[[:m:WikiLoop]]

WikiLoop at WikiDataCon

Contribute Knowledge to Open Knowledge World



zzn@, chaoyuel@, elanhr@, vrandecic@, tianyigeng@ with input from bianjiang@, weizhao@, wangyu@, banksb@ Copyright License: CC-BY-SA 2.0

[[:m:WikiLoop]]

Agenda

- Background
- Featured Project: WikiLoop Battlefield
- Portfolio Connections
- Discussion

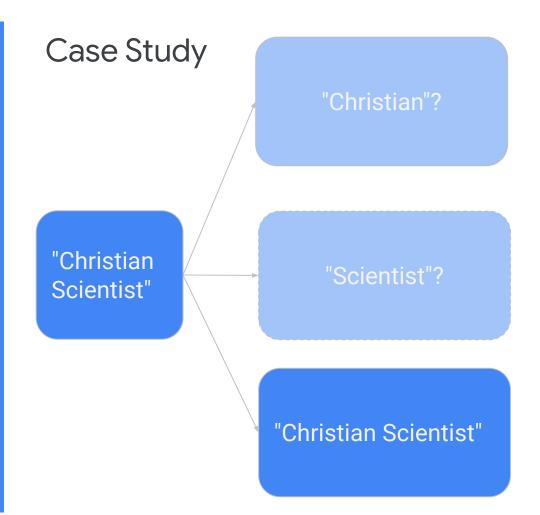
Background & Introduction

"Bezos can't afford a car" (Italian)



Sometimes **there are limitations** of how far algorithms and machine learning can tell

the "Bezos can't afford a car" vandalism case



Robot: how would I know?





Human: how do I fix them all?

Intention: Empower Human, Bootstrap Machines



"I can do it, but only some amount of time!"

"I can learn from you and do more"



the missing link in the data loop

We rely on the **knowledge sources** to be healthy.

We can and we need to help.



Google

(and any knowledge engines depending on open knowledge)

Wiki(p|m)edia

(and other open knowledge communities and movements)

[[:m:WikiLoop]]

Originated from Google, WikiLoop is an umbrella program for a series of **technical projects** intended to contribute **datasets and toolings** from technical industry back to open knowledge world.

What projects fits in WikiLoop



Source Improvements



Maximize Neutrality



Empower Human Bootstrap Machine

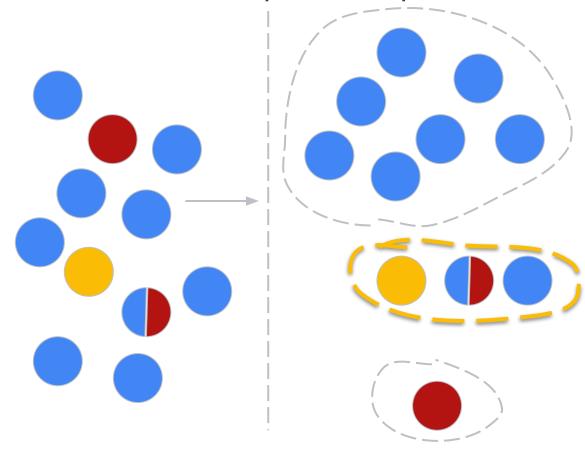
What kinds of projects fits in WikiLoop landscape



Source Improvements

In the area that algorithms has saturated the extraction maximum, the source noise becomes the dominant bottleneck

WikiLoop provide a platform to empower the community to reduce noise from source itself



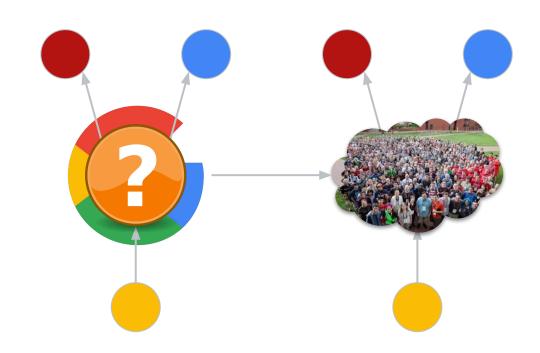
What projects fits in WikiLoop



Maximize Neutrality

In the areas when opinions are largely divided and controversial, any fact/opinion choice Google made could dissatisfy one side and deemed "biased".

WikiLoop provides a means to obtain consensus and maximize neutrality by relying on a peer-reviewed self-governed community.



What projects fits in WikiLoop



Desired Al precision level



"Bootstrap effect"

In the areas due to lack of training data, certain Al-based services are not good enough to serve general Google users and use-cases, but community members are doing hand-crafting creation anyway.

WikiLoop provides a platform to provide machine assistance for such knowledge creation, while imperfect, improving their productivity, and then send back training data to improve Al models, hopefully to a point that services quality will improve and bootstrap in this process

Current Al precision level



Al available to users (none)



WikiLoop Portfolio



WikiLoop **Bot** Deliver automatic communication





WikiLoop **Explorer** Browse all WikiLoop dataset portfolio



name TBD WikiLoop Battlefield v1

Web app for fighting vandals together



WikiLoop CatFacts From editor-marked categories to facts



WikiLoop PageRank

Is a source site considered reliable?



WikiLoop PreCheck

Crowdsource checked batch import



WikiLoop CitePool

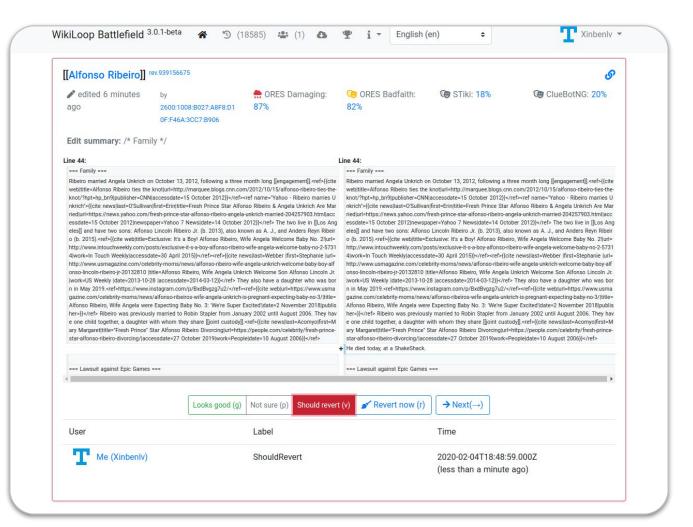
Citation candidate pool from the web



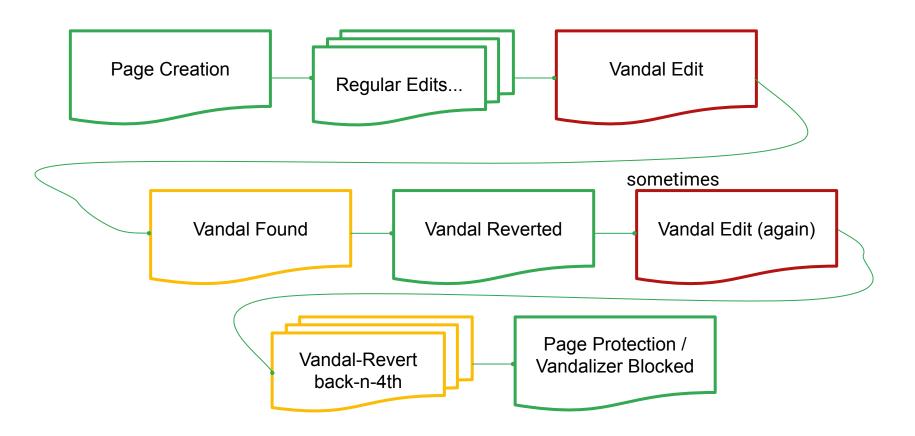
WikiLoop Babel

Bootstrap smaller language translations

WikiLoop Battlefield as an example



Wikipedia Fighting Vandalism Workflow



What is WikiLoop Battlefield and Why We Build it

for user (reviewer)

Responsive web app:

- no need to install
- works on phone as well
- can be used anywhere

Lower the barrier to review

- no need for WP:Rollback rights
- multiple people can label the same revision

Convenience features

- diff
- other people's label
- real-time
- (and building more)

for researcher

Easy to understand

- because becoming an user is easy
- reviewing a labeling is easy with diff

Easy to get data

- Client / API decoupled
- Microservice structured
- Open Sourced
- One click to download CSV

for developer

East to pickup

- repo link directly from app
- standard NodeJS setup
- documentation for quick start

Easy to build

- Continuous Integration setup
- Pull Request will generate review app

Easy to release

- standard release pipeline
- dev, staging and prod release stages

Wikipedia Community's Consideration for Tech Product



- Community Trust Comes First
- Usually Need to Be Open Sourced
- Non-Proprietary Tech Prefered
- Avoid Tech Lock-Down
- Community Approval Required for Certain Features

FrontEnd Pages

Pages

- index.vue
- marked.vue

Components

- RevisionCard
- DiffBox
- InteractionBtns.



What's Next for WikiLoop Battlefield

Grow User for Usage

Get
Developer
for Contrib.

Get Researcher for Adoption Operation / Outreach

RevCard Feed Filter and Ranking

Design for Better Research Development

Apply for WP Login

Google

WikiLoop Game

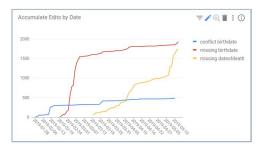
Wikidata Game

Missing Date of Death Wikipedia may contains information that wikidata doesn't have, this game helps add date of death claim to wikidata from wikipedia. · Click "Accept" to add a date of death(P570) claim to the wikidata entity, meanwhile set the wikipedia pages as references. description & instruction · Click "No" to refuse the suggestion. · If you're not sure, click "I don't know". • The suggested date of death comes from top ten wikipedia. There might be error due to outdated wikipedia page snapshots, program parsing error or wrong wikipedia info. Be sure to check the data before you make choice! • You can choose to show only entities that contains wiki reference from your primary language (the first language in your personal setting). Note that the primary language should be one of the top-ten languages [en,fr,de,zh,pt,es,ja,ru,it,pl]. · Bug reports and feedback should be sent to User:Chaoyuel. Language: Your primary language Predefined types, could use to load different dataset / generate different actions. 杨善集 [Q55707540] Auto zh main entity He was born in 1900. He studied at Communist University of the Toilers of the East. Possible date of death: Possible missing value by analyze wikijoins 1927-09-23 References: Which wikipedia page does this missing value comes from ZH Wikipedia Differenct actions will be made by different choices I don't know Accept

贺锦斋 [Q5570757

Next entity

Game Stats

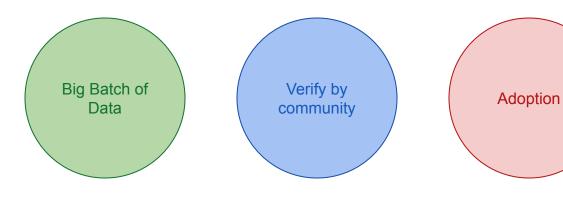






	Yes	Yes)	Others	Total	
Birthday Conflicts					483 (100%)	483	(0%)
[DEACTIVATED GAME]	132	(94%)	9	(6%)		141	(0%)
[DEACTIVATED GAME]	6	(100%)				6	(0%)
Depiction	8,766	(54%)	7,466	(46%)		16,232	(1%)
Missing Date of Birth	3,917	(98%)	74	(2%)		3,991	(0%)
[DEACTIVATED GAME]	32	(100%)				32	(0%)
Software-item matcher	61	(44%)	79	(56%)		140	(0%)
Missing Date of Death	2,525	(97%)	72	(3%)		2,597	(0%)
Born where	4,551	(56%)	3,624	(44%)		8,175	(1%)
Category Facts	872	(70%)			380 (30%)	1,252	(0%)

From WikiLoop Game - PreCheck



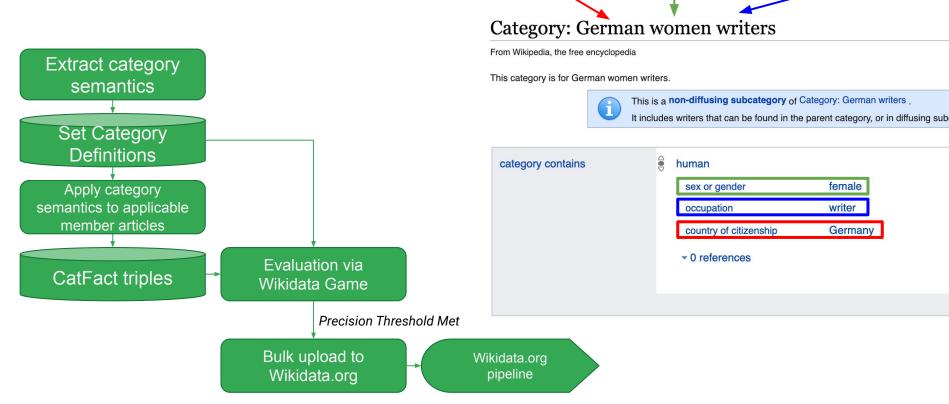


We can streamline this with
WikiLoop
PreCheck



Other Featured Projects

CatFacts (Early Stage)



nationality

gender

occupation

WikiLoop CatFacts (Early Stage)

- Uploading category-asserted facts to Wikidata.org is already semi-accepted by the community.
- Current process is not readily accessible/reviewable by the wider editor community.
- Community approach mainly uses bag-of-words which occasionally leads to some <u>unfortunate</u> <u>results</u>.
- Disproportionate impact by small number of users with few quality guarantees. We can help fix this last part.





This is embarrassing.

Google's knowledge graph draws data from various source, so I checked for the change that put Omar Kadh under "Canadian soldier."

Changes from a Russian account a month ago. \geq

TL;DR Andrew Scheer was manipulated by a Russian troll.

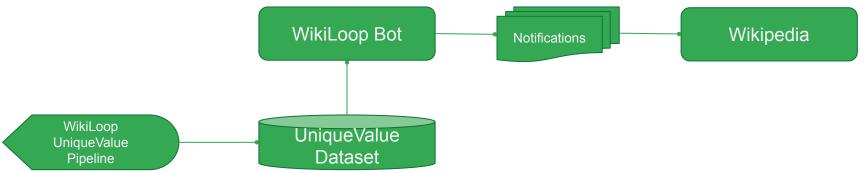
Not a troll, just an unfortunate category extraction.



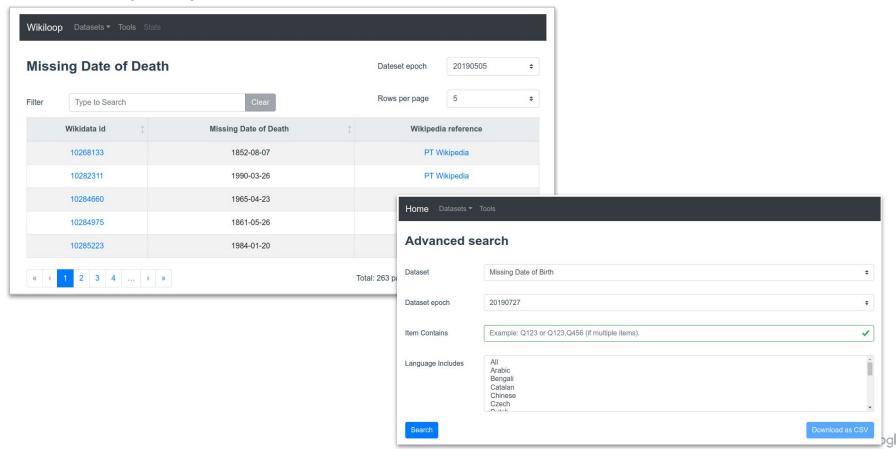
WikiLoop Bot

A notification tool to deliver message Can be used in other ways in the future, but we mostly focusing using them to deliver notice for editor to consider fixing suspected wrong information



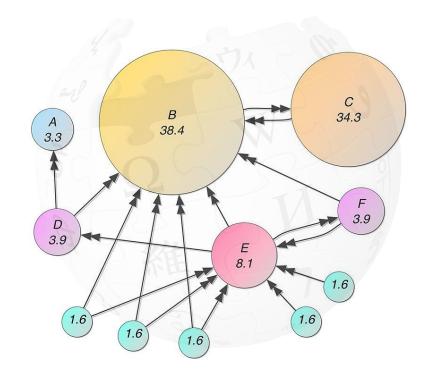


WikiLoop Explorer



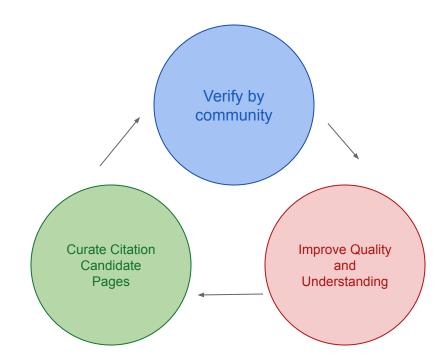
WikiLoop PageRank

- Provide a live version of PageRank scores of pages only computed through Wikipedia Links
- Can be used to assess
 - page quality
 - o page importance
 - page relevance
 - source reliability



WikiLoop CitePool

- Large scale curation of citation candidate link.
- Editor choose good quality and relevant reference link.
- Give us signal of that a page is a good reference or provide certain information.
- Good potential training data for Reconciliations and Fact Extractions



WikiLoop Babel

- Making i18n name translation easier for smaller language-speaking communities
- Creating training data to bootstrap machine translation, available to public researchers
- Bootstrap machine translations closer to launch





WikiLoop Portfolio



WikiLoop **Bot** Deliver automatic communication



WikiLoop Game WikiData Game for WikiLoop



WikiLoop **Explorer** Browse all WikiLoop dataset portfolio



WikiLoop Battlefield v1 Web app for fighting vandals together



WikiLoop CatFacts From editor-marked categories to facts



WikiLoop PageRank Is a source site considered reliable?



WikiLoop PreCheck

Crowdsource checked batch import



WikiLoop CitePool

Citation candidate pool from the web

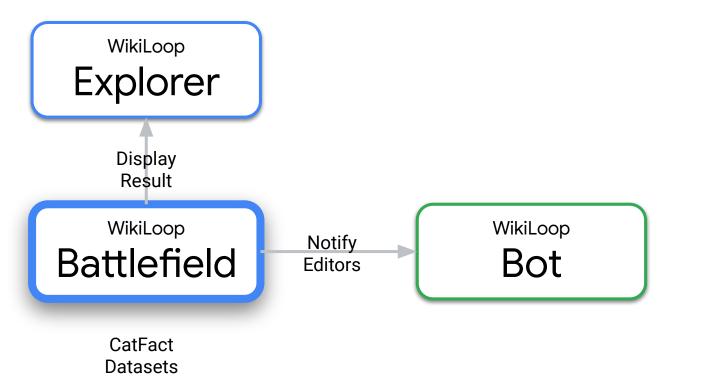


WikiLoop Babel

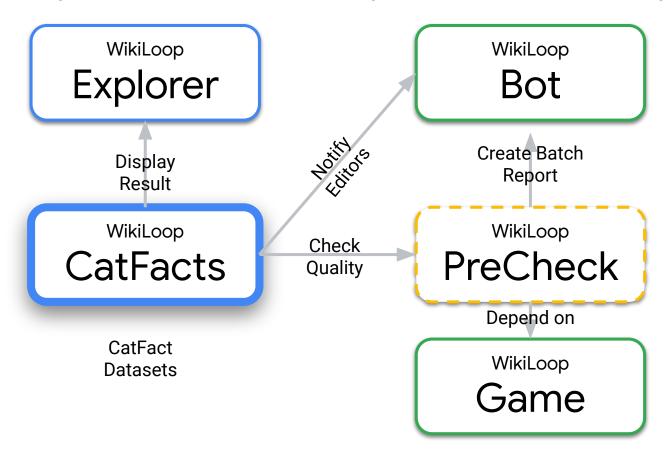
Bootstrap smaller language translations



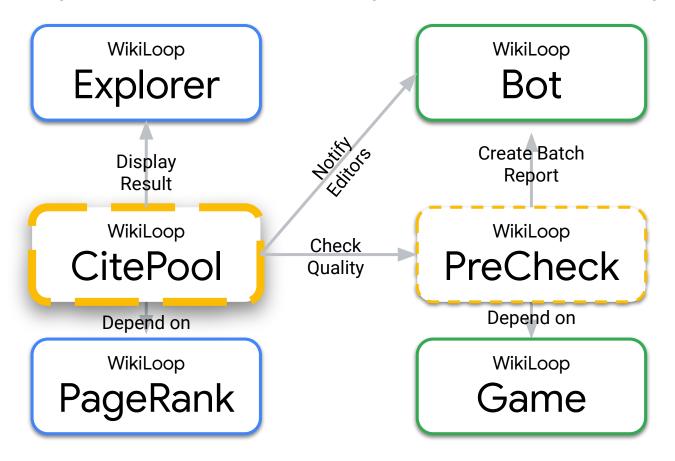
WikiLoop Portfolio Relationship: Battlefield as example



WikiLoop Portfolio Relationship: CatFacts as example



WikiLoop Portfolio Relationship: CitePool as example



Challenges

Roadmap



Ice-Break Phase #0

Small amount of data points *Q2/2018*



Root Phase #2

Prod-ize & identify opportunities *Q4/18-Q1/19*



Mature Phase #4

Long term sustainability *Q4/19-Q1/20*



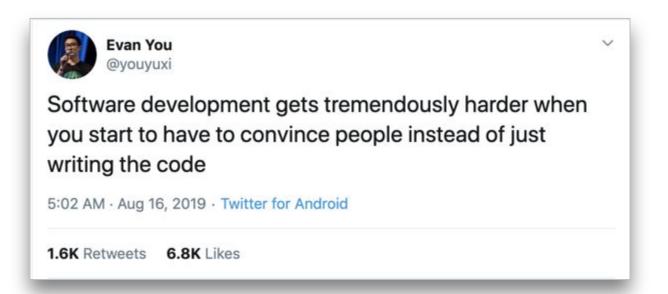
Seed Phase #1

Supply the first batch data *Q3/2018*



Flourish Phase #3

More data and tools *Q2/19-Q3/19*



https://twitter.com/youyuxi/status/1162198039861354496

Evan You, creator of VueJS

Challenges | Chicken-and-Egg of Trust Building





If you (WikiLoop team) want us to approve releasing data and code, how do we know if Wikipedia likes the idea?

If you(Google) don't show me(Wikipedia) what you have and what you can do, how do we know if it's good for us?

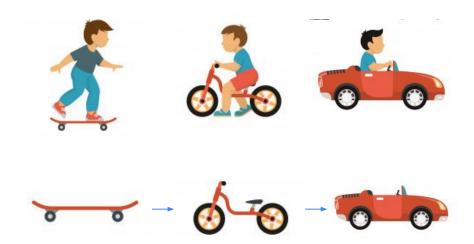




You are a for-profit, how do we know you are not here to influence our Neutrality?

We will make sure we help in a way that Wikipedia wants.





Build small tools and datasets in the beginning and bootstrap the trust

Value Proposition

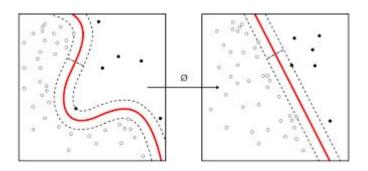
Value Proposition

Project WikiLoop aim at providing data improvements (quality, quantity, timeliness) that are not otherwise possible, merely through community efforts, or merely through machine/eng.

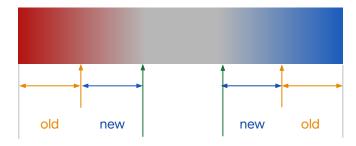
"Data points close to the **decision boundary** will greatly help the models."



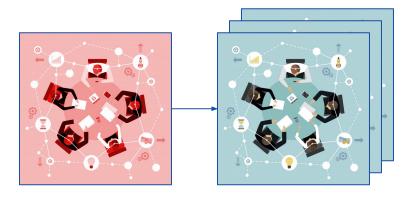
Opportunities Example



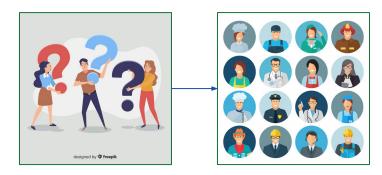
Generating data closer to the decision boundary



Improve range of input data in the greyer area



Crowdsource simple -> comprehensive



Editorial expertise with wide range of knowledge topic expertise



How you can get involved

Ways to get involved

- Users/Editors: Use our tools, play our games, give feedback! (on wikiproject talk page)
- Researchers: look at our datasets and leverage them in your research
- Developers: join the open source code repository, send issues and PRs

[[:m:WikiLoop]]

- https://meta.wikimedia.org/wiki/WikiProject_WikiLoop
- http://github.com/google/wikiloop-battlefield (more to come)

[[:m:WikiLoop]]

Thank you, and stay tuned!