Supporting Data Interlinking in Semantic Libraries with Microtask Crowdsourcing

Cristina Sarasua

SWIB 2014, Bonn
Supporting Data Interlinking in Semantic Libraries with Microtask Crowdsourcing

Cristina Sarasua

MARC 21
EDM
FRBR
Please share your thoughts on interlinking!
https://etherpad.mozilla.org/4IfZDaTBIe
Interlinking on the Web of Data

Cross-dataset links

(a,r,b) | a in D1, b in D2

d1:timbl owl:sameAs d2:timbernerslee;
d1:donostia owl:sameAs d2:sansebastian;
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 owl:sameAs ?city2 .
  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/lld/XGR-lld-20111025/
Generating links

Identify the resources to be connected with relation R

Decision boundary between link and non-link

Comparison criteria

Supporting Data Interlinking in Semantic Libraries with Microtask Crowdsourcing

Cristina Sarasua

https://etherpad.mozilla.org/4IfZDaTBlE

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

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, $30 per 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

Scalable
Fast

Cristina Sarasua
Supporting Data Interlinking in Semantic Libraries with Microtask Crowdsourcing
Supporting Data Interlinking in Semantic Libraries with Microtask Crowdsourcing

Cristina Sarasua
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

1. Query D1, D2

   | cl1: (s,p,o) |
   | cl2: (s,p,o) |
   | ... |
   | cln: (s,p,o) |

   candidate links

2. Analyse crowd workers

3. Collect crowd responses for the candidate links to be processed

4. Collect responses

   | cl5: (s,p,o) |
   | ... |
   | cln: (s,p,o) |

   crowd interlinking

Parse RDF links

Generate and publish microtasks

Collect responses

Generate RDF file with final links

Aggregated response
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

\[
\text{x:b rdfs:label "Berlin"; rdf:type o:City; }\quad \text{x:b2 rdfs:label "Berlinale"; rdf:type o:Event; } \\
\text{x:b rdfs:label "Córdoba"; rdf:type o:City; }\quad \text{x:b2 rdfs:label "Córdoba"; rdf:type o:City; } \\
\text{x:b rdfs:label "Córdoba"; rdf:type o:City; wgs84:lat -31.400; }\quad \text{x:b2 rdf:type o:City; wgs84:lat 37.883; }
\]
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

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!

Different matching cases

x:b rdfs:label "Berlin";
x:b2 rdfs:label "Berlinale";
x:b rdfs:label "Córdoba";
x:b2 rdf:type o:City;
x:b2 rdf:type o:City;
x:b2 rdfs:label "Córdoba";

x: b rdfs:label "Berlin";
x:b2 rdfs:label "Berlinale";
x:b rdfs:label "Córdoba";
x:b2 rdf:type o:City;
x:b2 rdf:type o:City;
x:b2 rdfs:label "Córdoba";
wgs84:lat -31.400;
x:b2 rdf:type o:City;
wgs84:lat 37.883;
Supporting Data Interlinking in Semantic Libraries with Microtask Crowdsourcing

Cristina Sarasua

1. Candidate links
   - cl1: (s,p,o)
   - cl2: (s,p,o)
   - ...
   - cln: (s,p,o)

2. Analyse crowd workers
   - Parse RDF links
   - Query D1, D2

3. Collect crowd responses for the candidate links to be processed
   - Collect responses
   - Generate RDF file with final links
   - Aggregated response
   - #workers per link agreement

Context information

Approach

D1
D2
Approach (II)

**Manual interlinking**

**HCOMP interlinking**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Number of Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eurostat (WBSC)</td>
<td>Provides information about European countries and regions.</td>
<td>140</td>
</tr>
<tr>
<td>CIA World Factbook</td>
<td>Provides information about countries.</td>
<td>550</td>
</tr>
<tr>
<td>flickr wrapper</td>
<td>A wrapper around flickr that tries to generate a photo collection for each DBpedia concept.</td>
<td>4,000,000</td>
</tr>
<tr>
<td>Freebase</td>
<td>A open-license database about millions of things from various domains.</td>
<td>3,900,000</td>
</tr>
<tr>
<td>GADM</td>
<td>Spatial database of the location of the world’s administrative areas.</td>
<td>39,000</td>
</tr>
<tr>
<td>GeoNames</td>
<td>Provides information about geographic features.</td>
<td>425,000</td>
</tr>
<tr>
<td>GeoSpecies</td>
<td>Information on biological orders, families, species as well as species occurrence records and related data.</td>
<td>16,000</td>
</tr>
<tr>
<td>Global Health Observatory</td>
<td>Provides access to statistical data about health problems.</td>
<td>200</td>
</tr>
<tr>
<td>Project Gutenberg</td>
<td>Provides information about authors and open access to their work.</td>
<td>2,500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5,800</td>
</tr>
<tr>
<td></td>
<td></td>
<td>104,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9,700</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27,000</td>
</tr>
</tbody>
</table>
Use cases
Mapping vocabularies

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

See also: [Sarasua et al., 2012]
Discovering links between instances

Object 1:
Name: 'Turner, Ted'

Object 2:
Name: 'Melanie Munch'

URL at New York Times: see related NYT web site

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' '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
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

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

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

**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?
3 Challenges
# Deciding whether to crowdsourc 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?
  - 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

*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
# 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

carasua@uni-koblenz.de
http://de.slideshare.net/cristinasarasua
https://github.com/criscod
References