Multilingual User Generated Content at Wikipedia

Alolita Sharma
Director of Language Engineering
Wikimedia Foundation
asharma@wikimedia.org

New Horizons for the Multilingual Web – W3C Multilingual Web Conference
ETSIT-UPM, Madrid, May 7 2013  #mlwmadrid

CC-BY-SA 3.0
Wikipedia-Scale

~31.5m Articles
~4.5m EN
~1.5m DE NL FR SE
~1m IT PL RU ES

532m Monthly uniques

1m–100k Articles
43 Languages
99k–10k Articles
73 Languages
10k–1k Articles
101 Languages
1k–100+ Articles
61 Languages

287 Languages
797 Production websites
567 Incubator websites

21b Monthly page views

4.8b Mobile monthly page views
Wikipedia: Growth by Region

495.18 Million

Mar 13 — Mar 14  -4.33%
Feb 14 — Mar 14  4.44%

Unique Visitors per Region (comScore)

<table>
<thead>
<tr>
<th>Region</th>
<th>Number of Visitors</th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
<td>532.70M</td>
</tr>
<tr>
<td>China</td>
<td>8.44M</td>
</tr>
<tr>
<td>Europe</td>
<td>202.10M</td>
</tr>
<tr>
<td>India</td>
<td>28.18M</td>
</tr>
<tr>
<td>Latin America</td>
<td>70.88M</td>
</tr>
<tr>
<td>Middle East/Africa</td>
<td>44.19M</td>
</tr>
<tr>
<td>North America</td>
<td>98.22M</td>
</tr>
<tr>
<td>Asia Pacific</td>
<td>119.32M</td>
</tr>
</tbody>
</table>
### Wikipedia today

<table>
<thead>
<tr>
<th>№</th>
<th>Language</th>
<th>Language (local)</th>
<th>Wiki</th>
<th>Articles</th>
<th>Total</th>
<th>Edits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>English</td>
<td>en</td>
<td>4,507,774</td>
<td>32,821,477</td>
<td>716,354,339</td>
</tr>
<tr>
<td>2</td>
<td>Dutch</td>
<td>Nederlands</td>
<td>nl</td>
<td>1,774,505</td>
<td>3,205,497</td>
<td>42,212,245</td>
</tr>
<tr>
<td>3</td>
<td>German</td>
<td>Deutsch</td>
<td>de</td>
<td>1,715,539</td>
<td>4,746,824</td>
<td>135,977,632</td>
</tr>
<tr>
<td>4</td>
<td>Swedish</td>
<td>Svenska</td>
<td>sv</td>
<td>1,618,424</td>
<td>3,610,936</td>
<td>26,566,975</td>
</tr>
<tr>
<td>5</td>
<td>French</td>
<td>Français</td>
<td>fr</td>
<td>1,502,818</td>
<td>6,470,166</td>
<td>105,861,687</td>
</tr>
<tr>
<td>6</td>
<td>Italian</td>
<td>Italiano</td>
<td>it</td>
<td>1,119,723</td>
<td>3,655,232</td>
<td>71,063,857</td>
</tr>
<tr>
<td>7</td>
<td>Russian</td>
<td>Русский</td>
<td>ru</td>
<td>1,110,690</td>
<td>3,739,694</td>
<td>75,205,855</td>
</tr>
<tr>
<td>8</td>
<td>Spanish</td>
<td>Español</td>
<td>es</td>
<td>1,098,911</td>
<td>4,559,462</td>
<td>79,701,638</td>
</tr>
<tr>
<td>9</td>
<td>Polish</td>
<td>Polski</td>
<td>pl</td>
<td>1,043,983</td>
<td>2,062,360</td>
<td>39,395,536</td>
</tr>
</tbody>
</table>
## Wikipedia today

<table>
<thead>
<tr>
<th>№</th>
<th>Language</th>
<th>Language (local)</th>
<th>Wiki</th>
<th>Articles</th>
<th>Total</th>
<th>Edits</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Waray-Waray</td>
<td>Winaray</td>
<td>war</td>
<td>959,771</td>
<td>1,980,590</td>
<td>4,867,190</td>
</tr>
<tr>
<td>11</td>
<td>Japanese</td>
<td>日本語</td>
<td>ja</td>
<td>907,711</td>
<td>2,539,048</td>
<td>52,609,371</td>
</tr>
<tr>
<td>12</td>
<td>Cebuano</td>
<td>Sinugboanong Binisaya</td>
<td>ceb</td>
<td>892,119</td>
<td>1,878,479</td>
<td>4,537,168</td>
</tr>
<tr>
<td>13</td>
<td>Vietnamese</td>
<td>Tiếng Việt</td>
<td>vi</td>
<td>888,199</td>
<td>2,298,929</td>
<td>16,384,752</td>
</tr>
<tr>
<td>14</td>
<td>Portuguese</td>
<td>Português</td>
<td>pt</td>
<td>826,676</td>
<td>3,509,946</td>
<td>39,885,533</td>
</tr>
<tr>
<td>15</td>
<td>Chinese</td>
<td>中文</td>
<td>zh</td>
<td>766,616</td>
<td>3,394,404</td>
<td>32,598,519</td>
</tr>
<tr>
<td>16</td>
<td>Ukrainian</td>
<td>Українська</td>
<td>uk</td>
<td>498,099</td>
<td>1,468,478</td>
<td>14,433,374</td>
</tr>
<tr>
<td>17</td>
<td>Catalan</td>
<td>Català</td>
<td>ca</td>
<td>427,345</td>
<td>1,066,616</td>
<td>13,707,545</td>
</tr>
<tr>
<td>18</td>
<td>Norwegian (Bokmål)</td>
<td>Norsk (Bokmål)</td>
<td>no</td>
<td>418,789</td>
<td>989,380</td>
<td>14,188,631</td>
</tr>
<tr>
<td>19</td>
<td>Persian</td>
<td>فارسی</td>
<td>fa</td>
<td>388,658</td>
<td>2,222,566</td>
<td>17,305,842</td>
</tr>
<tr>
<td>20</td>
<td>Indonesian</td>
<td>Bahasa Indonesia</td>
<td>id</td>
<td>351,399</td>
<td>1,366,544</td>
<td>8,890,468</td>
</tr>
<tr>
<td>21</td>
<td>Finnish</td>
<td>Suomi</td>
<td>fi</td>
<td>346,202</td>
<td>927,727</td>
<td>14,884,230</td>
</tr>
<tr>
<td>22</td>
<td>Czech</td>
<td>Čeština</td>
<td>cs</td>
<td>294,343</td>
<td>759,580</td>
<td>11,825,690</td>
</tr>
<tr>
<td>23</td>
<td>Arabic</td>
<td>العربية</td>
<td>ar</td>
<td>278,250</td>
<td>1,661,756</td>
<td>15,731,477</td>
</tr>
<tr>
<td>24</td>
<td>Korean</td>
<td>한국어</td>
<td>ko</td>
<td>275,927</td>
<td>899,566</td>
<td>14,213,400</td>
</tr>
<tr>
<td>25</td>
<td>Malay</td>
<td>Bahasa Melayu</td>
<td>ms</td>
<td>260,364</td>
<td>682,778</td>
<td>3,780,100</td>
</tr>
</tbody>
</table>
Who are our users

User generated content (UGC) is related to level of online activity of language communities.

**Early Adopters:** Large languages with large active online communities generating lots of content e.g. Latin languages (English, German, Dutch, French)

**Next Generation:** Large languages with small online communities are generating very little content e.g. Indic languages, Right-to-left languages, CJK languages

**Long tail** languages with tiny but passionate communities starting with little content e.g Native American Indigenous languages, Newari
Growing Content Contributions

Content has to grow at the same pace as rich delivery platforms

Access to content has to be free and pervasive

Virtuous cycle of contributions

Rich language tools and language assets for end users

The tablet is the platform

Let a thousand language web applications grow for contributing and consuming content to Wikipedia and other websites
Challenges

Accessible and open content

High quality content

Broken user experience for multilingual users

Inadequate language tools for Web and Mobile

Lack of reference data corpora

Growing contributor communities
What is Wikipedia doing

Language Selection
Universal Language Selector to set language preferences for display UI, fonts, input tools
Smart language selector search handling multiple scripts

Web Fonts
High-quality web fonts for 63 languages with 81 variants

Input Tools
Easy-to-use 139 input methods for 64 languages
Onscreen keymaps

Internationalization
Javascript and PHP i18n support for grammar, plurals, gender

Content Translation
Content translation platform integrated in Wikipedia, side-by-side translation editor
Machine translation, Translation Memories, Dictionaries, Glossaries, Wikidata

Software UI and Message Localization
Translation platform, side-by-side proofreading editor with translation aids
Crowdsourced web platform – translatewiki.net

Wikipedia Zero – Access for All
Where we are headed

As Wikipedia turns 14, it has become the most significant open content platform of this century.

Content Commons for the Web
Generate rich high-quality user content
Deliver first-class multilingual user experience
Engage new generation of users
Be mobile, be everywhere
Commoditize language software
Keep the Web Open and Free
Collaborate to make the Web Multilingual

Empower web and mobile platforms with language tools
Create and release high quality language fonts and typing tools
Collaborate with us to develop open tools, platforms and communities

Enable free access to content

Grow content communities
Build mobile web application developer ecosystem
Seed language applications to enable content contributions

Report problems you encounter when you’re reading or editing Wikipedia in your language
Thank you!

May 7: 10:45-11:15
Best Practices on the Design of Translation
Pau Giner, David Chan and Santhosh Thottingal

May 8: 14:15-15:00
Panel 1: Using Wikipedia for multilingual web content analytics across 287 languages.

May 8: 15:30-16:15
Panel 2: Growing Wikipedia editing with intelligent multi-language suggestion lists for article translation as well as other techniques and tools.