Facet Extraction, Annotation and Alignment in Dataspaces

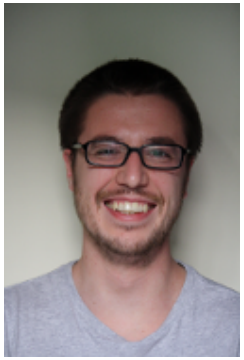
Riccardo Porrini

DISCo, University of Milano-Bicocca

7Pixel s.r.l.

riccardo.porrini@disco.unimib.it

About Me

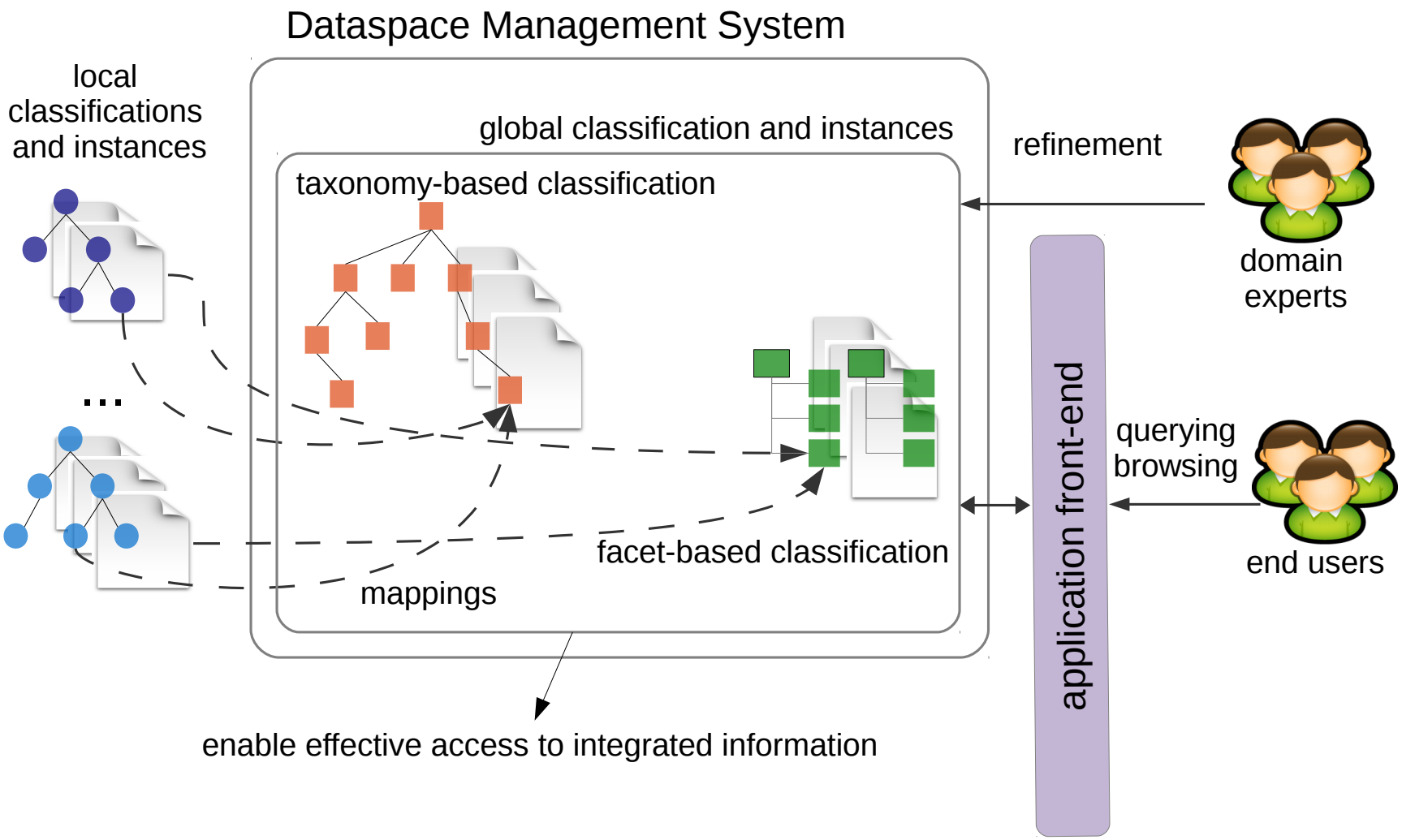


- PhD Student @ UniMiB (Milan) from January 2013
  supervisor: Matteo Palmonari
- currently visiting ADVIS Lab @ UIC
  under the supervision of Prof. Isabel Cruz
- 5 years as back-end software developer for 7Pixel
  eCommerce and price comparison domain
  mainly facing Data Integration issues

research interests (broadly speaking):

- information systems
- data integration
- semantic web
- linked data
- ...follow the rest of the presentation :)

Dataspaces with Multiple Classifications



Taxonomy-based Classification

taxonomy: categories organized through a hierarchical structure *(informal)*

1-24 of 8,933 results for **Grocery & Gourmet Food** : **Wine** : **Red**

coarse-grained classification helps end users to recall the "class" of instances

Country of Origin

- USA (320)
- France (91)
- Italy (40)
- Spain (18)
- Australia (17)
- Bulgaria (10)
- Chile (8)
- + See more

Vintage

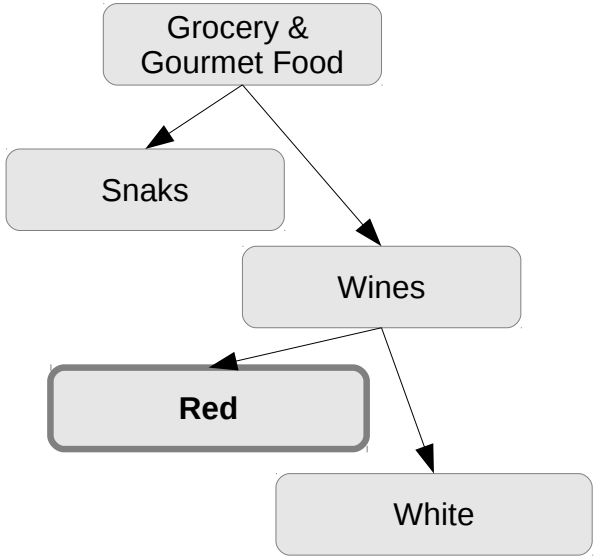
- No Vintage (96)
- 2013 (72)
- 2012 (102)
- 2011 (84)
- 2010 (50)
- + See more

Renwood Winter Reds Port, Syrah, Primitivo Mixed Pack, 3 x 750 mL

$63.91 ~~$79.89~~

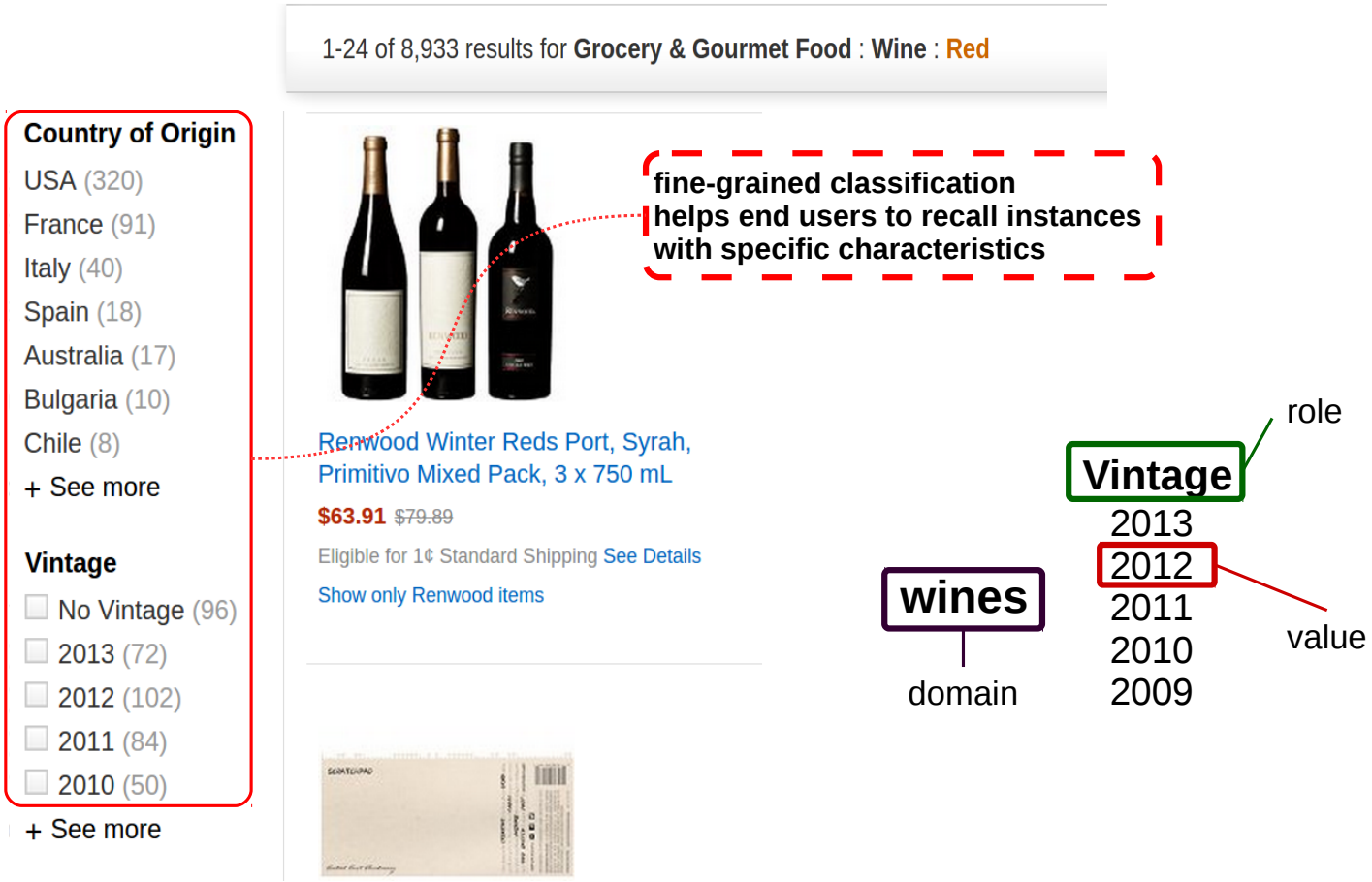
Eligible for 1¢ Standard Shipping See Details

Show only Renwood items



Facet-based Classification

facet: a clearly defined, mutually exclusive, and collectively exhaustive aspect, property, or characteristic of a class or specific subject [Taylor 2004]



[Taylor 2004] A.G. Taylor. Wynars introduction to cataloging and classification. *Libraries Unlimited*, 2004

Multiple Classifications in Action: Product Autocomplete

* translated from Italian

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

Multiple Classifications in Action: Product Autocomplete

* translated from Italian

support for explorative keyword queries

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

Multiple Classifications in Action: Product Autocomplete

* translated from Italian

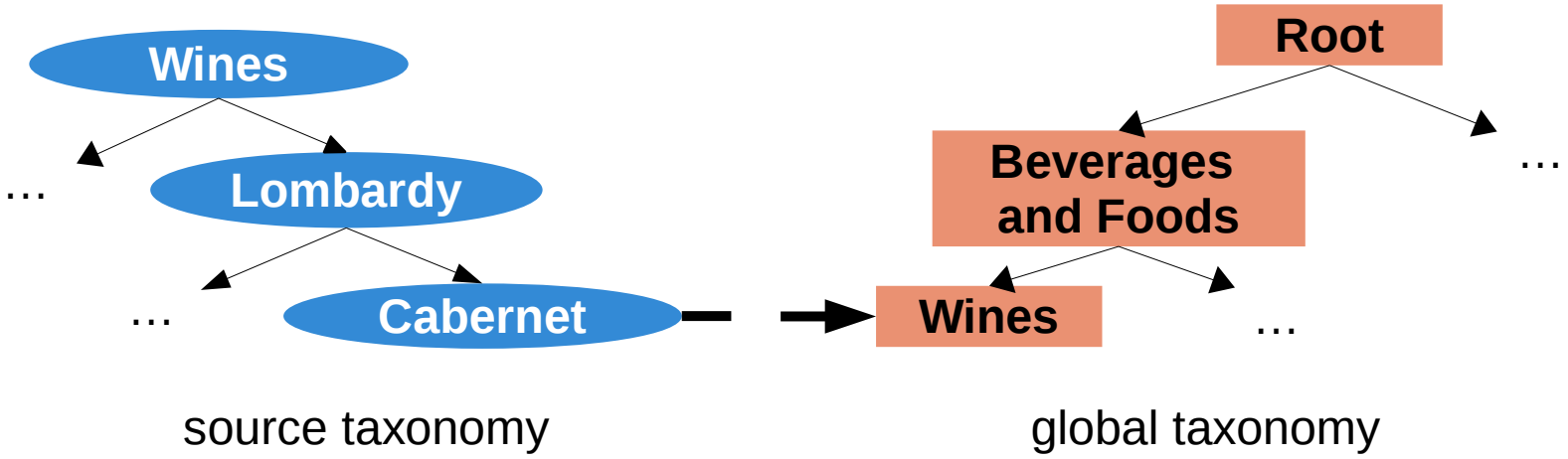
support for explorative keyword queries

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

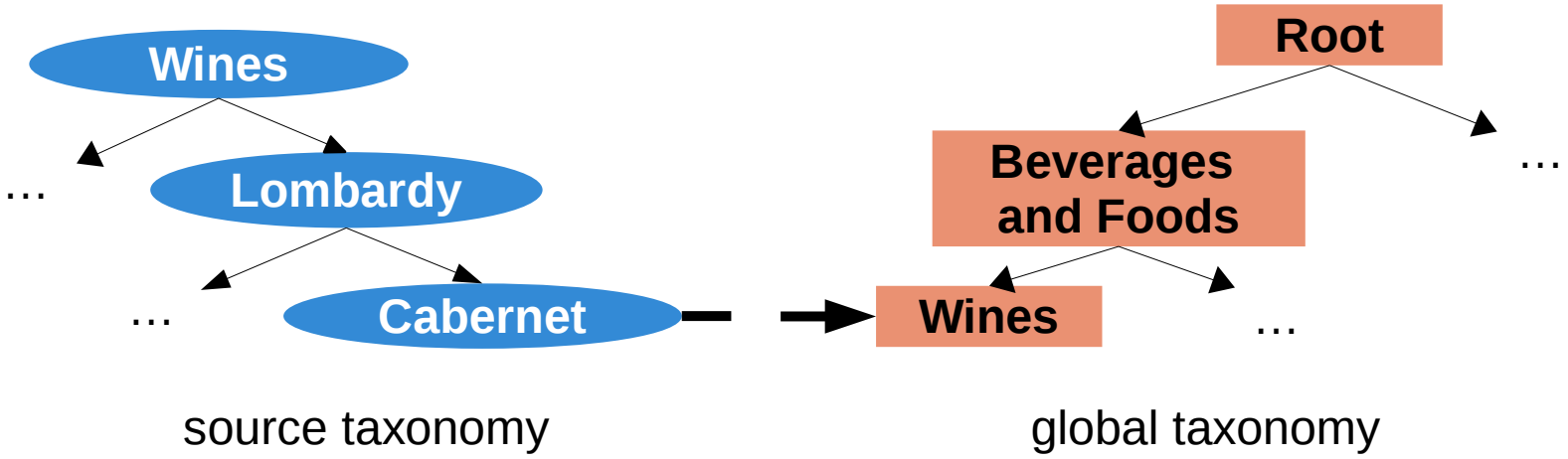
Popularity + Syntactic Relevance + **Semantic Relevance** + **Facet Salience**

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

Mappings


Wines
...
Lombardy
...
Cabernet
Root
Beverages and Foods
...
Wines
...
source taxonomy
global taxonomy

Mappings



usually, source taxonomies provide much more **granular classification**

Faceted Classifications Maintenance

faceted classification bootstrapping

- time and effort consuming
- requires detailed knowledge about dataspace instances

Faceted Classifications Maintenance

faceted classification bootstrapping

- time and effort consuming
- requires detailed knowledge about dataspace instances

however

- source taxonomies usually provide much more granular classification
- this granular information is lost when mapping specific source categories to generic global ones

Faceted Classifications Maintenance

faceted classification bootstrapping

- time and effort consuming
- requires detailed knowledge about dataspace instances

however

- source taxonomies usually provide much more granular classification
- this granular information is lost when mapping specific source categories to generic global ones

how about extracting facets from those lost fine-grained source taxonomies?

Problem Statement

Facet Extraction Problem

given a global category g, a set of mappings M from source categories $s_1, \ldots, s_n$ to g, extract a set $\mathcal{F}^g$ of facets F^g

Problem Statement

Facet Extraction Problem

given a global category g, a set of mappings M from source categories $s_1, \ldots, s_n$ to g, extract a set $\mathcal{F}^g$ of facets F^g

Wines

Problem Statement

Facet Extraction Problem

given a global category g, a set of mappings M from source categories $s_1, \ldots, s_n$ to g, extract a set $\mathcal{F}^g$ of facets F^g

Wines

Winery Country of Origin

- USA
- China
- Australia
- Italy

Wine Alcohol By Volume

- Under 10%
- 10% to 12%
- 12% to 14%
- 14% & Up

Grape Variety

- Blend - White
- Blend - Other
- Fruit
- Muscadine
- Cabernet Sauvignon
- Pinot Noir
- Chardonnay

Wine Bottle Volume

- 375 mL
- 500 mL
- 750 mL

Specialty Wine Type

- Sustainable
- Small Lot
- Kosher
- Gluten-Free

Wine Vintage

- 2011
- 2010
- 2009
- 2008
- 2007

Challenges

Source taxonomies are:

- **many**
  e.g., **3900** within the TrovaPrezzi Italian price comparison engine

Challenges

Source taxonomies are:

- **many**
  e.g., **3900** within the TrovaPrezzi Italian price comparison engine
- **noisy**
  type > white > by vine > chardonnay > producer > firriato

Challenges

Source taxonomies are:

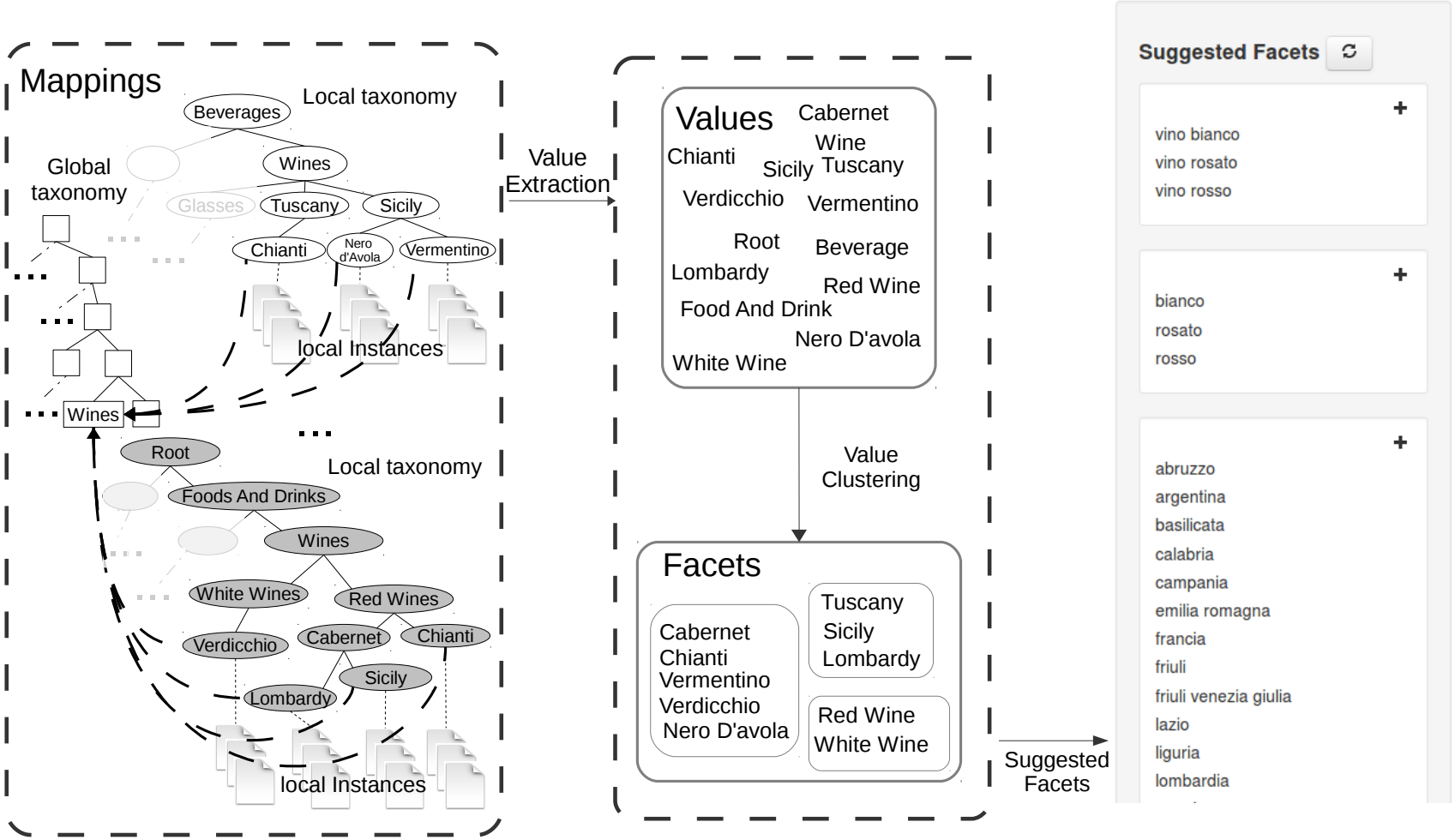
- **many**
  e.g., **3900** within the TrovaPrezzi Italian price comparison engine
- **noisy**
  type > white > by vine > chardonnay > producer > firriato
- **heterogeneous**
  type > white > by vine > chardonnay > producer > firriato
  wines > white wines > greco di tufo

Challenges

Source taxonomies are:

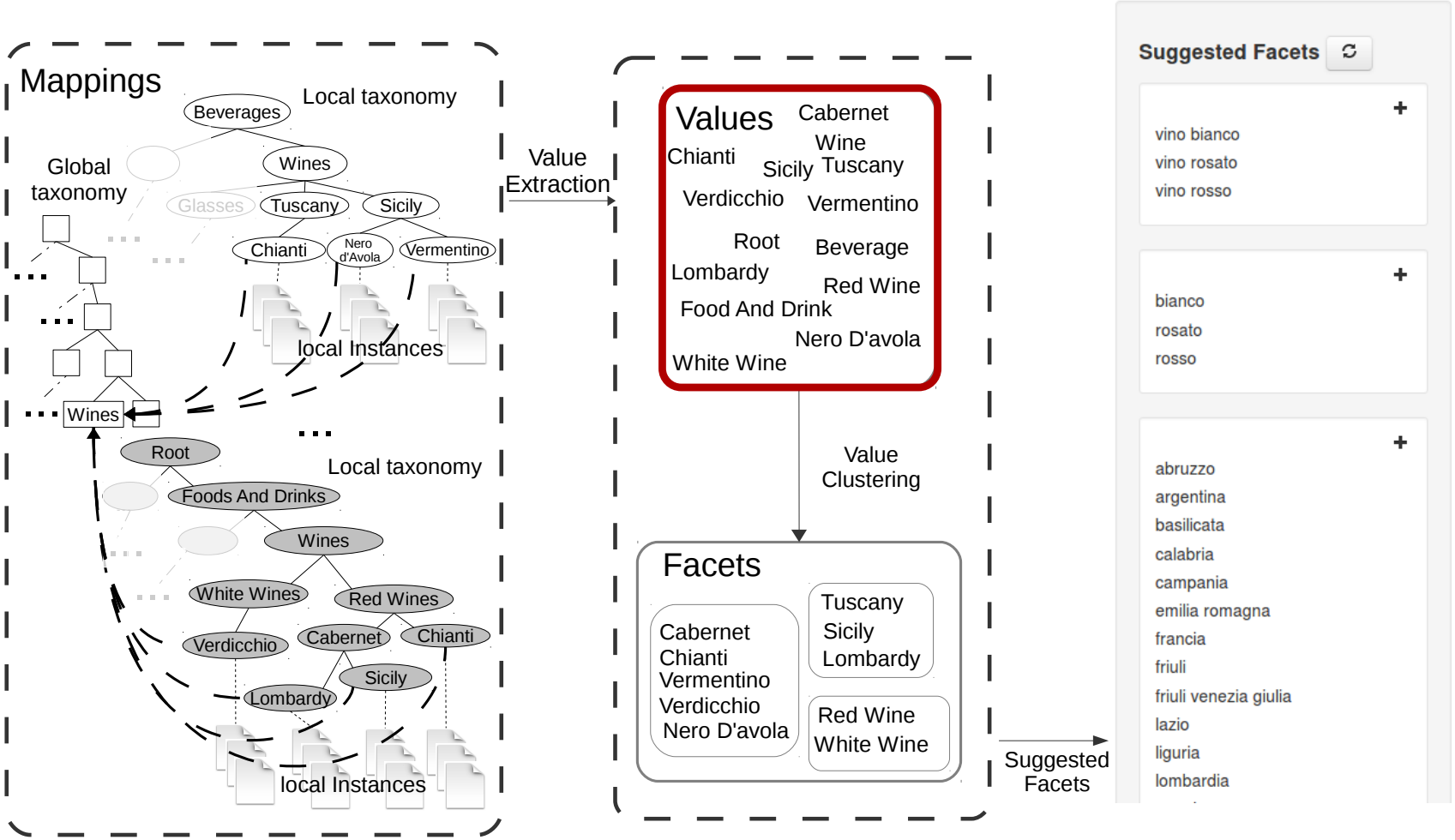
- **many**
  e.g., **3900** within the TrovaPrezzi Italian price comparison engine
- **noisy**
  type > white > by vine > chardonnay > producer > firriato
- **heterogeneous**
  type > white > by vine > chardonnay > producer > firriato
  wines > white wines > greco di tufo
- **ambiguous** different meaning in different domains
  red is a wine type for wines
  and a color for shirts

Facet extraction



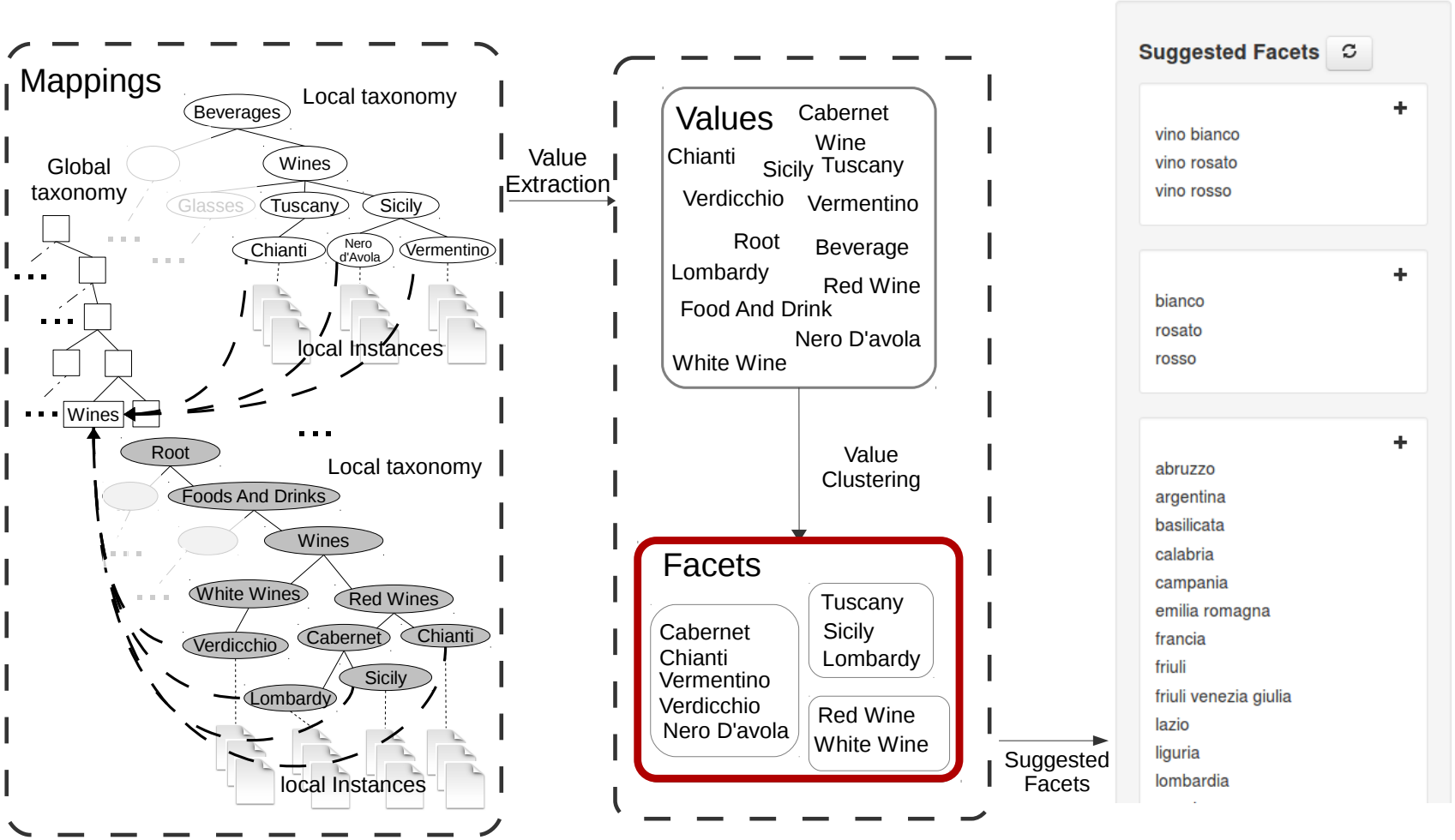
[Porrini et al. 2014] R. Porrini, M. Palmonari and C. Batini. Extracting Facets from Lost Fine-Grained Classifications in Dataspaces. *CAiSE*, 2014

Facet extraction



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

Facet extraction

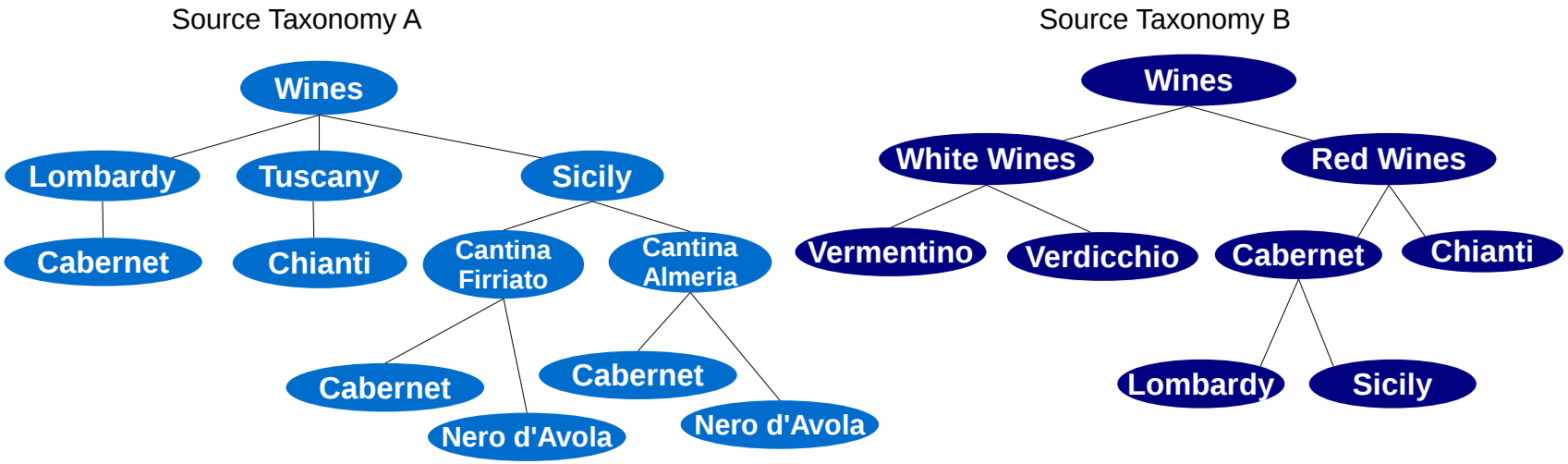


[Porrini et al. 2014] R. Porrini, M. Palmonari and C. Batini. Extracting Facets from Lost Fine-Grained Classifications in Dataspaces. *CAiSE*, 2014

Taxonomy Layer Distance

principle

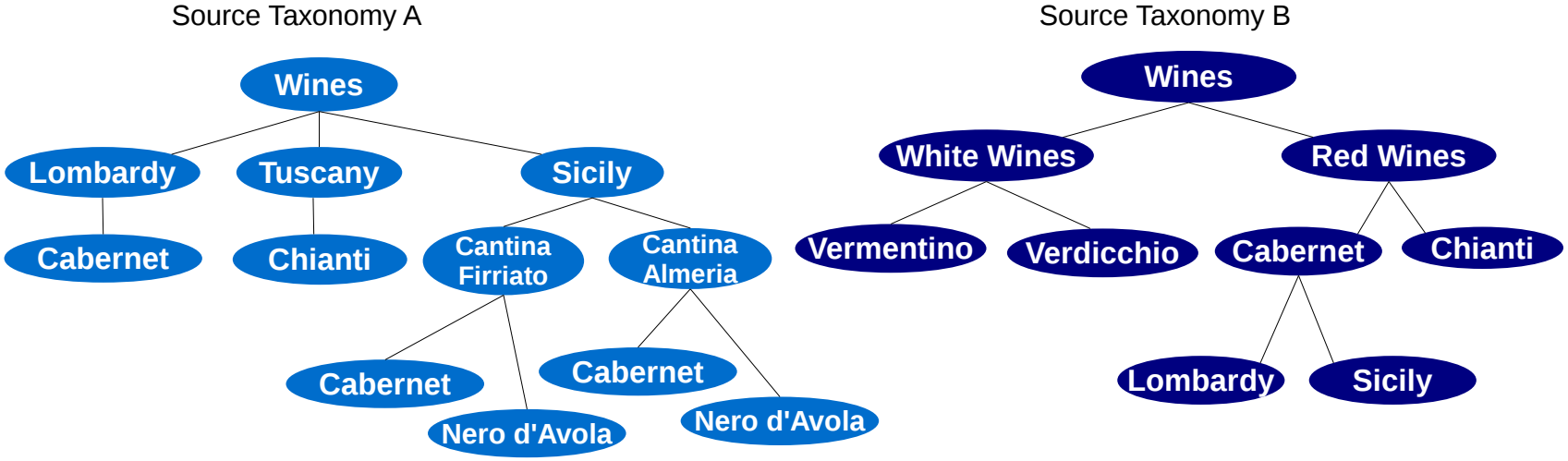
the more two values refer to mutually exclusive categories, the more they should be grouped together into the same facet



Taxonomy Layer Distance

$$\mathrm{TLD}(s_1, s_2) = 1 - \frac{|L_{s_1} \cap L_{s_2}|}{|L_{s_1} \cup L_{s_2}|}$$

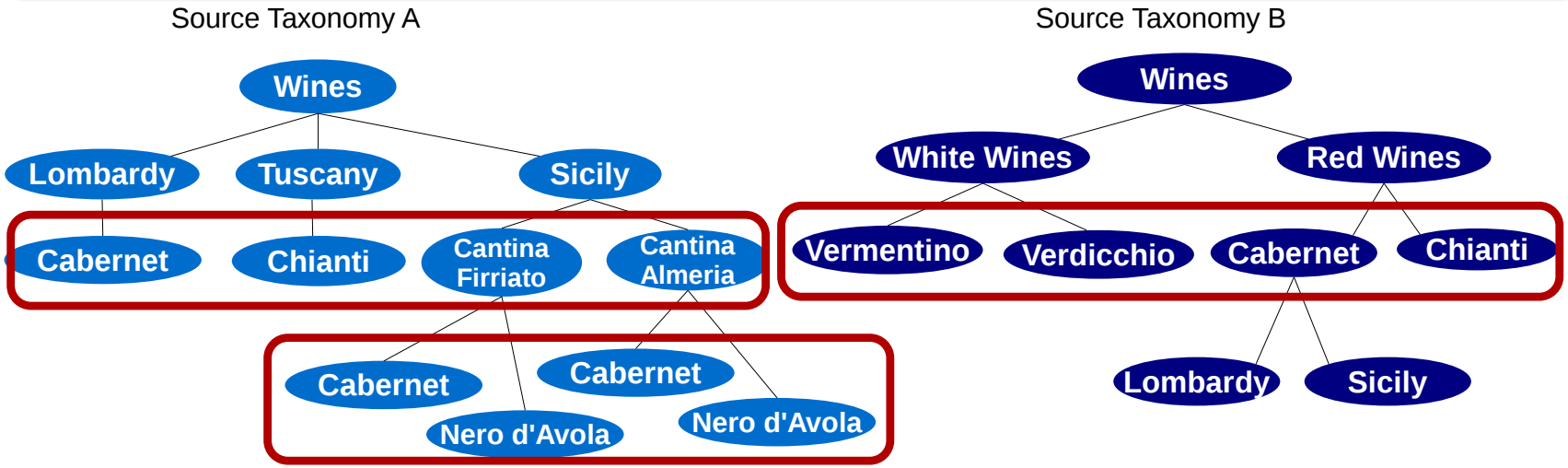
Jaccard Distance between the two sets of taxonomy layers where two categories s_1 and s_2 occur



Taxonomy Layer Distance

$$\mathrm{TLD}(s_1, s_2) = 1 - \frac{|L_{s_1} \cap L_{s_2}|}{|L_{s_1} \cup L_{s_2}|}$$

Jaccard Distance between the two sets of taxonomy layers where two categories s_1 and s_2 occur

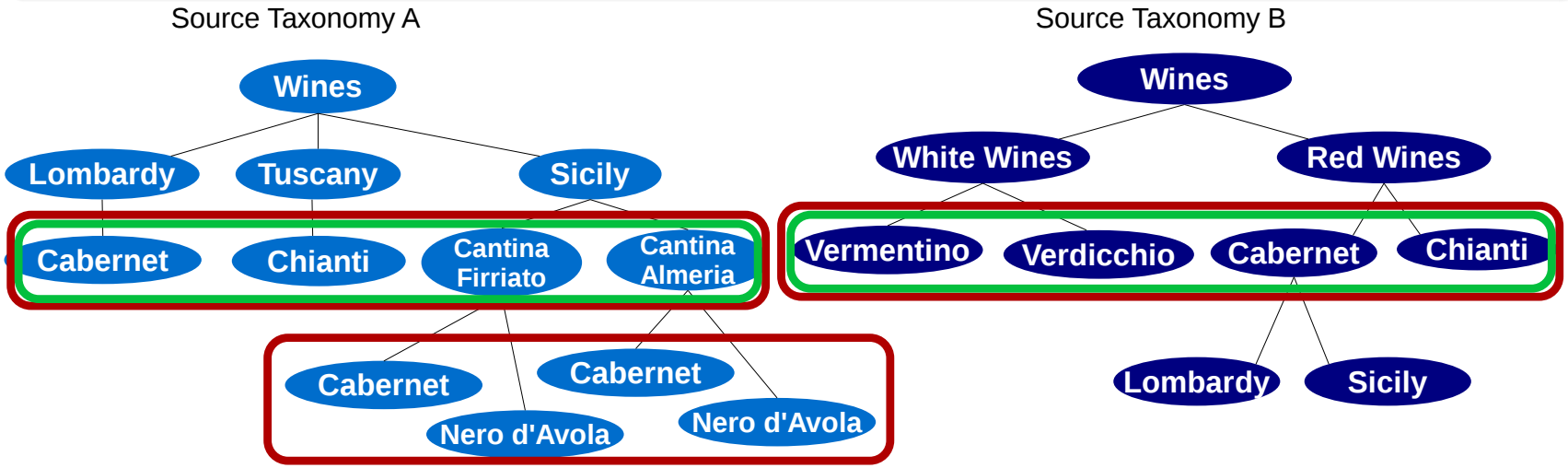


cabernet

Taxonomy Layer Distance

$$\mathrm{TLD}(s_1, s_2) = 1 - \frac{|L_{s_1} \cap L_{s_2}|}{|L_{s_1} \cup L_{s_2}|}$$

Jaccard Distance between the two sets of taxonomy layers where two categories s_1 and s_2 occur

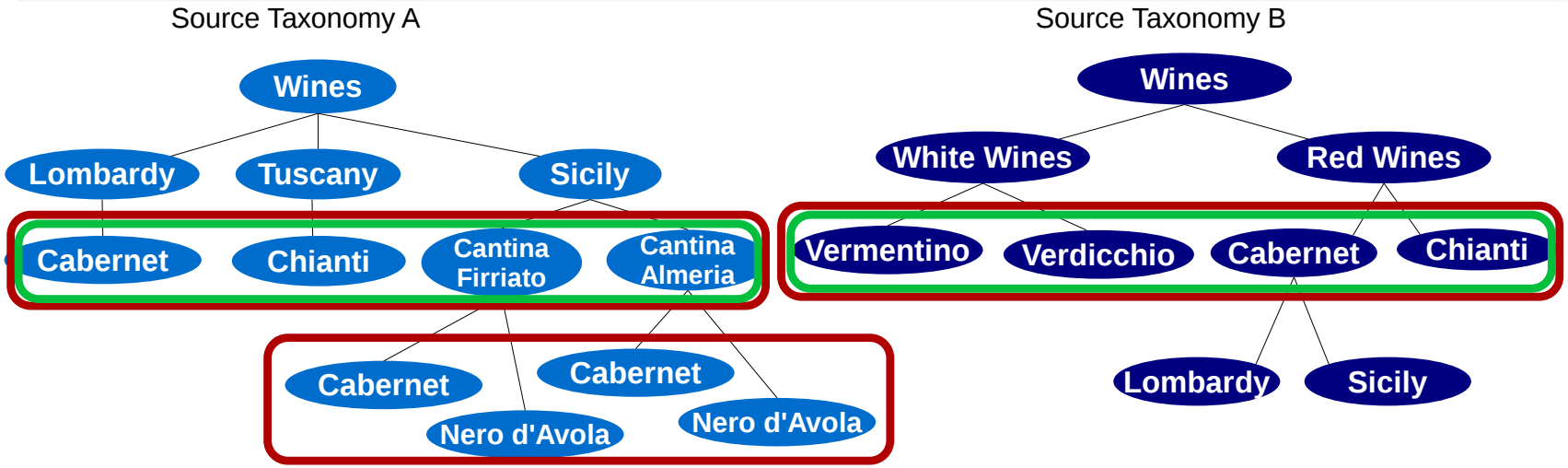


cabernet *chianti*

Taxonomy Layer Distance

$$\mathrm{TLD}(s_1, s_2) = 1 - \frac{|L_{s_1} \cap L_{s_2}|}{|L_{s_1} \cup L_{s_2}|}$$

Jaccard Distance between the two sets of taxonomy layers where two categories s_1 and s_2 occur



cabernet *chianti*

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

Example of Extracted Facets

LC	F_1^g = {Wine, Red Wine, White Wine, . . . , Piedmont, Lombardy, . . . , Sicily, Donnafugata, Cusumano, . . . , Alessandro di Camporeale, . . . , France} (98)
WP	F_1^g = {Wine, Red Wine, White Wine, . . . , Piedmont, Lombardy, . . . , Sicily, Donnafugata, Cusumano, . . . , France} (100)
TLD	F_1^g = { Piedmont, Tuscany, Sicily, , . . . , France } (14) F_2^g = { Red, White, Rosé } (3) F_3^g = { Red Wine, White Wine, Rosé Wine } (3) F_4^g = { Moscato, Chardonnay, . . . , Merlot } (13) F_5^g = { Tuscany Wine, Sicily Wine} (2) F_6^g = { Donnafugata, Cusumano, . . . , Principi di Butera } (27)
Gold Standard	F_1^g = { Piedmont, Lombardy, . . . , Sicily } (21) F_2^g = { Red Wine, White Wine, . . . , Rosé Wine } (14) F_3^g = { Donnafugata, Cusumano, . . . , Alessandro di Camporeale} (12)

- **LC**: Leacock and Chodorow similarity [Leacock and Chodorow 1998]
- **WP**: Wu and Palmer similarity [Wu and Palmer 1994]

Are we really done?

Suggested Facets

+

vino bianco
vino rosato
vino rosso

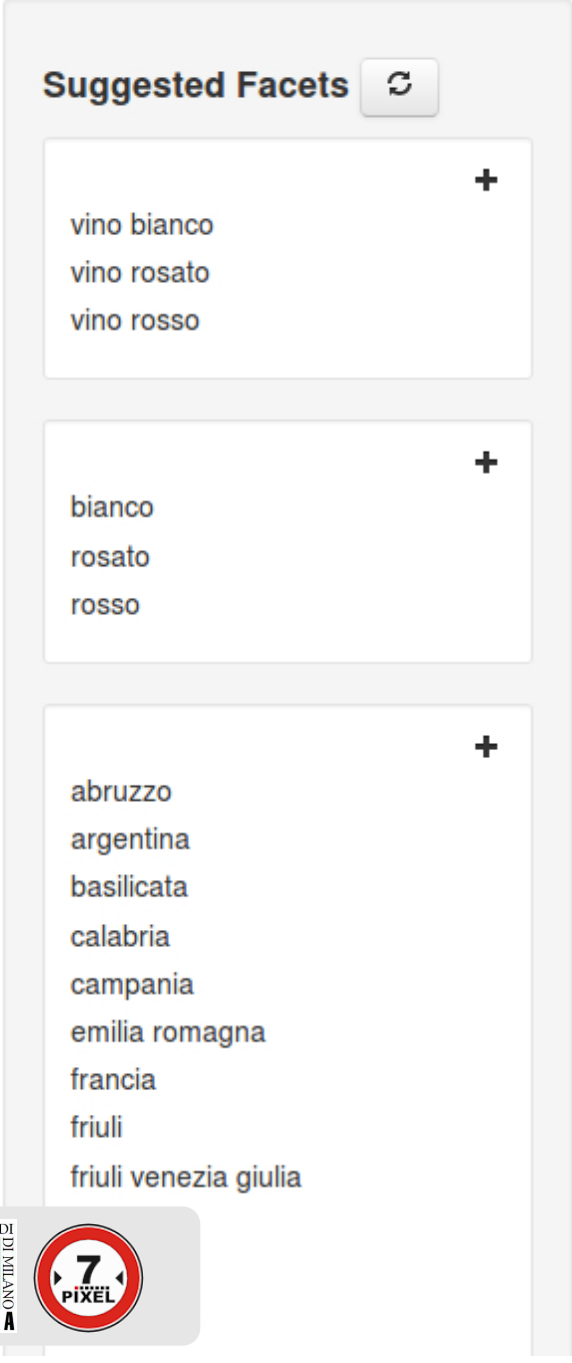
+

bianco
rosato
rosso

+

abruzzo
argentina
basilicata
calabria
campania
emilia romagna
francia
friuli
friuli venezia giulia

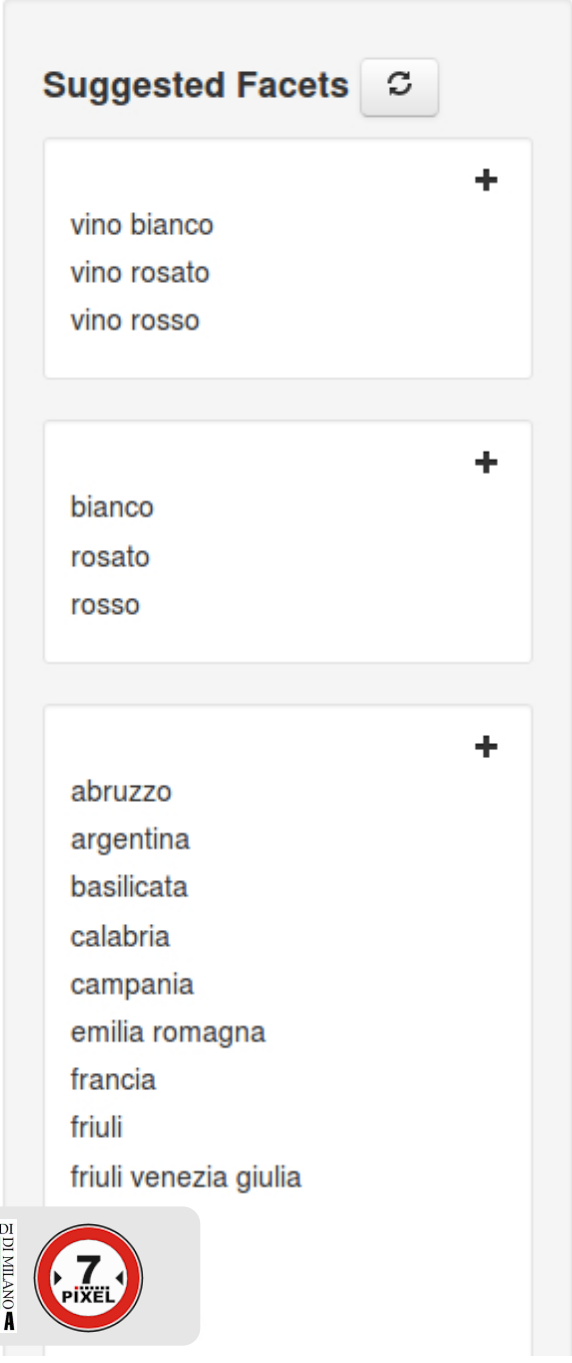
Are we really done?



so far

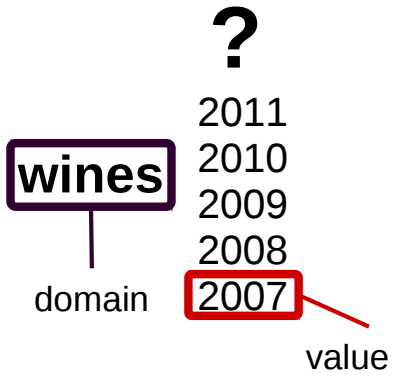
- only facet values were extracted

Are we really done?

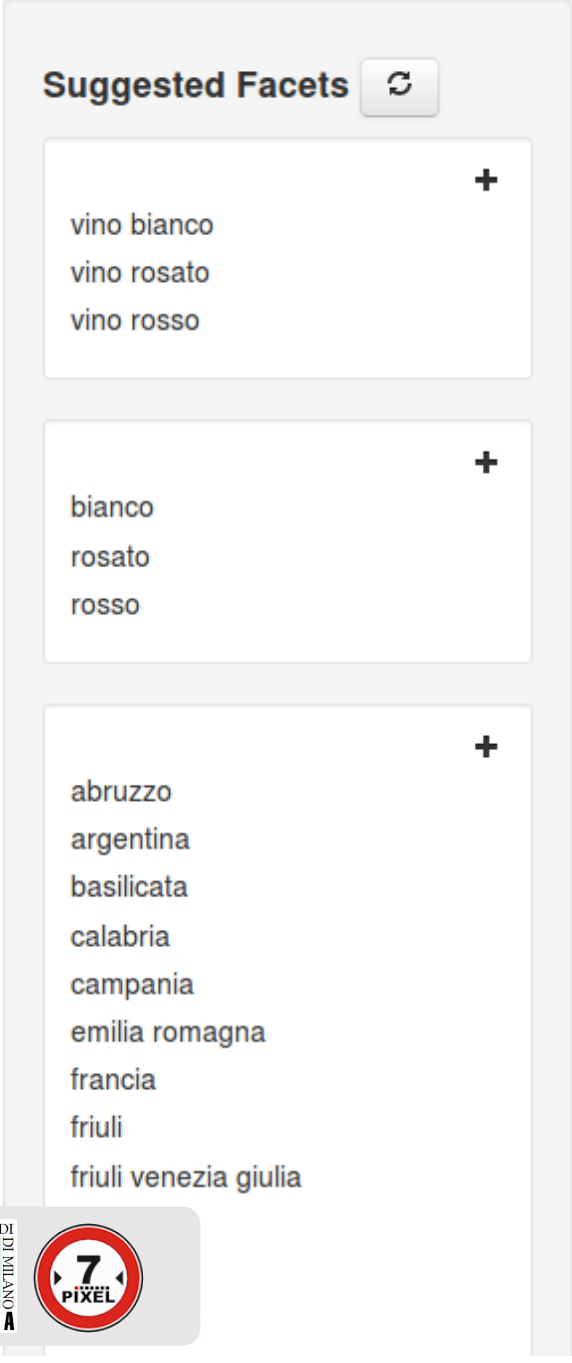


so far

- only facet values were extracted
- what about facet **roles**?

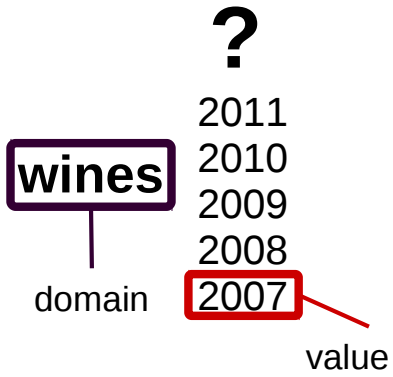


Are we really done?



so far

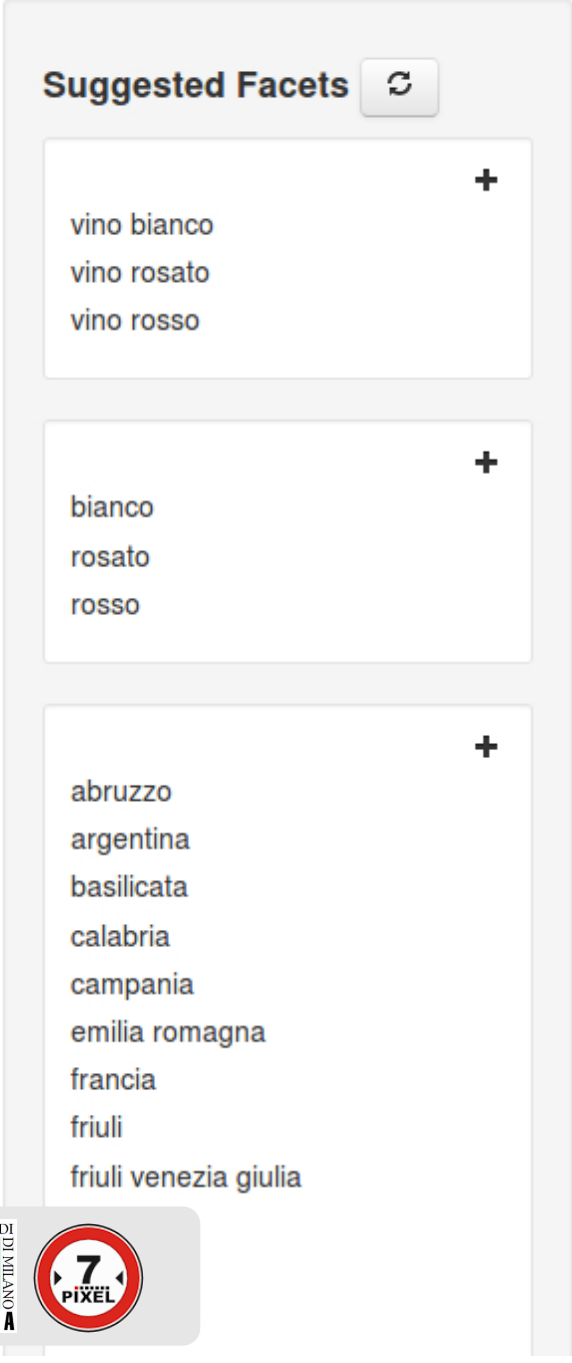
- only facet values were extracted
- what about facet **roles**?



recurrent problem in different domains

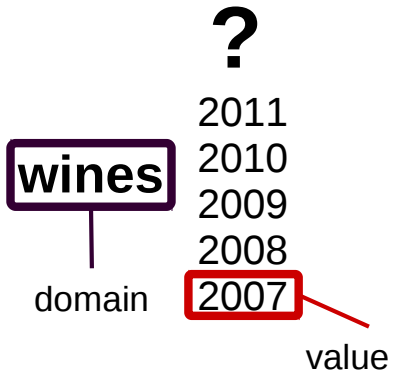
- only partially tackled by previous work on facet extraction [Dou et al. 2011, Kong and Allan 2013] ...

Are we really done?



so far

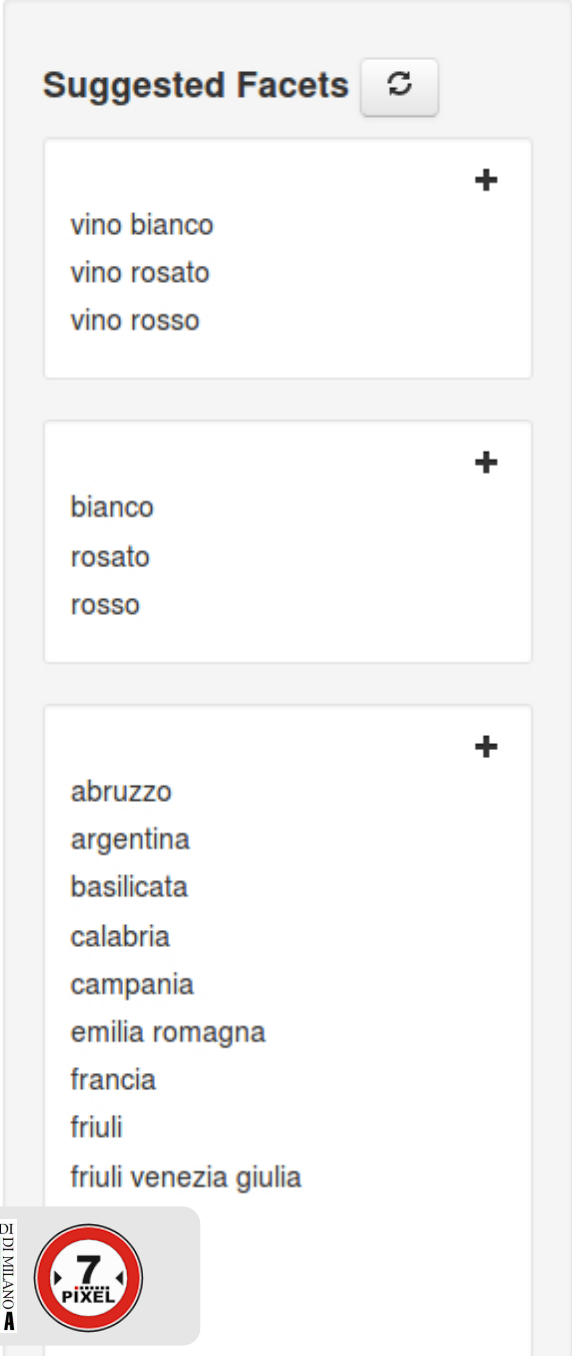
- only facet values were extracted
- what about facet **roles**?



recurrent problem in different domains

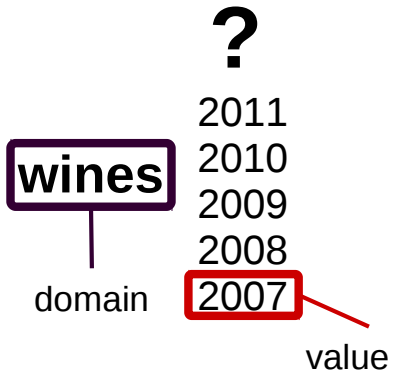
- only partially tackled by previous work on facet extraction [Dou et al. 2011, Kong and Allan 2013] ...
- interpretation of results from clustering algorithms [Carmel et al. 2009] ...

Are we really done?



so far

- only facet values were extracted
- what about facet **roles**?



recurrent problem in different domains

- only partially tackled by previous work on facet extraction [Dou et al. 2011, Kong and Allan 2013] ...
- interpretation of results from clustering algorithms [Carmel et al. 2009] ...
- Web table interpretation and annotation [Limaye et al. 2010, Venetis et al. 2011] ...

HammerSky 2010 Red Handed	Red	**2010**	U.S.A.
2009 Tignanello, Tuscany 750 mL	Red	**2009**	Italy
2012 Paulinshof Urstueck 750 mL	White	**2012**	Germany
2012 Sobremesa Vineyards VRM	White	**2012**	Argentina

Facet Role Linking

ongoing joint work with Prof. Isabel Cruz, Advis Lab - UIC

goal

link a **facet** F to a suitable representation of the role that **facet values** play in the characterization of instances from a specific **facet domain** D

Facet Role Linking

ongoing joint work with Prof. Isabel Cruz, Advis Lab - UIC

goal

link a **facet** F to a suitable representation of the role that **facet values** play in the characterization of instances from a specific **facet domain** D

again, **ambiguity** is challenging:

2011
2012
2013
...

Facet Role Linking

ongoing joint work with Prof. Isabel Cruz, Advis Lab - UIC

goal

link a **facet** F to a suitable representation of the role that **facet values** play in the characterization of instances from a specific **facet domain** D

again, **ambiguity** is challenging:

wines

vintage
2011
2012
2013
...

Facet Role Linking

ongoing joint work with Prof. Isabel Cruz, Advis Lab - UIC

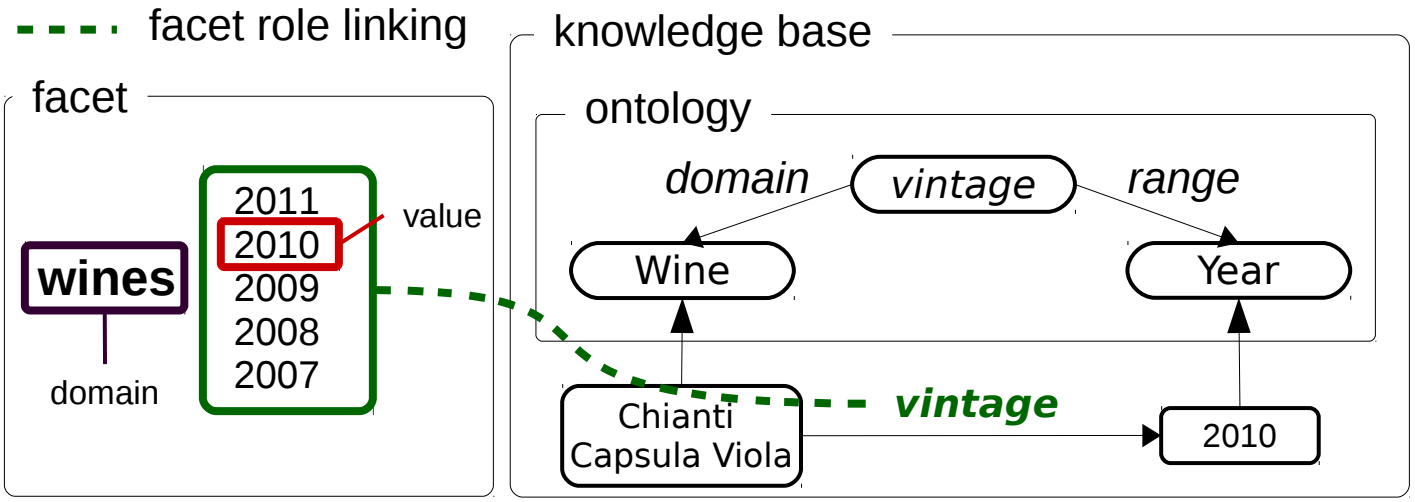
goal

link a **facet** F to a suitable representation of the role that **facet values** play in the characterization of instances from a specific **facet domain** D

again, **ambiguity** is challenging:

	release year
music albums	2011
	2012
	2013
	...

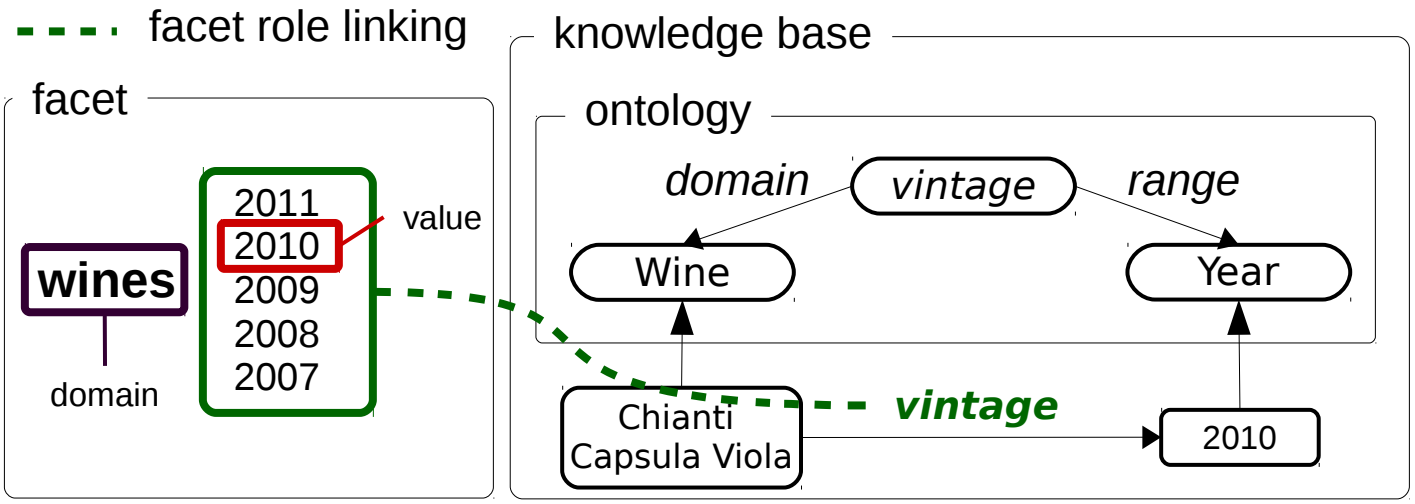
Facet Role Linking



intuitions adapted from Web table annotation approaches

- **properties** from existing ontologies provide the semantics that we are looking for
- a **knowledge base** can provide evidence of certain properties holding between entities from domains similar to the facet domain and facet values

Facet Role Linking



bonus points

- machine readable semantics
- facets published on the Web can be annotated with it (e.g., RDFa)

Approach Overview

basic idea

given a facet F with $f_1, \ldots, f_n$ facet values and a facet domain D

basic idea

given a facet F with $f_1, \ldots, f_n$ facet values and a facet domain D

select a set P of properties from triples $\langle s, p, o \rangle$ such that

- the facet domain D matches the type(s) of s
- one or more facet values $f_1, \ldots, f_n \in F$ match o (entity or literal)

Approach Overview

basic idea

given a facet F with $f_1, \ldots, f_n$ facet values and a facet domain D

select a set P of properties from triples $\langle s, p, o \rangle$ such that

- the facet domain D matches the type(s) of s
- one or more facet values $f_1, \ldots, f_n \in F$ match o (entity or literal)

then, rank $p \in P$ according to several criteria

- facet values coverage
- weighted frequency w.r.t. D
- specificity w.r.t. D

Future Work and Interesting Directions

facet annotation

- what if a facet role is already provided (in natual language)?

  syntactic-based property alignment techniques [Cheatham and Hitzler 2014]

  annotation of facets already published on the Web

Future Work and Interesting Directions

facet annotation

- what if a facet role is already provided (in natual language)?

  syntactic-based property alignment techniques [Cheatham and Hitzler 2014]

  annotation of facets already published on the Web

extension to Web table annotation

- apply the facet annotation technique to Web Table columns

  in literature, little study of relation annotation compared to type/entity

Future Work and Interesting Directions

extension to property aligment in LOD

- incorporate insights from facet annotation into property aligment techniques for LOD, by considering their usage

Future Work and Interesting Directions

extension to property aligment in LOD

- incorporate insights from facet annotation into property aligment techniques for LOD, by considering their usage

classification evolution

- study how to evolve facets (and also their values) over time

  all mobile phones have a digital camera: that facet is not important anymore

  a new operating system for mobile phones is released

Thank You

Questions?

riccardo.porrini@disco.unimib.it
http://rporrini.info

References

[Taylor 2004] A.G. Taylor. Wynars introduction to cataloging and classification. *Libraries Unlimited*, 2004

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

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

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

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

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

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

[Carmel et al. 2009] D. Carmel, H. Roitman and N. Zwerdling. Enhancing Cluster Labelling Using Wikipedia. *SIGIR*, 2009

[Limaye et al. 2010] G. Limaye, S. Sarawagy and S. Chakrabarti. Annotating and Searching Web Tables Using Entities, Types and Relationship. *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. *VLDB*, 2011

[Cheatham and Hitzler 2014] M. Cheatham and P. Hitzler. The Properties of Property Alignment. *OM*, 2014