Why standardize

Relevant
Measure quantities that describe important phenomena

Replicable
Make analyses easily reproducible and verifiable

Explicitly defined
Provide formal definitions, remove ambiguity

Consistent
Limit the proliferation of proprietary definitions; compare apples and apples

Easy to maintain
(ideally) replicable via multiple data sources at any point in time

Granular
Computable at different time scales
What to standardize

Participation metrics
- Newly registered user
- Active editor
- Edit session

Readership metrics
- Page view
- Unique visitor
- Visit
Use case: editor engagement vital signs

Standardize metrics used in editor engagement research

<table>
<thead>
<tr>
<th>New users</th>
<th>Community</th>
<th>Content creation</th>
<th>Content curation</th>
</tr>
</thead>
<tbody>
<tr>
<td>New users</td>
<td><strong>Active editors</strong> *</td>
<td>Edits</td>
<td>Deletions</td>
</tr>
<tr>
<td>Live new users</td>
<td>Very active editors</td>
<td>Uploads</td>
<td>Reverts</td>
</tr>
<tr>
<td><strong>New editors</strong> *</td>
<td>Page creators</td>
<td>Page creations</td>
<td></td>
</tr>
<tr>
<td>New page creators</td>
<td>Media uploaders</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Productive new users</td>
<td>IP editors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surviving new users *</td>
<td>Bots</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[https://www.mediawiki.org/wiki/Analytics/Epics/Editor_Engagement_Vital_Signs](https://www.mediawiki.org/wiki/Analytics/Epics/Editor_Engagement_Vital_Signs)
Newly registered user (acquisition)

- Genuine new registrations only
- No attached users
- No proxy registrations
- New user class

https://meta.wikimedia.org/wiki/Research:Newly_registered_user
Content edits in a timespan since registration
• Threshold-based metric: \( n \) content edits in \([T, T + t]\)
• Default: 1 ns0 edit in 24 hours
• Arbitrary thresholds per period
• Replace *New Wikipedian*

Productive new user (quality)

- Productive content edits in a timespan since registration
- Threshold-based metric: \( n \) content edits unreverted in time \( q \) in \([ T, T + t ]\)
- Default: 1 ns0 edit unreverted in 48 hours in 7 days
- New user class

https://meta.wikimedia.org/wiki/Research:Productive_new_user
Surviving new user (retention)

- Content edits after a survival period since registration and within a sunset period
- Threshold-based metric: \( n \) content edits in \([ T + t , T + t + s ]\)
- Default: 1 ns0 edit in 12\(^{th}\) month since registration
- Arbitrary thresholds per period
- Replace *Retained new Wikipedian*

https://meta.wikimedia.org/wiki/Research:Surviving_new_user
Active editor

- $n$ content edits in $\Delta T$
- Default: 5 no edits in 1 month
- Arbitrary thresholds per period (daily, weekly, monthly)
- Replace *Monthly active editor*

[Link to research page](https://meta.wikimedia.org/wiki/Research:Refining_the_definition_of_monthly_active_editors)
Guiding principles and open issues

robust vs responsive to configuration changes

- Project changes definition of a “countable” page
- Project creates new content namespace

stateless vs stateful

- Historical values change due to deletion, redaction, content redistribution
- Censored observations are flagged / excluded

easy to maintain vs accurate

- How do we identify a bot?

focus on conventional vs diagnostic parameters
Supportive analysis: new user activation

Supportive analysis: new user retention

https://meta.wikimedia.org/wiki/Research:Surviving_new_user
Supportive analysis: editor activity

https://meta.wikimedia.org/wiki/Research:Refining_the_definition_of_monthly_active_editors
Next steps

- fine-tune definitions + sample queries
  - community feedback
    - prototype dashboards
      - productization
Credits

Apples and Oranges • CC BY photo by Michael Johnson
https://flickr.com/photos/26176646@N04/2492945625/