I've been doing this for many years. I started in 2008 and have done this almost every single year since.

This began as an excuse for me to make sure I 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 my Wikimania 2008 Submission

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

“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 my 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 my Wikimania 2008 Submission

I tried to import the whole list into Zotero and managed to get banned for abusing the 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.

And my talk is even shorter.
“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 my Wikimania 2008 Submission

I tried to import the whole list into Zotero and managed to get banned for abusing the 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.

And my talk is even shorter.
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.
- **2968** Wikipedia-related publications in the Scopus database as of November 2013
- **191** recent publications reviewed or mentioned in the 12 issues of the Wikimedia Research Newsletter from July 2014 to June 2015.
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**.
Wikipedia as a Source of Data
Traditional methods rely on:

- Population of speakers
- Income or political power of speakers

Paper presents new network method based on measuring co-speakers of languages in several data sources including Wikipedia.
Two languages are connected when users that edit an article in one Wikipedia language edition are significantly more likely to also edit an article in the edition of the other language.

If an editor of Spanish is also likely to edit Galician, we’ll call those languages connected.
The number of people per language (born 1800–1950) with articles in at least 26 Wikipedia language editions as a function of their language’s eigenvector centrality.

The bottom row shows the number of people per language (born 1800–1950) listed in *Human Accomplishment* (a book by Charles Murray) as a function of their language’s eigenvector centrality.
Community and Organization
Perfect Alignment Hypothesis (PAH): There is an exact match between the supply of high-quality content and the demand for it.

<table>
<thead>
<tr>
<th></th>
<th>$PAH_1$</th>
<th>$PAH_2$</th>
<th>$PAH_3$</th>
<th>$PAH_4$</th>
<th>$PAH_5$</th>
<th>$PAH_6$</th>
<th>$PAH_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_1$</td>
<td>1,710,819</td>
<td>477,687</td>
<td>30,701</td>
<td>6,647</td>
<td>657</td>
<td>16</td>
<td>64</td>
</tr>
<tr>
<td>$Q_2$</td>
<td>454,270</td>
<td>477,547</td>
<td>92,585</td>
<td>37,148</td>
<td>6,130</td>
<td>190</td>
<td>852</td>
</tr>
<tr>
<td>$Q_3$</td>
<td>43,255</td>
<td>71,012</td>
<td>26,749</td>
<td>19,056</td>
<td>6,259</td>
<td>232</td>
<td>1,344</td>
</tr>
<tr>
<td>$Q_4$</td>
<td>14,408</td>
<td>30,669</td>
<td>13,707</td>
<td>12,102</td>
<td>5,447</td>
<td>262</td>
<td>1,351</td>
</tr>
<tr>
<td>$Q_5$</td>
<td>3,649</td>
<td>9,416</td>
<td>3,192</td>
<td>2,136</td>
<td>953</td>
<td>62</td>
<td>506</td>
</tr>
<tr>
<td>$Q_6$</td>
<td>132</td>
<td>398</td>
<td>128</td>
<td>92</td>
<td>31</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>$Q_7$</td>
<td>59</td>
<td>1,994</td>
<td>846</td>
<td>766</td>
<td>438</td>
<td>32</td>
<td>218</td>
</tr>
</tbody>
</table>

Quality: Stub, Start, C, B, Good Article, A, Featured Article
Popularity: equivalently sized buckets
Measure of the degree of misalignment can be used to build lists of categories that are relatively “overproduced” and “underproduced”:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Topic</th>
<th>N</th>
<th>Rel. Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Countries</td>
<td>144</td>
<td>506.9</td>
</tr>
<tr>
<td>2</td>
<td>Pop music</td>
<td>97</td>
<td>38.9</td>
</tr>
<tr>
<td>3</td>
<td>Internet</td>
<td>84</td>
<td>37.6</td>
</tr>
<tr>
<td>4</td>
<td>Comedy</td>
<td>134</td>
<td>21.9</td>
</tr>
<tr>
<td>5</td>
<td>Technology</td>
<td>58</td>
<td>15.8</td>
</tr>
<tr>
<td>6</td>
<td>Religion</td>
<td>121</td>
<td>15.8</td>
</tr>
<tr>
<td>7</td>
<td>Science Fiction</td>
<td>70</td>
<td>15.5</td>
</tr>
<tr>
<td>8</td>
<td>Rock music</td>
<td>84</td>
<td>11.4</td>
</tr>
<tr>
<td>9</td>
<td>Psychology</td>
<td>60</td>
<td>11.1</td>
</tr>
<tr>
<td>10</td>
<td>LGBT studies</td>
<td>136</td>
<td>9.1</td>
</tr>
</tbody>
</table>

Table 8: Topics most strongly over-represented in the Needs Improvement (NI) dataset, limited to topics w/ at least 50 NI articles. “N” column lists number of NI articles.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Topic</th>
<th>N</th>
<th>Rel. Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cricket</td>
<td>65</td>
<td>159.0</td>
</tr>
<tr>
<td>2</td>
<td>Tropical cyclones</td>
<td>112</td>
<td>99.3</td>
</tr>
<tr>
<td>3</td>
<td>Middle Ages</td>
<td>87</td>
<td>13.4</td>
</tr>
<tr>
<td>4</td>
<td>Politics</td>
<td>147</td>
<td>12.0</td>
</tr>
<tr>
<td>5</td>
<td>Fungi</td>
<td>53</td>
<td>9.1</td>
</tr>
<tr>
<td>6</td>
<td>Birds</td>
<td>78</td>
<td>8.2</td>
</tr>
<tr>
<td>7</td>
<td>Military history</td>
<td>404</td>
<td>5.3</td>
</tr>
<tr>
<td>8</td>
<td>Ships</td>
<td>88</td>
<td>5.0</td>
</tr>
<tr>
<td>9</td>
<td>England</td>
<td>72</td>
<td>4.9</td>
</tr>
<tr>
<td>10</td>
<td>Australia</td>
<td>258</td>
<td>4.3</td>
</tr>
</tbody>
</table>

Table 9: Topics most strongly over-represented in the Effort (SE) dataset, limited to topics w/ at least 50 SE articles. “N” column lists number of SE articles.
A decade after the landmark “Nature” study, there still aren’t too many systematic evaluations of the accuracy of Wikipedia’s content. Health articles continue to receive scrutiny, though. With good reason: Wikipedia is “the most frequently consulted online health care resource globally” [NEJM article].

The US Food and Drug Administration (FDA) frequently issues safety warnings about prescription drugs. How long does it take until these are reflected on English Wikipedia?

- 41% updated within two weeks (58% for high-prevalent diseases), but 36% still unchanged after more than a year.

Tilman

Articles about drugs used to treat high-prevalent diseases (affecting > 1 m Americans / year) were updated faster. But the result still caused concern. Authors find "there may be a benefit to enabling the FDA to update or automatically feed new safety communications to Wikipedia pages, as it does with WebMD". The paper raised awareness among WikiProject Medicine editors, but there's no systematic updating mechanism yet.
Selected 100 drugs from German undergrad curriculum in pharmacology

Extracted information from two standard textbooks

"Accuracy of drug information in [German] Wikipedia was 99.7%±0.2% when compared to the textbook data." Similar results for English Wikipedia
Completeness (as compared to the textbooks):
- 83.8% (of 224 statements) for German WP
- 87.2% for English WP

Completeness of contraindications information was 100% in the En WP sample.

English WP cited academic publications more often than German WP.

Quality "significantly improved" in drug articles assessed in a 2010 study.
Tilman

Starting to see more practical applications of AI methods to editing. Bots have been writing Wikipedia articles ever since back in 2002, User:Rambot covered US municipalities from US census data. Picked these two related papers for their somewhat unusual approach.
Banerjee et al., *Playscript Classification and Automatic Wikipedia Play Articles Generation*. 2014
22nd International Conference on Pattern Recognition (ICPR). pp. 3630–3635. DOI:10.1109/ICPR.2014.624
Author’s copy
Bot searches for playscripts and related documents on the web

Extract key information from them, e.g.
- The play’s main characters
- Relevant sentences from online synopses of the play
- Mentions in Google Books and Google News (as evidence that the play satisfies Wikipedia’s notability criteria)

Some heuristics to exclude non-encyclopedic sentences, e.g. first person statements

Tilman
NB: Most article creation bots work from well-defined databases (e.g. species, census data, geographical databases).
This bot finds article topics and online references itself, using an elaborate classifier algorithm to distinguish scripts from non-scripts.
15 articles submitted at Articles for Creation. Two accepted by Wikipedia editors. One of them without major changes.

Editors were unaware the articles had been automatically generated.

Related paper by some of the same authors:
Banerjee et al., *WikiKreator: Improving Wikipedia Stubs Automatically*. Preprint: 

Elaborate classifier method to find suitable web resources for expanding stubs - but copying sentences wholesale from these into articles landed the bot (User:MightyPepper) in a contributor copyright investigation
(https://en.wikipedia.org/wiki/Wikipedia:Contributor_copyright_investigations/Archive#2015...
Aaron:
Research focused on understanding gender dynamics in Wikipedia and their impact is another area of research that has continued to expand this year. A number of high quality papers came out, several of which analyzed how gender figures in the content of the encyclopedias.
Wagner, Claudia; David Garcia; Mohsen Jadidi; and Markus Strohmaier. 2015. “It’s a Man’s Wikipedia? Assessing Gender Inequality in an Online Encyclopedia.” Ninth International AAAI Conference on Web and Social Media (ICWSM).
We know there's a gender gap.

Need for more multidimensional analysis of how gender is represented in content of articles across Wikipedias.
Use data from three sources (Freebase, “Human Accomplishment,” and Pantheon) as baselines for comparison with six Wikipedias (EN, ES, DE, FR, IT, RU).

Examine multiple potential forms of bias: coverage, structure, lexical characteristics, visibility.
Some key findings:

- 1: Coverage of women (# articles, length) in WPs is generally better than other sources.
- 2: Articles about women tend to be less centrally connected in the network of articles than articles about men (Smurfette!)
- 3: (viz) Content of articles about women uses different words than those about men. Much higher incidence of language related to family, gender, and relationships.
Some key findings:

• 1: Coverage of women (# articles, length) in WPs is generally better than other sources.

• 2: Articles about women tend to be less centrally connected in the network of articles than articles about men (Smurfette!)

• 3: \textit{(viz)} Content of articles about women uses different words than those about men. Much higher incidence of language related to family, gender, and relationships.
Adopting Wikipedia as a Teaching Tool

Aaron:
Research focused on understanding how Wikipedia and related resources are adopted for classroom teaching. Growing area of work, still somewhat preliminary findings. Nonetheless, some of the papers in this domain make for entertaining reading...
Students use Wikipedia uncritically. Don’t understand how low quality much of the information may be or how it may be manipulated.

Professor (author) believes that WP is full of dubious information. Wants to unmask that for his students.

Through more in-depth exposure, students may understand the limitations of collaborative, open systems of knowledge production.
WP and the Wisdom of Crowds: Methods

- Require a Senior (college) composition class to work on editing WP articles (together and individually) throughout the semester.
- Incorporate assignments to help students learn about the history of WP as well as how to use it.
- Require students to reflect on their experiences in writing.
- Require students to analyze the pros/cons of open collaborative writing in their final projects.

This is all sort of fabulously in-line with exactly what the WikiEd Foundation recommends instructors do (!).
Both sources [crowds and experts] have different merits... My life experience since class pulls me in favor of the wisdom of the crowd. In my recent studies, I have found that I can learn much more from a group of my peers than from a single expert.

— Student 1

Not exactly what the instructor expected. Essentially, both he and the students came away with much more nuanced, and positive, views of the relative merits, possibilities, and limitations of open collaborative knowledge production. A happy ending :)}
Meta-Analyses


Those are my 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 20 minutes strikes me as increasingly crazy every year I try to do it.

The most important source, now going for a couple years, is the Wikimedia Research Newsletter which is published monthly in the (English) Signpost and syndicated on the Wikimedia Research.

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