Brainstorming

Language Gap metrics brainstorming with team

Phab task: https://phabricator.wikimedia.org/T376728

Notes from 2024-11-26

Miriam

- Should the schema follow the existing content gap dataset schemas? See below.
- Rows should be the same as previous schemas
    - i.e., each row should each be one language
- Focus on Wikipedias (for now).
    - Think abbott how to Include hosted *and* incubating and closed Wikipedias
- Think about multiple ways to describe the gap.
    - E.g., gender gap dataset has multiple variables for looking at the gap
- Talk to Fabian about the feasibility of these new pipelines and what's possible
    - Sync with Movement Insights

Notes from 2024-12-02

Martin

- Different from the other gaps because language gaps are across wikis as opposed to within wikis
- The buckets will just be redundant to the wikidb column
- Then use Content#Metrics_for_Aggregation (at the level of the language edition)
- Maybe higher-level beyond this, but it will need to be a different schema
- Long-term: Ethnologue data, metrics related to global/regional coverage
- Think about: Wiki Workshop 2-pager (call to come)
- Feel free to add to phab task

Notes from 2024-12-05

Yu-Ming

- Three different ways to conceptualize the metric(s):
    1. Thinking about what discussed with Martin
    2. Thinking about *if I have* Ethnologue data (for proposing metrics, don't yet need to have the data, but can go ahead and propose the metrics and outline the intended valuations, etc.)
    3. LANGUAGE: the language as the content (content ABOUT the language)

Notes from 2024-12-06

Tanja

- (Very meta) Metrics for coverage of information *about languages* across the wikis
    - coverage and quality of English_language on enwiki, frwiki, chrwiki, etc etc
    - coverage and quality of French_language on enwiki, frwiki, chrwiki, etc etc
    - coverage and quality of Cherokee_language on enwiki, frwiki, chrwiki, etc etc
    - Etc.
- (Very in-line with current schema) For each language edition of wikipedia…
    - number of articles per speaker
    - number of quality articles per speaker
    - articles that 'should' exist (in science, in x y z)

    - articles related to our 'topics for impact': e.g. women's health, climate
    - coverage of Vital Articles
- Ownership of the metrics?
    - Thinking about where it would live and where it would get used more

Notes from 2024-12-16–19

Isaac, async

- What the language gap metric(s) could/should look like?
    - I think the Language Gap in some ways was our least-well-defined Gap in the sense that it didn't fit the mold of the other gaps but also felt like it had to be addressed somewhere. With that in mind, don't be overly-constrained by what was stated in the Knowledge Gaps Taxonomy. For me, it's hard to think about the Content Language Gap without also thinking about the Important Topics gap. Just the presence of a language edition or content itself feels like a weak indicator of "coverage" of a language. Both the topic of the article and quality feel especially important. When I think about something like the gender gap, quality/topic are important and we seek to capture that but even just basic representation is a strong start.
    - As far as topic, Vital Articles and maybe eventually more is a strong start for the globally-important information. These are often areas where the content is relatively "static" in the sense that they aren't breaking news etc. and so it's reasonable for even a small language community to put the effort into creation without incurring a large maintenance overhead. The harder part is the local knowledge – things that likely won't be written about in other languages but are important to the speakers of that language.
    - We have good quality metrics but they are more focused on article structure and don't take into account the quality of the writing, which feels especially relevant for the Language gap. I wonder if this wouldn't be a good place to capture things like whether the article was initially a translation or bot-generated, how many different editors worked on the content (with there being value to greater diversity as that suggests the quality of writing is likely to be higher). Eventually
- What the language gap metric(s) should look like?
    - Right now, language access is essentially a binary in the sense that an article either exists locally or it does not. We know that readers use Google Translate etc. quite frequently from what little data we do have and LPL has been exploring MinT for readers, so let's make sure that whatever framework we choose can extend to at least: article exists, article is translated in, article translation not readily available.
    - We eventually would love to have a sense of what "unique" knowledge a given project is bringing to Wikipedia. A simple way of doing this is to look at whether an article has sitelinks to other language editions but that is prone to errors for smaller languages (missing sitelinks) and penalizes languages when a translation is created of their article. A bit more nuance might be tracking the order of creation for articles (similar to Diego's work). Even more nuance might go beyond the presence of an article and look at what sources are used in it and whether they are unique. This last piece is a good bit more difficult but to me is the more interesting aspect to eventually consider.
- Considerations
    - Think about the AI strategy and how to approach smaller languages
    - Think about what are the goals of the small wikis, and how are those going to be different from the larger wikis
    - How can we help wikis get to mid level stage

Proposals

Proposed language gap[1] metrics

1. Content availability (i.e. language representation)

Goal: Dataset helps track which languages have which Wikimedia projects, as well as level of representation via third-party data. Similar to canonical wikis dataset, but also takes into account test projects (e.g. Incubator projects). Overall, it also gives affiliates and user groups basic third-party information about languages, so that movement actors can also direct their work.

1.a. Schema based on one row per language.

Field name	Data type	Comment	Source (if third-party)
language_code	char	Language code (e.g., en)	-
language_name	char	Language name (e.g., English)	-
“hosted_since” or "start_date"	date	For hosted projects, date that the project edition first became hosted (i.e., graduated from Incubator or first edits having own domain and database)	-
“language_population” or something like “tech_literate_language_population”	integ	Approximate figures for the literate, functional population for each language in each territory: that is, the population that is able to read and write each language, and is comfortable enough to use it with computers — based on Unicode’s Territory-Language Information	Unicode CLDR
language_official	array	List of countries (country codes) where the language has official status — based on Unicode’s Territory-Language Information	Unicode CLDR
web_support?	char?	Website and mobile app support (level)?	STIL 2020?
unesco_status	char	The language’s endangerment status per UNESCO: ● (Ex) Extinct ● (CR) Critically endangered ● (SE) Severely endangered ● (DE) Definitely endangered ● (VU) Vulnerable ● (NE) Not endangered/ Safe	UNESCO (tbd)

[1] Specifically, language gaps within *content gaps*. Language gap metrics already exist for readership and contributor gaps.

language_is_indigenous	bool	Whether or not the language is an indigenous language	UNESCO (tbd)
wp_status	char	The Wikipedia status of the language: ● Hosted (Wikipedia edition in this language is hosted by the Foundation) ● Closed (Wikipedia edition in this language was previously hosted but is now closed) ● Test (Wikipedia edition in the language exists in Wikimedia Incubator)	-
ws_status	char	The Wikisource status of the language: ● Hosted (Wikisource edition in this language is hosted by the Foundation) ● Closed (Wikisource edition in this language was previously hosted but is now closed) ● Test (Wikisource edition in the language exists in Multilingual Wikisource)	-
wv_status	char	The Wikiversity status of the language ● Hosted (Wikiversity edition in this language is hosted by the Foundation) ● Closed (Wikiversity edition in this language was previously hosted but is now closed) ● Test (Wikipedia edition in the language exists in Wikiversity Beta)	-
wb_status	char	The Wikibooks status of the language: ● Hosted ● Closed ● Test	-
wq_status	char	The Wikiquote status of the language ● Hosted ● Closed ● Test	-
wn_status	char	The Wikinews status of the language ● Hosted ● Closed ● Test	-
wt_status	char	The Wiktionary status of the language ● Hosted ● Closed ● Test	-
wy_status	char	The Wikivoyage status of the language ● Hosted ● Closed ● Test	-

1.b. Same schema as above, except one row per wiki project edition, to align with the canonical wikis schema.

Field name	Data type	Comment	Source (if third-party)
wiki_db	char	Wikimedia database name (e.g., "enwiki")	
language_code	char	Language code (e.g., en)	-
language_name	char	Language name (e.g., English)	-
"language_population" or something like "tech_literate_language_population"	integ	Approximate figures for the literate, functional population for each language in each territory: that is, the population that is able to read and write each language, and is comfortable enough to use it with computers – based on Unicode's Territory-Language Information	Unicode
macroarea	char	Macro-area of the language's speakers/signers	Glottolog
etc	etc	etc	etc

2. Content coverage (i.e. vital article coverage)

Goal: Dataset helps track coverage of 1000 articles every Wikipedia should have in each Wikipedia language edition, including metrics (e.g. quality) for those articles. Follows Knowledge Gaps - Content Gaps schema.

2.a: Schema based on Research team's content gap metric dataset schema.

Field name	Data type	Comment
wiki_db	char	Wikimedia database name (e.g., "enwiki")
time_bucket	date	the time bucket, with monthly granularity (e.g. "2020-02")
content_gap	char	the content gap (e.g., "topic-gap")
category	char	the underlying categories for the gap; there will only be one category in this schema: it will be called "vital-articles" or "1000-articles-every-wp-should-have"
articles_created	integ	Number of articles (from the list of 1000) which have been created, at the time of the time bucket
pageviews_sum	integ	total number of pageviews for the vital articles that Wikipedia has
pageviews_mean	integ	mean number of pageviews for the vital articles that Wikipedia has, at the time of the time bucket
revision_count	integ	total number of edits for the vital articles that Wikipedia has, at the time of the time bucket
quality_score	integ	average article quality score for the vital articles that Wikipedia has, at the time of the time bucket
standard_quality	integ	percentage of vital articles that Wikipedia has (at the time of the time bucket) that satisfy the Standard Quality Criteria
standard_quality_count	integ	number of vital articles that Wikipedia has (at the time of the time bucket) that satisfy the Standard Quality Criteria

2.b: Same underlying data as in schema above, but at the level of the wikidata item associated with each article (i.e., the content_gap field will have the 1000 QIDs for the 1000 items associated with each of the 1000 articles every Wikipedia should have).

Field name	Data type	Comment
wiki_db	char	Wikimedia database name (e.g., “enwiki”)
time_bucket	date	the time bucket, with monthly granularity (e.g. “2020-02”)
content_gap	char	the content gap (e.g., “topic-gap”)
category	char	the underlying categories for the gap; there will be 1000 categories (one QID for each item associated with an article every wikipedia should have)
articles_created	integ	Number of articles from the category which have been created, at the time of the time bucket. Since each category only contains individual languages, • “1” will indicate that the associated article has been created in that wiki, at the time of the time bucket • “0” will indicate that the associated article has *not* been created in that wiki, at the time of the time bucket
pageviews_sum	integ	total number of pageviews for each category (i.e. total number of pageviews for each article associated with each language article) at the time of the time bucket
revision_count	integ	total number of edits for each category (i.e. total number of edits for each article associated with each language article) at the time of the time bucket
quality_score	integ	article quality score for each category (i.e. article quality score for each article associated with each language article) at the time of the time bucket
standard_quality_count	integ	number of language articles that Wikipedia has (at the time of the time bucket) that satisfy the Standard Quality Criteria. Since each category only contains individual language article, • “1” will indicate that the associated article satisfies the Standard Quality Criteria. • “0” will indicate that the associated article *does not* satisfy the Standard Quality Criteria.

3. Coverage of articles about languages

Goal: This dataset helps track articles about languages Wikipedia editions, including metrics (e.g. quality) for those articles. Follows Knowledge Gaps - Content Gaps schema.

Field name	**Data type**	**Comment**
wiki_db	char	Wikimedia database name (e.g., enwiki)
time_bucket	date	the time bucket, with monthly granularity (e.g. “2020-02”)
content_gap	char	the content gap (e.g., “language-gap”)
category	char	Language name of the language article (e.g., Tagalog)
is_language_of_wiki_db	boolean	Whether or not the language article is the language of the Wikipedia
articles_created	integ	Number of articles from the category which have been created, at the time of the time bucket. Since each category only contains individual QIDs, ● “1” will indicate that the associated article has been created in that wiki, at the time of the time bucket ● “0” will indicate that the associated article has *not* been created in that wiki, at the time of the time bucket
pageviews_sum	integ	total number of pageviews for each category (i.e. total number of pageviews for each article associated with each QID) at the time of the time bucket
revision_count	integ	total number of edits for each category (i.e. total number of edits for each article associated with each QID) at the time of the time bucket
quality_score	integ	article quality score for each category (i.e. article quality score for each article associated with each QID) at the time of the time bucket
standard_quality_count	integ	number of vital articles that Wikipedia has (at the time of the time bucket) that satisfy the Standard Quality Criteria. Since each category only contains individual QIDs, ● “1” will indicate that the associated article satisfies the Standard Quality Criteria. ● “0” will indicate that the associated article *does not* satisfy the Standard Quality Criteria.