Drafty: Enlisting Users to be Editors who Maintain Structured Data

Shaun Wallace
Department of Computer Science
Brown University
shaun_wallace@brown.edu

Lucy Van Kleunen
Department of Computer Science
Brown University
lucy_van_kleunen@brown.edu

Marianne Aubin-Le Quere
Department of Computer Science
Brown University
marianne_aubin_lequere@brown.edu

Abraham Peterkin
Department of Computer Science
Brown University
abraham_peterkin@brown.edu

Yirui Huang
Department of Computer Science
University of Toronto*
**Work performed at Brown University*

Jeff Huang
Department of Computer Science
Brown University
hcomp@jeffhuang.com

Abstract

Structured datasets are difficult to keep up-to-date since the underlying facts evolve over time; curated data about business financials, organizational hierarchies, or drug interactions are constantly changing. Drafty is a platform that enlists visitors of an editable dataset to become "user-editors" to help solve this problem. It records and analyzes user-editors' within-page interactions to construct user interest profiles, creating a cyclical feedback mechanism that enables Drafty to target requests for specific corrections from user-editors. To validate the automatically generated user interest profiles, we surveyed participants who performed self-created tasks with Drafty and found their user interest score was 3.2 higher on data they were interested in versus data they had no interest in. Next, a 7-month live experiment compared the efficacy of user-editor corrections depending on whether they were asked to review data that matched their interests. Our findings suggest that user-editors are approximately 3 times more likely to provide accurate corrections for data matching their interest profiles, and about 2 times more likely to provide corrections in the first place.

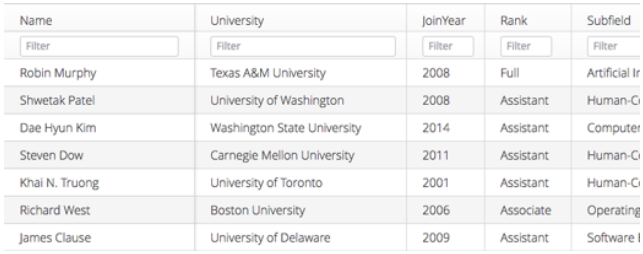
Introduction

Structured data represents information that can be in a constant state of change and needs to be updated to stay accurate and relevant. Popular examples of such structured datasets include company management information Crunchbase (Crunchbase Inc. 2007)), legal filings, soccer player career history (e.g. Crowdfill (Park and Widom 2014)), sponsored funding opportunities, socio-economic and law enforcement data (e.g. Communities and Crime Dataset (Redmond and Baveja 2002)), and country-level health statistics. There are many reasons data may change: new events occur, personnel changes, or existing data is revised. Without proper upkeep, these structured datasets degrade over time. This data often requires more monitoring and upkeep than a single individual can handle.

Our research combines the science of crowdsourcing with mining subtle forms of human-data interactions to create a sustainable human-in-the-loop system for maintaining structured datasets over time. Take as an example a spreadsheet of computer science professors with their academic background information or Crunchbase which lists start-up companies and their investment history (Crunchbase Inc. 2007). In both cases, the data rapidly expires and constantly needs updating to be kept relevant. However, traditional crowdsourcing efforts that rely on platforms such as Mechanical Turk are insufficient in several ways. First, crowd workers may not be familiar with the specialized vocabulary used in start-ups or academia. Crowd workers do not have the necessary context to provide accurate information (Papoutsaki et al. 2015), as revisions to this dataset require domain-specific knowledge. Second, systems powered by Mechanical Turk for upkeep of structured data pose long term financial commitments that might not be sustainable. Crowdfill (Park and Widom 2014) is an example of such a system that relies on crowd workers to successfully maintain structured datasets.

Drafty builds upon the fundamental ideas of Crowdfill in several ways by empowering its users. Drafty does not rely on paid crowd workers, but instead relies on *drafting* the regular users who visit the data. Drafty empowers the user by enabling them to edit the data using a Wikipedia-like model to score and weight suggested data points. Hence, users within the context of Drafty are referred to as "user-editors." Most importantly, Drafty is designed to harness the interests of its user-editors to maintain a structured dataset. It builds user interest profiles to target data points for user-editors to review and correct that match their interests. There has been research supporting the idea that crowd workers are more willing to do higher quality work with better retention if the task is relevant to their interests (Clauset, Arbesman, and Larremore 2015).

A user interest profile is automatically constructed for each user-editor based on their interactions with Drafty's interface. These interactions are collected remotely without disturbing the user (e.g., using methods from (Huang, White, and Dumais 2011)). We hypothesize that an interaction-based profiling approach provides higher quality data and reduces long term costs for maintaining structured datasets. As part of the system design of Drafty, the research challenge focuses on how to construct user interest profiles to use them to match user-editors to data they could best review.



Name	University	JoinYear	Rank	Subfield
Filter	Filter	Filter	Filter	Filter
Robin Murphy	Texas A&M University	2008	Full	Artificial In
Shwetak Patel	University of Washington	2008	Assistant	Human-C
Dae Hyun Kim	Washington State University	2014	Assistant	Computer
Steven Dow	Carnegie Mellon University	2011	Assistant	Human-C
Khai N. Truong	University of Toronto	2001	Assistant	Human-C
Richard West	Boston University	2006	Associate	Operating
James Clause	University of Delaware	2009	Assistant	Software I

Figure 1: Screenshot of the Drafty interface populated with profiles of computer science professors.

The research contribution is twofold: (1) we describe the implementation of Drafty, a platform that can host datasets that are self-sustaining, capturing and interpreting the footprints of user clicks, text highlighting, searches, and column sorting so others can host their own data or build off our work, and (2) we validate the ability of automatically-generated user profiles to reflect user interest, and show that users who are asked to review data that match their interests are more likely to volunteer and also provide more accurate suggestions.

While Drafty is a data-agnostic platform, for demonstration and experimentation, it was populated with a dataset consisting of over 50,000 data entries that are used to build academic profiles of over 3,600 computer science professors from across the USA and Canada[1]. This data came from Mechanical Turk crowdworkers recruited to build the dataset as part of an assignment in a Human-Computer Interaction seminar. Each professor's academic profile included: their affiliated university, the year they joined as faculty, their rank, subfield area of expertise, where they received their Bachelors, Masters, and Ph.D. degrees, where they did their PostDoc, a link to a profile photo, and links to sources for the aforementioned information types. This dataset received review and corrections by more than 50,000 visitors from 2014 to 2017. Hence, we believe the current quality and maturity of the dataset is representative of other structured datasets. As part of our contributions, we release Drafty and its dataset as an open source platform[2] so that curators of popular datasets can use it to maintain their own data, and other researchers can replicate or build off our work.

Related Work

Drafty builds on the foundational work in four areas: crowdsourcing, peer producton, learnersourcing, and recommender systems.

Crowd Powered Systems

Jeff Howe coined the term crowdsourcing (Howe 2006) as "an umbrella term for a highly varied group of approaches that share one obvious attribute in common: they all depend on some contribution from the crowd." (Howe 2008). He also categorized four main applications of crowdsourcing: 1) crowd wisdom or collective intelligence, 2) crowd creation or user-generated content, 3) crowd voting and 4) crowdfunding. Drafty is a combination of the first three categories, using the collective intelligence of its users by targeting them to review data based upon their interests while allowing them to add or edit new data. It empowers the crowd through an intelligent system that weights and scores user contributions to select the final data to display.

Traditional crowdsourcing platforms, such as Amazon Mechanical Turk, support an ecosystem where requesters can post micro-tasks for a number of use cases, such as data collection and verification. While this is practical to generate an initial dataset (Papoutsaki et al. 2015), the long term upkeep of the dataset can prove exponentially difficult. Crowdfill (Park and Widom 2014) is a crowdsourced system that maintains structured datasets using the microtask-based approach of Amazon Mechanical Turk. Rather than giving each worker a set of tasks to complete, workers are presented with one shared table of data which they can fill in however they want. Workers can also rate data submitted by other workers. This approach plays to the individual strengths of the workers and results in higher-quality submissions. In Crowdfill, Park and Widom mention that the system could potentially be improved by automatically recommending certain cells to individual workers based on their skills. Drafty builds upon this idea, matching unpaid volunteers to fix data that match their interests. Crowdfill's model, and other similar systems such as Wisteria (Haas et al. 2015b), may be hard to sustain long term. Long term repeatable tasks for maintaining data through Amazon Mechanical Turk are impractical due to increased cost, time, and accuracy risks (Mason and Watts 2010). The tasks and costs to employ crowd workers increase as the dataset grows larger. Drafty is a platform that does not require the continuous use of paid crowd workers to maintain data. By enlisting user-editors, Drafty relies on free purveyors of the data to maintain a structured dataset long term.

There is a history of crowd powered systems that seek to solve long-standing issues in crowdsourcing, such as bringing the benefits of crowd-powered work to an inexperienced audience. Soylent (Bernstein et al. 2015) and Fantasktic (Gutheim and Hartmann 2012) are novice-centric systems built to address common mistakes when crowdsourcing. For example, providing insufficient guidance to workers or not verifying the data. Drafty automatically addresses both of these shortcomings, and does not require a requester to post micro-tasks for workers. Drafty empowers its user-editors to perform these micro-tasks in a find and fix or find and verify pattern. Previous research has also shown that there is little difference between expert and non-expert workers for routine-tasks, but this changes for specialized tasks (See et al. 2013). This argument is further supported by research on knowledge-intensive tasks (De Boer et al. 2012) that are more successfully completed by crowds with specific knowledge (Oosterman et al. 2014; Haas et al. 2015a). Drafty seeks to solve this problem by exploring if user-editors interested in a specific dataset are more accurate at editing data for specialized fields. Crowd-

[1]This data is available at http://drafty.cs.brown.edu/professors/

[2]The software is available at http://drafty.cs.brown.edu/

fill, like many other similar systems, has not assessed sustainable low-cost solutions to this problem.

Peer Production

Research in peer production systems explore using user interactions and user interests for unstructured data upkeep. For example, SuggestBot uses Wikipedia editors' contribution histories to suggest editing tasks (Cosley et al. 2007). WikiTasks supports the creation of site-wide tasks and self-selection of personal tasks within Wikipedia (Krieger, Stark, and Klemmer 2009). Unlike SuggestBot and WikiTasks, Drafty hosts a structured dataset rather than unstructured Wikipedia articles, which allows for a different set of interactions and the data has a pre-specified attributes. Also, Drafty is a custom built system that is not dependent on another platform, such as Wikipedia, to work effectively.

Drafty has automated mechanisms to infer interests from interactions to make recommendations. Drafty automatically selects the fields to request edits for. In contrast, SuggestBot relies on downloading and analyzing large sets of Wikipedia articles. Additionally, WikiTasks relies on humans to manually create tasks. Drafty records interactions beyond those used by SuggestBot (search, click, sort), which allow Drafty to build interest profiles from a wider variety of interactions. This allows for more robust models and analysis to enhance system recommendations. SuggestBot and Kylin evaluate whether they increase submission rates of edits (Cosley et al. 2007; Apache 2015). Drafty evaluates this, and the interactions that increase submission rates. In addition, Drafty also evaluates recall and precision accuracy for edits.

Learnersourcing

Research in learnersourcing has explored using interactions from native system users to assist with upkeep of online content. Williams et al. developed AXIS, a system which combines feedback from learners in online courses with machine learning algorithms to improve problem explanations (Williams et al. 2016). Future students in the course use these explanations for assistance and engage in the same feedback mechanism. This loop over time helps these explanations adapt to new user habits through an automated system. The long term benefits of this feedback mechanism are a major motivation for Drafty's integration of user interest profiles to solicit user-editors to maintain data. To perform such complex analysis, Drafty relies on collecting and inferring interest from the interactions of its users. In a similar endeavor, Kim et al. collect interaction data from online learners and use it to improve the experience of navigating video lectures (Kim et al. 2014). Li et al. investigate how to optimally identify groups of workers based on their characteristics (Li, Zhao, and Fuxman 2014). For each task, their method identifies the subgroup of workers best suited for a particular task. Drafty shares this philosophy but applies it to individual users. Drafty's philosophy is that each user is personally interested in different areas, so finding the right question to ask is the key to a better response.

Recommendation Systems

Drafty employs user interest profiles to target user-editors for data upkeep. Targeting users based upon surveys, interactions, preset tasks, or interests is a common theme among recommender systems (Ricci, Rokach, and Shapira 2015). These interactive and intelligent systems are used to give recommendations that match users' preferences (Rashid et al. 2002). Drafty takes builds on this idea to assess if soliciting user-editors based upon their interests leads them to volunteer to review more data and do it more accurately.

GroupLens (Resnick et al. 1994) is a collaborative filtering system that predicts readers preference of an article based on ratings. This is accomplished through explicit feedback provided by the user, but implicit feedback from user interactions can be similarly useful (Huang, White, and Dumais 2011; Chen, Pavlov, and Canny 2009). Huang et al. collected fine-grained interaction data in the wild to understand web users behaviors and search patterns. Chen et al. also explore behavioral targeting in web users primarily using clicks. These types of behavioral interactions have shown positive results in building user interest profiles in related recommender systems (Zhao et al. 2015). Zhao et al. show how new content or data created by users can be used to enhance user interest profiles in recommender systems. Drafty builds on those ideas by inferring interest from users automatically from easy-to-collect online interactions. Drafty's user interest profile integrates these findings in addition to heavily weighting data edits in its model.

The Drafty Platform

Drafty is an interactive online platform for building and maintaining large structured datasets. The platform leverages the expertise of its users by drafting them to become user-editors. Drafty's user interface and basic functionality closely resemble that of a spreadsheet application (Figure 1), providing user-editors a minimal learning curve. This is in contrast to traditional database systems (e.g. Freebase (Bollacker et al. 2008) and Wikidata (Vrandečić and Krötzsch 2014)) that present structured data as articles, and require switching to an editor mode. User-editors can freely submit new data or peer evaluate existing data. Drafty uses interactions to provide a human-in-the-loop mechanism to allow normal users to edit data more effectively. Previous studies have demonstrated that data quality increases if users are aware that their work will be reviewed by others (Huang and Fu 2013). Drafty solicits user-editors to help validate or fix data using an additive model for interaction-based profiling (Figure 2). We refer to this model as the user interest profile.

As a user-editor performs more interactions, their user interest profile becomes more robust and Drafty can better personalize their review requests. Other research has validated methods for inferring interest from user interactions to build interest profiles (Kim, Oard, and Romanik 2000). The benefit to this system is that over time, solicited and unsolicited data fixes lead to a more accurate and mature dataset. If user-editors trust the data found in Drafty and feel empowered to correct inconsistencies, their long-term commitment to the system should increase.

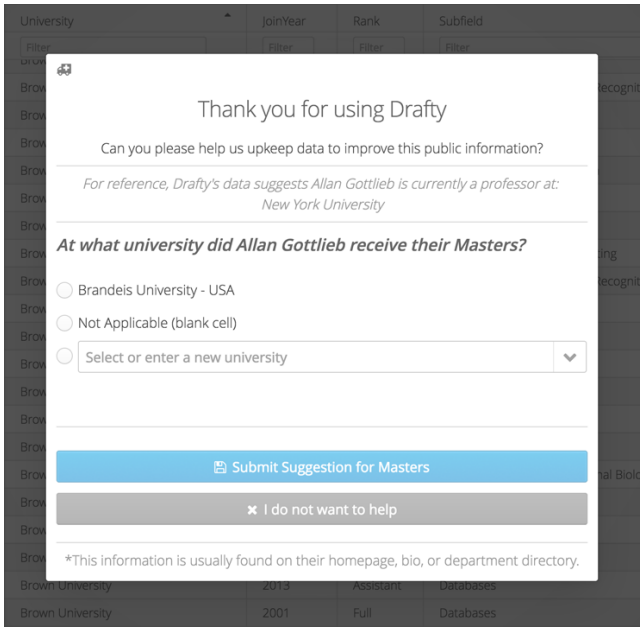


Figure 2: Drafty prompts user-editors to review data matching their interest profile. They can confirm a previous suggestion, submit an alternate suggestion, or close the prompt.

The following sections describe Drafty's methods for assigning confidence scores per data suggestion, user interest profiling, determing data for review, and a user study to validate the user interest profile.

Confidence Score per Cell For the purposes of this study Drafty uses a simple method to assign confidence scores per cell. The most recent submitted value per cell is assigned the highest confidence score among all possible values for that cell. The reasoning for this decision is twofold. First at the onset of this study, Drafty did not have enough data to create a valid learned model for assigning confidence scores. Therefore, the initial decision is to treat user-editors similar to administrators on Wikipedia (Krieger, Stark, and Klemmer 2009). User-editors are trusted to edit and maintain the dataset since they have an inherent interest in the content. This is also the same pattern followed in a normal business environment if users collaborated on Google Sheets. Second, Drafty focuses on providing user-editors a simple interface to edit cells. They select from pre-existing suggestions or submit a new one. This is in contrast to common upward and downward voting mechanisms found in other systems like Crowdfill (Park and Widom 2014). We feel Drafty's method mirrors mechanisms found in normal business environments and has a lower learning curve.

User Interest Profiles

Each user-editor has a unique profile computed for them. Interactions are recorded via the browser and stored in a central database. A user interest profile is either created or retrieved from the database when a new session starts, based on the user-editor's browser cookie.

Category	Interaction Type	Weight
Click	Click (Highlight)	1
	Row of Click	*1
	Double Click	*1
Search	Partial Search	2
	Complete Search	3
Edit	Suggestion (Validate)	4
	Suggestion (New Data)	5

Table 1: Weights (T_w) per interaction type used to build user interest profiles. (*A field is awarded an additional point when any field in that row is clicked.)

User Interest Weights Drafty's user interest weights mirror other approaches to establishing implicit user interest based on browsing information. Much of the literature on building user interest profiles based on implicit feedback focuses on the way that users interact with the web, and how to tailor web searches to pages of high interest. We used Chan's ideas about building web user profiles to inform our additive interaction model. Chan considers the frequency at which a user visits a page as the highest indicator of interest (Chan 1999). Our additive model performs similarly, where user-editors are determined to be more interested in a cell they more frequently interact with. Chan also notes that if a user clicks more links on a page, the page is likely to be of higher interest to that user. Similarly, Drafty reflects that if a user interacts multiple times with a category (e.g. clicks on three professors from the same university), this indicates general interest in that category.

Kim et al. outline five categories of observable behaviors users exhibit when interacting with websites (Kim, Oard, and Romanik 2000). In Drafty's user interest model, interactions are weighted by the interaction type, as shown in Table 1. Three categories adapted from their model are used to determine the weights for "Click," "Search," and "Edit." These are the main interactions that users perform on the data, in increasing order of demonstrated interest. The "Click" category is the weakest form of interaction that users exhibit, as clicks may be normal user behavior and thus may not show intent. The "Search" category shows that a user exhibited intent to engage with one or more of the results. The "Edit" category shows intent and knowledge of a precise cell, so is weighted the highest.

Recording Interactions A **click** on a specific cell might indicate interest in that particular data point. For example, a click on a cell that contains the value "Databases" might indicate a broader interest in that subfield. However, that interaction might also indicate interest in that specific row. Other potential factors could influence user interest in that row, such as the professor's university. Drafty handles these possibilities by first adding one point to the column's score for the value that was clicked. Then, Drafty adds one additional point to the additional column types in that row. **Double-clicks** are given an additional point because this indicates an attempt to edit data.

A **complete search** is recorded when a user-editor stops

typing and leaves a search field. This indicates they are satisfied with the result. A **partial search** is a search in progress where a user pauses typing but does not leave the search field. This indicates the user is examining the results, but is still in the process of searching. In both cases the user interest profile accumulates points for all possible data values the user-editor intends to search for.

Drafty records points for **edits** two different ways. If a user-editor selects an existing value of a cell, Drafty adds 4 points for the validation of the selected value. If a user-editor suggests a new value for the cell Drafty records 5 points for the relevant value. In the case of new values, Drafty takes this to be a stronger signal of interest given the proactive nature of the interaction.

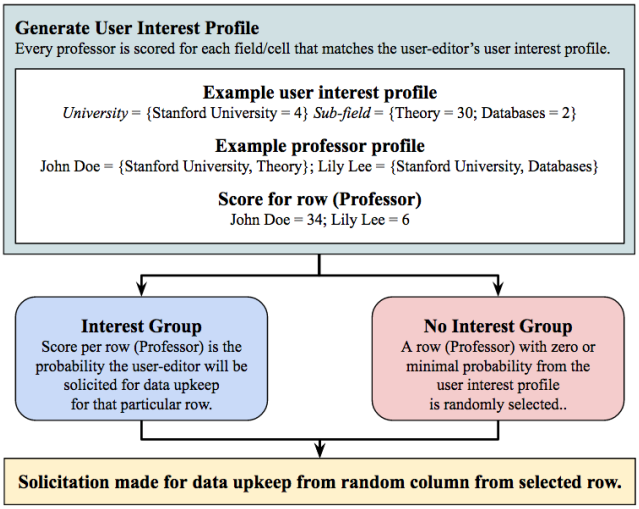


Figure 3: Drafty analyzes user interest profile to solicit data review based on a user-editor's experimental group.

Determining Data for Solicitation A user-editor's interest profile is used to determine what row will be used to solicit data review (Figure 3). Drafty will randomly select a column from that row for data upkeep. These solicitations are triggered randomly just before 10 interactions have been made. This number was derived from the average number of interactions made before data was submitted for upkeep during the system's pilot test phase. Drafty computes an interest score for each row in the dataset,

$$f_n(r) = \sum_{h=1}^{N} \sum_{i=0}^{T} (T_w \cdot x) \quad (1)$$

where N is the total number of columns. T is the number of interaction types (i.e. click, search, etc...), T_w is the weight of an interaction type (see Table 1). x is the number of interactions per value. The higher the computed score, the greater the inferred interest. A score of zero indicates no interest. The scores are used to compute the probability (Equation 1) that the row will be selected for review by a user-editor.

$$P(r) = \frac{f_n(r)}{f_o(r)} \quad (2)$$

$P(r)$ is the probability a row is solicited for review based on a user's interests. $f_o(r)$ is the total user interest score for all rows. $f_n(r)$ is the user interest score per row. To solicit data for review, Drafty randomly selects a row from the user interest profile. Rows with higher interest scores have a higher probability of solicitation. This ensures a measure of randomization so the same rows are not repeatedly selected.

Extending the User Interest Model Drafty's additive profiling model calculates relative interest based on a user-editor's interactions with the system. This technique relies on the assumption that over time interactions best representing user-editor interest will outweigh interaction "noise", such as errant clicks on non-relevant cells. The additive model acts as a catch-all for different interaction patterns. This approach is extensible to high activity users while still providing sufficient data from low activity users. A possible extension could be weighing recent interactions more strongly to give greater weight to current interests. Another possible extension is to track the peer-judged validity of a user-editor's data fixes. A user-editor whose suggestions are judged valid by their peers might be considered more of an expert and their contributions given higher confidence scores.

Another pertinent question is how to make interaction-based profiling data agnostic. For example, Drafty records interest in University, Bachelors, and Masters separately. However, it captures interest in universities from a broader perspective. The assessment of how certain columns in a structured dataset are dependent on other columns was not assessed in this current version of Drafty. For example, a user-editor interested in "University A" in the University column might reasonably be assumed to have interest in "University A" in the Bachelors column. However, Drafty does not currently aggregate interest in the model because that relationship might be too specific to this particular dataset.

Validating User Interest Profiles

A validation study was conducted to assess the validity of the user interest profile. The procedures were approved by our Human Subjects Protection (IRB) office. Following established human subject guidelines, all participants consented to the study. Participants were specifically recruited via convenience sampling from the computer science community, comprising 20 undergraduate students, graduate students, faculty, and staff with academic and professional backgrounds in Computer Science. Participants had current and past experiences with 20 different universities. Their range of expertise in Computer Science consisted of 16 different subfields. Participants included 8 (40%) females and 12 (60%) males with a mean age of 29 years per participant.

After a demographic questionnaire, participants were shown a 1 minute instructional video explaining what Drafty is and how its interface works. This video reviews the same instructions from the welcome screen Drafty shows to first-time user-editors. Then user-editors were given three training tasks, such as "Name a professor who joined their university between the years of 2000 and 2005." Participants performed each interaction type that comprise the user inter-

est profile across various columns to ensure they are familiar with the system.

Normal user-editors are free to use Drafty to explore and make edits freely. It was imperative that participants were given the same freedom to create and pursue tasks as normal user-editors. Participants were instructed to create three tasks to perform themselves, to simulate a natural inquiry of the data. Examples of tasks they created were:

- *Find my potential supervisor.*
- *Find all the professors at Carnegie Mellon who also obtained PhD at Carnegie Mellon.*
- *Count how many faculty members were hired in the past 10 years by UC Berkeley and Stanford, as well as their research areas.*

The interactions recorded during this part of the validation study were used to build user interest profiles for each participant. After the participant completed all three of their tasks they answered twelve randomly selected Likert scale questions. The four choices per question were: Not at all interested, Somewhat interested, Interested, or I do not know. Each question asks "How Interested are you in..." followed by one of the three following types: a *university name*, *subfield*, or a *professor name from a university*. The content of each question is directly determined by the participant's user interest profile. Each type will be asked four times. There are four methods for selecting the data to ask from the participant's user interest profile. *1) A random data point the participant showed no interest in. 2) Data with the highest interest score per type. 3) A random data point the participant showed some interest in. 4) Data randomly selected from the entire dataset.* This final part of the survey generated 240 total answers that are used to assess the validity of the user interest profile in the following section.

			Mean Interaction Score		
Answer	**N**	**All**	**Prof**	**University**	**Subfield**
Don't know	39	1.13	1.67	1.00	0.00
No Interest	76	0.84	1.47	0.86	0.25
Some Interest	51	1.49	4.18	0.71	0.83
Interest	74	2.73	4.56	1.80	2.54

Table 2: Mean interaction score per question for each participant from the validation study.

Validation Results The weights from Table 1 were used to calculate the mean interaction score per question type for each participant. The user interest profile is an additive model that relies on accumulating interactions. It works based on the premise that the more interactions the user makes with a specific field the greater their interest in that field. The results of this study validate this relationship between interactions and interest. Participants interaction score was on average 3.24 higher on the data they were interested in versus the data they had no interest in (Table 2). A Kruskal-Wallis test was conducted to evaluate the differences among answers from the four Likert scale questions[3]. The test, which was corrected for tied ranks, was significant $\chi_2(2, N = 240) = 29.6$, $p < 0.001$. This demonstrates that the interaction score from the user interest profile are a significant indicator of a user-editor's level of interest.

Live Experiment

A live experiment was conducted on a publicly-accessible version of Drafty to assess how user-editors behaved in the wild, outside of a lab study.

Drafty is a system that requires a large number of user-editors to collect a sufficient amount of data to answer pertinent research questions. So it was shared across various computer science forums such as Hacker News, TheGradCafe, and Reddit CompSci to attract users with an interest in Computer Science. This in contrast to other systems such as Crowdfill (Park and Widom 2014) and Soylent (Bernstein et al. 2015) that rely on enlisting and compensating crowdworkers. These workers may have varying levels of motivations to use the system and to perform tasks. An example title used on Reddit was "Records of 3,600 computer science professors at 70 top universities (US/Canada) help us keep it up to date!." Each post sharing Drafty contained the text: "Wanted to share a computer science resource a couple of us in the Brown University HCI Group have put together. It is a crowd-editable spreadsheet of data of approximately 3,600 computer science professors. For example, where they got their degrees, subfield of expertise, their join year and rank, etc... It might be useful if you're applying to Ph.D. programs or faculty positions, seeking external collaborators, or just to better understand hiring trends in CS departments."

New user-editors are shown a welcome screen on their first visit that includes key information such as: all interactions are captured and used anonymously for studies; double-click a cell to fix a piece of data; and that Drafty is a HCI research system. In addition to the welcome message, Drafty's footer reminds the user that "Drafty is a research project. All interactions are captured and used anonymously for studies." Each new user was randomly assigned to one of three experimental groups. In group 1, user-editors were asked to fix data for professors they showed no interest in. In group 2, user-editors were asked to fix data for professors they did show interest in. Only user-editors assigned to experimental groups 1 and 2 were solicited for data review.

Data

The live experiment, approved by our Human Subjects Protection office, ran over a 7 month period ranging from August 25, 2016 to March 25, 2017. During this time user-editors, could freely view, edit, and export various academic records using Drafty. Drafty recorded 41,426 interactions from 6,077 user-editors over 7,741 total visits. 809 user-editors had multiple visits at an average of 3 visits per user-editor. User-editors submitted data fixes when solicited or self-initiated 31.9% per attempt. Unsolicited user-editors

[3]Previous research (Allen and Seaman 2007; Clason and Dormody 1994) has shown a Kruskal-Wallis test is appropriate on Likert scale survey questions where group size is unequal

	Number of Attempts			*Mean per Attempt**		*Mean Totals per Visit*		
	Normal	**Exp.**	**Total**	**Interactions**	**Interest**	**Interactions**	**Interest**	**Visits**
All	1581	1482	3063	11.0	20.5	141.1	290.0	2.7
Not Completed	989	1389	2378	13.7	25.1	134.3	261.4	1.9
Completed	592	93	685	6.7	13.3	151.4	334.3	3.8
Incorrect (Interested)	-	17	17	15.3	64.1	26.8	116.8	1.7
Correct (Interested)	-	24	24	16.2	40.8	27.9	71.1	2.3
Incorrect (Uninterested)	-	4	4	13.5	25.3	71.8	132.8	1.3
Correct (Uninterested)	-	32	32	12.9	21.3	33.1	57.2	2.6
Incorrect (Normal)	121	-	121	5.9	12.3	30.0	61.3	2.6
Correct (Normal)	413	-	413	7.0	14.4	37.9	78.0	2.3

Table 3: Summary of correctness of participants' edits to the data. Incorrect/Correct submissions were manually verified by the authors. (*Per Attempt = between each attempt to upkeep data. Interest = cumulative score of user interest profile.)

submitted data fixes 37.5% per attempt. When solicited the uninterested group submitted data fixes 5.4% per attempt. The interested group submitted data fixes when solicited 8.8% per attempt. Submission rates can be dependent on the maturity of the dataset. Drafty does not collect user-editors personal or demographic information; such as name, age, or gender. Server logs capture IP addresses, but are not tied to the user profile nor examined for research.

Results

In the following section, we provide a detailed analysis that investigates the relationships between experimental groups, non-experimental groups, user-editor interactions, and accuracy (Table 3). Accuracy is determined by manually checking data submissions using online sources. Unless otherwise stated, non-parametric tests were conducted because the Shapiro-Wilk test of normality and Levene's test showed that the data was not normally distributed and the variances were unequal.

Active User-Editors

If a user-editor has multiple visits they are an active user-editors. It is important to show how acrtive user-editors make more effective contributions to Drafty that will ensure its long-term viability as a system. Table 4 contains a summary of interaction statistics for active versus inactive user-editors. Active user-editors perform 2.4 more interactions per visit than inactive user-editors. They also perform 3.1 more clicks than inactive user-editors. A click can indicate a user-editor has selected a cell to perform additional actions on. For example, to copy the cell contents to their clipboard. They also perform more searches, this can indicate an active interest to find specific data. Active user-editors also perform 5 times as many double clicks and create 6 times as many data submissions. These general patterns demonstrate how active user-editors are more engaged and make more contributions to upkeep a structured dataset.

Results indicate user-editors should be targeted for data review after their first visit. Multiple visits indicate an a higher level of interest and a commitment to the system and its contents. A one-tailed t-test of unequal variances was performed to compare the number of visits between user-editors who submitted accurate data when solicited vs user-editors who submitted inaccurate data when solicited (N = 56, M = 2.5, SD = 7.6) across all experimental groups. Results indicate a significant effect for visits, $t(64) = -1.8, p < 0.05$. By targeting active user-editors Drafty can make more effective interventions for data review.

User-editors who submit data have approximately 2 times more visits at the time they submit versus those who do not. A Mann-Whitney test was conducted to evaluate difference in total visits among user-editors who submitted suggestions (M = 4.68, SD = 11.78) against user-editors who did not (M = 2.38, SD = 5.85), for which the difference was significant at $p < 0.001$. This finding coincides with previous results showing that active and engaged user-editor will participate in data review. This demonstrates Drafty's potential to successfully maintain structured datasets over time.

Interaction Type	**Active**	**Inactive**
Click (Highlight)	3.4	1.1
Double Click	1.0	0.2
Partial Search	2.7	1.8
Complete Search	0.8	0.5
Sort Column	0.3	0.2
Submissions	0.6	0.1
Mean Interactions	8.8	3.7

Table 4: Average number of interactions by user-editors per visit segmented by the type of user. "Active" represents user-editors with multiple visits, who perform more interactions on average.

Interactions, Interest, and Edits

For the following section solicited user-editors are those who were asked to review data through an intervention by Drafty. They are part of the interested and uninterested experimental groups. Normal user-editors are those who made submissions for data review using the standard mechanisms within Drafty.

First we will observe the interaction habits that garner higher submission rates. User-editors solicited for data review showed more interesting results than normal user-editors. The following paragraph reviews the relationships between four groups:

- Interested - made submission
- Interested - did not submit
- Uninterested - made submission
- Uninterested - did not submit

A Kruskal-Wallis test was conducted to evaluate differences among four conditions/groups when user-editors were solicited to review data on the mean number of interactions and mean score of interactions a user-editor performed before being solicited to find missing data. User-editors with a higher number of interactions and interest score between solicitations, in addition to higher total interactions per visit are more likely to find and submit missing data. The number of interactions between solicitations is significantly different $\chi_2(3, N = 93) = 24.5$, $p < 0.001$. The cumulative interest score of interactions in between solicitations is significant $\chi_2(3, N = 93) = 17.8$, $p < 0.001$. Finally, the total number of interactions per visit is significantly different $\chi_2(3, N = 93) = 38.6$, $p < 0.001$. This indicates user-editors who are alerted too early could potentially have a negative reaction to the pop-up window used in solicitations. Soliciting reviews of data should be done after the user-editor has made a certain number of interactions to generate a robust user interest profile. To further support this result the total cumulative interest score is significant $\chi_2(3, N = 93) = 30.5$, $p < 0.001$. Follow-up tests were conducted to evaluate pairwise differences among the four groups. A Bonferroni correction was applied to control for Type I error. It showed no significant difference between the four groups. In general, the more intense and total interactions in a user-editor's session, the greater the chance they will edit data when solicited.

Condition	Precision	Recall
Uninterested group	57.1%	3.0%
Interested group	88.9%	9.2%
Normal unsolicited	75.8%	28.4%

Table 5: User-editors asked to review data they are interested in have higher precision accuracy on data submissions normal unsolicited and uninterested user-editors.

While submissions rates are useful to show an engaged user-editor population, accuracy is a better metric to ascertain Drafty's ability to maintain up-to-date structured data. We use use accuracy metrics, precision and recall, derived the information retrieval community. Precision is the number of correct submission over the number submissions per group. Recall is the number of correct submissions over the number of solicitations per group. Refer to Table 5 for a summary of precision and recall per data review condition. User-editors who were solicited to fix data they are interested in are three times more likely to submit accurate data than user-editors who are asked to fix data they are uninterested in. The highest precision is achieved when a user-editor is solicited to fix data based upon their interests. This demonstrates Drafty uses a more effective method for data collection and verification than traditional crowdsourcing methods such as those found in CrowdFill (Park and Widom 2014) and (Papoutsaki et al. 2015). In addition to these findings, user-editors submitted 81 fixes for subfield area of expertise. After manual verification, 92.6% of the submissions for subfield were deemed accurate. This is a substantial increase in accuracy for fields requiring domain-specific knowledge compared to traditional methods (Papoutsaki et al. 2015).

In general, solicited and normal user-editors who made accurate data submissions had more interactions per visit. This indicates the more engaged the user-editor, the better they are at maintaining a structured dataset. A one-tailed t-test of unequal variances was performed to compare the number of total interactions between user-editors who submitted accurate data when solicited (N = 21, M = 65.6, SD = 58.5) vs user-editors who submitted inaccurate data when solicited (N = 56, M = 36.4, SD = 39.9). The test determined that there was a significant difference in the total interactions per visit, $t(27) = 2.1, p < 0.05$. The same t-test was performed between user-editors who submitted accurate data when not solicited (N = 121, M = 30.0, SD = 23.2) vs user-editors who submitted inaccurate data when not solicited (N = 413, M = 37.9, SD = 38.8) across all experimental groups. Results indicate a significant effect for the number of total interaction score per visit, $t(333) = -2.8, p < 0.01$.

Interaction Type	Incorrect	Mixed	Correct	All
Click (Highlight)	7.0	5.8	6.1	6.2
Double Click	1.8	2.3	2.0	2.1
Partial Search	3.6	2.6	7.0	5.1
Complete Search	1.1	0.9	1.6	1.3
Sort Column	0.6	0.5	0.4	0.4
Submissions	3.2	2.6	0.4	2.5
Mean Interactions	17.5	14.7	19.6	17.8

Table 6: Mean number of interactions by user-editors per visit, separated by accuracy. (Incorrect = user-editor's who only made inaccurate submissions. Mixed = user-editors who made accurate and inaccurate submissions. Correct = user-editors who only made accurate submissions.)

User-editors who only submitted accurate data 2 across all conditions perform more partial and complete searches than other user-editors (Table 6). This matches earlier observations showing active user-editors perform more partial and complete searches than inactive user-editors. Both findings support previous research used to develop the user interest profile that validates giving higher weights to searches over clicks.

Discussion

This section reviews fundamental questions and observations about how to develop and use a system like Drafty

to enlist crowd editors in the upkeep of structured data. Drafty allows for structured data maintenance deploying both micro- (Park and Widom 2014) and macro- (Haas et al. 2015a) task based approaches. By combining the strengths of these systems in combination with validated user interest profile model, Drafty can maintain structured datasets over long periods of time at a lower cost and higher efficiency than traditional crowd powered systems and platforms. Current results show that interest-based profiling increases data accuracy by soliciting user-editors to find missing data they are interested in. Results also show Drafty is highly effective at sustaining data quality for fields that require domain-specific knowledge such as subfield.

Results demonstrated that user-editors with a higher number of visits to Drafty were more engaged. They submitted more fixes and had different interaction patterns and habits. In the future, predicting the most effective user-editors is essential. Drafty can be provide more effective interventions fewer times and achieve more accurate results. Attracting and keeping its base of successful user-editors is akin to other systems employing various methods to build trusted groups crowdsourced workers to ensure repeated efforts of high quality work (Bernstein et al. 2015).

Drafty can prioritize conflicted data for solicitation and upkeep. This will allow Drafty to scale its effectiveness for mature datasets, allowing for the feedback loop to better verify new data suggestions. For example, Drafty can prioritize cells with multiple suggestions whose standard deviation of confidence scores are within a certain threshold. This allows Drafty to preemptively identify uncertain data and make solicitations to fix it. Data verification is an open research area in crowdsourcing. In the future, Drafty can help answer these questions by providing an ecosystem to run controlled experiments on multiple versions of the same dataset. These experiments can lead to findings focused on what level of expertise is required to perform verification tasks.

Finally, the authors intend to investigate user-editor behavior in a wider range of datasets. There may be data where the incentives of the user-editors may not be aligned, such as if some of the data is the competitive advantage for some users (such as job openings or fellowship opportunities), or if the data has a higher level of subjectivity, or if there may be bad actors (Priedhorsky et al. 2007) with a political motivation manipulating the data. These questions combine the social nature of the user-editors with the technical capabilities of the structured data hosting platform.

Conclusion

In this paper, we report the design, implementation, and evaluation of Drafty, a platform that recruits editors from the users of structured data. Drafty's user-editors provide fixes for structured data through solicited and unsolicited methods. Their interactions were captured and analyzed to build user interest profiles, which were validated by a survey afterwards. In a longitudinal experiment in the wild, the interest profiles were used to solicit user-editors to fix data they were either interested in or had not shown interest in based on their interactions.

We found that user-editors who were asked to fix data they are interested in are more than three times more likely to submit accurate data than user-editors who are asked to find data they are uninterested in. User-editors who performed more interactions in between being prompted to fix data were not only more likely to submit data, but their submissions had higher levels of accuracy. Successful user-editors often engaged in active searching for data and would perform more search actions to find structured data to review. This experiment has shown that using user interest profiles help improve the accuracy of structured data and help build a self-sustaining dataset. Overall, our work brings a vision where users become custodians, and outdated data becomes an outdated concept.

Acknowledgments

This work would not be possible without the contributions of Brown University students who crowdsourced the initial data in the HCI seminar in Spring 2014 and Spring 2015. The authors also acknowledge Alexandra Papoutsaki who reviewed the data throughout 2014, as well as Lucas Kang who reviewed the data in the summer of 2014. Finally, we thank the many online visitors who submitted corrections to the data while the dataset was public, and the Amazon Mechanical Turk crowdworkers who acquired and validated the data itself.

References

Allen, I. E., and Seaman, C. A. 2007. Likert scales and data analyses. *Quality progress* 40(7):64.

Apache. 2015. Apache Kylin: Extreme olap engine for big data. http://kylin.apache.org/. [Online; accessed 2017-06-09].

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2015. Soylent: a word processor with a crowd inside. *Communications of the ACM* 58(8):85–94.

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 1247–1250. ACM.

Chan, P. K. 1999. A non-invasive learning approach to building web user profiles. In *KDD-99 Workshop on Web Usage Analysis and User Profiling*. Citeseer.

Chen, Y.; Pavlov, D.; and Canny, J. F. 2009. Large-scale behavioral targeting. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 209–218. ACM.

Clason, D. L., and Dormody, T. J. 1994. Analyzing data measured by individual likert-type items. *Journal of Agricultural Education* 35:4.

Clauset, A.; Arbesman, S.; and Larremore, D. B. 2015. Systematic inequality and hierarchy in faculty hiring networks. *Science Advances* 1(1):e1400005.

Cosley, D.; Frankowski, D.; Terveen, L.; and Riedl, J. 2007. Suggestbot: using intelligent task routing to help people find

work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*, 32–41. ACM.

Crunchbase Inc. 2007. Crunchbase accelerates innovation by bringing together data on companies and the people behind them. https://www.crunchbase.com/. [Online; accessed 2016-08-20].

De Boer, V.; Hildebrand, M.; Aroyo, L.; De Leenheer, P.; Dijkshoorn, C.; Tesfa, B.; and Schreiber, G. 2012. Nichesourcing: harnessing the power of crowds of experts. In *International Conference on Knowledge Engineering and Knowledge Management*, 16–20. Springer.

Gutheim, P., and Hartmann, B. 2012. Fantasktic: Improving quality of results for novice crowdsourcing users. *EECS Dept., Univ. California, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2012-112*.

Haas, D.; Ansel, J.; Gu, L.; and Marcus, A. 2015a. Argonaut: macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment* 8(12):1642–1653.

Haas, D.; Krishnan, S.; Wang, J.; Franklin, M. J.; and Wu, E. 2015b. Wisteria: Nurturing scalable data cleaning infrastructure. *Proceedings of the VLDB Endowment* 8(12):2004–2007.

Howe, J. 2006. The rise of crowdsourcing. Wired. https://www.wired.com/2006/06/crowds/. [Online; accessed 2016-04-21].

Howe, J. 2008. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. New York, NY, USA: Crown Publishing Group.

Huang, S.-W., and Fu, W.-T. 2013. Enhancing reliability using peer consistency evaluation in human computation. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 639–648. New York, NY, USA: ACM.

Huang, J.; White, R. W.; and Dumais, S. 2011. No clicks, no problem: Using cursor movements to understand and improve search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, 1225–1234. New York, NY, USA: ACM.

Kim, J.; Guo, P. J.; Cai, C. J.; Li, S.-W. D.; Gajos, K. Z.; and Miller, R. C. 2014. Data-driven interaction techniques for improving navigation of educational videos. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 563–572. ACM.

Kim, J.; Oard, D. W.; and Romanik, K. 2000. Using Implicit Feedback for User Modeling in Internet and Intranet Searching. *University of Maryland CLIS Technical Report* 1–21.

Krieger, M.; Stark, E. M.; and Klemmer, S. R. 2009. Coordinating tasks on the commons: designing for personal goals, expertise and serendipity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1485–1494. ACM.

Li, H.; Zhao, B.; and Fuxman, A. 2014. The wisdom of minority: discovering and targeting the right group of workers for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*, 165–176. ACM.

Mason, W., and Watts, D. J. 2010. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter* 11(2):100–108.

Oosterman, J.; Nottamkandath, A.; Dijkshoorn, C.; Bozzon, A.; Houben, G.-J.; and Aroyo, L. 2014. Crowdsourcing knowledge-intensive tasks in cultural heritage. In *Proceedings of the 2014 ACM conference on Web science*, 267–268. ACM.

Papoutsaki, A.; Guo, H.; Metaxa-Kakavouli, D.; Gramazio, C.; Rasley, J.; Xie, W.; Wang, G.; and Huang, J. 2015. Crowdsourcing from scratch: A pragmatic experiment in data collection by novice requesters. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

Park, H., and Widom, J. 2014. Crowdfill: collecting structured data from the crowd. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 577–588. ACM.

Priedhorsky, R.; Chen, J.; Lam, S. T. K.; Panciera, K.; Terveen, L.; and Riedl, J. 2007. Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work*, 259–268. ACM.

Rashid, A. M.; Albert, I.; Cosley, D.; Lam, S. K.; McNee, S. M.; Konstan, J. A.; and Riedl, J. 2002. Getting to know you: learning new user preferences in recommender systems. In *Proceedings of the 7th international conference on Intelligent user interfaces*, 127–134. ACM.

Redmond, M., and Baveja, A. 2002. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research* 141(3):660–678.

Resnick, P.; Iacovou, N.; Suchak, M.; Bergstrom, P.; and Riedl, J. 1994. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, 175–186. ACM.

Ricci, F.; Rokach, L.; and Shapira, B. 2015. Recommender systems: Introduction and challenges. In *Recommender Systems Handbook*. Springer. 1–34.

See, L.; Comber, A.; Salk, C.; Fritz, S.; van der Velde, M.; Perger, C.; Schill, C.; McCallum, I.; Kraxner, F.; and Obersteiner, M. 2013. Comparing the quality of crowdsourced data contributed by expert and non-experts. *PloS one* 8(7):e69958.

Vrandečić, D., and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM* 57(10):78–85.

Williams, J. J.; Kim, J.; Rafferty, A.; Maldonado, S.; Gajos, K. Z.; Lasecki, W. S.; and Heffernan, N. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning@Scale*, 379–388. ACM.

Zhao, Z.; Cheng, Z.; Hong, L.; and Chi, E. H. 2015. Improving user topic interest profiles by behavior factorization. In *Proceedings of the 24th International Conference on World Wide Web*, 1406–1416. ACM.