A Taxonomy of Knowledge Gaps for Wikimedia Projects (Summary and Motivation)

If you are looking for a shorter (than the full paper) write-up to have an overview of our thoughts on how to approach measuring knowledge gaps, please read on.

What is the problem?
Wikimedia projects aim to empower people from around the world to share in the sum of all knowledge. Knowledge gaps are a barrier to this goal. Today we do not have a shared definition and understanding of knowledge gaps on the Wikimedia projects. As a result, we do not have a way to measure the extent of knowledge gaps in the Wikimedia projects. We do not know how to measure certain types of gaps. We do not have established methods to understand the relationship between the different types of gaps.

Why is it important?
Our decisions can suffer from the risk of being sub-optimal as we do not see the full spectrum of knowledge gaps. As a result, we tend to invest in more widely discussed or understood types of knowledge gaps. We also have very limited ways to measure progress towards addressing specific gaps across many years. This makes learning and course-correction harder.

What is our solution?
We want to develop a Knowledge Gap Index that is a composite index which will measure the state of knowledge gaps across many Wikimedia projects.[1] The first step to do that is to build a taxonomy of knowledge gaps which gives us a shared framework for understanding the different types of knowledge gaps. Such a taxonomy also specifies what each gap type is (and is not). So far, we have built a taxonomy of knowledge gaps for the Wikimedia projects and we are seeking your input to improve it.

What do we mean by knowledge gaps?
Before this research, the phrase knowledge gaps equated to content gaps for many. We want to change this by expanding the definition of knowledge gaps to include gaps in contributorship and readership on top of content. For those of you interested in a formal definition of knowledge gaps, here is what we propose:

Knowledge gaps are major differences in participation or coverage of a specific group of readers, contributors, or content.

Why has developing the framework for understanding knowledge gaps been challenging?
Developing a framework for understanding and measuring knowledge gaps is hard because:

- We need to make judgement calls about what gaps to include and which ones to not include. We also need to decide how specific the gaps will need to be.
- The measurement of certain gap types can be complex due to technical reasons (how should we gather the data for it?) or privacy reasons. They can also be complex research topics that we know very little about from the methodological perspective.

[1] If this is the first time you are hearing about composite indices, don't be discouraged by the fancy name of it. A composite index captures a variety of indices that aim to measure specific dimensions of a problem and their relationship to the overall problem. For example, the gender equality index is a composite index that measures the state of gender equality considering work, money, power and other types of indices.

1 READERS

The readership dimension of knowledge gaps encompasses all those gaps related to readers' access to Wikimedia sites. We define readers as all users who connect directly to the projects to access Wikimedia content excluding how content consumption happens outside of Wikimedia, e.g., voice assistants, search engines, or third-party apps.

Facet	*Gap*	*Description*
Sociodemographics *Objective*: readers with different social status, demographics, and cultural background can easily and safely accessing free knowledge	*Gender*	Difference between readers of different gender identities in how and how much they access the sites.
	Age	Difference between readers of different age in how and how much they access the sites.
	Locale	Differences in readership between rural areas, towns, and cities
	Language	Differences in readership depending on readers' ability to read one or more languages
	Income	Difference on how readers with different income, wealth, or employment status access Wikimedia sites
	Education	Differences in readership depending on readers' educational background
	Background	Differences in readership among people with different cultural, political and sexual preferences
Information Need *Objective*: readers with different information needs can find and consume free knowledge	*Motivation*	Differences in readership depending on the reason behind readers' visit to the site
	Information Depth	Differences in readership depending on the depth of information for which a reader is looking
	Familiarity	Differences in readership depending on one's prior familiarity with a topic
Accessibility *Objective*: readers with different technical setup and skills can easily access Wikimedia projects	*Internet connectivity*	Contrasts among the ability of readers with different internet connections to access Wikimedia sites
	Device	Difference in accessibility to the site depending on readers' devices
	Tech Skills	Differences in readership depending on readers' general internet skill
	Disabilities	Disparities in ability to access the knowledge within Wikipedia depending on individual disabilities

2 CONTRIBUTORS

The contributor dimension of knowledge gaps covers all gaps related to categories of people contributing to Wikimedia sites. We define contributors as all individuals who *edit* or otherwise *maintain* Wikimedia content. For the purpose of this taxonomy, this definition does not include technical contributors—i.e. the individuals who build the MediaWiki software on which Wikimedia sites run—though the software and choices made in its design certainly are highly impactful on what types of contributors feel supported and what content is created.

Facet	*Gap*	*Description*
Sociodemographics *Objective*: contributors with different social status, demographics, and cultural background can easily and safely access and contribute to free knowledge	*Gender*	Differences between contributors of different gender identities in how and how much they contribute to the sites.
	Age	Differences between contributors of different ages in how and how much they contribute to the sites.
	Locale	Differences between contributors of different locales (urban, rural) in how and how much they contribute to the sites.
	Language	Differences between contributors of different reading abilities in a language in how and how much they contribute to the sites.
	Income	Differences between contributors with different income, wealth, or employment status in how and how much they contribute to the sites.
	Education	Differences between contributors of different educational backgrounds in how and how much they contribute to the sites.
Contextual *Objective*: contributors with different motivations and roles can access and contribute to free knowledge	*Motivation*	Differences in contribution depending on one's reason for contributing to the site.
	Role	Differences in contribution depending on the type of editing that one chooses to do.
Accessibility *Objective*: contributors with different technical resources and abilities can easily access and contribute to Wikimedia projects	*Internet connectivity*	Disparities in ability to contribute to the knowledge within Wikipedia depending on one's access to high-speed internet
	Device	Disparities in ability to contribute to the knowledge within Wikipedia depending on one's device.
	Tech Skills	Disparities in ability to contribute to the knowledge within Wikipedia depending on one's general internet skills
	Disabilities	Disparities in ability to contribute to the knowledge within Wikipedia depending on individual disabilities

3 CONTENT

Wikipedia is incomplete by design. The opportunity to share new information with the world is a major motivating factor among both new and established Wikipedia contributors. However, when important information about a topic is absent, incomplete, biased, or otherwise inaccessible to readers, these content gaps can undermine Wikipedia's ability to serve the needs of its global audience.

We characterize *gaps in content coverage* as follows.

Facet	***Gap***	***Description***
Policy *Objective*: content is consistent with core content policies	*Verifiability*	Differences in the use of reliable sources in order to verify content.
	Neutrality	Biases in the content across Wikipedia articles .
Accessibility *Objective*: content is accessible to different audiences	*Multimedia*	Differences in coverage with respect to the type of media used to share the content
	Structured Data	Differences in the use of information which is indexed and machine-readable
	Readability	Barriers for accessing or consuming information originating from content
Diversity *Objective*: content covers knowledge that is underrepresented, marginalized, and locally relevant	*Gender*	Differences in content coverage depending on the gender identity of subjects
	Geography	Differences in coverage of topics related to geographic regions or population distribution
	Impactful topics	Differences in coverage of topics that are of common interest
	Cultural context topics	Differences in coverage of topics related to the history, heritage, and characteristics of a current or former cultural group