Interpolating Quality Dynamics in Wikipedia and Demonstrating the Keilana Effect

Aaron Halfaker
Wikimedia Research
San Francisco, California, USA
ahalfaker@wikimedia.org

ABSTRACT

For open, volunteer generated content like Wikipedia, quality is a prominent concern. To measure Wikipedia's quality, researchers have historically relied on expert evaluation or assessments of article quality by Wikipedians themselves. While both of these methods have proven effective for answering many questions about Wikipedia's quality and processes, they are both problematic: expert evaluation is expensive and Wikipedian quality assessments are sporadic and unpredictable. Studies that explore Wikipedia's quality level or the processes that result in quality improvements have only examined small snapshots of Wikipedia and often rely on complex propensity models to deal with the unpredictable nature of Wikipedians' own assessments. In this paper, I describe a method for measuring article quality in Wikipedia historically and at a finer granularity than was previously possible. I use this method to demonstrate an important coverage dynamic in Wikipedia (specifically, articles about women scientists) and offer this method, dataset, and open API to the research community studying Wikipedia quality dynamics.

ACM Classification Keywords

C.4 PERFORMANCE OF SYSTEMS: Measurement techniques, Modeling techniques

Author Keywords

Wikipedia; Quality; Modeling; Predictive; Interpolation; Methods; Dataset

INTRODUCTION

Since its inception, quality has been the most prominent concern with regards to the future of Wikipedia. After all, how can high quality information artifacts be produced when there's literally no restriction on who is allowed to contribute? Over the past 12 years (as of 2017), the research literature around Wikipedia has advanced our understanding of the open encyclopedia's quality and the processes by which crowds of volunteers can manage such an information artifact.

Our first major leaps in understanding of Wikipedia's quality dynamics happened around the time that Jim Giles published a report in *Nature* (2005)[7] that surprised the world. This seminal report showed that Wikipedia's coverage of scientific content compared favorably (and in some ways, better) than dominant, traditional, print-based encyclopedias. Since that surprising result was published, researchers have been pushing toward greater understanding of how open, volunteer processes could have generated such a high quality information resource.

While we do know a lot about quality dynamics in Wikipedia, there are still many questions that remain. Where are Wikipedia's coverage gaps? What types of editing patterns lead to efficient quality improvements? These questions are important for the science and the practices of Wikipedians–the volunteers who write and curate the encyclopedia's content. In this paper, I detail the development of a measurement strategy and the release of a public dataset that I believe will make answering these questions far easier than ever before. In the following sections, I'll summarize the state of the art with regards to quality dynamics and measurement in Wikipedia, I'll explain my measurement methodology, and I'll provide a demonstration analysis that gives novel insights into the coverage gaps and quality dynamics of articles about women scientists.

The quality of English Wikipedia

Giles' study set in motion a series of studies examining the quality of several different subject spaces in the encyclopedia. Mesgari et al. provides an excellent survey of this work[15], but for the purposes of this paper, I'll summarize their key findings as follows:

- Wikipedia's coverage is broad and comprehensive.
- Wikipedia has a high level of currency – especially with regards to topics of interest to the public.
- Wikipedia's accuracy compares favorably with traditional encyclopedias.

Yet the story isn't all one of pure success. Given the conclusions drawn from years of research, Wikipedia's coverage of

OpenSym '17 August 23–25, 2017, Galway, Ireland

ACM ISBN 978-1-4503-5187-4/17/08.
DOI: https://doi.org/10.1145/3125433.3125475

most major topics that are covered in traditionally encyclopedias is hard to question. However, not all topics are covered as completely as others. Gaps in Wikipedia's content coverage are concerning because of the encyclopedia's dominance as an information resource. 600 million people read the encyclopedia every month[1]. With an information source of such popular use and apparent completeness, any topic that is covered less completely than others (or not at all) could imply less importance or relevance to certain types of knowledges. Based on past work, we know that Wikipedia lacks in coverage of topics about Women [17], of interest to Women [14], about rural and non-Western geographies [11, 10] and cultures [8]. Thus, understanding where Wikipedia's coverage is lacking and which initiatives are effective in closing coverage gaps is critical to the long term success of the project and is a concern for the preservation of *all* human knowledge.

Quality dynamics in Wikipedia

But how did Wikipedia arrive at such a high quality level? There are two major threads of research in this area: counter-vandalism and article quality dynamics.

Counter-vandalism. Due to it's open nature, Wikipedia is under constant threat of vandalism and other types of damaging changes. At it's most basic level, Wikipedia protects against damage by maintaining a history of all versions of every article. This allows "patrollers" to easily clean up damage whenever it is discovered. A *revert*[2] is a special type of edit that removes the changes of another edit – usually by restoring the last good version of the article.

But reverts are not the whole story of counter-vandalism in Wikipedia. There is complex network of communities of practices (like the conter-vandalism unit[3]), policies (like the 3-revert-rule[4]), and highly advanced automated tools that support editors in finding and quickly removing damaging contributions to Wikipedia[6][16]. Together this socio-technical system fills the infrastructural role of keeping out the bad stuff – ensuring that Wikipedia's open nature does not cause quality to decay into nonsense.

Article quality dynamics. While counter-vandalism and other types of patrolling work helps keep the bad out of the Wiki, there are other dynamics and processes at play that determine which articles will increase in quality efficiently. Stvilia et al. first hypothesized a framework for how high quality content was generated in Wikipedia[19]. Essentially, volunteers will allow their interests to drive where they contribute, and through building on to each others' work, articles grow and are gradually refined. This interest-driven pattern could likely explain how Wikipedia's demographic gaps have lead to coverage and quality gaps[14].

[1] https://tools.wmflabs.org/siteviews/?platform=all-sites&source=unique-devices&start=2016-04&end=2017-03&sites=en.wikipedia.org

[2] https://meta.wikimedia.org/wiki/Research:Revert

[3] https://en.wikipedia.org/wiki/Wikipedia:CVU

[4] https://en.wikipedia.org/wiki/Wikipedia:3RR

Table 1: Wikipedia 1.0 (wp10) assessment rating scale. The 6 levels of assessment are provided with the "readers experience" description copied from Wikipedia[5]. Each level also has a well-defined, detailed set of assessment criteria that involves discussion of formatting, coverage, and proper sourcing.

	summary
FA	Professional, outstanding, and thorough; a definitive source for encyclopedic information.
GA	Useful to nearly all readers, with no obvious problems; approaching (but not equalling) the quality of a professional encyclopedia.
B	Readers are not left wanting, although the content may not be complete enough to satisfy a serious student or researcher.
C	Useful to a casual reader, but would not provide a complete picture for even a moderately detailed study.
Start	Provides some meaningful content, but most readers will need more.
Stub	Provides very little meaningful content; may be little more than a dictionary definition.

Through Kittur et al.'s work, we know that group structure and dynamics play an important role in the efficiency of article improvement. They showed that articles where few editors make most of the edits (and the vast majority of editors contribute very little individually) tend to increase in quality more quickly than articles where the distribution of edits per editor is more uniform[13]. Arazy & Nov extend this conclusion by showing that certain types of editor experience are critical to article improvement – that it is not only important to have diversity in contribution rates but also diversity in the total experience level of Wikipedia editors[3]. In order to draw these conclusions, the authors of both of these studies needed to operationalize measurements of quality in Wikipedia and compare the configurations of editors and their contribution types to changes in measured quality.

Methods for measuring article quality

Wikipedia assessment ratings. Wikipedians assign quality assessments to articles based on a scale that was originally developed to produce an official "1.0" version of Wikipedia[6]. This scale has since been adopted by WikiProjects[7], self-organized subject-matter focused working groups on Wikipedia (*e.g.* WikiProject Video Games, WikiProject Medicine, and WikiProject Breakfast). Table 1 shows the rating scale with defunct old grades ("A", "B+", etc.) removed. Wikipedians use this scale to track progress towards content coverage goals and to build work lists (*e.g.* WikiProject Medicine's tasks[8]).

[6] https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team

[7] https://en.wikipedia.org/wiki/WikiProject

[8] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Tools#Tasks

Both Kittur et al.[13] and Arazy & Nov[3] used article quality assessments provided by Wikipedians to track article improvements and to find correlations with editor activity and experience level characteristics. They use these correlations to draw conclusions about successful collaboration patterns. Regretfully, the process by which Wikipedians assess and re-assess articles is unpredictable. We can be relatively sure of the quality level of an article at the time it was assessed, but there's no clear way to know when, exactly, the quality level of an article actually changed. The aforementioned studies use a set of complex propensity modeling strategies like Heckman Correction[9] to minimize the possibility that the correlations they observe were simply due to the *assessment* behavior of Wikipedians and not actual changes in article quality. The modeling of correction is difficult to optimize and evaluate independently of the effects that editor collaboration patterns may have had on those quality changes. Further, past analyses have been limited to the times at which Wikipedians were performing assessments of articles. Assessments didn't become common until 2006 (5 years after Wikipedia's inception) and the criteria by which articles are assessed has been undergoing changes since then. Articles that were "B" class in 2006 would likely now be classified as a "Start" class due to insufficient inline references. Table 2 shows the change in B-class criteria between 2006 and 2017.

Further, the fact that assessments are sparse and unpredictable also means that the assessment are often out of date. This is an operational issue for Wikipedians too. WikiProject groups organize re-assessment drives to bring the assessments of articles under their purview into compliance, but this is a never-ending process. Since Wikipedia contributions come from anyone (member of the WikiProject or not) and re-assessments are uncommon despite regular coordinated efforts, the overall assessments of Wikipedia articles are perpetually out of date and reassessments are a never-ending source of new work.

MODELING ACTIONABLE ARTICLE QUALITY

In order to enable better understanding of the dynamics of article quality in Wikipedia, I sought to develop a method that would allow for granular analysis of the quality level of articles at any point in time. Warncke-Wang et al.'s "actionable" modeling strategy seems particularly suitable to the task due to the high level of fitness they demonstrate (matching Wikipedians' own assessments) and its reliance on characteristics of the text content of the article (as opposed to external characteristics). I worked with Dr. Warncke-Wang to re-implement this model in the ORES system[12] and to implement a few minor improvements since he and his collaborators last published about the model [21]. Specifically, we added two features: the count of "[citation needed]" templates and the count of "Main article" linking templates. I then used the related open dataset [20] of assessments to train and test this model. For a full specification of the features used in prediction including scaling and controlling features, see the code [13]. ORES currently uses a GradientBoosting algorithm with estimators=700, max-depth=7, max-features=log2, and loss=deviance [14]. Table 3 presents the overall prediction fitness of the re-implemented model.


```
"prediction": "GA",
"probability": {
  "Stub": 0.0019,
  "Start": 0.0132,
  "C": 0.1252,
  "B": 0.2090,
  "GA": 0.3345,
  "FA": 0.3162
}
```


Figure 1: Quality prediction for Murie Curie. The article quality prediction of a revision of the article "Murie Curie" saved on April 4th, 2017 is presented. See https://ores.wikimedia.org/v2/scores/enwiki/wp10/773753742.

Modeling quality changes. I hypothesized that, at a large enough timescale, this "actionable model" would be able to track *quality dynamics* – the changes in article quality over time. After all, if the model is able to match Wikipedians' assessments, it should also be able to fill in the gaps between assessments as well. I suspected that a large timescale that covered several revisions would be necessary for measuring article quality dynamics accurately because it takes time for Wikipedia's counter-vandalism systems to work. *I.e.* vandalism and other damaging edits to an article might also change the features extracted favorably since they are arguably quite basic (*e.g. number of headers* and *number of image links* – see [22]). It would be inappropriate to consider such an edit to have increased the quality level of the article. By waiting a substantial time period between automatic quality predictions, Wikipedians' natural quality control processes can run its course.

Beyond the formal analysis of the prediction model (see **??**), there are two critical observations that *informally* support the conclusion that this strategy is working in practice: (1) I performed a set of basic spot-checks on hand-picked articles focusing on substantial quality changes and (2) I released this prediction model publicly 2 years ago and since then many users have told my collaborators and I that they find it to be effective and useful for supporting their work. I was actually quite surprised to find that the large timescale between assessments seems to not be necessary at all. Users report that the model is accurate and useful even when used to score every version of a article and that, in general, vandalism appears as either no change or a decrease in article quality, as intended [18].

One additional request that I received from Wikipedia editors was to provide a mechanism for getting between-class quality

[9] https://en.wikipedia.org/wiki/Heckman_correction

[12] The Objective Revision Evaluation System is a machine prediction platform used by Wikipedians to detect vandalism, measure article quality, and automate other useful activities in Wikipedia. See https://mediawiki.org/wiki/ORES

[13] https://github.com/wiki-ai/wikiclass/blob/950f693d789f8512e30f483f18e2d13483d13749/wikiclass/feature_lists/enwiki.py

[14] http://pythonhosted.org/revscoring/revscoring.scorer_models.html#revscoring.scorer_models.GradientBoosting

Table 2: B-class criteria comparison. The B-class criteria is compared between December of 2006[10] and April of 2017[11]. Note the increase in detail generally and the discussion of the importance of inline citations specifically. Also note that C-class didn't exist as of December 2006.

2006	Has several of the elements described in "start", usually a majority of the material needed for a completed article. Nonetheless, it has significant gaps or missing elements or references, needs substantial editing for English language usage and/or clarity, balance of content, or contains other policy problems such as copyright, Neutral Point Of View (NPOV) or No Original Research (NOR). With NPOV a well written B-class may correspond to the "Wikipedia 0.5" or "usable" standard. Articles that are close to GA status but don't meet the Good article criteria should be B- or Start-class articles.
2017	1. The article is suitably referenced, with inline citations. It has reliable sources, and any important or controversial material which is likely to be challenged is cited. Any format of inline citation is acceptable: the use of `<ref>` tags and citation templates such as `{{cite web}}` is optional. 2. The article reasonably covers the topic, and does not contain obvious omissions or inaccuracies. It contains a large proportion of the material necessary for an A-Class article, although some sections may need expansion, and some less important topics may be missing. 3. The article has a defined structure. Content should be organized into groups of related material, including a lead section and all the sections that can reasonably be included in an article of its kind. 4. The article is reasonably well-written. The prose contains no major grammatical errors and flows sensibly, but it does not need to be "brilliant". The Manual of Style does not need to be followed rigorously. 5. The article contains supporting materials where appropriate. Illustrations are encouraged, though not required. Diagrams and an infobox etc. should be included where they are relevant and useful to the content. 6. The article presents its content in an appropriately understandable way. It is written with as broad an audience in mind as possible. Although Wikipedia is more than just a general encyclopedia, the article should not assume unnecessary technical background and technical terms should be explained or avoided where possible.

Table 3: ORES wp10 fitness statistics. The fitness statistics of ORES "wp10" model based on 10-fold cross-validation are shown in table format. Note that "within-1" accuracy represents the per-class accuracy measure where being off by one is considered a successful prediction. The overall accuracy is 62.9% and the within-1 accuracy is 90.7%. These figures compare favorably to [21] of 58.2% and 89.5% respectively. I've included the ROC-AUC metric to demonstrate that the probability estimates made by the model are generally useful. As in [21], I also find that B- and C-class show lower overall fitness than other classes.

	n	ROC-AUC	acc (within-1)	Stub	Start	C	B	GA	FA
Stub	4968	98.5%	85.5% (99.2%)	4247	685	27	9	0	0
Start	4982	91.2%	64.3% (91.5%)	600	3205	754	358	58	7
C	4994	86.6%	48.9% (86.1%)	44	870	2443	986	558	93
B	4990	84.1%	40.3% (79.8%)	51	617	1258	2012	710	342
GA	5000	92.1%	62.7% (93.3%)	1	19	313	304	3135	1228
FA	4454	96.1%	77.4% (94.4%)	5	2	23	220	757	3447

predictions–essentially, being able to tell the difference between an article that is clearly a *Stub* and another article that is almost a *Start*. So I developed a basic strategy that I refer to as the *weighted sum* of class predictions. I assume that the ordinal article quality scale developed by Wikipedia editors is roughly cardinal and evenly spaced. To arrive at the *weighted sum* measurement, I multiply the prediction probability for each class by an enumeration of ordered classes starting at zero (0) for Stub and ending at five (5) for FA.

$$\text{weighed sum} = \sum_{c \in C} I(c)P(c) \tag{1}$$

This equation weighs each step in quality the same by multiplying the index of the class $I(c)$ by the prediction probability for that class $P(c)$.

When the model predicts that a version of an article is a Stub with 100% confidence, this *weighted sum* would be 0. If the prediction were split 50% Start and 50% C, the weighted sum would be 1.5. The *weighted sum* of the prediction demonstrated in figure 1 would be 3.8096 – slightly on the B side of GA-class. Again, my own spot checking and ORES' users confirm that this appears to be useful and consistent when applied to the history of articles.

Once I had concluded that the model operated consistently over time, I worked with my collaborators to generate a dataset that contains a predicted quality level for all articles in English Wikipedia at a monthly interval (600m article-month predictions). This dataset [9] contains the highest probability quality class ("prediction") as well as the *weighted sum* measurement for each article-month ("weighed_sum").

Aggregated quality measures. Using this dataset, we can move up a level from articles and assess the quality of Wikipedia at an aggregate level – as a whole or across interesting cross sections. I employed two aggregation strategies for comparing the quality of article aggregates: the *mean weighted sum*, and proportions of articles falling into each predicted class. In order to give these measurements a useful denominator I use the total number of articles in the aggregate as of the the most recent month of the dataset. Since the number of articles is clearly monotonically increasing (article creations always outnumber deletions), the count in the last month is always the *max* number of articles for any month in the cross section. Assuming that a Stub (the lowest quality class) is substantially more useful than no article at all, I assign a zero (0) *weighted sum* for all articles that have yet to be created in a particular month and increment the weighed sum for articles that exist by one (1) when generating the aggregate weighed sum. So if all articles were empty (like in the month before Jan 2001–the inception of Wikipedia), the *mean weighted sum* is zero. If all articles were predicted to be 100% FA class, the *mean weighted sum* would be 6.

These assumptions are bold and clearly provide an incomplete view of the reality of Wikipedia. *E.g.* if one were to draw a cross section of Wikipedia that included only one FA-class article, that cross section would get the maximum *mean weighted sum*. If in a following month, a Stub article were created, then the *mean weighted sum* would decrease even though there is more useful content. However, because this measure uses the number of articles in the last month as the denominator for all months, the *mean weighted sum* would also decrease historically, so this assumption is somewhat fair. Still, this assumption implies that Wikipedia will get no new articles after the date at which the dataset was generated. I feel that this is acceptable for comparing cross sections/aggregates that have more than a trivial number of articles and that are also generated using normal Wikipedian processes. I'll discuss this complicated problem and propose some solutions for addressing it in the Future work section. For the purposes of the demonstration analysis, I encourage you, dear reader, to judge this analysis by its interest as a demonstration of novel measurement capabilities and how this method enables us to visualize shifts in quality over time.

Demonstration: Coverage of Women Scientists in Wikipedia

When developing this measurement strategy, I found myself curious about the quality level of some cross sections of Wikipedia where there were once known content coverage gaps. I decided to focus on coverage of women scientists in Wikipedia. Regretfully, Wikipedia's category[15] structure is notoriously useless for drawing meaningful cross sections of content [12]. Luckily, WikiProject organizers are keen on making sure that all aticles about within their topic space are tagged and tracked by their project templates[16]. So searching for the presence of WikiProject Women Scientist templates is an efficient means for gathering this cross section[17]. Note that this method can be used to derive cross sections for any other WikiProject topic space.

As described in the previous section, I generated aggregate metrics both for the entire Wikipedia (about 5 million articles) and just the articles covering women scientists (5,681). I then compared the aggregated measures for *mean weighted sum* and proportion of articles in each quality class to look for shifts in the quality gap (when the apparent quality of articles about women scientists is lower than the apparent quality of the rest of the encyclopedia) and surplus (the opposite).

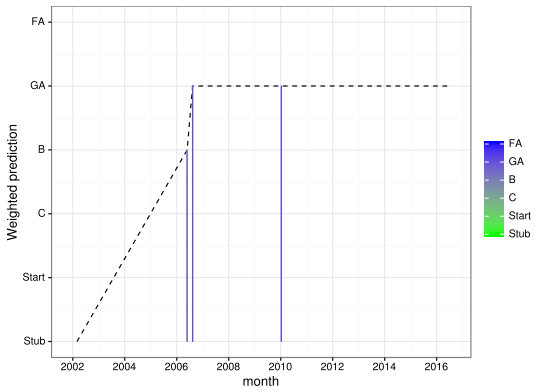
Figure 3a shows the difference between the "mean weighted sum" of article quality for all of Wikipedia and just the articles about women scientists. From the start of Wikipedia, a gap quickly develops and reaches the maximum of about a 15% a quality level around the middle of 2012. But after that point, the gap starts to rapidly close and a massive surplus begins to grow to about half of a quality class above the rest of the encyclopedia.

This dynamic roughly plays out in the same way for the proportions of articles that fall into the higher predicted quality classes. Figures 3b and 3c clearly show that the proportion of Start, C, and GA-class articles shows a similar gap and surplus

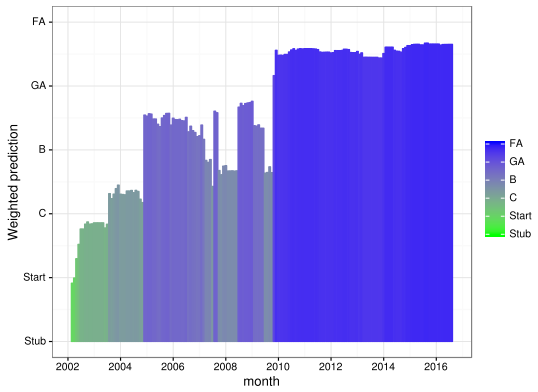
[15] See https://www.mediawiki.org/wiki/Help:Categories

[16] See https://www.mediawiki.org/wiki/Help:Templates

[17] See https://quarry.wmflabs.org/query/14033 for the query and its result.

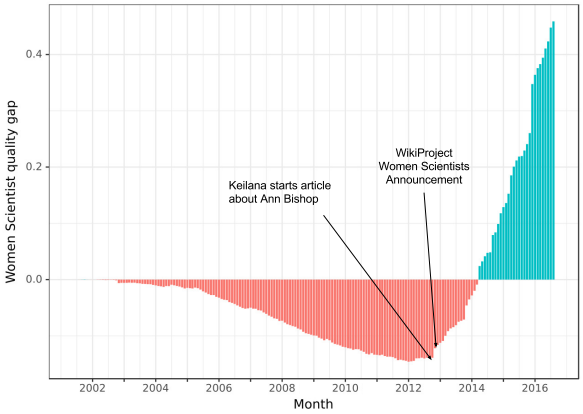


(a) Biology's assessments. The 3 manual assessments of the quality of Biology are plotted over time with a dashed line connecting them.

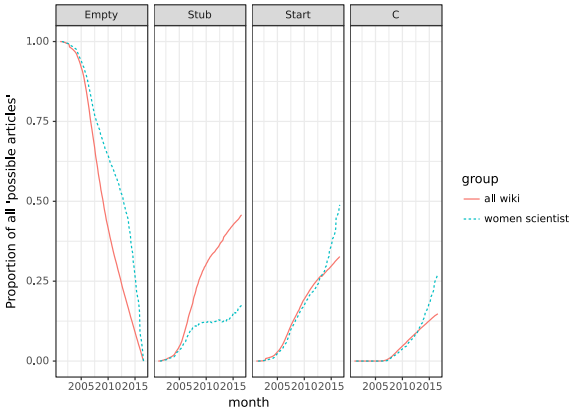


(b) Biology's predicted quality. The monthly *weighted sum* prediction for Biology is plotted.

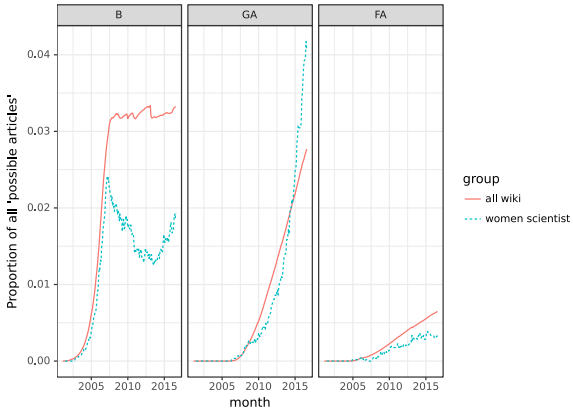
Figure 2: Comparing assessments to predicted quality for Wikipepdia's article titled "Biology". Note how 2b shows much more nuanced detail about the development of the article over time than 2a.



(a) The difference in *mean weighted sum* quality predictions for all wiki and articles about Women Scientists is plotted over time. Note the transition from red to blue represents the switch from a gap to a surplus. Important dates for User:Keilana's initiatives are annotated with arrows.



(b) The proportion of articles falling into Empty, Stub, Start, and C-class predictions is plotted for articles tagged by WikiProject Women Scientists and all of Wikipedia.



(c) The proportion of articles falling into B, GA, and FA-class predictions is plotted for articles tagged by WikiProject Women Scientists and all of Wikipedia.

Figure 3: The quality dynamics of biographies about women scientists vs. all of English Wikipedia.

period. Stub-class articles about women scientists seem to continue to fall behind the rest of the wiki, but there's still a noticeable increase in the proportion around the beginning of User:Keilana's initiatives (discussed in the Discussion section). It seems that the growth dynamics of B-class articles is shaped very differently from the rest of the quality prediction classes. Surprisingly, there seems to be no surplus in the proportion of articles that fall into FA-class. It seems that this is the only quality class where no noticeable shift occurs – the gap between women scientist biographies and the rest of Wikipedia only widens over time.

DISCUSSION

When I first saw this gap-to-surplus shift, I honestly had no idea what could have caused it. In an effort to share my analysis with a larger community, I presented the preliminary results in a prominent public forum for Wikimedia-related research projects – the Wikimedia Research Showcase[18]. During the presentation, Wikipedia editors who ran a series of outreach initiatives to bring attention to biographies about women scientists reached out to me to let me know that the beginning of their initiatives corresponded to the beginning of the period of the shift from gap to surplus. User:Keilana (aka Emily Temple-Wood) and her collaborators received substantial attention from the media for their work to increase coverage of women scientists in Wikipedia [4]. At the most basic level, my analysis of content quality and coverage of this topic in Wikipedia seems to suggest that their work has had a very large effect. However a causal conclusion must be left for future work which I discuss more substantially in the Future work section.

I haven't been able to determine what might explain the unusual shape of the trajectory for the proportion of B-class articles. One hypothesis is related to the process by which GA and FA-class articles are assessed. Unlike the lower quality classes, GA and FA-class articles go through a formal peer review process and are promoted in other places in the wiki. It's possible that there is a large incentive to bring any B-class article to the next stage (GA) since it is only one step away and achieving GA-class is rewarded with public recognition.

The smaller proportion of articles that are predicted to be at the FA-class level is maybe a bit more concerning for the coverage of content about women scientists. One possibility is that the model is biased against articles about women scientists. There's a well discussed bias against coverage of women scientists in the type of reference work that Wikipedia tends to cite. On one had, it could be that articles about women scientists are inherently shorter and just look to be lower quality that other articles in the wiki. It could also be that it is hard to write a truly comprehensive article about a woman scientist for the same reason. It turns out that Wikipedians' own assessments suggest that there is a similar proportion of FA-class articles about women scientists(7/5681 = 0.12% [2]) as there are across the entire wiki (6k/5m = 0.12% [1]), so it could be possible that the model is showing a slight bias here.

[18] https://www.mediawiki.org/wiki/Wikimedia_Research/Showcase

CONCLUSIONS & FUTURE WORK

In this paper, I introduce a novel measurement strategy. By repeatedly applying an extended version of the article quality model developed by Warncke-Wang et al. [22] to the historical versions of Wikipedia articles, the dynamics of their quality can be examined at a granularity that was not possible until now. Using this measurement strategy, I have demonstrated that the quality of coverage about topics of high importance to Wikimedia and Wikipedia editors (biographies about women scientists) can be examined in novel ways and new insights can be gained. I've also linked directly to a public release of open licensed data and an openly available API for generating new predictions that I hope will enable others to more easily explore similar analyses and analysis strategies.

To download the dataset, see https://doi.org/10.6084/m9.figshare.3859800
To access the prediction API, see https://ores.wikimedia.org/

Limitations

As in past work that used Wikipedia's quality assessments as an outcome measure, we're assuming that Wikipedians' notions of *quality* correspond to some general, fundamental *true quality*. But quality is a more complex concept that is arguably context and purpose dependent. For some users and uses of Wikipedia, a Stub-class may be perfectly acceptable (*e.g.* settling a bar bet), while for others an FA-class would be necessary to really get a complete overview of the subject. I have decided to model Wikipedians' own quality ratings because their context and purpose is the construction of an encyclopedia. While this may not fit all contexts and purposes, it does seem relevant to a substantial cross section of the research literature that is concerned about the efficiency of open production processes and it is clearly useful for ORES' intended audience: Wikipedians.

Another concern about his method is the *meaning* behind changes in the predicted quality level of an article over time. As discussed in the Methods for measuring article quality section, the article quality prediction model was formally evaluated against withheld data (wikipedians' assessments) and these statistics suggest a high level of fitness. However, Wikipedians direct their own assessment activities and that means there may be something special about the revisions that are generally assessed. Essentially, there are many unassessed versions of articles between the revisions that were directly assessed by Wikipedians and we don't have a ground truth about the quality of those revisions to compare against. One of the primary claims that I make in this paper is that the article quality model deployed in ORES (that was used to generate the linked dataset and perform the demonstration includedin this paper) is an effective means to assess these otherwise unassessed versions of article content. I justify this conclusion by (informally) observing that that the temporal dynamics of article quality as measured by the model seem to closely reflect reality. I also note that ORES' users and I have used the model *in practice*. It is from this that I conclude that the temporal dynamics of predicted quality represent useful information about real changes in article quality. While this type of

assessment may be reasonable for the purposes of this study, a formal analysis of the prediction model's ability to predict the quality inbetween natural assessments is desirable. Future work could ask Wikipedians to directly assess a random set of otherwise unassessed revisions and measure the fitness of the model against those revisions in to perform such a formal analysis.

As was mentioned in the Modeling actionable article quality section, measuring the *completeness* of a cross section of Wikipedia is difficult when it's not clear how many articles the cross section *should* have. I have chosen to operate on the assumption that the number of articles present in the cross section at the time of measurement is a useful denominator to use historically. Future work could explore the development of a more reasonable denominator by taking advantage of indexes of known notable topics that might eventually have an article in Wikipedia. Notably, User:Emijrp has already put substantial effort into just such an initiative [5]. He estimates that the current total number of articles in English Wikipedia (about 5 million) represents only about 5% of all of the articles that will eventually be covered in the encyclopedia.

Finally, it's important to note that this paper does not do a study that would be able to rigorously conclude the direct *causal relationship* between Keilana's initiatives and the coverage of women scientists in Wikipedia. Such a conclusion may be apparent, but there are other potential explanations for the observed correlation. For example, it could be that there was generally a sudden surge in interest around women scienists in the beginning of 2013 from which both Keilana's initiatives and the quality of articles about women scientists were independent effects. Future work may examine this by analysing the contributions in this content space and qualitatively studying the motivations of these volunteers to find out if the initiatives or something else had inspired them to do take on this work.

Future work

Beyond this demonstration, I would like to use this dataset and method to explore article quality inflection points that occur in other cross sections of Wikipedia. As was discussed, I was surprised to learn about the temporal proximity (and apparent effectiveness) of Keilana's efforts since I had discovered the inflection that started around her initiatives independently. Other "Keilana Effect"s should also become evident when re-applying this analysis method. Doing so would allow the Wikimedia Foundation and other organizations that support initiatives (like the grant that supported User:Keilana) to bring attention and resources to efforts that are already working far better than expected.

Beyond looking for inflection points, we can also simply look for gaps to target effort towards. In the past, determining where a coverage gap might be and how big large the gap is would require massive effort on the part of Wikipedians to assess thousands of articles. And worse, it would be nearly impossible to find out when the gap originated and whether or not it seemed to be widening or closing. Using this dataset and method, researchers can explore the largest gaps in coverage in Wikipedia and organizations like the Wikimedia Foundation can target new outreach campaigns (like Inspire campaigns[19]) to improve coverage in areas with known gaps.

This modeling approach also makes research around the nature of Wikipedia's quality dynamics easier. For example, using this modeling strategy, one could replicate or extend the modeling work performed by Kittur et al. [13] and Arazy & Nov [3] without resorting to propensity modeling, with far more observations, and in time scales where Wikipedian assessments are sparse. By developing and releasing this dataset, I intend to make exactly this type of work easier.

ACKNOWLEDGMENTS

I would like to thank Morten Warncke-Wang who has been a thoughtful collaborator in helping to re-implement his model in the ORES system. Thanks to Amir Sarabadani who contributed substantially to the generation of the article quality prediction dataset used in this study [9]. I would also like also acknowledge the Wikimedia Foundation who directly supported the research described in this study and Jonathan T. Morgan for feedback and proofreading of this paper.

REFERENCES

1. Wikipedia WP 1.0 bot statistics. `https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Statistics`, 2017. Accessed: 2017-04-19.
2. Women scientists WP 1.0 bot statistics. `https://en.wikipedia.org/wiki/User:WP_1.0_bot/Tables/Project/Women_scientists`, 2017. Accessed: 2017-04-19.
3. O. Arazy and O. Nov. Determinants of Wikipedia quality: The roles of global and local contribution inequality. In *CSCW*, pages 233–236. ACM, 2010.
4. R. Change. Emily temple-wood: A cool Wikipedian on a big mission. `https://en.wikipedia.org/wiki/User:Emijrp/All_human_knowledge`, 2013. Accessed: 2017-04-19.
5. Emijrp. All human knowledge. `https://en.wikipedia.org/wiki/User:Emijrp/All_human_knowledge`, 2016. Accessed: 2017-04-17.
6. R. S. Geiger and D. Ribes. The work of sustaining order in Wikipedia: The banning of a vandal. In *CSCW*, pages 117–126. ACM, 2010.
7. J. Giles. Internet encyclopaedias go head to head, 2005.
8. M. Graham, B. Hogan, R. K. Straumann, and A. Medhat. Uneven geographies of user-generated information: patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4):746–764, 2014.
9. A. Halfaker and A. Sarabadani. Monthly Wikipedia article quality predictions. `https://doi.org/10.6084/m9.figshare.3859800.v3`, 2016.

[19] https://meta.wikimedia.org/wiki/Grants:IdeaLab/Inspire

10. B. Hecht and D. Gergle. Measuring self-focus bias in community-maintained knowledge repositories. In *Communities and technologies*, pages 11–20. ACM, 2009.

11. I. L. Johnson, Y. Lin, T. J.-J. Li, A. Hall, A. Halfaker, J. Schöning, and B. Hecht. Not at home on the range: Peer production and the urban/rural divide. In *CHI*, pages 13–25. ACM, 2016.

12. A. Kittur, E. H. Chi, and B. Suh. What's in Wikipedia?: mapping topics and conflict using socially annotated category structure. In *CHI*, pages 1509–1512. ACM, 2009.

13. A. Kittur and R. E. Kraut. Harnessing the wisdom of crowds in Wikipedia: Quality through coordination. In *CSCW*, pages 37–46. ACM, 2008.

14. S. T. K. Lam, A. Uduwage, Z. Dong, S. Sen, D. R. Musicant, L. Terveen, and J. Riedl. Wp: clubhouse?: an exploration of Wikipedia's gender imbalance. In *WikiSym*, pages 1–10. ACM, 2011.

15. M. Mesgari, C. Okoli, M. Mehdi, F. Å. Nielsen, and A. Lanamäki. "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology*, 66(2):219–245, 2015.

16. R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in Wikipedia. In *GROUP*, pages 259–268. ACM, 2007.

17. J. Reagle and L. Rhue. Gender bias in Wikipedia and Britannica. *International Journal of Communication*, 5:21, 2011.

18. S. Ross. Visualizing article history with structure completeness. `https://wikiedu.org/blog/2016/09/16/visualizing-article-history-with-structural-completeness/`, 2016. Accessed: 2017-04-17.

19. B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Information quality work organization in Wikipedia. *Journal of the Association for Information Science and Technology*, 59(6):983–1001, 2008.

20. M. Warncke-Wang. English Wikipedia quality asssessment dataset. `https://doi.org/10.6084/m9.figshare.1375406.v1`, 2015.

21. M. Warncke-Wang, V. R. Ayukaev, B. Hecht, and L. G. Terveen. The success and failure of quality improvement projects in peer production communities. In *CSCW*, pages 743–756. ACM, 2015.

22. M. Warncke-Wang, D. Cosley, and J. Riedl. Tell me more: An actionable quality model for Wikipedia. In *OpemSym*, page 8. ACM, 2013.