Lessons learned building machine learning models for Wikidata

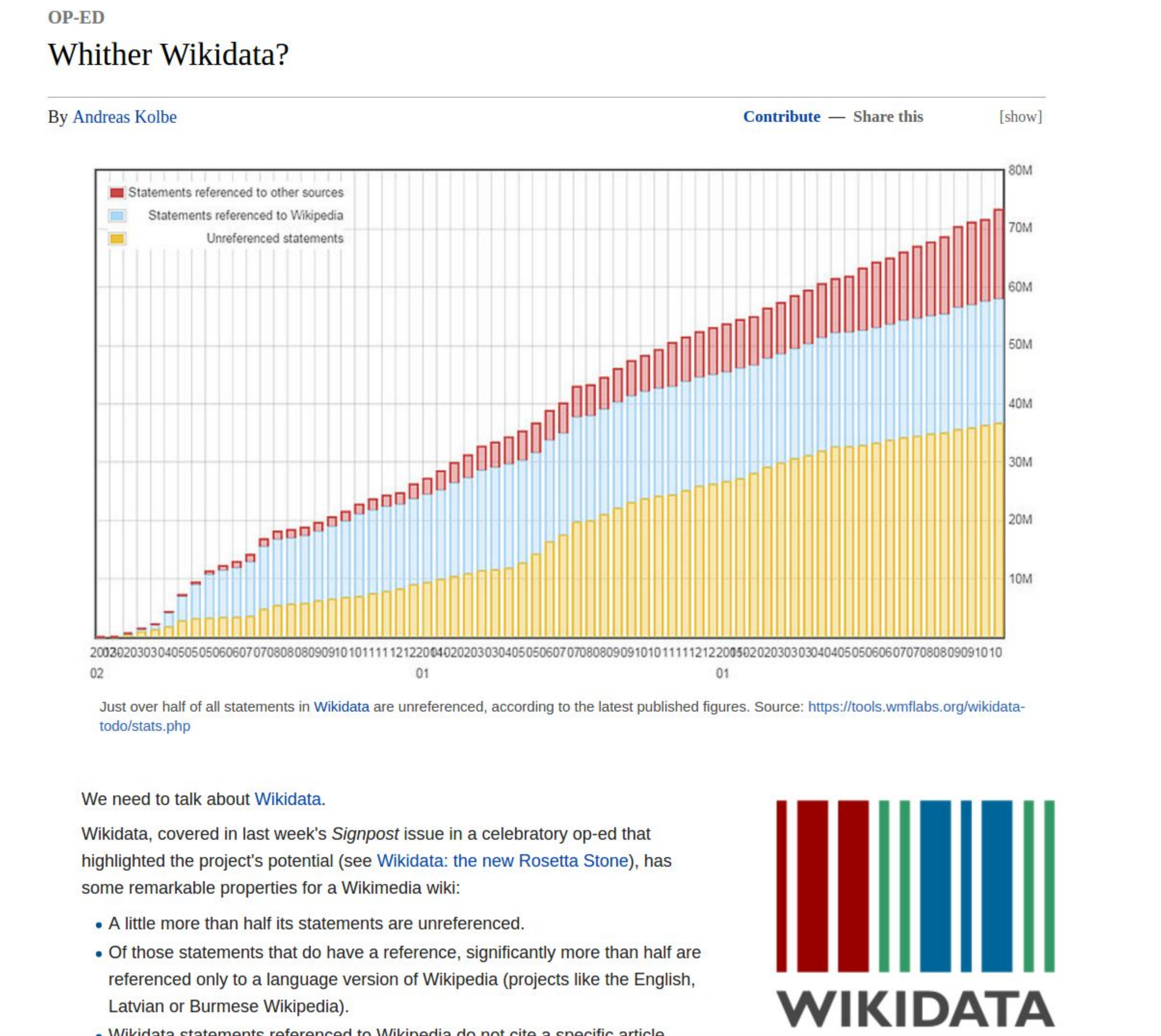
Wikimania 2016, Esino Lario, 24.06.20016
Amir Sarabadani User:Ladsgroup
@AmirSarabadani

Wikidata?

- free knowledge base by the Wikimedia movement
- structured data
- user created content
- linked data
- people, places, events, …
- powered by Wikibase (open source)
- data: CC-0

Quality control in Wikidata

- Quality of Wikidata is important
- Vandalism is a threat!
- But the backlog is too big



OP-ED

Whither Wikidata?

By Andreas Kolbe

Contribute — **Share this** [show]

Just over half of all statements in Wikidata are unreferenced, according to the latest published figures. Source: https://tools.wmflabs.org/wikidata-todo/stats.php

We need to talk about Wikidata.

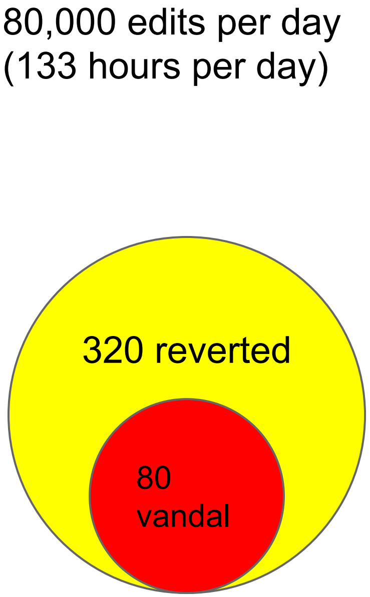
Wikidata, covered in last week's *Signpost* issue in a celebratory op-ed that highlighted the project's potential (see Wikidata: the new Rosetta Stone), has some remarkable properties for a Wikimedia wiki:

- A little more than half its statements are unreferenced.
- Of those statements that do have a reference, significantly more than half are referenced only to a language version of Wikipedia (projects like the English, Latvian or Burmese Wikipedia).
- Wikidata statements referenced to Wikipedia do not cite a specific article

WIKIDATA

Wikidata gets 80,000 human edits per day, 320 will be reverted and 80 of them will have been vandalism.

0.4% of edits are reverted, 0.1% of reverted edits are vandalism



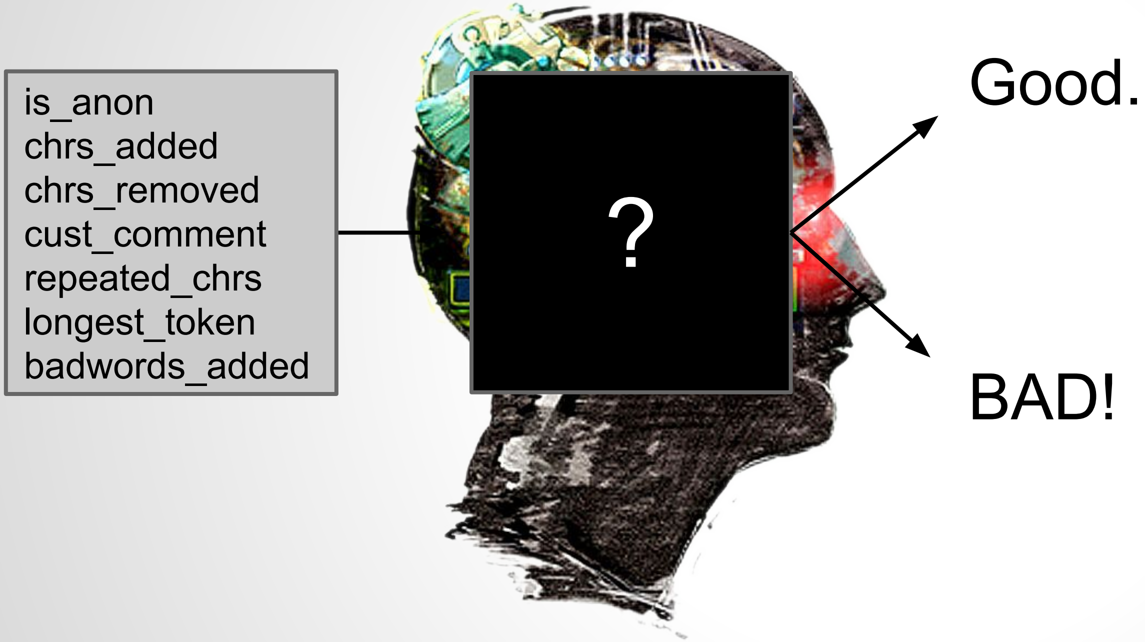
https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2015-12-09/Op-ed

“For Wikidata to truly give more people more access to more knowledge, the data in Wikidata needs to be of high quality.”

“And probably the most important part is machine-learning tools like ORES that help us find bad edits and other issues.”

Machine learning classifiers

- Machine learning in Wikipedia
- But there is no infrastructure
- Here it comes the ORES



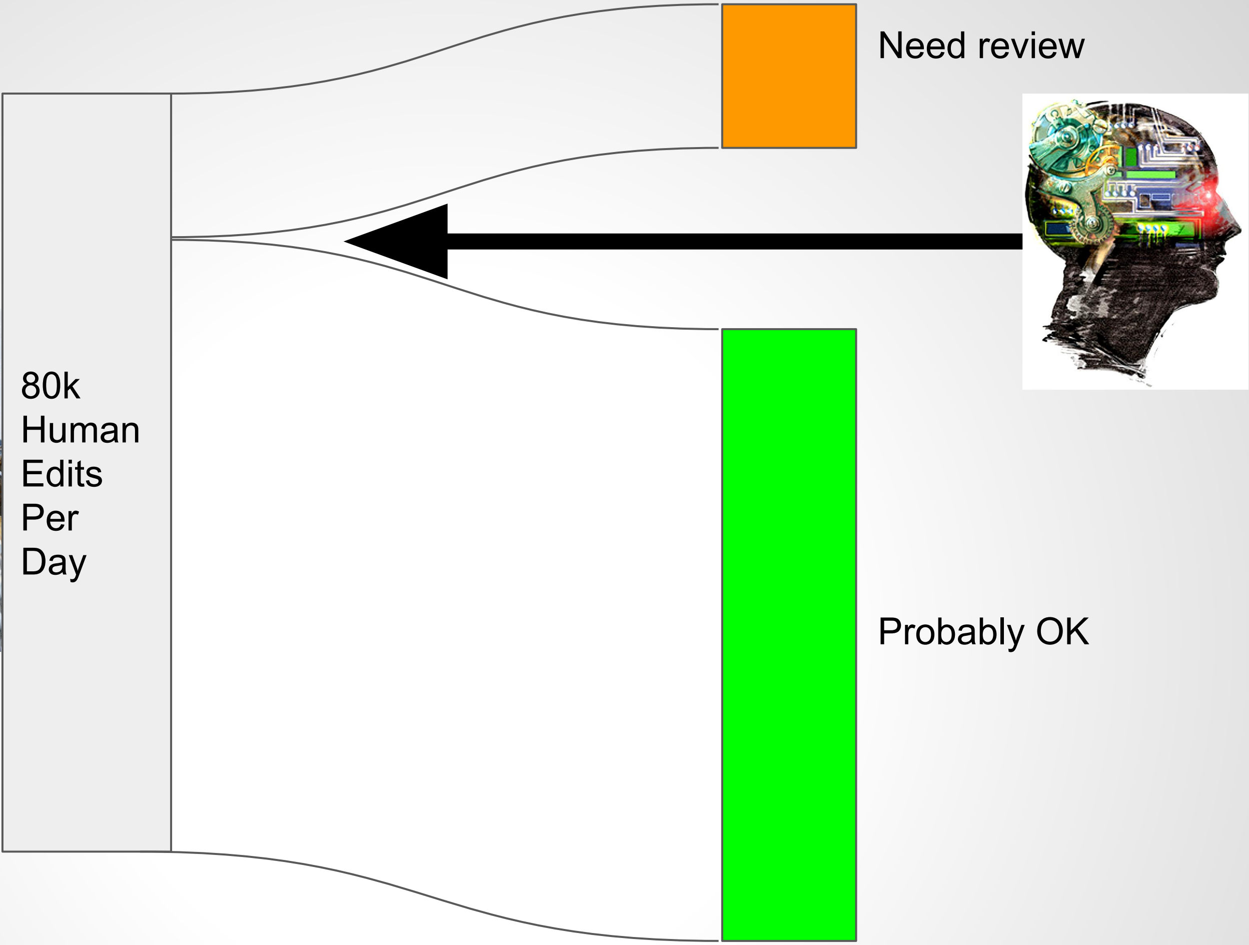
is_anon
chrs_added
chrs_removed
cust_comment
repeated_chrs
longest_token
badwords_added

?

Good.

BAD!

https://commons.wikimedia.org/wiki/File:Flickr_-_Official_U.S._Navy_Imagery_-_Sailor%27s_daughter_operates_a_fire_hose_with_crew_member_assistance..jpg

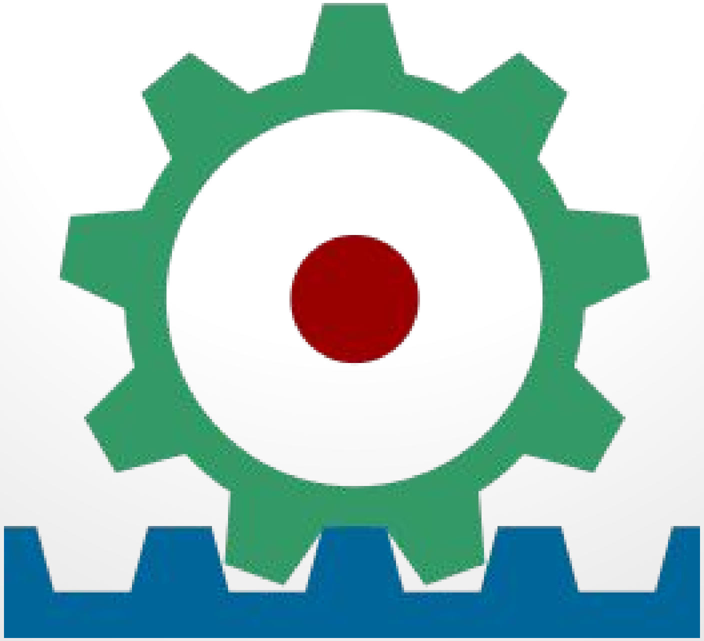


80k Human Edits Per Day

Need review

Probably OK

ORES



https://ores.wikimedia.org/scores/wikidatawiki/damaging/267233560

https://ores.wikimedia.org/scores/wikidatawiki/damaging/267233560

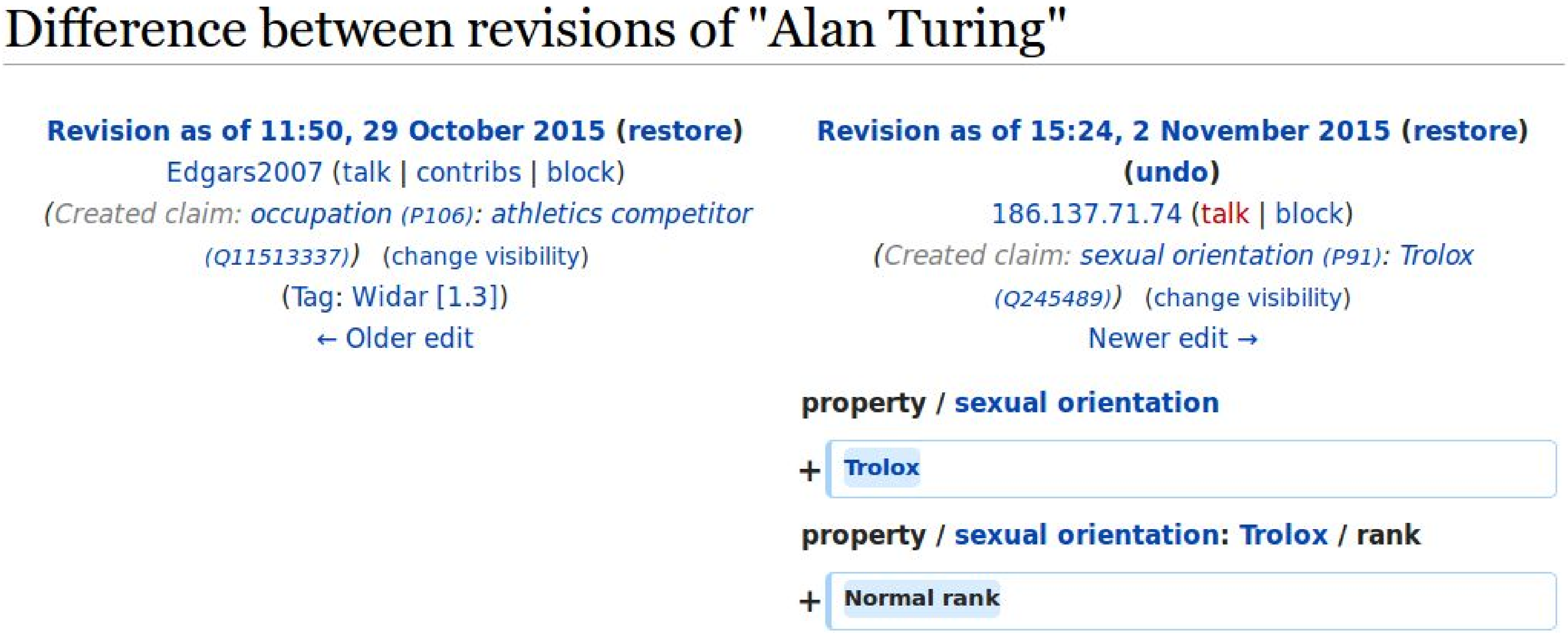
“Wikidata”

https://ores.wikimedia.org/scores/wikidatawiki/damaging/267233560

Is this edit damaging?

https://ores.wikimedia.org/scores/wikidatawiki/damaging/267233560

https://ores.wikimedia.org/scores/wikidatawiki/damaging/267233560



Difference between revisions of "Alan Turing"

Revision as of 11:50, 29 October 2015 (restore)
Edgars2007 (talk | contribs | block)
(Created claim: occupation (P106): athletics competitor (Q11513337)) (change visibility)
(Tag: Widar [1.3])
← Older edit

Revision as of 15:24, 2 November 2015 (restore) (undo)
186.137.71.74 (talk | block)
(Created claim: sexual orientation (P91): Trolox (Q245489)) (change visibility)
Newer edit →

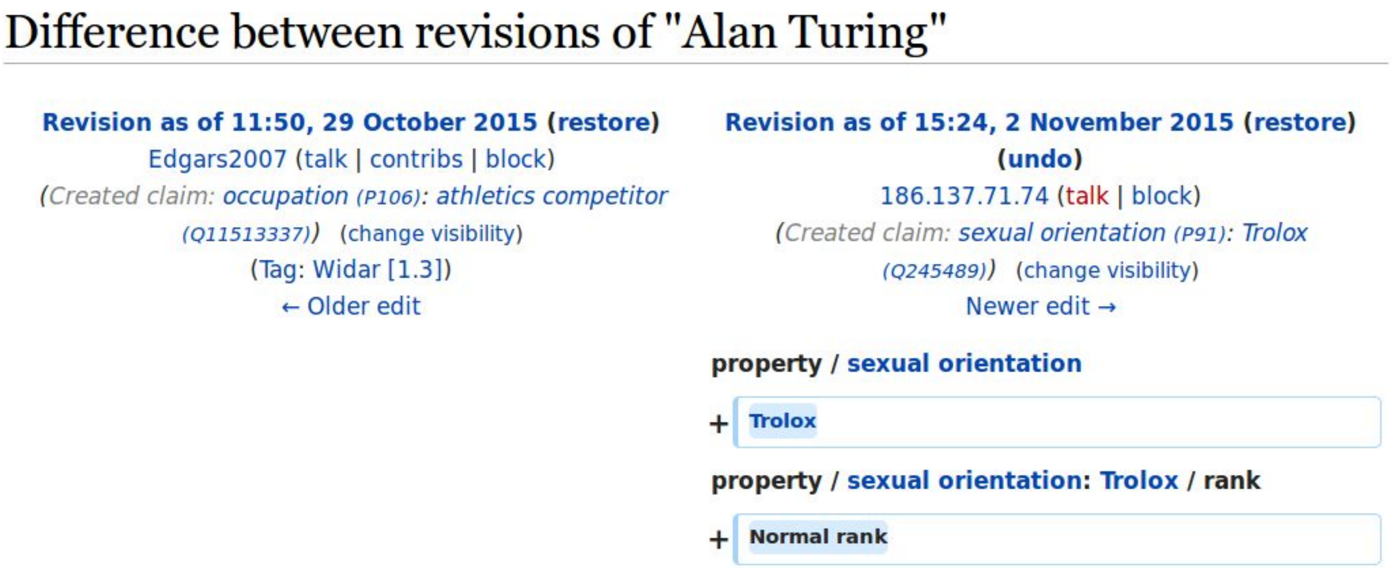
property / sexual orientation

+ **Trolox**

property / sexual orientation: Trolox / rank

+ **Normal rank**

https://ores.wikimedia.org/scores/wikidatawiki/damaging/267233560



Difference between revisions of "Alan Turing"

Revision as of 11:50, 29 October 2015 (restore)
Edgars2007 (talk | contribs | block)
(Created claim: occupation (P106): athletics competitor (Q11513337)) (change visibility)
(Tag: Widar [1.3])
← Older edit

Revision as of 15:24, 2 November 2015 (restore) (undo)
186.137.71.74 (talk | block)
(Created claim: sexual orientation (P91): Trolox (Q245489)) (change visibility)
Newer edit →

property / sexual orientation

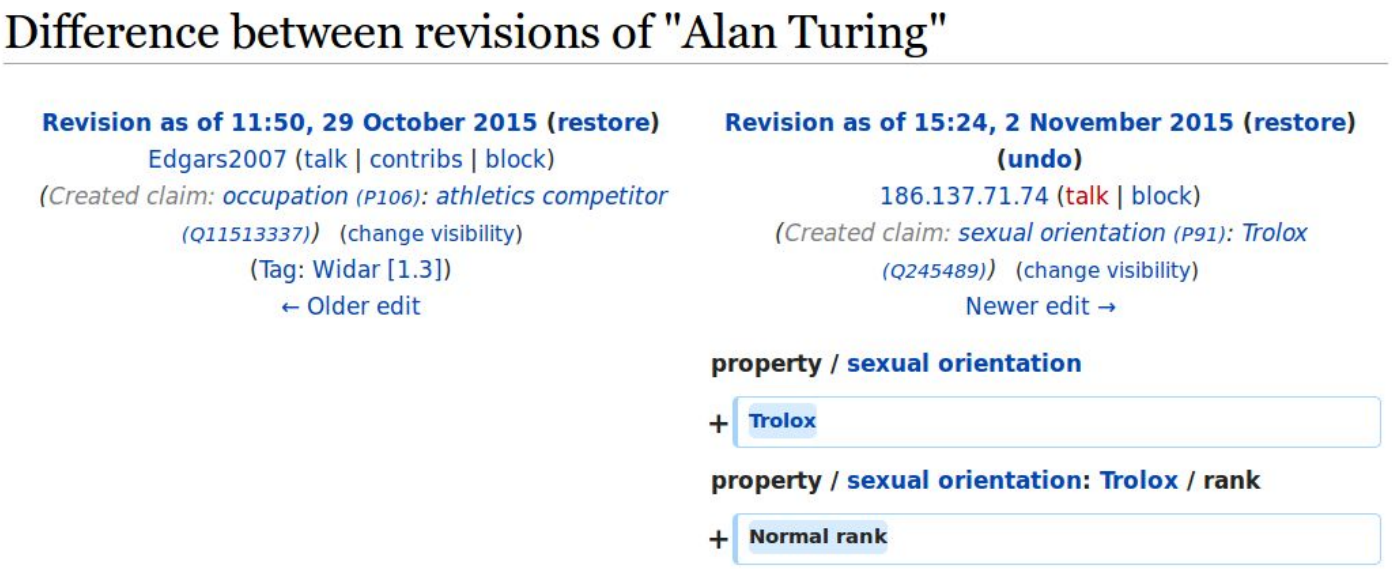
+ **Trolox**

property / sexual orientation: Trolox / rank

+ **Normal rank**

```
{
  "267233560": {
    "prediction": true,
    "probability": {
      "false": 0.111,
      "true": 0.889
    }
  }
}
```

https://ores.wikimedia.org/scores/wikidatawiki/damaging/267233560



Difference between revisions of "Alan Turing"

Revision as of 11:50, 29 October 2015 (restore)	Revision as of 15:24, 2 November 2015 (restore) (undo)
Edgars2007 (talk | contribs | block)	186.137.71.74 (talk | block)
(Created claim: occupation (P106): athletics competitor (Q11513337)) (change visibility)	(Created claim: sexual orientation (P91): Trolox (Q245489)) (change visibility)
(Tag: Widar [1.3])	
← Older edit	Newer edit →

property / sexual orientation

+ Trolox

property / sexual orientation: Trolox / rank

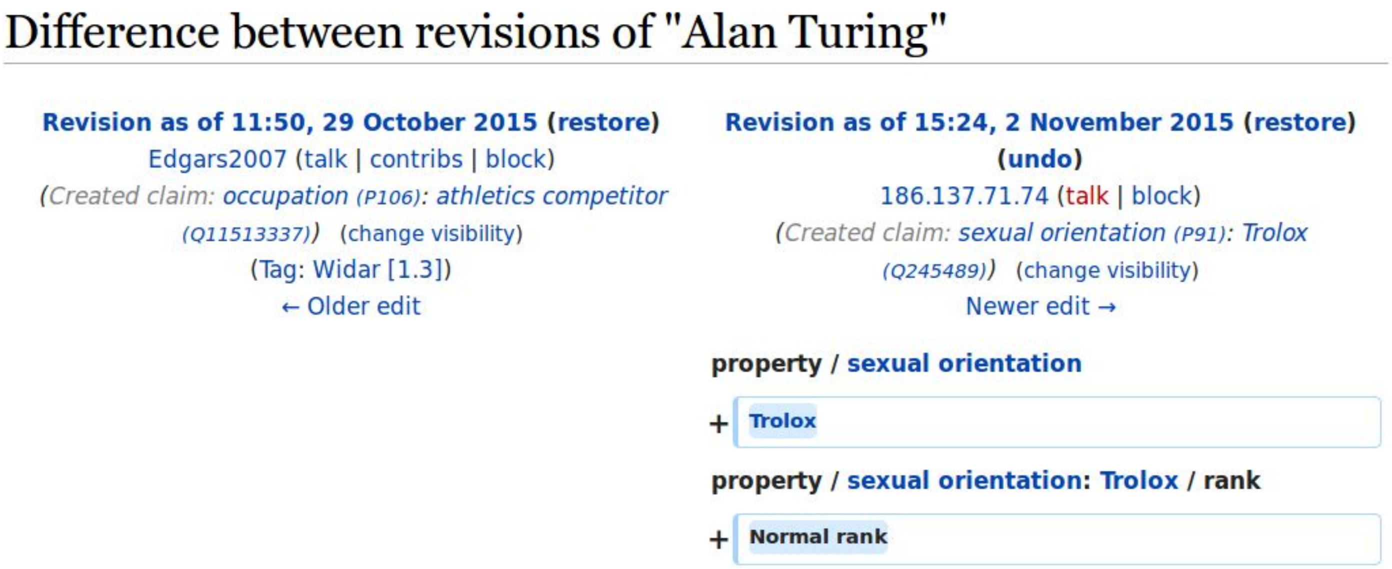
+ Normal rank

```
{
  "267233560": {
    "prediction": true,
    "probability": {
      "false": 0.111,
      "true": 0.889
    }
  }
}
```

https://ores.wikimedia.org/scores/wikidatawiki/damaging/286961313

```
{
  "286961313": {
    "prediction": false,
    "probability": {
      "false": 0.946,
      "true": 0.054
    }
  }
}
```

https://ores.wikimedia.org/scores/wikidatawiki/damaging/267233560



Difference between revisions of "Alan Turing"

Revision as of 11:50, 29 October 2015 (restore)
Edgars2007 (talk | contribs | block)
(*Created claim: occupation (P106): athletics competitor (Q11513337)*) (change visibility)
(Tag: Widar [1.3])
← Older edit

Revision as of 15:24, 2 November 2015 (restore) (undo)
186.137.71.74 (talk | block)
(*Created claim: sexual orientation (P91): Trolox (Q245489)*) (change visibility)
Newer edit →

property / sexual orientation

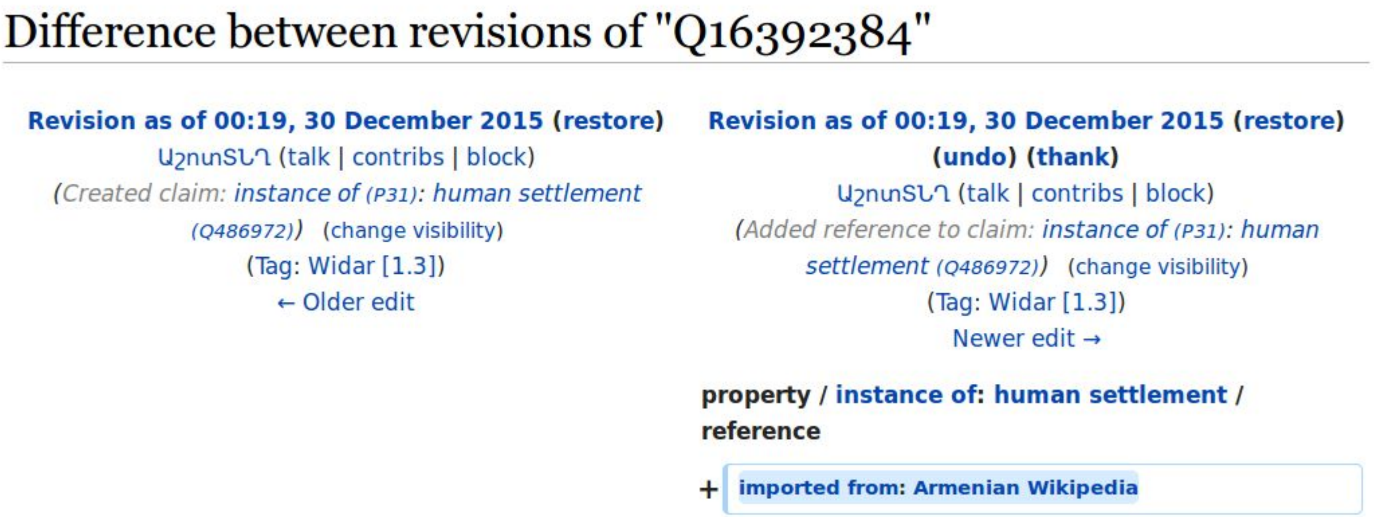
+ **Trolox**

property / sexual orientation: Trolox / rank

+ **Normal rank**

```
{
  "267233560": {
    "prediction": true,
    "probability": {
      "false": 0.111,
      "true": 0.889
    }
  }
}
```

https://ores.wikimedia.org/scores/wikidatawiki/damaging/286961313



Difference between revisions of "Q16392384"

Revision as of 00:19, 30 December 2015 (restore)
ԱշոտՏՆՂ (talk | contribs | block)
(*Created claim: instance of (P31): human settlement (Q486972)*) (change visibility)
(Tag: Widar [1.3])
← Older edit

Revision as of 00:19, 30 December 2015 (restore) (undo) (thank)
ԱշոտՏՆՂ (talk | contribs | block)
(*Added reference to claim: instance of (P31): human settlement (Q486972)*) (change visibility)
(Tag: Widar [1.3])
Newer edit →

property / instance of: human settlement / reference

+ **imported from: Armenian Wikipedia**

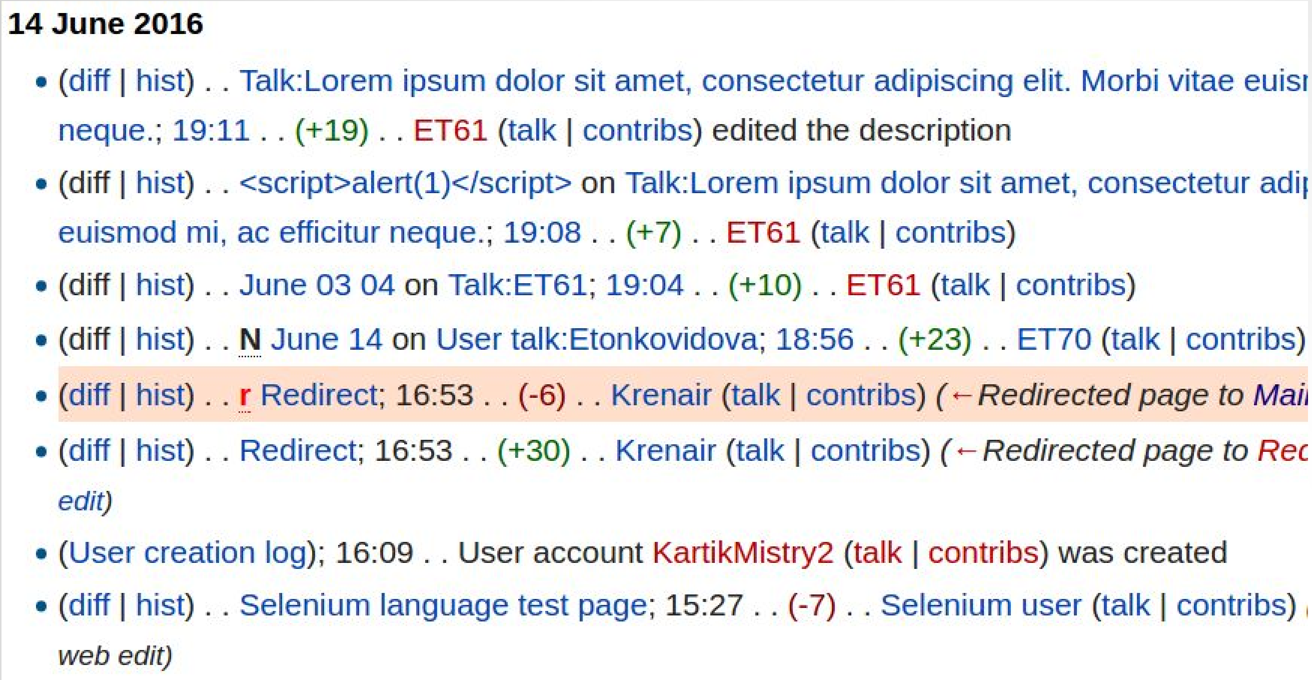
```
{
  "286961313": {
    "prediction": false,
    "probability": {
      "false": 0.946,
      "true": 0.054
    }
  }
}
```

ORES in Wikipedia vs. Wikidata

- Wikitext-based vandalism detection
- Wikibase-based vandalism detection
- Percentage of vandalism
  Wikipedia: 1-10% (Mostly ~7%)
  Wikidata: ~0.1%

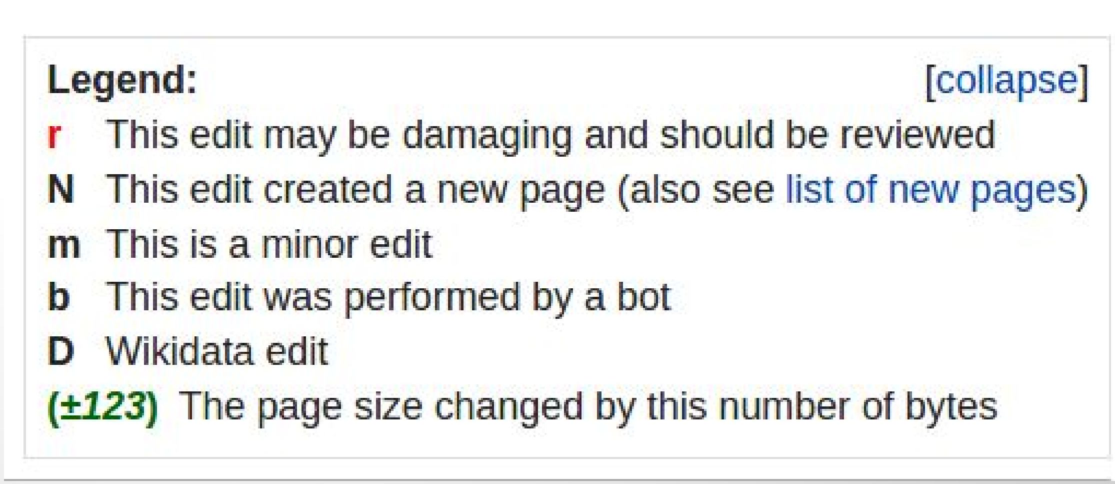
ORES extension

- Database cached = speed and filtering
- Hit record



14 June 2016

- (diff | hist) . . Talk:Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi vitae euisı neque.; 19:11 . . (+19) . . ET61 (talk | contribs) edited the description
- (diff | hist) . . <script>alert(1)</script> on Talk:Lorem ipsum dolor sit amet, consectetur adip euismod mi, ac efficitur neque.; 19:08 . . (+7) . . ET61 (talk | contribs)
- (diff | hist) . . June 03 04 on Talk:ET61; 19:04 . . (+10) . . ET61 (talk | contribs)
- (diff | hist) . . **N** June 14 on User talk:Etonkovidova; 18:56 . . (+23) . . ET70 (talk | contribs)
- (diff | hist) . . **r** Redirect; 16:53 . . (-6) . . Krenair (talk | contribs) *(←Redirected page to Mai*
- (diff | hist) . . Redirect; 16:53 . . (+30) . . Krenair (talk | contribs) *(←Redirected page to Rec edit)*
- (User creation log); 16:09 . . User account KartikMistry2 (talk | contribs) was created
- (diff | hist) . . Selenium language test page; 15:27 . . (-7) . . Selenium user (talk | contribs) *web edit)*



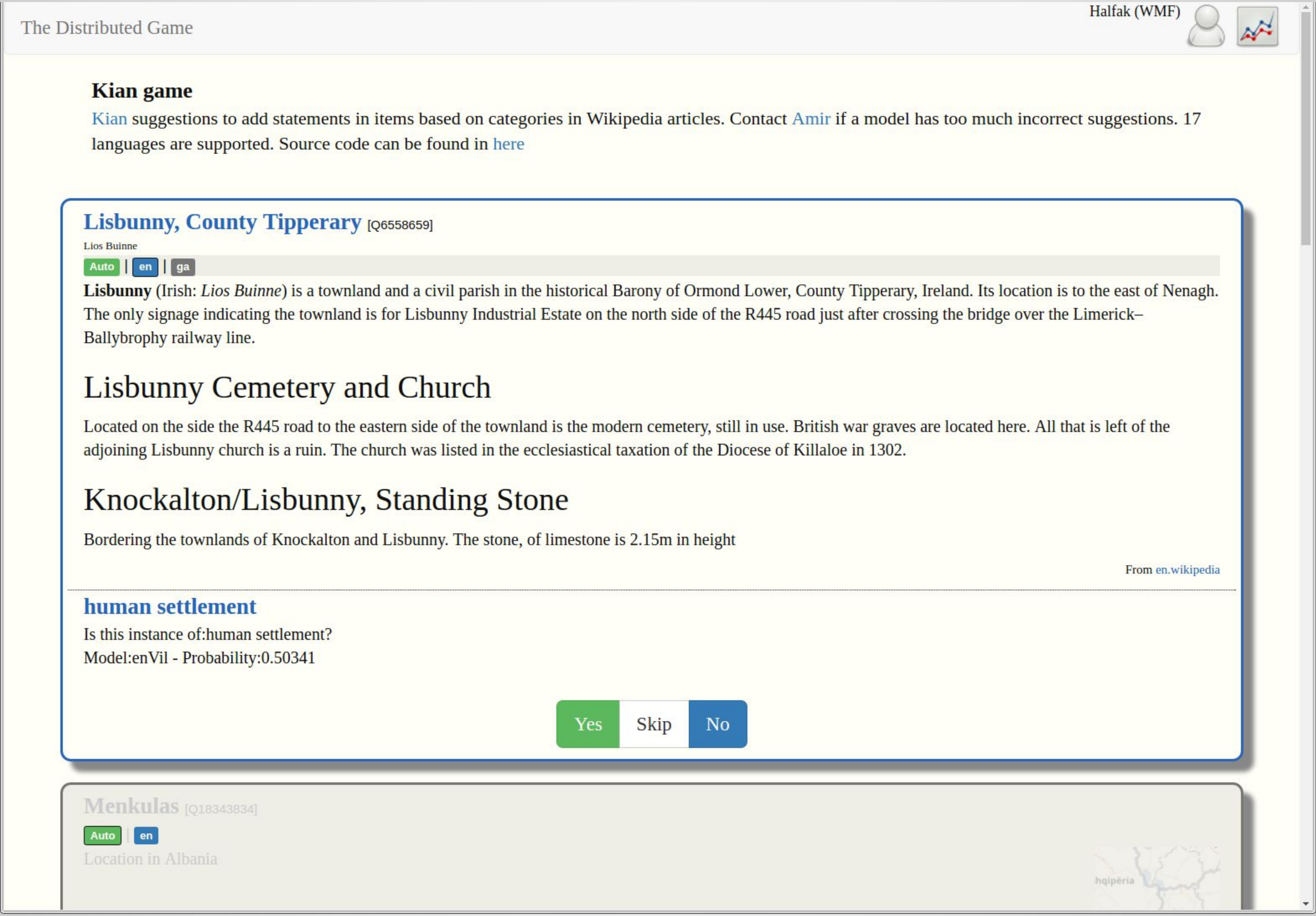
Legend: [collapse]

r	This edit may be damaging and should be reviewed
N	This edit created a new page (also see list of new pages)
m	This is a minor edit
b	This edit was performed by a bot
D	Wikidata edit
(*±123*)	The page size changed by this number of bytes

Kian

Problem of harvesting data

- Different languages
- Type of statements
- Why NLP and categories are not options



Kian game

Kian suggestions to add statements in items based on categories in Wikipedia articles. Contact Amir if a model has too much incorrect suggestions. 17 languages are supported. Source code can be found in here

Lisbunny, County Tipperary [Q6558659]

Lios Buinne

Lisbunny (Irish: *Lios Buinne*) is a townland and a civil parish in the historical Barony of Ormond Lower, County Tipperary, Ireland. Its location is to the east of Nenagh. The only signage indicating the townland is for Lisbunny Industrial Estate on the north side of the R445 road just after crossing the bridge over the Limerick–Ballybrophy railway line.

Lisbunny Cemetery and Church

Located on the side the R445 road to the eastern side of the townland is the modern cemetery, still in use. British war graves are located here. All that is left of the adjoining Lisbunny church is a ruin. The church was listed in the ecclesiastical taxation of the Diocese of Killaloe in 1302.

Knockalton/Lisbunny, Standing Stone

Bordering the townlands of Knockalton and Lisbunny. The stone, of limestone is 2.15m in height

From en.wikipedia

human settlement

Is this instance of:human settlement?

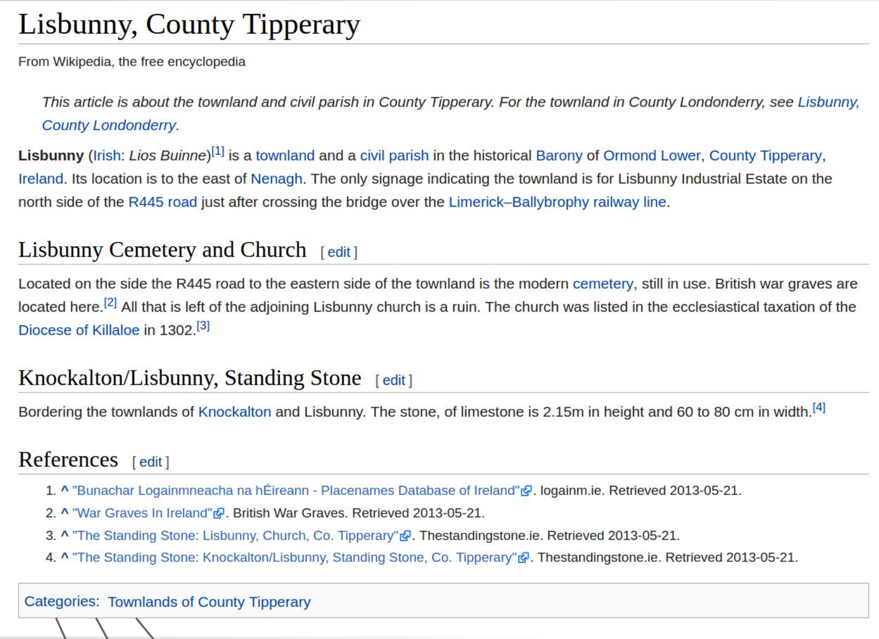
Model:enVil - Probability:0.50341

Menkulas [Q18343834]

Location in Albania

How Kian works

- Extracting features
- Training



Lisbunny, County Tipperary

From Wikipedia, the free encyclopedia

This article is about the townland and civil parish in County Tipperary. For the townland in County Londonderry, see Lisbunny, County Londonderry.

Lisbunny (Irish: *Lios Buinne*)[1] is a townland and a civil parish in the historical Barony of Ormond Lower, County Tipperary, Ireland. Its location is to the east of Nenagh. The only signage indicating the townland is for Lisbunny Industrial Estate on the north side of the R445 road just after crossing the bridge over the Limerick–Ballybrophy railway line.

Lisbunny Cemetery and Church [edit]

Located on the side the R445 road to the eastern side of the townland is the modern cemetery, still in use. British war graves are located here.[2] All that is left of the adjoining Lisbunny church is a ruin. The church was listed in the ecclesiastical taxation of the Diocese of Killaloe in 1302.[3]

Knockalton/Lisbunny, Standing Stone [edit]

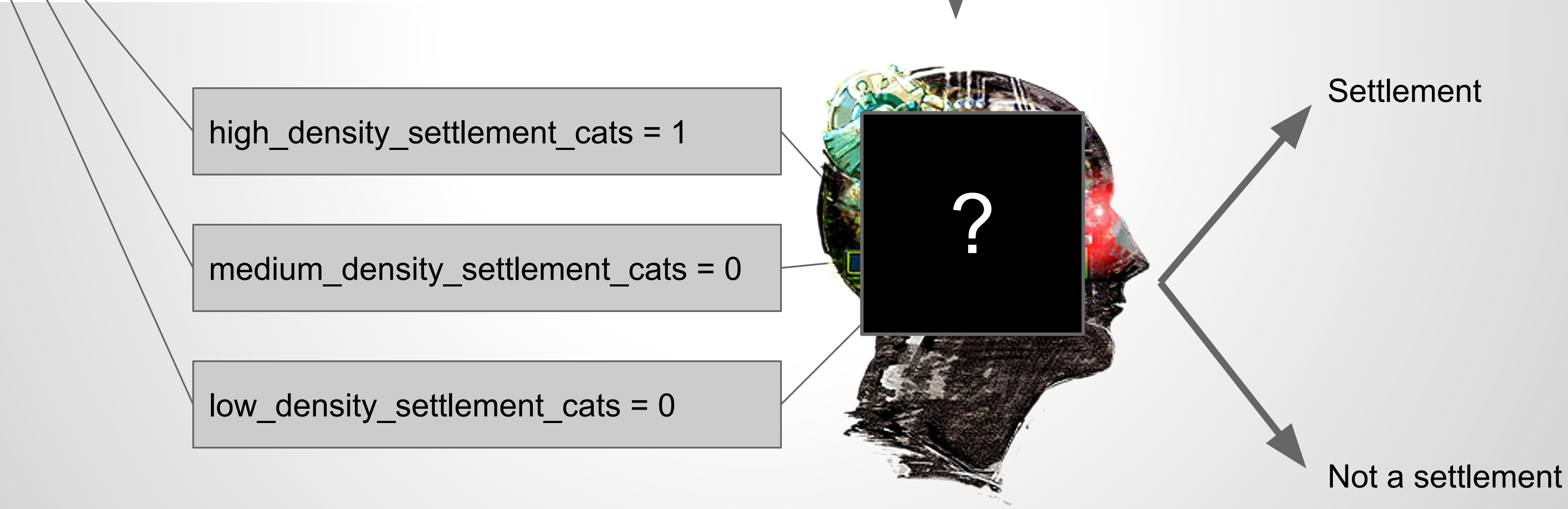
Bordering the townlands of Knockalton and Lisbunny. The stone, of limestone is 2.15m in height and 60 to 80 cm in width.[4]

References [edit]

1. ^ "Bunachar Logainmneacha na hÉireann - Placenames Database of Ireland". logainm.ie. Retrieved 2013-05-21.
2. ^ "War Graves In Ireland". British War Graves. Retrieved 2013-05-21.
3. ^ "The Standing Stone: Lisbunny, Church, Co. Tipperary". Thestandingstone.ie. Retrieved 2013-05-21.
4. ^ "The Standing Stone: Knockalton/Lisbunny, Standing Stone, Co. Tipperary". Thestandingstone.ie. Retrieved 2013-05-21.

Categories: Townlands of County Tipperary

Kian



high_density_settlement_cats = 1

medium_density_settlement_cats = 0

low_density_settlement_cats = 0

Questions?

Revision Scoring team:

Aaron Halfaker (ahalfaker@wikimedia.org)
Amir Sarabadani

Me?

User:Ladsgroup
Ladsgroup@gmail.com
@AmirSarabadani (twitter)