

Proposal *Shared Citations*

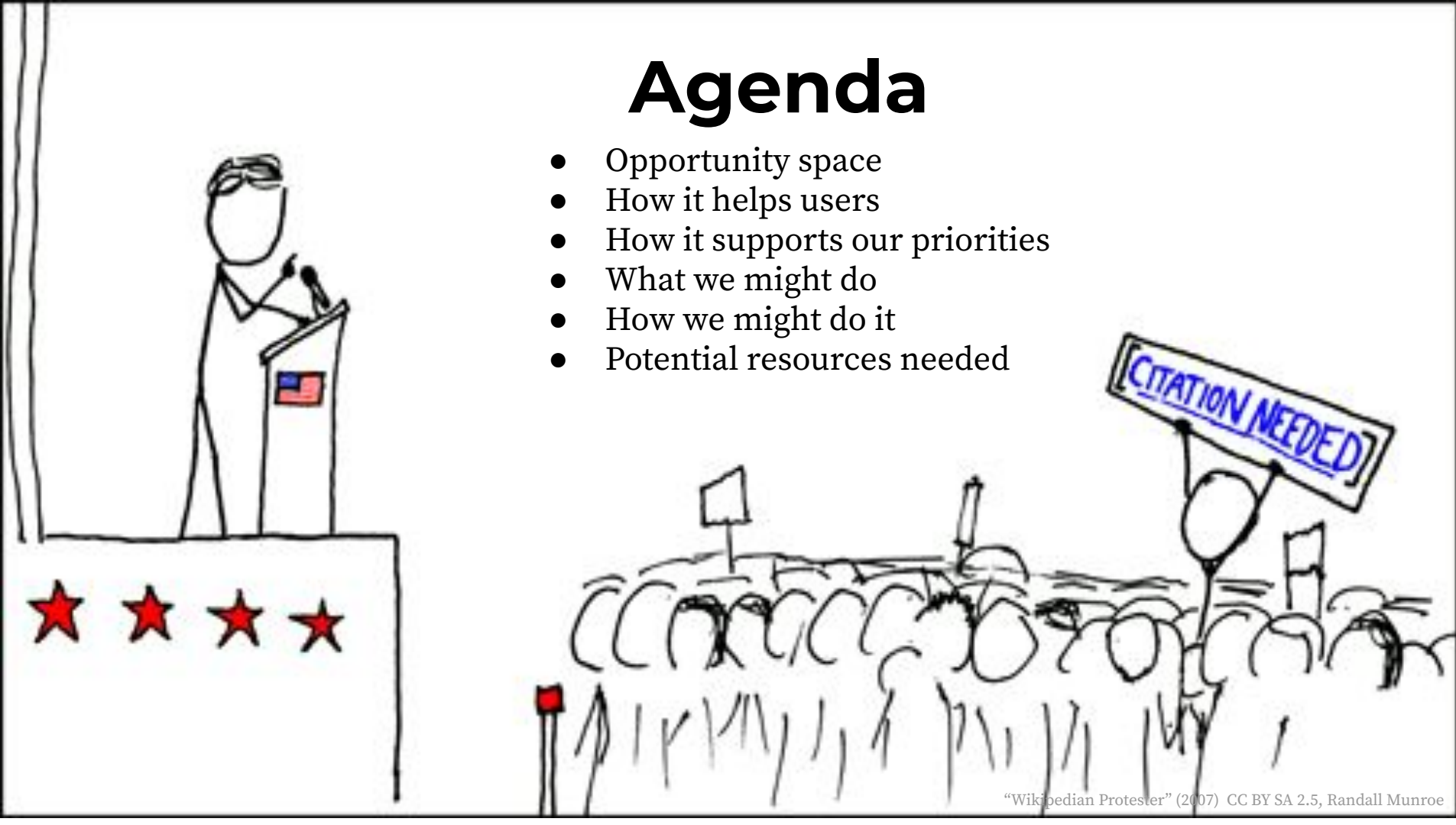
Prepared by Liam Wyatt / Wittylama
2020



WIKIMEDIA
FOUNDATION

Agenda

- Opportunity space
- How it helps users
- How it supports our priorities
- What we might do
- How we might do it
- Potential resources needed



Opportunity Space

- Citations
- The problem
- Current state

The stakes

Citations are the core of **anti-disinformation** and **knowledge integrity** in our movement. Our **Verifiability** policy has become the backbone of the reliable web.

“Citations are a simple, critical interconnection mechanism for all modern knowledge in the digital, Internet-connected world ... arguably the most important ingredient of open knowledge, sources and references have received little technical attention in the Wikimedia movement up until now.”

–2016 [WikiCite report](#)



The Objective

“Make citations easier for the editor,
more useful for the reader,
and more efficient for our architecture”



WIKIMEDIA
FOUNDATION

Photo by Diliff, CC BY SA 4.0

The Problem

Wikimedia's citations are one of our greatest assets, but by storing them as raw “inline” text in each content page, they are also one of our biggest burdens.

Our citations are high in maintenance, technical complexity, and duplication of effort.

This burden is shouldered by repetitive, manual, volunteer effort which is disproportionately felt by smaller communities.



Participants at WikiCite 2018 debating this problem.
We've known this is an issue for a *long* time.



References in MediaWiki (WP)

Closures and cancellations [[edit](#) | [edit source](#)]

Through the first quarter of 2020, arts and culture sector organisations around the world progressively restricted their public activities and then closed completely due to the pandemic. Starting with China, East Asia, and then worldwide, by late March most cultural heritage organisations had closed, and arts events were postponed or cancelled, either voluntarily or by government mandate. This included galleries, libraries,^[1] archives,^[2] and museums^{[3][4][5]} (collectively known as **GLAMs**), as well as film^[6] and television productions,^[7] theatre^[8] and orchestra performances,^[9] concert tours,^[10] zoos,^[11] and music^[12] and arts festivals.^{[8][13]}

==Closures and cancellations==

Through the first quarter of 2020, arts and culture sector organisations around the world progressively restricted their public activities and then closed completely due to the pandemic. Starting with China, East Asia, and then worldwide, by late March most cultural heritage organisations had closed, and arts events were postponed or cancelled, either voluntarily or by government mandate. This included galleries, libraries, <ref name="IFLA">{{cite web|url=https://www.ifla.org/covid-19-and-libraries|title=COVID-19 and the Global Library Field|publisher=[International Federation of Library Associations and Institutions|IFLA]]|access-date=3 April 2020|archive-url=https://web.archive.org/web/20200404123805/https://www.ifla.org/covid-19-and-libraries|archive-date=4 April 2020|url-status=live}}</ref> archives, <ref name="Accessible">{{Cite web|url=https://www.ica.org/en/what-archive/archives-are-accessible-search-the-map|title=Archives are Accessible|last=[first=]date=|website=www.ica.org|publisher=[International Council on Archives]]|url-status=live|archive-url=https://web.archive.org/web/20200414084458/https://www.ica.org/en/what-archive/archives-are-accessible-search-the-map|archive-date=14 April 2020|access-date=19 April 2020}}</ref> and museums<ref>{{cite web|url=http://www.theartnewspaper.com/news/here-are-the-museums-that-have-closed-due-to-coronavirus|title=Here are the museums that have closed (so far) due to coronavirus|website=www.theartnewspaper.com|access-date=26 March 2020|archive-url=https://web.archive.org/web/20200329062120/https://www.theartnewspaper.com/news/here-are-the-museums-that-have-closed-due-to-coronavirus|archive-date=29 March 2020|url-status=live}}</ref><ref>{{cite web|url=https://www.artnews.com/art-news/news/coronavirus-museum-closures-worldwide-1202680933/|title=See a List of Coronavirus-Related Closures at Museums Around the World|last1=Solomon|first1=Tessa|last2=Sylyin|first2=Claire|date=1 March 2020|website=ARTnews.com|language=en-US|access-date=26 March 2020|archive-url=https://web.archive.org/web/20200326171347/https://www.artnews.com/art-news/news/coronavirus-museum-closures-worldwide-1202680933/|archive-date=26 March 2020|url-status=live}}</ref><ref>{{Cite web|url=https://docs.google.com/document/d/1LW2hsErFzqdpZLS0Tf0Q8SsmQnRPhbYemOb1IAgw10/edit?usp=embed_facebook|title=Museums closures|website=Google Docs|language=en|access-date=27 April 2020}}</ref> (collectively known as [[GLAM (industry sector)|GLAM]]s), as well as film<ref>{{Cite news|last=Catherine|url=https://www.theguardian.com/film/2020/mar/20/over-one-hour-everything-cancelled-coronavirus-impact-film|title='Over one hour everything was cancelled' – how coronavirus devastated the film industry|date=2 March 2020|work=The Guardian|access-date=19 April 2020|url-status=live|language=en-GB|archive-url=https://web.archive.org/web/20200417102132/https://www.theguardian.com/film/2020/mar/20/over-one-hour-everything-cancelled-coronavirus-impact-film|archive-date=17 April 2020}}</ref> and television productions,<ref>{{Cite web|url=https://deadline.com/2020/03/coronavirus-tv-shows-production-delayed-1202881997/|title=Coronavirus: TV Shows That Have Halted Or Delayed Production Amid Outbreak|last=Pedersen|first=Erik|date=1 March 2020|website=Deadline|language=en|url-status=live|archive-url=https://web.archive.org/web/20200402092712/https://deadline.com/2020/03/coronavirus-tv-shows-production-delayed-1202881997/|archive-date=2 April 2020|access-date=19 April 2020}}</ref> theatre<ref name="List of major cancellations">{{Cite news|url=https://www.theguardian.com/culture/2020/mar/13/coronavirus-culture-arts-films-gigs-festivals-cancellations|title=Coronavirus and culture – a list of major cancellations|date=2 March 2020|work=The Guardian|access-date=26 March 2020|language=en-GB|issn=0261-3077|archive-url=https://web.archive.org/web/20200316052542/https://www.theguardian.com/culture/2020/mar/13/coronavirus-culture-arts-films-gigs-festivals-cancellations|archive-date=16 March 2020|url-status=live}}</ref> and [[orchestras|orchestra]] performances,<ref name="N Kenyon">{{Cite news|last=Kenyon|first=Nicholas|url=https://www.theguardian.com/music/2020/mar/21/home-listening-classical-music-online-livestream-at-home-coronavirus|title=Classical music: let the Berlin Phil come to you|date=2 March 2020|work=The Guardian|access-date=7 April 2020|language=en-GB|issn=0261-3077|archive-url=https://web.archive.org/web/20200407160747/https://www.theguardian.com/music/2020/mar/21/home-listening-classical-music-online-livestream-at-home-coronavirus|archive-date=7 April 2020|url-status=live}}</ref> [[concert tours]],<ref>{{Cite web|url=https://ew.com/music/bts-madonna-pearl-jam-more-artists-cancelling-shows-coronavirus/|title=BTS, Madonna, Khalid, Billie Eilish, and more artists canceling shows over coronavirus|website=EW.com|language=EN|access-date=19 April 2020|archive-url=https://web.archive.org/web/20200416104508/https://ew.com/music/bts-madonna-pearl-jam-more-artists-cancelling-shows-coronavirus/|archive-date=16 April 2020|url-status=live}}</ref> [[zoos]],<ref>{{Cite web|url=https://zahp.aza.org/covid-19-information-for-zoos-and-aquariums/|title=COVID-19: Information for Zoos and Aquariums|last=[first=]date=|website=zahp.aza.org|publisher=Zoo and Aquarium All Hazards Preparedness, Response, and Recovery (ZAHPR) Fusion Center|language=en-US|url-status=live|archive-url=https://web.archive.org/web/20200410124727/https://pitchfork.com/news/coronavirus-updated-list-of-tours-and-festivals-canceled-or-postponed-due-to-covid-19/|archive-date=10 April 2020|url-status=live}}</ref> and arts festivals.<ref name="List of major cancellations">{{cite web|url=http://www.theartnewspaper.com/news/here-are-the-museums-that-have-closed-due-to-coronavirus|title=Here are the museums that have closed (so far) due to coronavirus|website=www.theartnewspaper.com|access-date=27 March 2020|archive-url=https://web.archive.org/web/20200329062120/https://www.theartnewspaper.com/news/here-are-the-museums-that-have-closed-due-to-coronavirus|archive-date=29 March 2020|url-status=live}}</ref>



Source: [English Wikipedia: Impact of the COVID-19 pandemic on the arts and cultural heritage](#)

References in Wikipedia

```
<ref name = "koppen">
{{Cite journal
| title = Updated world map of the Köppen-Geiger climate classification
| url = http://www.hydrol-earth-syst-sci.net/11/1633/2007/hess-11-1633-2007.html
| year = 2007
| journal = Hydrology and Earth System Sciences
| volume = 11
| pages= 1633-1644
| <!--sid = 1633-1644→
| date =<!-- 01/30/2016→
| access-date=30 January 2016
| doi = 10.5194/hess-11-1633-2007
| first1 =M C| first2 = B L| first3 = T A| last1 = Peel| last2 = Finlayson| last3 =McMahon
| doi-access = free}}
</ref>
```

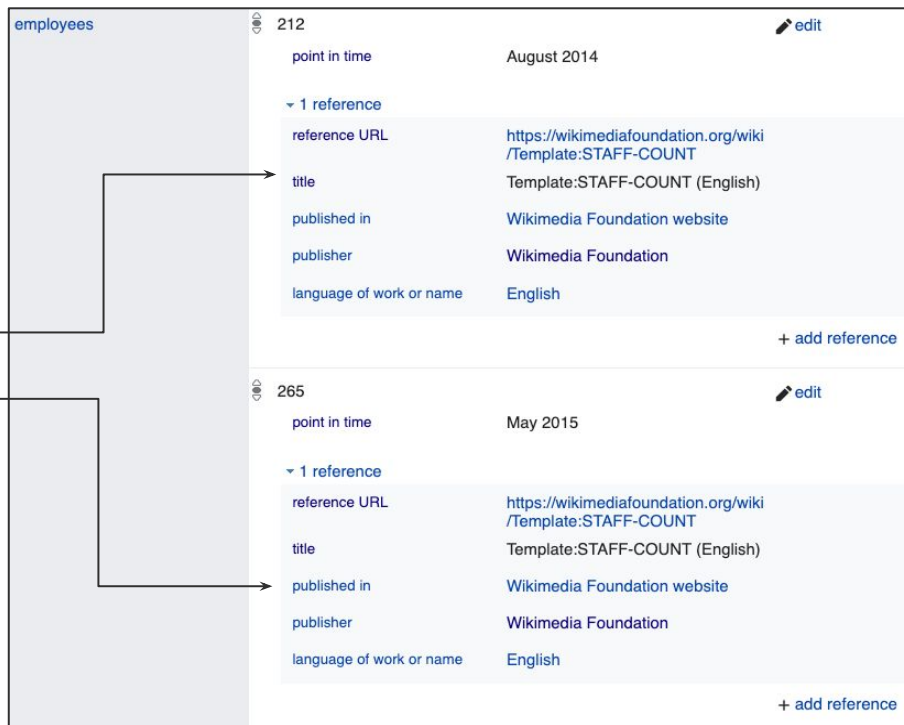


More
details
available

References in Wikibase (WD)

Reference fields are stored and edited independently.

This is even when the same information is identical, on the same property, on the same item.



employees	<div>212</div> <div>point in time</div> <div>August 2014</div> <div>edit</div> <div>1 reference</div> <div>reference URL</div> <div>https://wikimediafoundation.org/wiki/Template:STAFF-COUNT</div> <div>title</div> <div>Template:STAFF-COUNT (English)</div> <div>published in</div> <div>Wikimedia Foundation website</div> <div>publisher</div> <div>Wikimedia Foundation</div> <div>language of work or name</div> <div>English</div> <div>+ add reference</div>
	<div>265</div> <div>point in time</div> <div>May 2015</div> <div>edit</div> <div>1 reference</div> <div>reference URL</div> <div>https://wikimediafoundation.org/wiki/Template:STAFF-COUNT</div> <div>title</div> <div>Template:STAFF-COUNT (English)</div> <div>published in</div> <div>Wikimedia Foundation website</div> <div>publisher</div> <div>Wikimedia Foundation</div> <div>language of work or name</div> <div>English</div> <div>+ add reference</div>



More
details
available

Re-used references

Specific Source templates

Local templates for individual, frequently reused works

- English WP: [452 templates for “rail transport books”](#) e.g.

`{{Anchen-Iron Roads | page=136}}`

20. [▲] Anchen, Nick (2017). *Iron roads in the outback: the legendary Commonwealth Railways*. Ferntree Gully, Victoria: Sierra Publishing. p. 136. ISBN 9780992538828.

- French Wikipedia: [References Namespace](#). e.g.

- [↑] Pindare, *Odes* [détail des éditions] [lire en ligne [↗] [archive]], *Néméennes* VII, 1-5 et X, 18.
- [↑] Homère, *Iliade* [détail des éditions] [lire en ligne [↗] [archive]], IV, 2-3.

Links to manually compiled listing of bibliographic details of all major editions of [Pindare's Odes](#); [Homer's Iliad](#)

Cite Q

Lua module to display a Wikidata items as a reference

- English WP: [Used in 43,000 pages](#) (mostly inside infobox templates)

```
{{Cite Q|Q15625490|page=42|access-date=18 May 2017|quote=lorem ipsum}}
```

Jeffrey T. Williams; Kent E. Carpenter; James L. van Tassell; Paul Hoetjes; Wes Toller; Peter Etnoyer; Michael Smith (21 May 2010), "Biodiversity Assessment of the Fishes of Saba Bank Atoll, Netherlands Antilles" [↗], *PLOS ONE*, **5** (5): 42, doi:10.1371/JOURNAL.PONE.0010676 [↗], PMC 2873961 [↗], PMID 20505760 [↗], retrieved 18 May 2017, "lorem ipsum", Wikidata Q15625490

- Proof that *shared citations* is technically possible, and the challenges they still face

Case studies on English Wikipedia:

- [“South Pole Telescope”](#)
- [“Suffix Automaton”](#)



Citation duplication

“There are 4.5 million unique sources in the datasets. While on average, every source is cited 3.5 times, the vast majority of sources in this dataset are used less than 500 times across wikis. Nine “super publications” are used more than 10,000 times.”



– What are the ten most cited sources on Wikipedia? Let’s ask the data.
[WMF Blog](#) (2018)



<https://www.wired.com/story/wikipedia-most-cited-authors-no-idea/>
<https://www.nature.com/articles/d41586-018-05161-6>

Citation duplication

According to that 2018 research, the most reused citation is used 2,830,341 times:

^ Peel, M C; Finlayson, B L; McMahon, T A (2007). "Updated world map of the Köppen-Geiger climate classification" . *Hydrology and Earth System Sciences*. **11**: 1633–1644. doi:10.5194/hess-11-1633-2007 . Retrieved 30 January 2016.

2.8m citations @ 500 characters per citation = 1,4 billion characters ∴
= 1.4GB for a single citation.

This ~500 characters is hand curated by volunteers, often with semi-automated tools.



More
details
available

How it helps users (user stories)

1. Knowledge creators
2. Knowledge consumers

Knowledge creators

(Editors / Patrollers / Researchers)

want to easily access & curate citations,
to improve content quality & integrity



WIKIMEDIA
FOUNDATION

Knowledge creators

As a **Wikimedian editor**, I would like to

- Quickly and easily add references to my article by re-using references from another language Wikipedia
 - Benefit from the formatting done by others, so I can spend my time on research and writing, not on templates
 - Train new-users to add footnotes in one session, in a way they can continue without intense support
 - Create ‘redlists’ for Wikidata of authors and publications which are frequently cited but don’t have a Wikidata item
 - Have my work on disambiguating authors on Wikidata cascade through to their references on other Wikiprojects
-

As a **content patroller**, I would like to

- Tag all citations to an instance of *dis/misinformation* or a retracted publication across all projects
 - See, when someone is adding the same link across many different projects in quick succession, and revert
 - Identify and track any citations to predatory journals
-

As a **professional writer**, I would like to

- Have academic ‘impact factor’ reports include Wikipedia citations to my scholarly work
- Check when Wikidata is referencing my work to ensure its findings are accurately represented
- Be notified when a Wikipedia article cites my journalism and be able to share it on social media



Knowledge consumers

(Readers / Writers / Reusers)

want to easily understand and access citations, and to have confidence in them.

Knowledge consumers

As a reader, I would like to

- Understand if I can trust the citation I am reading
 - Be suggested other topics in Wikipedia which also reference this same author/book
 - See dictionary definitions which prioritise usage examples published in my country's vernacular
 - Generate a list of *primary*, or *newspaper*, or *local* sources used on this topic, for my high-school homework
-

As a researcher, I would like to

- Be able to extract parts of the citation corpus and analyse it, without massive pre-processing
 - Know how many references are behind paywalls, how much they cost, and if there are alternatives
 - See how many references are about the language/location, but published in another language/location
-

As a 'big tech' company, I would like to

- Be able to answer the question “says who?” when a customer asks me to verify a fact just given them
 - Train my algorithm to show more reliable sources for languages/topics where I have limited other data
-

As a library, I would like to

- Track that Wikimedia external links to our collection from Wikimedia are well maintained
- Notice which books on our key subjects are cited and ensure our library has holdings



How it supports our priorities

1. Strategic directions
2. OKRs
3. WMF teams

Knowledge equity

*No citations about us,
without us.*

“This will help us achieve epistemological decolonisation”

- João Peschanski (Wiki Movimento Brasil)

“If libraries could look at global Wikimedia citations, it would help break a self-reinforcing cycle of certain sources’ popularity in library holdings”

- Phoebe Ayers (MIT Libraries)

“Standardized templates and references benefit products designed for emerging markets, which need high interoperability of content form-factor, language, and projects.”

- Runa Bhattacharjee (WMF)

“Having systematic insight into where our knowledge comes from will help us to diagnose what kinds of sources, languages, voices are missing.”

- Ben Vershbow (WMF)

Citations as a service

“Commons makes it easy to use the same image on different wikis. But to copy a citation is very hard. Everything is *manual and slow*.” - Amir Aharoni (WMF)

“Imagine being able to *recommend* useful footnotes to editors, readers.” - Sam Walton (WMF)

“This adds *integrity* to our citations - “a fortified citation layer.” - Chris Albon (WMF)

“Abstract Wikipedia articles will be far more useful if references are formatted from structured data instead of plaintext.” - Denny Vrandečić (WMF)

Why us? Why now?

OKRs

- [Platform Evolution](#): Software platforms with structured data, reuse of code & content
- [Thriving Movement](#): We will support diverse content creation
- [Worldwide Readership](#): Substantially extend our core product experiences
- [Brand Awareness](#): Amplify the unique characteristics of Wikimedia
- [Global Advocacy](#): Build technical & community structures that exemplify our free knowledge policy agenda

Synergies

- We are prioritising development in **structured data**, product & platform **integration**
- addresses some existing needs directly (both in WMF & WMDE teams), lowering their **project costs** and **extending their functionality**



WMF use cases

Abstract WP – “if you can build the monitoring system, Abstract Wikipedia will be using it” – Denny Vrandečić

Architecture – “References as 1st class objects: This is consistent with a ‘software to systems’ model” – Kate Chapman & Moriel Schottlender

Campaigns – “Systematically leveraging relevant sources used elsewhere in our ecosystem to close knowledge gaps” – Felix Nartey

Citoid – “This could make creating better quality and more content-rich citations easier.” – Marielle Volz

Content Translation – “When adding refs in translated articles is like adding images, translators can focus on prose.” – Amir Aharoni

Okapi – “Many high volume reusers would like to be able to parse our citations consistently” - Lane Becker

ORES – “This would be a great training dataset, ‘Editors who add *these* kinds of footnotes do *those* kinds of actions’” - Chris Albon

Mobile – “This would make features like the ‘Featured reference’ prototype on iOS viable at scale” - Josh Minor

SDAW – “This integration would be directly applicable to *SDAW* and connecting reliable content across wikis.” – Amanda Bittaker

The Wikipedia Library – “We could have a field for Proxied URL allowing TWL cardholders direct access” – Sam Walton



More
details
available

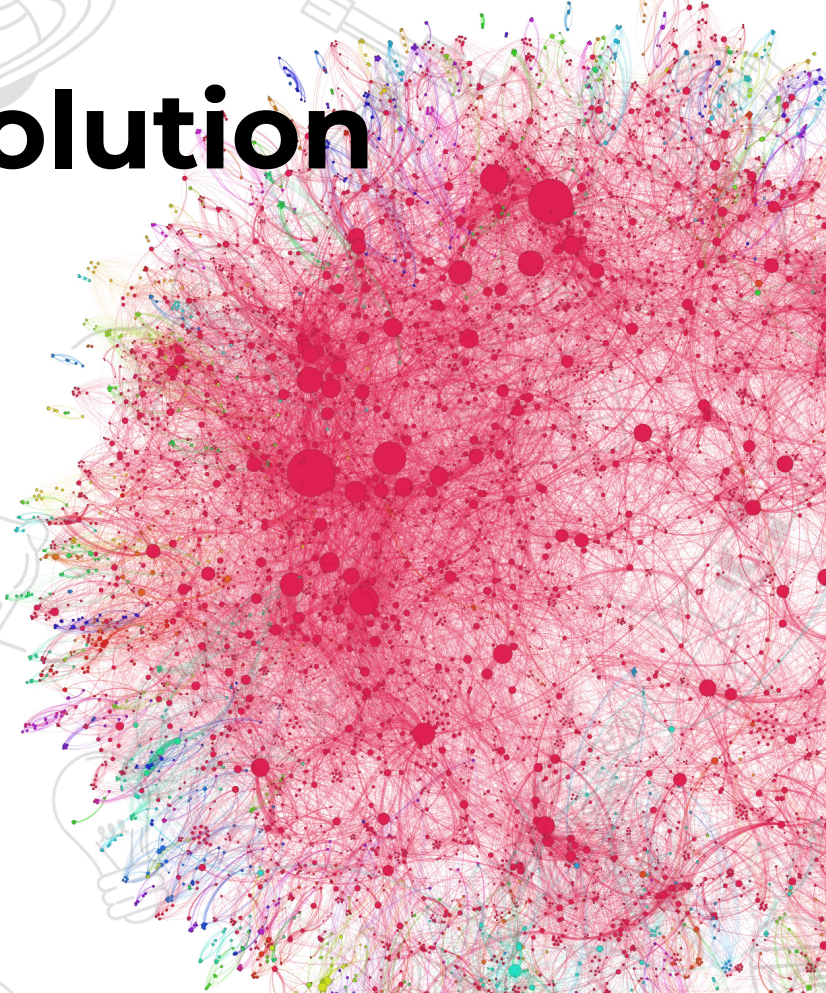
What we might do

1. The proposed solution
2. The principles
3. Examples

Proposed solution

- A service product: central **database** of Wikimedia citations.
- Software integration: improvements to inter-Wiki **monitoring** and **editing**.

These two pillars would empower community-managed workflows and tools to address all the aforementioned user-stories.



Principles

Add upon need

Citation items are created by community members, when they are being used for a reference on a wiki. Not pre-created by an automated process.

Enabled upon readiness

Communities should be supported to be ready to it for it to be enabled on their Wiki, and benefiting from a network effect.

Non-deprecation

Preserve existing referencing systems. Structured references being enabled on a Wiki does not prejudice 'traditional' references.

Style independence

Local wikis determine the display format of a reference for the reader and can be modified by local tools and templates.

Editorial independence

Local wikis determine their content standards. The existence of a shared citation in the database does not prejudice another project's policies.

Community Concerns

The two biggest pain-points expressed by the community, which hinder efforts to use structured-data in the Wikimedia ecosystem, are the [current inability to](#) easily:

Monitoring

Be informed of changes which affect how content is shown **here**, but was edited **over there**.

Editing

See (and ideally change) the content **over there**, without being required to leave **here**.

The status quo can be understood as [incomplete ecosystem integration](#)



More
details
available

1. Monitoring Principles

Granular

Watchlist and **Page-history** must record and reflect every change that affects what is shown to a reader of that page, and only those changes.

Arbitrary

The citation information being changed and displayed on the client wiki (e.g. Wikipedia) might come from any Wikidata item referenced in the citation.

Cascading

A dependency tracking system must propagate notifications through any and all affected items across sites (Wikipedia article, citation item, Wikidata item), regardless of the origin of the change.

Broader cross-wiki integration benefits

Building a unified “Dependency engine” for Shared Citations would be directly applicable other cross-wiki integration like Wikidata Bridge, WikiLambda, Global templates, and longstanding requests in Commons.



More
details
available

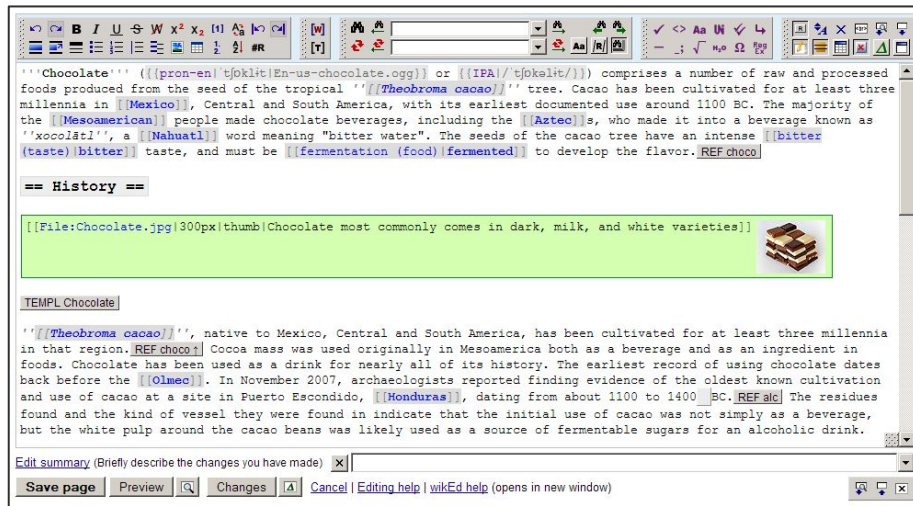
2. Editing Principles

Progressive discovery

An editor must be able, but not required, to know where the citation is hosted, and easily access it.

Visual editing

An editor should be able to search, view the content of the citation, and when possible, edit it in place – in all editing environments.



[Screenshot of “WikEd” gadget](#) (User:Cacycle)

A popular tool among longstanding editors, which features “code folding” references and embedded images – while in 2010 source editor.

Examples

Vindication: A Life of Mary Wollstonecraft (C60384) [English]

Citation

[APA](#) | [MLA](#) | [MHRA](#) | [Chicago](#) | [CSE](#) | [Bluebook](#) | [AMA](#) | [BibTeX](#) | [wiki](#)

Author

String as published: "Lyndall Gordon" First: Lyndall Last: Gordon Initial: L ([Q6708612](#)).

Publisher

String as published: "Harper Perennial" ([Q5663419](#))

Source type

Book – Secondary source

Date/Edition

2005. 3rd, Hardcover. ([4 other editions](#) cited across Wikimedia - with 12 citations)

Edition of

Wikidata item of overarching work: [Q123456789](#)

Citelinks

Total: 105. Current: 98.

By time

First: 11 October 2007, [User:CaptainSaru](#), on English Wikipedia - [Feminist philosophy](#) ([104 other links](#) [expand](#))

by section

Page 34: [Mary Wollstonecraft](#) on en.wikipedia.org ([38 more links to page 34](#) [expand](#))
Pagerange 56-60: [Mary Wollstonecraft](#) on de.wikipedia.org ([9 more links to page 56-60](#) [expand](#))
Chapter 3: [Feminist philosophy](#) on en.wikipedia.org ([12 more links to Chapter 3](#) [expand](#))

by project

En.Wikipedia.org ([9 links](#) [expand](#))
Wikidata.org ([54 links](#) [expand](#))

Fields uniquely possible to a Citation DB

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children (C44583)

Citation

[APA](#) | [MLA](#) | [MHRA](#) | [Chicago](#) | [CSE](#) | [Bluebook](#) | [AMA](#) | [BibTeX](#) | [wiki](#)

Author

String as published: "Dr AJ Wakefield, FRCS" ([Q508568](#)) First: Andrew Last: Wakefield Initial: J

URL

[https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(97\)11096-0/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(97)11096-0/fulltext)

Archive URL

[https://web.archive.org/web/*/https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(97\)11096-0/fulltext](https://web.archive.org/web/*/https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(97)11096-0/fulltext) Saved 207 times between [March 30, 2009](#) and [November 18, 2020](#).

Access

Closed repository [[link](#). Access via [The Wikipedia Library](#)]; OA Preprint [[link](#)]; Shadow Library [[link](#)]

Author

String as published: "Dr AJ Wakefield, FRCS" ([Q508568](#)) First: Andrew Last: Wakefield Initial: J

Source type

Scientific article – Primary source

Publication date

28 February 1998

Publication status

Retracted. Date: 6 February 2010

Citelinks

- zero current Wikimedia project citations - [Former links: 150 across 27 wikis [expand](#)]

*Fields uniquely
possible to a
Citation DB*

The Suffrage Cause Invades Men's Club (C1982)

Citation

[APA](#) | [MLA](#) | [MHRA](#) | [Chicago](#) | [CSE](#) | [Bluebook](#) | [AMA](#) | [BibTeX](#) | [wiki](#)

URL <https://www.nytimes.com/1910/05/25/archives/the-suffrage-cause-invades-mens-club-warm-debate-at-quaint-clubs.html>

Archive-URL (no matching Internet Archive record. [Create?](#))

Author String as published: "Greenhorn G. Reporter" (no matching Wikidata item. [Create?](#))

Source type **Newspaper article** – Primary source

Dateline Page: 3 Date: 25 May 1910 Location: New York

Publisher String as published: "The New York Times" ([Q9684](#))

Tags

Campaign [#1Lib1Ref](#);
Education Programs [#UniOfBolognaHistory101](#)

Lookups

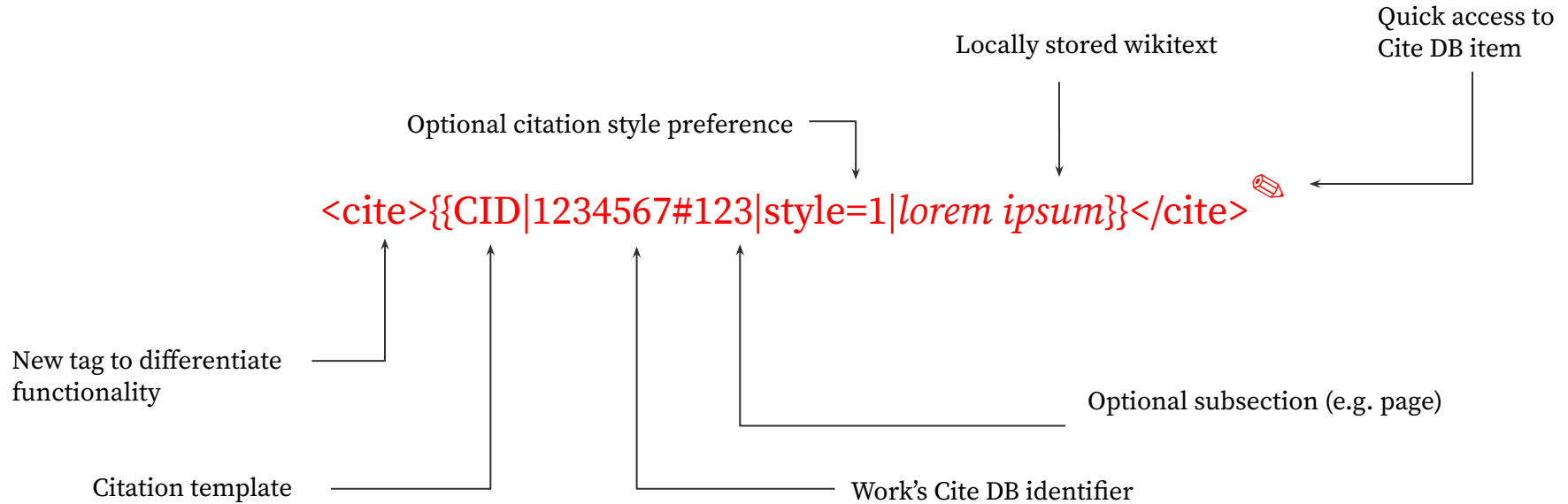
Other Shared Citation records from this:
Publication ([150,234 expand](#)); author (0 [expand](#)); date ([25 expand](#)); location ([1,000,302 expand](#))

*Fields uniquely
possible to a
Citation DB*



More
details
available

Hypothetical implementation of inspecting a shared citation in **source editor**:



Inspecting a citation in **VisualEditor**, and in Wikidata: Would look and feel *the same*



How we might do it

1. Scope
2. Resources

Scope

Citation database	Wikidata
Service product	Sister project
Records about specific publications: Scholarly publications, URLs, editions of books, news articles, archival records	Items about concepts: Authors, publishers, newspapers, websites, topics....
Only that which is cited as a Wikimedia reference. Only created upon their being used, individually.	Any works. Created in advance, en masse, for any use-case.
Examples: <ul style="list-style-type: none">• <i>Metamorphoses</i>. Ovid. Translated by A. D. Melville; introduction and notes by E. J. Kenney. Oxford: Oxford University Press. 2008. ISBN 978-0-19-953737-2.• https://www.theguardian.com/technology/2018/mar/13/youtube-wikipedia-flag-conspiracy• https://doi.org/10.1002/elps.200700396	Examples: <ul style="list-style-type: none">• <i>Democracy in America</i> (Q784882)• <i>Alexis de Tocqueville</i> (Q140694)• <i>Philosophy</i> (Q5891)

Estimated size: ~50m records



Current size: ~90m items

Roadmap

Product

Q3 Current fiscal

- Architectural exploration

1st 6 months

- Design research
- Engineering design and prototype release
- Monitoring + Editing backend research

2nd 6 months

- Launch database; Community building + properties
- Iteration on Monitoring + Editing integration
- First user actions in 1st round target wikis

3rd 6 months

- Completed workflow support for 1st round wikis
- Enable on 2nd round of Wikis
- Populating with content + support community tools

Resources

Time

18 months

Teams

Integrated Product-Platform team

1 PM, 2 front-end engineers, 2 platform engineers, .5 Designer, .5 QA, .5 Analyst, 1 PgM, .5 CL, .5 Eng Mgr.

Overlapping work with existing teams (inc. WMDE)

Community

The [community motivation](#) is strong with citation-focused volunteer projects in all Wikis, external orgs, and grant-supported initiatives.

Metrics

Wikis requesting it enabled; citation records created; Proportion of shared citations on a wiki; no of users creating/using a shared citation.



“Metadata is a love note to the future”

{ } wikicite



Satdeep Gill 2018 CC By-SA
& SLaporte 2017 CC By

Acknowledgements

The following people provided input as part of stakeholder research for this proposal

WMF

Amir Aharoni, Chris Albon, Asaf Bartov, Runa Bhattacharjee, Adam Baso, Amanda Bittaker, Carly Bogen, Kate Chapman, Bryan Davis, Carol Dunn, James Forrester, Satdeep Gill, Danny Horn, Ramsey Isler, Jon Katz, Daniel Kinzler, Samuel Klein, Carolyn Li-Madeo, Erica Litrenta, Josh Minor, Felix Nartey, Margeigh Novotny, Guillaume Paumier, Maryana Pinchuk, Evan Prodromou, Ed Sanders, Joseph Seddon, Moriel Schottlender, Ben Vershbow, Marielle Volz, Caitlin Virtue, Denny Vrandečić, Sam Walton, Leila Zia

Other organisations

WMDE – Lydia Pintscher, Amir Sarabadani, Adam Shorland;
Internet Archive – Mark Graham, Jake Orlowitz; *OpenCitations* – Silvio Peroni;

Community

Jean-Frédéric Berthelot, Anne Clin, Noé Gasparini, James Heilman, Toby Hudson, Andrew Lih, Andy Mabbett, Luca Martinelli, Mahir Morshed, user:MisterSynergy, user:NikkiMaria, Mike Peel, João Peschanski, Siobhan Leachman, Lane Rasberry, Thomas Shaffee, Doug Taylor, Nicolas Vigneron, Andra Waagmeester

And the **WikiCite Steering Committee**: Phoebe Ayers, Daniel Mietchen, Merrilee Proffitt, Alex Stinson, Dario Taraborelli



Appendix



WIKIMEDIA
FOUNDATION

References in Wiktionary*

⇒ **Adjectif** [modifier le wikicode]

vatique \va.tik\ masculin et féminin identiques

1. (*Très rare*) Relatif à la [poésie](#) divinement inspirée, prophétique.

- Et comme il était imbu d'un certain langage qu'on parlait autour de cette femme dans des milieux littéraires : « Elle a quelque chose de [sidéral](#) et même de **vatique**, tu comprends ce que je veux dire, le poète qui était presque un prêtre. » [...] Ces conversations aristocratiques avaient du reste, chez M^{me} de Guermantes, le charme de se tenir dans un excellent français. À cause de cela elles rendaient légitime, de la part de la duchesse, son hilarité devant les mots « **vatique** », « cosmique », « pythique », « suréminent », qu'employait Saint-Loup, — de même que devant ses meubles de chez Bing. — (Marcel PROUST, *À la recherche du temps perdu*, tome 6, page 152 et 205, 1919, Gallimard)

	Singulier	Pluriel
Masculin et féminin	vatique	vatiques
	\va.tik\	

```
{{fr-rég|mf=oui|va.tik}}
''vatique'' {{pron|va.tik|fr}} {{mf}}
# {{très rare|fr}} Relatif à la [[poésie]] [[divinement]] [[inspiré]]e, [[prophétique]].
#* ''Et comme il était imbu d'un certain langage qu'on parlait autour de cette femme dans des milieux littéraires : « Elle a quelque chose de sidéral et même de ''vatique'', tu comprends ce que je veux dire, le poète qui était presque un prêtre. » [...] Ces conversations aristocratiques avaient du reste, chez M<sup>me</sup> de Guermantes, le charme de se tenir dans un excellent français. À cause de cela elles rendaient légitime, de la part de la duchesse, son hilarité devant les mots « ''vatique'' », « cosmique », « pythique », « suréminent », qu'employait Saint-Loup, — de même que devant ses meubles de chez Bing.'' {{source|{{nom w pc|Marcel|Proust}}, ''{{w|À la recherche du temps perdu}}'', tome 6, [[s:Page:Proust - À la recherche du temps perdu édition 1919 tome 6.djvu/154|page 152]] et [[s:Page:Proust - À la recherche du temps perdu édition 1919 tome 8.djvu/207|205]], 1919, Gallimard}}
```

*The differences between the way in which different wiktionaries is possibly even bigger than the differences between wikipedias...



Source: [French Wiktionary - Vatique](#)

References in Wikidata

patronage	 3,738,189	point in time	January 2018
		▶ 1 reference	
	 3,543,007	point in time	October 2018
		▶ 1 reference	
	 3,580,960	point in time	November 2018
		▶ 1 reference	
	 3,860,239	point in time	December 2018
		▶ 1 reference	
	 3,047,596	point in time	February 2018
	▶ 1 reference		
 3,340,873	point in time	March 2018	
	▶ 1 reference		
 3,269,772	point in time	April 2018	
	▶ 1 reference		
 3,304,739	point in time	May 2018	

Each *field* of each reference is stored independently, even if it's the same.

[illegible]

Source: [Wikidata query for references on "São Paulo-Guarulhos International Airport \(Q385406\)"](#)

Internet Archive

Currently planning: *Turn All References Blue* [TARB]

- “a complete inventory of all cited references, to be continually updated.”
- “...A challenge to building this inventory is that cited references can be expressed in many different ways.”

A community-curated Wikimedia Citation DB would allow IA to:

- Focus on archiving, not on our template structures
- Leverage one database
- Identify gaps in their collection, or what needs re-archiving
- Freely access a de-duplicated, up-to-date, curated and filterable dataset of different types of publications, forming the backbone of the TARB project.

F	A	B	C	D	E	F	G	H	I	J	K	L	M
	Site	Language	Last state update	Number of Pages	Wayback Index	Wayback change monthly	Archives links	WebCite links	Other Archive links	Number of pages with Wayback links	Number of pages with Archive links	Number of pages with WebCite links	Number of pages with Other Archive links
1	ia.wikipedia.org	Armenian	20201117	27,386	14,400	1	129	0	41	9,538	117	21	43
2	ia.wikipedia.org	Azərbaycan	20201019	1,070,982	2,758,245	First run	26,261	52,800	3,029	152,778	24,771	43,889	1,207
3	ia.wikipedia.org	Azerbaijani	20201022	173,959	36,018	First run	3,287	11,823	92	14,555	1,021	3,751	77
4	ia.wikipedia.org	Bavarian	20201024	31,492	18,013	4	129	27	20	9,441	110	23	25
5	ia.wikipedia.org	Belarusian	20201027	197,138	39,629	+214	1,808	5,203	13	693	547	107	28
6	ia.wikipedia.org	Bengali	20200928	95,537	117,437	+4,122	4,510	3,743	486	42,104	3,533	1,309	342
7	ia.wikipedia.org	Bosnian	20201005	84,291	79,104	First run	729	373	140	31,275	520	100	125
8	ia.wikipedia.org	Catalan	20201012	3,602	2,646	253	29	0	1	1,494	25	0	3
9	ia.wikipedia.org	German-Kurdish	20201020	28,130	13,932	+530	248	331	27	5,144	157	128	23
10	ia.wikipedia.org	Czech	20201028	464,678	199,028	+15,188	4,634	2,038	430	100,199	3,415	1,428	336
11	ia.wikipedia.org	German	20201012	2,487,308	790,418	+1,588	7,824	11,497	2,726	393,411	6,248	3,561	2,200
12	ia.wikipedia.org	Greek	20201005	182,422	180,085	+1,556	7,340	6,019	314	62,796	4,405	2,097	416
13	ia.wikipedia.org	English	20201016	5,166,977	6,421,218	+47,356	30,858	337,278	45,303	1,528,839	28,641	11,384	28,093
14	ia.wikipedia.org	English	20201012	6,428,371	8,054	+157	1,020	43	7	6,899	1,016	96	6
15	ia.wikipedia.org	Spanish	20201012	1,578,696	1,379,007	+2,042	37,938	30,800	4,724	548,876	24,376	12,314	3,083
16	ia.wikipedia.org	Persian	20201019	749,865	320,411	+2,060	7,409	21,017	552	187,432	5,563	6,542	444
17	ia.wikipedia.org	French	20201008	3,879,064	5,446	+64	0	0	0	3,168	311	2	0
18	ia.wikipedia.org	Galician	20201010	107,439	98,813	+219	3,933	1,299	297	40,353	2,237	787	196
19	ia.wikipedia.org	Hebrew	20201003	276,802	80,399	+1,500	2,926	2,318	379	45,305	2,431	1,144	203
20	ia.wikipedia.org	Hindi	20201027	145,977	366,247	15	3,364	5,047	372	77,999	2,129	1,042	283
21	ia.wikipedia.org	Hungarian	20201021	477,209	325,217	+22,641	7,535	7,000	693	129,835	5,129	3,756	518
22	ia.wikipedia.org	Armenian	20201014	276,213	87,126	First run	4,399	12,342	897	21,015	2,483	9,470	318
23	ia.wikipedia.org	Indonesian	20201029	50,955	1,10	+45	149	0	0	1,243	113	56	629
24	ia.wikipedia.org	Italian	20201019	1,640,870	854,849	+2,719	52,487	19,212	3,677	359,825	18,288	10,838	2,808
25	ia.wikipedia.org	Georgian	20201010	143,175	18,196	First run	3,544	6,066	180	7,581	1,221	1,106	132
26	ia.wikipedia.org	Korean	20201014	523,087	182,175	+1,088	10,762	4,067	631	79,123	5,387	2,672	503
27	ia.wikipedia.org	Latvian	20201015	103,482	56,127	+261	916	1,099	125	25,107	798	987	131
28	ia.wikipedia.org	Lithuanian	20201008	98,421	4,708	First run	166	207	0	1,307	137	107	19
29	ia.wikipedia.org	Deutsch	20201009	14,314	34	First run	18	0	0	29	14	0	0
30	ia.wikipedia.org	Dutch	20201024	2,036,521	213,903	+286	6,816	1,807	373	115,020	5,132	1,090	884
31	ia.wikipedia.org	Norwegian	20201009	341,543	223,164	316	10,521	5,278	937	117,425	7,590	3,047	766
32	ia.wikipedia.org	Portuguese	20201025	1,044,344	489,490	+96,797	30,316	19,630	3,233	219,458	13,862	13,818	2,078
33	ia.wikipedia.org	Romanian	20201007	1,640,356	1,279,420	+112,803	19,246	147,339	4,938	488,132	31,774	13,774	4,066
34	ia.wikipedia.org	English	20201007	174,129	20,009	First run	1,226	5,760	231	11,099	691	2,204	107
35	ia.wikipedia.org	English	20201019	756,399	41,434	40	979	34	20	35,424	914	34	20
36	ia.wikipedia.org	Armenian	20201002	89,880	28,066	+316	1,110	872	127	12,195	877	285	109
37	ia.wikipedia.org	Serbian	20201026	639,428	246,529	+19,823	7,112	2,833	10,534	128,291	4,441	1,267	10,229
38	ia.wikipedia.org	Slovak	20201022	1,640,449	412,218	5,848	412,218	10,240	10,240	216,469	28,042	12,000	2,720
39	ia.wikipedia.org	Telugu	20201004	69,707	96,530	+98	928	944	124	26,831	680	287	89
40	ia.wikipedia.org	Turkish	20201019	375,728	614,742	+298,120	16,930	21,404	1,217	206,287	12,267	637	837
41	ia.wikipedia.org	Ukrainian	20201013	1,048,274	479,542	+2,381	19,366	185,526	1,138	238,962	11,114	87,103	1,033
42	ia.wikipedia.org	Yiddish	20201012	180,053	47,065	First run	189	676	35	10,570	398	298	72
43	ia.wikipedia.org	Vietnamese	20201029	1,257,925	128,047	+4,090	33,823	10,378	879	61,456	31,075	5,283	628

[The IA Bot dashboard](#), tracking the progress of extracting URLs from Wikipedias, and re-inserting archive-urls to the reference – a task undertaken on each wiki separately.



Knowledge equity

*No citations about us,
without us.*

Real time visibility into our citation graph allow us to:

- Help marginalized languages, communities and subject specialists to curate their reliable sources for easy reuse.
- Reduce the citation management workload, especially for smaller communities.

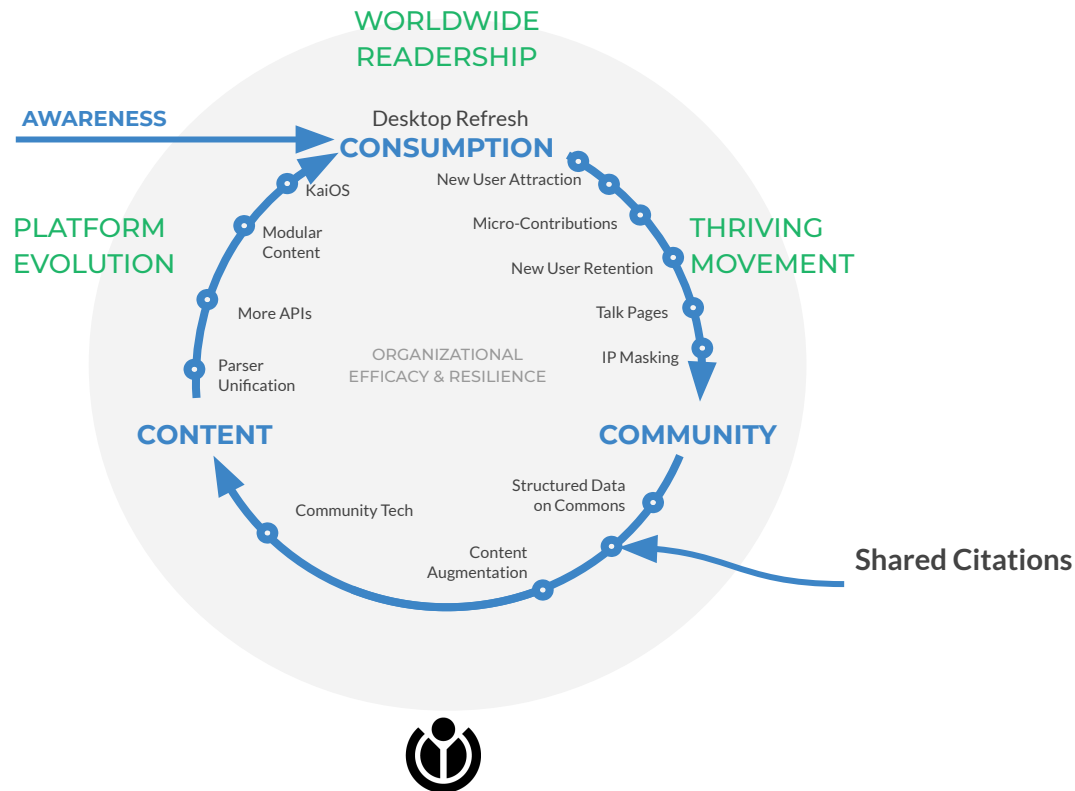
Key Opportunity: How might we identify citation gaps or imbalances, while not supercharging inequity?

Citations as a service

Structured citations contribute to the semanticization of our content and is core to ***Knowledge as a Service to the ecosystem*** around Wikimedia.

Key Problem: How might reusers take advantage of a language-agnostic citation graph?

The product flywheel



Addressing community concerns

[2014 English Wikipedia “delete” comments for Template:DOI](#)

- “We, as a wiki, have chosen to sacrifice the benefits of consistency and elimination of redundancy in exchange for ease of editing and the idea that hitting “edit” on an article actually allows you to edit the article, not just the meta-structure of the article...”
 - “Page watching is a major concern as well, as someone could change the cite doi child template, but that edit would go unnoticed and disconnected from the article itself...”
 - “I *hate* the giant clots of wikisnot left in article markup by cite templates and migrating citation data to wikidata can't come fast enough, but this was never a great solution...”
-

[2017 English Wikipedia “delete” comments for Template: Cite Q](#)

- “References are part of the page content, and editors should be able to view, edit and copy them by viewing the page source. Putting them on other pages made them harder to find and edit, harder to monitor for vandalism, harder to re-use by copy and paste....”
- “Maybe one day we will be able to edit Wikidata content within WP pages, and watch it for changes with other page changes. At that point it might make sense to host content such as references on Wikidata. But right now it makes the editing experience significantly worse and so should not be used...”
- {{cite doi}} was bad enough, but at least when an editor found the relevant page it was a recognisable citation, and could be examined, edited, copied as if it were in an article. Putting them on Wikidata is far worse, as it is a separate project and the way it works is completely unlike WP.



Malicious content

A service that centrally hosts content that is **visible to readers** of all sister-projects creates a new vector for malicious content contribution.

Mitigation:

- Consistent with [Risker's checklist for content-creation extensions](#), ensure the integration of:
 - checkuser & edit suppression/oversight [e.g. if name of a citation is changed to a libellous statement]
 - content logs, and user logs
 - #tags in edit history – indicating the username, the sisterproject originating the edit, and ‘via citations DB’
- Pan-Wikimedia project community-curation (via integration of monitoring tools) spreads the burden of patrolling

Reduced risk:

- Citation items are not ‘editorial’ content – there is minimised scope for edit warring
- Citation content and form are disaggregated – fields in the DB don’t need to be shown in the article
- Purely metadata, no mass-import, means very limited scope for copyright/database-rights violations



More
details
available

Spam

A service that centrally hosts content visible to **reusers** (e.g. search engines) creates a new vector for spam link promotion.

Mitigation:

- Centralising citations allows easier identification of poor quality links from common domains, especially useful to smaller language communities.
- Auto-delete Cite DB item when the reference was rapidly removed from its originating location
- Create Cite DB items only when used in mainspace (?)

Reduced risk:

- Being a service project, there is limited direct audience and therefore reduced spammer motivation
- Nofollow tags decreases SEO spam

Answers.com (WikiAnswers)		1 2 3 4	2010	Answers.com (previously known as WikiAnswers) is a Q&A site that incorporates user-generated content. In the past, Answers.com republished excerpts and summaries of tertiary sources, including D&B Hoovers, Gale, and HighBeam Research. Citations of republished content on Answers.com should point to the original source, with a note that the source was accessed "via Answers.com". Answers.com also previously served as a Wikipedia mirror; using republished Wikipedia content is considered circular sourcing.	1
Apple Daily		2020	2020	The consensus is that Apple Daily is often but not always reliable, and that it may be appropriate to use it in articles about Hong Kong, but subject to editorial judgment, particularly if the topic is controversial and/or Apple Daily is the only source for a contested claim. There is concern that historically, it was not necessarily as reliable as it is today.	1
Arab News		2020 1 2 3	2020	There is consensus that Arab News is a usable source for topics unrelated to the Saudi Arabian government. As Arab News is closely associated with the Saudi Arabian government and is published in a country with low press freedom, editors consider Arab News biased and non-independent for Saudi Arabian politics, and recommend attribution for its coverage in this area. Some editors consider Arab News unreliable for matters related to the Saudi Arabian government.	1
Ars Technica		1 2	2012	Ars Technica is considered generally reliable for science- and technology-related articles.	1 2
arXiv		1 2 3 4 A B	2015	arXiv is a preprint (and sometimes postprint) repository containing papers that have undergone moderation, but not necessarily peer review. There is consensus that arXiv is a self-published source, and is generally unreliable with the exception of papers authored by established subject-matter experts. Verify whether a paper on arXiv is also published in a peer-reviewed academic journal; in these cases, cite the more reliable journal and provide an open access link to the paper (which may be hosted on arXiv).	1
AskMen		1 2 3 4 5 6	2020	There is no consensus on the reliability of AskMen. See also: IGN.	1
Associated Press (AP)		1 2 3 4 5 6	2018	The Associated Press is a news agency. There is consensus that the Associated Press is generally reliable. Syndicated reports from the Associated Press that are published in other sources are also considered generally reliable.	1 2
The Atlantic (The Atlantic Monthly)		1 2 3	2019	The Atlantic is considered generally reliable.	1
The Australian		1 2	2020	The Australian is considered generally reliable. Some editors consider The Australian to be a partisan source. Opinion pieces are covered by WP:RSOPINION and WP:NEWSBLOG. Several editors expressed concern regarding their coverage of Climate Change related topics.	1
The A.V. Club		1 2 3 A	2014	The A.V. Club is considered generally reliable for film, music and TV reviews.	1
Axios		1 2	2020	There is consensus that Axios is generally reliable. Some editors consider Axios to be a biased or opinionated source. Statements of opinion should be attributed and evaluated for due weight.	1
Baidu Baike		2020 1 2 3 4	2020	Baidu Baike was deprecated in the 2020 RIC as it is similar to an open wiki, which is a type of self-published source. Although edits are reviewed by Baidu administrators before they are published, most editors believe the editorial standards of Baidu Baike to be very low, and do not see any evidence of fact-checking. The Baidu 10 Mythical Creatures kuso originated from Baidu Baike.	1 2

[English Wikipedia's "Perennial Sources"](#). Only very few Wikipedias have an equivalent page. Pooling resources would help communities identify [un]reliable sources in other languages.



Why *not* within Wikidata?

Wikidata is capable! There is a vibrant community, extensive data modelling, and corpus of existing citation content. For example:

Scale: A separate database means that *all* Wikimedia citations can “Fit”. E.g. the 10s of millions of citations to specific URLs.

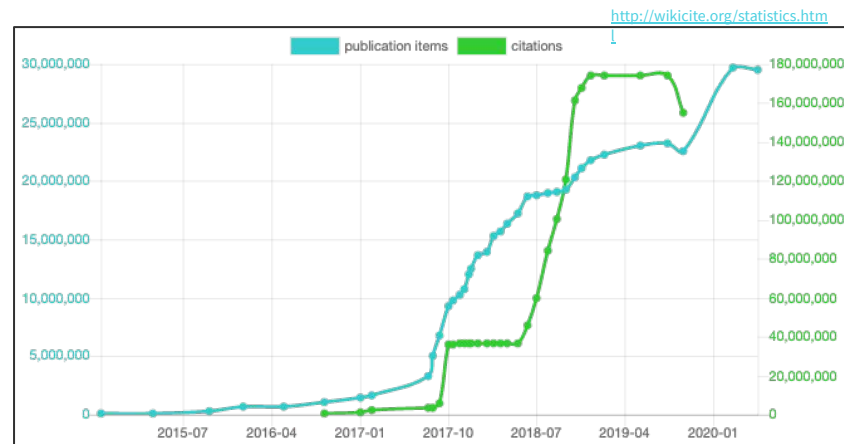
Service: By serving *all* Wikimedia *projects*, Wikidata itself can equally benefit from the shared citations, not merely host them.

Scope #1: Restricting to *only* Wikimedia citations ensures the ontology remains practical.

Scope #2: Restricting to *only* creation-upon-need ensures the community understands the difference from WD.

Sovereignty

WD and Cite DB would be editorially independent. **Some content overlap**, but would be overwhelmingly limited to individual scholarly journal articles, some book editions. WD does not normally have items about URLs, newspaper articles etc. Content policies would evolve.



“Let’s do it.” – Lydia Pintscher; “Sounds great.” – Adam Shorland; “It looks awesome to me” – Amir Sarabadani



Case study: Commons

Launched in 2004 – three years after Wikipedia – **Wikimedia Commons**’ role is to **reduce duplication of effort** across the Wikimedia projects.

Prior to this, a multimedia file had to be **locally hosted** in order to be able to be used.

- Many exactly matching files were uploaded to each Wikipedia language
- Metadata and captions were duplicated, but inconsistently and monolingually
- Curation of multimedia was handled by each community separately

The “service project” Commons.wikimedia.org [note: not WikiCommons.org, which implies a “sister project” status] centralised that work. The community later **expanded its scope** to include *any* freely-licensed multimedia.

Citations DB serves an equivalent role and, similar to continued “local upload” of multimedia, traditional citations would still operate in parallel. By contrast, the existence of **Wikidata**’s already fulfils the expanded scope use-case.

Commons *still* lacks the ability to notify Wikipedia when images-in-use are changed. That would become possible with the “monitoring” features in this proposal.



Wikimedia Commons mosaic, celebrating 1M files in 2006. Nux, CC By-SA



"communities" ensuring diversity built-in to the

structured format

ar with the format

es and content should be mapped to each other

languages

network effect'

Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch & Richard Cyganiak. <http://lod-cloud.net/>

2. Wikidata

- All content is a citation destination
- Most active community entirely among “e culture from the beginning
- Built on MediaWiki, the primary use-case

2. Wikidata

- Large number of citations already in a consistent structured format
- Multilingual and multicultural community, familiar with the format
- Necessary community discussion on how properties and content should be mapped to each other
- Built on WikiBase, requiring separate UI work

4. Wikipedias and other sister projects

- Community familiar with citation management
- Diversity of citation types, template formats, and languages

- #### 4. Wikipedias and other sister projects

- Largest block of content - most benefit from the ‘network effect’
- Largest variety of citation formats and edge-cases



Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch & Richard Cyganiak. <http://lod-cloud.net/>