

The sum of all human knowledge in the age of machines

A new research agenda for Wikimedia

Dario Taraborelli • Wikimedia Foundation

ICWSM 2015 Workshop, 26 May 2015



A conversation

Impact of Wikipedia research

contributor motivation

rise and decline of the editor population

gender gap

asymmetries in content and provenance of contributions

socio-technical systems governing quality control.

Human curated knowledge in the age of machines





the long-form encyclopedia

Outline

1. sourcing information
2. consuming information
3. distributing content

A new research agenda

Wikimedia as a platform for researchers

1. Sourcing information

Goats



Life expectancy

Life expectancy for goats is between fifteen and eighteen years.^[29] An instance of a goat reaching the age of 24 has been reported.^[30]

Several factors can reduce this average expectancy; problems during kidding can lower a doe's expected life span to ten or eleven, and stresses of going into rut can lower a buck's expected life span to eight to ten years.^[30]

https://en.wikipedia.org/wiki/Goat#Life_expectancy



average goat lifespan

Web

Shopping

Images

Videos

News

More ▾

Search tools

About 367,000 results (0.39 seconds)

15 – 18 y

Goat, Lifespan



Feedback

Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion

Xin Luna Dong^{*}, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Ni Lao,
Kevin Murphy[†], Thomas Strohmann, Shaohua Sun, Wei Zhang

Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043
{lunadong|gabr|geremy|wilko|nlao|kpmurphy|tstrohmann|sunsh|weizh}@google.com

ABSTRACT

Recent years have witnessed a proliferation of large-scale knowledge bases, including Wikipedia, Freebase, YAGO, Microsoft's Satori, and Google's Knowledge Graph. To increase the scale even further, we need to explore automatic methods for constructing knowledge bases. Previous approaches have primarily focused on text-based extraction, which can be very noisy. Here we introduce Knowledge Vault, a Web-scale probabilistic knowledge base that combines extractions from Web content (obtained via analysis of text, tabular data, page structure, and human annotations) with prior knowledge derived from existing knowledge repositories. We employ supervised machine learning methods for fusing these distinct information sources. The Knowledge Vault is substantially bigger than any previously published structured knowledge repository, and features a probabilistic inference system that computes calibrated probabilities of fact correctness. We report the results of multiple studies that explore the relative utility of the different information sources and extraction methods.

In recent years, several large-scale knowledge bases (KBs) have been constructed, including academic projects such as YAGO [39], NELL [8], DBpedia [3], and Elementary/ DeepDive [32], as well as commercial projects, such as those by Microsoft¹, Google², Facebook³, Walmart [9], and others. (See Section 7 for a detailed discussion of related work.) These knowledge repositories store millions of facts about the world, such as information about people, places and things (generically referred to as entities).

Despite their seemingly large size, these repositories are still far from complete. For example, consider Freebase, the largest open-source knowledge base [4]. 71% of people in Freebase have no known place of birth, and 75% have no known nationality⁴. Furthermore, coverage for less common relations/predicates can be even lower.

Previous approaches for building knowledge bases primarily relied on direct contributions from human volunteers as well as integration of existing repositories of structured knowledge (e.g., Wikipedia infoboxes). However, these methods are more likely to yield head content, namely, frequently mentioned properties of frequently mentioned entities. Suh

Douglas Adams (Q42)

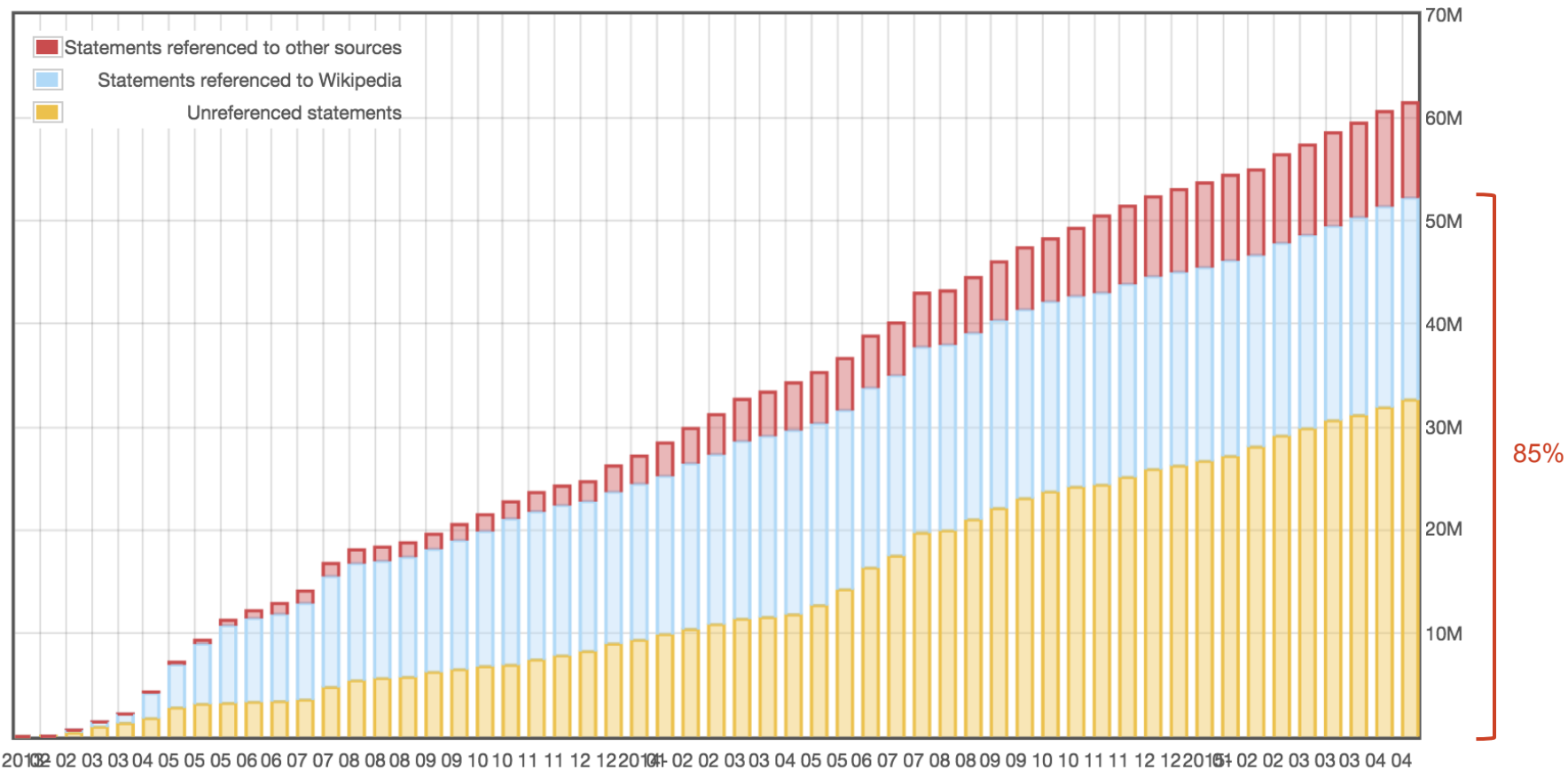
[\[edit\]](#)

English writer and humorist

Douglas Noël Adams | Douglas Noel Adams

[▶ In more languages](#)

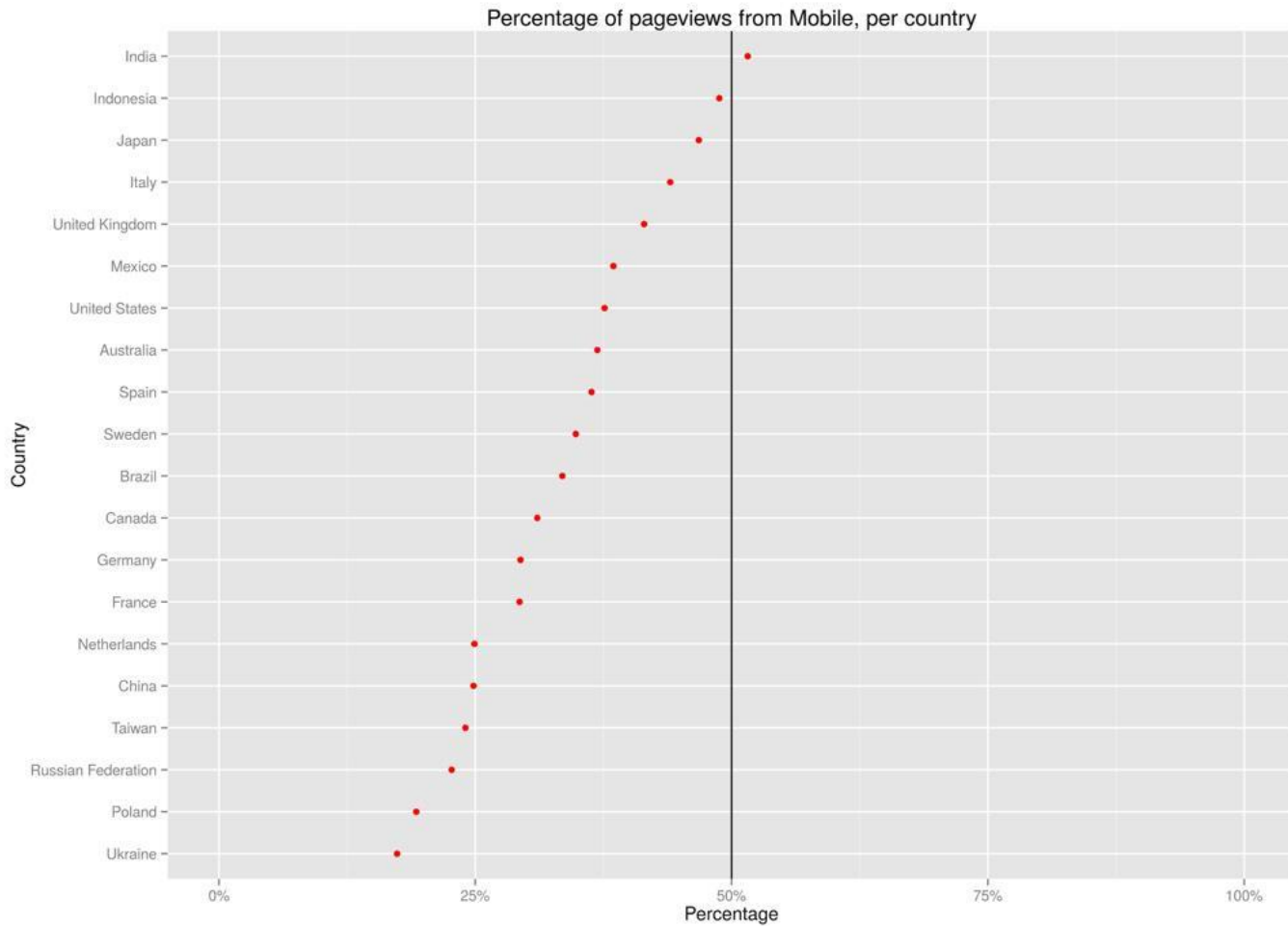
birth name	 Douglas Noël Adams (English) [edit]
	 ▼ 1 reference
	 [edit]
title	Obituary: Douglas Adams (English)
publisher	The Guardian
publication	15 May 2001
original language of work	English
reference URL	http://www.theguardian.com/news/2001/may/15/guardianobituaries.books
retrieved	7 December 2013
author	Nicholas Wroe
	[add reference]
	[add]



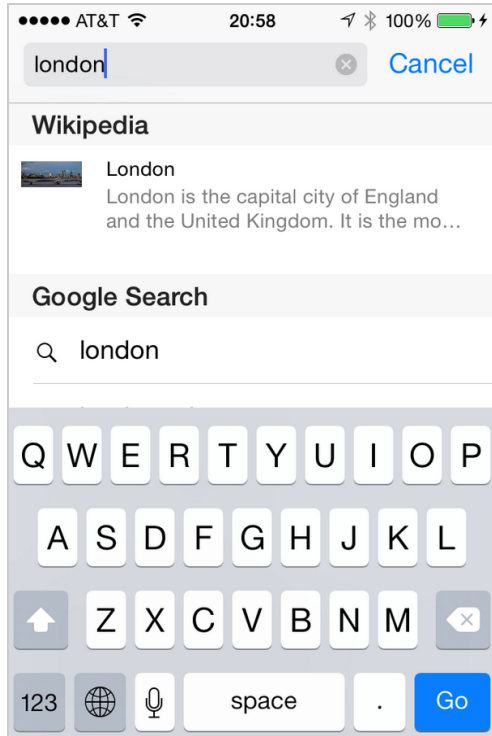
1. Sourcing information

- What role does information sourced by humans play when answers to most questions are readily available from search engines?
- Should Wikipedia start integrating algorithmically extracted sources in its contents?
- Should Wikipedia further invest in supporting human generated citations?

2. Consuming information



O. Keyes (2015) The Mobile Singularity is already here. Wikipedia and the Mobile Web, Part 1.



Bite size consumption



Andrea Mantegna

Andrea Mantegna (Italian: [anˈdrea manˈtɛŋɲa]; c. 1431 – September 13, 1506) was an Italian painter, a student of Roman archeology, and son-in-...

[Read more >](#)



Leonardo da Vinci

Leonardo di ser Piero da Vinci or Leonardo da Vinci (Italian: [leoˈnardo dá(v)vinˈtʃi]; 15 April 1452 – 2 May 1519) was an Italian polymath,...

[Read more >](#)



Bramantino

Bartolomeo Suardi, best known as Bramantino (c.1456 – c.1530), was an Italian painter and architect, mainly active in his native Milan. ...

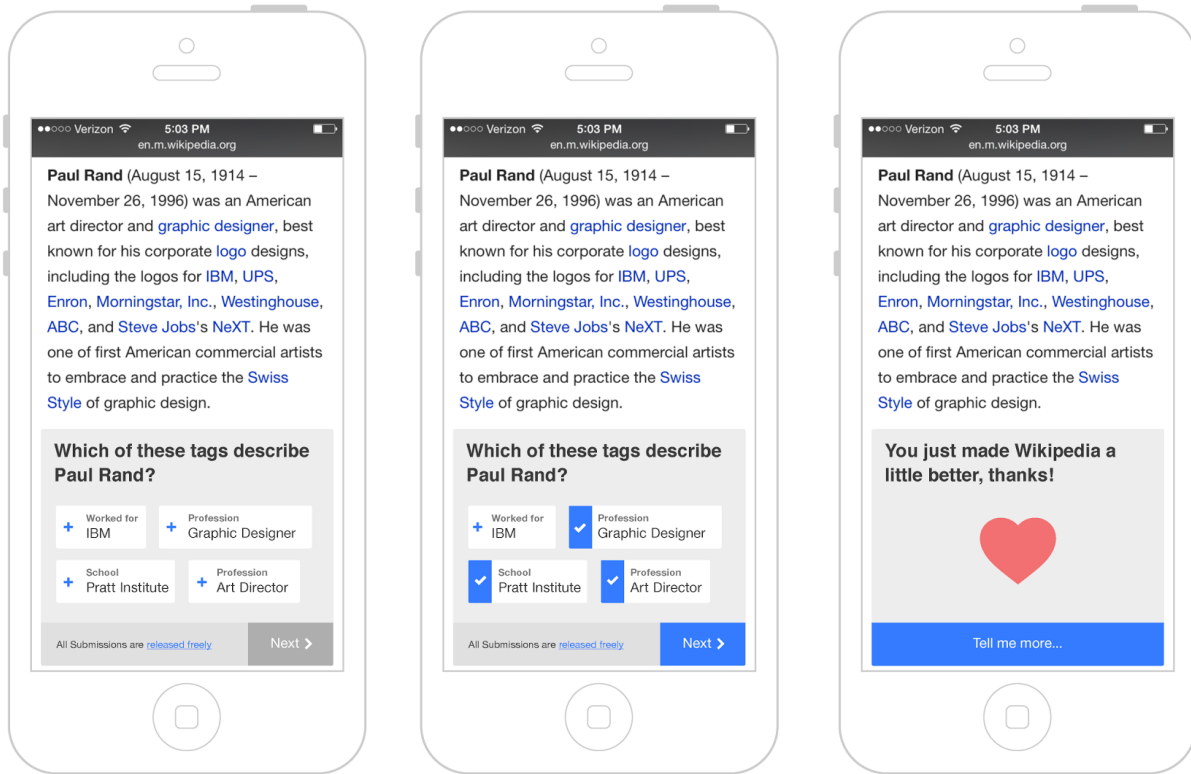
[Read more >](#)



Lorenzo Lotto

Lorenzo Lotto (c. 1480 – 1556/57) was a Northern Italian painter, draughtsman and illustrator, traditionally placed in the Venetian school. He...

[Read more >](#)



Structured contributions

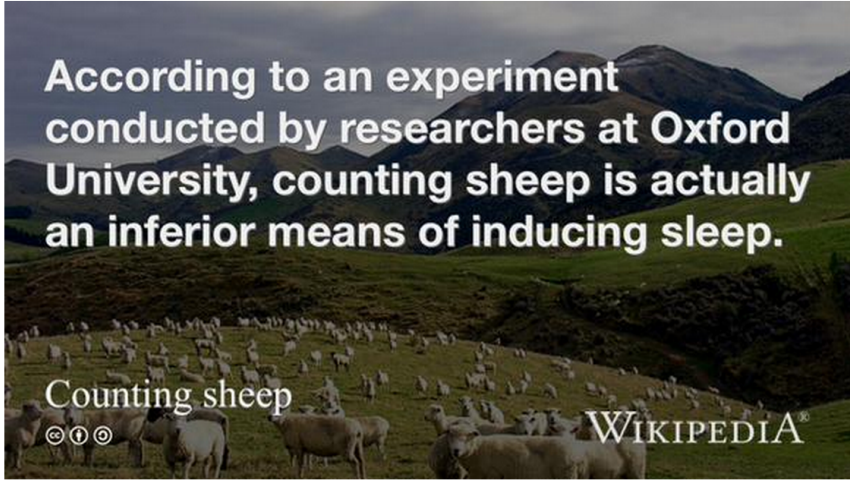


Jared Zimmerman
@JaredZimmerman



Following

Goodnight @Wikipedia
en.wikipedia.org/wiki/Counting_...



RETWEETS 3 FAVORITES 3



12:22 AM - 19 Feb 2015



Jared Zimmerman
@JaredZimmerman



Following

Not just for fashion crimes "glitter" on
@Wikipedia en.wikipedia.org/wiki/Glitter?s...



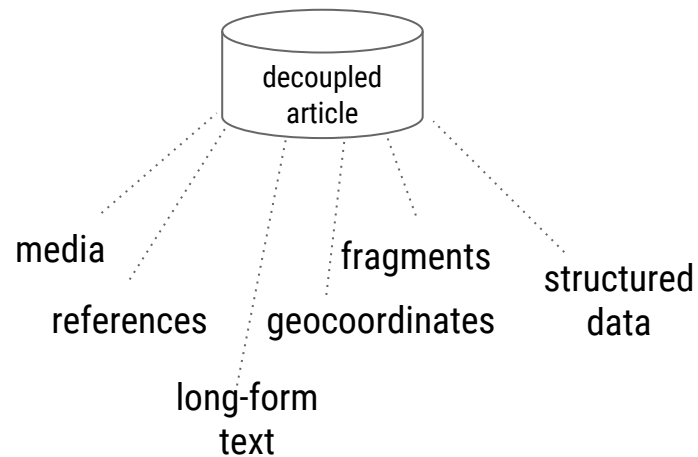
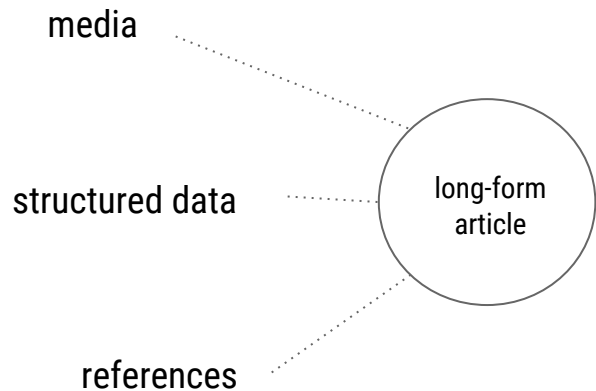
RETWEETS 2 FAVORITES 3



6:20 PM - 17 Feb 2015

Manipulating fragments

Decoupling the article



2. Consuming information

- How to transform Wikimedia contents to make them suitable to bite size consumption?
- How to accelerate extraction and coverage of structured data in Wikidata?
- How to design effective lightweight contribution funnels around structured data and fragments?
- How to support programmatic manipulation of content fragments?

3. Distributing content

Cooking 

Info [Related Posts](#) [Wikipedia](#)

Our goal is to make this Community Page the best collection of shared knowledge on this topic. If you have a passion for **Cooking**, sign up and we'll let you know when we're ready for your help.

Description

From Wikipedia, the free encyclopedia


Cooking is the process of preparing food by applying heat. Cooks select and combine ingredients using a wide range of tools and methods. In the process, the flavor, texture, appearance, and chemical properties of the ingredients can change. Cooking techniques and ingredients vary widely across the world, reflecting unique environmental, economic, and cultural traditions. Cooks themselves also vary widely in skill and training.

Preparing food with heat or fire is an activity unique to human beings, and some scientists believe the advent of cooking played an important role in human evolution. Most anthropologists believe that cooking fires first developed around 250,000 years ago. The development of agriculture, commerce and transportation between civilizations in different regions offered cooks many new ingredients. New inventions and technologies, such as pottery for holding and boiling water, expanded cooking techniques. Some modern cooks apply advanced scientific techniques to food preparation.


5 Friends Like This




 Kasey Galang  David Nguyen  Caitlin O'Farrell

The paradox of reuse

 **WolframAlpha** computational... knowledge engine

Enter what you want to calculate or know about:

    [Examples](#) [Random](#)

Assuming "London" is a city | Use as an administrative division or a surname or a given name instead

Assuming London (United Kingdom) | Use [London \(Canada\)](#) or [more](#) instead


Input interpretation:

London, Greater London, United Kingdom

Populations: [Show history](#)

city population	8.174 million people (country rank: 1 st) (2011 estimate)
metro area population	12.58 million people (London metro area) (2007 estimate)

Timeline:

European Renaissance 

Wikipedia summary:

The Renaissance was a cultural movement that spanned the period roughly from the 14th to the 17th century, beginning in Italy in the Late Middle Ages and later spreading to the rest of Europe. Though availability of paper and the invention of metal movable type sped the dissemination of ideas from the later 15th century, the changes of the Renaissance were not uniformly experienced across Europe.

[Full entry >](#)

Women in Science



The examples and perspective in this section **deal primarily with USA and do not represent a worldwide view of the subject**. Please [improve this article](#) and discuss the issue on the [talk page](#). (April 2013)

The Seattle Times
Winner of Nine Pulitzer Prizes

Books

[Home](#) | [News](#) | [Business & Tech](#) | [Sports](#) | [Entertainment](#) | [Food](#) | [Living](#) | [Homes](#) | [Travel](#) | [Opinion](#)

[Jobs](#) | [Autos](#) | [Shopping](#) | [Weekly Ads](#)

Originally published July 20, 2014 at 6:06 AM | Page modified July 21, 2014 at 11:59 AM

Beautiful minds: books that celebrate women in science

Seattle Times book editor Mary Ann Gwinn recommends the novels "The Signature of All Things" and "Remarkable Creatures," and the biographies "Obsessive Genius: The Inner World of Marie Curie" and "Rosalind Franklin: the Dark Lady of DNA."

Share:



[Recommend](#) 99

[1 Comments](#)

[E-mail article](#)

[Print](#)

Wikipedia needs your help

The English Wikipedia article [Women in Science](#) needs contributors from a more global perspective. [Help expand it!](#)



TRENDING WITH READERS

On [seattletimes.com](#)

[MORE TRENDING](#)

MOST READ

Routing attention



Baltimore Edits

@baltimoreedits

TWEETS

3,018

FOLLOWERS

5,328

FAVORITES

3



+ Follow



Baltimore Edits @baltimoreedits · 1h

"Death of Freddie Gray" #Wikipedia edit by Grand'mere Eugene #BaltimoreUprising
en.wikipedia.org/w/index.php?di...



Baltimore Edits @baltimoreedits · 1h

"Police brutality" #Wikipedia edit by Anonymous #BaltimoreUprising
en.wikipedia.org/w/index.php?di...



Baltimore Edits @baltimoreedits · 1h

"Death of Freddie Gray" #Wikipedia edit by Anonymous #BaltimoreUprising
en.wikipedia.org/w/index.php?di...



Routing attention



Wikipedia Stub Bot

@wpstubs FOLLOWS YOU

TWEETS

130K

FOLLOWING

5

FOLLOWERS

356



Wikipedia Stub Bot @wpstubs · 35m

Someone created a Wikipedia article about "Atlantsskip". Help expand it! [#Iceland](#)
[en.wikipedia.org/wiki/Atlantssk...](https://en.wikipedia.org/wiki/Atlantsskip)



Wikipedia Stub Bot @wpstubs · 40m

Someone created a Wikipedia article about "Kjell Nilsson (cyclist)". Help expand it!
[#Biography](#) [#Olympics](#) [#Sweden](#)
[en.wikipedia.org/wiki/Kjell_Nil...](https://en.wikipedia.org/wiki/Kjell_Nilsson)



Wikipedia Stub Bot @wpstubs · 47m

Someone created a Wikipedia article about "Frank Grant (boxer)". Help expand it!
[#Biography](#) [#Boxing](#) [#England](#)
[en.wikipedia.org/wiki/Frank_Gra...](https://en.wikipedia.org/wiki/Frank_Grant)



Routing attention

3. Distributing content

- How can we design content distribution systems that *do not intermediate* Wikipedia?
- How do we leverage content syndication to route (expert) attention to the source?

A new research agenda

Designing and evaluating systems to:

1. preserve and increase transparent sourcing of information
2. break down long-form articles into their constituents
3. optimize content fruition, as a function of access
4. enable lightweight contribution/manipulation of structured data / fragments
5. leverage content distributed / syndicated by 3rd parties
6. prioritize work and route contributors to the site, as a function of demand

Wikimedia Research as a platform

Wikimedia Research as a platform

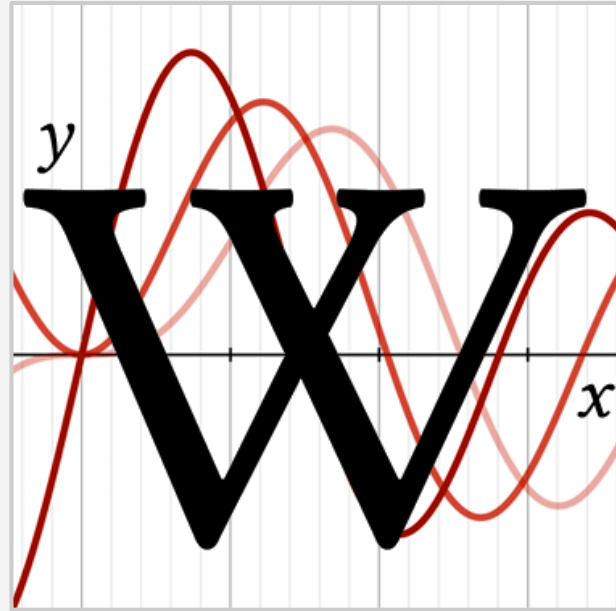
Wikimedia Research & Data team

Edit quality classifiers

Automated link recommendations

Article translation recommendations

Fundraiser optimization



Scaling Wikimedia research

1:100,000,000

Approximate ratio of full-time researchers at WMF by monthly unique visitors

Formal collaborations

Fellows and NDA'ed collaborators

Stanford University

GroupLens, University of Minnesota

Oxford Internet Institute (2015)

Los Alamos National Laboratory (2015)

Open access policy

The Wikimedia Foundation's [mission](#) is to disseminate open knowledge effectively and globally. In keeping with this mission, the Wikimedia Foundation supports research in areas that benefit the Wikimedia community. We aim to make any work produced with our support [openly available](#) to the public and reusable on Wikimedia projects.

1. Expectations

Researchers will need to provide unrestricted access to and reuse of all their research output if their research receives support from the Wikimedia Foundation in the form of:

- funds;
- letter of endorsement;
- equipment, hosting, or office space;
- access to non-public data or special API privileges; or
- other support under an agreement between researchers and the Wikimedia Foundation.

Conclusions

Questions?

dario@wikimedia.org

[@readermeter](#)

[@wikiresearch](#)



Image credits

Election Night Crowd, Wellington, 1931

https://www.flickr.com/photos/nationallibrarynz_commons/3326203787

CC0

King Billy of Dalkey Island

<https://www.flickr.com/photos/paulodonnell/5937678226>

CC BY

Secretary at typewriter, 1912

https://www.flickr.com/photos/muohio_digital_collections/3192197470

CC0

"Getting em up" at U.S.Naval Training Camp, Seattle, Washington. ca. 1917 - ca. 1918

<https://www.flickr.com/photos/usnationalarchives/5505933145>

CC0