



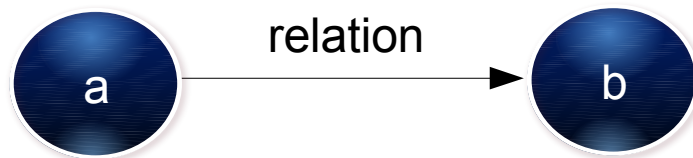
Supporting Data Interlinking in Semantic Libraries with Microtask Crowdsourcing

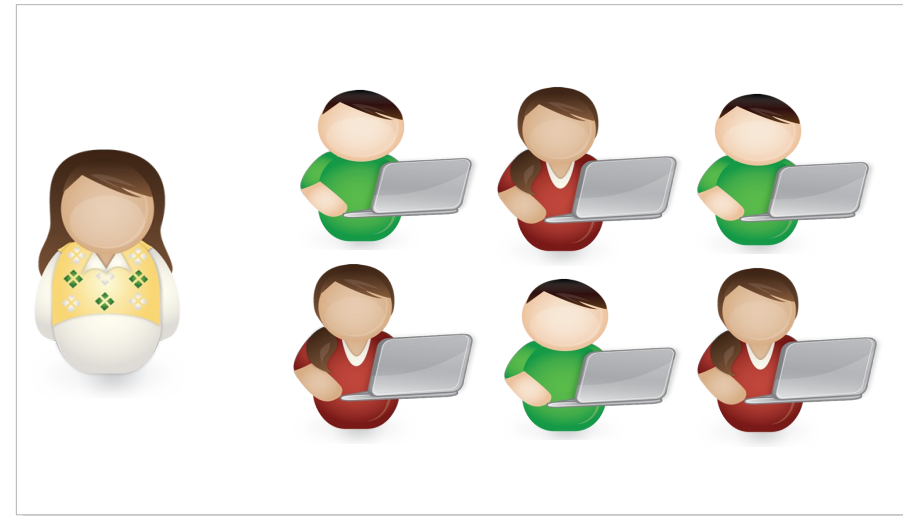
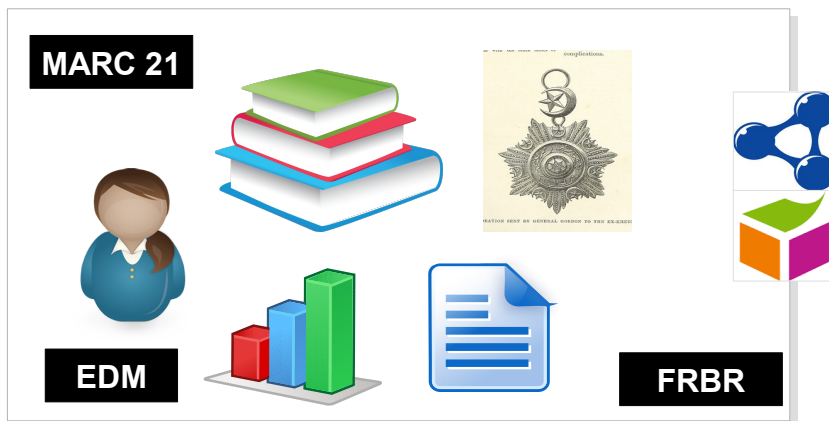
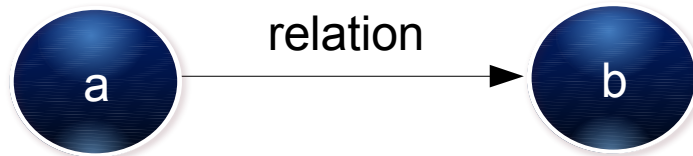
Cristina Sarasua

SWIB 2014, Bonn









Please share your thoughts on interlinking!
<https://etherpad.mozilla.org/4IfZDaTBle>

<https://etherpad.mozilla.org/4IfZDaTBle>



Cross-dataset links

<https://etherpad.mozilla.org/4IfZDaTBle>



$(a,r,b) \mid a \text{ in } D1, b \text{ in } D2$



```
d1:timbl owl:sameAs d2:timbernerslee;  
d1:donostia owl:sameAs d2:sansebastian;
```

<sameAs>
interlinking the Web of Data

```
d1:bjork dc:creator d2:volta;  
d1:Bonn wgs84:location d2:Germany;  
d1:work2012 o:inspiredBy d2:song1900;
```

```
o1:Conference owl:equivalentClass o2:Congress;  
o1:Democracy skos:related o2:Government;  
o1:Publication skos:broader o2:JournalArticle;  
o1:ImpressionistPainting rdfs:subClassOf o2:Painting;
```

Why is interlinking important?

What is known about Berlin?

```
x:berlin owl:sameAs
  dbpedia:Berlin;
  tour:berlin;
x:berlin o:homeOf
  authors:berlin;
x:img09112014
  lode:atPlace geo:brandtor;
```

```
SELECT ?city
WHERE {
  ?city1 gov:population ?pop .
  ?city1 owl:sameAs ?city2 .
  ?city2 unesco:count ?mon .
  FILTER (?pop > 1000000
    ?mon > 50) }
```

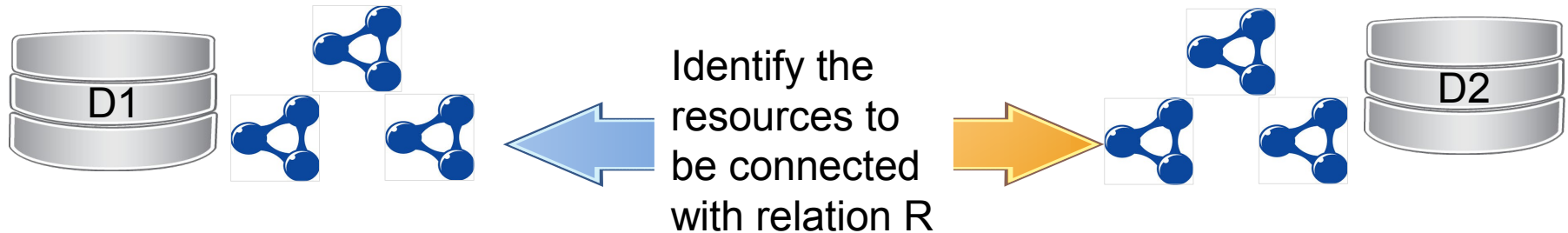
- Enhance the description of local entities
- Richer queries over aggregated data
- Cross-data set browsing

3.2.3 Linking across datasets has begun but requires further effort and coordination

<http://www.w3.org/2005/Incubator/ld/XGR-ld-20111025/>

Generating links

<https://etherpad.mozilla.org/4IfZDaTBle>



```
<Compare metric="jaccard" threshold="0.2">
  <TransformInput function="alphaReduce">
    <TransformInput function="tokenize">
      <Input path="?a/skos:prefLabel[@lang = 'en']">
    </TransformInput>
  </TransformInput>
  <TransformInput function="alphaReduce">
    <TransformInput function="tokenize">
      <Input path="?b/rdfs:label[@lang = 'en']" />
    </TransformInput>
  </TransformInput>
</Compare>
```

Comparison criteria

```
<Output type="file" minConfidence="0.95">
  <Param name="file" value="accepted_links.nt" />
  <Param name="format" value="ntriples" />
</Output>
```

Decision boundary between link and non-link

Silk Workbench

Workspace: Example Editor: drugs Generate Links Reference Links About

Positive Negative Import Reference Links Help

Expand All Collapse All Prev 1 Next Filter:

Source	Target	Confidence	Status	Correct?
http://www4.wiwiss.fu-berlin.de/sider/resource/drugs/4052	http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00...	100.0%	found	✖
Aggregation(max) 100.0% <ul style="list-style-type: none"> Comparison(levenshteinDistance) 100.0% <ul style="list-style-type: none"> Input ?skos:label meloxicam Input ?bdrfs:label meloxicam Comparison(levenshteinDistance) 0.0% <ul style="list-style-type: none"> Input ?skos:label meloxicam Input ?bdrugbanksynonym meloxicamum [latin] 				
http://www4.wiwiss.fu-berlin.de/sider/resource/drugs/51577	http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00...	100.0%	found	✖
http://www4.wiwiss.fu-berlin.de/sider/resource/drugs/16231	http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00...	100.0%	found	✖

Picture:
https://www.assembla.com/spaces/silk/wiki/Managing_Reference_Links

He is already busy



He is already busy

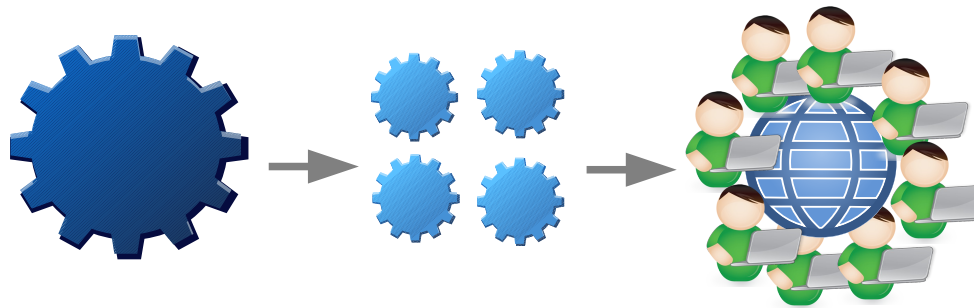
**... but still would like
correct and useful links**

Crowdsourced Interlinking

Crowdsourcing

*“Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an **open call**”*

Jeff Howe, 2006



Scalable

Fast



Microtask crowdsourcing



- E.g. tweet sentiment analysis
- Seconds, reward cents
- Crowd workers register with simple profile, limited filtering

Macrotask crowdsourcing



- E.g. writing an E-Book
- Months, \$30per hour / hundreds or thousands of dollars
- Freelancers recruitment, interviews

Contest-based crowdsourcing



- E.g. NLP algorithm for a particular challenging scenario
- Months, up to thousands of dollars
- Final evaluation and winner selection

Citizen Science



- E.g. classify galaxies in pictures
- seconds/minutes, no money
- Open to everyone

All HITS

1-10 of 2136 Results

Sort by: HITS Available (most first) GO

[Show all details](#) | [Hide all details](#)1 2 3 4 5 > [Next](#) >> [Last](#)

Geo Result Relevance-Sat Nov 29 21:39:03 PST 2014

[View a HIT in this group](#)

Requester: [Amazon Requester Inc.](#) HIT Expiration Date: Dec 30, 2014 (3 weeks 6 days) Reward: \$0.00
Time Allotted: 60 minutes HITS Available: 26320

Description: HIT to judge the relevancy of a given geo search result for a specified query.

Keywords: [search](#)

Qualifications Required:

Geo Result Relevancy Qualification Test is not less than 100

Inv_B_2

[View a HIT in this group](#)

Requester: [rohzt0d](#) HIT Expiration Date: Dec 21, 2014 (2 weeks 4 days) Reward: \$0.00
Time Allotted: 48 minutes HITS Available: 25525

Description: New Inv_B hit type.

Keywords: [inv_b](#)

Qualifications Required:

Inv_B has been granted

Extract purchased items from a shopping receipt

[View a HIT in this group](#)

Requester: [Jon Brellig](#) HIT Expiration Date: Dec 9, 2014 (6 days 23 hours) Reward: \$0.09
Time Allotted: 2 hours HITS Available: 13308

Description: Transcribe all of the purchased items and total from a shopping receipt

Keywords: [image](#), [receipt](#), [categorize](#), [transcribe](#), [extract](#), [data](#), [entry](#), [transcription](#), [text](#), [easy](#), [qualification](#), [secure](#), [jon](#), [brellig](#), [prod](#)

An interlinking microtask

Object 1:

Name: 'Turner, Ted'

URL at New York Times: [see related NYT web site](#)

Object 2:


Name: 'Melanie Munch'

Description: 'Melanie Münch, (born August 4, 1981), better known by the stage name Mell, is a German singer, best known for being the lead singer of trance group Groove Coverage.'

Category: 'musical artist' 'living people' 'Living people' 'Person' 'somebody' 'individual' 'someone' 'mortal' 'soul' 'person' 'artist' 'German singers' 'Thing' 'agent' 'person'


Question 1 - Is Object 1 the same as Object 2?

- ☐ no
☐ yes


 Please select only one of the answers

Question 2 - Select the name of Object 2

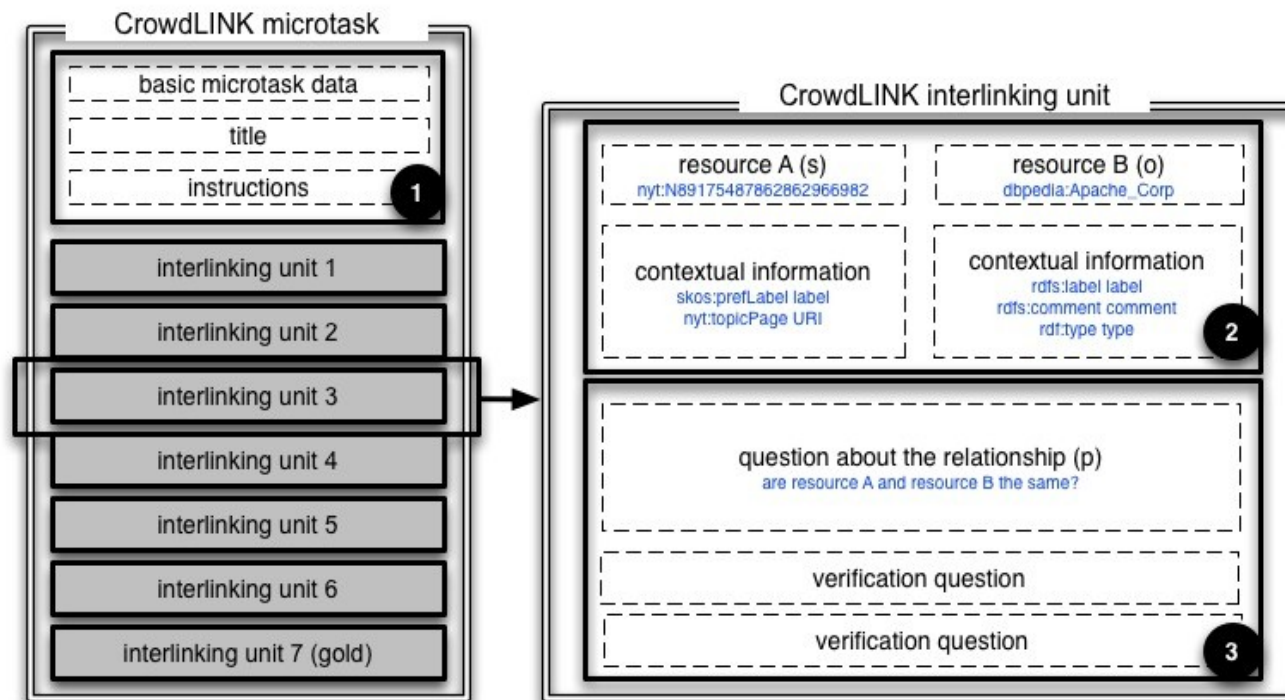
- ☐ Turner, Ted
☐ Melanie Munch

 Please select only one of the answers

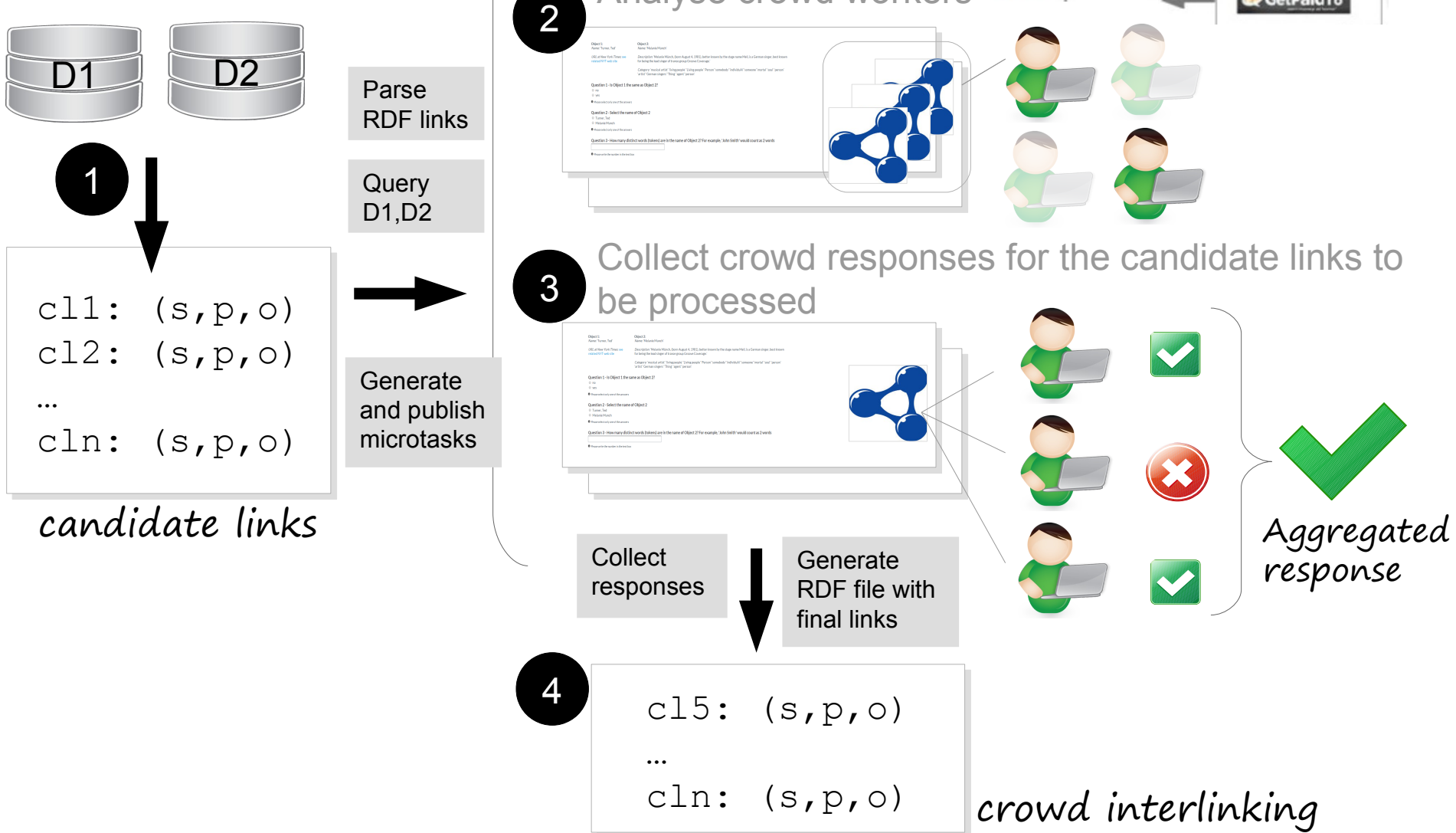
Question 3 - How many distinct words (tokens) are in the name of Object 2? For example, 'John Smith' would count as 2 words

 Please write the number in the text box

An interlinking microtask



Approach



Approach (II)

- Analyse crowd workers to filter out people
 - With bad intentions (i.e. scammers)
 - Who do not have enough knowledge
- Select representative links from which the answer is known (ground truth) and assess people → domain expert useful

```
x:b rdfs:label "Berlin";  
rdf:type o:City;
```

```
x:b2 rdfs:label "Berlinale";  
rdf:type o:Event;
```

Measure
difficulty based
on data
heuristics

```
x:b rdfs:label "Córdoba";  
rdf:type o:City;
```

```
x:b2 rdfs:label "Córdoba";  
rdf:type o:City;
```

Select
different
matching
cases

```
x:b rdfs:label "Córdoba";  
rdf:type o:City;  
wgs84:lat -31.400;
```

```
x:b2 rdf:type o:City;  
wgs84:lat 37.883;
```


Approach (II)

- Analyse crowd workers to filter out people
 - With bad intentions (i.e. scammers)
 - Who do not have enough knowledge
- Select representative links from which the answer is known (ground truth) and assess people → domain expert useful

```
x:b rdfs:label "Berlin"; x:b2 rdfs:label "Berlinale";
```

Measure
based

Two-way feedback

For the question titled "Question 1 - Is Object 1 the same as Object 2?" you answered: **yes** but the correct answer was: **no**.

The reason for this is: **"["They do not refer to the same element"]"**

If you believe that this test question is unfair or incorrect, please let us know below. We'll review these items for fairness and accuracy.

☒ That's ok
☐ This test question is unfair or incorrect!

```
x:b rdfs:label "Córdoba";  
rdf:type o:City;  
wgs84:lat -31.400;
```

```
x:b2 rdf:type o:City;  
wgs84:lat 37.883;
```

different
matching
cases

Approach



1

c11: (s,p,o)
c12: (s,p,o)
...
c1n: (s,p,o)

candidate links

Parse
RDF links

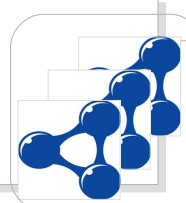
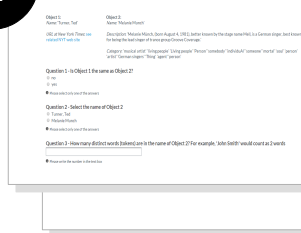
Query
D1,D2

Generate
and publish
microtasks

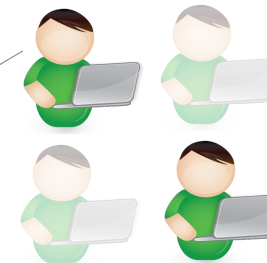
Context information

Analyse crowd workers

2

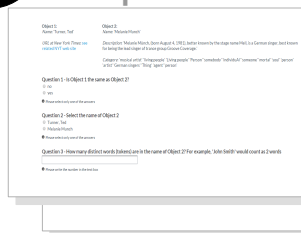


CrowdFlower



3

Collect crowd responses for the candidate links to be processed



*Aggregated
response*

Collect
responses

Generate
RDF file with
final links

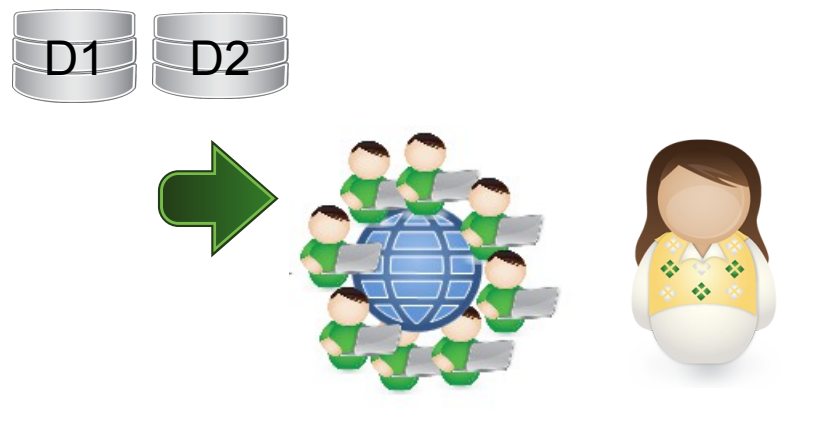
4

c15: (s,p,o)
...
c1n: (s,p,o)

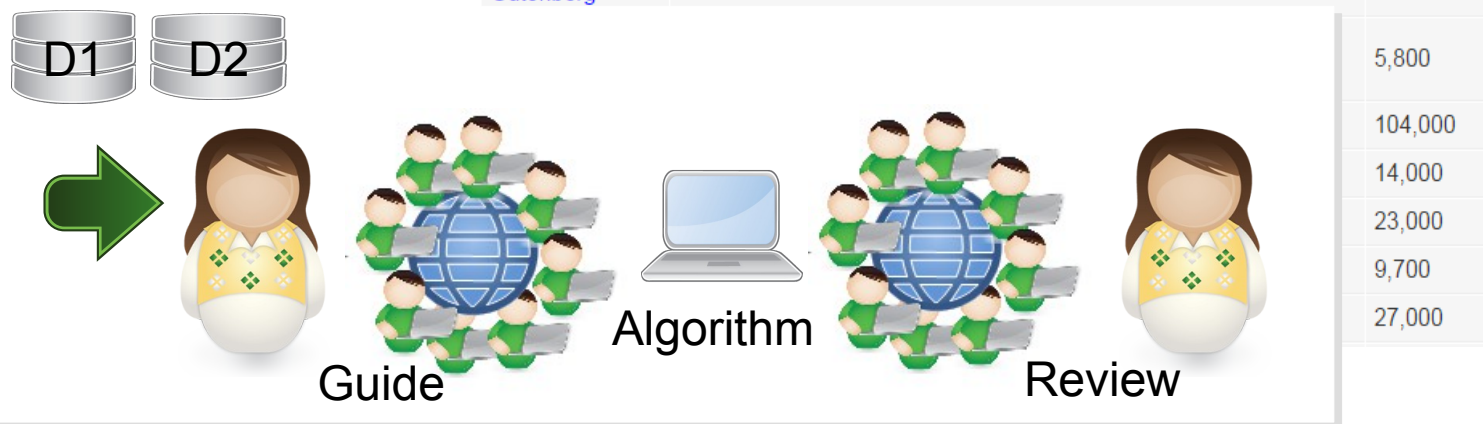
crowd interlinking

#workers per link
agreement

Approach (II)



Manual interlinking



HCOMP interlinking

Eurostat (WBSG)	Provides information about European countries and regions.	140
CIA World Factbook	Provides information about countries.	550
flickr wrappr	A wrapper around flickr that tries to generate a photo collection for each DBpedia concept.	4,000,000
Freebase	A open-license database about millions of things from various domains.	3,900,000
GADM	Spatial database of the location of the world's administrative areas.	39,000
GeoNames	Provides information about geographic features.	425,000
GeoSpecies	Information on biological orders, families, species as well as species occurrence records and related data.	16,000
Global Health Observatory	Provides access to statistical data about health problems.	200
Project Gutenberg	Provides information about authors and open access to their work.	2,500

Use cases

Mapping vocabularies

Concept A: Booklet

Definition (English): A work that is printed and bound, but without a named publisher or sponsoring institution.


Concept B: TechReport

Definition (English): A report published by a school or other institution, usually numbered within a series.

TechReport is a kind of: Informal


Are Concept A and Concept B the same?

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


 Please select only one of the answers

Select the name of Concept A

- ☐ Booklet
- ☐ TechReport

 Please select only one of the answers

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

 Please write the number in the text box

Run an automatic ontology alignment tool and post-process the results with the crowd

See also: [Sarasua et al., 2012]

Concept A: Phdthesis

Definition (English): A PhD thesis.

Concept B: Thesis


Definition (English): A thesis (either Master or PhD).

Thesis is a kind of: Academic

Other elements that are of kind Academic: 'LectureNotes'


Do you see any connection between Phdthesis and Thesis?

- ☐ There is no relation between Phdthesis and Thesis
- ☐ Thesis is a kind of Phdthesis
- ☐ Phdthesis is a kind of Thesis
- ☐ Phdthesis is the same as Thesis


 Please select only one of the answers

Select the name of Concept B

- ☐ Phdthesis
- ☐ Thesis

 Please select only one of the answers

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

 Please write the number in the text box

Context information
pre-configured

Discovering links between instances

Object 1:

Name: 'Turner, Ted'

URL at New York Times: [see related NYT web site](#)

Object 2:


Name: 'Melanie Munch'

Description: 'Melanie Münch, (born August 4, 1981), better known by the stage name Mell, is a German singer, best known for being the lead singer of trance group Groove Coverage.'

Category: 'musical artist' 'living people' 'Living people' 'Person' 'somebody' 'individuAl' 'someone' 'mortal' 'soul' 'person' 'artist' 'German singers' 'Thing' 'agent' 'person'


Question 1 - Is Object 1 the same as Object 2?

- ☐ no
☐ yes


 Please select only one of the answers

Question 2 - Select the name of Object 2

- ☐ Turner, Ted
☐ Melanie Munch

 Please select only one of the answers

Question 3 - How many distinct words (tokens) are in the name of Object 2? For example, 'John Smith' would count as 2 words

 Please write the number in the text box

- a) To extract the patterns of the linkage rules (i.e. labelling)
- b) To post-process irregular multilingual values, different name versions
- c) To automatically identify patterns of errors in a resulting set of links, which may be afterwards reviewed by the experts

Curating mapping extensions to authority files

Sam is a famous Actress. She has many facets and has been named with different names.

[see her related Wikipedia web site](#)

Quality control can be done by giving these answers to other crowd workers

Question 1 - Find and write another name that has been used to refer to Sam.

Question 2 - Where did you find it? Type in the URL of the Web site where you found it

Checking usefulness of links with library users

- There are different possible targets for the interlinking of a dataset: which possibility to select for the Web portal?
- Embed Web site in a microtask and ask for specific information or observe next Web site opened

3 Challenges

Deciding whether to crowdsource or not

- Depends to a large extent on the data
 - Specific domains require more crowd management effort
 - Benefit compared to automatically generated links may vary
 - Availability of workers may change in time

Libraries and the cultural heritage domain have high potential (multilinguality, different naming conventions, knowledge exploration)

- What should be processed by the crowd
 - Criteria for selecting subsets of the data (e.g. confidence of machine)
 - > Trial, error and assessment

Building a loyal workforce

- Attracting good crowd workers
 - Microtasks are constantly being published
 - Higher reward may also attract more malicious workers
- Working with people repeatedly is not supported by majority of crowdsourcing platforms
- How to make crowd workers keep on working in these microtasks without them getting demotivated?

It's really easy to change people's motivations, [at Zooniverse] we find people are motivated by wanting to contribute, they want a sense that this is something real. And in adding game-like elements you can destroy that quite quickly” Chris Lintott, Zooniverse
<http://www.wired.co.uk/news/archive/2013-09/12/fraxinus-gamifying-science/viewgallery/307960>

- > Be fair (see also *Guidelines on Crowd Work for Academic Researchers*, 2014)
- > Listen to crowd workers (e.g. direct comments, twitter, ratings, monitor online discussions)
- > Recognize their work
- > Be aware that gamification is not always the best solution

Working with unknown humans

- Open call can be a problem and an opportunity at the same time: people have diverse
 - Motivation and dedication
 - Context and profile
 - Background knowledge
- Crowdsourcing platforms have limited support for personalisation
- Working with suitable crowd
 - Identify what they can do best
 - Type of task / data level
 - Competences vs experience cross platform analysis
 - Assign work accordingly
 - Weight vs reject

>Towards a Crowd Work CV
See also: [Sarasua et al., 2014]

Plea to this community

- Interlinking is much more than deduplication, consider using also other relations
- Consider connecting library datasets to different complementary domains
- Interlinking to non editorial data can also be enriching
- The more datasets you connect the better
- Document your interlinking on the VoiD description of your dataset
- Query and make use of available links

If you need humans to process data while interlinking datasets, consider crowd intervention because it can be very valuable for enhancing your results.



Thank you for your attention!

Cristina Sarasua
Institute for Web Science and Technologies
Universität Koblenz-Landau



csarasua@uni-koblenz.de
<http://de.slideshare.net/cristinasarasua>
<https://github.com/criscod>



References

- Sarasua, C., Simperl, E., Noy, N.F.: CrowdMAP: Crowdsourcing ontology alignment with microtasks. In: Proceedings of the 11th International Semantic Web Conference (ISWC). (2012)
- Sarasua, C., Thimm, M. Crowd Work CV: Recognition for Micro Work. In: SoHuman workshop, co-located with Social Informatics (SocInfo). (2014)
- Guidelines on Crowd Work for Academic Researchers (2014). http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters