Quality Control Mechanisms for Crowdsourcing: Peer Review, Arbitration, & Expertise at FamilySearch Indexing

Derek L. Hansen
Brigham Young University
Provo, UT
dlhansen@byu.edu

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

Douglas Corey
Brigham Young University
Provo, UT
corey@mathed.byu.edu

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

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

ABSTRACT

The FamilySearch Indexing project has enabled hundreds of thousands of volunteers to transcribe billions of records, making it one of the largest crowdsourcing initiatives in the world. Assuring high quality transcriptions (i.e., indexes) with a reasonable amount of volunteer effort is essential to keep pace with the mounds of newly digitized documents. Using historical data, we show the relationship between prior experience and native language on transcriber agreement. We then present a field experiment comparing the effectiveness (accuracy) and efficiency (time) of two quality control mechanisms: (1) Arbitration – the existing mechanism wherein two volunteers independently transcribe records and disagreements go to an arbitrator, and (2) Peer Review – a mechanism wherein one volunteer's work is reviewed by another volunteer. Peer Review is significantly more efficient, though not as effective for certain fields as Arbitration. Design suggestions for FamilySearch Indexing and related crowdsourcing initiatives are provided.

Author Keywords

Crowdsourcing, transcription, historical documents, quality control, peer review, FamilySearch Indexing, genealogy.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Crowdsourcing projects such as Wikipedia, IMDB, Project Gutenberg, and FamilySearch Indexing (FSI) are responsible for the creation of millions of information artifacts. Their output is used daily by the masses and increasingly integrated into other systems. Despite our increasing reliance on peer-produced content, our understanding of how to assure its quality is still in its infancy. The work presented here sheds light on this topic by examining FSI, the world's largest historical document transcription service. Specifically, we examine the role of expertise and different quality control mechanisms on accuracy and efficiency.

FSI volunteers have been responsible for transcribing (i.e., indexing) billions of historical records making them freely available on the Internet in machine-readable format for genealogists and historians around the globe to search. Documents include census records, vital records (e.g., birth, death, marriage, burial), church records (e.g., christening), military records, legal records, cemetery records, and migration records from countries around the globe. These documents are often hand-written, making their transcription challenging, particularly when the document authors used cursive letters that differ from those used today. These challenges make the role of experience and learning vital.

The volunteer workforce has included nearly 400,000 contributors with over 500 new volunteers signing up each day. Volunteers use special client software that shows an image alongside data entry fields (Figure 1). The software and funding for the project are provided by the Church of Jesus Christ of Latter-day Saints. Despite their tremendous effort, users cannot keep pace with the mass of newly digitized documents prompting the need for more efficient, yet equally effective transcription processes.

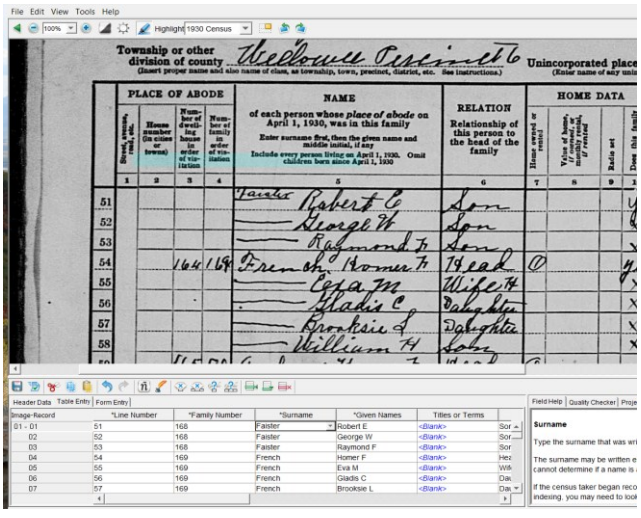

Figure 1. FamilySearch Indexing client software. Original image on top and structured data entry fields and notes on bottom.

CSCW '13, February 23–27, 2013, San Antonio, Texas, USA.

Since its inception, FSI has sought to create high-quality transcriptions. To increase accuracy, FSI has used an *arbitration-based* quality assurance process, whereby two individuals (A and B) independently transcribe an image and any discrepancies between their outputs are passed to an experienced arbitrator (ARB) who makes the final decision. We refer to this model as A-B-ARB. While this process presumably provides a high standard of quality, it comes at a high cost, since each image must be fully transcribed by two people plus pass through the arbitration stage.

An as-yet untried, alternate method is to use a *peer review* quality assurance process. This entails the transcription of a record by one person (A) which is passed along to a reviewer (R) who identifies and fixes any errors she can find. We refer to this model as A-R. Optionally, changes made by the reviewer can be arbitrated by a third party contributor (RARB). We refer to this model as A-R-RARB.

If reviewing a document takes significantly less time than transcribing it from scratch, the peer review method will more efficiently allocate volunteer efforts. More documents can be transcribed in the same amount of time. However, peer review's impact on quality is not clear. On the one hand, reviewers may be too prone to agree with the original transcriber, which would lead to mistakes that would have been caught had the reviewer independently transcribed the work. Alternatively, reviewers may spend more time on difficult cases leading to a more careful review.

This paper explores quality control within FSI using two complementary approaches. First, we analyze historical data created using the arbitration mechanism to better understand the relationship of prior experience and native language on transcriber agreement and time spent. Then, using a subset of data from the 1930 US Census, we present a field experiment comparing the effectiveness (accuracy) and efficiency (time spent) of A-B-ARB, A-R, and A-R-RARB. We finish by discussing design implications, many of which can be applied to other crowdsourcing projects.

BACKGROUND

Crowdsourcing, a term coined by Jeff Howe in 2006, broadly describes "the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call" [13]. Though exact definitions of crowdsourcing vary, they typically involve the completion of discrete tasks by voluntary contributors who have varying levels of expertise [10]. Many crowdsourcing initiatives, such as Wikipedia, IMDB, Project Gutenberg, and FSI create open access information resources freely available to the public. Crowdsourcing systems must recruit users, orchestrate their activity (e.g., break tasks into modular components), aggregate their contributions, and discourage deviant behavior [8]. FSI is a prototypical example of crowdsourcing, as it takes a task formerly performed by professional transcribers, breaks it up into discrete tasks completed by volunteers, and aggregates their contributions into a final product.

FamilySearch Indexing work is also a good example of human computation, because the transcription work (1) might someday be solvable by computers (i.e., it is a "computational" task), and (2) is completed by humans who are directed by a computational system or process [21]. Other human computation systems include Games With A Purpose (GWAP) such as the ESP game [25], services such as Mechanical Turk, and a range of others focused on topics as diverse as translation, annotation [14], document editing, transcription, and image recognition [21]. As is typical of many human computation systems (as well as "lightweight peer production" systems [12]), FSI is currently organized in such a way that largely anonymous contributors independently complete discrete, repetitive tasks provided by authorities. Though members of local religious congregations and genealogical societies help train and motivate one another to participate in FSI, the FSI system does not yet support "heavy-weight peer production" wherein contributors develop strong-tie relationships and internally-negotiated community norms and policies [12].

Despite the success of many crowdsourcing and human computation initiatives, the processes that lead to high quality output are not fully understood. Quinn and Bederson describe nine types of quality control mechanisms for human computation (e.g., redundancy, multi-level review) [21], though most systems rely on a single mechanism. Paul Duguid shows that existing "laws of quality" invoked by peer production systems (e.g., given enough eyeballs all bugs are shallow) do not always hold up under scrutiny at Project Gutenberg, Wikipedia, and other peer production systems [9]. Work on optical character recognition, a task similar to transcription, illustrated how simple techniques like independent agreement don't work well for difficult tasks or ones with many possible answers [18].

The strategies used to assure high quality depend largely on the tasks and outputs of crowdsourcing projects. For complex unstructured tasks, such as contributing to a Wikipedia article, quality is largely affected by the number and diversity of contributors, coordination practices, and policies [15,22]. For concrete, structured tasks, such as those at FSI, automated methods can be used to aggregate user contributions in a way that promotes high quality [8]. For example, some systems use the reputations of contributors as weights in determining which of multiple contributions is most likely accurate (e.g., [19]). Others use peer or expert oversight to increase quality and quantity of contributions [5]. Some new techniques, like the Find-Fix-Verify pattern [3] or the tournament selection approach [23], work well in some conditions, but come at a cost of more total effort needed and people involved. Despite the recognized importance of producing high quality output from crowd-sourcing initiatives, few studies have compared alternate quality control processes in a large-scale project.

METHODS

Study Design

We use a two-pronged approach to better understand the effect of experience and quality control mechanisms at FSI. First, we perform an analysis of historical agreement data between independent transcribers (A and B) to understand the impact of experience on quality (i.e., agreement) and effort (i.e., time). Second, we perform a field experiment of a proposed peer review quality control process and compare it to the current arbitration model.

Data Sources

Data for the historical analysis and the field experiment was drawn from the database containing data associated with the FSI transcription software. The database includes the transcription text entered for each field (e.g., surname, given name, birthyear), name of the transcriber, her role (e.g., A transcriber, B transcriber, arbitrator), her native language, time spent transcribing, time spent idle (i.e., more than 30 seconds without any activity), number of keystrokes entered, and meta-data associated with the image being transcribed (e.g., language, collection).

The historical dataset includes all transcribed records from various FSI data collections (e.g., U.S. Census records, Canadian Census records) completed by February 2011. The field experiment dataset is based on a random sample of 2,000 images pulled from 25 of the 48 states that participated in the 1930 Federal U.S. Census for which data were available. Each census page includes up to 50 records, each of which describes an individual person and their information (e.g., surname, given name, birthplace). Images with less than half of the 50 rows completed were excluded to assure sufficient data.

A truth set transcription of the 2,000 images was created to calculate the accuracy of the two quality control mechanisms tested in the field experiment. A professional transcription company was hired to create a truth set with at least 99% accuracy. Two FamilySearch internal auditors reviewed the truth set. They tagged one row from 380 of the images (which included 18 fields each) and identified errors in the truth set. They found a total of 17 errors (e.g., Allan vs Allon; Evins vs Enins), which results in an estimated 99.75% accuracy. Errors in the field experiment were identified when a transcribed field did not match the truth set.

Historical Data Analysis

The goal of the historical data analysis was to understand the impact of a transcriber's experience and native language on quality and efficiency. Efficiency was measured as the total amount of non-idle time spent indexing a particular image. Native language is self-reported by users.

Users' experience was measured by totaling the number of images a volunteer had transcribed up until a given point in time. In February 2012, the top five transcribers of all time had each transcribed between 206,000 and 269,000 full images. In comparison, 26,459 of the nearly 400,000 transcribers only annotated a single image. To account for the skewed distribution, we used a log transformation of number of pages transcribed. For clarity, users are grouped into nine experience levels based on the following formula:

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

Where U represents the transcriber, N(U) is the number of images that U has transcribed, and EL(U) is the experience level of U. A volunteer who has transcribed only 1 page has an experience level of 0, while the top transcriber has an experience level of 8. The median number of images transcribed is 50, which equates to a skill level of 2.

Since no truth set exists for the large historical data corpus, accuracy is estimated by examining *agreement* between A and B transcribers. In contrast, the field experiment (described below) compares each transcription with the truth set to determine *accuracy*. An analysis of the 2,000 image sample showed that A and B agreement was strongly correlated with accuracy as compared with the truth set (when A and B agreed, the transcription matched the truth set 98% of the time on average for all fields in the dataset). This suggests that patterns observed in the historical data using agreement between A and B likely mirror patterns that would be observed based on actual accuracy.

Experimental Design

The goal of the field experiment was to evaluate three different quality control mechanisms as outlined below:

- A-B-ARB Condition: A and B transcribe a page independent of one another and any discrepancies are passed to a third arbitrator (ARB) who makes the final decisions. This is the currently used model at FSI. The FSI user interface has been optimized for this process. For example, when transcribers type in a location not on a pre-approved list it is highlighted to indicate it may be an error. Likewise, arbitrators have a special view that highlights fields that are different in the A and B transcriptions and allows them to quickly choose one or the other, or (as rarely happens) enter something else.
- A-R Condition: A transcribes a page and R reviews the transcription by identifying and correcting any errors he finds. The same FSI software was used to perform the review, with the original A transcription being flagged as "incomplete" so that R could have access to it. No specialized tools were developed to support peer review (e.g., no highlighting of suspected errors). Instructions were provided to volunteers on how to perform peer review, since this process was new to FSI volunteers. The newness of the peer review process to volunteers and the lack of optimized tools to support peer review likely underestimate the potential of this process. In other words, our findings provide a lower bound on the quality of peer review compared to arbitration.

- A-R-RARB Condition: A transcribes a page, R reviews the page, and any discrepancies are reviewed by a third arbitrator (RARB). Data from the A-R Condition above was passed to an arbitrator using the same arbitration software as is typically used. The arbitrator was presented with both A and R fields, without knowing which was which, since the standard arbitration tool does not give precedence to one transcription over another.

Recruitment and Limitations

Historical data from 2011 was used for the A-B-ARB condition. The original A transcription was used in the A-R and A-R-RARB conditions as well. However, FSI volunteers were recruited from five active U.S.-based user groups in January and February of 2012 to play the role of R and RARB. Each group had a known leader who helped notify their volunteers, point them to the instructions for the new peer review process, and answer any questions they had. Volunteers were asked to only complete a handful of pages each, to solicit a wide range of experience levels among volunteers and more independent assessments.

Despite our efforts to recruit a representative sample for the R and RARB roles, analysis of the experience distribution of our volunteers showed that R and RARB were on average less experienced than A, B, and ARB transcribers (see Figure 2). Even with the narrower spectrum of experience for R and RARB, there is evidence that R's experience is positively associated with higher accuracy. As a result, our A-R and A-R-RARB accuracy numbers may be lower than expected from a set of volunteers who directly matched the experience distribution of A-B-ARB volunteers. However, the distribution of R was spread out enough to see the expected positive relationship of higher experience and accuracy, though this was not true for RARB (see Field Experiment Results). For this reason, we control for experience of the original transcribers, arbitrators, and reviewers in our statistical models (see next section).

Some of the data fields in the B transcription data were not created via the typical transcription method, requiring us to make a special adjustment. Some fields, including surname and given name, were transcribed initially by a company, after which FSI volunteers transcribed the remaining fields.

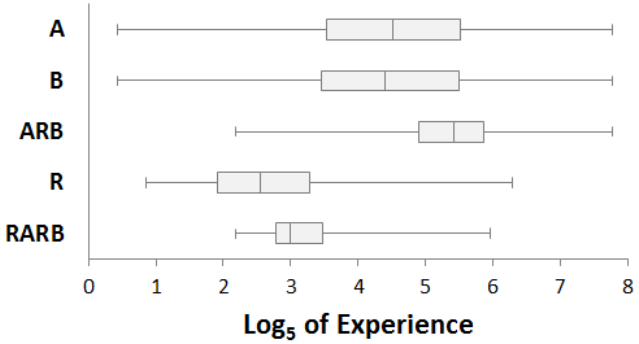

Figure 2. Box plots of experience distributions of A (n=1,550), B (n=1,588), ARB (n=1,010), R (n=344), and RARB (n=47)

The result was that for most fields, the B transcription was slightly, but consistently higher quality (compared to the truth set) than the A transcription which was created completely by FSI transcribers. We report the raw A-B-ARB results, as well as adjusted A-B-ARB results which estimate the amount of error that would have occurred had the B transcription been the same quality as the A transcription.

To estimate the adjusted A-B-ARB transcription quality, the following method was used. First, calculate how often there is a discrepancy between A and B and only one of them is correct. For example, for surname, A is correct 5.3% of the time when B is wrong, and B is correct 7.1% of the time when A is wrong, totaling 12.4%. Next, calculate how often ARB chooses the correct answer among the 2 options (e.g., 81% of the time for surname). Then, adjust B down to the same level as A (i.e., 5.3%) and multiply 10.6 (5.3% + 5.3%) by the percent of the time ARB is correct for this field (81%). This gives us an estimate of the adjusted value-added of ARB (8.7%) assuming that B is the same quality as A. Subtracting this from the actual value-added of ARB (10.1%) gives the adjustment amount (i.e., accuracy goes down by 1.4%). This approach is reasonable because ARB only reviews records where A and B disagree, ignoring records where A and B agree (whether or not their agreed value was correct).

Statistical Analysis

The historical data analysis is descriptive, while our analysis of the field experiment uses a statistical model to identify significant differences between experimental conditions. The dependent variable in each of our models is the volunteer's accuracy, a binary variable that indicates whether the volunteer's transcription matched the truth set or not. We measure a match in 3 different ways as described in the following section. Our independent variables include the experimental condition (described below) and the experience of the volunteer. The experience variable was used to control for possible differences among volunteers in different conditions. This is particularly important given the difference in expertise levels between the conditions as discussed earlier and in Figure 2.

The data has a complicated structure that violates key assumptions of standard t-test or OLS regression. Volunteers transcribed an entire census page (i.e., image) of records at a time, each of which included up to 50 records in what is called a batch. Batches of records were transcribed by several individuals playing the role of A, B, ARB, R, and RARB. Thus, the independence assumption needed for simpler analyses was unreasonable because records on the same census page have more highly correlated accuracy rates than records on different pages (due to the common handwriting of the census taker, repeated surnames and placenames, etc.). Also, records transcribed by the same person have more highly correlated accuracy rates than those transcribed by different people. To account for these unique properties, we used a mixed model [6] with random

effects that included records nested in batches and cross-nested between two transcribers (or a transcriber and a peer reviewer) and an arbitrator.

Each model was used to estimate or compare accuracy rates of different quality control processes (A vs A-R; A vs A-B-ARB; A-R vs A-B-ARB; and A-R vs A-R-RARB) controlling for the expertise of the volunteers as well as the random effects for batch, transcribers (or reviewer), and arbitrator. Expertise was measured, as mentioned before, by a log transformation of the number of batches that had been transcribed by an individual up to the time of their transcription of the experimental records. For A, B, and ARB this was February, 2011. For R and RARB this was February, 2012 since their data was collected a year later. Each model was run separately for each field (surname, birthplace, etc). We used a logistic regression model because we were testing accuracy rates with an outcome of each model as a 1 (correctly transcribed) or 0 (incorrectly transcribed).

Accuracy Measures

Agreement between two transcriptions (e.g., A and the truth set) was calculated in 3 ways for the field experiment:

- Exact Match: A standard string comparison, where each string (including punctuation and spaces) had to exactly match the other one ("Washington D.C." did not match "Washington DC")
- Known Variants: Matches were counted in cases where the transcriptions were exact matches or effectively equivalent as determined by a 1930 Census expert working for FamilySearch. These included obvious abbreviations (e.g., W in the Race column matched with White; "Wash DC" in a location column for "Washington DC" and other variants with periods), extra spaces in names (e.g., "Mc Arthur" matched "McArthur"), and unnecessary details added by the census taker (e.g., "Montgomery County" matched "Montgomery"). The equivalency matches were developed by looking at the most common errors.
- Authorities Variants: Matches were counted in cases where the transcriptions were known variants or where a surname, given name, or place-name was identified as equivalent in the authorities tables used by FamilySearch's search engine. These tables have been iteratively developed over 40 years by FamilySearch, primarily from observed variations where there were multiple records which seemed to refer to the same person but with name/place differences. For example, Katia and Katherine match. The authorities variants are not statistically motivated, so there can be names that occur with relatively high frequency but which are lumped together because of some limited number of variants (e.g., Mike, Michael, and Chiel, which are translations of each other).

Transcribed Field	Agreement
Gender	98.8%
Given Name	82.5%
Surname	74.7%
Birthplace	96.1%
Relationship to Head of House	95.0%
Age	91.6%
Birth Date	97.8%
Father's Birthplace	96.7%
Mother's Birthplace	96.7%
Immigration Year	90.0%

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

HISTORICAL ANALYSIS RESULTS

A-B Agreement Analysis

Agreement between A and B varied dramatically depending upon the field of data that was transcribed. Table 1 shows the percent agreement for each field in the U.S. Censuses from 1850 to 1920. Other collections included similar differences between fields. As expected, fields with few possible values had extremely high agreement (e.g., gender = 99%), while fields with many possible values had much lower agreement (surname = 75%).

Differences in accuracy were also observed based on the language of the census taker. The Canadian 1871 Census data used the exact same census forms to record data in both English and French, allowing for a comparison. Agreement was much higher for English-language transcriptions than for French-language transcriptions. Comparing just the given name and surname fields, 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 nearly all transcribers are native English speakers.

Finally, results from our analysis of agreement and prior experience of the transcribers is presented. The heatmap presented in Figure 3 shows agreement levels between all possible experience matchups. The lowest agreement values (shown in blue) occur when inexperienced transcribers are matched up with other inexperienced transcribers (in the bottom-left-region of the graph), while the highest agreement occurs between experienced transcribers (in the upper-right region of the graph). Though the general trend is not surprising, the continued improvement even at the very high levels is worth noting. It is consistent with lab experiments conducted by cognitive psychologists who show continued improvement at tasks even at extremely high levels of practice, though with diminishing returns [1].

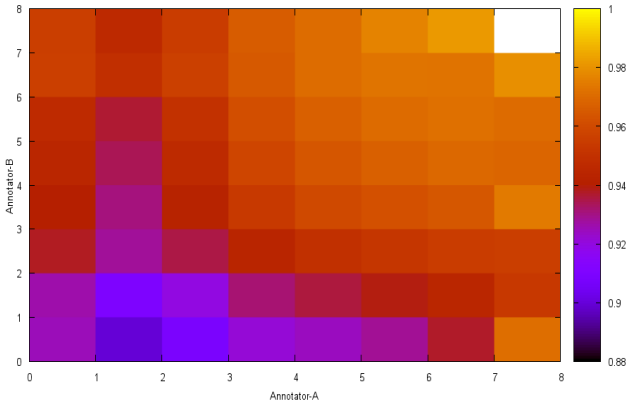

Figure 3. Heatmap of A-B agreement by Experience Level for Birthplace in all US Censuses in our corpus.

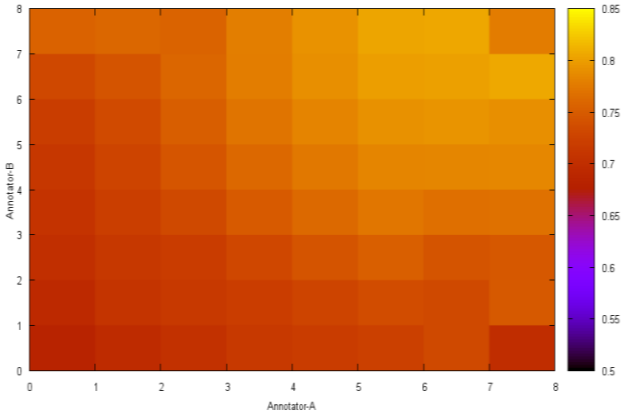

Figure 4. Heatmap of A-B agreement by Experience Level for Surname in all US Censuses in our corpus.

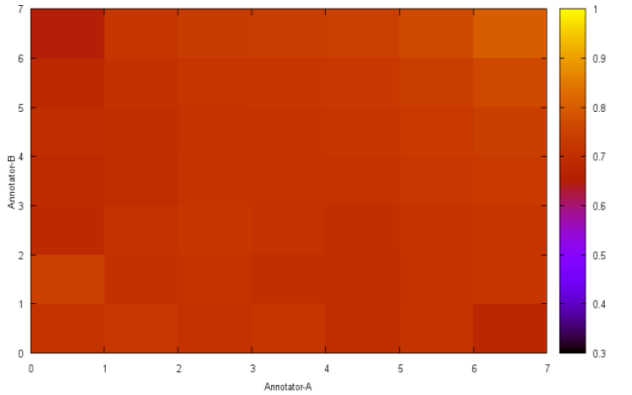

Figure 5. Heatmap of A-B agreement by Experience Level for Birthplace in the English-speaking Canadian Census.

This pattern of continuous improvement was found in other fields such as surname (Figure 4) and given name. Improvements in agreement for gender were noticeable but very small since agreement was so much higher for all experience levels. Agreement for birthplace did not improve for certain other datasets including the English-speaking Canadian Censuses (Figure 5). For this field, unlike with gender, the agreement remained relatively low. It is likely that the predominantly U.S. transcribers were unfamiliar with Canadian places, contributing to this effect. This suggests that transcription expertise is partially tied to domain knowledge, and that expertise in one domain (e.g., U.S. locations) does not necessarily translate into expertise in another domain (e.g., Canadian locations).

Transcription Time Analysis

Data on non-idle time and keystrokes was not available for specific fields, since the time was based on completion of an entire image. There are cases where some transcribers may be completing other tasks on their computer concurrently or take frequent breaks while working on projects, which would skew time results. For this reason outlier times that were in excess of two hours or less than 20 seconds were discarded.

The number of keystrokes and the time spent decreased for those with more experience (Table 2). The estimated average of keystrokes per line for the US records in our dataset had a small but consistent downward trend, except for those transcribing their very first record (Experience Level 0 people who were given easy batches to get them acquainted with the system). Since the same amount of data needed to be transcribed on each line irrespective of experience level, it is likely that experienced transcribers revised their entries less often (e.g., re-type in a surname after realizing they got it wrong the first time).

Changes in time per line (measured in seconds) are far more dramatic. Experienced volunteers 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 FSI, is double the weighted average of the three highest experience levels (6-8). This lowers the average Time per Keystroke.

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

Table 2. Time and Keystroke data by Experience Level for all US Censuses in our corpora

Experience Level of Transcriber	Avg Key-strokes of ARB	Avg Time of ARB	Avg Time per Key-stroke
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

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

Differences in time and keystrokes based on language were also observed. The English-speaking 1871 Canadian Census was 2.68 seconds faster per line than the French version, despite the fact that it required more keystrokes. Though 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 a year's work by one full-time employee. This difference was likely due to a lack of native French-speaking volunteers.

As expected, arbitrators spent more time reviewing content transcribed by novices. Table 3 shows this using data from the 1910 US Census. All US Censuses were not combined in this analysis because each Census had a different number of columns and rows. Because there were fewer errors to correct in records transcribed by experienced users, fewer keystrokes and less time were required. The average time per keystroke was not dramatically different, though there was a slight upward trend as experience levels increased. This was likely because cases where experienced volunteers differed (particularly amongst themselves) were often the hardest cases that required more time to evaluate.

FIELD EXPERIMENT RESULTS

Accuracy Analysis

Consistent with findings from the historical analysis of A and B agreement, there were dramatic differences between accuracy across fields, with the lowest accuracy in surname and given name fields (Table 4). This pattern is consistent for all 3 quality metrics (exact match, known variants, and authorities variants) and experimental conditions.

Surprisingly, the impact of the additional arbitration (RARB) of the peer reviewed content (R) failed to improve the quality. No statistically significant results were found at the .05 level. Furthermore, arbitration of peer reviewed content actually lowered the accuracy for all fields, though only slightly. In other words, when presented with A and R discrepancies RARB chose the wrong one more times than they chose the right one. Since the majority of R's changes improved accuracy (54-86% depending on field), it appears that RARB chose A instead of R too often. As mentioned earlier, the RARB volunteers were not as experienced as expected, though they were more experienced than R (see Figure 2). Even so, the fact that R's edits were mostly improvements suggests that the need for RARB does not seem justified by the extra effort required. For this reason we now focus on A-B-ARB and A-R quality control models.

Results for the Known Variants quality metric are shown in Figure 6. They show lower error rates than those based on Exact Match (Table 4), since field values such as "Washington D.C." match with known variants such as "Washington DC". Both quality control mechanisms (A-B-ARB and A-R) resulted in statistically significant reductions to A's original error rates ($p<.001$ for all fields). For most fields, A-B-ARB (even with the adjustment) outperforms A-R, though only 3 of the fields found the improvement to be statistically significant at the .001 level (see Figure 6).

Another key observation relates to the two main types of error: missed error and introduced error. Missed error refers to errors that were made in the original dataset (A) that the quality control mechanism did not catch. These make up the vast majority of errors. Introduced errors refer to the errors that the quality control mechanism introduced. In other words, A had it right, but the quality control mechanism changed the correct answer into an incorrect answer. In nearly all cases the A-R process introduces fewer errors than the A-B-ARB process, though R also missed more errors than B-ARB. As mentioned earlier, no customized tools were provided to R in order to help identify potential errors, so the R results should be interpreted as a minimum threshold of quality.

	A	ARB	R	RARB	N
Surname	82.2%	87.0%*	84.2%	83.9%	97,600
Given Name	84.0%	89.1%*	86.5%	86.1%	98,496
Age	96.9%	98.2%*	98.3%	97.7%	98,402
Relationship	97.9%	99.3%*	98.6%	98.2%	98,188
Race	99.8%	99.9%*	99.9%	99.8%	98,516
Birthplace	97.8%	98.6%*	98.4%	98.0%	98,512
F_Birthplace	96.5%	97.3%	97.0%	96.6%	98,499
M_Birthplace	96.5%	97.3%	97.0%	96.7%	98,497
Gender	99.2%	99.7%	99.5%	99.2%	98,517
Marital	99.2%	99.7%	99.5%	99.1%	98,316

Table 4. Unadjusted Exact Match accuracy by field for A, ARB, R, and RARB. The B transcription was transcribed by a company and B for those fields with *s.

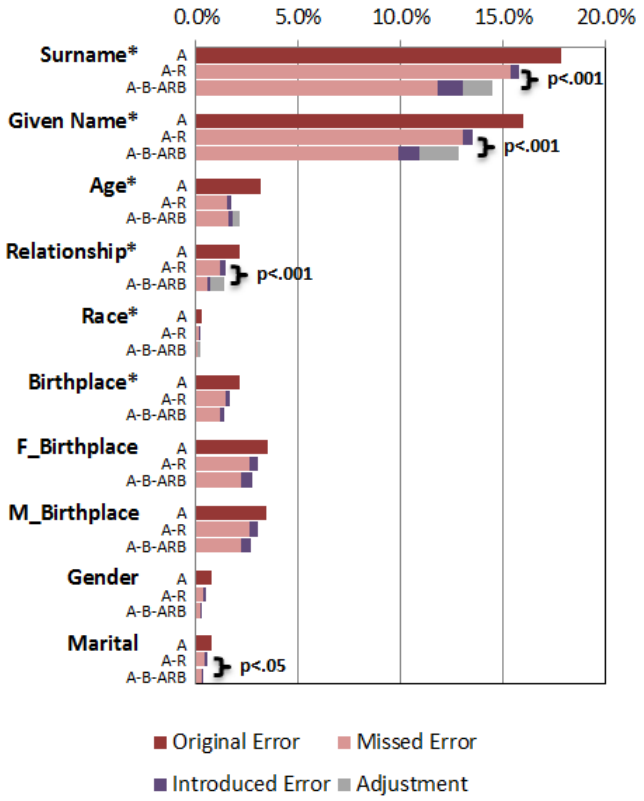

Figure 6. Errors based on Known Variant matching by field and quality control mechanism. Fields with a * require an adjustment (shown in gray) because B was transcribed by a company and B rather than just B volunteers.

Adding Known Variant matches to Exact Matches made the biggest difference in location-based fields (e.g., birthplace fields), where the error rate was reduced by just over a third. Adding Search Variant matches reduced surname error rates by11% for ARB and R, and given name error rates by 21% for ARB and R. Changes to location-based fields were negligible.

The statistical models allowed us to parse the variation in accuracy rates between image, transcriber(s), reviewer (if present), and arbitrator. Using this data, it was possible to identify which factors contributed the most to the variability in accuracy. Table 5 displays the upper and lower bounds based on the estimated random effects of the model for the surname field. These bounds were found by taking two standard deviations above and below the log-odds mean and translating those values into percentages (since it was not a linear transformation the interval was not centered around the mean). The values represent reasonable upper and lower bounds in a category assuming other categories contributed their average effect.

For example, for a typical transcriber (A), reviewer, and arbitrator, a particular image could end up with 50.8% accuracy at the low end or 98.7% accuracy at the high end.

	Lower Bound	Upper Bound
Image	50.8%	98.7%
A	81.5%	94.9%
Reviewer	88.8%	91.2%
Arbitrator	87.7%	92.0%

Table 5. Lower and Upper bounds for different sources of variation in accuracy of the surname field. These bounds assume an average result for the other sources of variation.

This a dramatic effect, likely driven by factors such as differences in handwriting of the census taker and the complexity of surnames and placenames that show up together in the same image. In contrast, with an average difficulty image, the best transcribers would get 94.9% accuracy while the worst would get 81.5% accuracy. Overall, the image is by far the most important driver of accuracy variance followed by the transcriber. The effect of reviewer and arbitrator are much smaller.

The effect of experience on the review process is consistent with findings from the historical analysis. For instance, the experience of the initial transcribers and the reviewers is significantly related to the accuracy of most fields, but the effects, although significant, are relatively small. Because most of the fields are transcribed with high accuracy, comparing experienced transcribers to novice transcribers gives a difference in accuracy of less than 1% – even when the effect is significant.

The effect of experience on more difficult fields (surname and given name) is more dramatic. For example, the effect (in log odds) on the surname field is .018 per unit increase in the log of experience. A volunteer who has completed 22,000 batches has about a 1.9% increase in accuracy over a person who only transcribed 9 batches. The effect of experience on accuracy is about 4% for the same comparison when looking at the given name field. The experience of a reviewer has about the same effect on accuracy as the experience of the initial transcriber. We did not find evidence that the experience of arbitrators matters, perhaps because all arbitrators must meet a minimum level of experience and expertise to be accepted in that role.

Time Analysis

As expected, A-R (and A-R-RARB) required significantly less time to complete than the A-B-ARB process. Overall, peer reviewing a document takes about half as long as transcribing a document from scratch. The end result is that the A-R process takes about 70% as long as the A-B-ARB process (Figure 7), which is significantly different at the .001 level. Data based on medians instead of averages was similar, though faster (e.g., Median for A-B-ARB process is 61 min; Median for A-R-ARB process is 46 min).

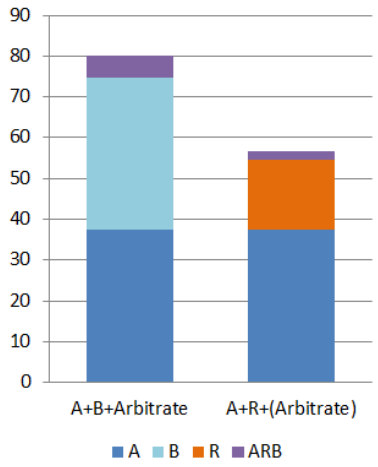

Figure 7. Average minutes per page for different quality control mechanisms.

DISCUSSION & DESIGN IMPLICATIONS

Below is a discussion of our findings and their design implications. Though we focus on improving transcription work, many of our suggestions have natural applications to other crowdsourcing endeavors such as image labeling [20], truth set creation [7], and database entry (e.g., IMDB) [5].

The role of Expertise on Quality & Efficiency

Both our historical and experimental data showed significant differences between experienced and novice volunteer efforts. Not only did quality increase for more experienced transcribers, but the time to complete the work was much less. As is typical of crowdsourcing projects [27], the distribution of expertise and the amount of work contributed is highly skewed with the majority of contributions made by a minority of highly experienced contributors. The findings suggest the importance of retaining experienced members and helping motivate less experienced members to contribute more (see [16,17] for a discussion of design strategies to promote contributions).

However, the historical analysis also gives us reason to be cautious about how prior transcription experience does and does not impact current tasks. Overall experience was not strongly related to fields such as Canadian placenames, which are likely unfamiliar to the largely U.S. based volunteers. This may indicate the importance of the transcribers having some contextual knowledge (e.g., knowledge of Canadian placenames) that may indirectly benefit them when performing transcription work. This knowledge may be completely independent of their transcription experience (e.g., ability to read cursive handwriting and use the FSI system). Tools that leveraged people's contextual knowledge by, for example, intelligently recommending images to transcribe based on a volunteer profile, may improve quality. Additionally, tools that helped volunteers gain contextual knowledge (e.g., by providing county maps and sending sets of images from the same county) could improve quality. Though not tested in this paper, it is likely that those who work extensively with Canadian Census records would come to know the Canadian placenames and improve over time. Likewise, those working with certain languages will likely improve over time, though recruiting native speakers to transcribe within their own language seems essential given our findings on language accuracy.

Systems, such as the new FSI mobile application or reCAPTCHA [26], which allow users to transcribe single records (as opposed to entire images) should take this into consideration. If they serve a user randomly selected images, the user's prior experience is less likely to benefit their future performance. Conversely, if they serve a user similar fields (e.g., placenames) from the same collection (e.g., Canadian census), that user's experience will help him with future transcriptions. More generally, a system that breaks down large heterogeneous tasks into their constituent subtasks (i.e., a census image broken into specific field entries) can promote the development of specialist experts who are served up similar subtasks. This will likely improve quality as well as reduce time to complete, though its effect on user satisfaction is unclear. Future work should examine learning effects in crowdsourcing environments more directly. Interventions designed to more rapidly improve expertise could be developed, such as sending novices common errors identified through arbitration or peer review.

Improving Peer Review

One of the most surprising results was the lack of a need for arbitration of peer reviewed content (i.e., the lack of benefit provided by RARB). Peer reviewers seemed to make changes only when they were certain that their change would fix a mistake, as shown by the low number of introduced errors by reviewers. Adding a verification step (as proposed in the Find-Fix-Verify pattern [3]) was unnecessary in the FSI context, since it increased the amount of time without improving quality.

Our findings related to strict peer review (A-R) are mixed. On the one hand, it took considerably less time to complete than the A-B-ARB method. On the other hand, the resulting quality was, in some cases statistically inferior to the A-B-ARB method. As discussed earlier, the peer review data from our study likely understates the quality of the mechanism because (a) the tools are not customized for reviewing, while they are customized for A-B-ARB, (b) the peer review process was new to the volunteers, while the A-B-ARB process was not, and (c) the average experience levels of the reviewers was low due to a selection bias. Given these factors, our findings suggest that peer review should be carefully considered as a viable alternative to the current A-B-ARB model, particularly if steps are taken to improve the tools and experience of reviewers.

Several techniques could be used to improve the quality of peer review, all of which could be empirically tested in future work. First, reviewers could be provided with an estimated number of errors per page (or field) in order to set

their expectations within a reasonable range. The high variability in quality among pages and initial transcribers suggests that such estimates should be customized based on the page and expertise of the original transcriber.

Second, tools could highlight entries that have a high probability of error. The current system does this for transcribers by literally highlighting entries that do not match other entries in existing databases (e.g., county names or surnames that are not in an authorities table). This is a good start, but more sophisticated techniques may prove useful. For example, prior A-R data could be used as a training set to help identify potentially erroneous entries (see discussion in [7] of various techniques). Alternatively, the original transcriber could flag items that they have low confidence in. This allows for the decoupling of the identification of potential problems with the provision of the solution, which was found to be a useful strategy in the Find-Fix-Verify pattern [3] (though without the heavy overhead of including as many people in the process). It may be that novices can accurately identify difficult records, even if novices can't accurately transcribe them. While these approaches may slow down the review process, they have the potential to pay off in increased quality.

Third, entries that are likely to contain errors (based on the earlier techniques mentioned) could be routed to more experienced reviewers, while entries estimated to be accurate could be routed to novice reviewers. Caution would need to be applied in this approach, since it may be in conflict with the goals of training newcomers to become more expert (see prior section).

Alternate Quality Control Mechanisms

Though A-B-ARB and A-R both provide significantly better results than using a single transcription, there is significant room for improvement for difficult fields such as surname and given name. Our data provides some hints at alternate mechanisms that could be used to improve quality without dramatically increasing the amount of effort.

One alternative quality control method would be double review where an original transcriber is reviewed by one person and the updated file is passed on to another reviewer for a final pass. Our data shows that reviewing a completed transcript takes about half as long as transcribing one from scratch. Thus, an A-R_1-R_2 process would take about the same amount of time as the current A-B-ARB process. Yet, a single review pass is almost as good as the A-B-ARB process, so it may prove that a second round of review would continue to improve the accuracy. This seems likely, since, for some fields (e.g., surname, given name), there are many missed errors left for the second reviewer to catch. It may make sense to assure that the second reviewer is an expert.

The A-R_1-R_2 method could be combined with other recommendations mentioned earlier. For example, the first reviewer could flag fields that they want the second reviewer to review. Furthermore, some fields with a low probability of error (e.g., gender) could skip the second stage of review to improve efficiency. Indeed, even the first round of review could be skipped for some fields if there is a high enough probability that the original transcriber was accurate (based on her expertise and/or machine learning-based predictions like those used in other contexts [7]).

Another promising approach, which could provide increased efficiency, is to blend human transcription with algorithmic transcriptions (e.g., [24]). A growing body of literature uses image processing techniques to automatically transcribe historical texts [2]. Though accuracy of interpreting cursive writing in documents such as U.S. Censuses is not nearly as good as human transcription, for some fields with a limited set of entries (e.g., gender, relationship) it is becoming viable [24]. Automatic transcription could be used in several ways. It could perform the initial transcription that is reviewed by a human. It could act as the reviewer. Or, it could help intelligently determine if human peer review is needed on a given entry.

One exciting possibility would be the creation of a large-scale machine learning training dataset, which could be developed as part of the standard workflow. For example, if automatic transcription algorithms served as the original transcriber (A), they could be reviewed by humans as part of the A-B-ARB, A-R, or A-R_1-R_2 quality control process. Data on the correct and incorrect transcriptions could be fed back into learning algorithms, helping to improve future (or even prior) classifications. Such a process could create a training set of millions and possibly billions of records, orders of magnitude larger than existing sets. While this approach would not be based on a truth set, for many of the simpler fields, it would create a training set that is above 95% accurate. If successful, such approaches could increase the rate of transcription by orders of magnitude.

Another approach that could augment existing quality control mechanisms would be to allow users of the transcriptions to make corrections. Like a wiki, the existing transcription would become a living document that is constantly updated. While this introduces more opportunities for vandalism, it also provides people with the most content knowledge (e.g., the grandson of a person in a census record) the ability to fix it, an approach that helps sites like Wikipedia [22]. One difficulty associated with this approach would be dealing with the many definitions of accuracy. In this paper we focused on accurately transcribing what was written on the historical page. However, those writing the historical documents were not perfect, particularly in time periods when spelling norms were not established. Allowing users to fix, or annotate, direct transcriptions would help genealogists and historians in their efforts to identify individuals whose names may be written in several different forms.

Finally, models that made transcription more social and intrinsically rewarding may motivate more people to partic-

ipate. Currently, transcription work is an isolating experience, where tools outside of the existing software must be used to get help. This is contrary to the highly collaborative work that genealogists enjoy. Allowing users to request help on a difficult record from an expert, or even their Facebook friends may add enjoyment, speed up the learning process, and attract new volunteers. Making the transcription task part of an intrinsically rewarding game may also attract a different audience. There is precedence in games that help identify optical character recognition errors (i.e., scannos) [4] and image recognition errors [11]. Additionally, adding a more explicit community component may also allow volunteers to take a more active role in improving the process as a whole, which is currently dictated almost entirely by FSI. This could move the project into a "heavyweight peer production" environment, which raises new challenges and opportunities [12].

CONCLUSION

This paper has provided an in-depth look at quality control mechanisms within FamilySearch Indexing, one of the world's largest crowdsourcing projects. Analysis of historical agreement data showed that experienced transcribers are more likely to agree than novices, as are those working on English language transcriptions compared to other languages. Furthermore, experts take less time, suggesting the need to retain experts and train novices as opposed to treating all users as interchangeable.

A field experiment comparing the existing arbitration-based process (A-B-ARB) and two proposed peer review processes (A-R and A-R-RARB) showed that peer review took dramatically less time, though it did not achieve quite as high accuracy. This may be in part due to the lack of customized tools to support peer review and other limitations in our experimental data. Interestingly, arbitration (i.e., validation) of the peer reviewer's edits did not increase quality, suggesting that a simple A-R process is preferable in this context. Tools customized to better support peer review (e.g., help reviewers identify potential errors or set their expectations) could likely lead to more efficient quality control that achieves comparable or better quality over existing methods. Alternate models such as double peer review (A-R_1-R_2), integration of human and machine transcription, and intelligent routing of difficult records may improve quality and efficiency above existing golden standard approaches.

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 field experiment.

REFERENCES

1. Anderson, J. *Cognitive psychology and its implications*, 7th edition. Worth Publishers, 2009.
2. Barret, B., Brown, M.S., Manmatha, R., and Gehring, J. *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, ACM Press (2011).
3. Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. Soylent: a word processor with crowd inside. Proc. UIST '10, ACM Press (2010), 313-322.
4. Chrons, O. and Sundell, S. Digitalkoot: Making old archives accessible using crowdsourcing. *Human Computation: Papers from the 2011 AAAI Workshop*, (2011).
5. Cosley, D., Frankowski, D. Kiesler, S., Terveen, L., and Riedl, J. 2005. How oversight improves member-maintained communities. In *Proc. CHI 2005*, ACM Press (2005), 11-20.
6. Demidenko, E. *Mixed Models: Theory and Applications*. Wiley-Interscience, 2004.
7. Dligach, D., and Palmer, M. 2011. Reducing the need for double annotation. In *Proc. of the 5th Linguistic Annotation Workshop* (LAW V '11), (2011), 65-73.
8. Doan, A., Ramakrishnan, R., and Halevy, A. Y. Crowdsourcing systems on the World-Wide Web. *Commun. ACM* 54, 4 (April 2011), 86-96.
9. Duguid, P. 2006. Limits of self-organization: Peer production and "laws of quality". *First Monday*, 11 (10).
10. Estellés-Arolas, E., and González-Ladrón-de-Guevara, F. Towards an integrated crowdsourcing definition. *J. of Information Science*, 20, 10 (2012), 1-14.
11. Hansen, D., Jacobs, D., Lewis, D., Biswas, A., Preece, J., Rotman, D., and Stevens, E. Odd Leaf Out: Improving visual recognition with games. In *Proc. SocialCom '11*, IEEE (2011), 87-94.
12. Haythornthwaite, C. Crowds and Communities: Light and Heavyweight Models of Peer Production. In *Proc. HICSS '09*, (2009), 1-10.
13. Howe, J. Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business http://www.crowdsourcing.com/
14. Hsueh, P., Melville, P., and Sindhwani, V. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proc. NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing (HLT '09)*, (2009), 27-35.
15. Kittur, A., and Kraut, R. E. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proc. CSCW '08*, ACM Press (2008), 37-46.
16. Kraut, R.E. and Resnick, P. Building Successful Online Communities: Evidence-Based Social Design. The MIT Press, 2012.
17. Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Li, X., Cosley, D., Frakowski, D., Terveen, L., Rashid, A.M., Resnick, P., and Kraut, R. Using social psycholo-

gy to motivate contributions to online communities. *JCMC*, 10, 4 (2005).

18. Little, G. and Sun, Y. Human OCR: insights from a complex human computation process. *Workshop on Crowdsourcing and Human Computation, Services, Studies and Platforms, ACM CHI*, (2011).
19. McCann, R., Doan, A, Varadajan, V., and Kramnik, A. Building Data Integration Systems via Mass Collaboration. In *International Workshop on the Web and Database*, (2003).
20. Nowak, S., and Rüger, S. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proc. MIR '10*. ACM Press (2010), 557-566.
21. Quinn, A. J. and Bederson, B. B. Human computation: a survey and taxonomy of a growing field. In *Proc. CHI 2011*, ACM Press (2011), 1403-1412.
22. Stvilia, B., Twidale, M., Smith, L. C., Gasser, L. Information quality work organization in Wikipedia. *JASIST*, 59(6), (2008), 983–1001.
23. Sun, Y., Roy, S., and Little, G. D. Beyond independent agreement: a tournament selection approach for quality assurance of human computation tasks. *Papers from the 2011 AAAI Workshop on Human Computation*, (2011).
24. Toselli, A.H., Romero, V., and Vidal, E. Rodriguez, L. Computer Assisted Transcription of Handwritten Text Images. *Proc. ICDAR '07*. IEEE (2007), 944-948.
25. von Ahn, L. Games with a purpose. *Computer, 39*(6), (2006), 92-94.
26. von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. reCAPTCHA: Human-based character recognition via web security measures. *Science* (2008),1465-1468.
27. Wilkinson, D.M. Strong regularities in online peer production. In *Proc. EC '08*. ACM Press (2008), 302-309.