“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
“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
“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
Number of citation, per year, with the term “wikipedia” in the title.
(Source: Google scholar results. Accessed: 2013-08-06)
- **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
Wikipedia as a Source of Data
How to measure the global influence of languages?

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.
Wikipedia as a source of data: Ronen et al.

Language Family
- Afro-Asiatic
- Altaic
- Amerindian
- Austronesian
- Caucasian
- Creoles & pidgins
- Dravidian
- Indo-European
- Niger-Congo
- Other
- Sino-Tibetan
- Tai
- Uralic

Population
- 1 million
- 10 million
- 100 million
- 1 billion

Link Weight and Color
- t-statistic: 2.59
- min: 6 (users, editors, translations)
- max: 994,682 (users, editors, translations)
- max: 49,637 (wikipedia)
- max: 183,329 (book translations)
Wikipedia as a source of data: Ronen et al.

\[ R^2 = 0.755 \]
\[ p\text{-value} < 0.001 \]
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>PAH₁</th>
<th>PAH₂</th>
<th>PAH₃</th>
<th>PAH₄</th>
<th>PAH₅</th>
<th>PAH₆</th>
<th>PAH₇</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q₁</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>Q₂</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>Q₃</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>Q₄</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>Q₅</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>Q₆</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>Q₇</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>
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.
Content quality
Quality of drug articles


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.
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.
Automation in Wikipedia
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
Fourteen is a play by Alice Gerstenberg. This one act play was originally published in the February issue of Drama Magazine, 1920. It is now a public domain work and may be performed without royalties.

- 15 articles submitted at Articles for Creation. Two accepted by Wikipedia editors. One of them without major changes.
Gender Beyond the Gap
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.
It’s a Man’s Wikipedia: Results
It’s a Man’s Wikipedia: Results

(a) Proportion of the 150 most discriminative words for women per category

(b) Proportion of the N most discriminative words for women per category

Figure 8: **Lexical Bias**: The proportion of the 150 most discriminative words of articles about women that belong to different categories. In all language editions between 32% and 23% of the 150 most indicative words for women belong to one of the three categories, while only between 0% and 4% of the most discriminative words for men belong to one of these categories. In some language edition, like the Russian (RU), the English (EN) and the German (DE) one, the proportion of the most discriminative words that belong to one of these three categories is especially high among the top words.
Adopting Wikipedia as a Teaching Tool
WP and the Wisdom of Crowds: Motivation

- 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.
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
Meta-Analyses


More Resources

- **Wikimedia Research Newsletter**
  [[:meta:Research:Newsletter]] / @WikiResearch
- **WikiSym/OpenSym** (This August in San Francisco!)
- **WikiPapers Repository** [http://wikipapers.referata.com]
- **Much More**