Karlsruhe Institute of Technology

AIFB

CrowdMAP: Crowdsourcing Ontology Alignment with Microtasks

Cristina Sarasua, Elena Simperl, Natasha Noy

Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

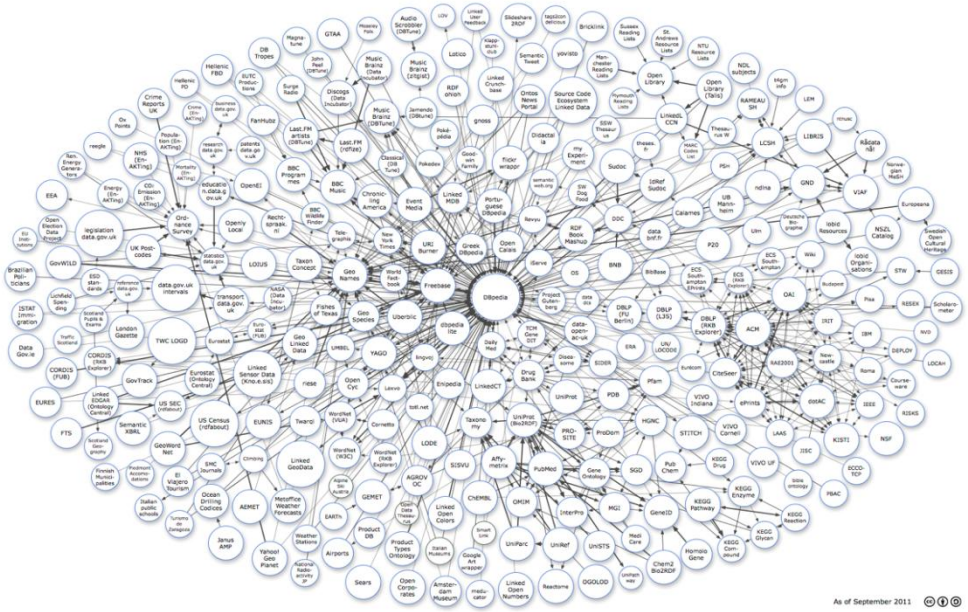
Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

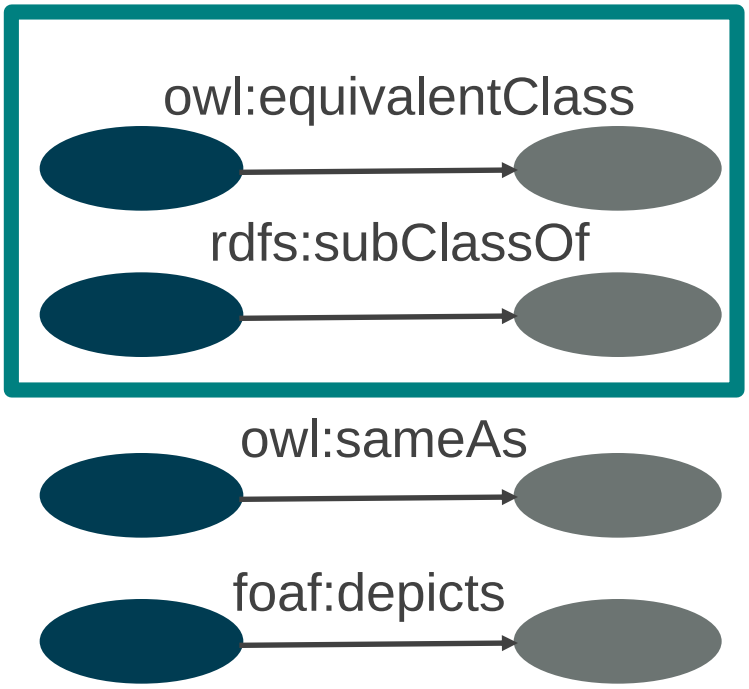
STANFORD
SCHOOL OF MEDICINE

Why?

Motivation

- Interlinks – typed links between resources of different data sources – are the key enablers of data integration





"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

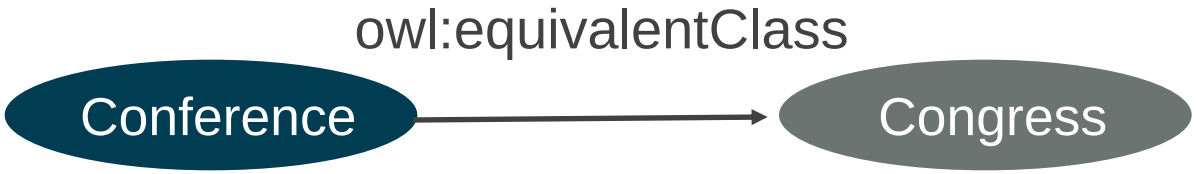
"64.41 % data sources use proprietary vocabulary terms and 7.89 % data sources that use proprietary terms provide mappings to other vocabularies for their terms." ***Bizer et al., State of the LOD Cloud, 2011***

The problem

- After a decade …

PROMPT	MAFRA	FCA- MERGE	Chimaera	ONION
AROMA	YAM++	AgrMaker	CIDER	LogMap
BLOOMS	MeLinDa	HCM		

- Automatic techniques have problems in discovering some mappings, and some of the discovered mappings are not correct



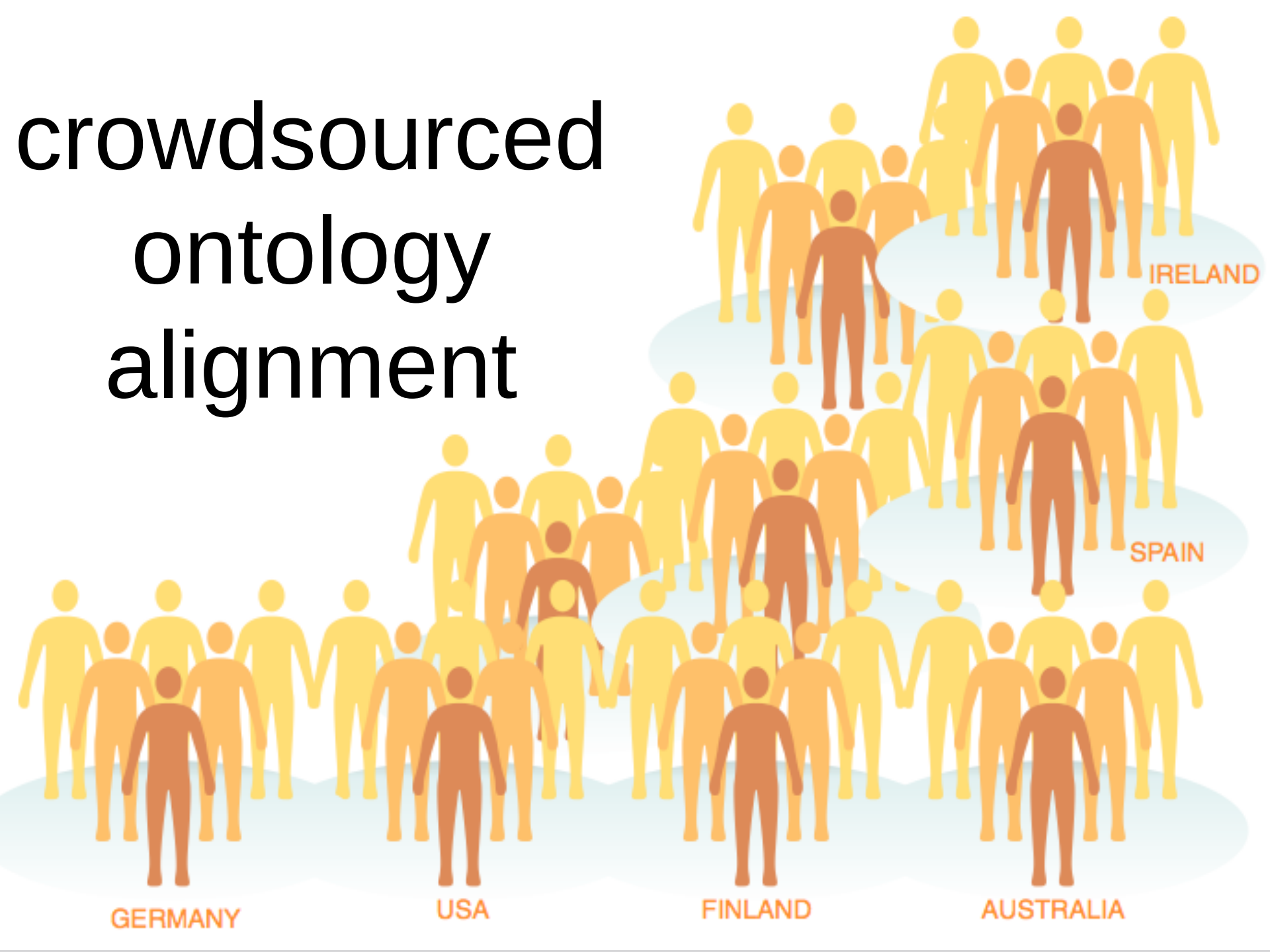
KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)

How can we use human intervention?

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)

Goal

crowdsourced ontology alignment



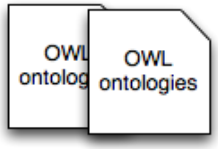
The concept

- Extend current technology with human computation
- Alignment microtasks
- Pay $ to contributors
- Publish at marketplace

http://www.mturk.com

Scenario

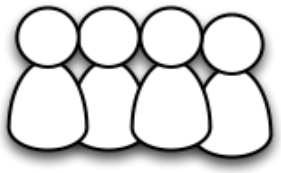
Henry
the ontology engineer



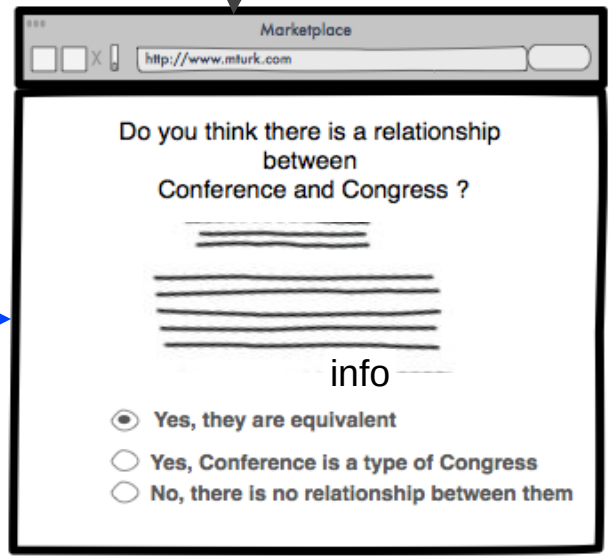
ontology alignment

candidate mappings

CrowdMAP



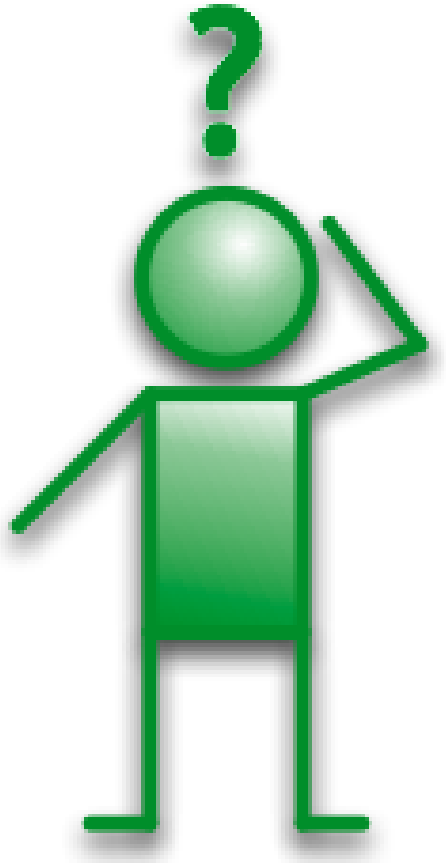
Crowd



Mapping

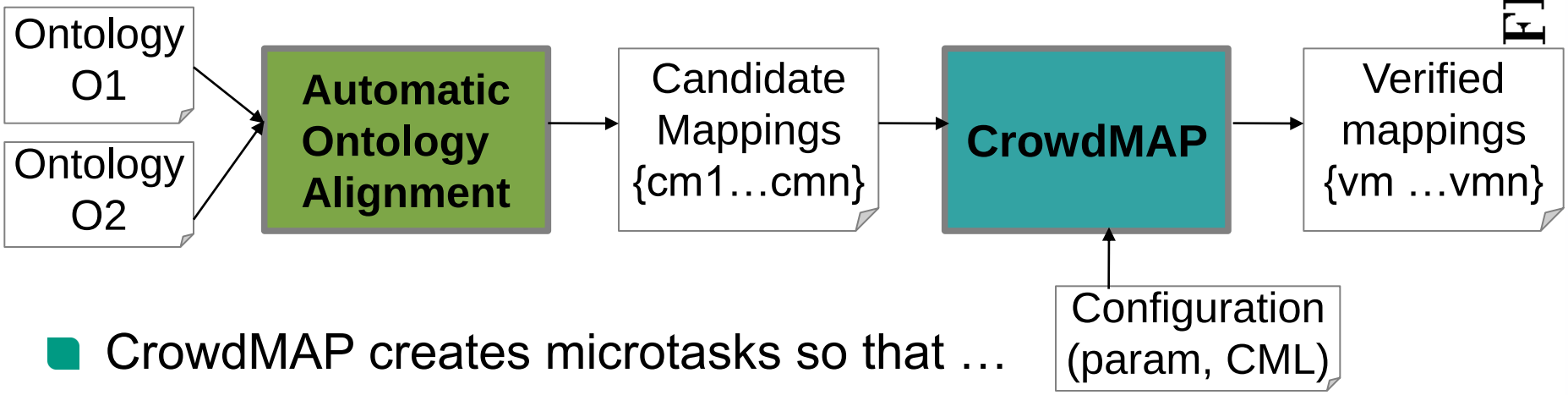
m = {e1, e2, r, conf}

Research questions



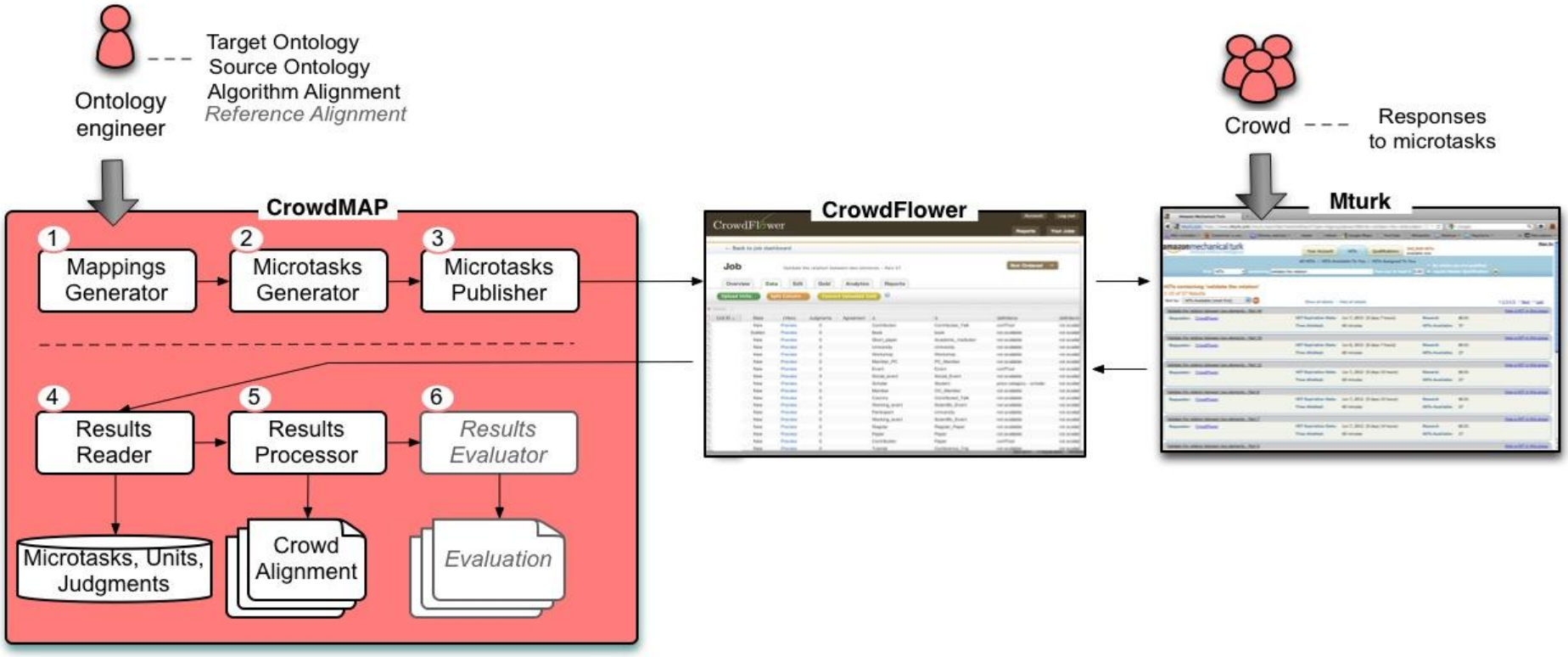
- **RQ1**: Is ontology alignment amenable to microtask crowdsourcing?
- **RQ2**: How does such a human-driven approach compare to (semi-)automatic techniques, and can improve their results?

The CrowdMAP workflow



- CrowdMAP creates microtasks so that …
  - Can ask workers to validate a given mapping relationship (=) or to identify it from a list (=, >=, <=)
  - If candidate mappings only = then 3 workers are asked to verify a mapping
  - If else candidate mappings contain =, >=, <= then 7 workers are asked to select a relationship for each mapping

The CrowdMAP architecture



Microtask UI design

Concept A: Misc
Definition (English): Use this type when nothing else fits.

Misc *is a kind of:* Reference

Other elements that are of kind Reference: 'Academic' 'Informal' 'MotionPicture'

Concept B: Misc.
Definition (English): Use this type when nothing else fits.

Misc. *is a kind of:* REFERENCE

Other elements that are of kind REFERENCE: 'Book' 'Academic' 'Motion_picture'

Is Concept A the same as Concept B? (required)

- ○ yes
- ○ no

Please select only one of the answers

Select the name of Concept A (required)

- ○ Misc
- ○ Misc.

Please select only one of the answers

How many distinct words are in the name of Concept A? (required)

[ ]

Please write the number in the text box

Microtask UI design

Concept A: Misc
Definition (English): Use this type when nothing else fits.

Misc *is a kind of:* Reference

Other elements that are of kind Reference: 'Academic' 'Informal' 'MotionPicture'

Concept B: Misc.
Definition (English): Use this type when nothing else fits.

Misc. *is a kind of:* REFERENCE

Other elements that are of kind REFERENCE: 'Book' 'Academic' 'Motion_picture'

Is Concept A the same as Concept B? (required)
- ○ yes
- ○ no

Please select only one of the answers

Select the name of Concept A (required)
- ○ Misc
- ○ Misc.

Please select only one of the answers

How many distinct words are in the name of Concept A? (required)

Please write the number in the text box

Microtask UI design

Concept A: Misc
Definition (English): Use this type when nothing else fits.

Misc *is a kind of:* Reference

Other elements that are of kind Reference: 'Academic' 'Informal' 'MotionPicture'

Concept B: Misc.
Definition (English): Use this type when nothing else fits.

Misc. *is a kind of:* REFERENCE

Other elements that are of kind REFERENCE: 'Book' 'Academic' 'Motion_picture'

Is Concept A the same as Concept B? (required)

- ○ yes
- ○ no

Please select only one of the answers

Do you see any connection between Concept A and Concept B? (required)

- ○ Concept A is the same as Concept B
- ○ Concept A is a kind of Concept B
- ○ Concept B is a kind of Concept A
- ○ There is no relation between Concept A and Concept B

Please select only one of the answers

Select the name of Concept A (required)

- ○ Misc
- ○ Misc.

Please select only one of the answers

How many distinct words are in the name of Concept A? (required)

Please write the number in the text box

Microtask UI design

Concept A: Misc
Definition (English): Use this type when nothing else fits.

Misc *is a kind of:* Reference

Other elements that are of kind Reference: 'Academic' 'Informal' 'MotionPicture'

Concept B: Misc.
Definition (English): Use this type when nothing else fits.

Misc. *is a kind of:* REFERENCE

Other elements that are of kind REFERENCE: 'Book' 'Academic' 'Motion_picture'

Is Concept A the same as Concept B? (required)

- ○ yes
- ○ no

Please select only one of the answers

Select the name of Concept A (required)

- ○ Misc
- ○ Misc.

Please select only one of the answers

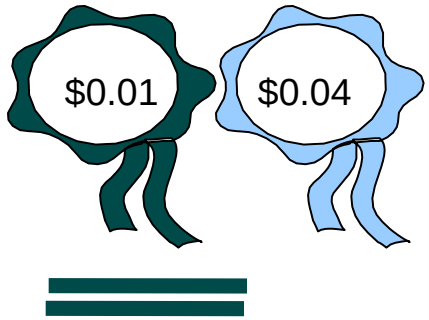
How many distinct words are in the name of Concept A? (required)

[ ]

Please write the number in the text box

Evaluation – CrowdFlower & Mturk setup

- Templates + data units (mappings)
- Set of golden mapping units
- Microtask = 6 mappings + 1 gold mapping
- After_gold = 2 (trusted worker)
- Mturk batches of max 50 HITs
- 50% HITs template X, 50% HITs template X_turned



Evaluation - data

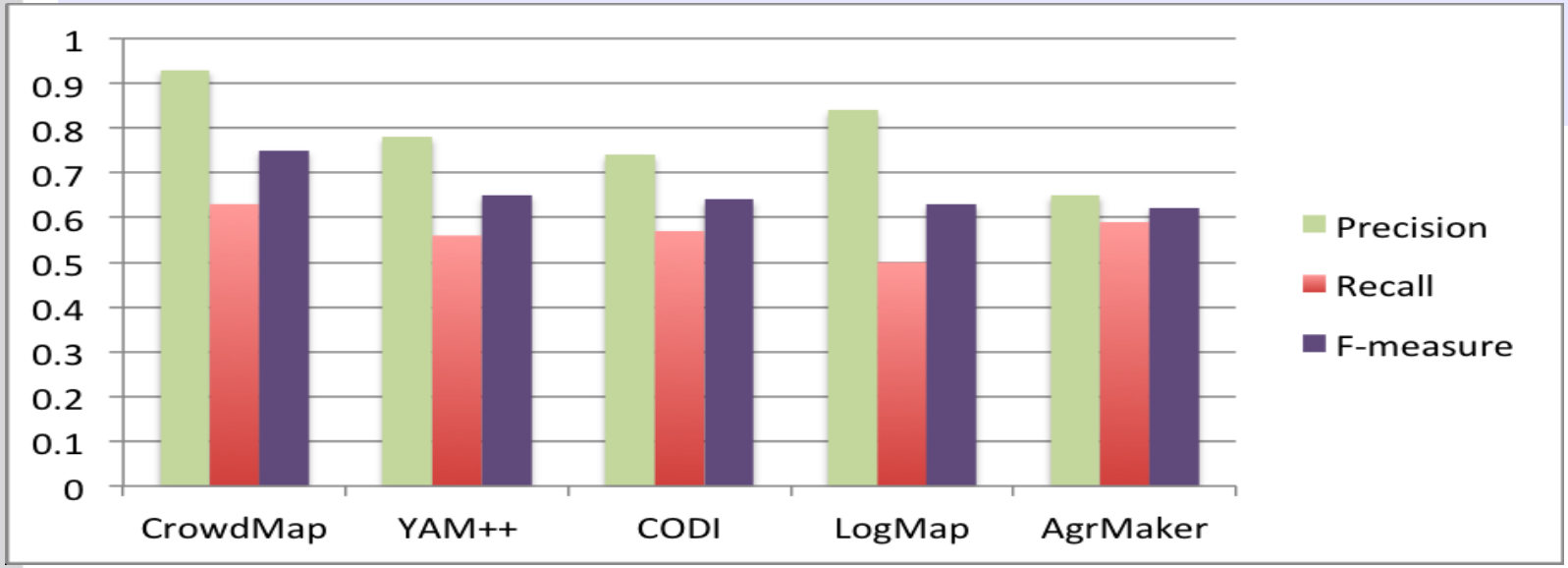
- OAEI 2011 – Conference & oriented matching ontologies
- Experiments – set of mappings
  - **CartP**: cartesian product between classes of both ontologies
  - **Imp**: candidate mappings from the OA algorithm (simulated taking subset of CartP)
  - **100R50P**: correct mappings from reference alignment + same number of incorrect mappings

Evaluation - results

	CartP 301-304	100R50P Edas-lasted	100R50P Ekaw-lasted	100R50P Cmt-Ekaw	100R50P ConfOf - Ekaw	Imp AROMA 301-304
P	0.53	0.8	1.0	1.0	0.93	0.73
R	1.0	0.42	0.8	0.75	0.65	1.0

Evaluation - results

CartP 301-304	100R50P Edas-lasted	100R50P Ekaw-lasted	100R50P Cmt-Ekaw	100R50P ConfOf - Ekaw	Imp AROMA 301-304
					0.73
					1.0



Evaluation – results (2)

- Average number of users per batch of HITs was 11
- Average agreement was 87,40%
- We spent $50 in total, including pilots
- The set of mappings had size ~ 600 / 40 / 20

Evaluation – analysis & lessons learned

- Amenable?

From CartP we see that it is feasible, the crowd gives meaningful answers

- Improvement vs automatic techniques?

We can improve precision without loosing recall

- 2 Step potential CrowdMAP algorithm

Evaluation – analysis & lessons learned (2)

- Small pilots different wording (e.g. equivalent/same;concept A)
- Golden units help avoiding spammers
- The crowd identified missing valid mappings
- CrowdFlower as an alternative to using Mturk directly

Evaluation – analysis & lessons learned (2)

Do you see any connection between Concept A and Concept B? (required)

- ○ Concept A and Concept B are not the same
- ○ Concept A is the same as Concept B

Please select only one of the following possible answers

Select the name of Concept B (required)

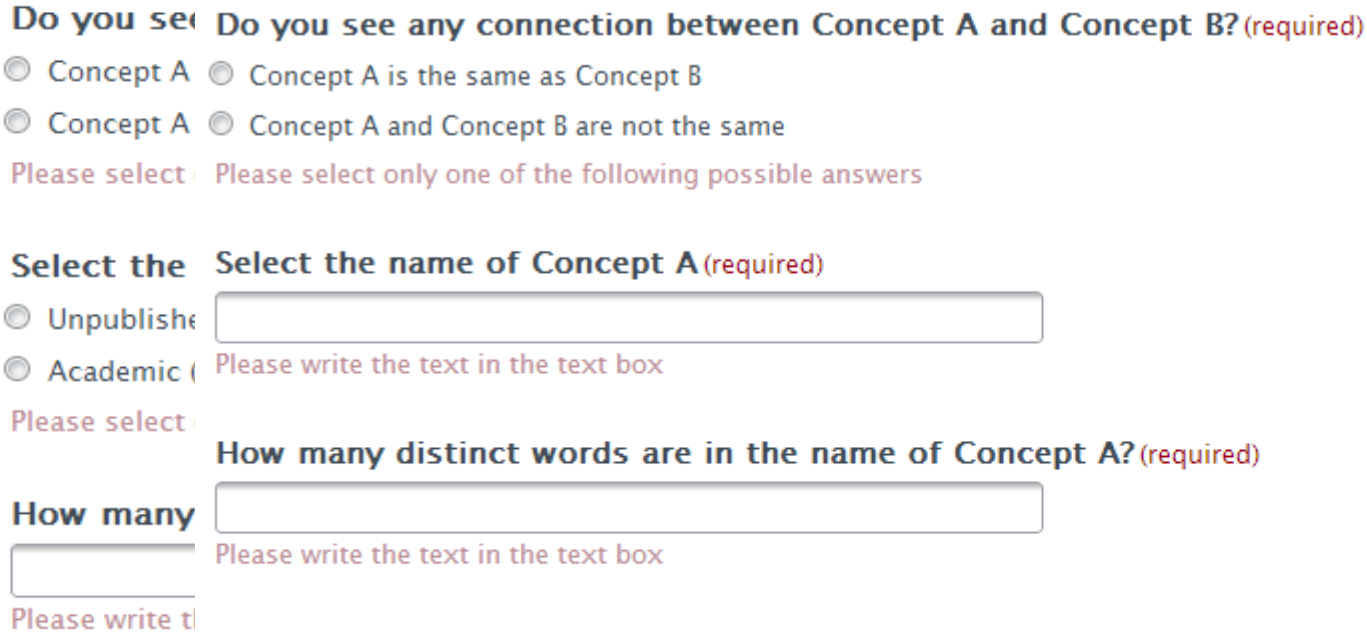
- ○ Unpublished (Concept A)
- ○ Academic (Concept B)

Please select only one of the following possible answers

How many distinct words are in the name of Concept B? (required)

Please write the text in the text box

Evaluation – analysis & lessons learned (2)



Next steps?

Ongoing – future work

- Which is the most useful information?
  - Definition
  - Siblings
  - Superclasses / superproperties
  - Subclasses / subproperties
  - Instances
- Can we apply CrowdMAP to ontologies that require specific domain knowledge?
- How can we extend the approach to data interlinking?

So …

Conclusions

- With crowdsourcing techniques it is easy and cheap to obtain human contributions
- Good results require appropriate management of microtasks and workers
- Crowdsourced ontology alignment seems to be feasible and could improve precision and recall of current automatic techniques

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)

Asking the **crowd** about **mappings** in order to **improve** the perfomance of **current** alignment technology