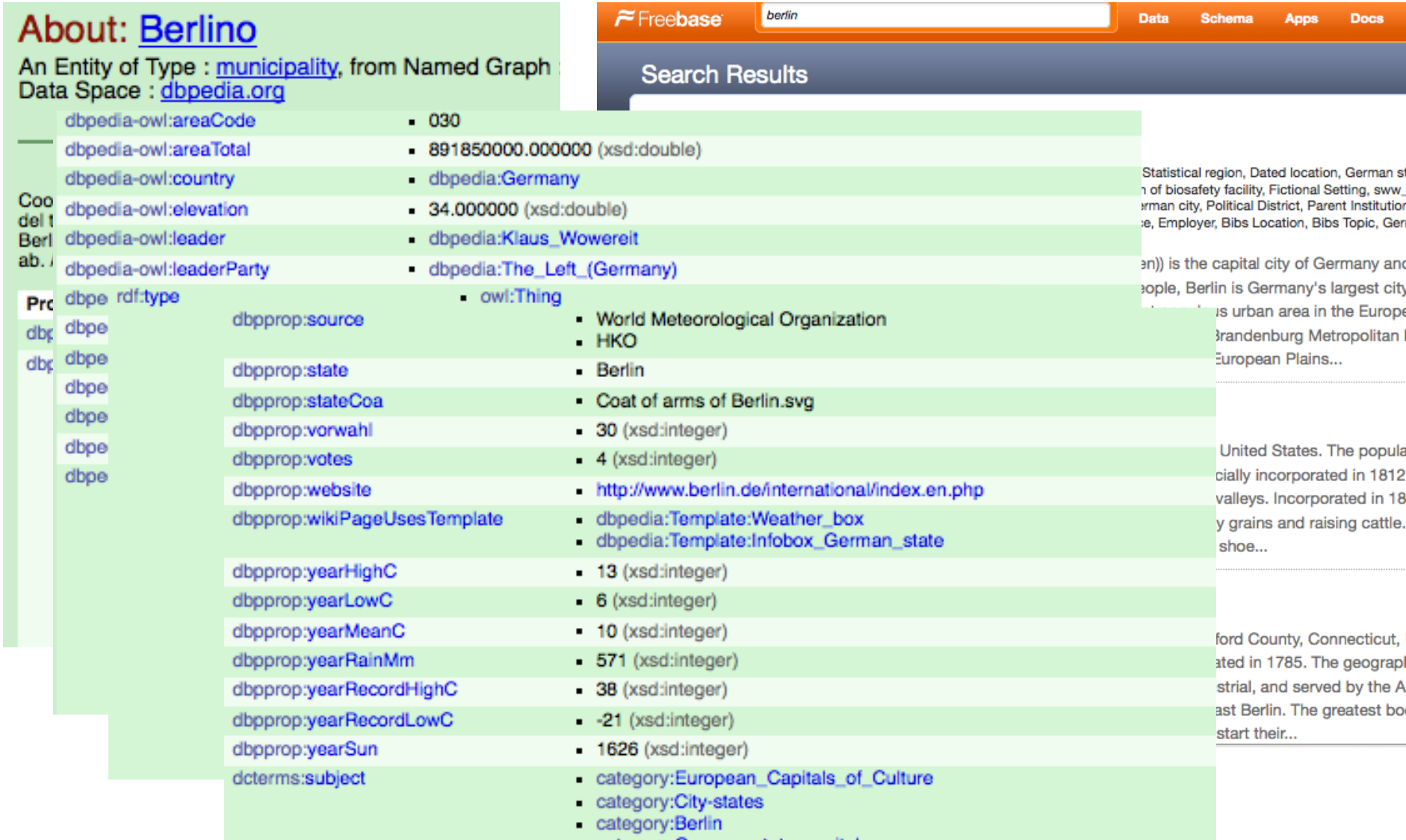
- A similarity measure σ_i for each property p_i
- A threshold θ

Create the link (s,t) if $\sigma(s,t) \geq \theta$

Instance Matching (also called Entity Co-resolution)

Source Dataset

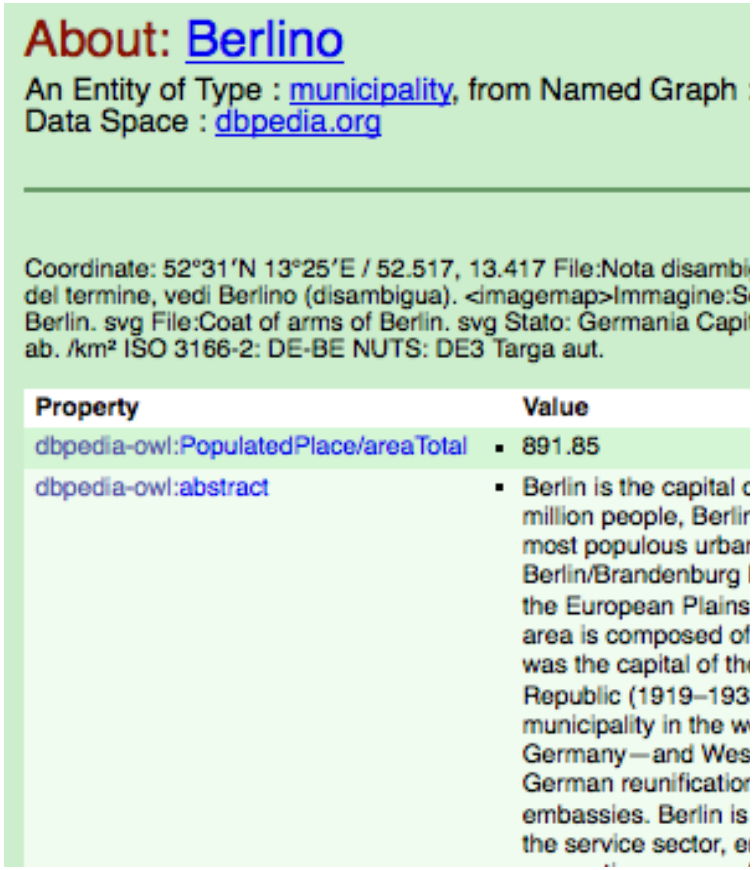
Target Dataset

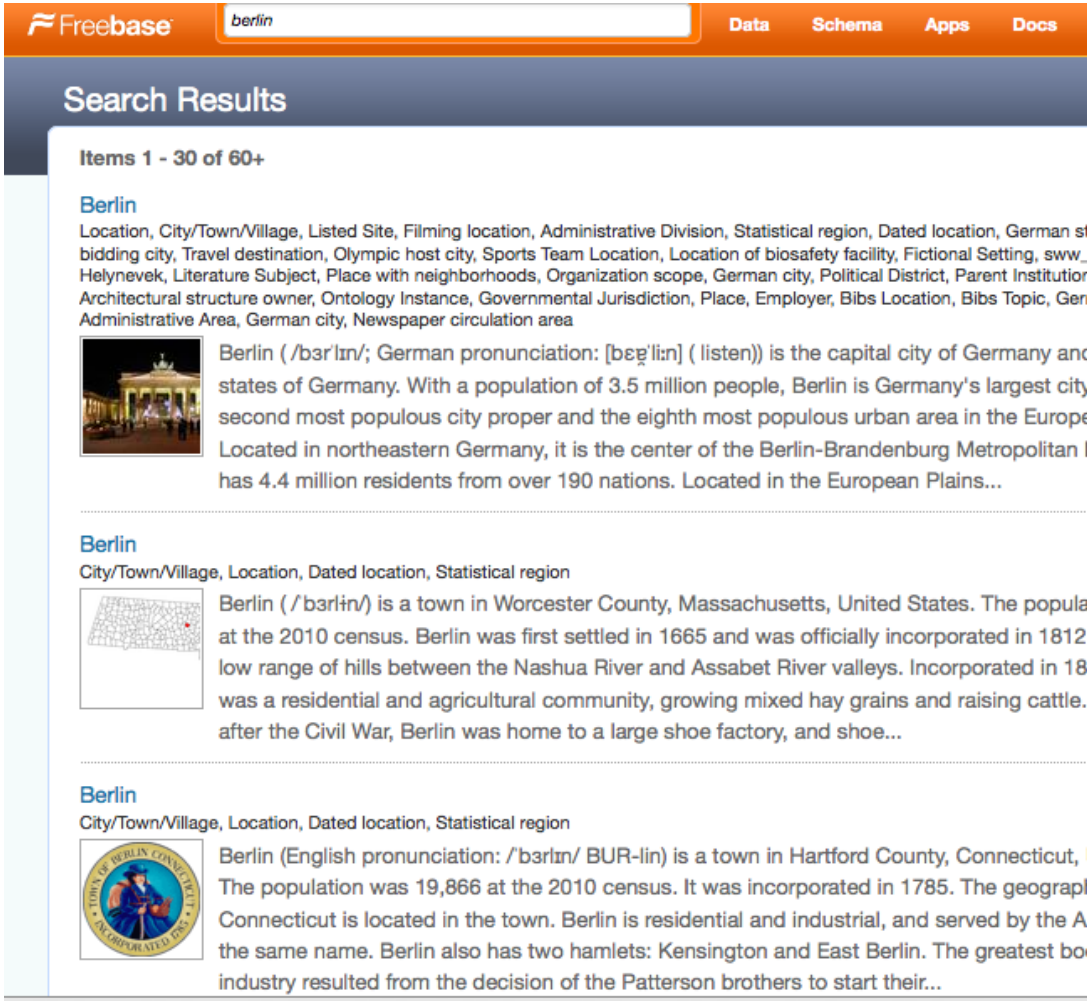


Instance Matching (also called Entity Co-resolution)

Source Dataset

Target Dataset

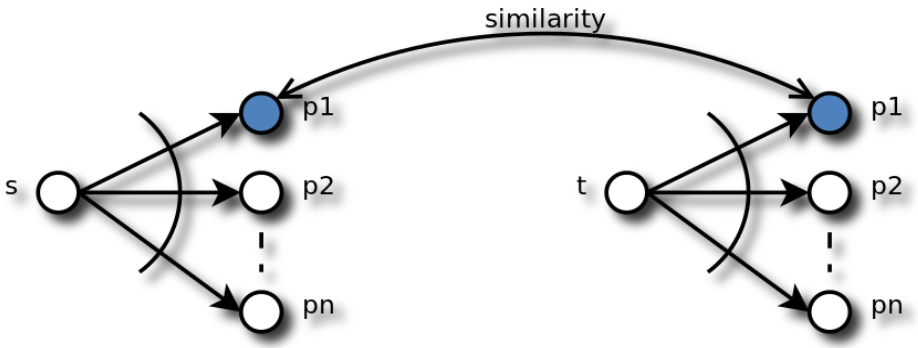




Instance Matching (also called Entity Co-resolution)

Scalability problems: *blocking techniques*

Link Discovery



$$\sigma(s,t) = \sigma_1(s.p_1, t.p_1) \cdot w_1 + \ldots + \sigma_n(s.p_n, t.p_n) \cdot w_n$$

Given:

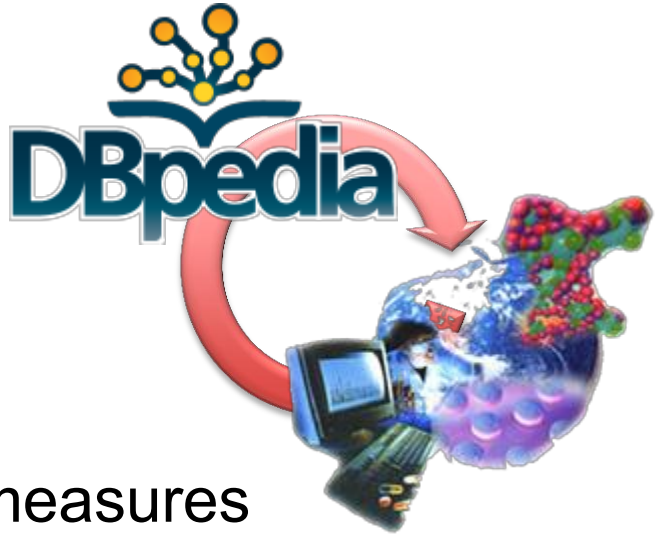
- A similarity measure σ_i for each property p_i
- A threshold θ

Create the link (s,t) if $\sigma(s,t) \geq \theta$

Learning weights or more *complex specifications*

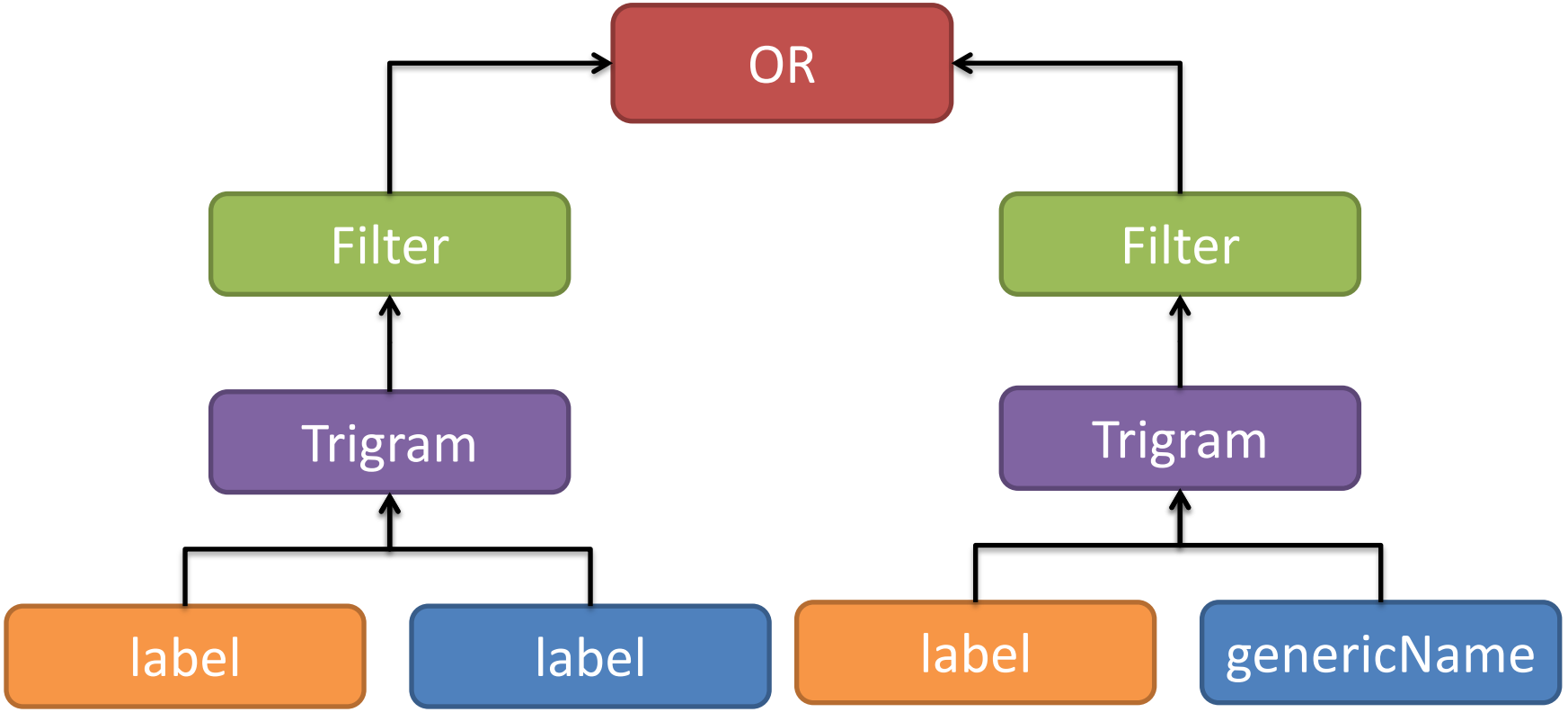
LIMES Native Interface

- **Task**: Link drugs across knowledge bases
  - Source: DBpedia
  - Target: Drugbank
  - Features
    - Definition of complex measures



LIMES Native Interface

- **OR(trigram(x.rdfs:label, y.drugbank:genericName)|0.8, trigram(x.rdfs:label, y.rdfs:label)|0.8)**



Semantic Matching: KG Integration and Construction

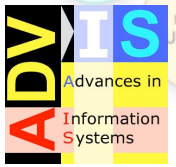
Part IV: Ontology Matching

→ See Isabel Cruz's slides on AgreementMaker

+ next slides

Ontology Matching for Linking LOD ontologies

Matteo Palmonari



Department of Computer Science, Systems and Communication,
University of Milan-Bicocca

Visiting at ADVIS Lab, UIC, Sept 2010 – Gen 2011

AgreementMaker for Aligning Linked Open Data Ontologies

- Aligning LOD ontologies ≠ aligning OAEI ontologies [Jain *et al.* ISWC 2010]
  - *Towards "On the Go" Matching Techniques for Information Matching [LDH 2011]**
    - Capability to accurately align LOD ontologies in a timely fashion
  - *Building linked ontologies with high precision using subclass mapping discovery [Artif. Intell. Rev. 40(2): 127-145 (2013)]*
    - *Improved algorithms and results*

*joint work with Isabel Cruz, Federico Caimi, Cosmin Stroe

- Research questions for LOD
  - New semantic relationships
    - e.g., subClassOf
  - New algorithms
    - e.g., use of imported concepts in the ontologies
  - Scalability
    - e.g. lightweight disambiguation techniques

Semantic Matching: KG Integration and Construction

Part V: More Matching Problems

Bridging the language gap

User Interaction

Table Annotation

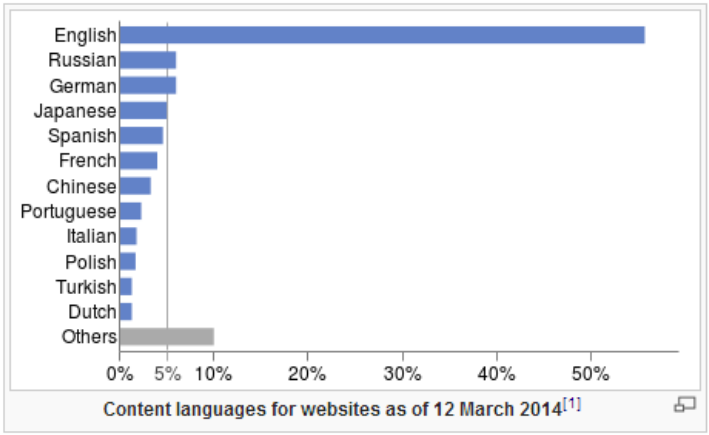
Geospatial matching

Cross-lingual Ontology Matching

A Web of Multilingual Data

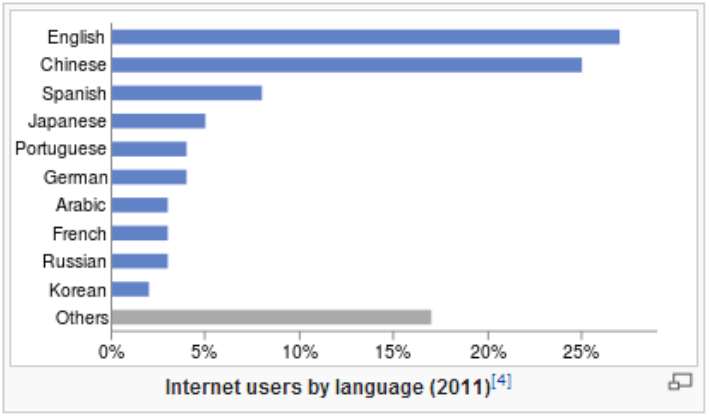
Multilingual Web, why is it challenging?

- **"Textual" data are growing fast**[2]
  - 14+ Trillion Webpages
  - 672*10^9 GB of accessible data
  - 166 different languages



Content languages for websites as of 12 March 2014[1]

- **Data expressed in a certain language is not easily accessible to speakers of other languages**
  - 1M+ Open Government Datasets in 24 languages (As of march 2015[3])



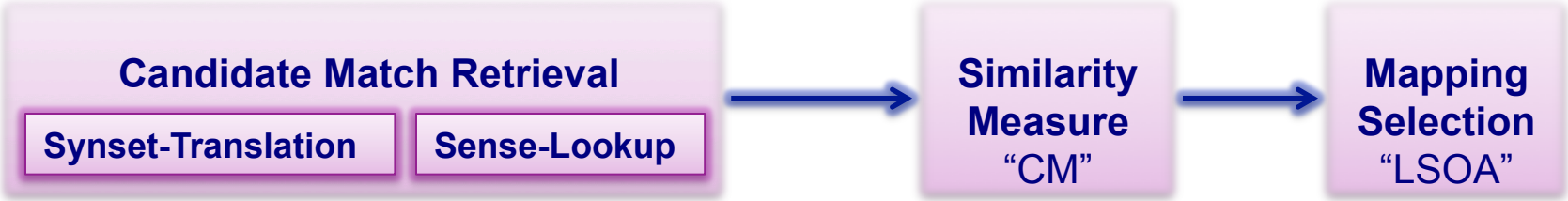
Internet users by language (2011)[4]

What makes Multilingual Web of data true?

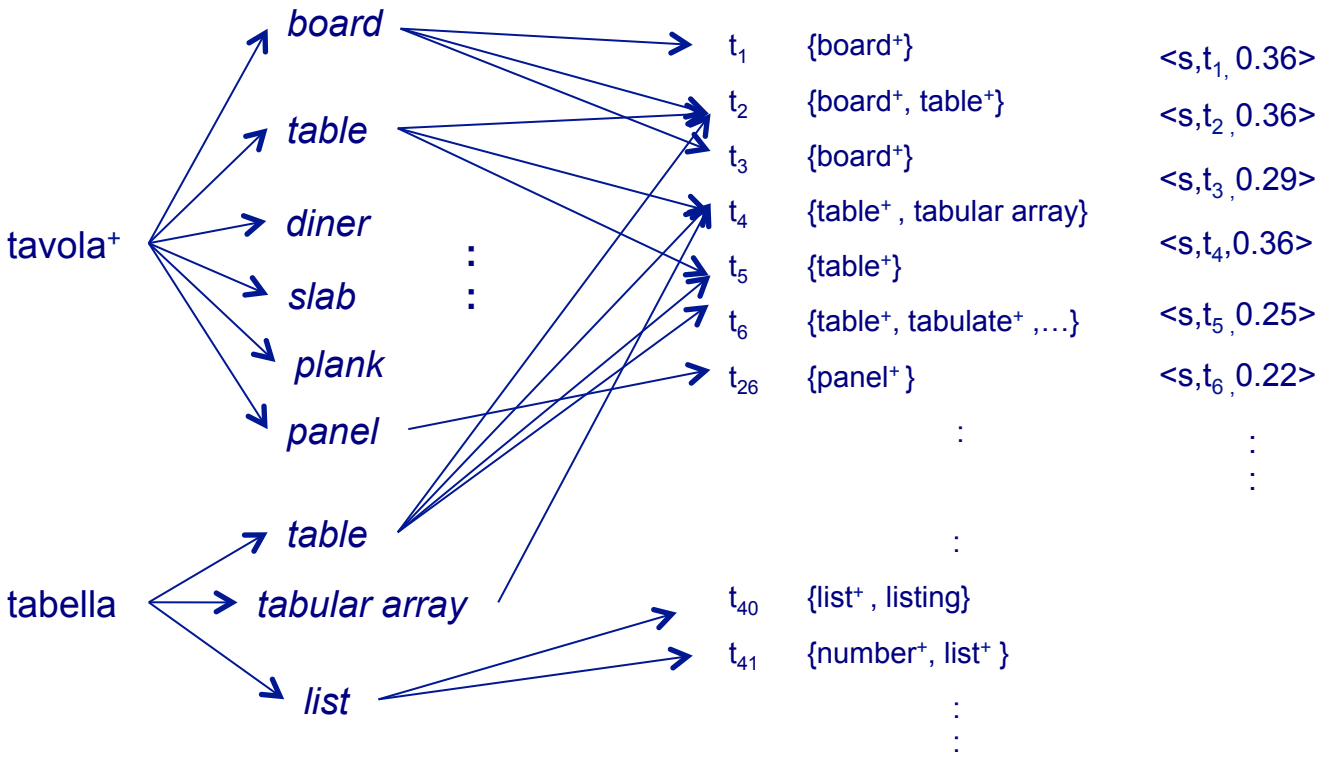
- **Cross-Lingual Mapping** [Gracia et al, 2012]
  - Multilingual lexically-rich resources
    - For poor-resource languages
    - Inter-lingual links between resources

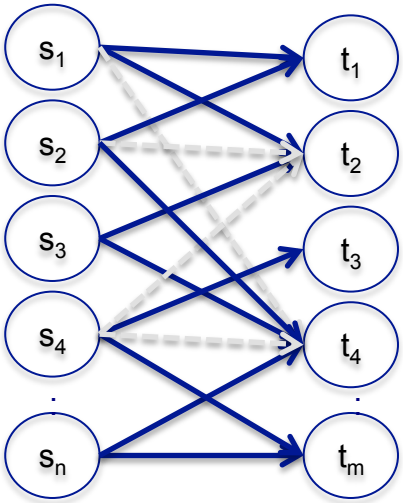
[1][4] http://w3techs.com/technologies/overview/content_language/all
[2] http://www.factshunt.com/2014/01/total-number-of-websites-size-of.html
[3] http://logd.tw.rpi.edu/page/international_dataset_catalog_search

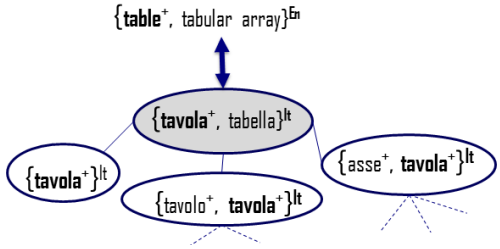
CLM Method : Overview



s = {tavola⁺ , tabella}







⁺ : Polysemous word

Candidate Matches Retrieval : Observations

- Upper bound for cross-lingual concept mapping with external translation resources
  Abu Helou & Palmonari, In the 20th NLDB –poster (2015)
- Effectiveness of Automatic Translations for Cross-lingual Ontology Mapping
  Abu Helou, Palmonari & Jarrar. In JAIR Journal (2015)

- **14 Observations, …**

Main results of the study suggest that:

- **Monosemous Heuristic** strategy **fails** for a significant number of mappings (**~20**%)

  is mapped to {**flat**}, "flat" has 9 senses in English }شقّة{ •

  - {**forchetta**} is mapped to {**fork**}, "fork" has 24 senses in English

- **Polysemous but synonymless** synsets are **harder** to filter out

  }is mapped to {**thing** }شيء{ •

)has 6 senses (5 are synonymless "شيء" •

  - "thing" has 12 senses (11 are synonymless**)**

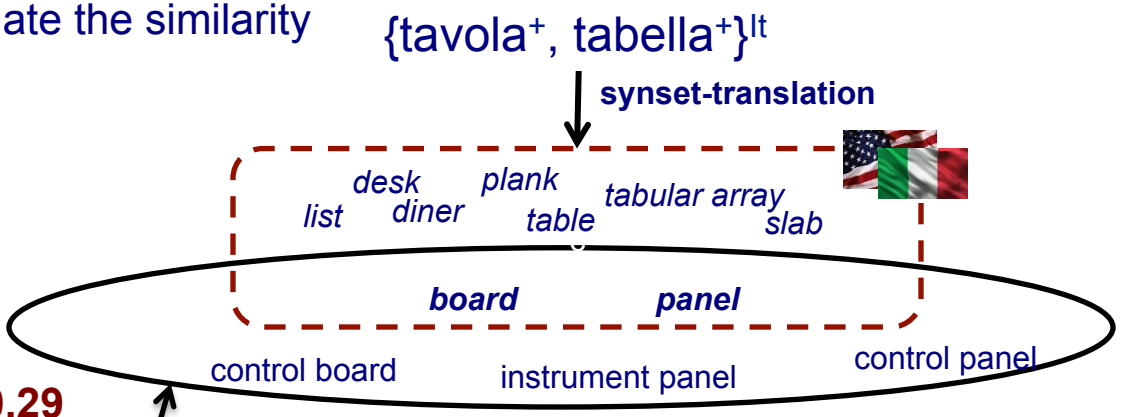
  - {**cosa**} is mapped to {**thing**}

    - "cosa" has 13 senses (5 are synonymless)

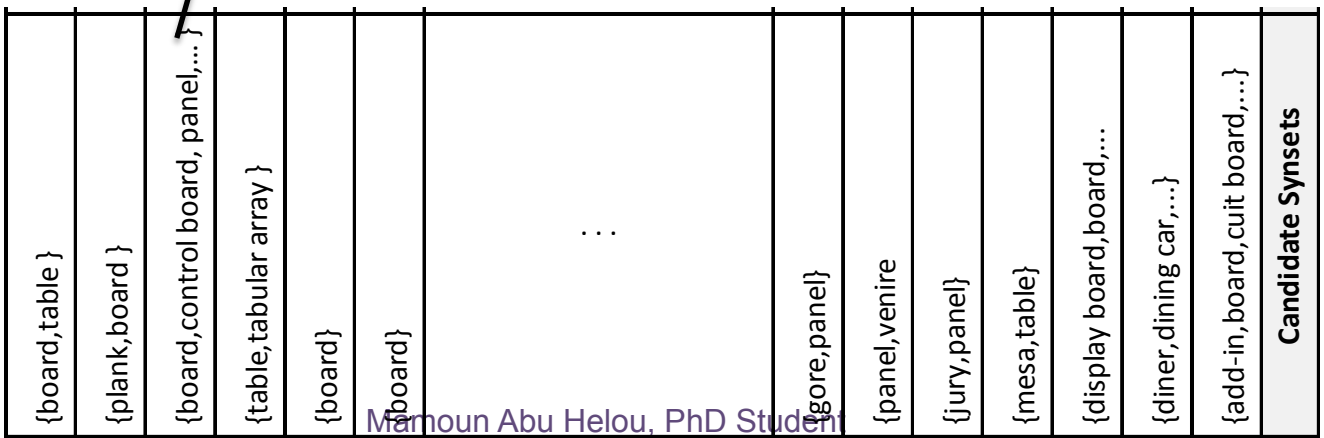
CLM Similarity with Multi-Word Translation (CM)

Cross-Lingual Lexical Matching with Word Translation and Local Similarity Optimization
Abu Helou and Palmonari. In the 11th SEMANTiCS (2015)

Inspired by the **classification-based interpretation of mappings**' semantics
Use **Recall** and **Precision** of the translation of the source concepts over the target concept to evaluate the similarity



***Recall* = 2/5 ;**
***Precision*= 2/9**
➔ ***F_1-measure*=0.29**



Mapping Selection with Local Similarity Optimization Algorithm (LSOA)

Cross-Lingual Lexical Matching with Word Translation and Local Similarity Optimization
Abu Helou and Palmonari. In the 11th SEMANTiCS (2015)

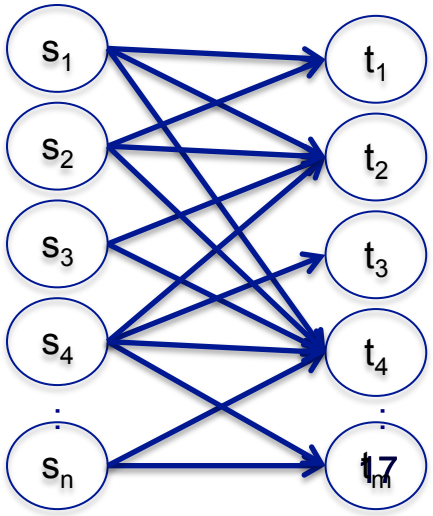
View mapping selection as an instance of the **Assignment Problem** [Cruz et al, 2009]

- A mapping can be **influenced by other** mappings that are relevant for the decision
- May solve **ties** in top-ranked mapping sets
- Maximize the similarity of the final alignment
  - Optimal, e.g., the Hungarian Method [Kuhn, 1955]
    - Computationally expensive

    execution time: , memory size:
  - Efficient and suboptimal, e.g., a bipartite graph [Cruz et at, 2009]
    - Still difficult to scale to very large mapping problems (e.g., 100K+)

Tie

1	1	0.4	1	1	1
0.22	0.22	0.22	0.22	0.11	0.11
0.36	**0.36**	**0.29**	**0.36**	**0.20**	**0.20**
{board,table }	{plank,board }	{board,control board, panel,…}	{table,tabular array }	{board}	{board}



Mapping Selection with Local Similarity Optimization Algorithm (LSOA)

Cross-Lingual Lexical Matching with Word Translation and Local Similarity Optimization
Abu Helou and Palmonari. In the 11th SEMANTiCS (2015)

View mapping selection as an instance of the **Assignment Problem** [Cruz et al, 2009]

- A mapping can be **influenced by other** mappings that are relevant for the decision
- May solve **ties** in top-ranked mapping sets
- Maximize the similarity of the final alignment
  - Optimal, e.g., the Hungarian Method [Kuhn, 1955]
    - Computationally expensive

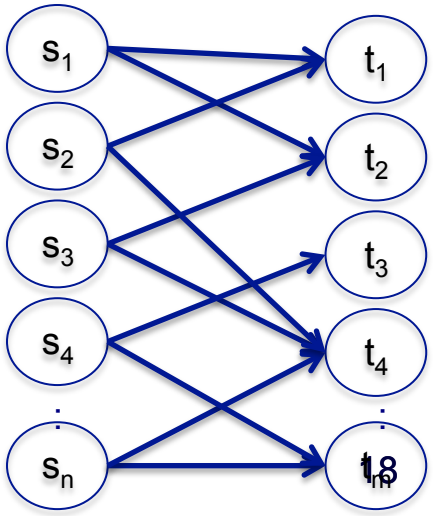
    execution time: , memory size:
  - Efficient and suboptimal, e.g., a bipartite graph [Cruz et at, 2009]
    - Still difficult to scale to very large mapping problems (e.g., 100K+)

Tie

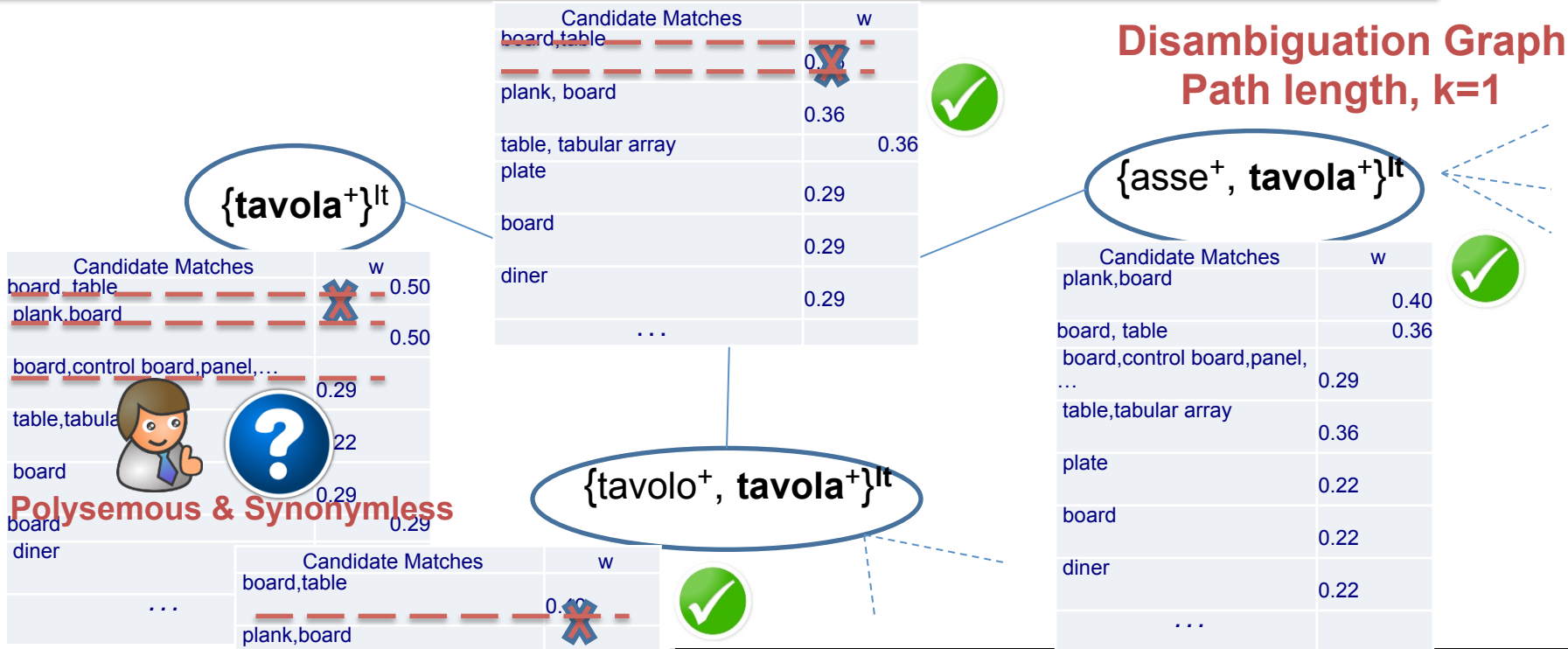
1	1	0.4	1	1	1
0.22	0.22	0.22	0.22	0.11	0.11
0.36	**0.36**	**0.29**	**0.36**	**0.20**	**0.20**
{board,table }	{plank,board }	{board,control board, panel,…}	{table,tabular array }	{board}	{board}

Local Similarity Optimization Algorithm (LSOA)

- **A lexical-based disambiguation graph**
- **A locally optimal solution** (running the Hungarian method)



Mapping Selection with Local Similarity Optimization Algorithm (LSOA)



Candidate Synsets

Source Synsets		{board,table }	{plank,board }	{board,control board, panel,... }	{table,tabular array }	{desktop}	{board}	{board}	{diner}	{panel}	{panel}	{panel}	{plank}	{plate }	{table}	{table}	{table}	{slab}	axis,bloc	{axis,axis vertebra}	{axis,axis of rotation}	{board,gameboard}	{dialog box,panel }	{dining table,board}	{gore,panel}	{panel,venire	{jury,panel}	{mesa,table}	{display board,board,...	{diner,dining car,...}	{add-in,board,cuit board,...}
	{tavola+, tabella}	0.36	0.36	0.29	**0.36**	0	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0	0	0	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.17	0.15	0.13
	{asse+, tavola+}	0.36	**0.40**	0.29	0.36	0	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.17	0.15	0.13
	{tavolo+, tavola+}	**0.40**	0.40	0.31	0.20	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0	0	0	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.22	0.20	0.17
	{tavola+}	0.50	0.50	0.22	0.25	0	**0.29**	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0	0	0	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.22	0.20

Weight Matrix
Hungarian Method

Experiment 1: Quality of the Alignment

Cross-Lingual Lexical Matching with Word Translation and Local Similarity Optimization
Abu Helou and Palmonari. In the 11th SEMANTiCS (2015)

- Evaluate the performance of the matcher in selecting the correct match
  - **CM** : top-1 selection over the cross-lingual similarity measure
  - **CM+LSOA$_{L=k}$**: LSOA selection using CM similarity (with path length = k)

- **Baselines**
  - Monosemous Words Heuristic (**MWH**)
  - Majority Voting (**MV**)
    - Every sense returned by *Sense-Lookup* receive a vote

- REMARK: If there is a **tie**, the mapping is considered **undecidable (thus incorrect)**

Matcher performance w.r.t gold standards

Lang.	Synsets	MWH R	MWH P	MWH F1	MV R	MV P	MV F1	CM R	CM P	CM F1	CM+LSOA$_{p=1}$ R	CM+LSOA$_{p=1}$ P	CM+LSOA$_{p=1}$ F1	CM+LSOA$_{p=2}$ R	CM+LSOA$_{p=2}$ P	CM+LSOA$_{p=2}$ F1
w.r.t. gold standards																
Ar	Monos.	14.0	44.1	21.3	25.4	53.6	34.4	29.2	47.5	36.2	45.2	73.0	55.8	**52.1**	**73.6**	**61.0**
	All	–	–	–	17.7	37.2	24.0	20.5	32.7	25.2	29.9	**69.6**	41.8	**38.7**	66.8	**49.0**
It	Monos.	45.8	**95.4**	60.3	47.5	82.0	60.2	55.2	75.0	63.6	66.4	89.6	76.3	**73.3**	90.0	**80.8**
	All	–	–	–	32.1	63.8	42.7	40.0	58.4	47.5	43.7	**88.3**	58.5	**55.4**	85.3	**67.2**
Slv	Monos.	48.9	89.6	63.3	58.2	91.6	71.2	61.4	83.1	70.6	65.1	**91.9**	76.2	**69.0**	90.5	**78.3**
	All	–	–	–	34.5	73.7	47.0	38.9	62.1	47.9	36.8	**91.1**	52.4	**45.1**	85.6	**59.0**
Es	Monos.	38.1	82.6	52.2	49.4	83.8	62.2	60.3	79.3	68.5	60.8	**89.0**	72.3	**64.6**	88.7	**74.7**
	All	–	–	–	31.9	60.7	41.9	37.9	53.0	44.2	37.8	**87.3**	52.7	**44.2**	83.6	**57.8**

Cross-lingual Instance Matching

Cross-language Linking of eGov Services

Why it is Challenging

≈ sameAs links

Semantic heterogeneity

- not a mere "translation" problem
- cultural bias

- Challenging **cross-language matching** problem
- Most of the approaches:
  - use **structural information** [Spohr et al. 2011,Fu et al. 2011, Wang et al. 2009] or **long textual descriptions** [Knoth et al. 2011]
  - or report problems when automatic **translation** return descriptions **with heterogeneous vocabulary** [Hertling & Paulheim 2012]

Ultra-short descriptions

Explicit Semantic Analysis in CroSeR

Wikipedia-based representation of natural language expressions with the ESA matrix

Wikipedia articles

Terms occurring in Wikipedia articles

ESA	Job Interview	Employment Agency	...	Unemployment benefits
unemployment	0,65	0,84	...	0,92
...	TF-IDF	TF-IDF	TF-IDF	TF-IDF
Term k	TF-IDF	TF-IDF	Tf-IDF	TF-IDF

- A **set of terms** is represented by the centroid of the vectors associated with the individual terms
  - *E.g.: "Unemployment Support"* → *Job Interview* (0,42), *Employment Agency* (0.55), …, *Unemployment Benefits*(0.62)
- Feature generation + light-weight disambiguation

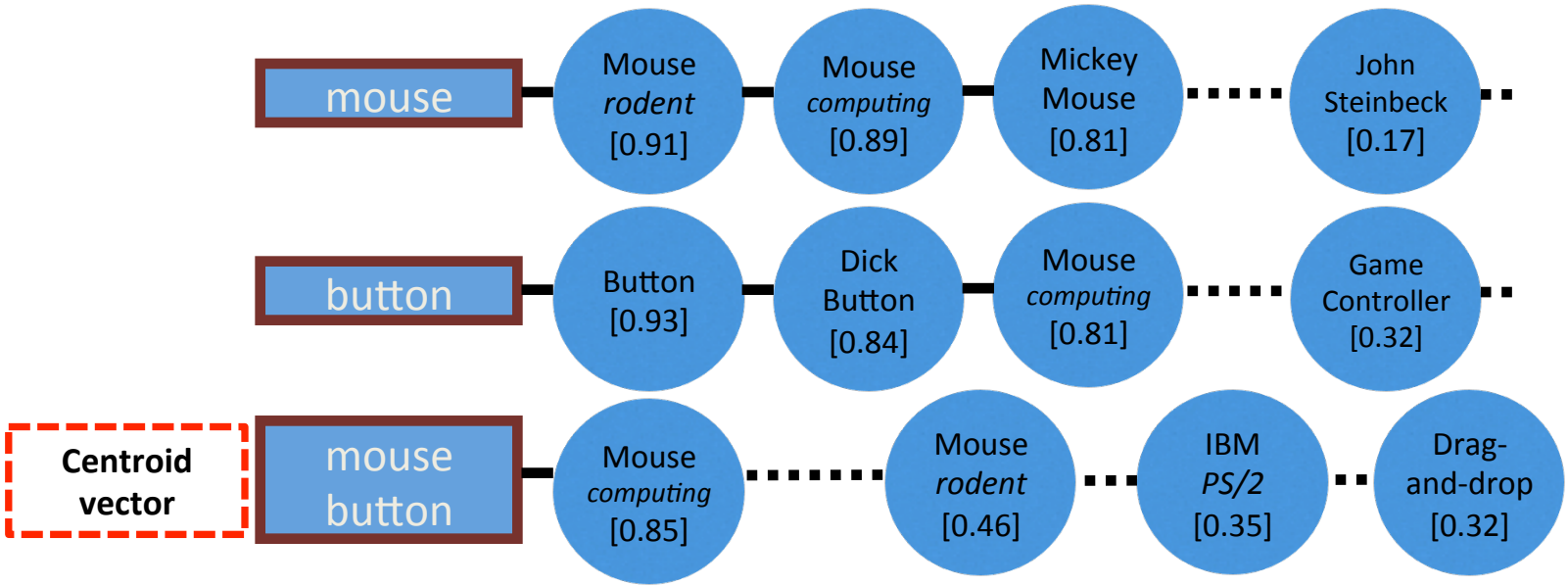
ESA: Wikipedia Concepts and Terms

Every Wikipedia **article** represents a **concept**

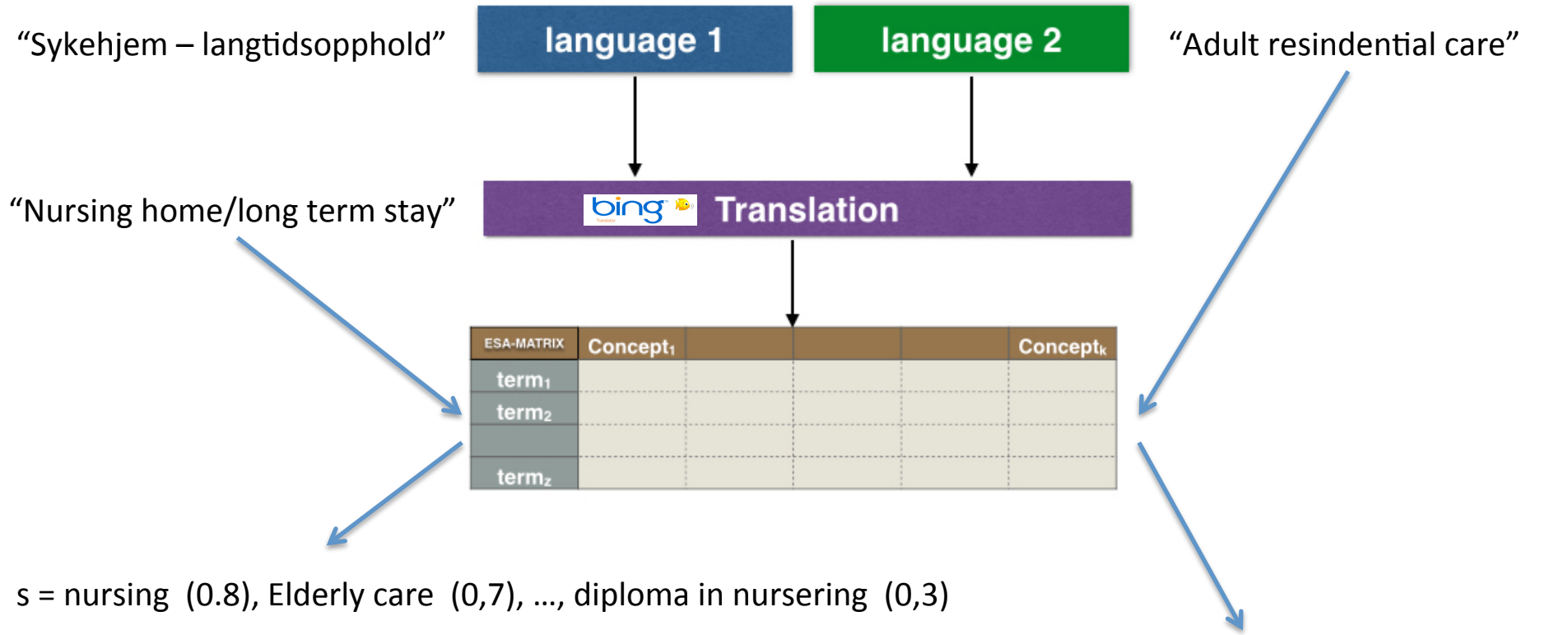
Article **words (terms)** are associated with the **concept** (TF-IDF)

ESA: Text Fragment Interpretation

The **sematic interpretation vector** of a text fragment is the **centroid** vector of the terms occurring in



TR-ESA: Translation-based ESA



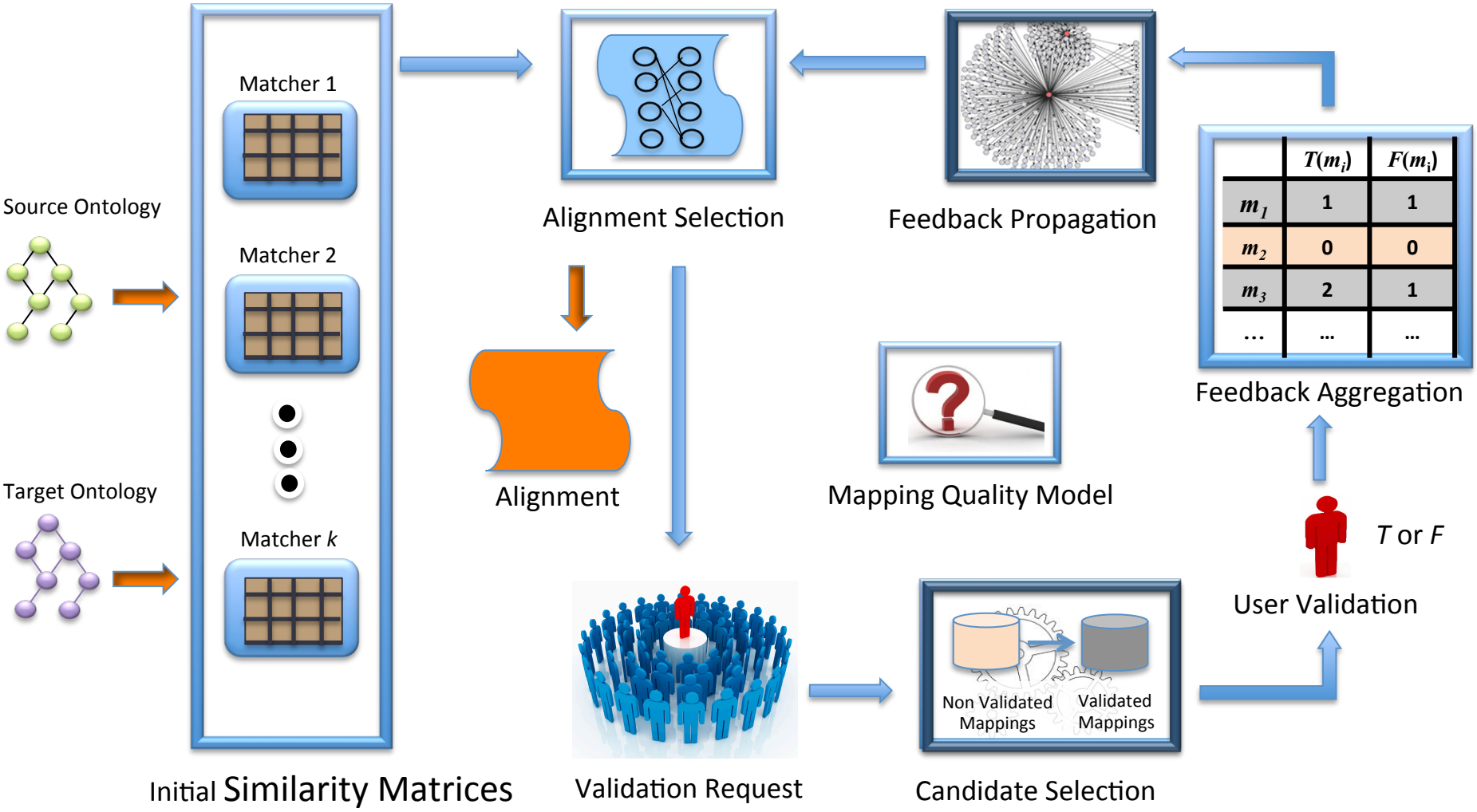
$$cosSim(s, s') = \frac{\sum_{i=1}^{n} s_i * s'_i}{\sqrt{\sum_{i=1}^{n} (s_i)^2} * \sqrt{\sum_{i=1}^{n} (s'_i)^2}}$$

Interactive Matching Methods

Interactive Matching

- Correct errors of automatic matching methods
- Complement automatic matching methods by adding mappings not found by them
- Optimize the automatic matching methods by learning how to match
  - User feedback loop: user feedback → improve the alignment; more feedback → higher improvement
    - How to propagate the user feedback?
- Try to correct/complement/optimize faster
    - On which mappings should we ask the feedback first?

Pay-as-you-go Ontology Matching with Multi-User Feedback



Cruz, Loprete, Palmonari, Stroe, & Taheri (2014). Pay-As-You-Go Multi-user Feedback Model for Ontology Matching, EKAW.

Quality-Based Candidate Selection (text)

- Candidate selection strategies: combine different mapping quality measures and rank mappings in decreasing order of quality

1. Disagreement and Indefiniteness Average (*DIA*)
   - Selects mappings with the most disagreement by the automatic matchers and most indefinite similarity values

Non Validated Mappings

2. Revalidation (*REV*)
   - Selects mappings with the lowest consensus, highest feedback instability and highest conflict with other mappings

Validated Mappings

- Meta-strategy: combine DIA and REV
  - Revalidation Rate (RR) determines the proportion of mappings selected from the two ranked lists for a sequence of validated mappings
  - E.g., for RR = 0.3, every ten iterations three mappings are picked from the REV ranked list

Quality-Based Candidate Selection (figure)

Disagreement and Indefiniteness Average (*DIA*)
Selects mappings with the most disagreement by the automatic matchers and most indefinite similarity values

Revalidation (*REV*)
Selects mappings with the lowest consensus, highest feedback instability and highest conflict with other mappings

Mapping Quality Model

Non Validated Mappings

Validated Mappings

DIA

REV

… M4 M3 M2 M1

… N5 N4 N3 N2 N1

Ranked list of mappings

Meta Strategy

$P_{REV} \in [0,1]$

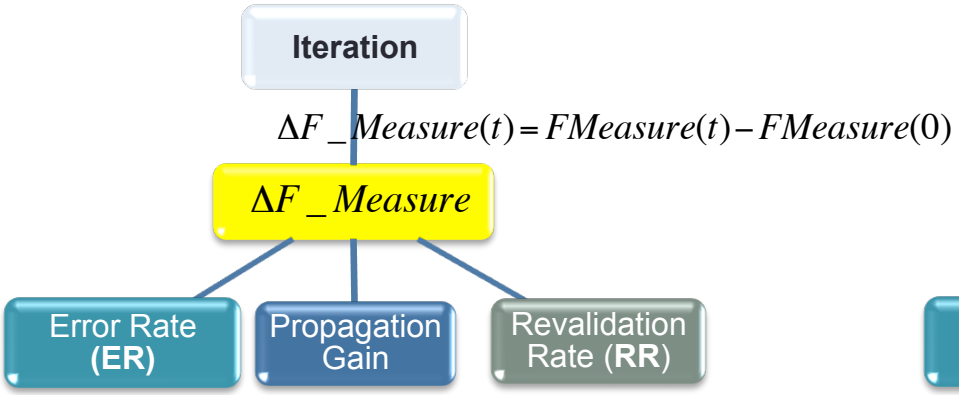
Revalidation Rate (RR)

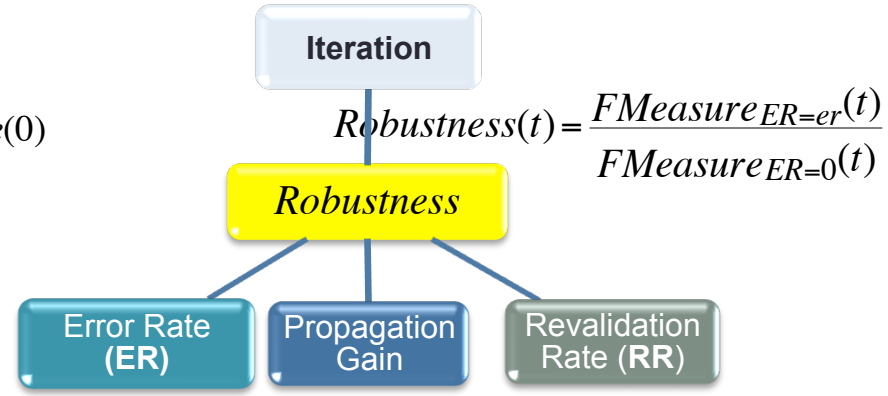
M1

Meta-strategy: combine DIA and REV
Revalidation Rate (RR) determines the proportion of mappings selected from the two ranked lists for a sequence of validated mappings
E.g., for RR = 0.3, every ten iterations three mappings are picked from the REV ranked list

Experimental Setup

- Dataset: benchmark track of OAEI 2010 (101-301, 101-302, **101-303**, 101-304)
  - Comparison with Baseline (ORFL): user feedback is propagated when consensus is reached



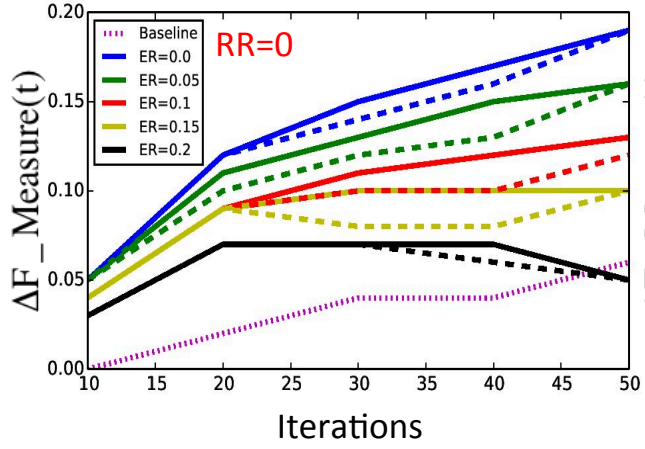


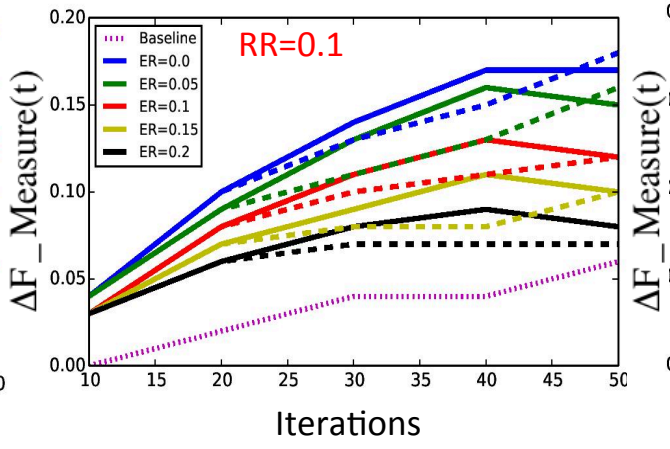
- Simulation of users
  - Error rate (ER): 0.0, 0.05, 0.1, 0.15, 0.2
  - Number of users: 10
- AgreementMaker
  - matchers, alignment selection

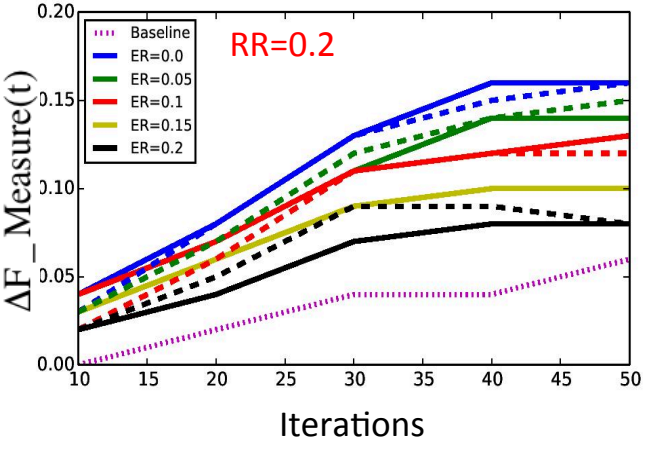
- Propagation gain (g)
  - 0.0 (no gain), 0.5
- Revalidation rate
  - 0.0, 0.1, 0.2, 0.3, 0.4, 0.5

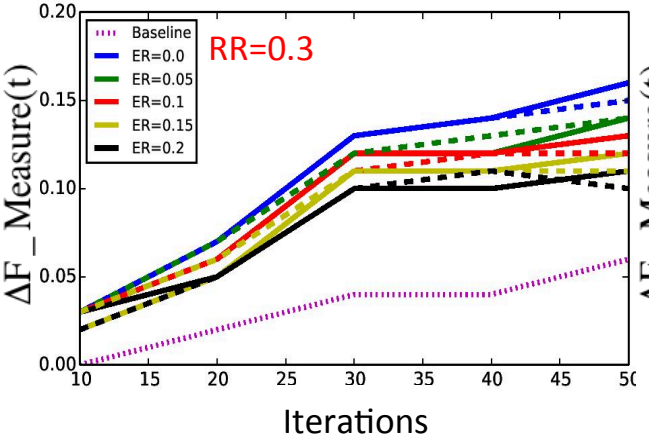
Evaluation

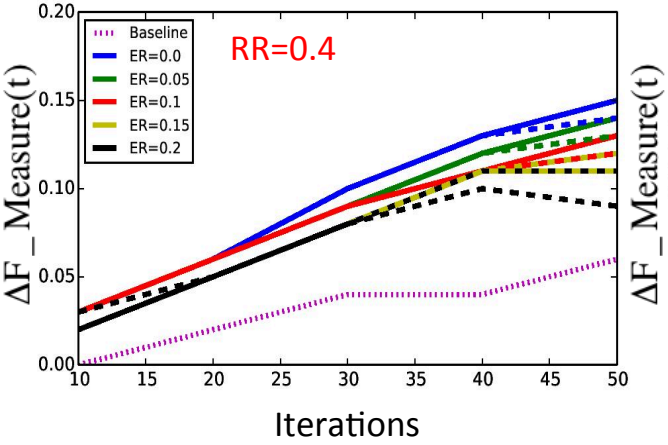
OAEI Benchmark Track 101-303

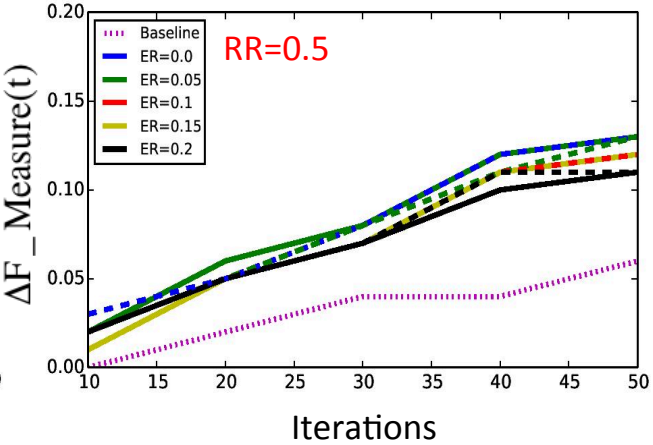










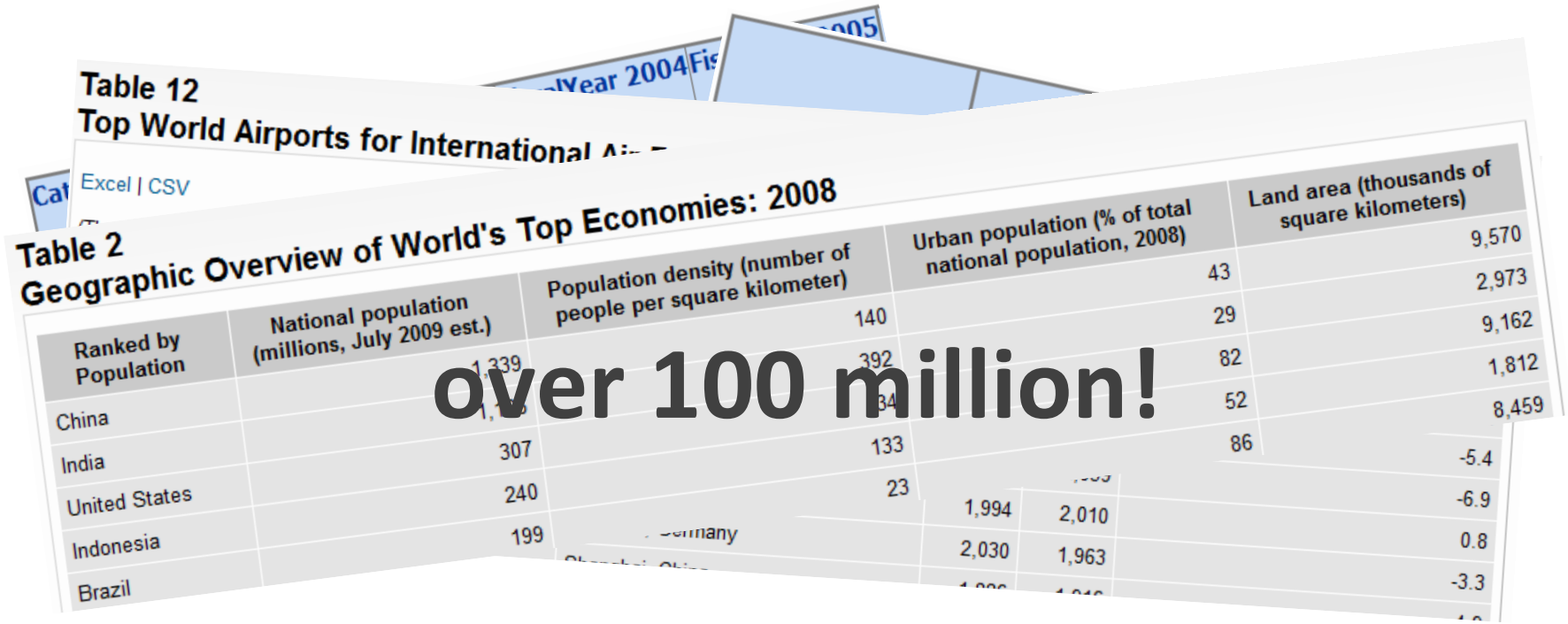


The dashed lines represent a propagation gain equal to zero.
The dotted pink line represents ORFL (Optimally Robust Feedback Loop).

Initial F-Measure=72.73

Semantic Table Annotation

Semi-structured Data Extraction



- **Challenge: Assigning semantics to each table cell**

Table Annotation

N.	Ritratto	Nome (Nascita-morte)	Mandato		Partito	Governo e composizione		Leg.	Presidente della Repubblica
			Inizio	Fine					
		Romano Prodi (1939-)	17 maggio 2006	8 maggio 2008	L'Ulivo; Partito Democratico	Prodi II	L'Unione DS-DL/PD-PRC-RnP (SDI-RI) -PdCI-IdV-FdV UDEUR-SI-DCU-AL-SD-LD-MRE	XV (2006)	
		Silvio Berlusconi (1936-)	8 maggio 2008	16 novembre 2011	Il Popolo della Libertà	Berlusconi IV	Centro-destra PdL-LN-MpA-CN-PT-FdS-DC	XVI (2008)	Giorgio Napolitano  (2006-2015)[4]
		Mario Monti (1943-)	16 novembre 2011	28 aprile 2013	Indipendente	Monti	Governo tecnico con l'appoggio esterno di: PdL-PD-UdC-FLI-ApI		
		Enrico Letta (1966-)	28 aprile 2013	22 febbraio 2014	Partito Democratico	Letta	Grande coalizione PD-PdL/NCD-SC-UdC-PI-RI	XVII (2013)	
		Matteo Renzi (1975-)	22 febbraio 2014	*in carica*	Partito Democratico	Renzi	PD-NCD-SC-UdC-Dem. Solidale-PSI		Sergio Mattarella  (2015-)

Figura 1.1: Un esempio di tabella estratta da Wikipedia relativa ai presidenti del consiglio italiani.

Table Annotation

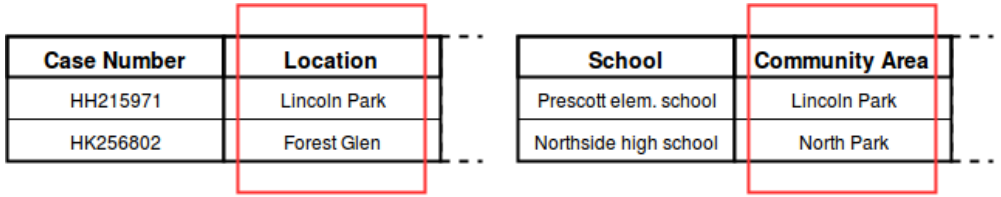
The following table lists the total coffee production of each coffee exporting country in the year 2006[1].

Country	60 kilogram bags	Kilograms	Pounds
Brazil	42,512,000	2,550,720,000	5,611,584,000
Vietnam	15,000,000	900,000,000	1,980,000,000
Colombia	11,600,000	696,000,000	1,531,200,000
Indonesia	6,850,000	411,000,000	904,200,000
Ethiopia	5,500,000	330,000,000	726,000,000
India	5,005,000	300,300,000	660,660,000
Mexico	4,500,000	270,000,000	594,000,000
Guatemala	4,000,000	240,000,000	528,000,000
Peru	3,500,000	210,000,000	462,000,000
Honduras	2,700,000	162,000,000	356,400,000
Uganda	2,500,000	150,000,000	330,000,000
Ivory Coast	2,350,000	141,000,000	310,200,000
Costa Rica	1,808,000	108,480,000	238,656,000

SKU	Marca	Prodotto	Prezzo	Categoria	Disponibilità
90183	Rolex	Rolex Date	2300	Orologeria	7
67127	Omega	Omega Seamaster	1700	Orologeria	10
09972	Zenith	El Primero	3000	Orologeria	5
09972	Pandora	Anello Eternity	349	Gioielleria	20
09972	Swarovski	Christie Oval	79	Gioielleria	20

Tabella 2.1: Esempio di uno dei listini commerciali di 7Pixel.

Table Annotation per Integrazione di Sorgenti Locali in una KB



Case Number	Location
HH215971	Lincoln Park
HK256802	Forest Glen

School	Community Area
Prescott elem. school	Lincoln Park
Northside high school	North Park

Figura 2.2: Esempio di integrazione tra due tabelle del dataset del caso di studio di Chicago

Table Annotation

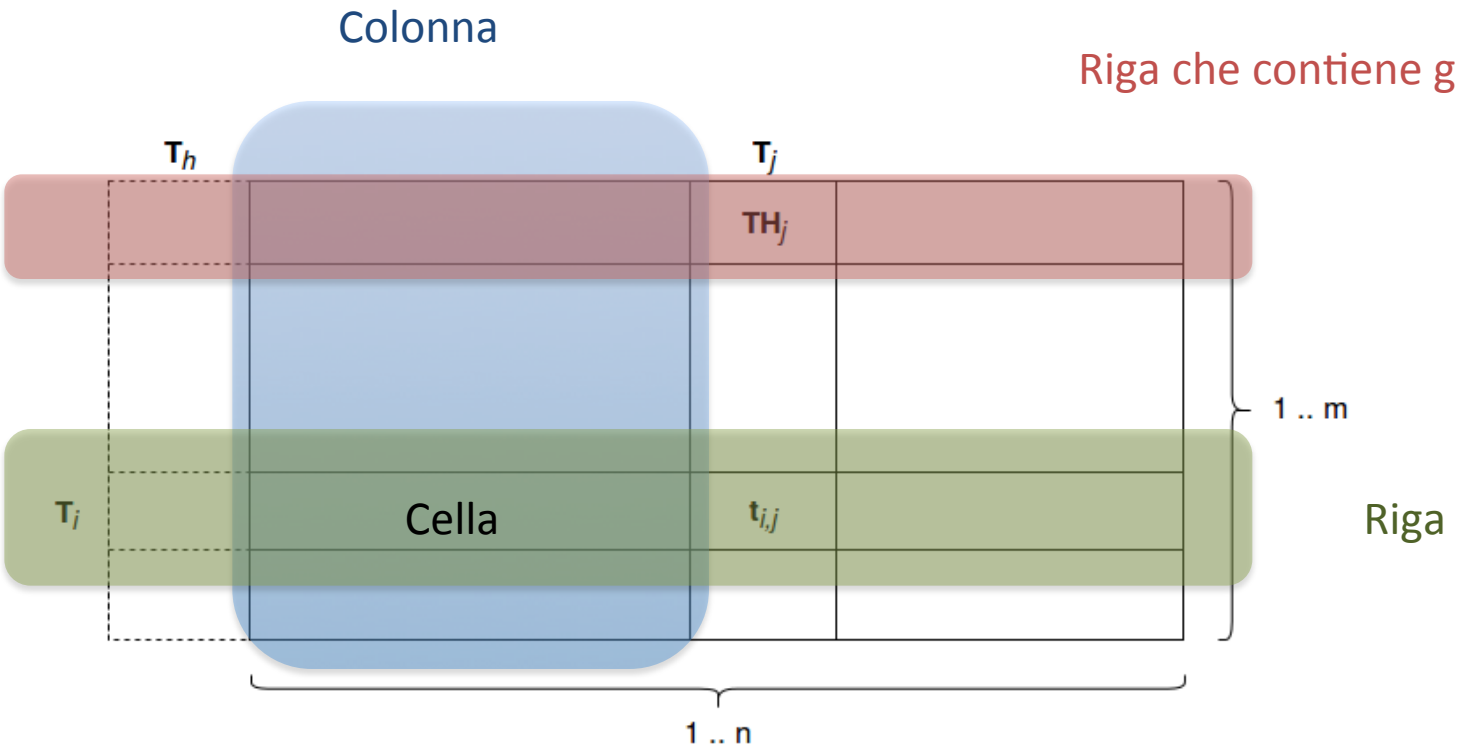
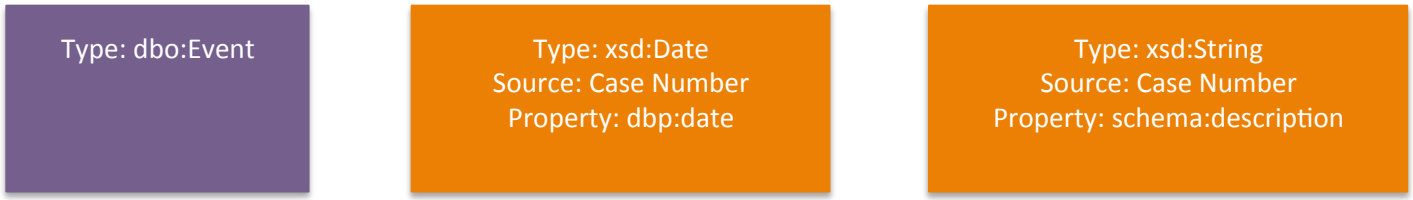


Figura 3.1: Schematizzazione di una tabella T.

Star Model



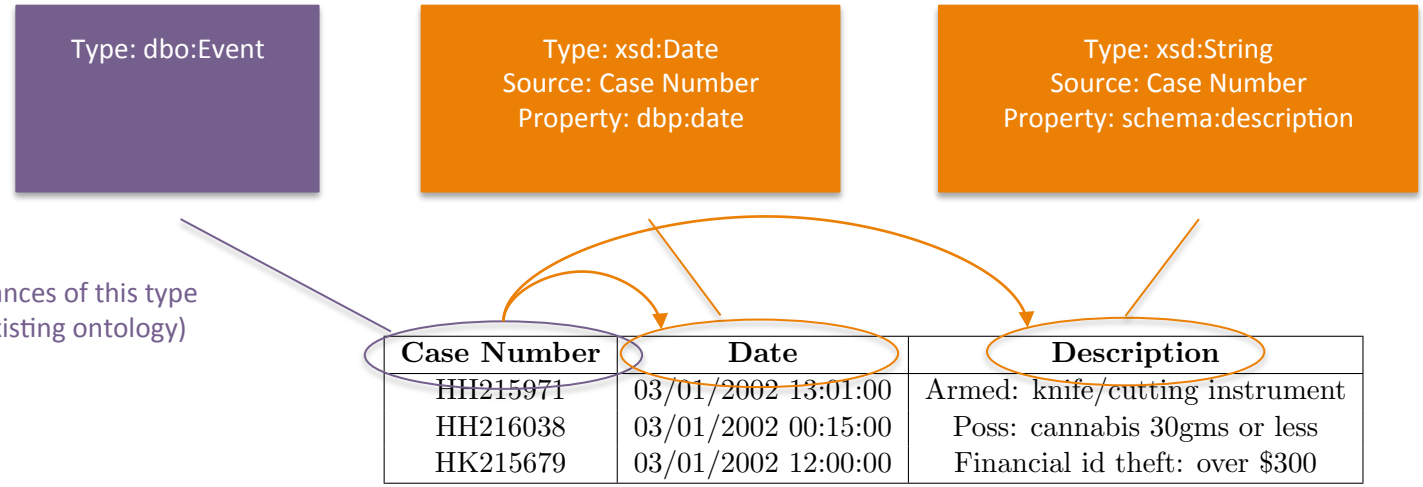
1 subject column
- define the **type**
  - values will be instances of this type
- (New type? Link it to an existing ontology)

Case Number	Date	Description
HH215971	03/01/2002 13:01:00	Armed: knife/cutting instrument
HH216038	03/01/2002 00:15:00	Poss: cannabis 30gms or less
HK215679	03/01/2002 12:00:00	Financial id theft: over $300

Tabella 3.1: Estratto della tabella relativa ai crimini commessi a Chicago.

n non-subject columns
- define the **type**
  - values will be instances of this type
- define the **source column**
  - values will be objects of triples whose subjects will be taken from this column
- define the **property**
  - values will be objects of triples with this property
- (New property/type? Link it to an existing ontology)

Star Model



1 subject column
- define the **type**
  - values will be instances of this type
- (New type? Link it to an existing ontology)

n non-subject columns
- define the **type**
  - values will be instances of this type
- define the **source column**
  - values will be objects of triples whose subjects will be taken from this column
- define the **property**
  - values will be objects of triples with this property
- (New property/type? Link it to an existing ontology)

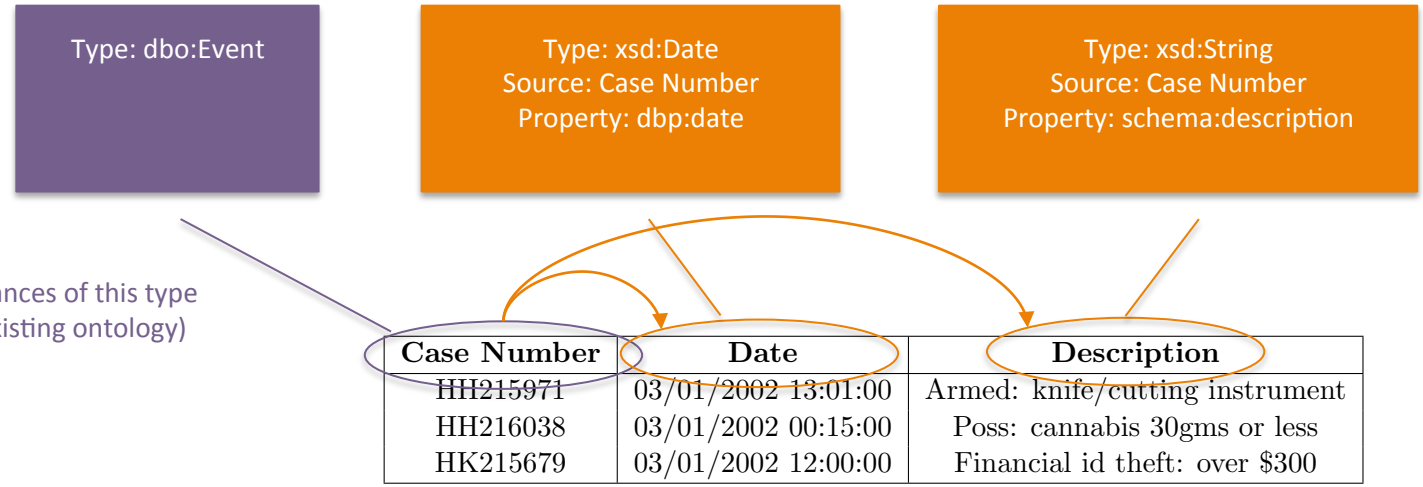
Case Number	Date	Description
HH215971	03/01/2002 13:01:00	Armed: knife/cutting instrument
HH216038	03/01/2002 00:15:00	Poss: cannabis 30gms or less
HK215679	03/01/2002 12:00:00	Financial id theft: over $300

Tabella 3.1: Estratto della tabella relativa ai crimini commessi a Chicago.

Star Model

1 subject column
- define the **type**
  - values will be instances of this type
- (New type? Link it to an existing ontology)

n non-subject columns
- define the **type**
  - values will be instances of this type
- define the **source column**
  - values will be objects of triples whose subjects will be taken from this column
- define the **property**
  - values will be objects of triples with this property
- (New property/type? Link it to an existing ontology)



Case Number	Date	Description
HH215971	03/01/2002 13:01:00	Armed: knife/cutting instrument
HH216038	03/01/2002 00:15:00	Poss: cannabis 30gms or less
HK215679	03/01/2002 12:00:00	Financial id theft: over $300

Tabella 3.1: Estratto della tabella relativa ai crimini commessi a Chicago.

Star Model

Type: Crime Case
(Crime Case
subclassOf dbo:Event)

Type: xsd:Date
Source: Case Number
Property: occurredOn
(occurredOn subProp dbp:date)

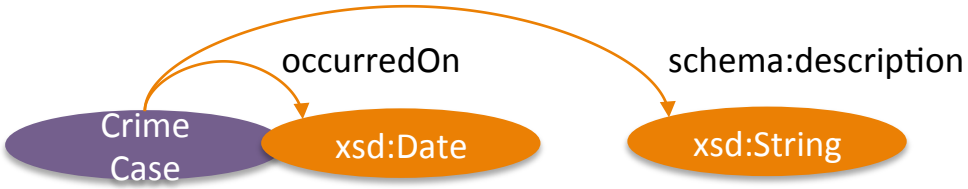
Type: xsd:String
Source: Case Number
Property: schema:description

1 subject column
- define the **type**
  - values will be instances of this type
- (New type? Link it to an existing ontology)

n non-subject columns
- define the **type**
  - values will be instances of this type
- define the **source column**
  - values will be objects of triples whose subjects will be taken from this column
- define the **property**
  - values will be objects of triples with this property
- (New property/type? Link it to an existing ontology)

Case Number	Date	Description
HH215971	03/01/2002 13:01:00	Armed: knife/cutting instrument
HH216038	03/01/2002 00:15:00	Poss: cannabis 30gms or less
HK215679	03/01/2002 12:00:00	Financial id theft: over $300

Tabella 3.1: Estratto della tabella relativa ai crimini commessi a Chicago.



Composed Star Model

Case Number	Latitude	Longitude
HH215971	41.765839537	-87.645508246
HH216038	41.906690424	-87.640736236
HK215679	42.006627736	-87.675994307

Tabella 3.2: Estratto della tabella relativa ai crimini commessi a Chicago con relative coordinate geografiche.

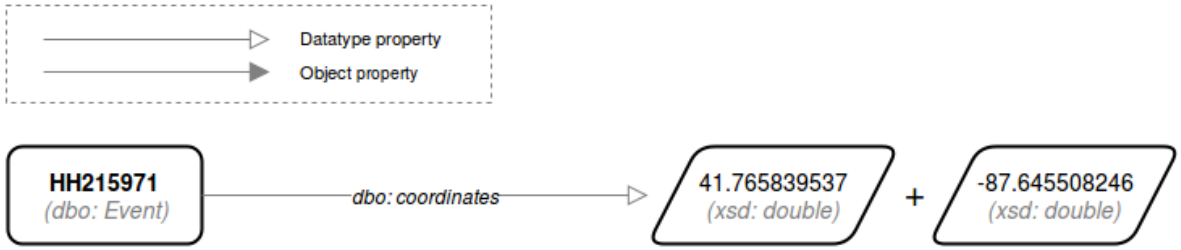


Figura 3.3: Esempio di modello Composed Star.

Snow-Flake

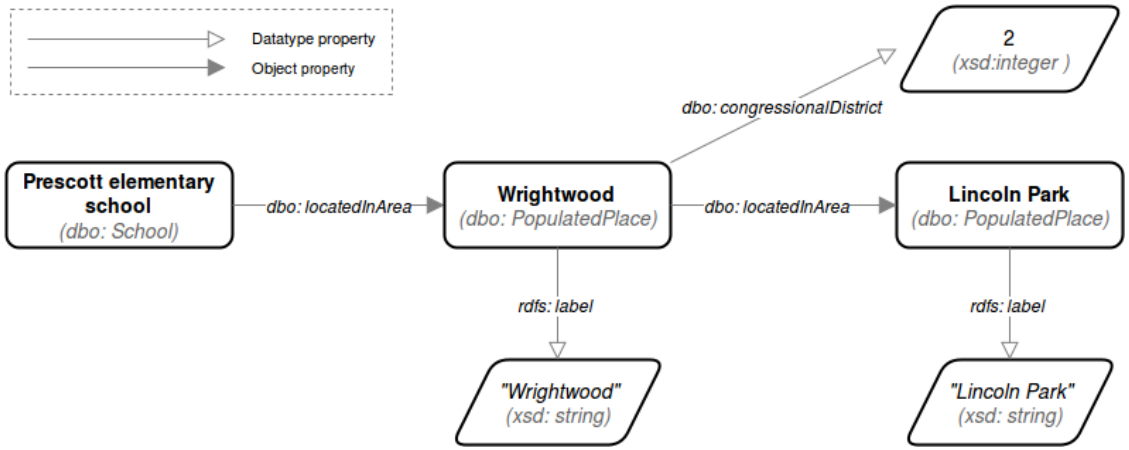


Figura 3.4: Esempio di modello Snow-Flake.

School Name	Community Area	Neighborhood	Congr. District
Prescott elementary school	Lincoln Park	Wrightwood	2
Edgebrook elementary school	Forest Glen	Edgebrook	1
Northside Prep high school	North Park	River's Edge	5

Tabella 3.3: Estratto della tabella relativa alle scuole di Chicago.

STAN Alpha

Semantic Table Annotation Tool

Home / CPS_Schools_2013-2014_Academic_Year.csv

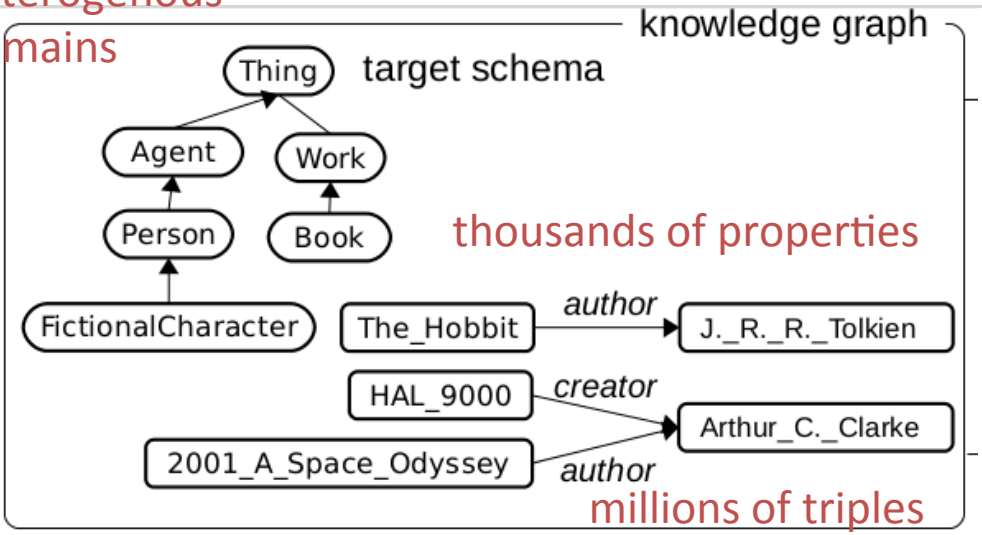
Annotation | Ontologies | Namespace

Save | Export

School	name	FullName	SchoolName2	ISBE Name	address	Street Direction	address	city	S
400010	Ace Technical Chtr HS	Architecture, Construction, and Engineering(ACE)Technical Charter School	Ace Technical Chtr HS	Ace Technical Charter High School	5410	S	State St	Chicago	IL
609772	Addams	Jane Addams Elementary School	Addams	Addams Elem School	10810	S	Avenue H	Chicago	IL
609773	Agassiz	Louis A Agassiz Elementary School	Agassiz	Agassiz Elem School	2851	N	Seminary Ave	Chicago	IL
610513	Air Force HS	Air Force Academy High School	Air Force HS	Air Force Acad High School	3630	S	Wells St	Chicago	IL
610212	Albany Park	Albany Park Multicultural Academy	Albany Park	Albany Park Multicultural Elem	4929	N	Sawyer Ave	Chicago	IL
609774	Alcott ES	Louisa May Alcott	Alcott ES	Alcott Elem	2625	N	Orchard St	Chicago	IL

Table Annotation against a Knowledge Base

heterogenous domains



consider the semantics of properties in the table and in the KG

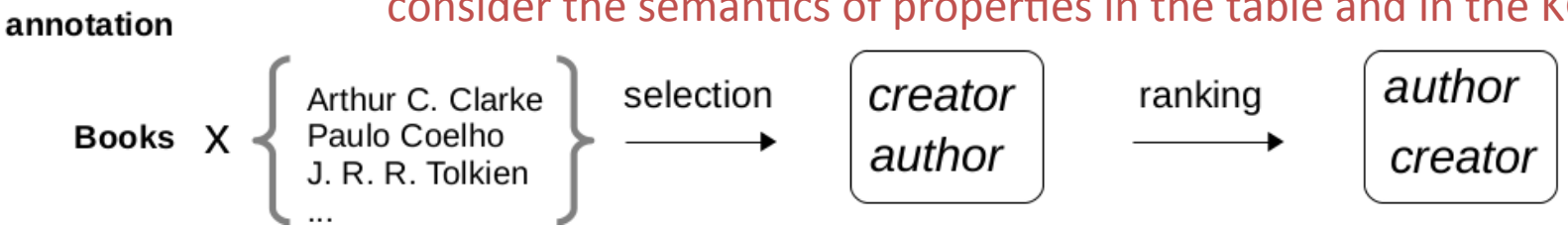
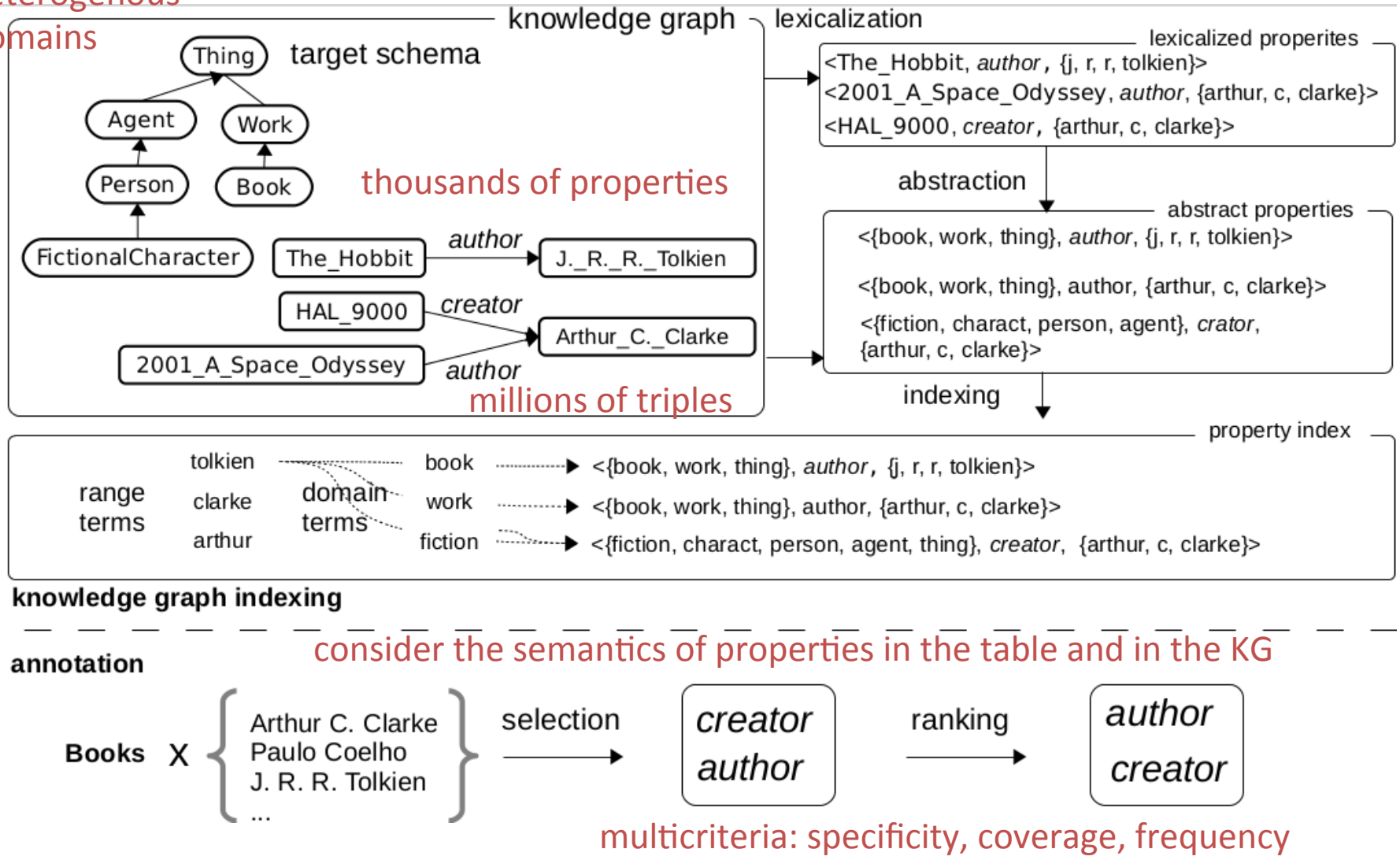


Table Annotation against a Knowledge Base

heterogenous domains



eCommerce app based on large-scale information integration systems

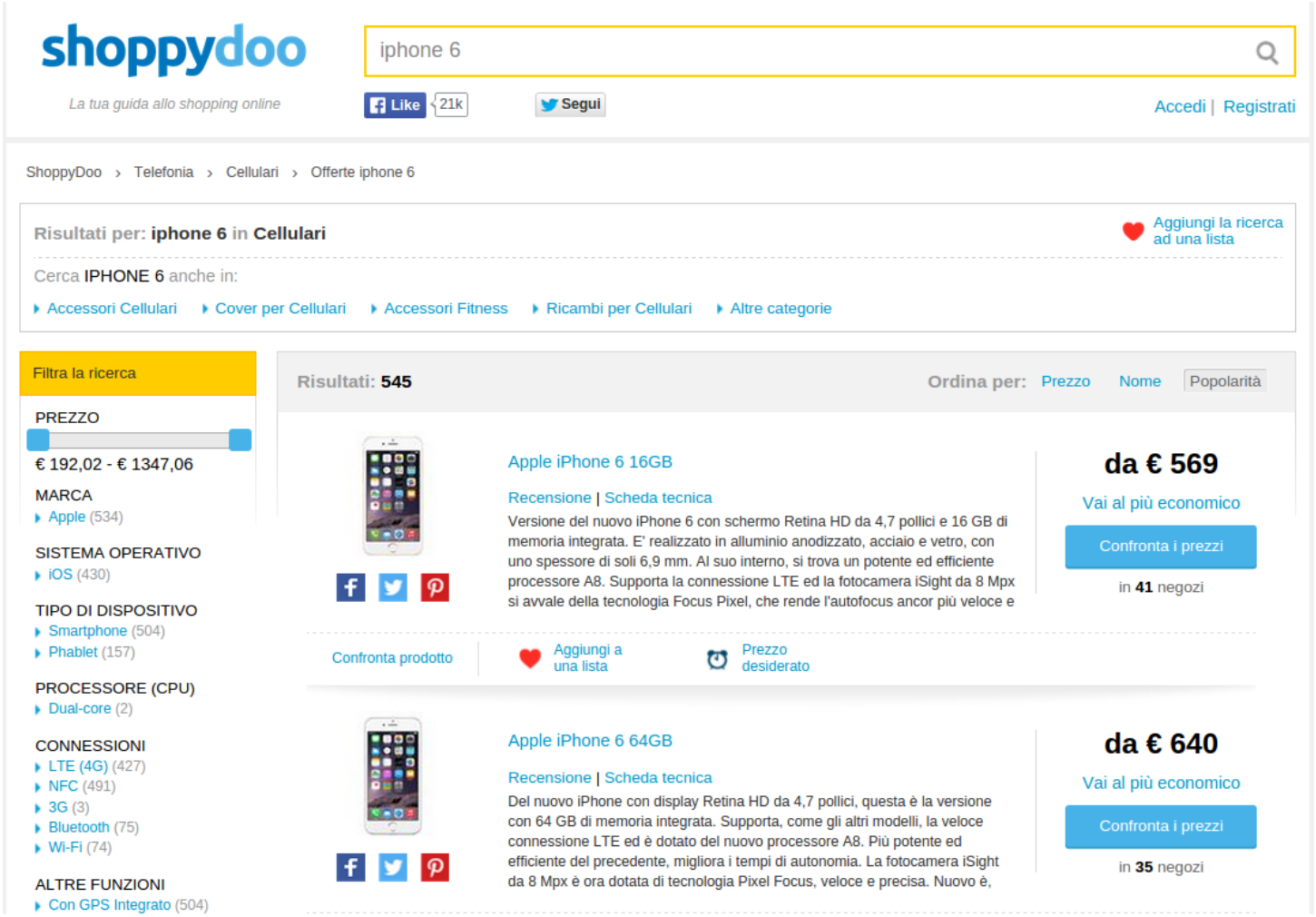
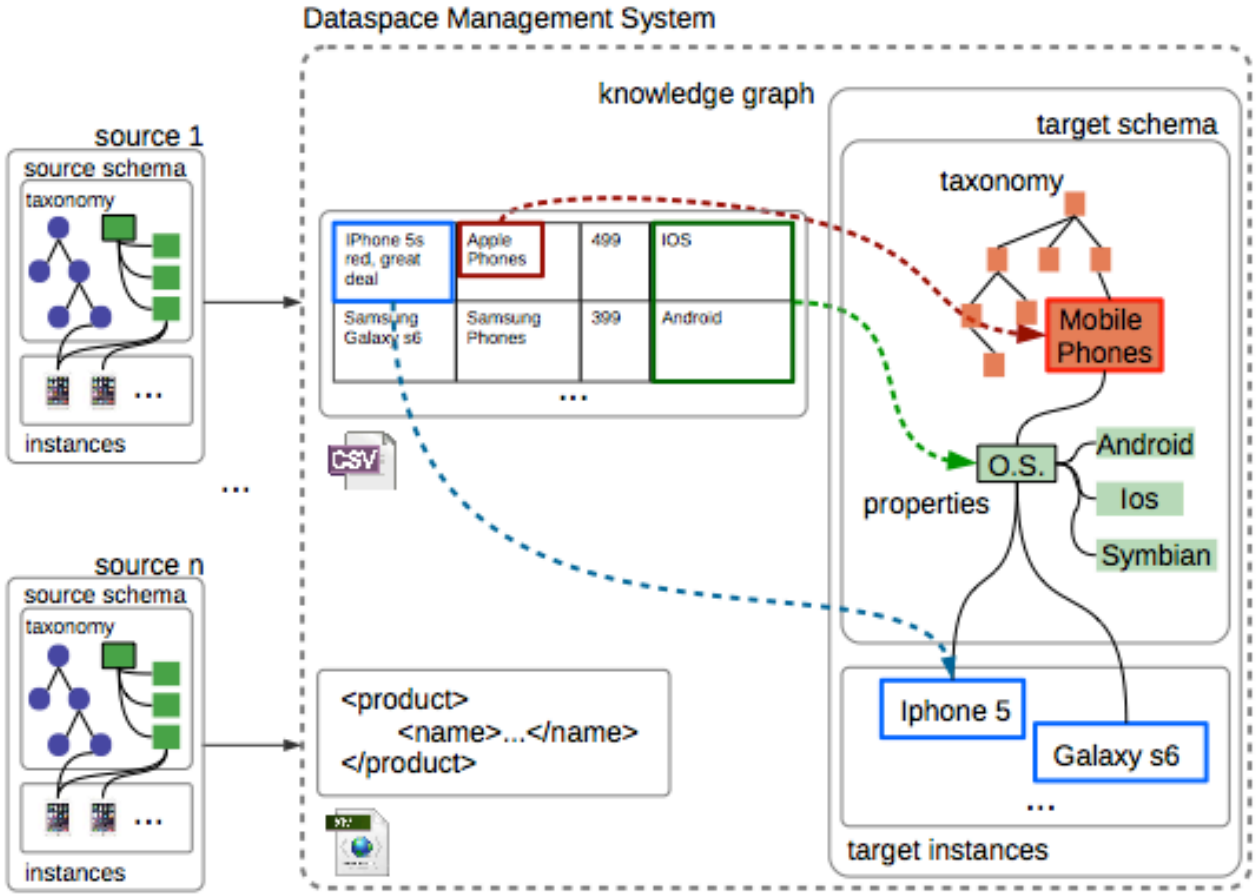


Table Annotation for Real-world Data Integration Systems

Dataspace, Knowledge Graph, Mappings



3K sources

8M source instances
200K target instances

30K source properties
4K target properties

10K source categories
500 target categories

Millions of mappings

Enrich Knowledge Graphs with Local Information via Smart Table Annotation / CHALLENGES

- Coupling instance-based matching methods with schema-based matching methods
- Including entity linking methods to support top-to-bottom data integration
- Incremental table annotation learning from previous user annotations
- Link against the whole LOD

Enrich Knowledge Graphs with Local Information via Smart Table Annotation / CHALLENGES

- **Coupling instance-based matching methods with schema-based matching methods**
- Including entity linking methods to support top-to-bottom data integration
- Incremental table annotation learning from previous user annotations
- Link against the whole LOD

Table Annotation against a Knowledge Base

header string matching

	author	language	No. of installments	releaseDate
n	Jean M. Auel	English	6	1980 - 2011
ibrary Of Learning	various authors	English	29	1980 -
	Barbara Park	English	30	1992 -
ime	Robert Jordan, Brandon Sanderson	English	15	1990 - 2013
	Michael Connelly	English	15	1992 -
	Jo Nesbø	Norwegian	9	1997–present
击队 (Picture-and-story book Railway Guerilla)	original author: Liu Zhixia	Chinese	10	1955–1962
ar	Michael Bond	English	70	1958–present
Cycle	Christopher Paolini	English	4	2002–2011
gawa Ieyasu)	Sohachi Yamaoka	Japanese	26	1950–1967
	Beverly Cleary	English	8	1955–1999
r	Stephen King	English	8	1982-2012
	Warren Murphy and Richard Sapir, various authors	English	150	1971–present

Enrich Knowledge Graphs with Local Information via Smart Table Annotation / CHALLENGES

- Coupling instance-based matching methods with schema-based matching methods
- **Including entity linking methods to support top-to-bottom data integration**
- Incremental table annotation learning from the past user annotations
- Link against the whole LOD

Table Annotation against a Knowledge Base

	author			
	Jean M. Auel			
ibrary Of Learning	various authors			
	Barbara Park			
ime	Robert Jordan, Brandon Sanderson	English	15	1990 - 2013
	Michael Connelly	English	15	1992 -
	Jo Nesbø	Norwegian	9	1997–present
站队 (Picture-and-story book Railway Guerilla)	original author: Liu Zhixia	Chinese	10	1955–1962
ar	Michael Bond	English	70	1958–present
Cycle	Christopher Paolini	English	4	2002–2011
gawa Ieyasu)	Sohachi Yamaoka	Japanese	26	1950–1967
	Beverly Cleary	English	8	1955–1999
r	Stephen King	English	8	1982-2012
	Warren Murphy and Richard Sapir, various authors	English	150	1971–present

entity linkage

Michael Connelly

Author · michaelconnelly.com

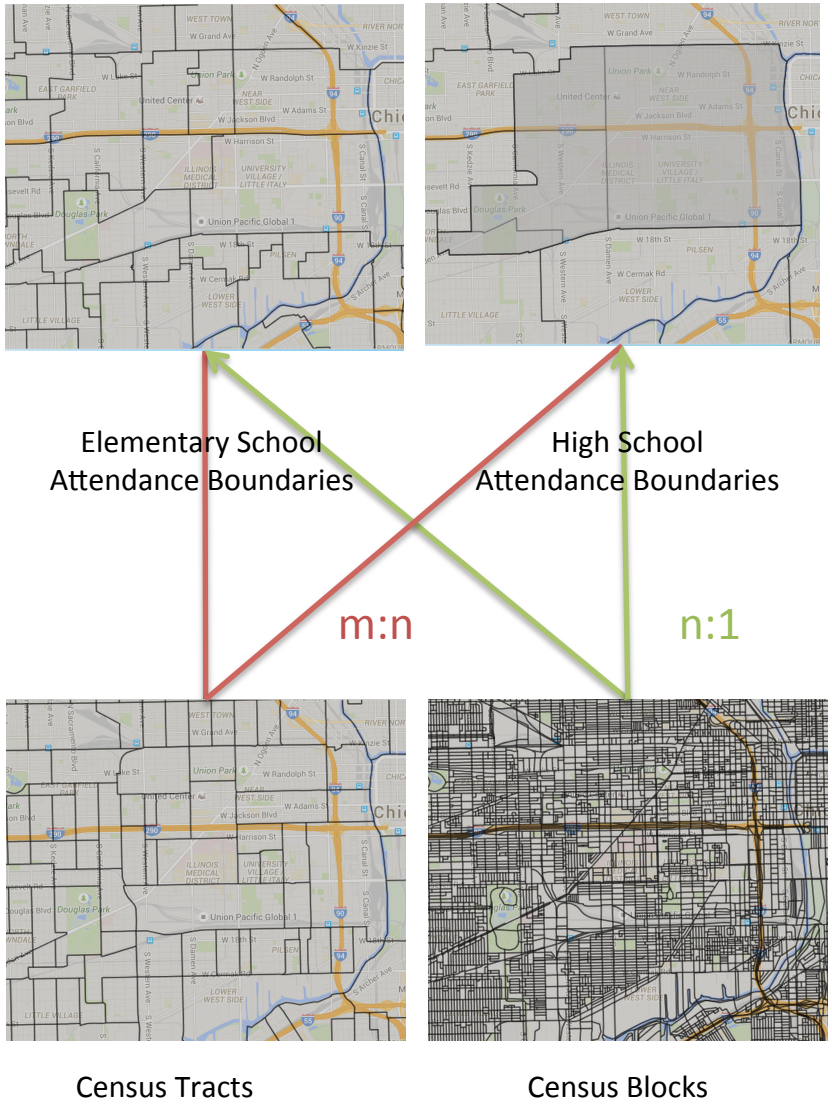
Michael Connelly is an American author of detective novels and other crime fiction, notably those featuring LAPD Detective Hieronymus "Harry" Bosch and criminal defense attorney Mickey Haller. Wikipedia

Born: July 21, 1956 (age 59), Philadelphia, Pennsylvania, United States

Geospatial Information Matching

Geospatial Integration with Transformation

"Is there a correlation between poverty and school performance?"



1. Transform format (to obtain harmonized Coordinate Reference System (CRS) or geometry serialization)

   E.g., Convert every point in the geometries of School Boundaries to WGS 84 CRS

2. Compute geometric similarity for 1:1 mappings (e.g., with Hausdorff Distance) or discover spatial relations for n:1/1:n/m:n mappings using RCC8 or DE-9IM (e.g., contains, within, intersects)

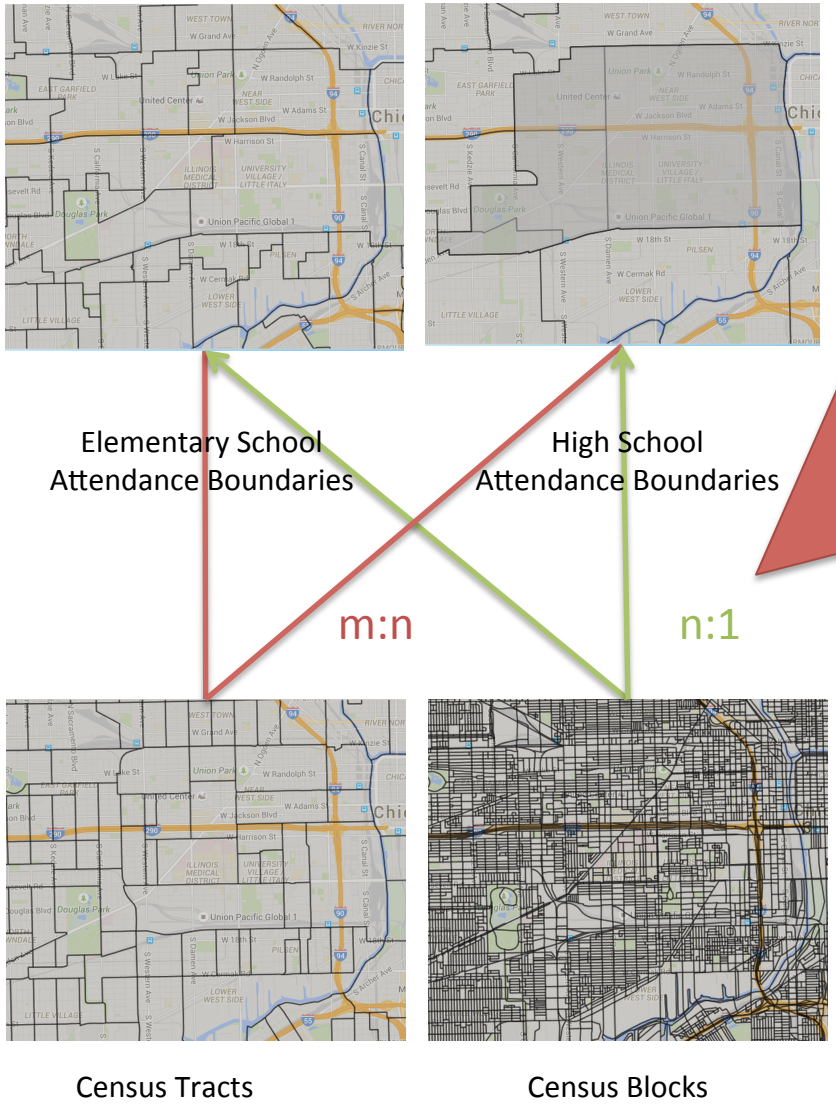
   E.g., compute containment between blocks and attendance boundaries

3. Transform data values referred to the

Mapping specification Spatial matching and execution

Geospatial Integration with Transformation

"Is there a correlation between poverty and school performance?"



Resolution vs. mapping cardinality vs. integration

- n:1 mapping facilitates the specification of aggregation functions compared to m:n mapping, e.g., block resolution is preferred to tract resolution for the integration of census data
- m:n mapping may be difficult and require complicated transformations, e,g,, disaggregation of overlapping areas

Uncertainty

- Approximate matching might be needed, e.g., if few blocks overlap two different attendance boundaries
- Approximate matching required to deal with uncertainty when variables are aggregated

for n:1/1:n/m:n mappings using RCC8 or DE-9IM (e.g., contains, within, intersects)

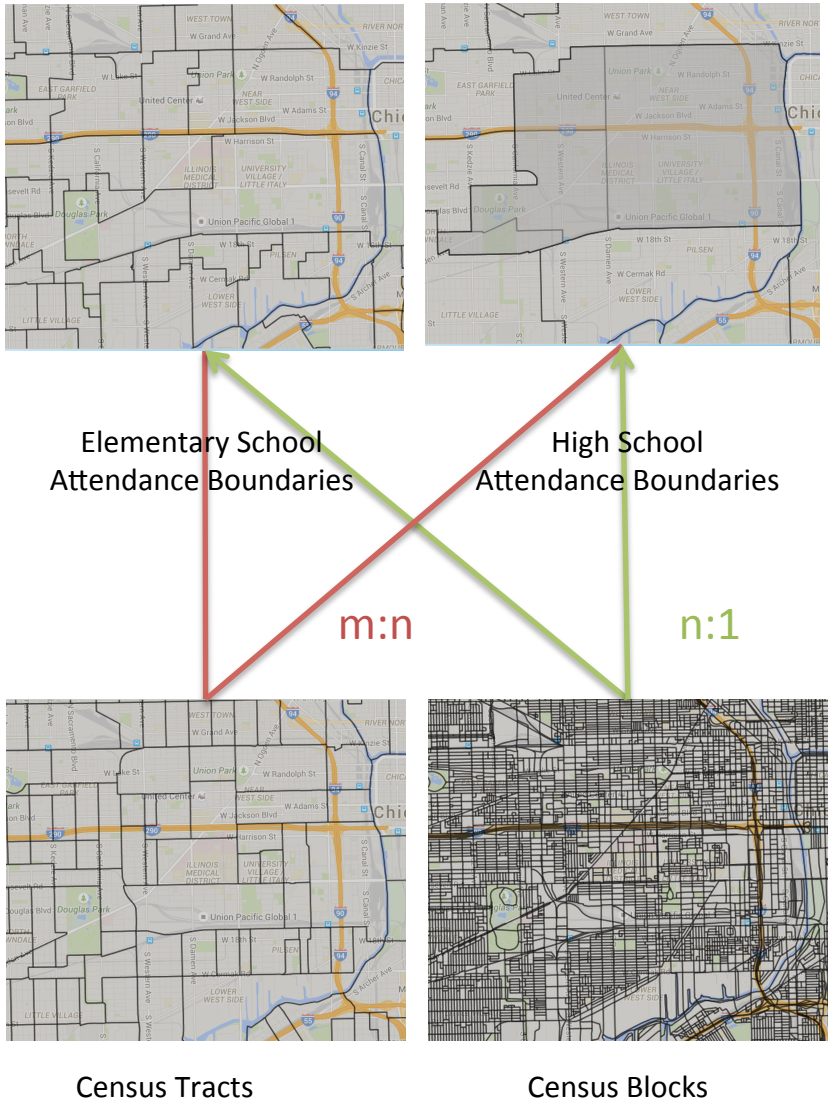
E.g., compute containment between blocks and attendance boundaries

3. Transform data values referred to the

Mapping specification and execution

Geospatial Integration with Transformation

"Is there a correlation between poverty and school performance?"



Domain-dependent vs general data transformations

- Data transformation is often domain dependent, e.g., population (absolute value), demography (distribution), income (average value)
- Subject to mathematical and domain constraints, e.g., average values cannot be added up, spatial interpolation of crime data is discouraged by many experts
- Expert involvement is needed

User support for mapping specification

- Towards semantic-assisted mapping specification, e.g., data semantics and analytics used to recommend possible transformation, support for hypothesis verification (Kandel et al. 2012)

for n:1/1:n/m:n mappings using RCC8 or DE-9IM (e.g., contains, within, intersects)

E.g., compute containment between blocks and attendance boundaries

3. Transform data values referred to the

Mapping specification and execution