Knowledge equity: Addressing Knowledge Gaps

Taxonomy of Knowledge Gaps

“Flywheel” model for engagement

Learn more: https://w.wiki/deC
Research’ Readership Roadmap

1: Who and why?
Surveys on demographics and motivations of readers (FY19,20)

2: Navigation paths
How are readers navigating content (articles, citations, images, etc)?
(FY20, FY21)

3: Learning
How do readers learn on Wikipedia? What makes readers curious/inquisitive?
(FY22+)

We are HERE
Observing reader navigation from logs

- **Scale**: [Webrequest](#) data from one month (enwiki 2021-03)
  - 6.25B pageloads: >2000 pageloads/sec vs ~1.5 edits/sec

- **Complexity**: Reconstructing paths from individual pageloads
  - 1.47B unique readers from “fingerprinting” IP+`user_agent` (no cookies)
#1 We are able to learn about readers’ needs

- Navigation is short
  - ~73% with only 1 pageloads
  - still >10M sessions with 10+ pageloads
- Navigation is usually fast
  - median 74 s time between pageloads
- Navigation stops at low-quality articles
  - short text, few links, etc

Article maintenance and improving quality is important
#2 Reader interactions are diverse

- Mobile vs Desktop
  - Mobile: Spike in no. of sessions during evening
  - Desktop: more pageloads per session (2.4 vs 1.99)

- Topics
  - Length: longer (entertainment) vs shorter (STEM)
  - Strategy: breadth (entertainment) vs depth (STEM)

Illustrated by Jasmina El Bouamraoui and Karabo Poppy Moletsane, CC0

There are no one-size-fits-all solutions to address readers’ needs
#3 External Search is crucial for readers’ navigation

- ~75% of reading sessions start from external search
- Using search to navigate between pages
  - 40.1% of pairs of consecutive pageloads
  - in 30% of these cases internal link available

How can we improve Wikimedia’s internal search?
#4 Images are important for readers

- Engagement with images is high
  - Click-through rate: ~1/30 (citations: ~1/300)
- Images complement missing information
  - Higher engagement with images in shorter articles
  - Lower click-through rate for page previews with images
- Types of images with high engagement
  - Geographic locations (monuments), biology, complex objects (visual arts, etc), unfamiliar (!) faces

Improve infrastructure and tools for multimedia content
#5 Our public data is extremely valuable

- Research on navigation limited
  - **Webrequest data** (private): complete navigation logs
  - **Clickstream data** (public): aggregate counts how often a link from a source-to-target article was clicked

- Navigation in *webrequest* can be sufficiently approximated by the *clickstream* dataset
  - Differences are statistically significant but have a small effect size (<10%) for many research types

Build public resources to empower community of researchers

More examples: Search-referral data ([link](https://wikinav.toolforge.org/)), User-script for in-situ visualization of WikiNav ([link](https://wikinav.toolforge.org/)), top-read articles by country from DataEng ([link](https://wikinav.toolforge.org/)), Differential Privacy project from Privacy ([link](https://wikinav.toolforge.org/))
Future: How are readers learning?

The role of images

How are readers curious?

Dani Bassett+team
(March 2021 showcase https://w.wiki/4y$S)
Want to learn more

● If you have questions, please reach out: #research-and-data, mgerlach@

● Read more about our projects on readers
  ○ Research report: https://research.wikimedia.org/report.html
  ○ Knowledge gaps taxonomy https://w.wiki/deC
  ○ Reader surveys and demographics https://w.wiki/4uho
  ○ Reader engagement with citations https://w.wiki/4uhq
  ○ Reader engagement with images https://w.wiki/4uhu
  ○ Reader navigation paths https://w.wiki/4uhv
  ○ Learning with images https://w.wiki/4r38
  ○ Curiosity of readers https://w.wiki/4r39

● Thanks: Data Engineering, Privacy, Legal, External collaborators