Automatic Key Selection for Data Linking*

Manel Achichi, Mohamed Ben Ellefi, Danai Symeonidou, Konstantin Todorov

LIRMM / University of Montpellier, France

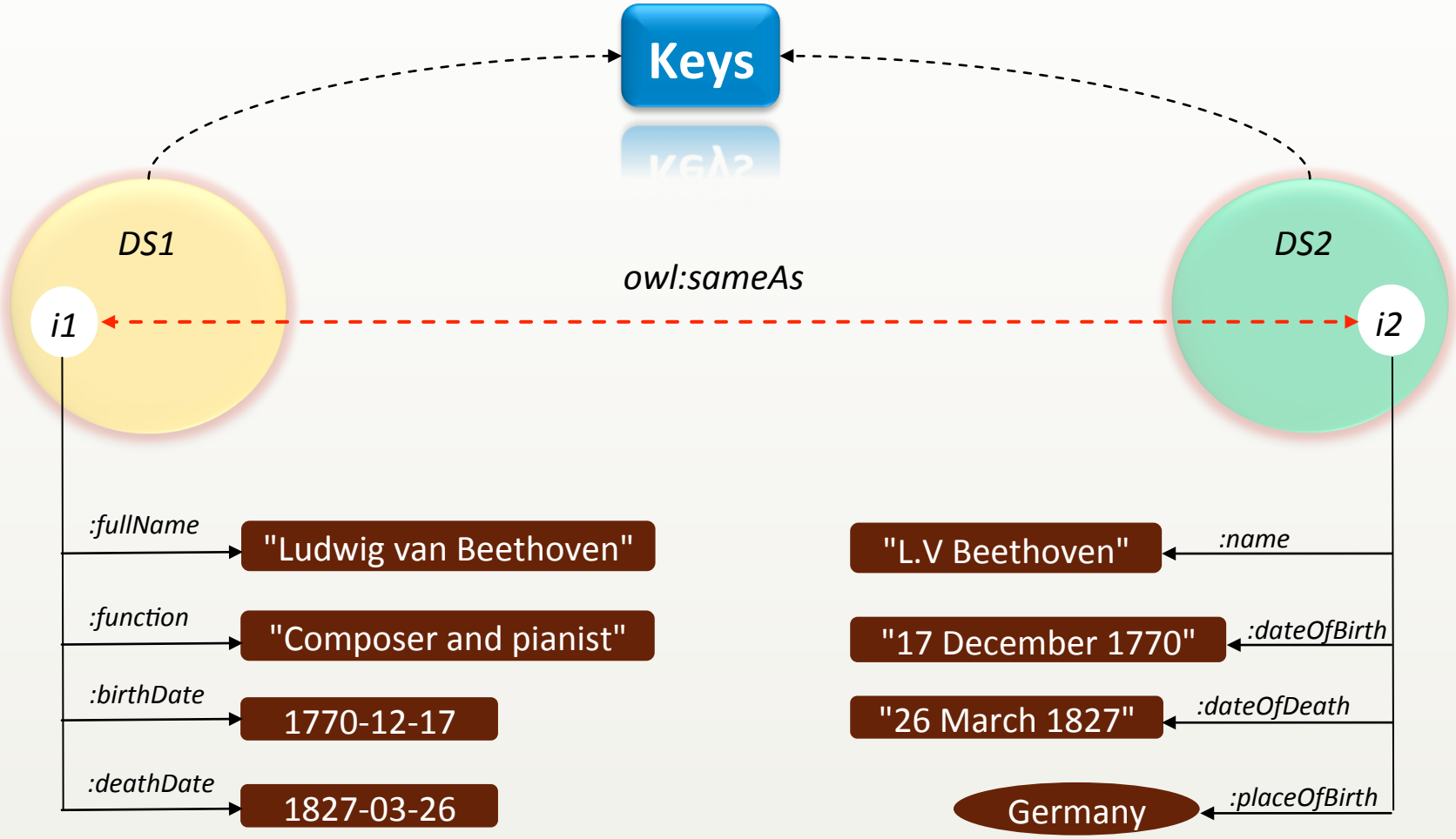
INRA, MISTEA Joint Research Unit, UMR729, F-34060 Montpellier, France

IN-OVIVE, 02/12/16

**Presented in the conference EKAW 2016*

Data Linking

Data Linking



Problem Statement

- A *key discovery* tool generates a large number of keys

Which of the keys achieves the best *performance* respect to data linking results?

***Main problem*:** no approach provides a strategy to rank the discovered keys, by considering their effectiveness for the matching task

Key Discovery

	FirstName	LastName	Nationality
p1	"Mary"	"Tompson"	**British**
p2	"John"	"Tompson"	American
p3	"Vincent"	"Dupont"	French
p4	"Kate"	"Martin"	**British, Greek**
p5	"Michael"	"Kinard"	USA

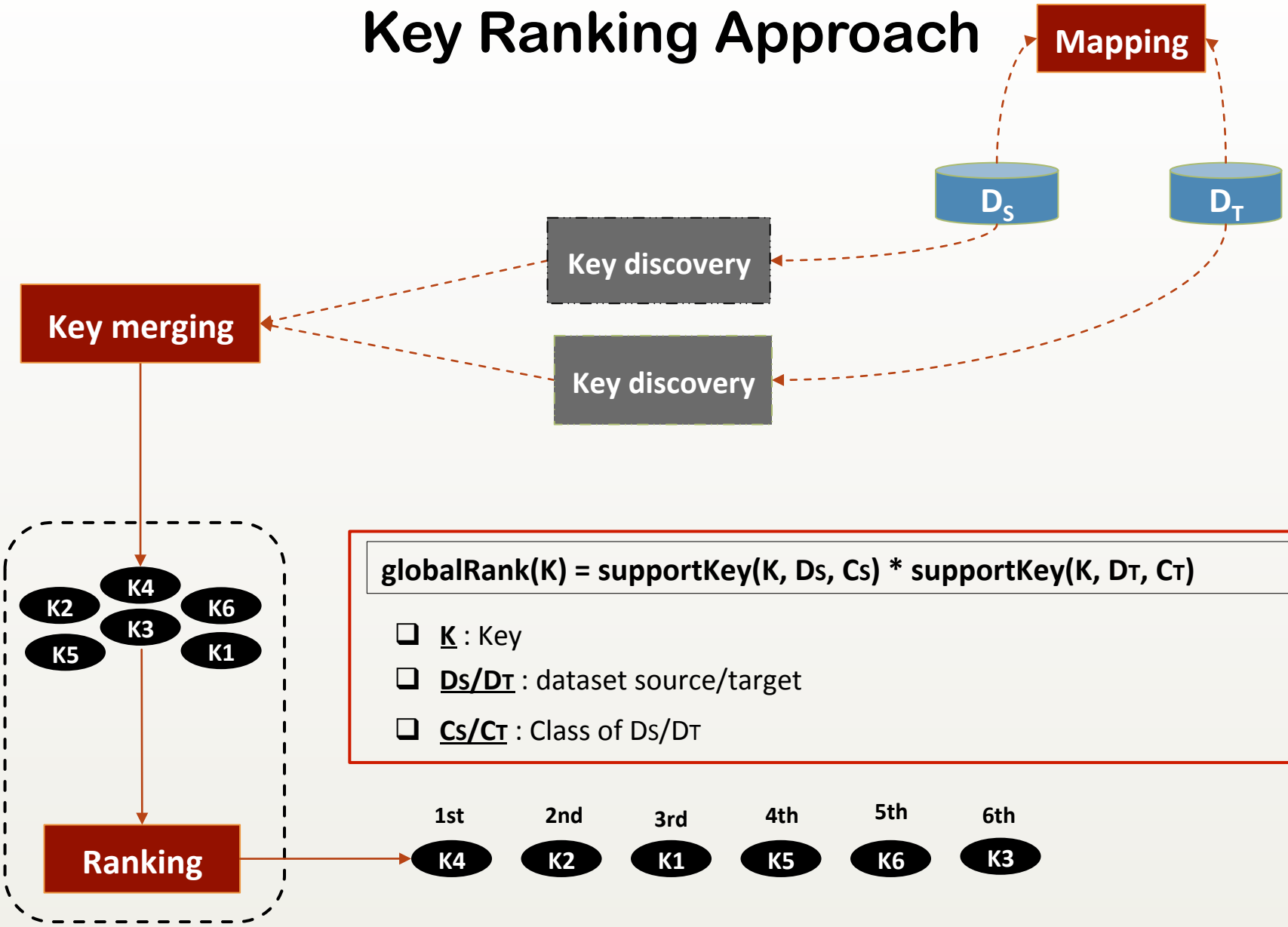
SAKey (Symeonidou et al., 2014)
- Open World Assumption (OWA)
- Discovers *n-almost keys*

ROCKER (Soru et al., 2015)
- Closed World Assumption (CWA)
- Discovers *keys*

SAKey keys: {FirstName},
{LastName, Nationality}

ROCKER keys: {FirstName},
{Nationality}

Key Ranking Approach



globalRank(K) = supportKey(K, Ds, Cs) * supportKey(K, DT, CT)

- **K** : Key
- **Ds/DT** : dataset source/target
- **Cs/CT** : Class of Ds/DT

Key Ranking Approach: Example

- DS = 300 instances
- DT = 100 instances
- Ki / Kj = 2 merged keys

K_i K_j

- supportKey(Ki, DS, Cs) = 160/300
- supportKey(Ki, DT, CT) = 40/100
- globalRank(Ki) = 0.21

- supportKey(Kj, DS, Cs) = 110/300
- supportKey(Kj, DT, CT) = 90/100
- globalRank(Kj) = **0.33**

The key **Kj** gives leads to better data linking results than the key **Ki**

Evaluation

Goal

- Evaluate the effectiveness of the ranking function inside the link discovery task

Tools

- Keys identificator: SAKey and ROCKER
- Linking tool: SILK (Volz et al., 2009)

Datasets

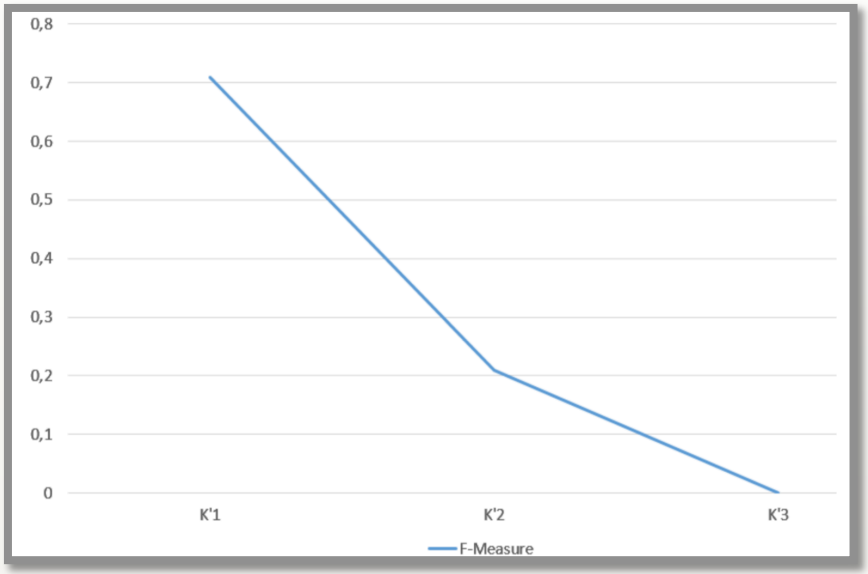
- Real world test cases from DOREMUS (IM@OAEI2016)
- Synthetic benchmark (IM@OAEI2010)

Evaluation metrics

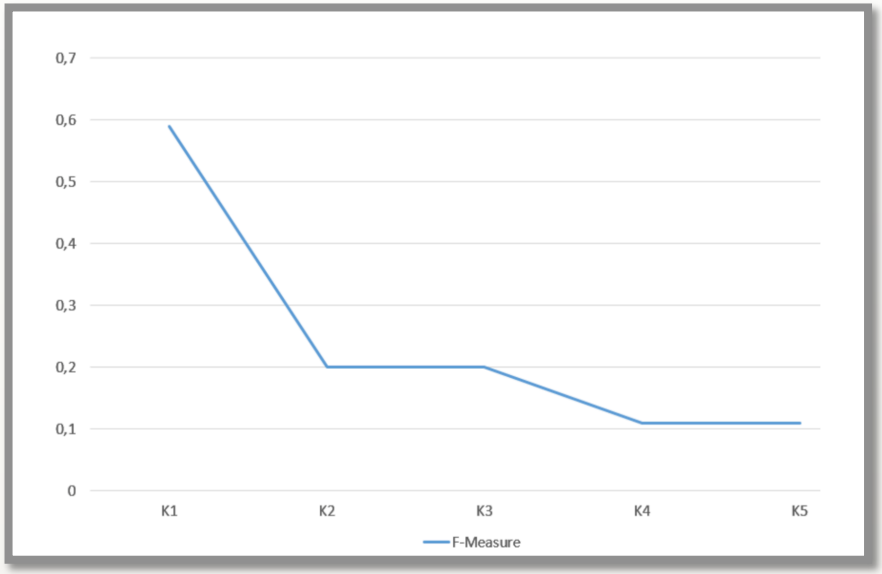
- F-Measure, Precision and Recall

[1] *http://www.doremus.org/*

Evaluation Results on DOREMUS-OAEI2016

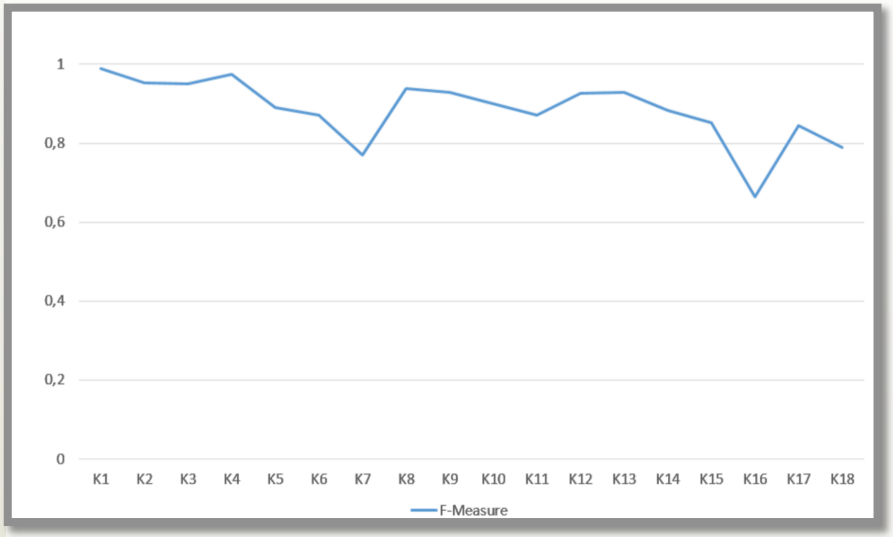


SAKey

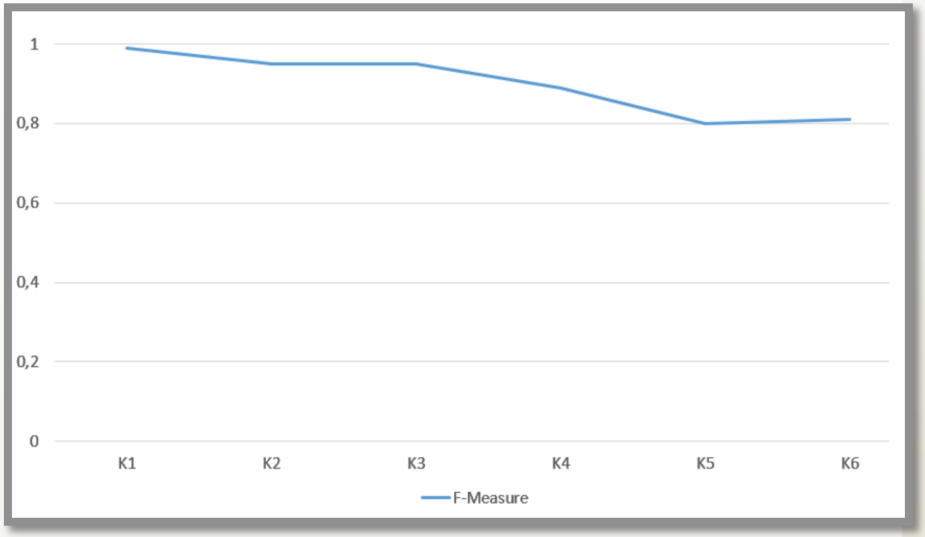


ROCKER

Evaluation Results on OAEI-Person



SAKey



ROCKER

Top Ranked Keys Complementarity

- The top ranked keys with low recall: ***heterogeneous nature of data***
- Can the combined ***k*** top-ranked keys improve the linking scores ?

	SAKey						**ROCKER**			
	DS1			DS2			DS1			DS2
	F	P	R	F	P	R	F	P	R	No merged key has been identified.
K1	0.12	0.12	0.11	0.5	0.75	0.37	0.59	0.8	0.47	
K2	0.71	0.9	0.58	0.48	0.7	0.37	0.2	0.66	0.11	
K3	0.52	1	0.35	0.37	0.56	0.28	0.2	0.66	0.11	
K1+K2+K3	0.54	0.44	0.7	0.51	0.63	0.43	0.62	0.75	0.52	

Summary

- **WHAT?** Discovering the best keys for data linking
- **HOW?** Providing a strategy to rank the keys
  - Key *discovery*
  - Key *merging*
  - Key *ranking*
- **WHY?** Bridging the gap between key discovery and data linking approaches

Future Work

- Improve the ranking function
- Define other criteria than the support

Thank you!

Manel Achichi[1], Mohamed Ben Ellefi[1], Danai Symeonidou[2] and Konstantin Todorov[1]

{achichi,benellefi,todorov}@lirmm.fr, danai.symeonidou@supagro.inra.fr

[1]LIRMM / University of Montpellier, France.

[2]INRA, MISTEA Joint Research Unit, UMR729, F-34060 Montpellier, France.

References

- Symeonidou, D., Armant, V., Pernelle, N., and Saïs, F. (2014, October). Sakey: Scalable almost key discovery in rdf data.. In International Semantic Web Conference (pp. 33-49). Springer International Publishing.

- Soru, T., Marx, E., and Ngonga Ngomo, A. C. (2015, May). ROCKER: A refinement operator for key discovery. In Proceedings of the 24th International Conference onWorldWideWeb (pp. 1025-1033). ACM.

- Volz, J., Bizer, C., Gaedke, M., and Kobilarov, G. (2009). Silk-A Link Discovery Framework for the Web of Data.. LDOW, 538.