# Quality-driven Evolution in Information Integration Systems

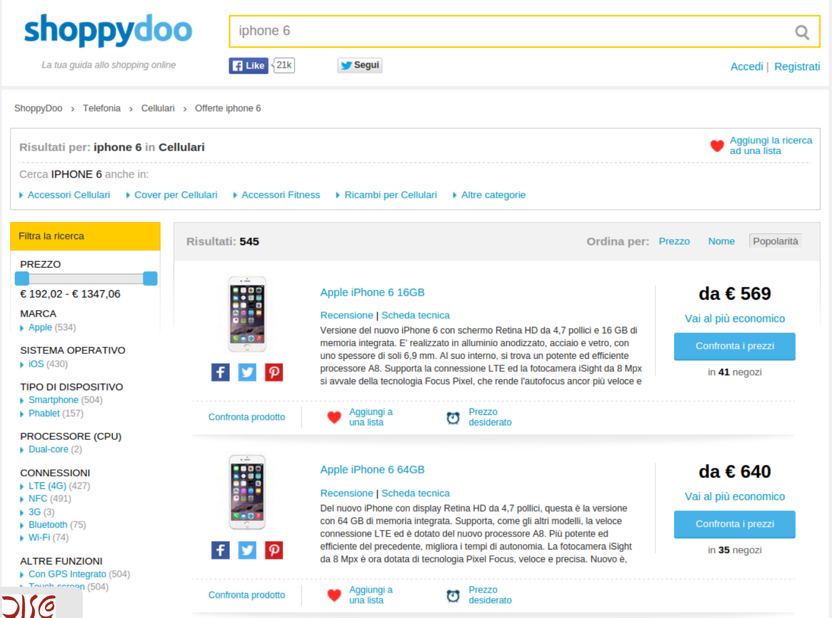
Ph.D. in Computer Science Course - XXVIII Series

**Riccardo Porrini**

**Supervisor:** Dott. Palmonari

**Tutor:** Prof. Messina

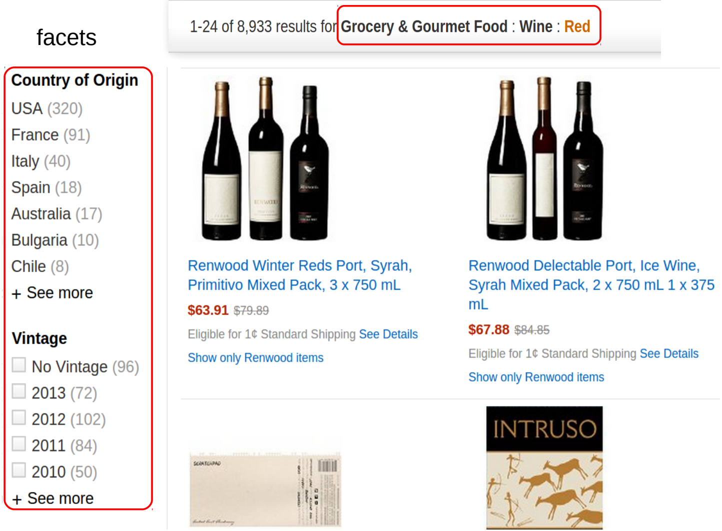
# Web Information Integration



# Multiple Classifications

categories

facets



# Multiple Classifications in Action - Product Autocomplete


* translated from Italian

demo at http://autocomplete.shoppydoo.com/demo.html

[Porrini et al. WIAS 2014] R. Porrini, M. Palmonari and G. Vizzari. Composite Match Autocompletion (COMMA): a Semantic Result-Oriented Autocompletion Technique for e-Marketplaces. In *Web Intelligence and Agent Systems Journal*, 2014

[Palmonari et al. WI 2012] M. Palmonari, G. Vizzari, R. Porrini, A. Broglia, N. Lamberti. Comma: A Result-Oriented Composite Autocompletion Method for e-Marketplaces. In *Web Intelligence*, 2012

# Multiple Classifications in Action - Product Autocomplete


* translated from Italian

support for explorative keyword based queries

demo at http://autocomplete.shoppydoo.com/demo.html

[Porrini et al. WIAS 2014] R. Porrini, M. Palmonari and G. Vizzari. Composite Match Autocompletion (COMMA): a Semantic Result-Oriented Autocompletion Technique for e-Marketplaces. In *Web Intelligence and Agent Systems Journal*, 2014

[Palmonari et al. WI 2012] M. Palmonari, G. Vizzari, R. Porrini, A. Broglia, N. Lamberti. Comma: A Result-Oriented Composite Autocompletion Method for e-Marketplaces. In *Web Intelligence*, 2012

# Multiple Classifications in Action - Product Autocomplete


* translated from Italian

support for explorative keyword based queries

**facets** and **categories** are considered in result-driven completion of the query

popularity + syntactic relevance + **semantic relevance** + **facet salience**

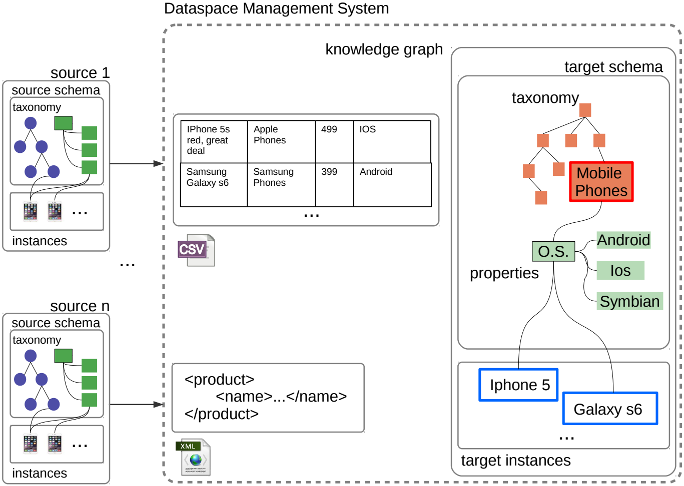
demo at http://autocomplete.shoppydoo.com/demo.html


[Porrini et al. WIAS 2014] R. Porrini, M. Palmonari and G. Vizzari. Composite Match Autocompletion (COMMA): a Semantic Result-Oriented Autocompletion Technique for e-Marketplaces. In *Web Intelligence and Agent Systems Journal*, 2014

[Palmonari et al. WI 2012] M. Palmonari, G. Vizzari, R. Porrini, A. Broglia, N. Lamberti. Comma: A Result-Oriented Composite Autocompletion Method for e-Marketplaces. In *Web Intelligence*, 2012

# Dataspace, Knowledge Graphs and Mappings

## real-world example from eCommerce




**3K** sources

**8M** source instances
**200K** target instances

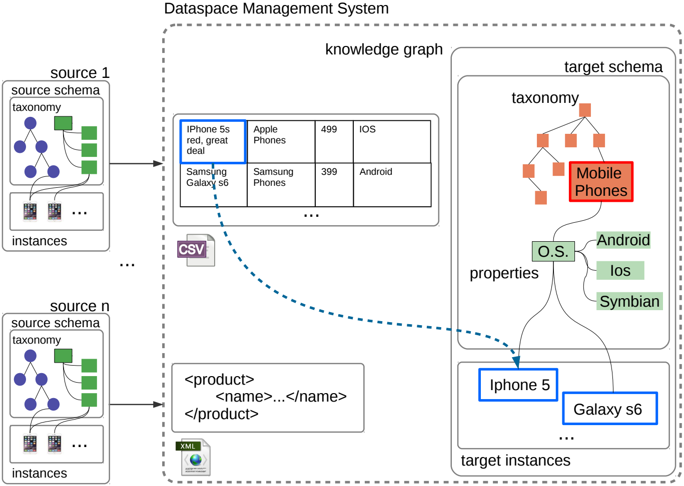
**30K** source properties
**4K** target properties

**10K** source categories
**500** target categories

**Millions** of mappings

# Dataspace, Knowledge Graphs and Mappings

## real-world example from eCommerce




**3K** sources

**8M** source instances
**200K** target instances

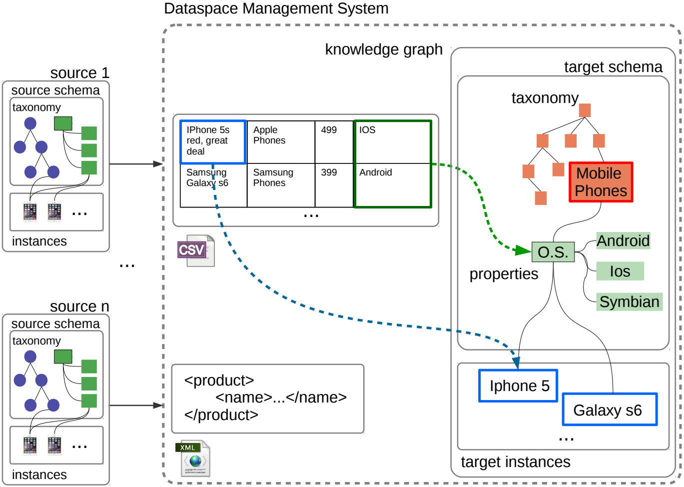
**30K** source properties
**4K** target properties

**10K** source categories
**500** target categories

**Millions** of mappings

# Dataspace, Knowledge Graphs and Mappings

## real-world example from eCommerce




**3K** sources

**8M** source instances
**200K** target instances

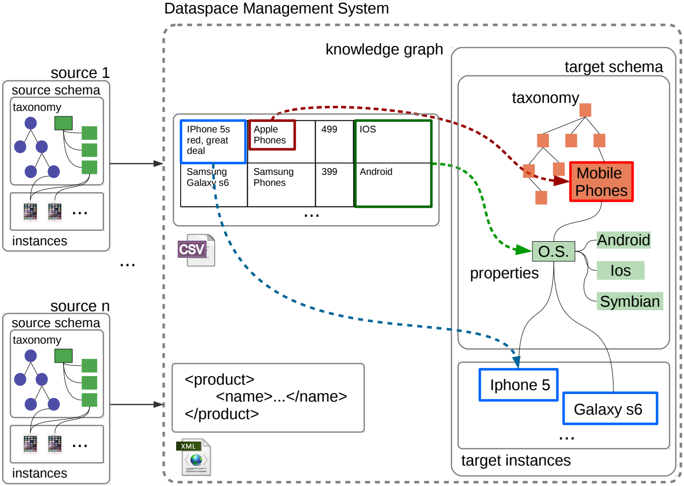
**30K** source properties
**4K** target properties

**10K** source categories
**500** target categories

**Millions** of mappings

# Dataspace, Knowledge Graphs and Mappings

real-world example from eCommerce




**3K** sources

**8M** source instances
**200K** target instances

**30K** source properties
**4K** target properties

**10K** source categories
**500** target categories

**Millions** of mappings

**target schema and mapping maintenance is hard**

- need for extensive knowledge of disparate domains

  wines and clothes

- need for domain experts supervision

  quality issues

## Challenges

**target schema and mapping maintenance is hard**

- need for extensive knowledge of disparate domains

  wines and clothes

- need for domain experts supervision

  quality issues

**crucial for**

- inclusion of new data sources covering different domains
- pay-as-you-go refinement of the integration

  from wines

  to italian, cabernet wine bottles, from 2011

## Research Challenges

highlighted by seminal works on Dataspaces [Franklin et al. 2005]

relevant to this thesis

- **schema enrichment**
  - extraction of properties and categories
    [Pound et al. 2011, Medelyan et al. 2013, Kong and Allan 2013] . . .
  - category and property profiling
    [Presutti et al. 2011, Jarrar et al. 2012] . . .
- **mapping discovery**
  - schema and ontology matching (p-to-p, c-to-c)
    [Bernstein et al. 2011, Shvaiko and Euzenat 2013] . . .
  - web table annotation (sources with weak structure)
    [Limaye et al. 2010, Venetis et al. 2011] . . .

**Dataspace** $\Delta = < \mathcal{S},\ T,\ M >$

- $\mathcal{S}$ set of sources $S$ with respective schema and instances
- $T$ target knowledge graph (schema and instances)
- $M$ set of mappings between the sources and the target

**Dataspace $\Delta =< \mathcal{S},\ T,\ M >$**

- $\mathcal{S}$ set of sources $S$ with respective schema and instances
- $T$ target knowledge graph (schema and instances)
- $M$ set of mappings between the sources and the target

| | | |
|---|---|---|
| **schema** | $\mathcal{A}^S =< C^{\mathcal{S}}, P^{\mathcal{S}} >$ | $\mathcal{A}^T =< C^T, P^T >$ |
| **instances** | $X^S = \{x_1^S, \ldots x_h^S\}$ | $X^T = \{x_1^T, \ldots x_k^T\}$ |
| **categories** | $C^S = \{c_1^S, \ldots, c_m^S\}$ | $C^T = \{c_1^T, \ldots, c_n^T\}$ |
| **properties** | $P^S = \{p_1^S, \ldots, p_w^S\}$ | $P^T = \{p_1^T, \ldots, p_j^T\}$ |

- categories as FOL unary predicates from taxonomy or a lattice
  $MobilePhones(iphone\ 6)$
- properties $p_i \subseteq X \times V_i$ as FOL binary predicates
  $price(iphone\ 6, \text{“32 } euro\text{”})$

## Formalization

### Dataspace $\Delta =< \mathcal{S},\ T,\ M >$

- $\mathcal{S}$ set of sources $S$ with respective schema and instances
- $T$ target knowledge graph (schema and instances)
- $M$ set of mappings between the sources and the target

Horn clauses encoded **mappings** from sources to target schema

$$a^T(\bar{x}) \leftarrow a_1^S(\bar{x}), \ldots, a_i^S(\bar{x}), c_1, \ldots, c_n$$

$a^T \in \mathcal{A}^T,\ a^S \in \mathcal{A}^S$

$c_1, \ldots, c_n$ constraints

**c-to-c** (category-to-category)

$Phones(x) \leftarrow ApplePhones(x)$

**p-to-p** (property-to-property)

$year(x, v) \leftarrow yearOfProduction(x, v)$

**c-to-p** (category-to-property)

$year(x, \text{“2013”}) \leftarrow WinesFrom2013(x)$

# Contributions

# Contributions

## schema enrichment

- extraction of domain specific properties from sources

  [Porrini et al. CAiSE 2014]

- analysis of property usage within the dataspace

  [Palmonari et al. ESWC 2015]

# Contributions

## schema enrichment

- extraction of domain specific properties from sources

  [Porrini et al. CAiSE 2014]

- analysis of property usage within the dataspace

  [Palmonari et al. ESWC 2015]

## mapping discovery

- establishment of category-to-property mappings

  [Porrini et al. CAiSE 2014]

- establishment of property-to-property mappings

  Paper to be submitted

## Contributions

### schema enrichment

- extraction of domain specific properties from sources

  [Porrini et al. CAiSE 2014]

- analysis of property usage within the dataspace

  [Palmonari et al. ESWC 2015]

### mapping discovery

- establishment of category-to-property mappings

  [Porrini et al. CAiSE 2014]

- establishment of property-to-property mappings

  Paper to be submitted

### case study from the eCommerce domain

- usage of target categories and properties for product autocompletion

  [Palmonari et al. WI 2012, Porrini et al. WIAS 2014]

# Property Extraction and Mapping Discovery

## goal

granular characterization of dataspace instances by extracting domain specific properties

[Porrini et al. CAiSE 2014] R. Porrini, M. Palmonari and C. Batini. Extracting Facets from Lost Fine-Grained Classifications in Dataspaces. In *CAiSE*, 2014

## goal

granular characterization of dataspace instances by extracting domain specific properties

## Wines

[Porrini et al. CAiSE 2014] R. Porrini, M. Palmonari and C. Batini. Extracting Facets from Lost Fine-Grained Classifications in Dataspaces. In *CAiSE*, 2014

# Property Extraction and Mapping Discovery

## goal

granular characterization of dataspace instances by extracting domain specific properties

## Wines

**Winery Country of Origin**
- USA
- China
- Australia
- Italy

**Specialty Wine Type**
- Sustainable
- Small Lot
- Kosher
- Gluten-Free

**Wine Alcohol By Volume**
- Under 10%
- 10% to 12%
- 12% to 14%
- 14% & Up

**Wine Vintage**
- 2011
- 2010
- 2009
- 2008
- 2007

**Grape Variety**
- Blend - White
- Blend - Other
- Fruit
- Muscadine
- Cabernet Sauvignon
- Pinot Noir
- Chardonnay

**Wine Bottle Volume**
- 375 mL
- 500 mL
- 750 mL

[Porrini et al. CAiSE 2014] R. Porrini, M. Palmonari and C. Batini. Extracting Facets from Lost Fine-Grained Classifications in Dataspaces. In *CAiSE*, 2014

## goal

granular characterization of dataspace instances by extracting domain specific properties

## observations about source categories

- source categories often come from specialized sources

  (e.g., emarketplaces selling only wine bottles)

- c-to-c mappings typically map specialized to generic categories

  $Wines(x) \leftarrow Barolo(x)$

[Porrini et al. CAiSE 2014] R. Porrini, M. Palmonari and C. Batini. Extracting Facets from Lost Fine-Grained Classifications in Dataspaces. In *CAiSE*, 2014

# Property Extraction and Mapping Discovery

**goal**

granular characterization of dataspace instances by extracting domain specific properties

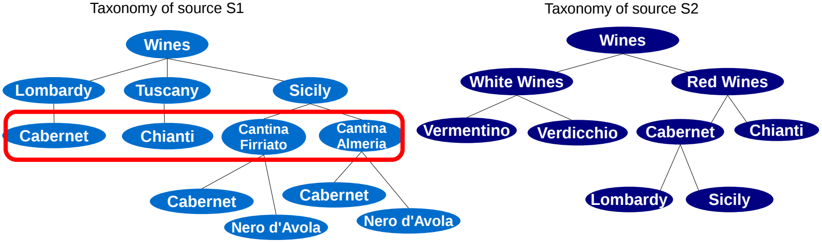
**idea:** leverage already defined c-to-c mappings

1. fix a target category $c^T$
2. pick all the source categories $c^S$ such that $c^T(x) \leftarrow c^S(x)$
3. form clusters of **homogeneous** property values $V^1, \ldots, V^n$
4. output properties $p_i \subseteq X \times V^i$
5. output c-to-p mappings: $p_i(x, v) \leftarrow v(x), v \in V^i$

[Porrini et al. CAiSE 2014] R. Porrini, M. Palmonari and C. Batini. Extracting Facets from Lost Fine-Grained Classifications in Dataspaces. In *CAiSE*, 2014

# Source Category Mutual Exclusivity Principle

the more two source categories are mutually exclusive, the more they should be clustered together into the same property value set

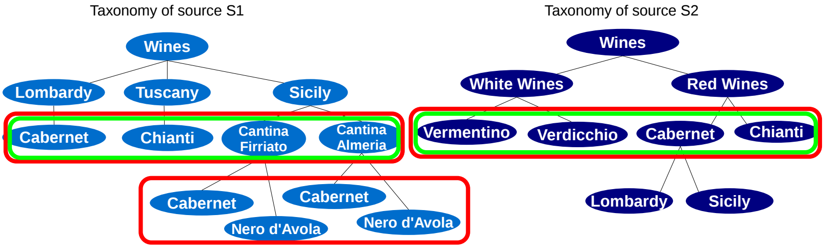



given two source categories $c_1$ and $c_2$, their occurrence as siblings indicates that $c_1$ and $c_2$ are mutually exclusive

# Taxonomy Layer Distance

$$TLD(c_1, c_2) = 1 - \frac{|L_{c_1} \cap L_{c_2}|}{|L_{c_1} \cup L_{c_2}|}$$

Jaccard distance between the two sets of taxonomy layers where two categories $c_1$ and $c_2$ occur




*cabernet chianti*

$$TLD(cabernet, chianti) = 1 - \frac{|L_{cabernet} \cap L_{chianti}|}{|L_{cabernet} \cup L_{chianti}|} = 1 - \frac{2}{3} = \frac{1}{3}$$

# Evaluation

| | #Taxonomies | #Mappings |
|---|---|---|
| Wines | 184 | 8967 |
| Musical Instruments | 128 | 1306 |
| Grappe, Liquors, Aperitives | 115 | 1254 |
| Beers | 58 | 156 |
| DVD Movies | 164 | 2042 |
| Blu-Ray Movies | 55 | 395 |
| Dogs and Cats Food | 80 | 5592 |
| Rings | 138 | 936 |
| Ski and Snowboards | 55 | 790 |
| Necklaces | 148 | 1156 |
| **Overall** | **688** | **22594** |

evaluation using real world data from the Italian PCE TrovaPrezzi

- property values manually grouped by domain experts
- comparison with Leacock and Chodorow similarity
  [Leacock and Chodorow 1998]
- and with Wu and Palmer similarity
  [Wu and Palmer 1994]

# Quantitative Evaluation

| | *Value Effectiveness* | | | *Clustering Effectiveness* | | | | *Quality* |
|---|---|---|---|---|---|---|---|---|
| | $P$ | $R$ | $F_1$ | $F^*$ | $NMI^*$ | Purity | $E^*$ | $PRF^*$ |
| LC | 0.394 | 0.953 | 0.537 | 0.666 | 0.709 | 0.220 | 0.685 | 0.531 |
| WP | 0.377 | **0.984** | 0.525 | 0.682 | 0.714 | 0.210 | 0.744 | 0.520 |
| TLD | **0.416** | 0.901 | **0.541** | **0.719** | **0.746** | **0.286** | **0.416** | **0.558** |

# Quantitative Evaluation

| | *Value Effectiveness* | | | *Clustering Effectiveness* | | | | *Quality* |
|---|---|---|---|---|---|---|---|---|
| | $P$ | $R$ | $F_1$ | $F^*$ | $NMI^*$ | Purity | $E^*$ | $PRF^*$ |
| LC | 0.394 | 0.953 | 0.537 | 0.666 | 0.709 | 0.220 | 0.685 | 0.531 |
| WP | 0.377 | **0.984** | 0.525 | 0.682 | 0.714 | 0.210 | 0.744 | 0.520 |
| TLD | **0.416** | 0.901 | **0.541** | **0.719** | **0.746** | **0.286** | **0.416** | **0.558** |

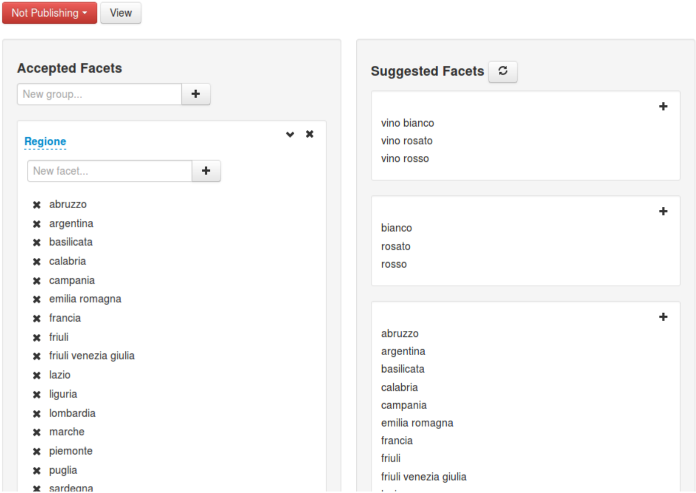
- TLD more effective in finding relevant property values and discarding noisy ones (high $F_1$)

# Quantitative Evaluation

| | *Value Effectiveness* | | | *Clustering Effectiveness* | | | | *Quality* |
|---|---|---|---|---|---|---|---|---|
| | $P$ | $R$ | $F_1$ | $F^*$ | $NMI^*$ | Purity | $E^*$ | $PRF^*$ |
| LC | 0.394 | 0.953 | 0.537 | 0.666 | 0.709 | 0.220 | 0.685 | 0.531 |
| WP | 0.377 | **0.984** | 0.525 | 0.682 | 0.714 | 0.210 | 0.744 | 0.520 |
| TLD | **0.416** | 0.901 | **0.541** | **0.719** | **0.746** | **0.286** | **0.416** | **0.558** |

- TLD more effective in finding relevant property values and discarding noisy ones (high $F_1$)
- TLD more effective in clustering homogeneous values (high *clustering effectiveness*)

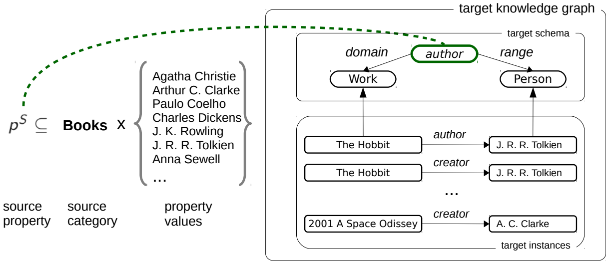
# Real World Deployment




in production within the Italian PCE TrovaPrezzi dataspace

extracted properties for **16** target categories and counting

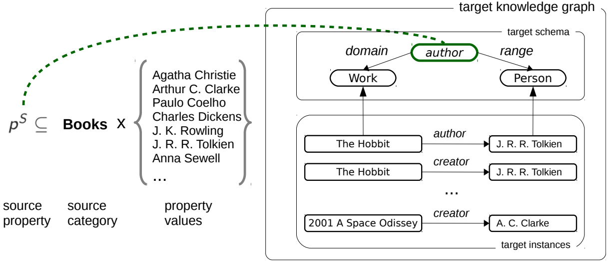
# Property-to-Property Mapping




## goal

given a source property $p^S$ select a property $p^T$ from the target schema such that the semantics of $p^S$ is compatible with $p^T$

# Property-to-Property Mapping




## challenges

- more than one target property suitable for mapping

  e.g., ∼50000 properties from the DBPedia knowledge graph

- how to resolve ambiguities?

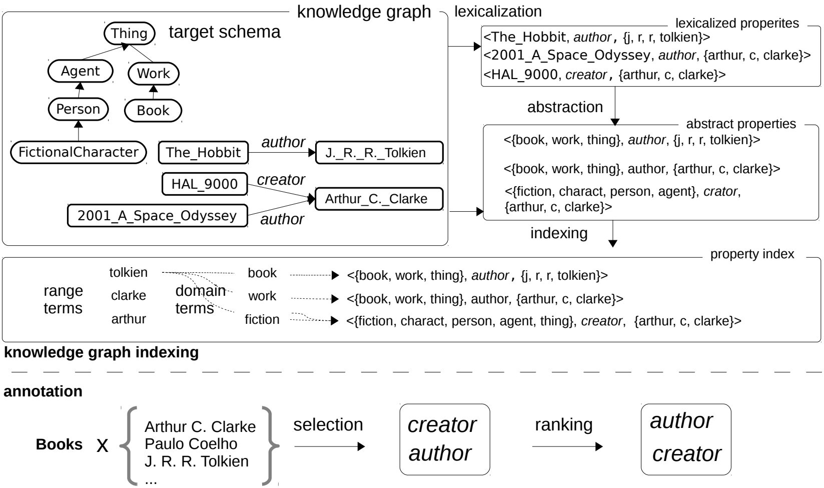
  $p^S \subseteq X^S \times \{2010, 2011, 2012, \ldots\}$

  *releaseYear* for music albums

  *vintage* for wines

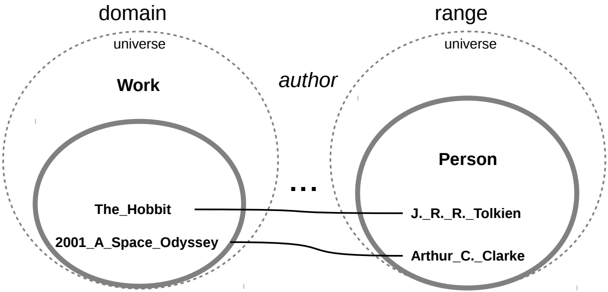
- how to capture semantic compatibility between properties?

# Approach Overview



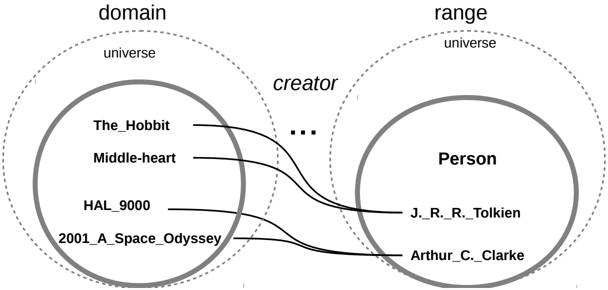
## Specificity

$$p^S \subseteq \textbf{Books} \times \left\{ \begin{array}{l} \text{Agatha Christie} \\ \text{Arthur C. Clarke} \\ \text{Paulo Coelho} \\ \text{Charles Dickens} \\ \text{J. K. Rowling} \\ \text{J. R. R. Tolkien} \\ \text{Anna Sewell} \\ \ldots \end{array} \right\}$$



## Specificity

$$p^S \subseteq \textbf{Books} \times \left\{ \begin{array}{l} \text{Agatha Christie} \\ \text{Arthur C. Clarke} \\ \text{Paulo Coelho} \\ \text{Charles Dickens} \\ \text{J. K. Rowling} \\ \text{J. R. R. Tolkien} \\ \text{Anna Sewell} \\ \ldots \end{array} \right\}$$

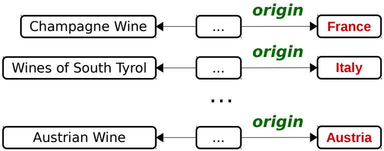


## Coverage

**Wines** X {

USA, **France**, **Italy**, **Argentina**, **Spain**, Chile, Australia, **Germany**,
**Portugal**, New Zealand, **South Africa**, **Austria**, **Greece**, Canada, **Israel**, Uruguay,
**Bulgaria**, **Hungary**, Mexico, **Switzerland**, Lebanon, Japan, **Georgia**, Morocco, **Slovenia**

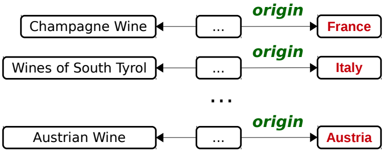
}



# Semantic Compatibility

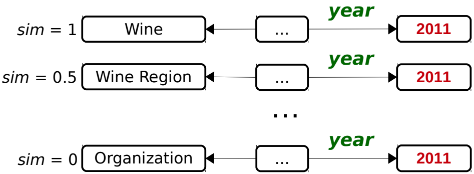
## Coverage

**Wines** X { USA, **France**, **Italy**, **Argentina**, **Spain**, Chile, Australia, **Germany**, **Portugal**, New Zealand, **South Africa**, **Austria**, **Greece**, Canada, **Israel**, Uruguay, **Bulgaria**, **Hungary**, Mexico, **Switzerland**, Lebanon, Japan, **Georgia**, Morocco, **Slovenia** }




## Frequency

**Wines** X { **2011**, 2010, 2009, 2008, 2007, 2013, 2012, 2006, 2005, 2004, 2003, 2002, 2001, 2000, 1999, 1998, 1997, 1996, 1995, 1994, 1993, 1992, 1991, 1990 }



# Evaluation

| | source properties | | target properties | | |
|---|---|---|---|---|---|
| | **#** | **categories** | **#** | **relevant** | **fair** |
| **dbpedia-numbers** | 8 | 7 | 53195 | $\sim$4 | $\sim$19 |
| **dbpedia-entities** | 31 | 13 | 53195 | $\sim$7 | $\sim$7 |
| **dbpedia** | 39 | 13 | 53195 | $\sim$3 | $\sim$9 |
| **yago-full** | 83 | 17 | 89 | 1 | - |
| **yago-abstract** | 83 | 10 | 89 | 1 | - |

- DBPedia and YAGO as target knowledge graphs
- DBPedia based gold standard created through a questionnaire
- YAGO based state-of-the-art gold standard
  [Limaye et al. 2010, Venetis et al. 2011]

# Evaluation Results

Mean Average Precision on the DBPedia Gold Standard

| | **dbpedia-numbers** | **dbpedia-entities** | **dbpedia** |
|---|---|---|---|
| majority | 0.22 | 0.50 | 0.44 |
| maximum likelihood | 0.16 | 0.39 | 0.34 |
| proposed approach | **0.25** | **0.55*** | **0.49*** |

Mean Reciprocal Rank on the YAGO Gold Standard

| | **yago-full** | **yago-abstract** |
|---|---|---|
| majority | 0.76 | 0.86 |
| maximum likelihood | 0.81 | 0.85 |
| proposed approach | **0.88*** | **0.90*** |

* $p < 0.05$

- comparison with the well known majority voting scheme and
- with maximum likelihood based approach [Venetis et al. 2011]

# Research Prototype: STAN

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 |

demo @ http://stan.disco.unimib.it

algorithm source code @ http://bitbucket.org/rporrini/cluster-labelling

credits to Brando Preda (MSc student) for the web app development

**goal**

support the analysis of property usage in a knowledge graph

- what categories are described in the knowledge graph?
- what properties are used to characterize the instances?
- how frequent is the use of a given property?

[Palmonari et al. ESWC 2015] M. Palmonari, A. Rula, R. Porrini, A. Maurino, B. Spahiu and V. Ferme. ASBTAT: Linked Data Summaries with ABstraction and STATistics. In *ESWC Posters and Demos*, 2015

# Research Prototype: ABSTAT

ABSTAT linked data summaries

browse the **abstract knowledge patterns** found into http://ld-summaries.org/dbpedia-2014 view

downloads | sparql endpoint | ld-summaries ontology | 97.58% Uptime

| | subject type (occurrences) | predicate (occurrences) | object type (occurrences) | frequency |
|---|---|---|---|---|
| filter | subject | predicate | object | |
| | http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#Agent (1688266) | http://xmlns.com/foaf/0.1/name (3546622) | http://www.w3.org/2000/01/rdf-schema#Literal (10407268) | 2270174 |
| | http://xmlns.com/foaf/0.1/Person (1445775) | http://xmlns.com/foaf/0.1/name (3546622) | http://www.w3.org/2000/01/rdf-schema#Literal (10407268) | 2058043 |
| | http://schema.org/Person (1445106) | http://xmlns.com/foaf/0.1/name (3546622) | http://www.w3.org/2000/01/rdf-schema#Literal (10407268) | 2057209 |
| | http://wikidata.dbpedia.org/resource/Q215627 (1445106) | http://xmlns.com/foaf/0.1/name (3546622) | http://www.w3.org/2000/01/rdf-schema#Literal (10407268) | 2057209 |
| | http://wikidata.dbpedia.org/resource/Q5 (1445106) | http://xmlns.com/foaf/0.1/name (3546622) | http://www.w3.org/2000/01/rdf-schema#Literal (10407268) | 2057209 |
| | http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#NaturalPerson (1445106) | http://xmlns.com/foaf/0.1/name (3546622) | http://www.w3.org/2000/01/rdf-schema#Literal (10407268) | 1980978 |
| | http://xmlns.com/foaf/0.1/Person (1445775) | http://purl.org/dc/elements/1.1/description (1738736) | http://www.w3.org/2000/01/rdf-schema#Literal (10407268) | 1738440 |
| | http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#Agent (1688266) | http://purl.org/dc/elements/1.1/description (1738736) | http://www.w3.org/2000/01/rdf-schema#Literal (10407268) | 1738356 |
| | http://wikidata.dbpedia.org/resource/Q215627 (1445106) | http://purl.org/dc/elements/1.1/description (1738736) | http://www.w3.org/2000/01/rdf-schema#Literal (10407268) | 1737966 |
| | http://schema.org/Person (1445106) | http://purl.org/dc/elements/1.1/description (1738736) | http://www.w3.org/2000/01/rdf-schema#Literal (10407268) | 1737966 |

demo @ http://abstat.disco.unimib.it

source code @ http://github.com/rporrini/abstat

# Conclusions

studied the problem of **property management** in dataspaces

- extraction of domain specific properties
- establishment of c-to-p and p-to-p mappings
- evaluation on different application domains

  eCommerce - LOD
- algorithms deployed in production or in research prototypes

**last mile(s)**

- refine the formal framework of the thesis
- experiment p-to-p mapping approach in the eCommerce domain

# Scientific Activities

## Publications

[1] M. Palmonari, A. Rula, R. Porrini, A. Maurino, B. Spahiu and V. Ferme. ASBTAT: Linked Data Summaries with ABstraction and STATistics. In *ESWC*, 2015

[2] R. Porrini, M. Palmonari and C. Batini. Extracting Facets from Lost Fine-Grained Classifications in Dataspaces. In *CAiSE*, 2014

[3] R. Porrini, M. Palmonari and G. Vizzari. Composite Match Autocompletion (COMMA): a Semantic Result-Oriented Autocompletion Technique for e-Marketplaces. In *Web Intelligence and Agent Systems Journal*, 2014

[4] M. Palmonari, G. Vizzari, R. Porrini, A. Broglia, N. Lamberti. Comma: A Result-Oriented Composite Autocompletion Method for e-Marketplaces. In *Web Intelligence*, 2012

## Conference Paper Reviews (as sub-reviewer)

- **ISWC 2015**: 15th International Semantic Web Conference
- **ESWC 2015**: 12th Extended Semantic Web Conference
- **EDBT 2015**: 18th International Conference on Extending Database Technology
- **AAAI-15**: 23th AAAI Conference on Artificial Intelligence
- **EKAW 2014**: 19th International Conference on Knowledge Engineering and Knowledge Management
- **ISWC 2014**: 14th International Semantic Web Conference
- **WI 2014**: 2014 IEEE/WIC/ACM International Conference on Web Intelligence
- **AAAI-14**: 22th AAAI Conference on Artificial Intelligence
- **ESWC 2014**: 11th Extended Semantic Web Conference
- **CAISE 2014**: 26th International Conference on Advanced Information Systems Engineering
- **ODBASE 2013**: 12th International Conference on Ontologies, DataBases, and Applications of Semantics
- **WI 2013**: 2013 IEEE/WIC/ACM International Conference on Web Intelligence

# Educational Activities

## Courses and Schools

- The Impact of Logic: from Proof Systems to Databases - Politecnico di Milano *(evaluation pending)*
- Recommender Systems - DISCo - (*final report submitted*)
- Cluster Analysis - DISCo
- Third ESWC Summer School - Kalamaky - Crete (GR)
- Foundations of Data Exchange and Integration - Politecnico di Milano
- Advanced Analytics and Behavior Informatics - DISCo

## Seminars

- From Sentiment Analysis to Continuous Learning - DISCo
- Introduzione alla Logica nella Rappresentazione della Conoscenza - DISCo
- La Misurazione Della Felicità al Tempo dei Big Data - DISCo
- Optimum Hyperpaths in Directed Hypergraphs - DISCo
- Phase Transitions in Social and Economic Systems - DISCo
- Progettare e Fare Open Data. Metodologia e tools sviluppati in Evodevo a partire dall'esperienza Open Data INPS - DISCo
- Semantic Constraints for Data Quality Assessment and Cleaning - DISCo
- Big Data e la forza degli eventi - DISCo
- WOA 2012: $13^{th}$ National Workshop "Dagli Oggetti agli Agenti" - DISCo

## Teaching

- Tutor for *Data and Web Semantics* couse (MSc and PhD, Fall 2014) - University of Illinois at Chicago
- Co-Advisor of two BSc thesis and two MSc thesis
- Lecturer (2 lectures) for the *"Artificial Intelligence"* course (MSc A.A. 2012/2013 and 2013/2014) - DISCo
- Tutor for *Distributed Systems* course (BSc A.A. 2012/2013) - DISCo

# Questions?

# Challenges

## Source taxonomies are:

- **many**
  **3900** within the TrovaPrezzi italian price comparison engine
- **noisy**
  type > white > by vine > chardonnay > producer > firriato
- **heterogeneous**
  type > white > by vine > chardonnay > producer > firriato
  wines > white wines > greco di tufo
- **ambiguous** different semantics for different contexts
  *red* is a wine type for wines
  and a color for shirts

# Related Work - Property Extraction

- *document corpora*
  focus on property hierarchies - specific for unstructured data
  [Stoica et al. 2007, Dakka and Ipeirotis 2008, Wei et al. 2013, Medelyan et al. 2013]
- search engines' *query logs* and *documents*
  user search queries as a primary source of information
  [Li et al. 2009, Pasca and Alfonseca 2009, Pound et al. 2011]
- search engines' *query results*
  integrate and rank properties already present in web documents
  [Yan et al. 2010, Dou et al. 2011, Kawano et al. 2012, Kong and Allan 2013]

## Similarity-Relatedness between taxonomy categories

- Leacock and Chodorow similarity [Leacock and Chodorow 1998]
- Wu and Palmer similarity [Wu and Palmer 1994]
- ...
  not designed for taxonomies

# Related Work - p-to-p Mapping

## from ontologies

- intensional matchers - focus on schema, neglect instances
  [Cheatham and Hitzler 2014] . . .
- extensional matchers - focus on the instances, neglect the schema
  [Zhang et al. 2015] . . .

## from tabular data

- custom knowlege graphs - knowledge graphs with special features
  [Venetis et al. 2011, Wang et al. 2012] . . .
- holistic approaches - annotate the table as a whole
  [Limaye et al. 2010, Mulwad et al. 2013, Zhang 2015] . . .

# References I

## General References

[Franklin et al. 2005] M. Franklin, A. Halevy and D. Maier. From Databases to Dataspaces: A New Abstraction for Information Management. In *SIGMOD Record*, 2005

[Bernstein et al. 2011] P. A. Bernstein, J. Madhavan and E. Rahm. Generic Schema Matching, Ten Years Later. In *PVLDB*, 2011

[Presutti et al. 2011] V. Presutti, L. Aroyo, A. Adamou, B. A. C. Schopman, A. Gangemi and G. Schreiber. Extracting Core Knowledge from Linked Data. In *COLD*, 2011

[Jarrar et al. 2012] M. Jarrar and M.D. Dikaiakos. A Query Formulation Language for the Data Web. In *IEEE Trans. Knowl. Data Eng*, 2012

[Shvaiko and Euzenat 2013] Pavel Shvaiko and Jérôme Euzenat. Ontology Matching: State of the Art and Future Challenges. In *IEEE Trans. Knowl. Data Eng.*, 2013

## Property Extraction References

[Wu and Palmer 1994] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *ACL*, 1994

[Leacock and Chodorow 1998] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. In *MIT Press*, 1998

[Stoica et al. 2007] E. Stoica, M.A. Hearst and M. Richardson. Automating creation of hierarchical faceted metadata structures. In *HLT-NAACL*, 2007

[Dakka and Ipeirotis 2008] W. Dakka and P.G. Ipeirotis. Automatic extraction of useful facet hierarchies from text databases. In *ICDE*, 2008

[Li et al. 2009] X. Li, Y.Y. Wang, A. Acero. Extracting structured information from user queries with semi-supervised conditional random fields. In *SIGIR*, 2009

[Pasca and Alfonseca 2009] M. Pasca and E. Alfonseca. Web-derived resources for web information retrieval: from conceptual hierarchies to attribute hierarchies. In *SIGIR*, 2009

# References II


[Yan et al. 2010] N. Yan, C. Li, S.B. Roy, R. Ramegowda and G. Das. Facetedpedia: enabling query-dependent faceted search for wikipedia. In *CIKM*, 2010

[Pound et al. 2011] J. Pound, S. Paparizos and P. Tsaparas. Facet discovery for structured web search: a query-log mining approach. In *SIGMOD*, 2011

[Dou et al. 2011] Z. Dou, S. Hu, Y. Luo, R. Song and J.R. Wen. Finding dimensions for queries. In *CIKM*, 2011

[Kawano et al. 2012] Y. Kawano, H. Ohshima and K. Tanaka. On-the-fly generation of facets as navigation signs for web objects. In *DASFAA*, 2012

[Wei et al. 2013] B. Wei, J. Liu, J. Ma, Q. Zheng, W. Zhang and B. Feng. Dft-extractor: a system to extract domain-specific faceted taxonomies from wikipedia. In *WWW*, 2013

[Medelyan et al. 2013] O. Medelyan, S. Manion, J. Broekstra, A. Divoli, A.L. Huang and I.H. Witten. Constructing a focused taxonomy from a document collection. In *ESWC*, 2013

[Kong and Allan 2013] W. Kong and J. Allan. Extracting query facets from search results. In *SIGIR*, 2013


## Property Mapping References


[Limaye et al. 2010] G. Limaye, S. Sarawagy and S. Chakrabarti. Annotating and Searching Web Tables Using Entities, Types and Relationship. In *VLDB*, 2010

[Venetis et al. 2011] P. Venetis, A. Halevy, J. Madhavan, M. Pasca, W. Shen, F Wu, G. Miao and C. Wu. Recovering Semantics of Tables on the Web. In *VLDB*, 2011

[Wang et al. 2012] J. Wang, H. Wang, Z. Wang, and K. Q. Zhu. Understanding tables on the web. In *ER*, 2012

[Mulwad et al. 2013] V. Mulwad, T. Finin, and A. Joshi. Semantic message passing for generating linked data from tables. In *ISWC*, 2013

[Cheatham and Hitzler 2014] M. Cheatham and P. Hitzler. The properties of property alignment. In *ISWC*, 2014

[Zhang et al. 2015] Z. Zhang, A. L. Gentile, I. Augenstein, E. Blomqvist, and F. Ciravegna. An Unsupervised Data-driven Method to Discover Equivalent Relations in Large Linked Datasets. In *Sem. Web - accepted for publication*, 2015

# References III

[Zhang 2015] Z. Zhang. Effective and Efficient Semantic Table Interpretation using TableMiner+. In *Sem. Web - under transparend review*, 2015