Improving Indexing Efficiency & Quality: Comparing A-B-Arbitrate and Peer Review

Derek Hansen
Brigham Young
University, Provo, UT
shakmatt@gmail.com

Jake Gehring
FamilySearch
Salt Lake City, UT
GehringJG@familysearch.org

Patrick Schone
FamilySearch
Salt Lake City, UT
BoiseBound@aol.com

Matthew Reid
Brigham Young
University, Provo, UT
matthewreid007@gmail.com

ABSTRACT

The Family Search Indexing project has enabled the manual indexing of millions of records by hundreds of thousands of volunteers making it one of the largest crowdsourcing initiatives in the world. Currently to assure the quality of indexing, each image (e.g., census page) is indexed by two independent indexers and any discrepancies are reviewed by an arbitrator. An alternate, yet untested peer-review indexing process would use only one indexer, one reviewer of their work, and optionally an arbitrator who looks at differences. This method will likely lead to higher efficiency, but its effect on quality is not known. In this paper we analyze historical data that uses the existing A-B-Arbitrate process and describe an experiment that is underway to compare it with the proposed peer review process. The historical data analysis shows that agreement between independent indexers increases as their prior indexing experience increases; agreement is higher in English-speaking languages than foreign languages; and that agreement varies considerably based on field type (e.g., surname, county, gender). Implications of these findings are discussed.

Keywords

Family Search Indexing, A-B-Arbitrate, arbitration, peer review

1. INTRODUCTION

FamilySearch Indexing volunteers have been responsible for indexing nearly 700 million records making them searchable by genealogists around the globe. The volunteer workforce consists of hundreds of thousands of volunteers with over 500 new volunteers signing up each day. Their job is to transform historical documents, such as handwritten census records, into machine readable text using specialized software that displays digitized images alongside data entry fields. Despite their tremendous effort, volunteers struggle to keep pace with the mass of documents being digitized at rates never before possible thanks to new digitization techniques.

Since its inception, FamilySearch Indexing has been dedicated to creating high-quality indexes. To assure accurate indexing, FamilySearch has used the A-B-Arbitrate quality assurance process, whereby two individuals (A and B) independently index a page and any discrepancies between their work is passed to an expert arbitrator (ARB) who is responsible for the final indexing decision. While this process presumably provides a high standard of quality, it comes at a cost, since each image must be fully indexed by 2 people plus pass through the quality-assurance stage.

An untried alternate method is to use a peer review quality assurance process. This entails the indexing of a record by one person (A) which is passed along to a reviewer (R) who looks at the image and index and fixes any errors they can find. Optionally, the changes that are made could be reviewed by an independent, third-party arbitrator (ARB). We call this the A-R-Arbitrate quality assurance process. If reviewing takes significantly less time than indexing from scratch, this method will more efficiently allocate volunteer efforts. More documents can be indexed for the same amount of time expended. However, its impact on quality is not clear. On one hand, reviewers may be too prone to agree with the original indexer, which would lead to mistakes that would have been caught had the reviewer independently indexed the work. On the other hand, reviewers may take a bit more time to focus on the difficult cases leading to a more careful review.

The authors are currently assessing these two methods using two different approaches. First, we are performing an analysis of the historical data to understand the variations in quality and efficiency based on field (e.g., Surname, Gender, County), language, project (e.g., U.S. 1880 Census versus U.S. 1930 Census), and the expertise levels of the indexers (A, B, and ARB). In this paper we present some of our preliminary findings and discuss their implications for improving A-B-ARB as well as their potential implications for using an A-R-ARB model.

Second, we are currently conducting a field experiment to test the viability of the A-R-ARB model. The study will allow us to compare the effect of the different methods on quality (measured by comparison to a truth set) and efficiency (measures by keystroke time capture logs). In this paper we describe our methodology, though we do not yet have results to share.

While this study will provide direct insights into the experience of FamilySearch Indexing, the findings of our study are of interest to other related indexing projects being conducted by companies such as Ancestry.com and governments such as France. More generally, the findings may provide benchmarking data and insights to other related crowdsourcing efforts such as Project Gutenberg, which uses volunteers to proofread digitized text from out-of-copyright works. As the number of "commons-based peer production" projects such as Wikipedia, FamilySearchIndexing, Encyclopedia of Life, and Project Gutenberg increase [1], it is increasingly important to systematically evaluate the quality assurance processes to assure that the volunteer efforts so generously provided are put to good use.

2. HISTORICAL DATA ANALYSIS

When images are indexed through the FamilySearchIndexing platform, there is a record that is made which saves several fields that are useful for performance analyses. In particular, in addition to the transcript that is created, it will save the name of the indexer, the time that the indexer invests in annotation of the image (as opposed to time being "idle"), and the number of keystrokes that were required to index the image.

Using a compilation of all these indexing records created from inception through January of 2011, we can reasonably determine how experienced each indexer is by the number of images that he or she has transcribed. The top five indexers of all time have each indexed between 205K and 267K full images. On the other hand, 12,931 indexers only annotated a single image.

In our experiments, we want to see how the experience level of the annotators affects the accuracy, speed, and the number of keystrokes they use while indexing. Therefore, we have assigned an experience level that directly correlates with the number of images an indexer has transcribed. Let *U* represent the user/indexer, and let N(*U*) be the number of images that *U* has transcribed. Then, for simplicity, we assign *U* to an experience level, EL(*U*), based on the formula

$$EL(U) = \mathrm{round}(\log_5(N(U)))$$

where "round(X)" is the usual rounding operation. This formula means that an indexer who has annotated only one image will be assigned to a skill level of 0, whereas the top indexers will be assigned to a skill level of 8. The median number of images indexed is 50, which means that the median skill level is a 2.

Given an estimation of annotator experience, we can perform a number of analyses. Our analyses are broken into two parts. First, we focus on the variation in the percent agreement between A and B. This measure of consistency is assumed to correlate with actual quality. Future work will test this assumption. Next we focus on variations in time spent indexing and the number of keystrokes.

2.1 A-B Agreement Analysis

Since we do not have a truth set for all of the data, we will focus on the percent agreement between A and B. While this is an imperfect measure of quality (and is not a measure of accuracy), we anticipate that fields with high agreement are likely to be of higher quality than fields with lower agreement. Future work will test this assumption. However, we can gain several important insights from the A-B agreement data. Table 1 shows the A-B agreement on key genealogical fields from the U.S. Census records in our data collection.

Table 1. A-B Agreement percent by Field for all U.S. Census records in our corpus.

Indexed Field	Agreement
Gender	98.8%
Census County	70.3%
Given Name	82.5%
Surname	74.7%
Birth Place	96.1%
Relation to Head of House	95.0%
Age	91.6%
Birth Date	97.8%
Father's Birth Place	96.7%
Mother's Birth Place	96.7%
Immigration Year	90.0%

As expected, we find that agreement varies significantly depending upon the field of data being indexed. Table 1 shows the percent agreement for each field in the available U.S. Censuses from 1850 to 1920. Other datasets include similar differences between fields. As expected, fields with few possible values (e.g., gender) have extremely high agreement, while fields with many possible values (e.g., Surname) have a much lower agreement.

The Canadian 1871 Census provides us with a nice opportunity to compare agreement based on language, since the exact same forms were used in both English and French. The comparison shows that agreement is significantly higher for English-language indexing than for French-language indexing. If we consider just the given name and surname fields, we find that the French Canadian census records on average have only a mere 62.7% and 48.8% accuracy compared to 79.8% and 66.4% for the English Canadian records. This may be due to the fact that most indexers are native English speakers. This finding suggests the importance of recruiting volunteers who can index records in their native tongue.

Finally, we evaluate agreement based on the prior experience (as described earlier) of the indexers. Figure 1 shows this comparison in the form of a heatmap which shows the agreement levels between all possible experience matchups. The lowest agreement (shown in blue) occurs when inexperienced annotators are matched up with other inexperienced annotators, while the highest agreement occurs between experienced annotators. While the general trend is not surprising, the continued improvement even at the very high levels is worth noting. It is consistent with other work on expertise that shows that people continue to get better tasks though with diminishing returns [2].

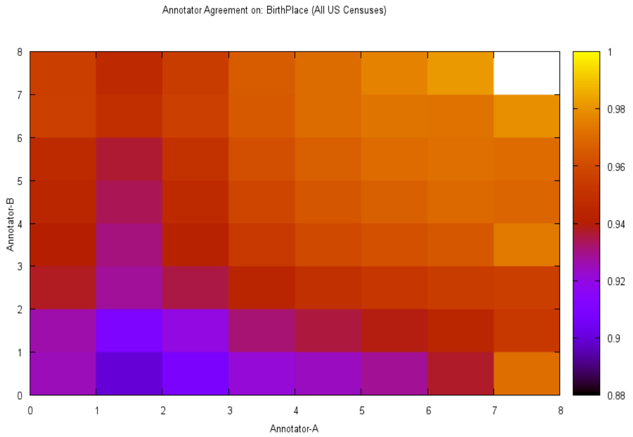

Figure 1. Heatmap of A-B agreement by Experience Level for Birth Place in all U.S. Censuses in our corpus.

This pattern of continuous improvement is found in other fields such as Given Name and Surname, though it is less dramatic in the case of Surname (see Figure 2). Improvements in agreement for Gender were noticeable but very small since agreement was so much higher for all of the expertise levels. Counter intuitively, agreement for Birth Place did not seem to improve for certain other datasets including the English-speaking Canadian Censuses. For this field, unlike with Gender, the agreement remained relatively low. Further analysis is needed to determine why this is the case, but it is likely that prior experience in one project does not necessarily translate into improved expertise because indexers are not familiar with the place names that are involved.

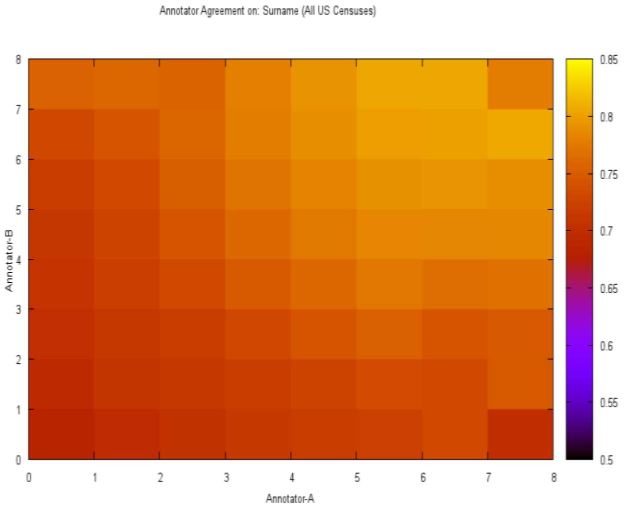

Figure 2. Heatmap of A-B agreement by Experience Level for Surname in all U.S. Censuses in our corpus.

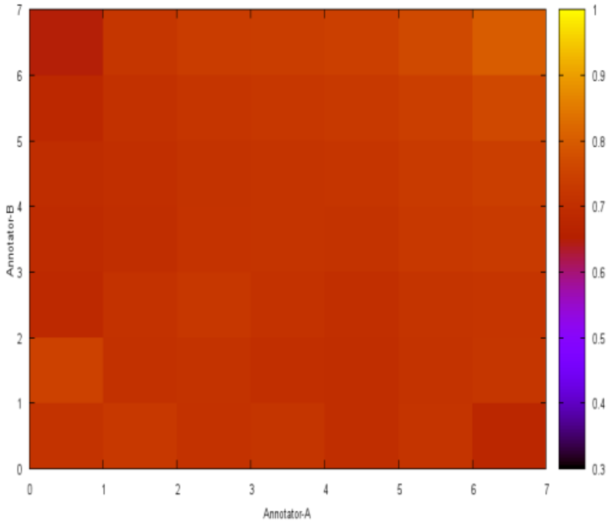

Figure 2. Heatmap of A-B agreement by Experience Level for Birth Place in the English-speaking Canadian Census.

Taken as a whole, these finding suggests that the selection of the A and the B indexer may be an important factor influencing the quality of the work and/or the effort required by the arbitrators.

Indexing Time Analysis

The next analysis is based on the keystroke and non-idle time captured by the indexing software. Data on specific fields is not available, since the time is based on completion of an entire page. In computing time analyses, we recognize that some indexers may take a break before finishing a project which could skew results, so we discard outlier times that are in excess of two hours or that are less that 20 seconds.

The number of keystrokes and the time spent by experts is less than that of novices -- as shown in Table 2. The estimated average of keystrokes per line for the US records in our dataset has a small but consistent downward trend, except for those indexing their very first record (Experience Level 0 people who are given easy batches to get them acquainted with the indexing system). Since the same amount of data needs to be indexed on each line by experts and non-experts, it is likely the case that experts revise their entries less often (e.g., change a surname in a previous field because they notice they got it wrong the first time).

Table 2. Time and Keystroke data by Experience Level

Experience Level	Avg Keystrokes per Line	Avg Time per Line	Avg Time per Keystroke
0	18.74	65.79	4.31
1	19.25	63.10	3.96
2	19.42	55.54	3.47
3	18.53	48.21	3.22
4	18.03	41.53	2.92
5	17.67	34.71	2.57
6	17.50	28.87	2.22
7	17.44	23.16	1.82
8	17.65	14.95	1.18

Changes in Time per line (measured in seconds) are far more dramatic. Experts can be up to 4 times faster than novices. The weighted average of the times for the first three experience levels (0-2), which includes the skill level of the median contributor to FamilySearchIndexing, is double the weighted average of the three highest Experience Levels (6-8). The average Time per Keystroke also goes down considerably since there are such improvements in time. This finding suggests that FamilySearch-Indexing should work hard to cultivate continued participation among indexers who become much faster with experience. This is particularly true, since not only does their efficiency increase, but as we saw before, their agreement levels also increase.

We also compared the time per line and keystroke per line for the French-speaking and English-speaking 1871 Canadian Census. The English-speaking census was 2.68 seconds faster per line. While this sounds small, when aggregated over the 3 million+ French-language census lines, it amounts to over 2,000 hours of additional time, or about one person working full time for a year. As discussed earlier, it is likely that most indexers are not French-speaking natives, which suggests that having more people index in their native language would increase efficiency as well. Interestingly, there were an average of 2.64 more keystrokes in Canadian English than in Canadian French, which may be attributable to differences in the length of words between the languages. In summary, English-language indexing was faster than French-language indexing even though more keystrokes were used.

Finally, we look at the time and keystrokes that were spent by the arbitrator of records that were indexed by people with different experience levels. Table 3 shows this data from annotations of the 1910 US Census. All US Censuses were not combined in this analysis because each has a different number of columns and rows. The table emphasizes that arbitrators spend less time and keystrokes when reviewing data from more experienced indexers, presumably because there are fewer corrections to be made. The average time per keystroke is not dramatically different, though there is a slight upward trend as expertise level increases. This makes sense since cases where experts differ (particularly amongst themselves) are likely the hardest cases.

Table 3. Time and Keystroke data by Experience Level for the 1910 US Census

Experience Level of Indexer	Avg Keystrokes of Arbitrator	Avg Time of Arbitrator	Avg Time per Keystroke
0	35.1	466.6	13.3
1	35.0	492.1	14.1
2	35.4	482.1	13.6
3	33.8	467.9	13.8
4	31.6	445.0	14.1
5	28.9	421.7	14.6
6	27.4	399.7	14.6
7	26.7	386.7	14.5
8	24.1	370.8	15.4

3. FIELD EXPERIMENT DESIGN

We are currently conducting a field experiment to compare the results of three different quality assurance processes. The key variables we are interested in are time (our measure of efficiency) and accuracy as compared to a truth set (our measure of quality). We are focusing on 2,000 randomly selected pages from the 1930 U.S. Census. The truth set for these images was generated by an indexing company who assured a 99.9% accuracy rate. FamilySearchIndexing experts audited a subset of the work and found it to meet the criteria. Below is a discussion of the 3 conditions:

A-B-Arbitrate Condition: A and B index a page independent of one another and any discrepancies are passed to a third Arbitrator who makes the final decisions. Historical data will be used since this method was already implemented to index the records.

A-R Condition: A indexes a page and R reviews the index by identifying and correcting any errors they find. Historical data will be used for the original A index and new volunteers will serve as the reviewers (R).

A-R-Arbitrate Condition: A indexes a page and R reviews the index and corrects any errors they find. A third party Arbitrator reviews all places where A and R differ and makes the final decision. The same data for the A-R Condition will be used, but new Arbitrator data will be collected based on the differences between A and R.

We hypothesize that the A-R and A-R-Arbitrate processes will take less time because reviewing (R) should take less time than independently indexing a record (B). The effect on quality is less clear. We are hopeful that there is no difference in quality between the methods, in which case the A-R Condition could be implemented with its anticipated gains in efficiency. However, it is possible that reviewing leads to lower quality indexing, in which case the tradeoffs would need to be assessed.

The historical data analysis presented in Section 2 above showed the importance of expertise on time and our imperfect estimate of quality (A-B agreement). We are ensuring that the reviewers (R) come from a wide range of expertise levels, so that we can control for the impact of expertise when comparing the results between groups. We also will be able to use the data to model different scenarios. For example, we could model the impact of allocating more experts as reviewers versus original indexers.

4. DISCUSSION

Our preliminary study has used historical data to show the significant impact that indexing experience has on reducing time and improving agreement among indexers. As mentioned previously, these findings suggest the importance of retaining experienced indexers and motivating those already doing some indexing to continue to index more. Other findings suggest the need to recruit native speakers of foreign language indexing projects to improve efficiency and quality.

Our future experimental study will provide data that will assess the viability of a new peer review process that will likely improve efficiency, but whose impact on quality is not clear. Modeling how to use experts and novices in the new model will be important.

One potential way of improving the indexing process would be to split the full page images into individual fields that could be indexed independent of the rest of the page. This would allow easy fields (e.g., gender) to be indexed by novices who can later transition into indexing more advanced fields (e.g., surname) as they demonstrate their competency. Furthermore, it would enable the use of image recognition software that uses algorithms to perform the indexing. While such systems are not currently able to replace human effort entirely, they could augment human endeavors. For example, they may be able to perform an initial peer review of a human indexed record. If the algorithm had a high enough confidence level a human review would not be needed, whereas if it had a low confidence level it could be passed on to a human reviewer. Alternatively, in some cases it may be able to perform the initial indexing (e.g., of a field like gender), which would be reviewed by a human.

Integrating humans and algorithms into the same process has the added advantage that machine learning techniques can improve over time as humans continue to provide feedback on the algorithms' successes and failures. This hybrid approach may have the ability to increase indexing capacity by orders of magnitude, something that is desperately needed if we are to keep pace with the wealth of vital data that is being digitized each day.

5. ACKNOWLEDGMENTS

We would like to express our appreciation to Rose Pierson, Zane Jacobson, Katie Gale, Stephen Pew, Paul Starkey, and LDS missionaries for their tremendous assistance in helping to create and implement the method for testing the new A-R-Arbitrate paradigm.

6. REFERENCES

[1] Benkler, Y. 2002. Coase's penguin, or, Linux and the nature of the firm. Yale Law Review, 112, 3, 369-446.

[2] Anderson, J. 2009. *Cognitive psychology and its implications, 7th edition*. Worth Publishers.