Mako has been doing this for many years. It started in 2008 and has happened almost every single year since.

This began as an excuse for Mako to make sure he was up to date on Wikimedia Research.
“This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – of the last year's academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project.”

– From Mako’s Wikimania 2008 Submission

Back at Wikimania 2008, Mako set out to run a session that would provide a comprehensive literature review of articles in Wikipedia published in the last year. Quote Mako:

“This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – of the last year's academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project.”

– From Mako’s Wikimania 2008 Submission

“Then, about two weeks before Wikimania, I did the scholar search so I could build the literature.”
“This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – of the last year's academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project.”

– From Mako’s Wikimania 2008 Submission
“This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – of the last year's academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project.”

– From Mako's Wikimania 2008 Submission

“I tried to import the whole list into Zotero and managed to get banned for abusing Google Scholar because they thought that no human being could realistically consume the amount of material published on Wikipedia that year.

So anyway, I had a 45 minute talk so it worked out to 3.45 seconds to per paper...

And believe it or not, this year is even bigger.”
Academics have written a lot of papers about Wikipedia. There are more than 500 papers published about Wikipedia each year and although we've reached and moved past a peak it seems, it's not slowing by much.

(Source: Google scholar results. Accessed: 2019-08-17)
• 8,700 Wikipedia-related publications in the Scopus database as of this week (August 14, 2019)

• 109 recent publications covered in the 8 issues of the Wikimedia Research Newsletter from June 2017 to June 2018 (and hundreds more on our list!)

The newsletter aims to be comprehensive, but mostly ignores papers that use Wikipedia as a corpus only (which is popular e.g. in NLP research).
In selecting papers for this session, the goal is always to choose examples of work that:

- Represent **important themes** from Wikipedia in the last year.
- Research that is likely to be of **interest** to Wikimedians.
- Research by people who are **not at Wikimania**.
- ...with a bias towards **peer-reviewed** publications.

Within these goals, the selections are **incomplete**, and **wrong**.
Speaker: Aaron

Gender bias in coverage, content, deletion, etc. versus gender gaps (usually in participation, engagement, etc.)
• Sample ~3,000 U.S. sociology faculty in 2014.
• Matched to living sociologist articles on EN:WP.
• Compare who has articles across demographic categories, scholarly status indicators, and accomplishments.
• Also analyze article subject gender and deletion decisions.
Explain figure

- Nonwhite and female sociology faculty disproportionately less likely to be represented on EN:WP.
- Seniority and citations matter a lot too.
- No difference in deletion rate by gender.
Adams et al.: Gender bias

"Supply" vs. "Demand" side explanations & responses

- supply: issues w underrepresentation in the world
- demand: issues w gatekeeping, exclusion in the community.
- importance of WP as space for representation
- call for more systematic efforts to make representation more equal (500 women scientists)
Speaker: Isaac

Long-standing area of research on what leads to quality information on Wikimedia projects. I'm going to try to get two papers in so bear with me – one from Aaron Halfaker that is very applied and looks at supporting editors in evaluating new articles – so how do we ensure that high quality information makes it into the wikis – and another that is motivated by the many discussions around political polarization and whether that is damaging article quality on Wikipedia – so what processes lead to high quality articles.
https://doi.org/10.1145/3274290
Problem: large backlog in New Page Patrol (NPP) and Articles for Creation (AfC)

Goal: route new article (drafts) to relevant experts based on WikiProjects

Developed set of 43 mid-level categories based on WikiProject directory

Extracted dataset of 93,445 EN:WP articles and associated WikiProject labels

Built machine learning (GradientBoosting) models to predict labels from article text (word embeddings)
• Achieved >90% accuracy
• More importantly, ease of iteration and improving on labeling and model

https://doi.org/10.1038/s41562-019-0541-6
Shi et al.: Information Integrity / Quality

- Goal: understand whether Wikipedia editors with high political polarity can create high-quality articles
- Gathered edit history of EN:WP articles in categories related to politics (20,947 articles), social issues (162,085 articles), and science (49,530)
- Calculated political polarity of editors based on how much they edit liberal / conservative articles
- Evaluated political polarity scores through surveys of active editors (500 requests; 118 responses)
- For each category, build ordinal logistic regression relating article quality to the political polarity of editors
Relationship between article length and political polarity by topic
Table 1: Odds ratios from ordinal logistic regression models predicting article quality

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Politics</th>
<th>Social issues</th>
<th>Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>polarization</td>
<td>18.88 ***</td>
<td>2.06 ***</td>
<td>1.79 **</td>
</tr>
<tr>
<td>alignment</td>
<td>0.30 ***</td>
<td>0.49 ***</td>
<td>0.65 **</td>
</tr>
<tr>
<td>editing experience</td>
<td>1.05 *</td>
<td>1.06 ***</td>
<td>1.01</td>
</tr>
<tr>
<td>number of editors</td>
<td>0.41 ***</td>
<td>0.51 ***</td>
<td>0.56 ***</td>
</tr>
<tr>
<td>article length</td>
<td>33.55 ***</td>
<td>47.83 ***</td>
<td>56.54 ***</td>
</tr>
<tr>
<td>number of edits</td>
<td>3.26 ***</td>
<td>1.71 ***</td>
<td>1.69 ***</td>
</tr>
<tr>
<td>N</td>
<td>12,570</td>
<td>161,070</td>
<td>49,995</td>
</tr>
</tbody>
</table>

Note: *, **, *** denote statistical significance levels of 0.1, 0.01 and 0.001, respectively. The columns present odds ratios estimated on Political, Social issues and Science articles, separately.
How people read Wikipedia
How people read Wikipedia

Goal: identify “prerequisite relations between learning resources”, i.e.

- To learn about A, I first need to learn about B
- To learn about B, I first need to learn about C and D
- etc.

Important especially for self-learners using online materials such as Wikipedia articles

Explain the image

(Image source: Figure 1 from the paper, “Sample learning path from Metacademy”, cropped)
Goal: automatically generate prerequisite relations between concepts, by using navigation patterns of Wikipedia readers.

Used the “Wikipedia Clickstream” dataset - a sample of referrer data from Wikipedia's weblogs, showing how many readers clicked from one article to another.

Built classifier, trained on prerequisite relations from Metacademy (online learning platform) and from Carnegie Mellon University.

Tested various features drawn from the Clickstream data. As might be expected, the number of backwards clicks from a more advanced concept to one of its prerequisites turned out to be an important feature.

This dataset has been at the base of multiple other research papers since it was first released by WMF some years ago.

Skipping over a lot of ML detail here (e.g. they also had to match Wikipedia articles to Metacademy concepts).
Result: A well-performing algorithm that (according to the authors) improves on previous research and generates data that should be useful for self-directed learners using web resources in general.

How could this data be used on Wikipedia itself, to make it easier for readers to plan their learning process?
Economists have previously attempted to calculate the monetary value of Wikipedia and other free content used across the internet.

An interesting discussion as well surrounds the value of Wikimedia Commons.

• This paper attempts to quantify the monetary value of Wikimedia Commons by tracking reuse of images from Wikimedia Commons across the internet.

• They did so by asking - how much the licensing of images on WC would generate if it was operated under a for-profit model such as that of Getty Images.
Impact of Wikimedia projects on the world

- The authors applied an automated reverse-image search to a dataset of 10,000 random images on WC to determine where they had been used on the internet.
- The domain of each re-use was then evaluated to determine whether, for instance, it was a commercial entity (i.e. .com) or a non-commercial entity like (i.e. .org)
- Some 15% of the sample consisted of freely-licensed images that were pulled from Flickr.
Impact of Wikimedia projects on the world

• Assumptions
  • USD $175 for commercial
  • USD $60 for non-commercial use
• Results
  • 34.8% of images used externally
  • USD $28.9 billion for Wikimedia Commons

Using Getty’s licensing model of USD $175 for commercial use and USD $60 for non-commercial use,
• 34.8% images in the sample of 10,000 (excluding images previously used on Flickr) were used at once.
• they extrapolate out how often on average each image is used (and where) to reach a total estimate of USD $28.9 billion for Wikimedia Commons.
The assumptions to this conclusion are not without question.

How much would people have paid if Wikimedia Commons was not operated under a for-profit model?

This research has a lot of limitations and assumptions that are made regarding how much money individuals would pay to re-use Commons images if they were not openly available.

These assumptions are common to much work in this field of economics though.

This is also an early attempt at estimating a value, and we expect to see a lot more papers of this variety in the coming years with the rise of Wikidata and its CC0 license and more direct incorporation of Wikipedia into search and voice assistants.

So we see this as one attempt to value Commons, but hopefully more estimates will be generated with complementary approaches that will help us get a better sense of the scale of the answer for Commons as well as Wikipedia, Wikidata, and any other commons-based peer production products.
Community governance & participation

Speaker: Reem & Matej

- Governance and participation have been studied for a long time.
- New focus on changes over time and new projects.
What was studied:

• 62 policy pages (incl. pillars)
• 175 guideline pages
• 1476 essay pages
• 311 failed rule pages
• 265 248 revisions
• 460 124 talk page revisions
How have rule-making patterns on Wikipedia changed over time?

• More than a decade after the first policies were created, people are still heavily discussing and making rules.

• However, more attention is given to older rules than newer rules. They get more discussion but fewer changes.
How have patterns of user contributions to Wikipedia rules changed over time?

- Editors who are involved in rule-making are not always focused only on editing the rules and also they edit different kinds of rules.
- Also new rule-making edits started to take place in talk pages (again deliberation) more than policy pages themselves.
How does participation in Wikipedia rulemaking change editors’ behavior?

• Getting involved in rule-making is a chance to rediscover shared values or to talk about best practices.
• Making and discussing rules doesn’t distract users from writing articles.
Figure 3: The total number of revisions to the policy (red) and policy talk (blue) pages by year.
Figure 6: The average annual editor revisions to male pages, stratified by year the editor began editing male (colored line).

https://dl.acm.org/citation.cfm?doid=3290265.3274410
Fig. 5. Proportion of contributions per user type and by yearly cohort over time and percentage of users per type. The count of anonymous users is based on unique IP addresses, as these users are only known through them. Nothing prevents editors to connect from different addresses, though. Years in (c) refer to the period between October of the previous year and September of the following (e.g. 2013 means Oct. 2012–Sep. 2013).
Those are our six exemplary studies from the past year. There has been just tons and tons of work in this area. Trying to talk about this in 45 (or even 25) minutes seems increasingly crazy every year we try to do it. A major sin of omission this year is more research about Wikidata and research about more language communities (although both have been covered elsewhere at the conference!).

The most important source is the Wikimedia Research Newsletter which has since 2011 been published monthly in the (English) Signpost and syndicated on the Wikimedia Research space on Meta-Wiki. (Special thanks to Miriam Redi, Dario Taraborelli, Tilman Bayer and User:Masssly for finding and cataloguing new publications throughout the year!)

But there are other resources as well. And we encourage you to get involved.