Applications of Microtasks & Final Remarks

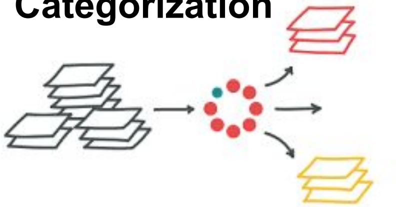
Maribel Acosta, Amrapali Zaveri

KIT Karlsruher Institut für Technologie

Maastricht University
Institute of Data Science

Applications of Microtask Crowdsourcing

Classification and Categorization



Finding Metadata

Ranking

Promoting

Data Collection and Enhancement

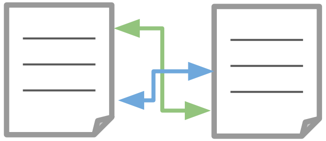
Sentiment Analysis

Media Transcription

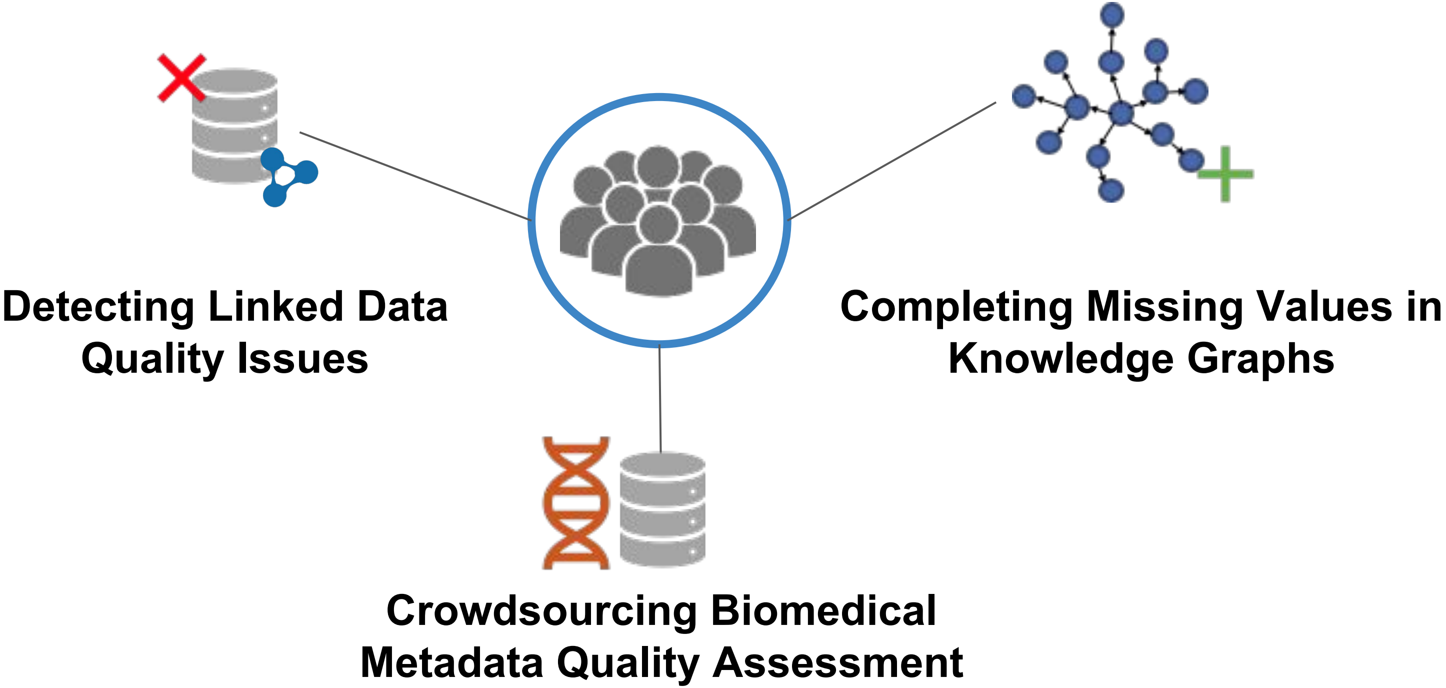
Content Feedback

Content Moderation

Content Verification



Examples of Crowdsourcing Applications in Data Science



Detecting Linked Data Quality Issues

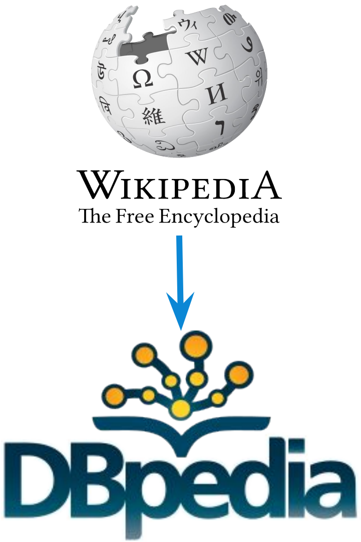
Completing Missing Values in Knowledge Graphs

Crowdsourcing Biomedical Metadata Quality Assessment

Outline

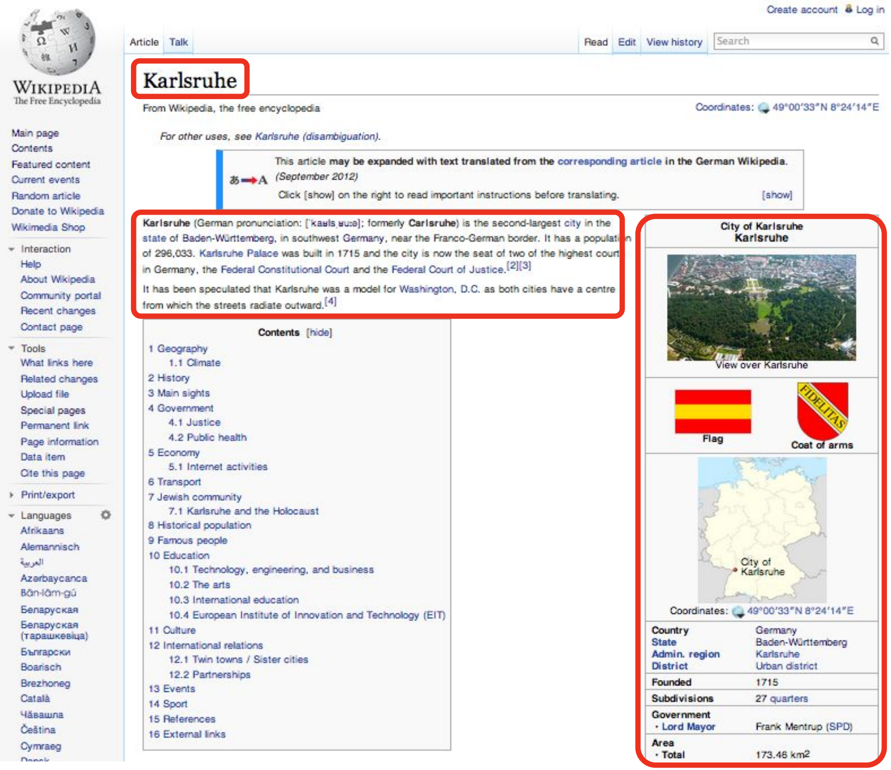
1. **Detecting Linked Data Quality Issues**
2. Completing Missing Values in Knowledge Graphs
3. Crowdsourcing Biomedical Metadata Quality Assessment
4. Other Applications (Open Discussion)
5. Final Remarks

The DBpedia Knowledge Graph



Semi-structured data from Wikipedia

http://en.wikipedia.org/wiki/Karlsruhe



Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Data item
Cite this page

Print/export

Languages
Afrikaans
Alemannisch
العربية
Azərbaycanca
Bân-lâm-gú
Беларуская
Беларуская (тарашкевіца)
Български
Boarisch
Brezhoneg
Català
Чăвашла
Čeština
Cymraeg

Karlsruhe

From Wikipedia, the free encyclopedia

Coordinates: 49°00′33″N 8°24′14″E

For other uses, see Karlsruhe (disambiguation).

あ→A This article **may be expanded with text translated from the corresponding article in the German Wikipedia.** *(September 2012)*
Click [show] on the right to read important instructions before translating. [show]

Karlsruhe (German pronunciation: [ˈkaʁlsˌʁuːə]; formerly **Carlsruhe**) is the second-largest city in the state of Baden-Württemberg, in southwest Germany, near the Franco-German border. It has a population of 296,033. Karlsruhe Palace was built in 1715 and the city is now the seat of two of the highest courts in Germany, the Federal Constitutional Court and the Federal Court of Justice.[2][3]

It has been speculated that Karlsruhe was a model for Washington, D.C. as both cities have a centre from which the streets radiate outward.[4]

Contents [hide]

1 Geography
  1.1 Climate
2 History
3 Main sights
4 Government
  4.1 Justice
  4.2 Public health
5 Economy
  5.1 Internet activities
6 Transport
7 Jewish community
  7.1 Karlsruhe and the Holocaust
8 Historical population
9 Famous people
10 Education
  10.1 Technology, engineering, and business
  10.2 The arts
  10.3 International education
  10.4 European Institute of Innovation and Technology (EIT)
11 Culture
12 International relations
  12.1 Twin towns / Sister cities
  12.2 Partnerships
13 Events
14 Sport
15 References
16 External links

City of Karlsruhe
Karlsruhe

View over Karlsruhe

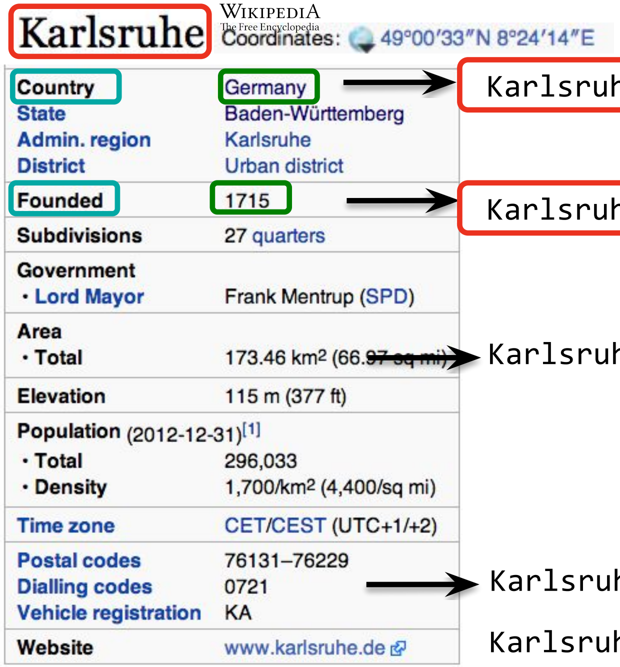
Flag

Coat of arms

Coordinates: 49°00′33″N 8°24′14″E

Country	Germany
State	Baden-Württemberg
Admin. region	Karlsruhe
District	Urban district
Founded	1715
Subdivisions	27 quarters
Government	
· **Lord Mayor**	Frank Mentrup (SPD)
Area	
· **Total**	173.46 km2

The DBpedia Knowledge Graph



Karlsruhe Coordinates: 49°00′33″N 8°24′14″E

Country	Germany
State	Baden-Württemberg
Admin. region	Karlsruhe
District	Urban district
Founded	1715
Subdivisions	27 quarters
Government	
• Lord Mayor	Frank Mentrup (SPD)
Area	
• Total	173.46 km2 (66.97 sq mi)
Elevation	115 m (377 ft)
Population (2012-12-31)[1]	
• Total	296,033
• Density	1,700/km2 (4,400/sq mi)
Time zone	CET/CEST (UTC+1/+2)
Postal codes	76131–76229
Dialling codes	0721
Vehicle registration	KA
Website	www.karlsruhe.de

Triples!

```
Karlsruhe  country  Germany

Karlsruhe  founded  “1715” (year)
```

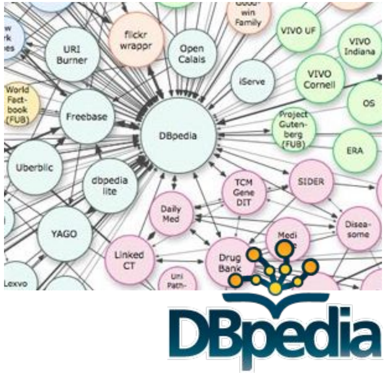
Karlsruhe area “173.46” (km^2)

Karlsruhe dialing_code “0721”

Karlsruhe website www.karlsruhe.de

Quality Issues to Crowdsource

Three categories of quality problems occur
in DBpedia [Zaveri2013] and can be crowdsourced:



- **Incorrect object**

  ```
  dbr:Dave_Dobbyn dbp:dateOfBirth “3” .
  ```

- **Incorrect data type or language tags**

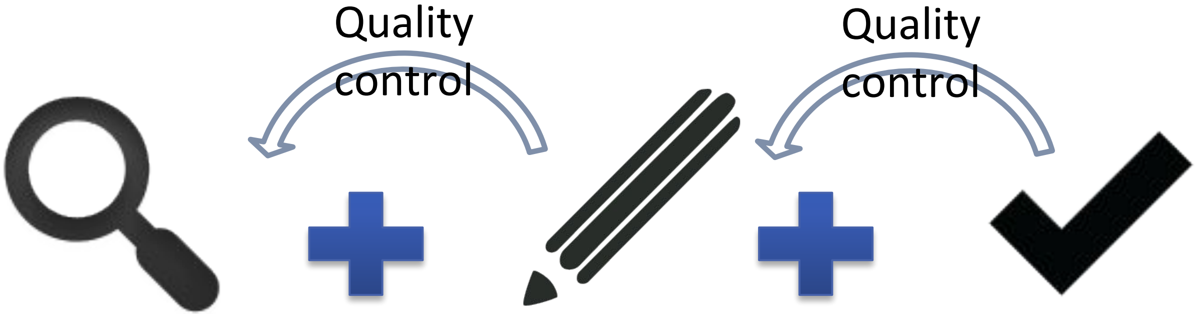
  ```
  dbr:Torishima_Izu_Islands foaf:name “鳥島”@en .
  ```

- **Incorrect link to “external Web pages”**

  ```
  dbr:John-Two-Hawks dbo:wikiPageExternalLink <http://cedarlakedvd.com>.
  ```

Crowdsourcing Approach

Find-Fix-Verify Pattern [Bernstein2010]



Find

Identify problematic elements within a data source.

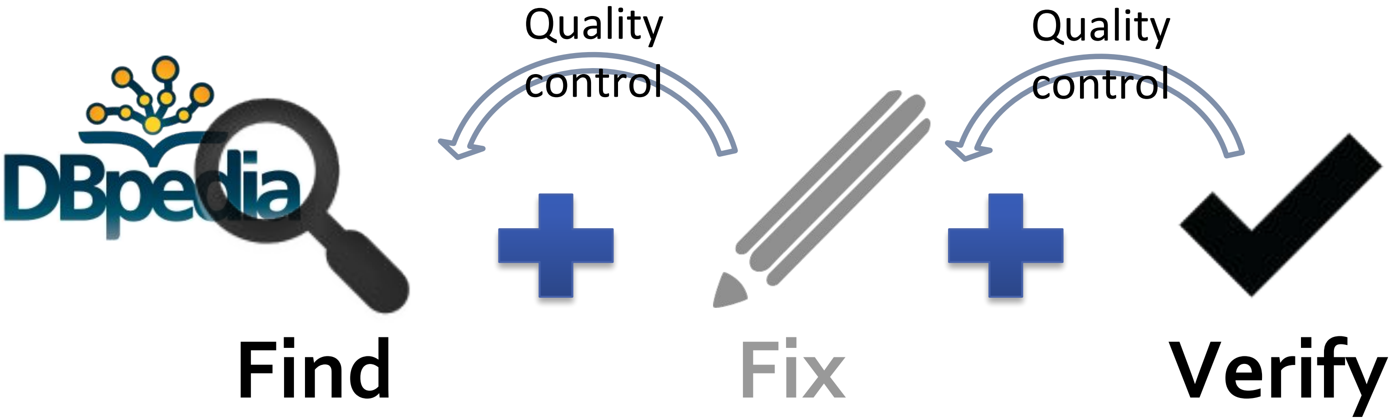
Fix

Correct the elements identified in the previous stage.

Verify

Confirm the output from the previous stage.

Applying Find-Fix-Verify to our Case Study: DBpedia-DQ



Quality control

Quality control

Find

Fix

Verify

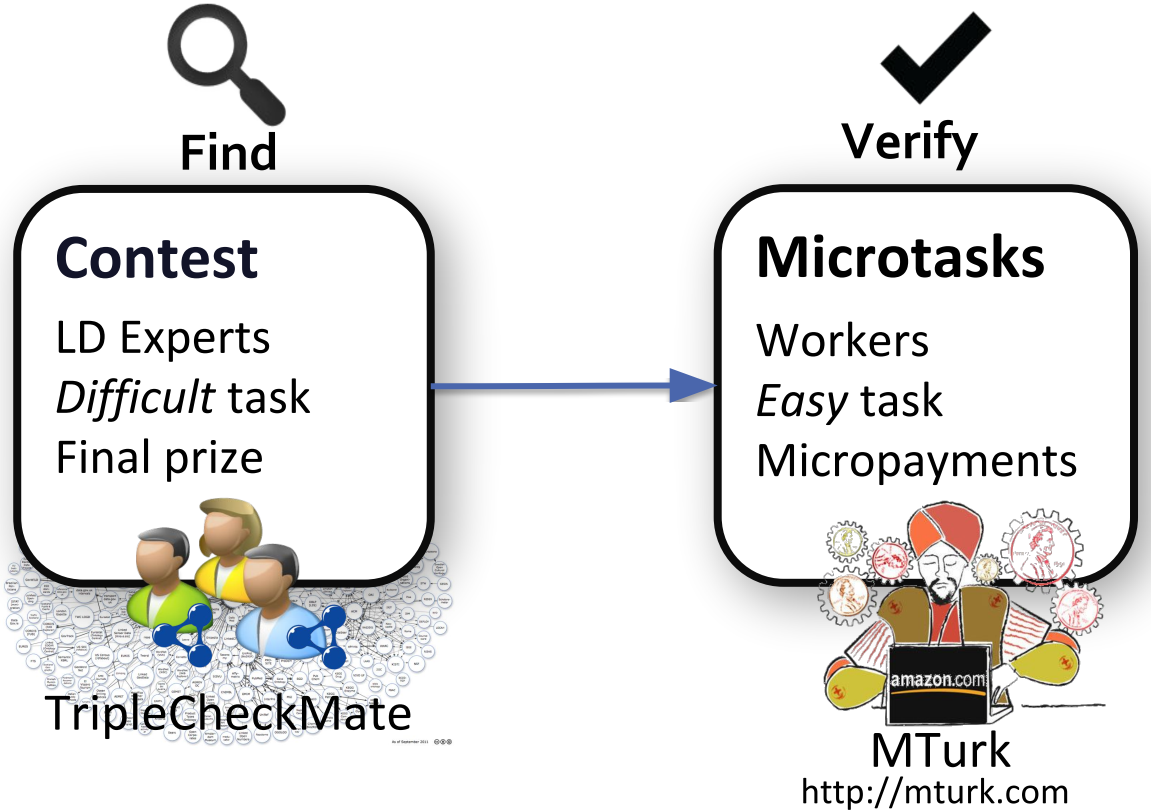
Identify erroneous triples and classify them according to the error found.

Confirm the output from the previous stage.

DBpedia-DQ has two variants:

1. Combining experts + non-experts crowds (workers)
2. Using non-experts crowds (workers) in both stages

Combining Experts + Workers (EW)



Find

Contest

LD Experts

Difficult task

Final prize

TripleCheckMate

Verify

Microtasks

Workers

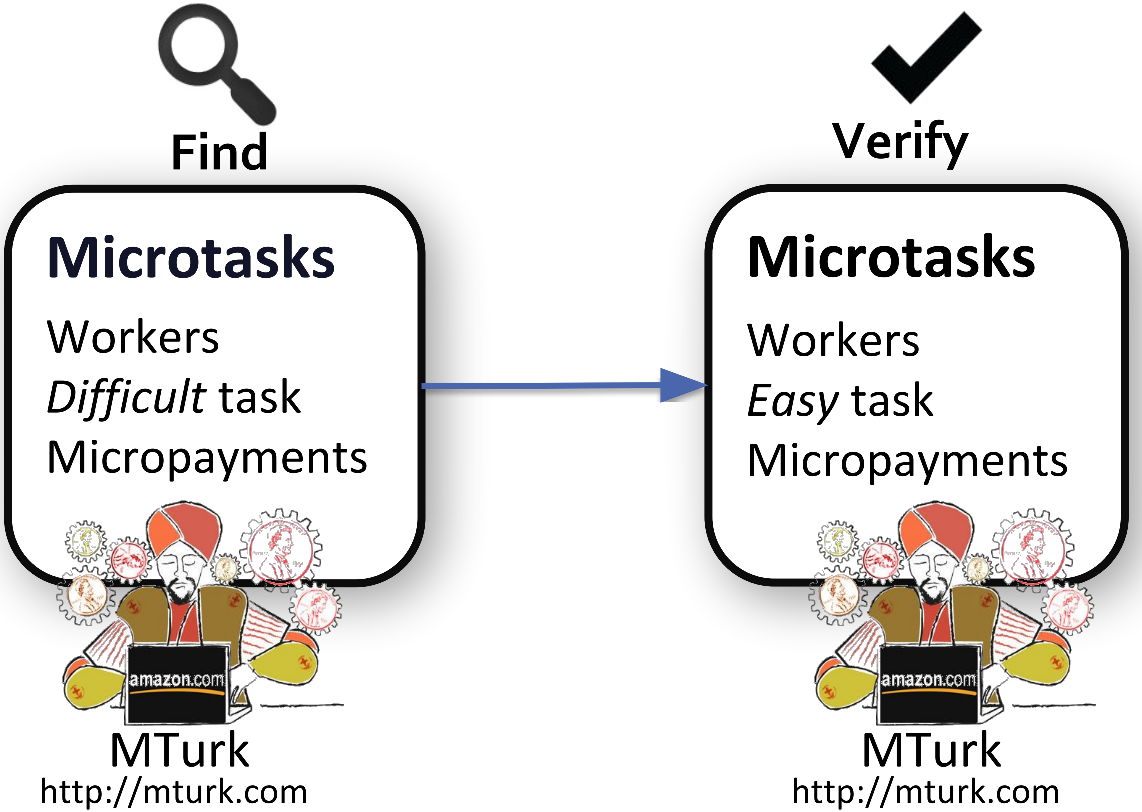
Easy task

Micropayments

MTurk

http://mturk.com

Combining Workers + Workers (WW)



Find

Microtasks

Workers

Difficult task

Micropayments

MTurk

http://mturk.com

Verify

Microtasks

Workers

Easy task

Micropayments

MTurk

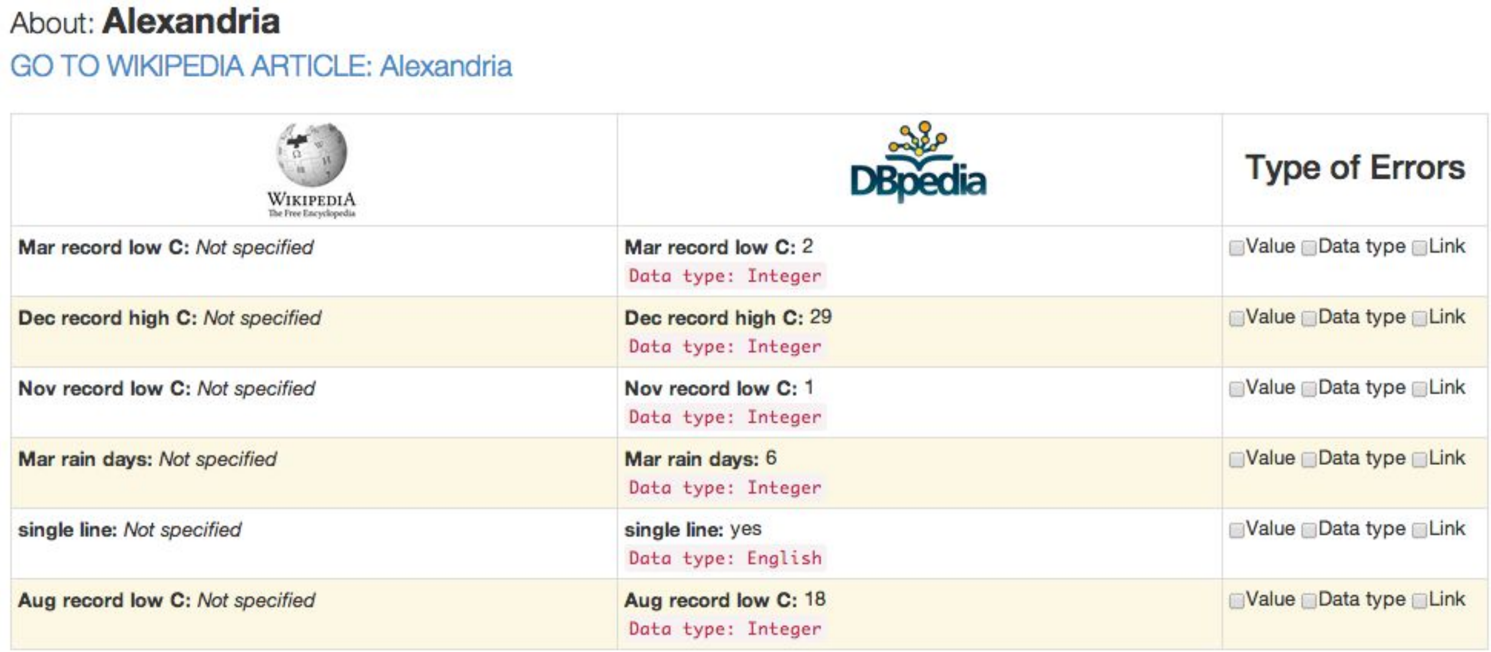
http://mturk.com

DBpedia-DQ Microtask Interfaces

Find stage with workers: MTurk Tasks

Microtask settings:

- Max. 30 questions per microtask
- Payment: 0.06 US$ per microtask (Nov. 2014)



About: **Alexandria**

GO TO WIKIPEDIA ARTICLE: Alexandria

WIKIPEDIA The Free Encyclopedia	DBpedia	Type of Errors
Mar record low C: *Not specified*	**Mar record low C:** 2 `Data type: Integer`	□Value □Data type □Link
Dec record high C: *Not specified*	**Dec record high C:** 29 `Data type: Integer`	□Value □Data type □Link
Nov record low C: *Not specified*	**Nov record low C:** 1 `Data type: Integer`	□Value □Data type □Link
Mar rain days: *Not specified*	**Mar rain days:** 6 `Data type: Integer`	□Value □Data type □Link
single line: *Not specified*	**single line:** yes `Data type: English`	□Value □Data type □Link
Aug record low C: *Not specified*	**Aug record low C:** 18 `Data type: Integer`	□Value □Data type □Link

DBpedia-DQ Microtask Interfaces

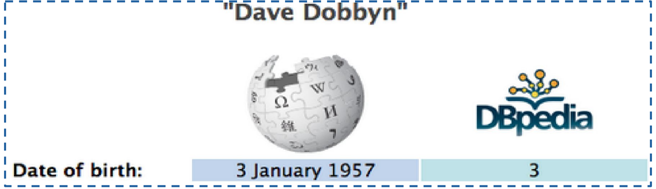
Verify stage with workers: MTurk Tasks

Microtask settings:

- 5 questions per microtask
- Payment: 0.04 US$ per microtask (Feb. 2013)

```
dbr:Dave_Dobbyn dbp:dateOfBirth “3” .
```

Incorrect object

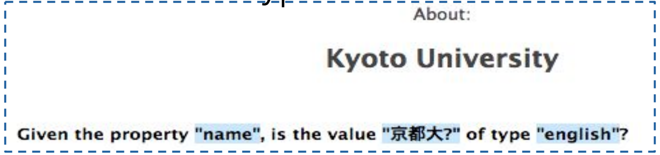


"Dave Dobbyn"

Date of birth:	3 January 1957	3

```
dbr:Torishima_Izu_Islands foaf:name “鳥島”@en .
```

Incorrect data type



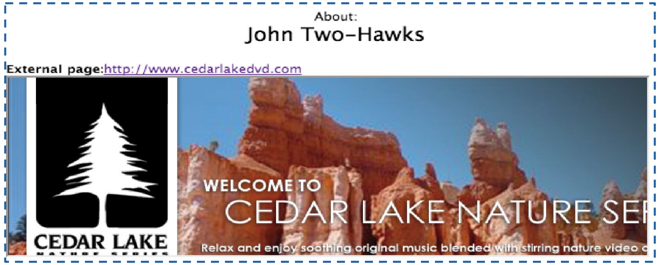
About:

Kyoto University

Given the property "name", is the value "京都大?" of type "english"?

```
dbr:John-Two-Hawks dbo:wikiPageExternalLink
    <http://cedarlakedvd.com>.
```

Incorrect outlink



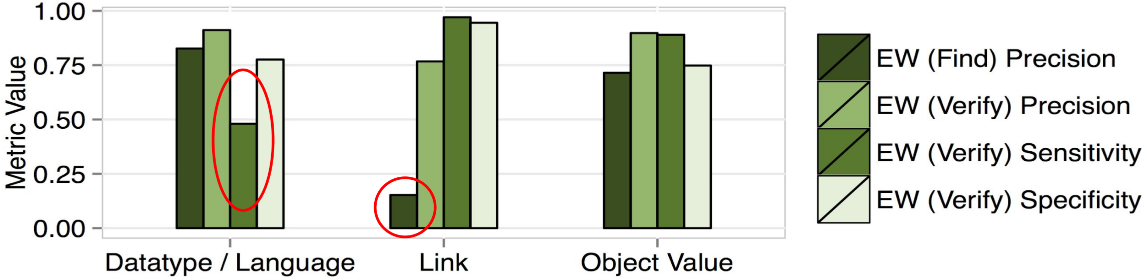
About:

John Two-Hawks

External page:http://www.cedarlakedvd.com

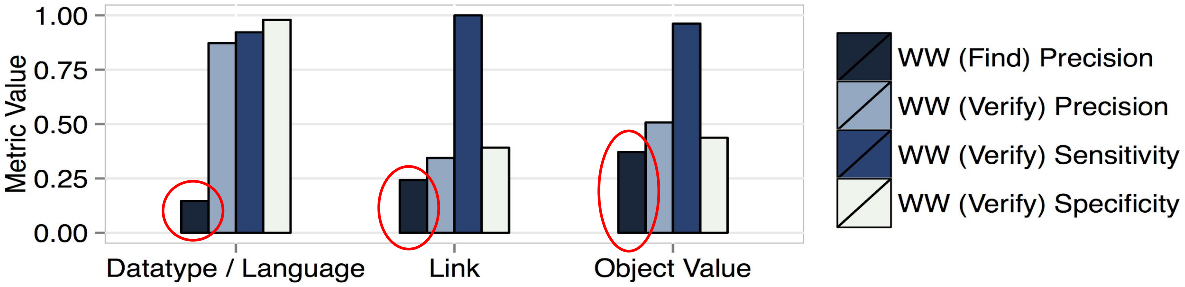
CEDAR LAKE
WELCOME TO
CEDAR LAKE NATURE SEI
Relax and enjoy soothing original music blended with stirring nature video c

DBpedia-DQ: Experimental Results



Main findings:

- It is difficult for the workers to assess datatypes
- Experts are not good in assessing external links
- Two-step validation increases the overall quality



Main findings:

- It is difficult for the workers to execute the find stage
- Workers are exceptionally good at identifying incorrect triples (high sensitivity)

M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, F. Flöck, J. Lehmann. Detecting Linked Data Quality Issues via Crowdsourcing.Semantic Web Journal, 2018.

Experimental Results:
Crowd-based vs. Automatic Data Quality Assessment

Main finding:

Humans (experts and workers) detected quality issues that were not detected via RDFUnit (automatic tool) and vice versa.

M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, F. Flöck, J. Lehmann. Detecting Linked Data Quality Issues via Crowdsourcing.Semantic Web Journal, 2018.

Outline

1. Detecting Linked Data Quality Issues in DBpedia
2. **Completing Missing Values in Knowledge Graphs**
3. Crowdsourcing Biomedical Metadata Quality Assessment
4. Other Applications (Open Discussion)
5. Summary & Final Remarks

Missing Values in Knowledge Graphs (1)

Retrieve drugs that are annotated with the prefix "C01" (Cardiac Therapy) in the Anatomical Therapeutic Chemical (ATC) classification system and which have known routes of administration.

```
SELECT DISTINCT ?drug WHERE {
 ?drug rdf:type dbo:Drug .
 ?drug dbo:atcPrefix “C01” .
 ?drug dbp:routesOfAdmainistration ?route .
}
```

(v. 2016)

47 drugs

Missing Values in Knowledge Graphs (2)

Retrieve drugs that are annotated with the prefix “C01” (Cardiac Therapy) in the Anatomical Therapeutic Chemical (ATC) classification system ~~and which have known routes of administration~~.

```
SELECT DISTINCT ?drug WHERE {
  ?drug rdf:type dbo:Drug .
  ?drug dbo:atcPrefix “C01” .
  ~~?drug dbp:routesOfAdministration ?route .~~
}
```

DBpedia

(v. 2016)

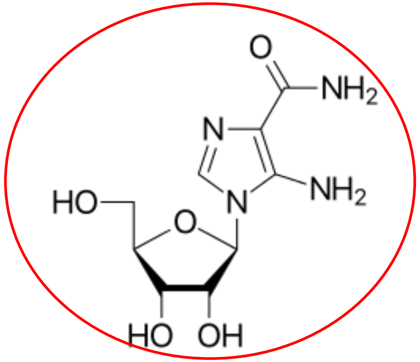
⇨

98 drugs

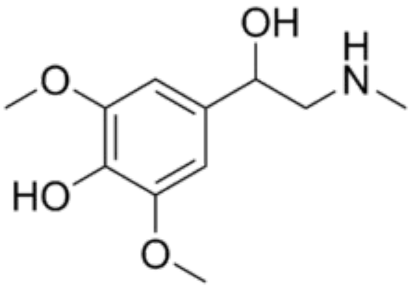
(There are 48 drugs without routes of administration)

Missing Values in Knowledge Graphs (3)

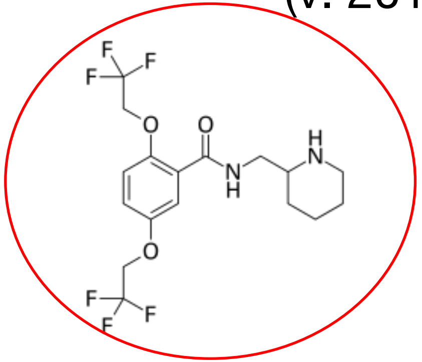
Examples of drugs (with ATC prefix “C01”) with no route as of administration in DBpedia (v. 2016)



dbr:Acadesine



dbr:Dimetofrine



dbr:Flecainide

Intravenous administration, for treating leukemia.
Source: PubChem

Also used in doping (sports).
Source: PubMed

No route found.

Oral administration,
Source: DrugBank

All images licensed under Fair use via Wikipedia.

Our Approach:
HARE - Hybrid SPARQL Query Engine

- A hybrid **machine/human SPARQL query engine** that is able to enhance the size of query answers.

- Based on a novel RDF completeness model, HARE implements querying techniques able to **detect missing values in knowledge graphs on-the-fly**

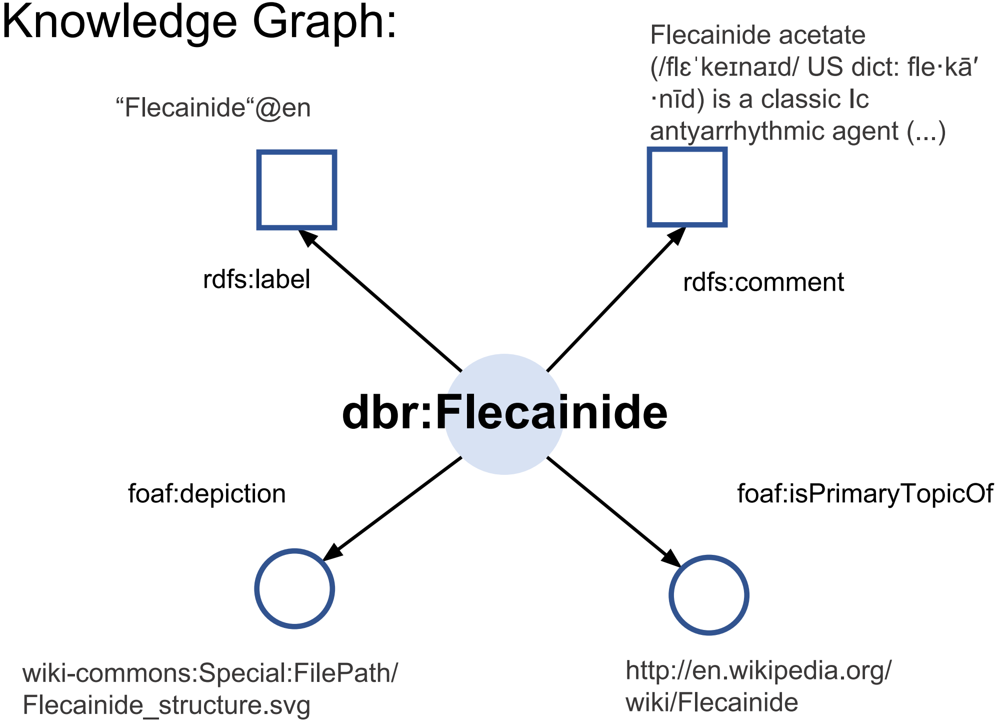
- Resorts to microtask **crowdsourcing** to resolve missing values:
    - The HARE **microtask manager** generates task interfaces automatically
    - HARE exploits the **semantics of resources** in knowledge graphs to generate task interfaces

CrowdFlower

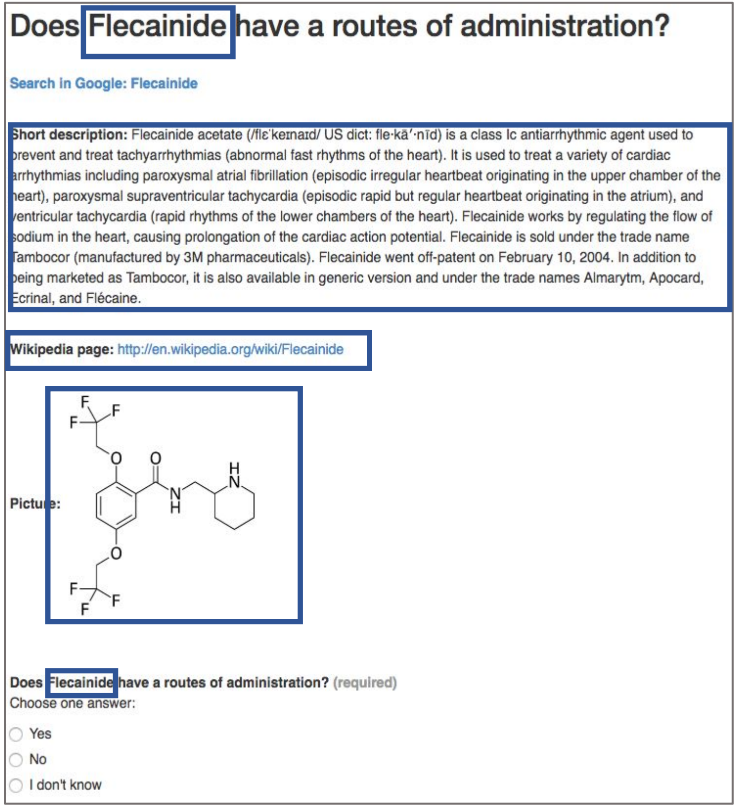
M. Acosta, E. Simperl, F. Flöck, M.E. Vidal. HARE: A Hybrid SPARQL Engine to Enhance Query Answers via Crowdsourcing. International Conference on Knowledge Capture, 2015. **Best Student Paper Award.**

HARE Microtask Manager

Knowledge Graph:



"Flecainide"@en
Flecainide acetate (/flɛˈkeɪnaɪd/ US dict: fle·kā′·nīd) is a classic Ic antyarrhythmic agent (...)
rdfs:label
rdfs:comment
dbr:Flecainide
foaf:depiction
foaf:isPrimaryTopicOf
wiki-commons:Special:FilePath/Flecainide_structure.svg
http://en.wikipedia.org/wiki/Flecainide



Does Flecainide have a routes of administration?

Search in Google: Flecainide

Short description: Flecainide acetate (/flɛˈkeɪnaɪd/ US dict: fle·kā′·nīd) is a class Ic antiarrhythmic agent used to prevent and treat tachyarrhythmias (abnormal fast rhythms of the heart). It is used to treat a variety of cardiac arrhythmias including paroxysmal atrial fibrillation (episodic irregular heartbeat originating in the upper chamber of the heart), paroxysmal supraventricular tachycardia (episodic rapid but regular heartbeat originating in the atrium), and ventricular tachycardia (rapid rhythms of the lower chambers of the heart). Flecainide works by regulating the flow of sodium in the heart, causing prolongation of the cardiac action potential. Flecainide is sold under the trade name Tambocor (manufactured by 3M pharmaceuticals). Flecainide went off-patent on February 10, 2004. In addition to being marketed as Tambocor, it is also available in generic version and under the trade names Almarytm, Apocard, Ecrinal, and Flécaine.

Wikipedia page: http://en.wikipedia.org/wiki/Flecainide

Picture:

Does Flecainide have a routes of administration? (required)

Choose one answer:

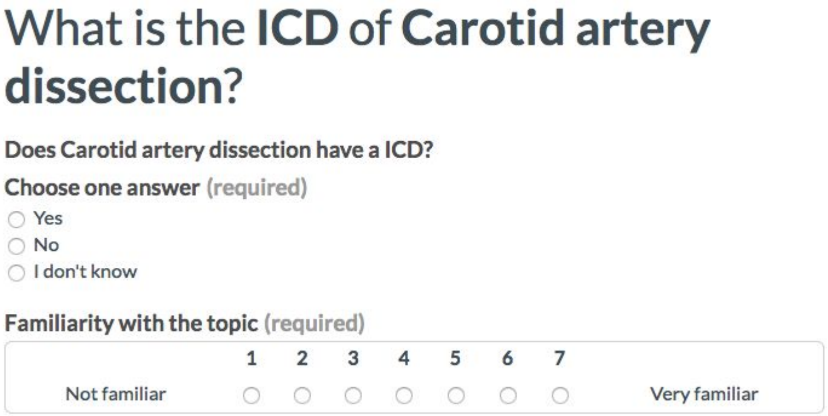
○ Yes

○ No

○ I don't know

HARE: Experimental Results

Simple Task Interface: HARE-BL



What is the **ICD** of **Carotid artery dissection**?

Does Carotid artery dissection have a ICD?

Choose one answer (required)

○ Yes
○ No
○ I don't know

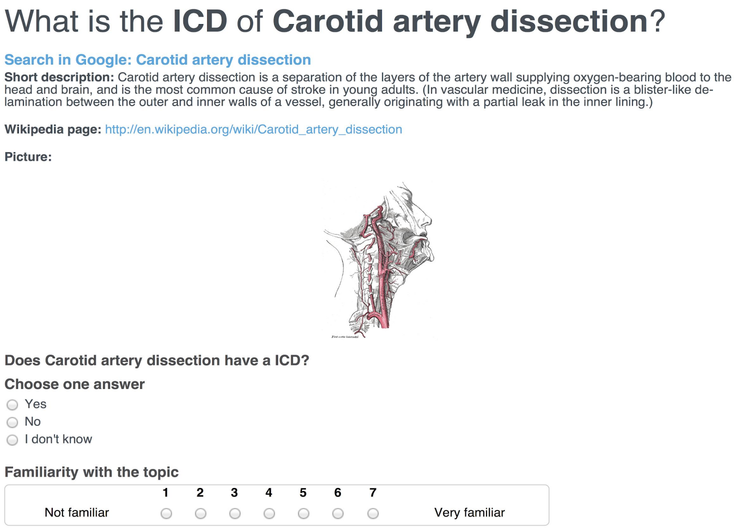
Familiarity with the topic (required)

	1	2	3	4	5	6	7	
Not familiar	○	○	○	○	○	○	○	Very familiar

Microtask settings:

- Four questions per microtask
- Payment: 0.07 US$ per microtask (Sep. 2015)

"Enriched" Task Interface: HARE



What is the **ICD** of **Carotid artery dissection**?

Search in Google: Carotid artery dissection

Short description: Carotid artery dissection is a separation of the layers of the artery wall supplying oxygen-bearing blood to the head and brain, and is the most common cause of stroke in young adults. (In vascular medicine, dissection is a blister-like de-lamination between the outer and inner walls of a vessel, generally originating with a partial leak in the inner lining.)

Wikipedia page: http://en.wikipedia.org/wiki/Carotid_artery_dissection

Picture:

Does Carotid artery dissection have a ICD?

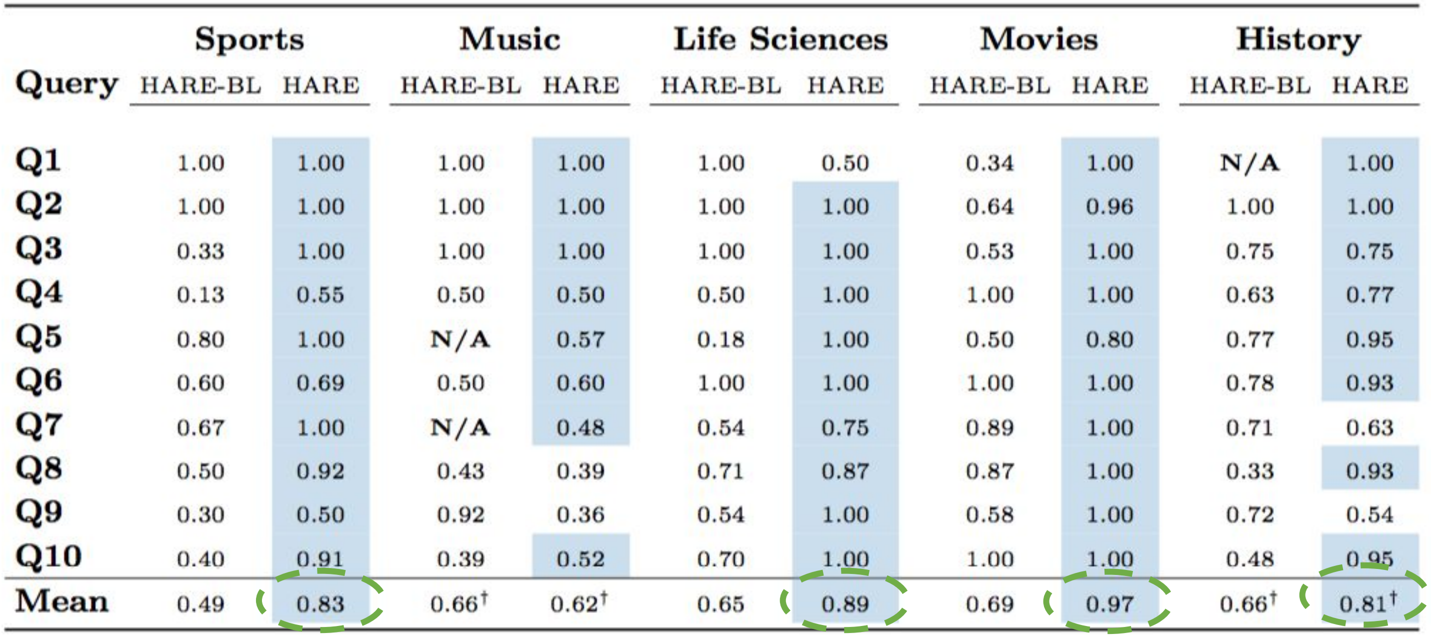
Choose one answer

○ Yes
○ No
○ I don't know

Familiarity with the topic

	1	2	3	4	5	6	7	
Not familiar	○	○	○	○	○	○	○	Very familiar

Quality of Crowd Answers: Precision



Query	Sports		Music		Life Sciences		Movies		History	
	HARE-BL	HARE	HARE-BL	HARE	HARE-BL	HARE	HARE-BL	HARE	HARE-BL	HARE
Q1	1.00	1.00	1.00	1.00	1.00	0.50	0.34	1.00	N/A	1.00
Q2	1.00	1.00	1.00	1.00	1.00	1.00	0.64	0.96	1.00	1.00
Q3	0.33	1.00	1.00	1.00	1.00	1.00	0.53	1.00	0.75	0.75
Q4	0.13	0.55	0.50	0.50	0.50	1.00	1.00	1.00	0.63	0.77
Q5	0.80	1.00	N/A	0.57	0.18	1.00	0.50	0.80	0.77	0.95
Q6	0.60	0.69	0.50	0.60	1.00	1.00	1.00	1.00	0.78	0.93
Q7	0.67	1.00	N/A	0.48	0.54	0.75	0.89	1.00	0.71	0.63
Q8	0.50	0.92	0.43	0.39	0.71	0.87	0.87	1.00	0.33	0.93
Q9	0.30	0.50	0.92	0.36	0.54	1.00	0.58	1.00	0.72	0.54
Q10	0.40	0.91	0.39	0.52	0.70	1.00	1.00	1.00	0.48	0.95
Mean	0.49	0.83	0.66†	0.62†	0.65	0.89	0.69	0.97	0.66†	0.81†

The precision of the crowd answers is in general higher when crowdsourcing semantically enriched tasks.

M. Acosta, E. Simperl, F. Flöck, M.E. Vidal. *Enhancing Answer Completeness of SPARQL Queries via Crowdsourcing.* Journal of Web Semantics, 2017.

Outline

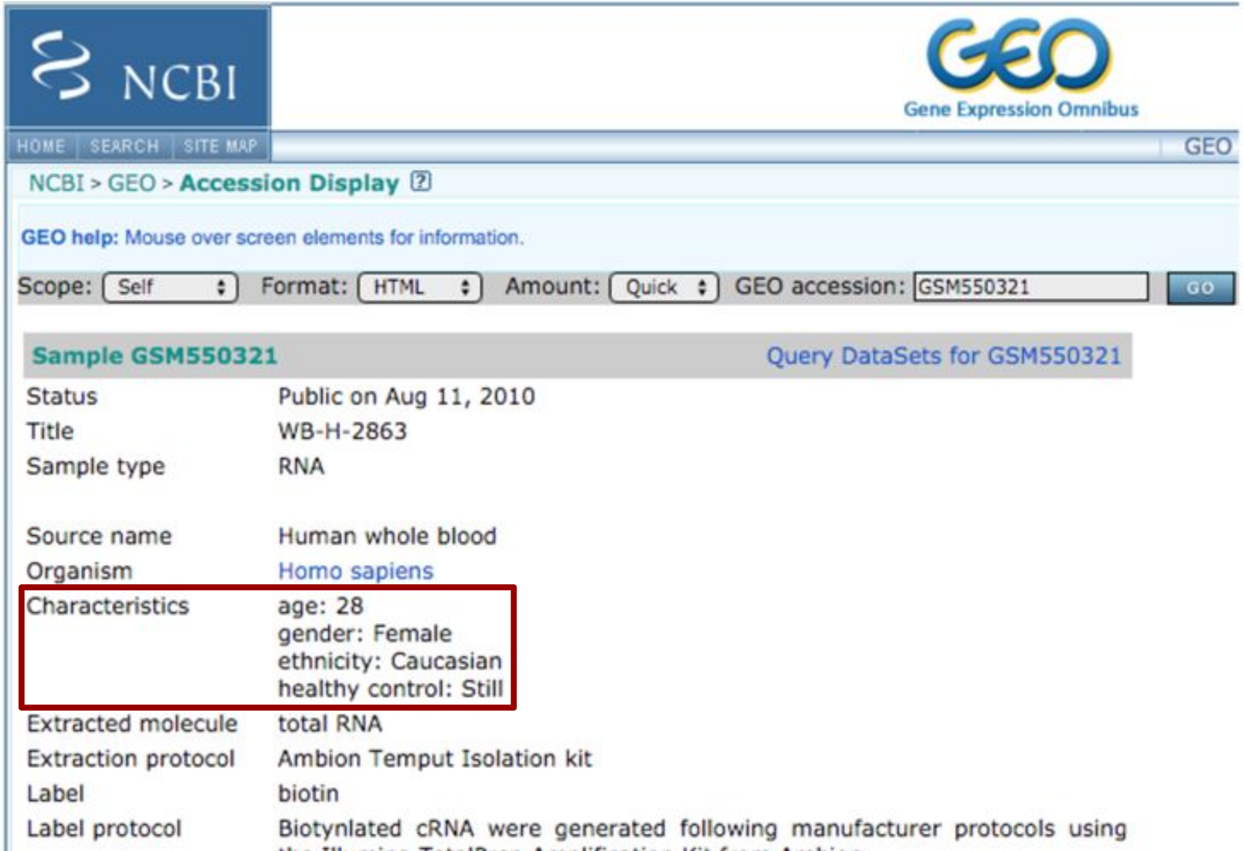
1. Detecting Linked Data Quality Issues in DBpedia
2. Completing Missing Values in Knowledge Graphs
3. **Crowdsourcing Biomedical Metadata Quality Assessment**
4. Other Applications (Open Discussion)
5. Summary & Final Remarks

Crowdsourcing Biomedical Data Quality Assessment

Hypothesis:

Crowdsourcing i.e. *non-expert workers* can be used to curate large-scale digital biomedical data on the Web.

Use Case -- Gene Expression Metadata Quality Issues



NCBI > GEO > **Accession Display**

GEO help: Mouse over screen elements for information.

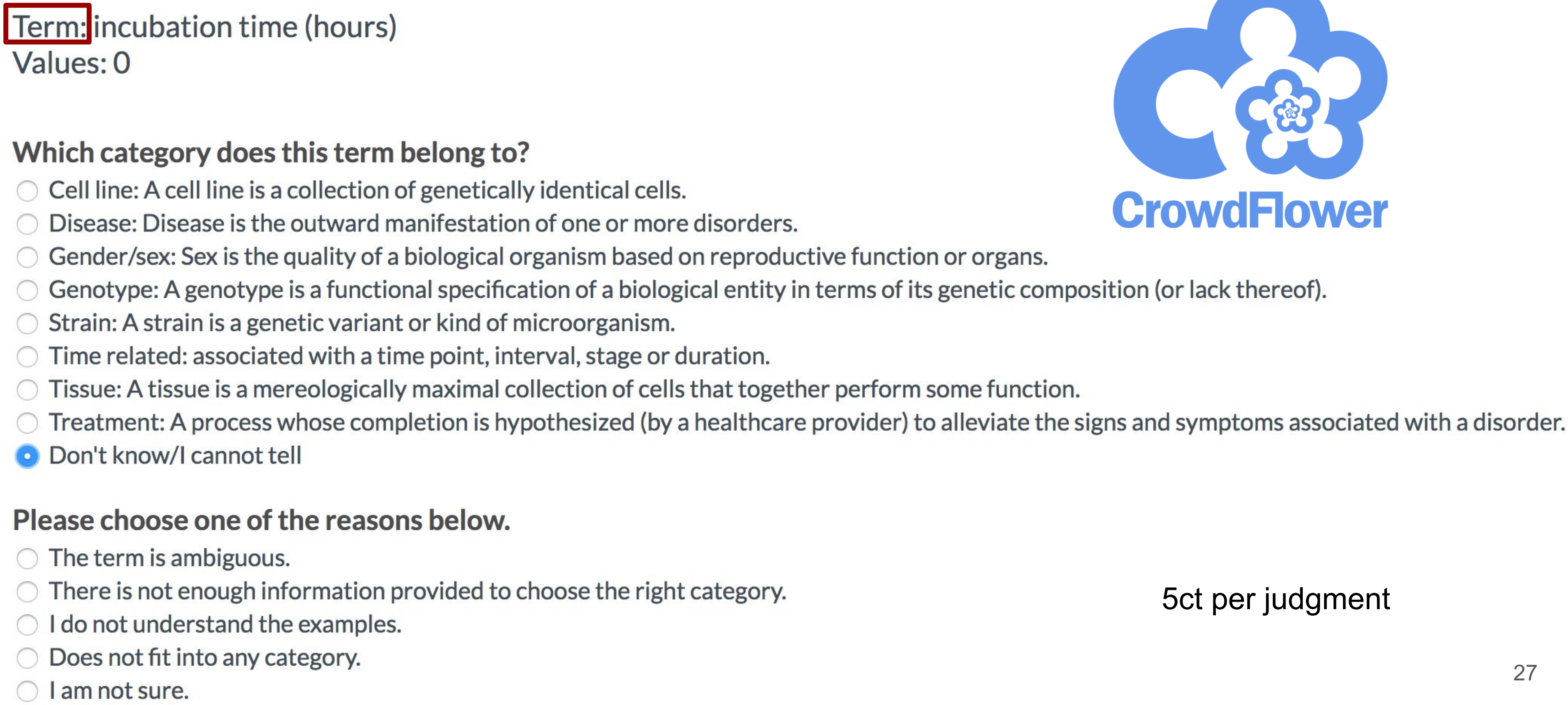
Scope: Self   Format: HTML   Amount: Quick   GEO accession: GSM550321   GO

Sample GSM550321   Query DataSets for GSM550321

Status	Public on Aug 11, 2010
Title	WB-H-2863
Sample type	RNA
Source name	Human whole blood
Organism	Homo sapiens
Characteristics	age: 28 gender: Female ethnicity: Caucasian healthy control: Still
Extracted molecule	total RNA
Extraction protocol	Ambion Temput Isolation kit
Label	biotin
Label protocol	Biotynlated cRNA were generated following manufacturer protocols using the Illumina TotalPrep Amplification Kit from Ambion

Key

Crowdsourcing GEO Metadata Keys -- Microtask



Term: incubation time (hours)
Values: 0

Which category does this term belong to?

- ○ Cell line: A cell line is a collection of genetically identical cells.
- ○ Disease: Disease is the outward manifestation of one or more disorders.
- ○ Gender/sex: Sex is the quality of a biological organism based on reproductive function or organs.
- ○ Genotype: A genotype is a functional specification of a biological entity in terms of its genetic composition (or lack thereof).
- ○ Strain: A strain is a genetic variant or kind of microorganism.
- ○ Time related: associated with a time point, interval, stage or duration.
- ○ Tissue: A tissue is a mereologically maximal collection of cells that together perform some function.
- ○ Treatment: A process whose completion is hypothesized (by a healthcare provider) to alleviate the signs and symptoms associated with a disorder.
- ● Don't know/I cannot tell

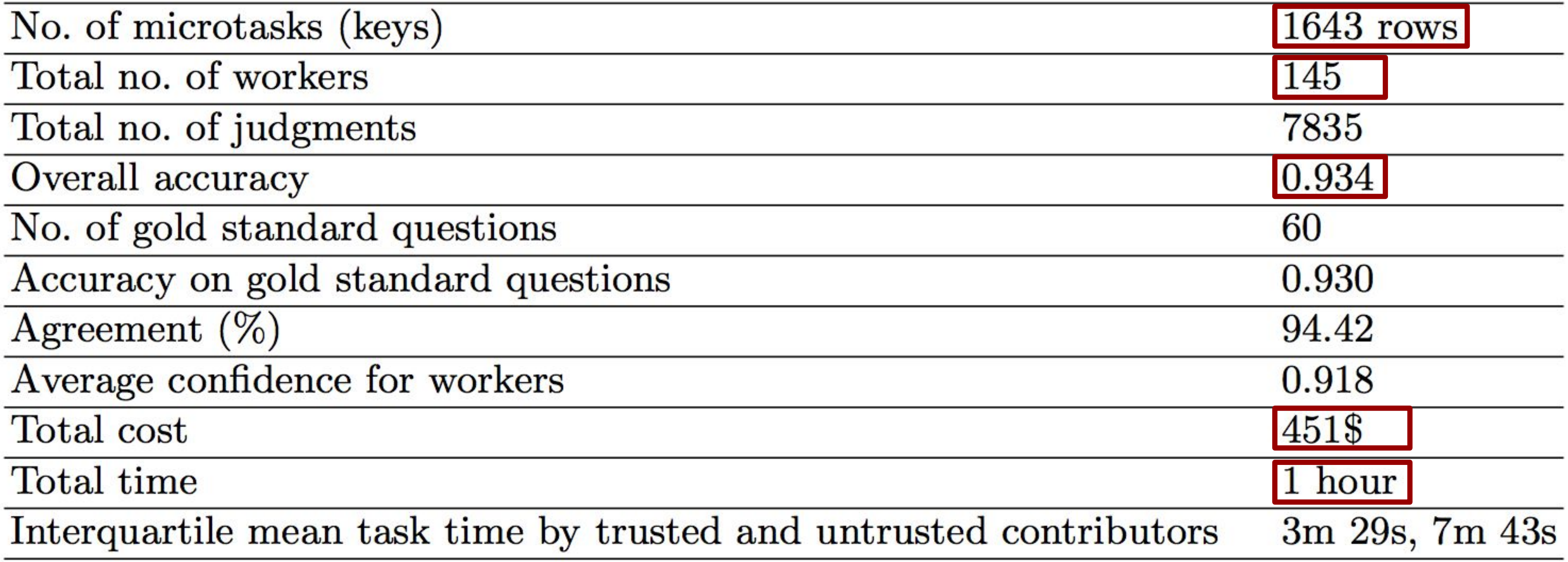
Please choose one of the reasons below.

- ○ The term is ambiguous.
- ○ There is not enough information provided to choose the right category.
- ○ I do not understand the examples.
- ○ Does not fit into any category.
- ○ I am not sure.

CrowdFlower

5ct per judgment

Crowdsourcing GEO Metadata Keys -- Results



No. of microtasks (keys)	1643 rows
Total no. of workers	145
Total no. of judgments	7835
Overall accuracy	0.934
No. of gold standard questions	60
Accuracy on gold standard questions	0.930
Agreement (%)	94.42
Average confidence for workers	0.918
Total cost	451$
Total time	1 hour
Interquartile mean task time by trusted and untrusted contributors	3m 29s, 7m 43s

Outline

1. Detecting Linked Data Quality Issues in DBpedia
2. Completing Missing Values in Knowledge Graphs
3. Crowdsourcing Biomedical Metadata Quality Assessment
4. **Other Applications (Open Discussion)**
5. Final Remarks

Other Applications

Outline

1. Detecting Linked Data Quality Issues in DBpedia
2. Completing Missing Values in Knowledge Graphs
3. Crowdsourcing Biomedical Metadata Quality Assessment
4. Other Applications (Open Discussion)
5. **Final Remarks**

Final Remarks

- Main assumption for applying microtask crowdsourcing:
  the problem can be **divided** into microtasks

- Exploiting the right **incentives** for increasing the quality of crowd answers:
  - Monetary rewards are not the only incentives for the crowd
  - Altruism and fun can be other incentives in microtask crowdsourcing

- Combining **machine-human solutions** produce high quality results

- Consider **ethical conditions**: we are working with **people**

Feedback Please !

https://bit.ly/2HhxvYi