

ABSTRACT

Title of Thesis: REVISITING SHAKESPEARE'S WORLD:
OPTIMIZING DATA OUTCOMES AND
INVESTIGATING CONTRIBUTOR DYNAMICS

ZhiCheng Wang, Master of Science, Human-
Computer Interaction, 2024

Thesis Directed By: Dr., Victoria Van Hyning, College of Information
Studies (iSchool)

In this study, we present our work processing data output from *Shakespeare's World* (2015-2019), an early transcription project hosted on the Zooniverse online crowdsourcing platform. We refined the dataset to make it more amenable to low-code tools such as OpenRefine, enabling easier data exploration and reuse. Utilizing the cleaned dataset, we also explored *Shakespeare's World* volunteers' contribution patterns. By documenting our process of cleaning the outcome dataset, we provide steps and insights that may be useful for other transcription projects working with data derived from the Zooniverse platform. In addition to offering one plausible way to clean and analyze Zooniverse outcome data, our study also reveals the significant contributions from both anonymous and registered *Shakespeare's World* volunteers; the challenges in maintaining participation over the project's lifespan; and how the original aggregation protocol, which was designed specifically to combine multiple transcriptions by *Shakespeare's World* volunteers, resulted in fewer successfully transcribed

lines than expected. These findings have broader implications for project design, volunteer engagement and data management practices in online crowdsourced transcription projects.

REVISITING SHAKESPEARE'S WORLD:
OPTIMIZING DATA OUTCOMES AND INVESTIGATING CONTRIBUTOR
DYNAMICS

by

ZhiCheng Wang

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2024

Advisory Committee:
Assistant Professor, Victoria Van Hyning, Chair
Assistant Professor, Katrina Fenlon
Assistant Professor, Giovanni Luca Ciampaglia

© Copyright by
ZhiCheng Wang
2024

Acknowledgements

I would like to first extend my heartfelt thanks to Prof. Victoria Van Hyning for introducing me to the concept of crowdsourcing and for giving me the invaluable opportunity to work on the *Shakespeare's World* outcome dataset. The weekly discussions under her mentorship have provided me with crucial insights and essential details that are essential to the completion of this work. Without her expertise, encouragement, and belief in my abilities, this project would not have been possible. I am profoundly grateful for her role in my academic and personal growth during this endeavor.

I also want to extend my sincere gratitude to Prof. Katrina Fenlon and Prof. Giovanni Luca Ciampaglia for their willingness to serve on my committee and for providing invaluable critical and constructive feedback and suggestions.

Further, I wish to give a special shoutout to Prof. Eun Kyoung Choe for introducing me to the world of data visualization, and to Prof. Zach Drake for his systematic approach to teaching data analytics and visualization in his class. Their passion has been a constant source of inspiration throughout the duration of this project.

I would also like to thank everyone on the Zooniverse and *Shakespeare's World* team and everyone who participated in *Shakespeare's World* between 2015 and 2019. Their prior contributions have been the foundation of this project.

Most importantly, I would like to express my appreciation for my friends and family, whose invaluable support has propelled me through the most difficult times in my life. Without them, I would not even be here to write down these words today.

Portions of this paper have been presented at the “Past Meets Future” workshop at the International Conference on Intelligent User Interfaces (IUI) 2024 in Greenville, South Carolina. This opportunity not only allowed me to showcase the progress of my work but also to receive valuable feedback and insights from other scholars in the field.

This study uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including from the National Science Foundation, NASA, the Institute of Museum and Library Services, UKRI, a Global Impact Award from Google, and the Alfred P. Sloan Foundation.

Table of Contents

Acknowledgements.....	ii
Table of Contents.....	iv
List of Tables	vii
List of Figures	viii
1. Introduction.....	1
1.1 Crowdsourcing and Zooniverse.....	1
1.2 Shakespeare’s World and Its Workflow	3
1.4 Shakespeare’s World’s Outcomes	8
1.5 Project Motivation	10
1.6 Project Goal	11
2. Prior and Related Work	13
2.1 Continuous Efforts in Shakespeare’s World.....	13
2.2 Sharing Crowdsourced Outcome.....	14
2.3 Zooniverse Contribution Pattern	15
3. Methods	16
3.1 Low-programming Tools for Analysis	16
3.1.1 OpenRefine	17
3.1.2 Tableau.....	19

3.2 Approaches to Data Cleaning	19
3.3 Cleaning Shakespeare’s World Outcome Dataset	21
3.3.1 Normalizing Dataset.....	23
3.3.2 Ethics and Privacy	33
3.3.3 Validating Data and Handling Missing or Incomplete Data	34
3.4 Preparing Dataset for Analysis	36
4. Preliminary Data Analysis & Visualizations	41
4.1 Summary Statistics of Contributions.....	41
4.2 Contribution Pattern – Volume.....	43
4.3 Contribution Pattern – Logged-in Volunteers per Month.....	47
4.4 Contribution Pattern – Macro	51
4.5 Subject Retirement Pattern & Estimated Transcribed Ratio	53
4.6 Subject Retirement Pattern – Macro	58
5. Findings	60
5.1 Volunteer Contribution Patterns.....	60
5.2 Subject Retirement Patterns.....	63
5.3 Consensus and Estimated Transcribed Ratio.....	67
6. Discussion.....	68
6.1 Implications	68
6.1.1 Reuse Potential.....	68
6.1.2 Volunteers in <i>Shakespeare’s World</i> and Similar Crowdsourced Projects.....	70
6.1.3 Online Crowdsourced Projects Management	73
6.3 Limitations.....	74

6.4 Future Work.....	75
6.4 Conclusion.....	76
Appendices.....	78
Bibliography	81

List of Tables

Table 1 – Hierarchy of Raw Output Dataset Attributes	22
Table 2 – Attribute Name Mapping Table	31
Table 3 – Hierarchy of Cleaned Output Dataset Attributes	32
Table 4 – Summary Statistics for the Number of Classification Submitted per Month.....	42
Table 5 – Summary Statistics for Est. Transcribed Ratio under Different Subject Retirement Reasons	55

List of Figures

Figure 1 – Screenshot of Shakespeare's World Transcription Interface [17].....	5
Figure 2 – Screenshot of Shakespeare's World Talk Forum.....	6
Figure 3 – Screenshot of Shakespeare's World Blog.....	6
Figure 4 – Screenshot of Shakespeare's World Twitter Page	7
Figure 5 – Treemap of Total Classifications Contributed per Volunteer throughout Shakespeare's World	8
Figure 6 – All three of these fields were actual subject_data appeared at different stages of Shakespeare's World	18
Figure 7 – Distribution of Positive Session Durations.....	38
Figure 8 – Distribution of Session Durations without Outliers.....	39
Figure 9 – Total Number of Classifications Submitted per Month.....	44
Figure 10 – Total Number of Classifications Submitted per Month - Varied by Task Category	45
Figure 11 – Total Number of Classifications Submitted per Month - Varied by Subject Genre	46
Figure 12 – Number of Distinct Active Volunteers per Month	48
Figure 13 – Number of Distinct Active Volunteers Per Month Each Year	49
Figure 14 – Number of Distinct Registered and Active Volunteers & Number of Classifications Contributed by Them per Month	50
Figure 15 – Classifications Submitted by Volunteer Over Time.....	52
Figure 16 – Subject Retirement Timespan with Different Retirement Reasons	56

Figure 17 – Subject’s Estimated Transcribed Ratio.....	57
Figure 18 – Subject Retirement Timespan and Estimated Transcribed Ratio with Classification Counts and Retirement Reasons	59
Figure 19 – John Ward Diaries, Vol. 14 [47].....	66
Figure 20 – Shakespeare’s World Top 100 Contributing Registered Volunteer Classifications Pattern, Sorted by Their user_ids.....	71
Figure 21 – Shakespeare’s World Top 10 Contributing Registered Volunteer Classifications Pattern, Sorted by Total Classification Counts	72
Figure 22 – Screenshot of the Google Form Distributed to Collect Alternative Usernames..	78
Figure 23 – Data Structure of the Raw Output Dataset	79
Figure 24 - Data Structure of the Cleaned Dataset	80

1. Introduction

1.1 Crowdsourcing and Zooniverse

Initially coined by Jeff Howe in 2006 [1], crowdsourcing gained significant traction in the mid-2000s amid the rapid expansion of the internet. From capturing, organizing, and sharing critical information through the Ushahidi Platform during the 2010 earthquake in Haiti [2] to enabling community-driven product development via LEGO Ideas, crowdsourcing has now become a transformative methodology of harnessing collective talents and engaging communities, widely used across different domains for various purposes. Crowdsourcing's definition has since been debated over the years [3], with Estellés-Arolas and González-Ladrón-de-Guevara's being the most all-encompassing [4]:

Crowdsourcing is a type of participative online activity in which an individual, and institution, a non-profit organisation, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while

the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken.

First established in 2007 with the Galaxy Zoo project [5] in astrophysics, Zooniverse has now become one of the world’s largest and most popular platforms for online crowdsourced research [6]. While the majority of the more than 200 projects hosted on the platform are still within the scientific domains, humanities-focused projects started to appear on the platform in 2011 following the launch of the *Ancient Lives* project [7].

Zooniverse projects invite volunteers to perform classification tasks on individual images, audio, or video files — referred to as subjects. These individual efforts are stored and aggregated for “consensus” through a series of algorithms developed by individual projects or the Zooniverse research group. The subjects are only “retired” (removed from the queue of subjects shown to volunteers for classification) once enough volunteers have contributed to an individual subject. The exact thresholds for retirement are determined by project owners and typically decided based on task complexity [8]. The consensus-based outcomes of Zooniverse projects are generally considered high-quality in the scientific community and have been used in more than 490 peer-reviewed academic papers across numerous disciplines as of March 2024 [9].

1.2 Shakespeare's World and Its Workflow

Shakespeare's World [10] launched on the Zooniverse platform in December 2015 as a collaboration between the Folger Shakespeare Library, Zooniverse and the Oxford English Dictionary (OED) to create base transcriptions for Folger's Early Modern Manuscripts Online (EMMO) collection [11]. The project called upon volunteers to take part in the transcription of manuscripts created during or approximate to William Shakespeare's lifetime (1564–1616). The project initially featured early modern manuscripts of letters and recipes [12] and later expanded to include newsletters [13].

Utilizing Zooniverse's project framework, *Shakespeare's World* presented digital images of manuscripts from Folger's collections to volunteers one at a time. As the sixth transcription project within the humanities domain on Zooniverse, *Shakespeare's World* experimented with an innovative approach by permitting volunteers to transcribe only the text they felt confident in reading. By placing location markers (dots) on the digital manuscripts, volunteers could select individual words, phrases or complete lines that they felt confident with and then enter the corresponding digitized text in the pop-up transcription boxes for the selected segments (see Figure 1) [8]. To ensure volunteers' transcriptions are not influenced by others, *Shakespeare's World* intentionally concealed the contributions made by others and opted instead to display dots placed by previous transcribers only when three or more people had placed dots in roughly the same place. Once an individual transcription was submitted, it was aggregated with previous records in real-time: a clustering algorithm was first performed to

align the classification with previous records stored in the database using its positional information (coordinates of dots marked by the transcriber), followed by a genetic sequencing algorithm, Multiple Alignment using Fast Fourier Transform (MAFFT), on the transcriptions to aggregate text strings. By design, a text string was only supposed to retire once three different volunteers had transcribed the same fragment with at least 95% similarity to each other [14]. Before the launch of *Shakespeare's World* the aggregation process happened offline after the project had collected a set number of *classifications per subject*. *Shakespeare's World* was unusual in the deployment of real-time aggregation. Though not seeking 100% alignment between the text strings submitted by volunteers, this outcome of this aggregation was described as “consensus” in keeping with the Zooniverse naming conventions [15].

Shakespeare's World was also designed to cater to better assist volunteers with limited paleography experience. Several learning aids were provided on the project's website. Shortcut buttons for some common contractions and abbreviations, like superscripts and brevigraphs, were also integrated into the transcription interface [12], [16].

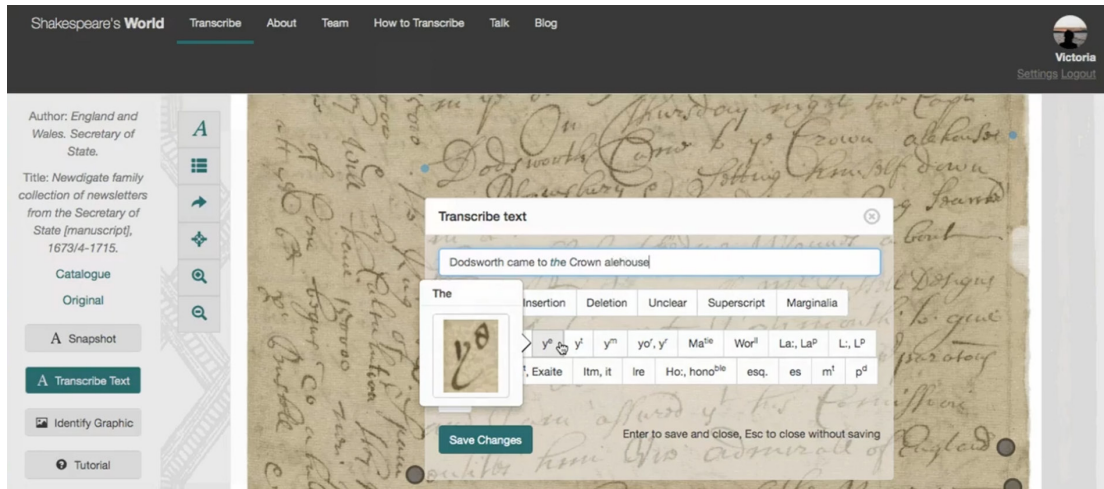


Figure 1 – Screenshot of Shakespeare's World Transcription Interface [17]

In August 2016, an internal testing was initiated, reflecting a proposal by the Folger and Zooniverse team and highly advocated by volunteers, to distribute subjects in their original document sequence, hoping to create better contextual integrity and immersion during transcription workflows. However, this feature faced technical difficulties and could not be reliably implemented. This feature was ultimately not integrated into the public workflow.

To enhance community engagement, the *Shakespeare's World* team also leveraged Zooniverse's Talk forum (see Figure 2). Volunteers were invited to discuss the page they had just transcribed or to use hashtags for summarization after saving their transcriptions. Additionally, the team also actively maintained an official blog (see Figure 3) and a Twitter account (see Figure 4) to keep the community informed about interesting discoveries and the latest updates regarding the project [12].

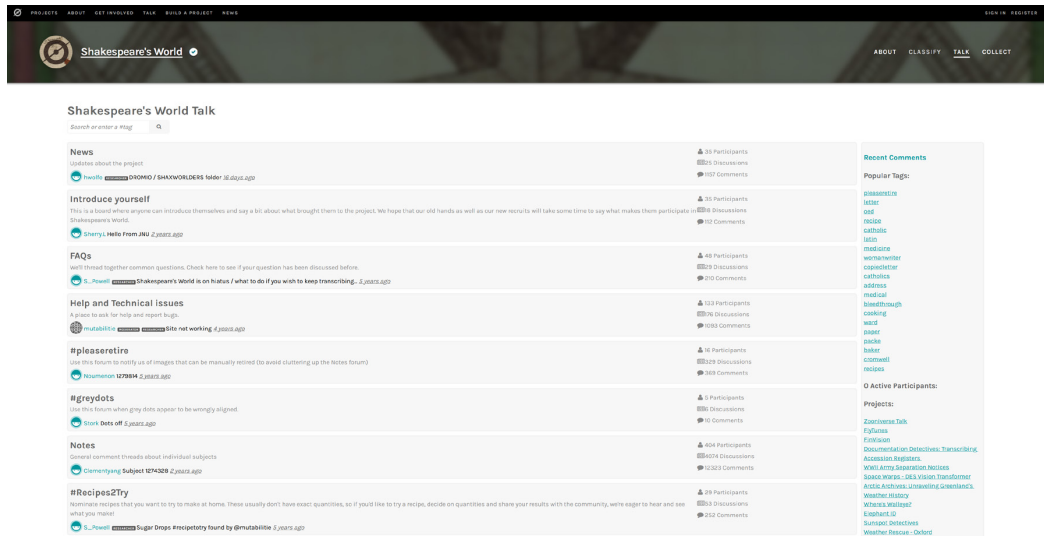


Figure 2 – Screenshot of Shakespeare's World Talk Forum

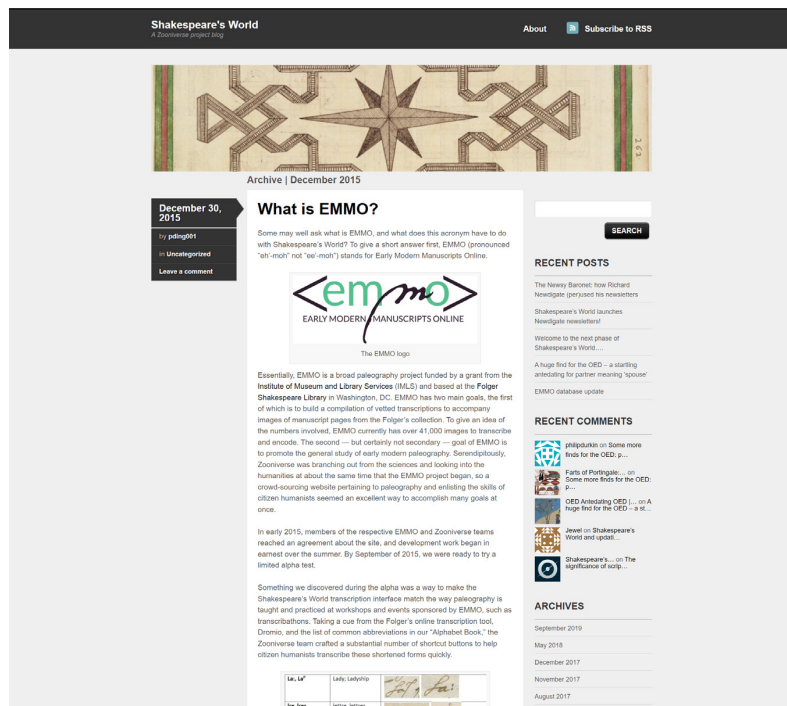


Figure 3 – Screenshot of Shakespeare's World Blog



Figure 4 – Screenshot of Shakespeare's World Twitter Page

1.4 Shakespeare's World's Outcomes

Shakespeare's World went on hiatus in October 2019. Throughout its operational phase, the project received contributions from 3,926 registered Zooniverse volunteers along with more than 94,570 anonymous sessions (representing an unknown number of individuals). The volunteers together contributed 203,390 classifications in total. The following treemap (Figure 5) shows the distribution of volunteer contributions towards *Shakespeare's World*. The anonymous (not-logged-in) volunteers' contribution was considered as one single collective entity and the other squares each represents one registered and logged-in volunteer.

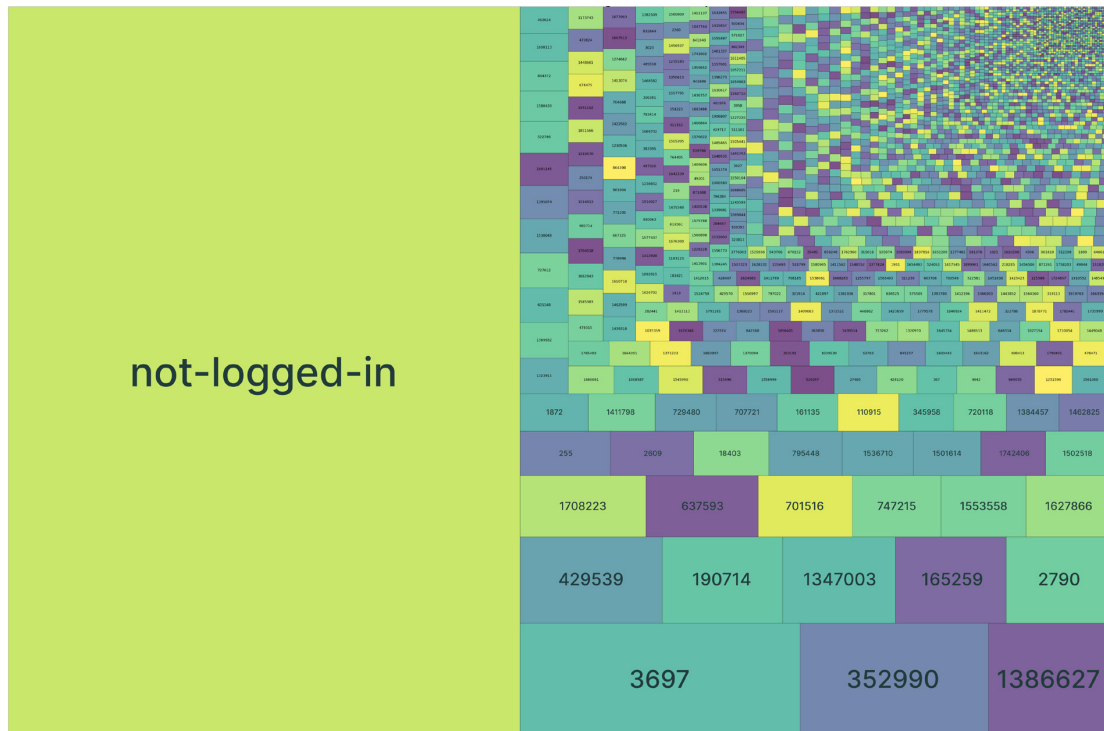


Figure 5 – Treemap of Total Classifications Contributed per Volunteer throughout *Shakespeare's World*

Shakespeare's World's outcome dataset contains both the transcription data made by the volunteers and metadata about volunteers' contribution acts. The outcome of the project has been impactful in several instances. Some of the crowdsourced transcriptions have been further cleaned and verified by expert paleographers at Folger and updated into EMMO.

Since most of the manuscripts provided in the project were never transcribed [12][11], some volunteers also helped uncover some previously undocumented words and new definitions as well as earlier attestations of words that were recorded in the OED. In one instance, a volunteer identified a nearly 200-year antedating use of "white lie" in Folger L.a.2 "Letter from Ralph Adderley to Sir Nicholas Bagnall, Ireland, 1567 April 10" [18]. This occurrence was later recorded by OED [19].

A significant part of *Shakespeare's World's* outcome was also volunteers' engagement with the materials [16]. Ridge et al. argued that the outcome of crowdsourcing projects is not limited to tangible outcomes such as organizing crowd efforts to successfully transcribe manuscripts, they also help to engage the volunteers with their heritage and learn more about their language and history [20]. Volunteer engagement has been in part captured by qualitative information via volunteers' interactions on the Talk forum and the qualitative information in the outcome dataset through the form of classification timestamps (time volunteers spent on the materials) and number of classifications contributed (number of materials volunteers engaged with). Prior research has utilized *Shakespeare's World's* Talk forum data to investigate the role of hashtags and "pings" as organizational tools within the project. Rohden et al. posited that the forum

extended beyond merely motivating engagement; it also acted as an interface through which contributors could not only generate but also deeply immerse themselves in knowledge-based practices [21].

1.5 Project Motivation

While crowdsourced transcription projects on Zooniverse promise great potential, they also present unique challenges. Since different projects employ different workflows, aggregation and retirement methodologies, the accuracy of the transcriptions must be meticulously verified individually to ensure the data's reliability, which usually requires intensive labor from professional paleographers (in the case of the Folger) or other subject specialists. Due to the scale and complexity of the data gathered, even just cleaning and preparing the datasets themselves could be a challenging and time-consuming task. As a result, not every team behind crowdsourced transcription projects on Zooniverse has made their outcome public and the rate of data reuse is significantly lower compared to other projects utilizing different task variants, such as object classifications or object identifications, hosted on the same platform [9].

Shakespeare's World suffered from similar challenges. While some outcomes of *Shakespeare's World* have been used by other scholars, only the volunteer transcriptions were partially processed and integrated into LUNA, Folger's digital image collection platform. Another obstacle is the rich data collected by the project, resulting in complex classification data, aggravated by complex metadata automatically generated by the Zooniverse system. As a result, the output dataset contains both intricate terminologies and nested and semi-structured

data. This information could be difficult to work with for scholars in both the humanities and computer science domains. The complex data structure made it difficult to surface patterns and find specific information, especially for those with limited programming experience and using low-programming tools. On the other hand, the archival terminologies and sometimes inconsistent terms in metadata have made the dataset difficult to comprehend for those with limited knowledge about the project.

While there are many studies investigating the quality of annotation tasks on paid micro-tasking platforms such as Mechanical Turk [22], [23], very little research has been conducted to study the quality of annotations produced by non-paid volunteer macro-tasking projects like *Shakespeare's World*. Victoria Van Hying, who was a Co-Principal Investigator of *Shakespeare's World* with Heather Wolfe, also expressed concerns about only having a broad understanding of the project and very specific anecdotes about the project but lacking nuanced quantitative insights into its performance and quality (conversation with the author).

1.6 Project Goal

To mitigate the complexity of the raw project outcome, we decided to flatten the dataset in order to make it easier to derive new quantitative insights into volunteer participation and quality of the resulting aggregations in *Shakespeare's World*. We intended for this work to help provide a template for how the data outcome of online crowdsourced transcription projects could be prepared and used. To these ends, we established the following research goals:

1. We will closely document our process of cleaning *Shakespeare's World* outcome dataset for public use and reuse as a case study. This will not only help to ensure the credibility and understandability of the outcome dataset but could also serve as a case study for other transcription projects hosted on Zooniverse to prepare their outcome datasets.
2. Using the cleaned dataset, we will investigate the contributions made by the volunteers and try to answer the following questions:
 - What were *Shakespeare's World's* volunteer contribution patterns like? How did this contribution pattern change over time and across different subjects?
 - Could we spot any previously unknown anomalies in volunteer contribution patterns? What can these anomalies tell us about the structure of the *Shakespeare's World* project and perhaps about transcription approaches on the Zooniverse platform?
 - Did the aggregation protocol work as intended?

2. Prior and Related Work

Since *Shakespeare's World* went on hiatus in 2019, various efforts have been made to enhance and utilize the outcomes of its dataset. Recently, other similar crowdsourced transcription projects on the Zooniverse platform have also shared their project outcome with the public. Previous research has already examined the general volunteer contribution patterns on the Zooniverse platform. These works provide a foundation for understanding data handling practices and volunteer engagement profiles on Zooniverse.

2.1 Continuous Efforts in Shakespeare's World

While *Shakespeare's World* had achieved some notable success, its experimental approach of letting volunteers transcribe what they were comfortable with had also led to inconsistency in the aggregated results which required significant editing. From 2016 to 2018, the *Shakespeare's World* team, including Van Hyning, web developers at Zooniverse, and both grant-funded and permanent staff at Folger, worked on improving the aggregated transcriptions. Their efforts led to professionally edited transcriptions that were successfully integrated into Folger's EMMO collection [15]. However, this process was effort-consuming and was not replicable on all transcriptions generated during the project. In a separate attempt in 2021, *Shakespeare's World* Co-Principle Investigators Heather Wolfe and Van Hyning worked with colleagues at Folger and undergraduate capstone students from the University of Maryland iSchool's iConsultancy on a data cleaning project to remove duplicated lines, the most common error of the aggregated data provided by Zooniverse at the time [24]. These

partially cleaned transcriptions were integrated into Folger’s image collection LUNA to facilitated research, and some of the project’s output was further edited by Folger docents, volunteers, and staff, while others were uploaded to another crowdsourcing platform, FromThePage [25] for a further phase of crowdsourced editing [15].

2.2 Sharing Crowdsourced Outcome

Besides *Shakespeare’s World*, other online crowdsourced transcription initiatives on Zooniverse have also been actively trying to make the outcomes of their projects publicly accessible. *Engaging Crowds: citizen research and heritage data at scale*, for example, is a project led by The National Archives (TNA) in the UK, with the Zooniverse team at the University of Oxford, Royal Botanic Garden Edinburgh (RBGE), and Royal Museums Greenwich (RMG) [26]. The project explores the potential of democratizing the engagement with cultural heritage collections through three different crowdsourced projects set up on Zooniverse: *HMS NHS: The Nautical Health Service* (ongoing as of March 2024), *Scarlets and Blues* (complete) and *The RBGE Herbarium: Exploring Gesneriaceae, the African violet family* (complete) [27] – [29]. As part of the *Engaging Crowds* project, the team hosted a GitHub page containing Zooniverse links to the document images and the processed outcome datasets of the transcriptions made by the volunteers for all three projects [30]. *Engaging Crowds* team preserved most of the original Zooniverse data export but flattened both record metadata and classification metadata fields. The team also replaced usernames with pseudonyms for both logged-in volunteers’ usernames and approximated anonymous volunteers using their hashed

IP addresses. A separate user engagement analysis was also performed on each of the three crowdsourced projects to gain insights into user behavior [31].

2.3 Zooniverse Contribution Pattern

Researchers have also tried to understand the contribution patterns of volunteers in crowdsourcing projects. In their 2014 study, Ponciano and Brasileiro studied the engagement patterns of volunteers in two crowdsourced projects hosted on Zooniverse, *Galaxy Zoo* and *The Milky Way Project* [32]. They presented a detailed quantitative analysis of volunteer contribution profiles within the two projects and identified five distinct engagement profiles (hardworking, spasmodic, persistent, lasting, and moderate), each characterized by a unique contribution pattern, duration, and intensity. Their study found that most of the volunteers fall into the moderate profile and only a few volunteers exhibit a persistent engagement profile. Persistent volunteers stay involved with the projects over a long period but are only actively contributing on few selected days within that timeframe. Despite having a smaller population and contributing only on few days, persistent volunteers were found to account for most of the human efforts contributed towards those projects.

3. Methods

3.1 Low-programming Tools for Analysis

Zooniverse suggests that it's best practice for projects to make classification data public after a proprietary period and allow volunteers to contribute in more diverse ways [33]. While Zooniverse provides resources to analyze project outputs, they are typically quite intimidating for people with limited programming experience to use. We believe it is also important to make the outcome of crowdsourced projects, especially transcription projects, more accessible to the public with limited programming experience. Lots of volunteers contributing to crowdsourced transcription projects like *Shakespeare's World* are drawn in by the source materials (manuscripts themselves) and might have limited programming experience, making it challenging for them to understand the raw output generated by Zooniverse. To effectively communicate to these volunteers what they have helped to achieve, it is not enough to only provide the raw outcome dataset. Similar to how *Shakespeare's World* provided learning aids and preset abbreviations for people with limited paleographic experience, we should also account for volunteers' variations in data literacy and programming skills and provide appropriate scaffoldings when sharing project outcomes. These supports would allow a broader audience, including all volunteers who participated in the project and the public, to meaningfully understand and explore what they have helped to achieve. By pre-processing the output dataset in Python, we hope to take one step in the direction of making project outcomes more accessible, especially to those utilizing low-programming tools.

3.1.1 OpenRefine

OpenRefine (formerly known as Google Refine) is a powerful free, open-source tool for working with messy data [34]. Operating as a local Java-based application, it is compatible with both Mac and Windows systems, providing flexible computing environments for different users. It utilized a graphical interface for interacting with data, similar to Microsoft Excel, but runs closer to a database. OpenRefine provides better support for complex data, includes more advanced built-in functions, like clustering, and supports multiple extensions to further enhance its capability, especially in specific domains like GIS [35]. Custom data transformations could also be performed by using General Refine Expression Language (GREL), Python or Clojure. This capability not only enables power users to perform more complex tasks, but it could also be advantageous for amateur programmers as OpenRefine provides a smooth pathway from direct data manipulation through graphical interface to more powerful data manipulation using programming functions.

These advantages of OpenRefine have made it a popular tool in GLAM (Galleries, Libraries, Archives, and Museums) [36] – [39]. Similar to other datasets in GLAM, the outcome datasets of online crowdsourced transcription projects are usually complex and messy. Due to the crowdsourced nature of the projects, the outcome datasets usually contain a large number of records and each record would contain numerous attributes, like volunteer transcriptions and subject metadata. The data outcome of *Shakespeare's World*, for example, was too large to properly open in Excel. By leveraging OpenRefine, owners of online crowdsourced

transcription projects could explore and clean their data through various manipulations like filtering and field-splitting, and ultimately acquire quantitative information to help them better understand the state of their project and iterate workflows for their future projects.

Despite OpenRefine's great potential to support online crowdsourced transcription projects, it is not very effective at handling semi-structured data fields, such as JSON. While users could use relatively simple GREL syntax to parse JSON formats in OpenRefine, this process usually requires the users to have a comprehensive understanding of all attributes used in the dataset. This pre-condition can be really challenging to achieve using low-programming tools like OpenRefine alone, especially for long-term, large-scale online crowdsourced projects. The following screenshot (Figure 6) provides three sample volunteer classifications recorded for *Shakespeare's World* during different phases of the project. Each of them contains a different attribute naming convention and records different numbers of attributes. It will be challenging to manually inspect the dataset and find all attribute names and potential variations accurately using only low-programming tools.

```
{
  "1071520": {
    "retired": null,
    "Year": "1585",
    "Author": "Philip Sidney",
    "Fol./p.": "3v-4r",
    "RF/JPG#": "50769.jpg",
    "Call Number": "H.b.1",
    "Primary Genre": "Literary Works",
    "Title and/or Description": "Countess of Pembroke's Arcadia"
  }
}

{"1120243": {"retired": null, "Filename": "015236.jpg"}}

{"1279559": {
  "retired": {"id": "1456562", "workflow_id": "205", "classifications_count": "19", "created_at": "2015-12-10T15:44:29.849Z", "updated_at": "2016-09-24T01:47:47.255Z", "retired_at": "2016-09-24T01:47:47.236Z", "subject_id": "1279559", "retirement_reason": null},
  "Genre": "Letters",
  "Title": "State letters from 1533 to ca. 1630 [manuscript], copied ca. 1650.",
  "Author": "",
  "Origin": "",
  "Filename": "82350.jpg",
  "Luna URL": "https://luna.folger.edu/luna/servlet/detail/FOLGERCM1~6~6~1071141~176475",
  "Priority": "3430",
  "Page Sort": "210",
  "Hamnet URL": "https://hamnet.folger.edu/cgi-bin/Pwebrecon.cgi?BBID=233143",
  "Call Number": "V.a.239",
  "Page Number": "p. 175 [i.e. 174]"
}}
```

Figure 6 – All three of these fields were actual subject_data appeared at different stages of *Shakespeare's World*

3.1.2 Tableau

Tableau is another low-programming tool that supports data visualization and a wide range of capabilities for converting complex data into insights. While OpenRefine focuses on browsing and cleaning large and complex datasets, Tableau emphasizes the visualization aspect, enabling users to surface patterns in the datasets. Like OpenRefine, Tableau also supports filtering datasets via different fields and more advanced data manipulations using Tableau Calculation Language. We believe it is a valuable complement to OpenRefine, specifically addressing the visualization aspect of data analysis.

Together, these tools offer robust capabilities for browsing, analyzing, and visualizing data and support sophisticated data manipulation through approachable graphical user interfaces.

3.2 Approaches to Data Cleaning

To address the complexity of *Shakespeare's World* outcome dataset and make it more usable in low-programming tools like OpenRefine and Tableau, we defined a data cleaning procedure for the dataset. This approach was informed by Arthur D. Chapman's argument that data cleaning should encompass three to four sub-processes "Data checking and error detection; Data validation; Error correction and Improvement of the error prevention process" [40] to make them "'fit for use' by users through reducing errors in the data and improving their documentation and presentation" [41]. He also argued that correcting errors and eliminating bad records are essential steps despite being time-consuming and tedious and that corrections should be documented and traceable [40].

Researchers like Rawson and Muñoz have argued against “data cleaning” [42], specifically because preserving “messy” data could potentially reveal deeper insights and help foster a more critical engagement with the material. This is especially true when cleaning outcome datasets for online crowdsourced transcription projects. Aroyo and Welty also suggested that in the context of big data and human annotation, truth is subjective and best understood as a consensus among diverse perspectives, a concept they defined as “crowd truth” [43]. The goal of our cleaning process should not be seeking a “perfect” transcription of the manuscripts, but to accurately capture the actual interactions of volunteers on the platform with the materials. This approach underscores the value of inaccuracy, disagreements, and nuances in individual transcriptions as reflections of volunteers' unique engagements with the materials and could potentially be used to help better understand their language and cultural identities. On the project level, these inaccuracies could also provide insights into the significant events that happened during the lifetime of *Shakespeare's World*. The most accurate interpretations of *Shakespeare's World* volunteer contribution need to take both the project's trajectory and volunteers' contribution patterns into consideration. So, while complete and accurate aggregated transcriptions were a goal of the project and are a key part of *Shakespeare's World's* outcome dataset, the goal of this project was to make the raw data more amenable for exploration and visualization.

Based on the discussions about data cleaning with consideration of the current state of *Shakespeare's World* outcome dataset and the limited time frame of the project, the following data cleaning goals were made:

1. Improve terminology consistency: ensuring all attribute names were consistent throughout the dataset,
2. Validate data and check for duplications,
3. Flatten key attributes to ensure easy information retrieval,
4. Address ethical and privacy concerns,
5. Retain detailed documentation of data cleaning steps,
6. and Document errors we encountered and use that information to infer rules to prevent errors in other online crowdsourced transcription projects or future relaunch of *Shakespeare's World*.

3.3 Cleaning Shakespeare's World Outcome Dataset

Using Zooniverse's data export function, we requested a "classification export" representing the latest state of the project on September 20, 2023 (refers to as raw dataset). The raw dataset contains 14 fields and 203,390 entries. However, the number of fields in the raw dataset does not accurately represent the amount of information documented. Within the raw output generated by the Zooniverse platform, multiple fields, such as metadata and annotation, are nested and/or in a semi-structured format while other fields contain duplicated information (see Table 1, Figure 23).

Hierarchy Level	Attribute Names
First Level	classification_id, user_name, user_id, user_ip, workflow_id, workflow_version, created_at, gold_standard, expert, metadata, annotation, subject_data, subject_ids
Second Level	metadata.started_at, metadata.user_agent, metadata.finished_at, metadata.user_language, metadata.subject_dimensions, metadata.viewport, metadata.utc_offset, metadata.user_group_ids, metadata.seen_before, metadata.session, annotation.T0, annotation.T2, annotation.T3, annotation.init, subject_data.subject_id
Third Level	annotation.T2.text, annotation.T2.type, annotation.T2.complete, annotation.T2.endPoint, annotation.T2.startPoint, annotation.T2.x, annotation.T2.y, annotation.T2.type, annotation.T2.width, annotation.T2.height, annotation.T2.tag, subject_data.subject_id.retired, subject_data.subject_id.Author, subject_data.subject_id.Genre, subject_data.subject_id.Filename, subject_data.subject_id.Page Number, subject_data.subject_id.Origin, subject_data.subject_id.Call Number, subject_data.subject_id.Luna URL, subject_data.subject_id.Hamnet URL, subject_data.subject_id.Primary Genre, subject_data.subject_id.Tile and/or Description, subject_data.subject_id. Fol./p., subject_data.subject_id.Year, subject_data.subject_id.RF/JPG#, subject_data.subject_id.file name, subject_data.subject_id.imgname, subject_data.subject_id.# dr7objid, subject_data.subject_id.Priority, subject_data.subject_id.Page Sort
Forth Level	annotation.T2.endPoint.x, annotation.T2.endPoint.y, annotation.T2.startPoint.x, annotation.T2.startPoint.y, subject_data.subject_id.retired.created_at, subject_data.subject_id.retired.retired_at, subject_data.subject_id.retired.classification_count, subject_data.subject_id.retired.retirement_reason, subject_data.subject_id.retired.workflow_id, subject_data.subject_id.retired.subject_id, subject_data.subject_id.retired.updated_at, subject_data.subject_id.retired.id

Table 1 – Hierarchy of Raw Output Dataset Attributes

To facilitate future reuse, and especially empower analysis using low-programming tools, we used Python (3.9.18) and pandas DataFrame to process the raw dataset. The following steps were taken to improve the structure of *Shakespeare's World* raw output dataset while preserving individual classifications recorded for every submission and essential metadata about volunteers' contribution patterns and subjects.

3.3.1 Normalizing Dataset

The first step we took was to identify and break down the nested attributes so they could be easily accessible in low-programming tools. Within the raw dataset, three fields were identified to contain nested values: `annotation`, `metadata` and `subject_data`.

3.3.1.1 `annotation`

The `annotation` field from the raw dataset encapsulates the classifications contributed by volunteers for each submission under three distinct task labels: `T0`, `T2`, and `T3`. The `T1` label was not found in the raw output dataset. The task labels are defined as follows:

- `T0` – identifying if the subject (typically a digital copy of a single manuscript leaf) is blank. Each classification result is captured as a Boolean value indicating blank (`true`) or not (`false`),

- T2 – the primary task for annotating graphical and/or text elements within the subject, this field contains an attribute (`type`) for the type of classifications (`graphic`, `image` or `text`), the starting (`x` and `y`) coordinates and the width and height of the element for all graphical elements or the starting and the starting and ending (`x` and `y`) coordinates for each text transcription segment and a Boolean attribute (`complete`),
- and T3 – identifying if the subject has been fully transcribed, and the result is captured as a Boolean value indicating if the page is finished (`True`) or if there are still things to transcribe (`False`)

During the early alpha and beta and later testing phases, an additional task labeled `init`, was also used. This task asked volunteers to classify the subject as text-only, image-only, or mixed. The response was recorded as a text value encapsulated within a list. Through testing, this task was deemed unnecessary, and so is not present in the post-public launch data.

Shakespeare's World allowed volunteers to submit a diverse range of classifications, from straightforward single classifications to complex combinations of tasks among all possible task categories (T0, T2, T3, and `init`). This variability presents a challenge in data restructuring, particularly when considering how to preserve structural consistency for all variations of classification tasks completed per submission and how they are recorded within each annotation in the final dataset. Fully flattening the classifications into each category's subfields could potentially result in fragmented lists of partial information, further complicating data comprehension instead of simplifying it for potential users we are targeting. Taking this

potential drawback into consideration, we decided to partially preserve the nested structure within the dataset. By reorganizing the `annotation` field into the four fields according to their task categories—`T0`, `T2`, `T3`, and `init`, the revised dataset preserves the intrinsic structure of detailed information nested within each task while also enhancing its comprehensibility by offering clear cues for all task variants that were made available to the volunteers during *Shakespeare's World*.

3.3.1.2 metadata

The `metadata` field in the raw dataset is structured as key-value pairs and mostly contains information captured by the Zooniverse platform for submitted classification sessions. It records the following metadata for each submitted session:

- `started_at` and `finished_at` – session starting and ending timestamps in ISO 8601 format,
- `user_agent` – the browser or interface used for the session,
- `live_project` – the status of a project, stored as a Boolean value indicating whether the project was open to the public for participation (`True`) or the project was in a development or testing phase (`False`)
- and `user_language` – the volunteer's preferred system languages.

During alpha and beta testing for the sequential distribution of manuscripts, three additional attributes were recorded for 620 classification sessions:

- `subject_dimensions` – the dimensions of digitized manuscripts presented to volunteers during the session
- `viewport` – the dimensions of the content viewing area for the session
- and `utc_offset` – the time zone offset from Coordinated Universal Time (UTC).

During the testing phases and early public phase of the project, an attribute `seen_before` was recorded for 4415 classification sessions to indicate if the subject has already been presented to the same registered volunteer.

`user_group_ids` was an attribute presented in 89885 classifications sessions during the public phase and later testing for the sequential distribution of manuscripts. However, this attribute was only used to store values 122 times, and all of them were classifications made by Zooniverse employees.

Another attribute that was infrequently used was `session`, it was only recorded for five classifications to support the testing for sequential distribution of manuscripts.

The `metadata` field features multiple attributes with different purposes. Abiding by the cleaning principle we've decided on, the following steps were taken to flatten the metadata field. Four attributes that were consistent throughout the project were separated into their own independent fields. This allows for easy filtering by the following key information about the submitted sessions: timestamps, project status, browser platform, and system language. Other attributes predominantly used during testing periods (including `subject_dimensions`,

viewport, and session) which do not directly contribute to the core analyses we have identified, were consequently omitted to simplify the dataset and minimize potential confusion for future users. Given the sporadic appearance of `user_group_ids`, only recording valid values for sessions submitted by Zooniverse employees and limited impact on the broader project goal, we decided to drop the attribute in the revised dataset. `utc_offset` was incorporated into the time stamps and removed from the final dataset, the exact steps and rationale are discussed in the data validation section.

3.3.1.3 `subject_data`

The `subject_data` field in the raw dataset is the most complicated among the three nested fields, primarily because two different sets of information within one key-value pair are included in this field, identified by their `subject_id`. The first set includes retirement information, which is automatically generated by the Zooniverse platform. This includes information like the retirement status of the subject and its retirement reason. The second set comprises archival data provided by Folger, containing essential information about each subject, like the title and author of the original manuscript and the genre of the original document. When examining the `subject_data` field, attributes from both sets of information contained variations in their naming conventions throughout the project's duration, presenting issues with data consistency and challenges for future interpretation.

The retirement information was stored as a key-value pair nested within the field with the identifier `retired`. Some of the subjects were not retired throughout the project and would have

a null value under the retired key. Subjects that were successfully retired usually contain the following metadata about their retirement:

- `created_at` – the timestamp when the subject was first transcribed, converted to UTC and in ISO 8601 format,
- `retired_at` – the timestamp when the subject was retired, converted to UTC and in ISO 8601 format,
- `classification_count` – total number of classifications contributed towards the subject before its retirement,
- and `retirement_reason` – the retirement reason recorded by the Zooniverse system when the subject was retired. Possible values from this attribute contained `consensus`, `other` and a couple of alternatives. However, we will discuss later that the term “consensus” here does not necessarily mean the volunteers’ classifications have resulted in unanimous transcriptions of the original manuscripts.

Besides the previous information, several redundant pieces of information about the subject were also recorded. This includes information already covered by other attributes in the dataset like `workflow_id` and `subject_id` and information that does not provide much additional context for the target users (like `updated_at`, which contains the timestamp when the subject was last updated before retirement). Another `id`, different from `subject_id`, was also stored within this key-value pair. Due to its lack of utilization in other contexts of *Shakespeare’s*

World dataset and the lack of documentation, we were not able to accurately describe the meaning of this attribute.

To make critical information about the subject's retirement more accessible, we first isolated the entire key-value pair into a separate field called `retirement_info`. To make it easier to identify and filter retired subjects, we also added a `retired` attribute and assigned `True` to all subjects with valid retirement information. We then parsed the most valuable attributes for reuse like `created_at`, `retired_at`, `classification_count` and `retirement_reason` into independent fields. Two redundant attributes, `workflow_id` and `subject_id`, were then dropped to streamline the dataset, and the remaining attributes were preserved within the `retirement_info` field.

The archival data was stored within the original `subject_data` field. Some of the subjects from the alpha and beta testing periods used different terminologies or lacked detailed archival information altogether. The original field contained straightforward attributes like the title, author, genre and filename of the original manuscript, as well as the following attributes:

- `Page Number` – page number, uses abbreviations to represent the front and back sides of a sheet, `r` for recto (front side) and `v` for verso (back side),
- `Origin` – century, year or exact date of the original manuscript
- `Call Number` – the unique identification code of the original manuscript assigned by Folger,

- `Luna URL` – web address to the original digital copy of the subject on LUNA, Folger's digital image collection platform,
- and `Hamnet URL` – web address to the detailed record of the subject on the Hamnet catalog.

Some attributes used during internal testing periods employed a slightly different naming schema. We examined all those attributes and were able to find the following correspondence:

- `Primary Genre` was used similar to `genre` (but indicating a wider range of materials used for alpha and beta testing period),
- `Title and/or Description` was used similar to `Title`,
- `Fol./p.` was used similar to `Page Number`,
- `Year` was used similarly to `Origin`,
- and `RF/JPG#` and `file name` were used similarly to `Filename`.

One subject (`subject_id: 458922`) also used attributes `# dr7objid` and `imgname`, both attributes were only used once. Two other attributes, `Priority` and `Page Sort`, used during the internal testing for the sequential distribution of subjects were also included in `subject_data` field for all subjects.

In February 2024, Folger announced that they would be migrating their digital collections from LUNA to a new site built on Islandora 2 [44]. Back in 2022, Folger had already announced migrating its catalog from Hamnet to TIND ILS [45]. And in the 2024 announcement, LUNA was declared to retire by the end of March 2024. To ensure the catalog and digital copies of the

manuscripts used in *Shakespeare's World* project could still be accessible by future users, we worked with Emily Wahl, Folger Metadata and Digital Asset Management Librarian, to obtain the latest URLs for every subject in the dataset.

To address the nested nature of the archival information and the inconsistent naming conventions, the following steps were taken. We first dropped the classification for `subject_id` 458922, considering the subject was only classified once during the testing period and used completely different attribute conventions. We then renamed and updated the variables based on the following schema:

Original Attribute Names	Revised Attribute Names
Title	title
Author	author
Genre	genre
Primary Genre	
Filename	filename
RF/JPG#	
file name	
Page Number	folio/page
Fol./p,	
Origin	
Year	origin
Call Number	call number
Luna URL	image url
Hamnet URL	catalog url

Table 2 – Attribute Name Mapping Table

After standardizing these attributes, we isolated frequently queried information such as `title`, `author`, `genre`, `folio/page` and `origin` into independent attributes. We decided to keep

call numbers in the `subject_data` field, because their unique structural format, such as “L.a.846”, makes it unlikely to be mistakenly queried. We also chose to leave the updated Islandora URL and TIND URL in the original field, because these URLs are easily distinguishable in the graphical interface while rarely being filtered directly. The two priority attributes for sequential distribution testing were also preserved.

By taking the previous data cleaning steps, we ensured that the cleaned dataset (see Table 3, Figure 24) retained most of its original information while being more focused, accessible, and more immediately relevant to various user groups, especially those using low-programming tools for analysis.

Hierarchy Level	Attribute Names
First Level	Subject_ids, classification_id, user_id, user_ip, started_at, finished_at, T0, T2, T3, T-init, origin, folio/page, genre, retired, retirement_reason, retirement_info, created_at, user_agent, user_language, subject_data, live_project, workflow_id, workflow_version
Second Level	T2.text, T2.type, T2.complete, T2.endPoint, T2.startPoint, T2.x, T2.y, T2.type, T2.width, T2.height, T2.tag, retirement_info.id, retirement_info.classification_count, retirement_info.updated_at, retirement_info.retired_at, subject_data.call number, subject_data.image url, subject_data.catalog url, subject_data.Priority, subject_data.Page Sort
Third Level	T2.endPoint.x, T2.endPoint.y, T2.startPoint.x, T2.startPoint.y

Table 3 – Hierarchy of Cleaned Output Dataset Attributes

3.3.2 Ethics and Privacy

Initially, we had hoped to recognize the contributions of the volunteers by preserving their `user_name` in the outcome dataset. However, when going over the dataset, we noticed several occasions where `user_name` explicitly reveals potential registration emails of volunteers. Considering the cleaned dataset will be published online for public access, this was caught as a privacy violation concern.

To thoroughly assess the situation, we filtered the `user_name` fields according to common domain names. From the filtered username list, we then cross-examined the potential links on search engines, like Google. This process led us to 22 usernames bearing close resemblance to valid email addresses. After collecting the list of potential email addresses, we drafted a Google Form (see Figure 22) to inquire how these volunteers wish to be acknowledged in the public dataset. This form was distributed to the implicated individuals using Blind Carbon Copy (BCC) in February 2024, ensuring confidentiality. In the Google Form, we invited the volunteers to enter alternative usernames; if the volunteers did not respond to the inquiry, we would, by default, replace their usernames with an arbitrary, platform-generated Zooniverse ID in the Zooniverse generated output.

By the deadline of February 20th, 2024, set in the update form, only one response was received, opting to use their Zooniverse ID instead. Hoping to receive more responses, we left the form open. However, as of March 2024, no additional responses had been received.

Since *Shakespeare's World* was an early project launched on Zooniverse, the platform had yet to implement the option for volunteers to provide a “Publishable Name” upon registration, so we cannot reliably determine how `user_name` should be presented in the public dataset. During our discussion, Van Hyning also raised an argument that volunteers may not fully comprehend the implications of their contributions being made public under their usernames, especially when those contributions are less frequent or perceived as “suboptimal”. Publicly attributing contributions could inadvertently introduce a competitive element and future study of engagement patterns of *Shakespeare's World* might render volunteers with a negative image, which defeats the initial premise of allowing volunteers to contribute only what they felt comfortable with. Given the arbitrary `user_id` was already generated for every user at registration by the Zooniverse platform and could be used as a unique identifier across different projects, we eventually decided to drop the `user_name` field altogether to eliminate the risk of inadvertently disclosing personally identifiable information and violating volunteers' privacy.

3.3.3 Validating Data and Handling Missing or Incomplete Data

After normalizing and anonymizing the outcome dataset, we validated it by examining all fields for missing or inconsistent data. Since null values are often intentionally employed to encode specific meanings and variations in naming conventions across different phases of the project, the presence of null values in the dataset is anticipated, serving as an indicator for data that may have been unrecorded or inconsistently labeled.

The first inconsistency we addressed was the different representation of time for time stamps between the early testing periods and the public phase of the project. During the early testing periods, the time stamps were recorded in UTC (denoted with a `Z` after the ISO 8601 timestamps) with a `utc_offset`. While the retirement timestamps also employed the UTC format, we decided to follow the more streamlined timestamp format used during the public period (local time with UTC offset). This format avoids the need for an additional independent `utc_offset` attribute, while also preserving the detailed information about the session's time zone. To assist analysis of the retirement timespan and contribution pattern, we also parsed `finished_date` from `finished_at` for easier analysis.

After consolidating the timestamp formats, we validated the dates for all classifications to check if all of them fell within the project timeline. By checking the earliest and latest `finished_date` against *Shakespeare's World's* beginning and end dates, we spotted an anomaly in the earliest `finished_date`. The earliest `finished_date` returned was September 03, 2012. Considering that the project's internal testing periods only started in 2015, and its `workflow_version` indicating the public period, this was clearly an error. To find all classifications with an invalid `finished_date`, we filtered the dataset by `finished_date` earlier than 2015. This returned 9 classifications, all showing a `finished_date` of September 03, 2012. Since this was an extremely small subset of the classifications and the annotations submitted with the classifications were all valid, we decided to record the

`classification_ids` in an `invalid_date_id_list` and only exclude them from date/time-sensitive analysis.

3.4 Preparing Dataset for Analysis

To analyze the contribution patterns of volunteers, we first decided to calculate the session durations of each classification. Using the `finished_at` and `started_at` attributes, we could easily calculate the session duration by subtracting the timestamp of `started_at` from that of `finished_at`. The time differences were then converted to seconds using `.total_seconds()` method and stored as `duration`.

To validate this `duration` value, we checked its maximum and minimum values. Anomalies were spotted on both ends. The shortest session duration was -894 seconds while the longest session duration was 47,693,564 seconds (which is more than 552 days!). To better understand the irregular values, we extracted the `classification_ids` of both extreme values and examined their exact `started_at` and `finished_at` timestamps in OpenRefine. Both extreme values reflect the timestamps from the dataset correctly. We decided to first tackle the unreasonably low values by creating two separate lists of `classification_ids` for negative durations and zero duration.

Only three classifications from the public period between 2016 and 2017 resulted in negative durations, and only two classifications resulted in zero duration. While the classification tasks submitted for these zero-duration sessions were both binary classification T0 and T3 tasks, it

is extremely unlikely for a classification to be made within milliseconds (considering the precision of the timestamp used by the Zooniverse system was 0.001 seconds). It is very likely that these five classifications all ran into a timestamp error. Considering these sessions recorded valid classifications and reasonable dates, their `classification_ids` were stored in `negative_zero_duration_id_list` and will only be excluded from any analysis that is session duration sensitive.

The unreasonably high values, on the other hand, are a little hard to conclude. We first print out a list of the top 10 highest durations and they are all at least 38,000,000 seconds long. After validating them in OpenRefine, we found that their timestamp was indeed pointing to these extreme values. To get a better understanding of the situation, we also examined the top 100 durations. Considering that the high values did not fall rapidly past the top 10 and even the top 100th value was still at a duration of more than 192 days, it was still difficult to conclude which retirement timespan should be considered reasonable. To get a better understanding of the distribution of duration, we plotted a histogram for the distribution of all positive session durations (see Figure 7).

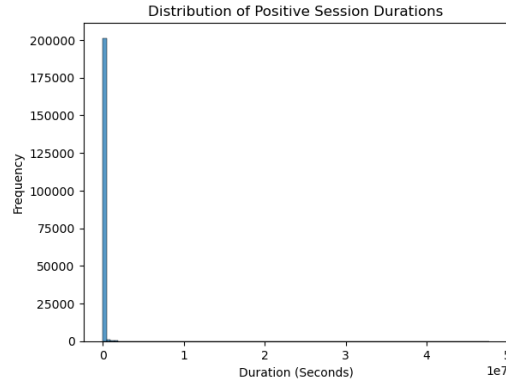


Figure 7 – Distribution of Positive Session Durations

The histogram showed extremely skewed distributions, with the majority of the duration being smaller than 100 seconds and a long tail. To account for this extremely skewed data, we calculated the statistical maximum with outliers ($Q3 + 1.5 \text{ IQR}$), which is 1741. This is significantly smaller than the previous maximum values we've found. Temporarily removing the outliers, we replot the histogram to observe the distribution. While the distribution is still very skewed, with the majority of the duration resting within 500 seconds, the variation between maximum and minimum durations is not as extreme as the visualization with outliers (see Figure 8).

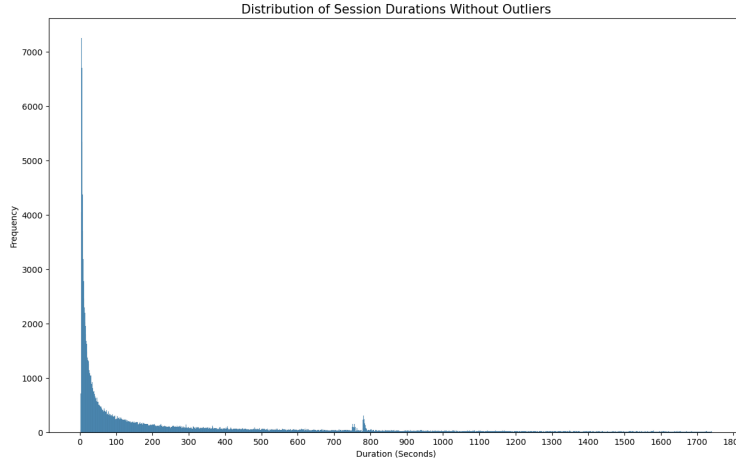


Figure 8 – Distribution of Session Durations without Outliers

To better understand the retirement pattern of the subjects, the retirement timespan of each subject was calculated. We calculated this value using the difference between the timestamps for `created_at` and `retirement_at`. To mitigate potential variations in retirement timespans, we chose to calculate both the retirement timespan in hours and days and preserved the precision to 0.01. After calculating the retirement timespans, we validated the fields by checking the minimum retirement timespan in hours and the maximum retirement timespan in days. The maximum retirement timespan is a reasonable value, just a little over 1,103 days – below *Shakespeare's World's* lifespan. The minimum retirement timespan, on the other hand, was zero. To investigate this retirement timespan, we extracted its corresponding `subject_id` (11207034) and examined the three classifications contributed to this page.

Here, we found an interesting loophole for potential error. By design, the Zooniverse platform would prevent multiple submissions from the same users. However, since the platform allows

non-logged-in users, this identity verification could not be reliably enforced. The three classifications 113834653, 113834654 and 113834657 were all submitted milliseconds away from each other. Their salted and hashed IP addresses were also the same, indicating a high possibility that these submissions were made on the same machine. When examining the individual transcriptions by the three classifications, we found that they were all the same. This means the three classifications were actually a false resubmission accidentally allowed by the Zooniverse platform.

Because the subject contains only very little text, the anonymous volunteer might have felt confident about annotating the whole document and submitted a classification with all three tasks. Something went wrong during the submission that pushed three identical classifications to Zooniverse's backend. Because these classifications were essentially the same, the system would happily retire the subject as "consensus". In reality, this transcription was never vetted against any other volunteers' transcriptions. This is an edge case, considering it would require a submission to complete all three classification tasks and be submitted three times and this situation only occurred once during the project's lifespan. However, this might be an instructive case for other Zooniverse projects or workflows which consist of fewer task variations.

4. Preliminary Data Analysis & Visualizations

To better understand *Shakespeare's World* volunteer contribution pattern and discover potential anomalies during the project, we performed the following preliminary analysis from the raw output's metadata. As a validation of the usability of the cleaned dataset for people with limited programming experience, most of the following analysis and visualizations were performed using OpenRefine and Tableau on the cleaned outcome dataset.

4.1 Summary Statistics of Contributions

The first topic we examined was the number of classifications submitted by volunteers per month. The nine classifications with false finished dates in 2012 were excluded from this analysis. We explored the contribution pattern difference between anonymous volunteers and logged-in registered volunteers and the contribution breakdown for each task category (T0, T2 and T3). The following table contains the summary statistics about the classifications submitted by volunteers per month (all decimal values have been rounded). Note that T0, T2 and T3 do not equal the number of classifications because each classification can consist of only one of those task categories or any combination of them. T2 only accounts for the number of submissions instead of counting the total words, sentences or lines transcribed for each submission. For statistics excluding testing classifications, both classifications from alpha and beta testing and later sequential subject delivery testing were excluded from the calculation.

			Task Category			
			T0	T2	T3	Classifications
Statistics and Volunteer Status Category	Minimum	Minimum (Anonymous)	31	7	31	31
		Minimum excl. testing (Anonymous)	107	62	103	107
		Minimum (Logged-in)	1	0	1	1
		Minimum excl. testing (Logged-in)	112	99	111	112
	Maximum	Maximum (Anonymous)	9,864	4,315	8,665	9,864
		Maximum excl. testing (Anonymous)	9,820	4,282	8,622	9,820
		Maximum (Logged-in)	10,787	5,252	9,397	10,787
		Maximum excl. testing (Logged-in)	10,761	5,242	9,372	10,761
	Median	Median (Anonymous)	1,412	604	1,387	1,412
		Median excl. testing (Anonymous)	1,541	672	1,513	1,541
		Median (Logged-in)	1,749	1,139	1,707	1,749
		Median excl. testing (Logged-in)	1,781	1,171	1,770	1,781
	Average	Average (Anonymous)	1,818	790	1,731	1,818
		Average excl. testing (Anonymous)	1,995	866	1,898	1,995
		Average (Logged-in)	2,122	1,207	2,030	2,122
		Average excl. testing (Logged-in)	2,292	1,305	2,193	2,292

Table 4 – Summary Statistics for the Number of Classification Submitted per Month

4.2 Contribution Pattern – Volume

Using Tableau, we explored how volunteers' contributions volume changed throughout the project and across different genres and task categories through a series of three visualizations to highlight the shifts in the number of classifications submitted over time and explore how that pattern might be influenced by task categories and subject genres. The nine classifications with false finished dates in 2012 were also excluded from this analysis.

The first visualization (*Figure 9*) depicts the month-to-month variations in the total classifications submitted for different user groups (anonymous versus logged-in volunteers). The second visualization (*Figure 10*) portrays how the total classifications contributed by volunteers changed over time for different task categories (T0, T2 and T3) changed over time. The last visualization (*Figure 11*) in this series examines the impact of subjects' genres on volunteer's contribution patterns over time. Because a different set of genres was included in the alpha and beta testing period compared to the ones included for the public phase, only classifications from the public workflow were included.

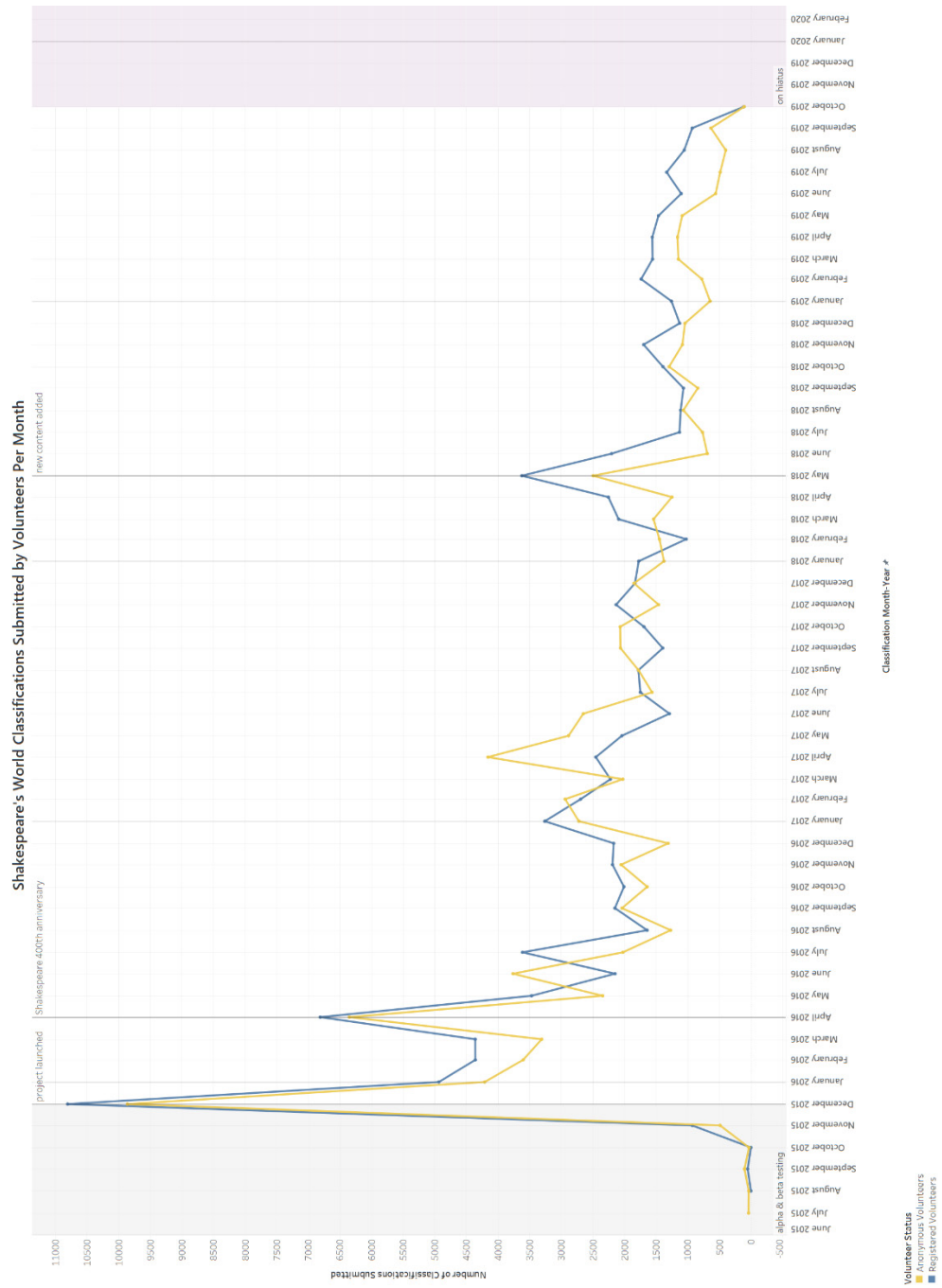


Figure 9 – Total Number of Classifications Submitted per Month

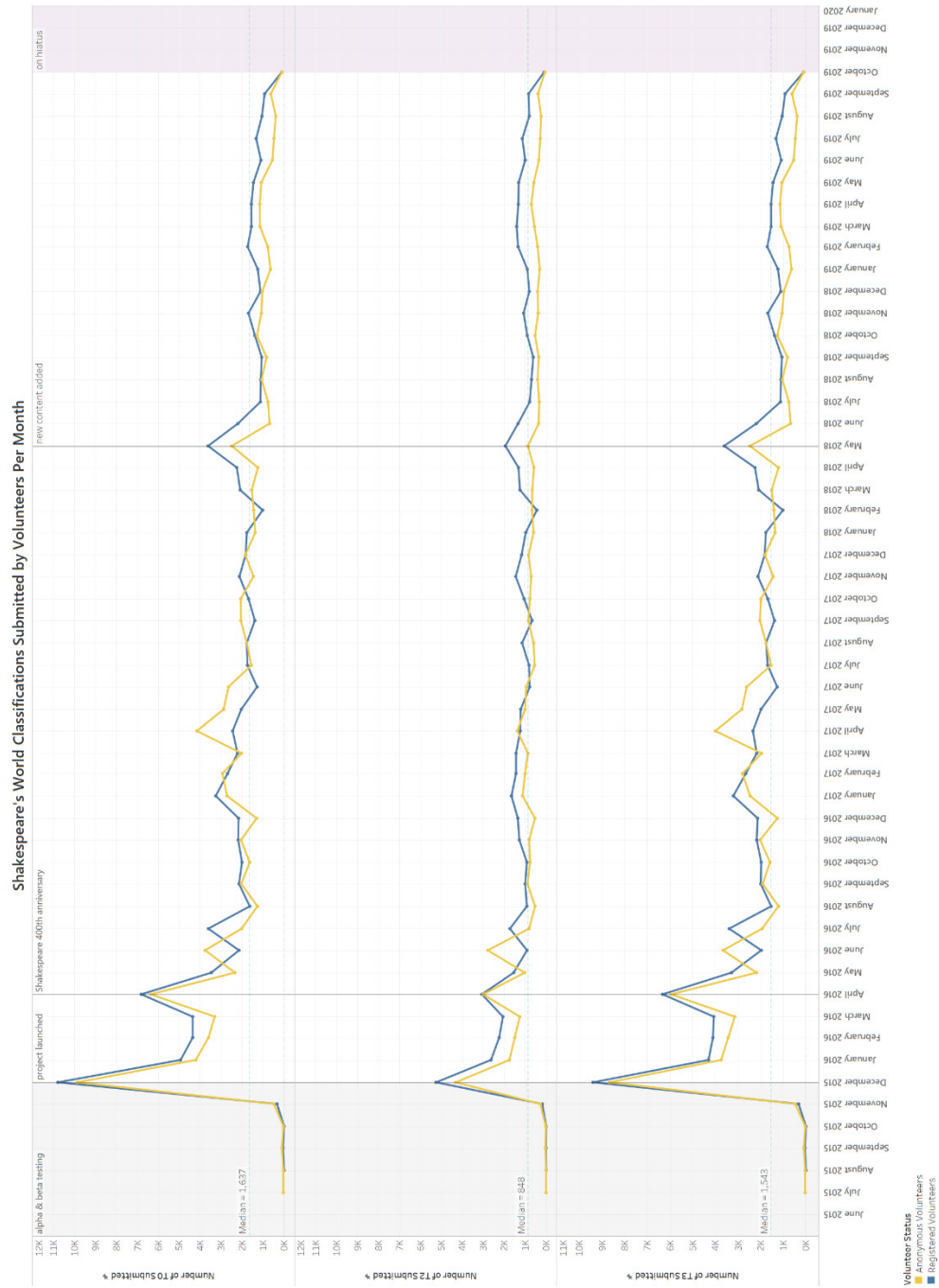


Figure 10 – Total Number of Classifications Submitted per Month - Varied by Task Category

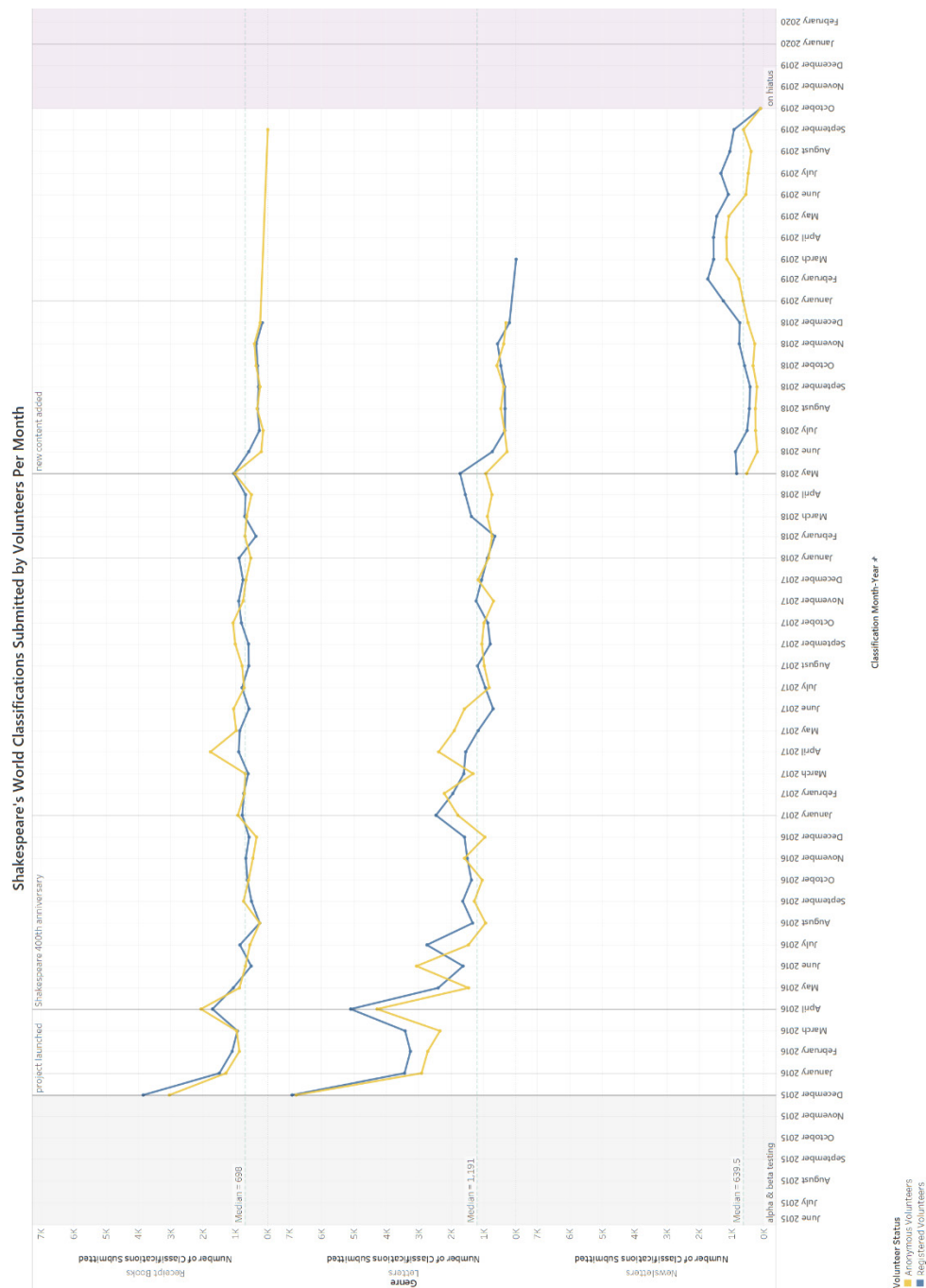


Figure 11 – Total Number of Classifications Submitted per Month - Varied by Subject Genre

4.3 Contribution Pattern – Logged-in Volunteers per Month

To better understand how volunteers' level of engagement, we further examined the contribution pattern of distinct logged-in volunteers. A series of three visualizations were created in Tableau to demonstrate the change in logged-in volunteers' engagement towards *Shakespeare's World*. Since we could not reliably identify anonymous volunteers based on the information collected for the project, only logged-in volunteers were included in this analysis. The first visualization (*Figure 12*) presents how the number of unique logged-in volunteers changed throughout the project. A supplementary bar chart (*Figure 13*) was created to show the number of unique logged-in volunteers who contributed each month. The last visualization (*Figure 14*) reveals the changing dynamic between the number of unique logged-in volunteers and the number of classifications contributed per month throughout the project.

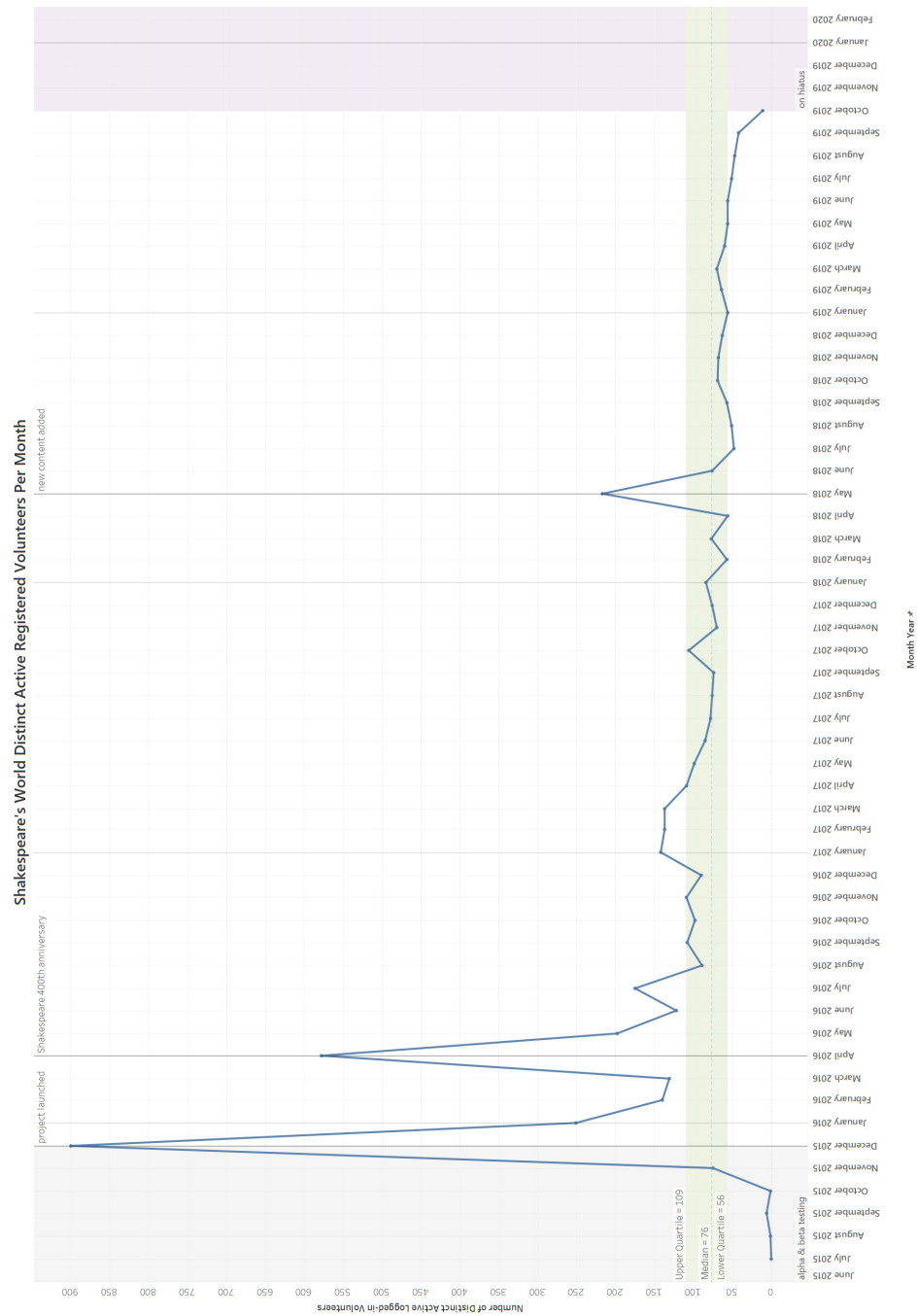


Figure 12 – Number of Distinct Active Volunteers per Month

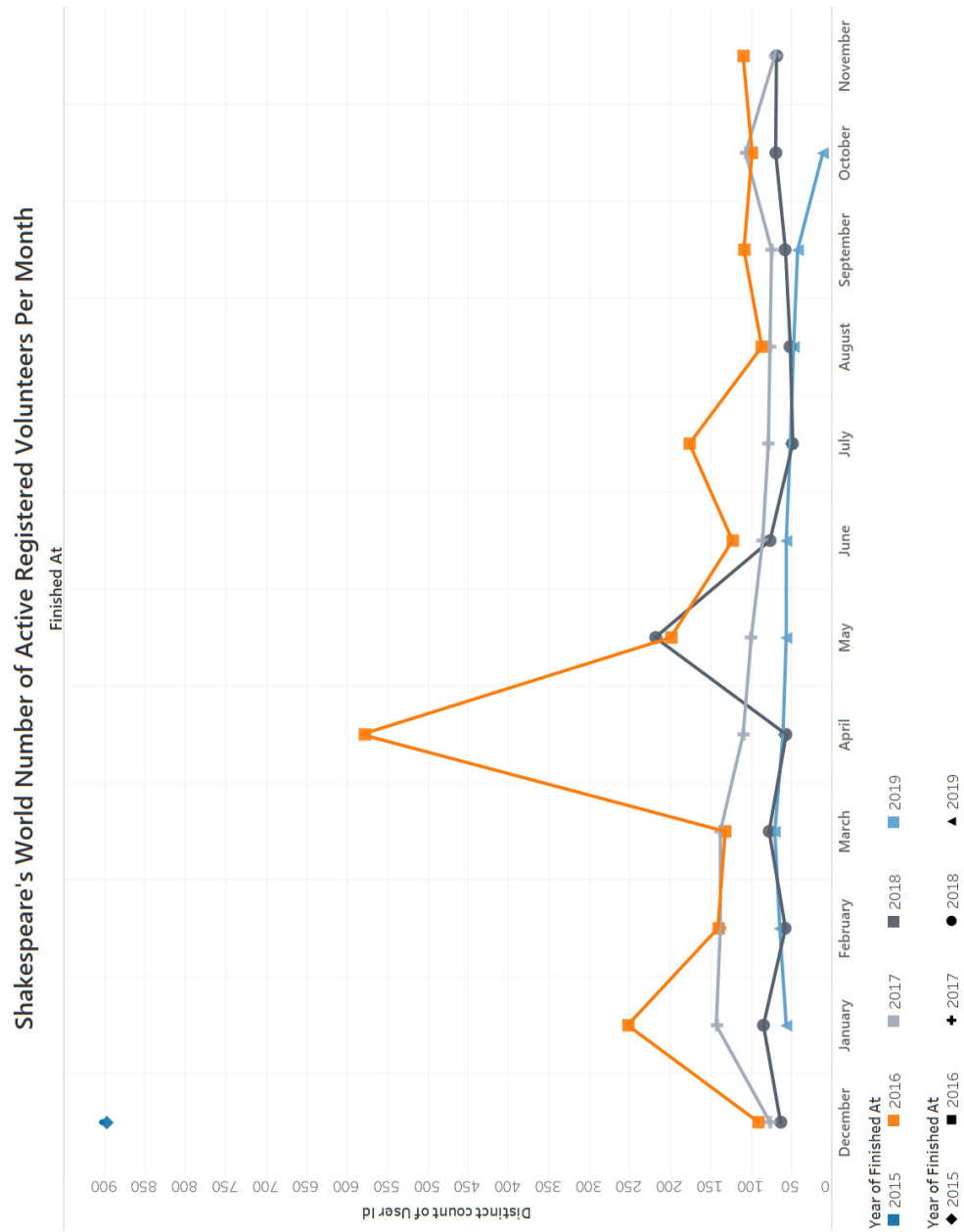


Figure 13 – Number of Distinct Active Volunteers Per Month Each Year

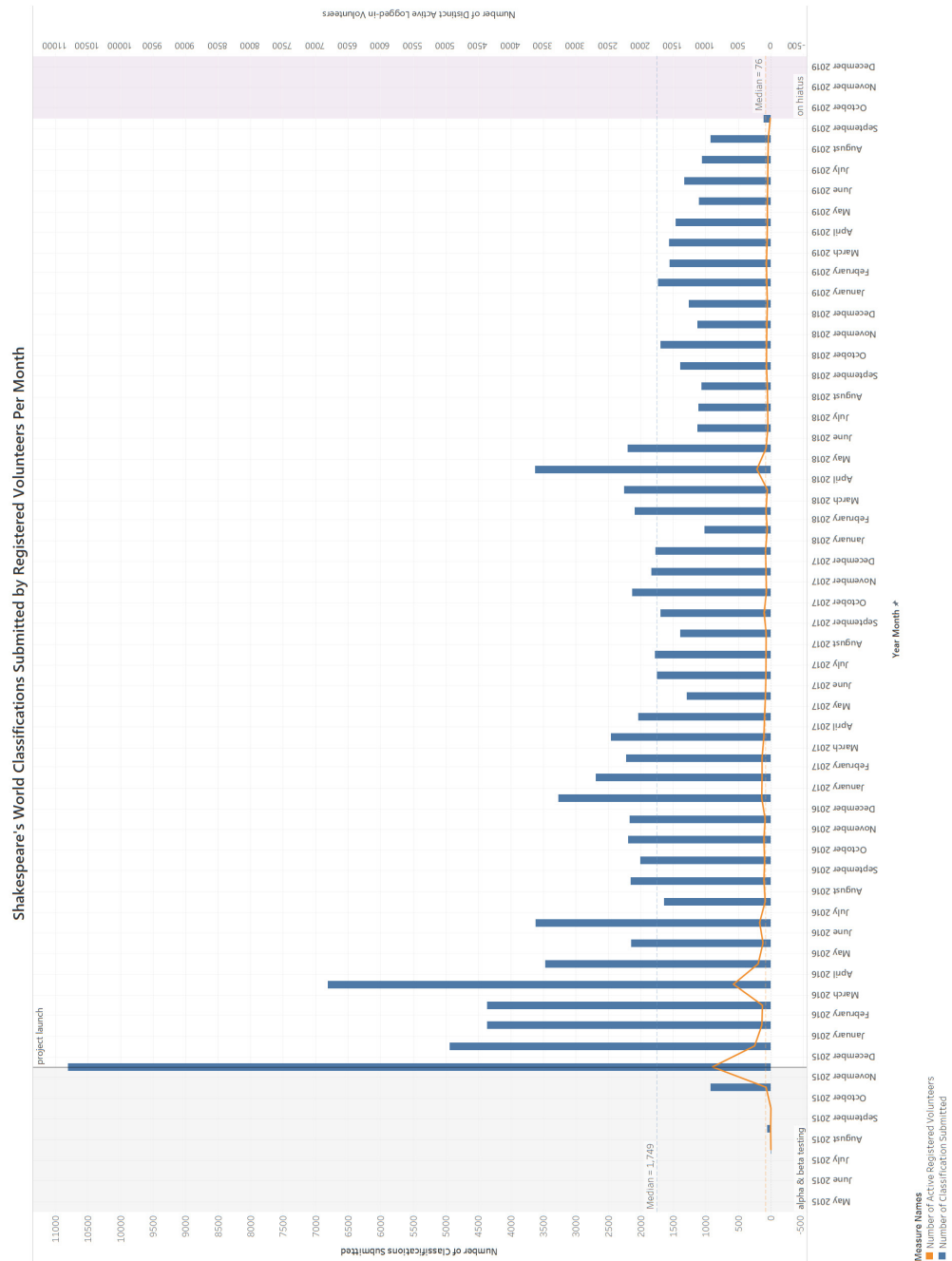


Figure 14 – Number of Distinct Registered and Active Volunteers & Number of Classifications Contributed by Them per Month

4.4 Contribution Pattern – Macro

Hoping to understand the macro contribution pattern of all *Shakespeare's World* volunteers throughout the project, a final visualization was created for volunteer classification counts (*Figure 15*). Each individual registered and logged-in volunteer was represented by a series on the x-axis according to their Zooniverse user id. While this id is arbitrary, it loosely encodes the sequence in which volunteers first registered for the platform. Since we could not reliably identify individual anonymous volunteers, they were presented as one entity. The anonymous volunteers were moved close to the center of the plot for visual clarity. The y-axis represents the continuous flow of time throughout the project with an increment of day. Each little dot represents the number of classifications submitted by individual volunteers per day. The color of the dots was used to encode the login status of the two volunteer groups (individual logged-in volunteers and the anonymous volunteers as a whole). We also picked the dots to be semi-transparent so the overall pattern could properly highlight more contributors' frequency. The nine classifications with false finished dates in 2012 were excluded from this analysis.

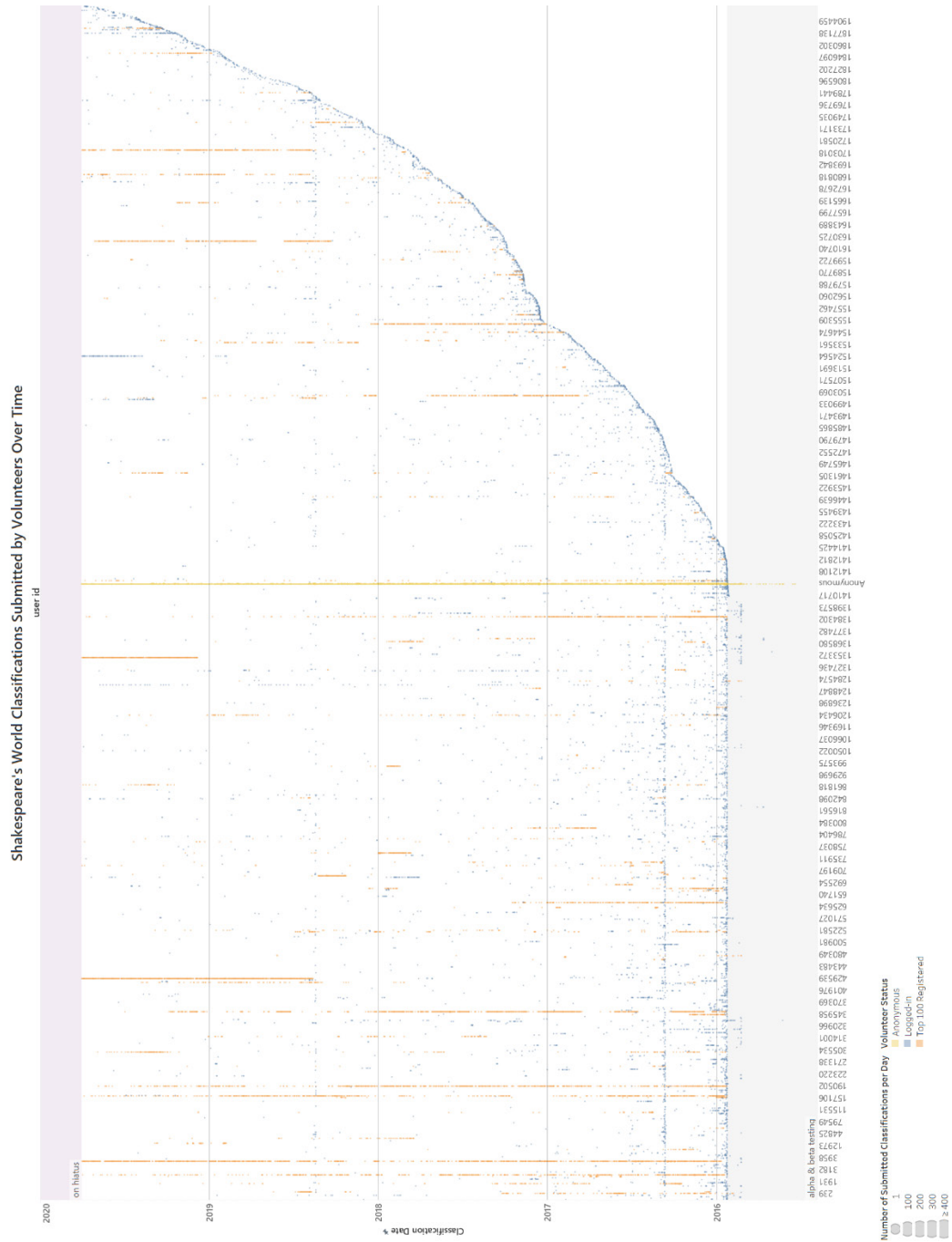


Figure 15 – Classifications Submitted by Volunteer Over Time

4.5 Subject Retirement Pattern & Estimated Transcribed Ratio

Using the calculated retirement timespan (days), we were able to plot a bar chart for the retirement pattern of all subjects transcribed during the project using Tableau. The initial visualization revealed an abrupt peak between 250 and 300 days. This anomaly coincides with Van Hyning's recollection of a forced retirement of subjects implemented by the *Shakespeare's World* team during the early stage of the project due to the low subject retirement rate. To better understand the scale and effect of this forced retirement, we performed the following analysis.

We first included additional context, retirement reason, within the bar chart (Figure 16). This shows a clear cluster of subjects with `null` retirement reasons within 300 days, with the bin between 250 and 300 being the most noticeable. Considering the forced retirement was made around September 2016 (about 300 days after the initial project launch), there is a high probability that this abrupt spike around 300 days was caused by the forced retirement.

To better understand the quality of transcription generated by these forced retired subjects, the aggregated transcriptions by Zooniverse were examined. This aggregated transcription was generated after *Shakespeare's World* went on hiatus and the data was organized in folders, named after each subject. Each subject folder contains three files:

- `transcription.txt`: the aggregated output transcriptions (if available) of the subject,
- `aggregation_data.json`: the raw aggregated transcriptions and metadata about the subject's aggregation in JSON structure,

- and `line_metadata.csv`: the metadata for each transcribed line.

We wrote a Python script to iterate through each subject folder and extract the `transcribed_lines` and `low_consensus_lines` from each subject's aggregation metadata. Then, we calculated a heuristic estimation of the transcribed ratio per subject using the following formula:

$$\text{est. transcribed ratio} = \frac{\text{transcribed lines}}{\text{transcribed lines} + \text{low consensus lines}}$$

Utilizing the calculated estimated transcribed ratio, a visualization (Figure 17) was created in Tableau to show the distribution of estimated transcribed ratios for all retired subjects. The graph is split according to four different retirement reasons. Some subjects' classifications could not be reliably aggregated and would result in zero transcribed lines and zero low consensus lines. We assigned zero to those `null` values and included them in this visualization to demonstrate the scale of retired subjects that did not result in any aggregated lines.

The following summary statistics (Table 5) about these estimated transcribed ratios were also calculated to provide additional insights into subjects' retirement patterns, and we excluded the null transcription ratio from this calculation of the following statistics.

		Statistics Category			
		Maximum	Minimum	Mean	Median
Subjects Retirement Reason	null likely forced retirement (n = 2,653)	1.000	0.500	0.652	0.610
	consensus (n = 2,998)	1.000	0.500	0.593	0.556
	nothing_here (n = 5)	0.634	0.506	0.570	0.543
	other (n = 3,061)	1.000	0.500	0.641	0.614
	Retired Subjects excl. subjects with zero transcribed lines (n = 7,497)	1.000	0.500	0.624	0.585

Table 5 – Summary Statistics for Est. Transcribed Ratio under Different Subject Retirement Reasons

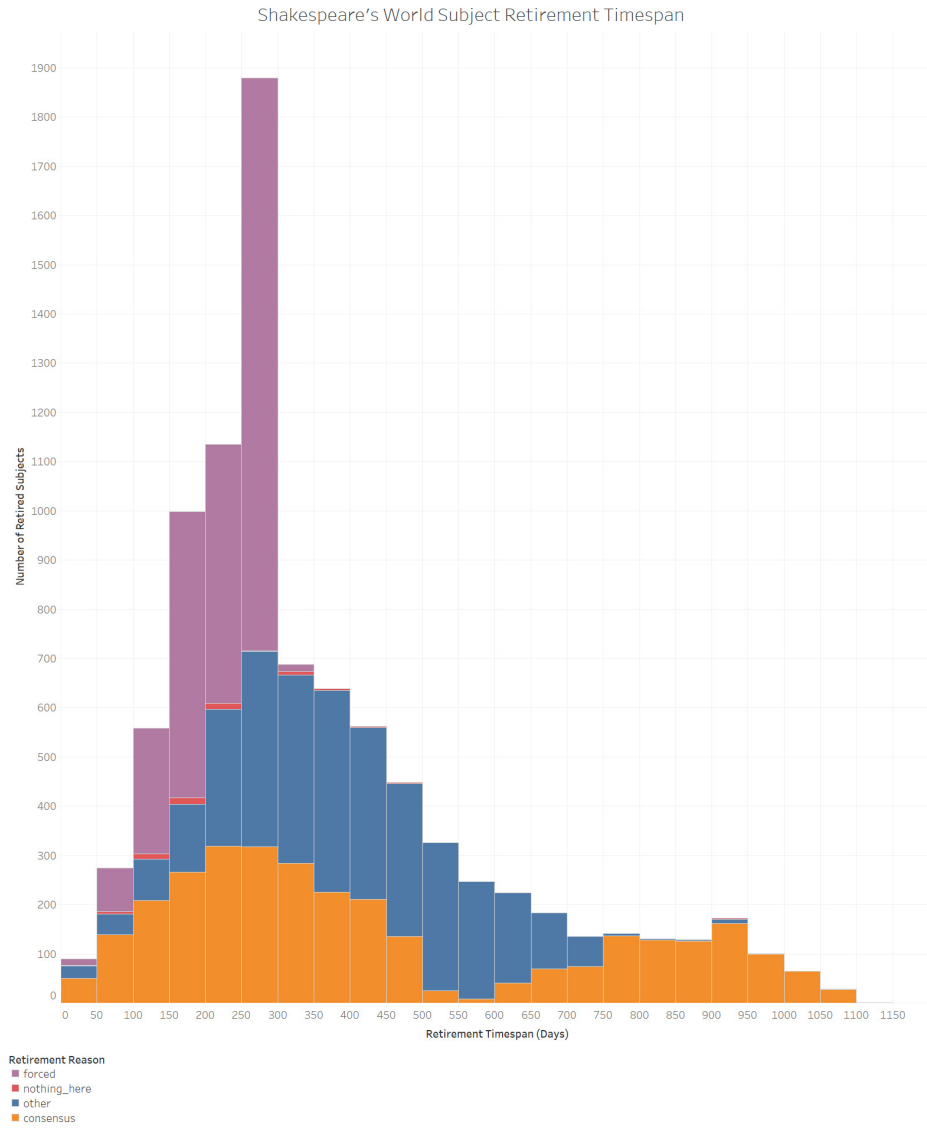


Figure 16 – Subject Retirement Timespan with Different Retirement Reasons

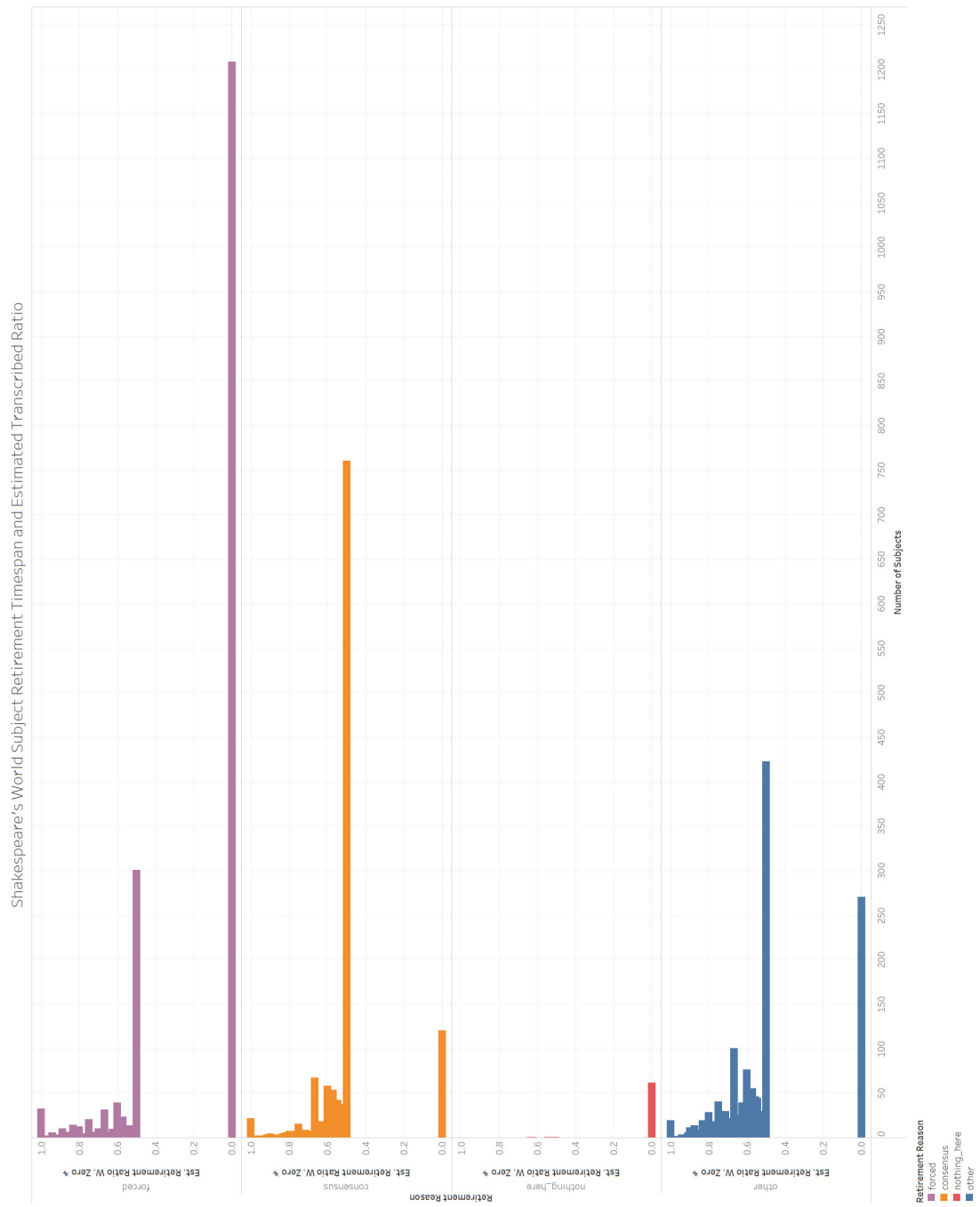


Figure 17 – Subject's Estimated Transcribed Ratio

4.6 Subject Retirement Pattern – Macro

To better examine the macro subject retirement pattern, we created a visualization for retired subjects from *Shakespeare's World* (Figure 18). This visualization examines the relationship between retirement timespan (days) and the estimated transcribed ratio. Each dot represents a distinct subject, and the size of the dots shows the total number of classifications submitted toward the subject. Their color was used to denote different retirement reasons. The median and quartiles of the estimated transcribed ratio were also included. Subjects that could not produce any line aggregations were excluded from this visualization.

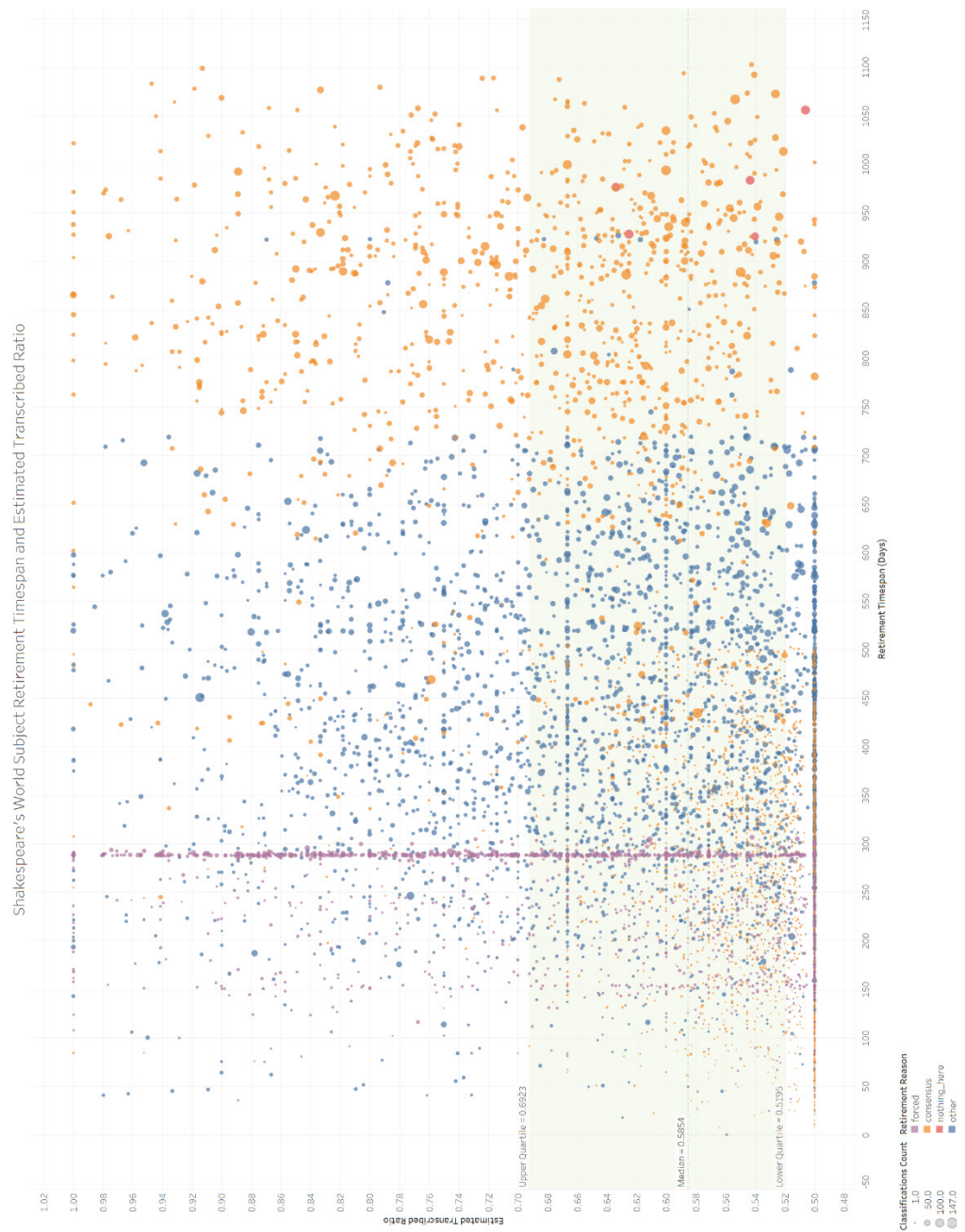


Figure 18 – Subject Retirement Timespan and Estimated Transcribed Ratio with Classification Counts and Retirement Reasons

5. Findings

5.1 Volunteer Contribution Patterns

The summary statistics and visualizations of volunteer contribution patterns over time provide some preliminary insights into *Shakespeare's World's* volunteer engagement and task preference. It shows the variations in contribution levels between anonymous volunteers and those who are registered and logged in as well as variations in participation across the three task categories.

The noticeable contribution pattern in the summary statistics is that logged-in volunteers consistently show higher minimum, maximum, medium, and average contributions across all task categories when compared to anonymous volunteers (besides low submission during the initial alpha and beta testing periods). While this phenomenon holds true when we examine the classifications submitted per month over time, there are some interesting outliers. During multiple months, anonymous volunteers were contributing at levels equal to, and even surpassing, logged-in users. From April 2017 to June 2017, anonymous users were contributing at a significantly higher level compared to logged-in users for three consecutive months (see Figure 9). Overall, however, registered and logged-in volunteers contributed more sustainably throughout the span of the project. This is especially true when we examine the contribution pattern after new content dropped in May 2018, logged-in volunteers' contribution levels

maintained at a higher level compared to anonymous volunteers until *Shakespeare's World* went on hiatus in late 2019.

Another pattern was the contribution differences in different task categories. The two binary classification tasks T0 and T3 received similar levels of engagement, while the more involved and challenging text transcription task T2 received significantly fewer submissions (see Figure 10). Logged-in volunteers contributed significantly more to the challenging transcription task as well. This pattern is more salient when looking at the median number of submitted T2 classifications per month, where logged-in volunteers were contributing almost twice as much (see Table 4). This trend suggests that registered volunteers are not only more committed to the project over time, but they are also more willing to take on the project's more demanding tasks.

Contribution patterns towards different genres also provided some interesting insight. In May 2018, "Newdigate Newsletters" were added to *Shakespeare's World* as new content. While this introduction of new materials drove up the overall number of classifications contributed that month, most of them were submitted for recipe books and letters, both genres were available since the start of the project (see Figure 11). Newsletters only see a significant growth in classifications after December 2018, likely caused by the gradual retirement of the earlier materials from the project.

Logged-in volunteers' contribution pattern revealed some other interesting insights about the community around *Shakespeare's World*. We noticed large variations between the most

popular months and the rest (see Figure 12). Those peaks usually correlated with major events related to Shakespeare and/or the project. The launch month of the project (December 2015), for example, drew the highest number of unique registered volunteers (900). The coverage of *Shakespeare's World* in the *New Yorker* in 2017 [46] also helped to drive up the engagement level for registered volunteers for three months. While volunteers' momentum dropped over time, the total number of distinct active registered volunteers per month remained above 50.

The overall macro volunteer contribution pattern provides some significant additional insights into the engagement of *Shakespeare's World* community. While being relatively sparse, the contributions from registered users, denoted in both blue and orange (for the top 100 contributing registered volunteers), can be seen throughout the project's timeline (see Figure 15), suggesting a consistent level of engagement from a dedicated subset of volunteers. Frequent contributors leave a trail of vertical streaks, indicative of their sustained involvement over time. This effort of the core contributing group was instrumental to the project's outcome.

Streaks of contributions were also spotted towards the very end of the axis, indicating some sustained interest and contributions from newcomers even as *Shakespeare's World* is heading toward hiatus. In fact, one of the top 100 contributing volunteers (`user_id` 1347003) only joined the project in 2019. Darker clusters not only represent periods of heightened activity from individual contributors but also highlight how the project gained momentum from significant events through the resemblance of horizontal streaks.

Despite not being able to draw firm conclusions about the performance of individual anonymous volunteers, their combined contributions towards the project outcome were evidently consistent throughout the project's duration. This pattern reveals their participation was also a significant part of the project in terms of classifications submitted, underlining the value of volunteers who prefer to remain anonymous, affirming their essential involvement in this years-long online collaborative endeavor.

5.2 Subject Retirement Patterns

Previous visualizations of retirement timespan help provide a more comprehensive picture of the time it takes for subjects to retire in *Shakespeare's World*. Besides that, the visualizations and summary statistics using estimated transcribed ratios could also help provide insights into the potential quality of aggregated transcription produced by the volunteers.

The detailed distribution of the subject retirement timespan provides additional context besides the forced retirement during the early stage of the project. Although overall retirement timespan distribution still peaks around 250 to 300 days excluding forced retirements, it is primarily caused by retirement reasons other than reaching the designated 95% consensus among at least three individual transcriptions. When assessing the retirement timespan of subjects that reached consensus, we noticed two bumps around 200-300 days and 750-950 days (see Figure 16). These variations suggest differences in complexity among the subjects, which is understandable given the diverse range of subjects transcribed for the project. Several factors

could contribute to this complexity, including whether a subject contains only images or is completely empty, the number of lines on each page, and the manuscripts' legibility.

Using estimated transcribed ratios, we could also approximate the quality of aggregated transcription for retired subjects. Despite *Shakespeare's World* workflow being designed to only retire subjects that have reached 95% similarities among at least three individual transcriptions, Zooniverse aggregations seem to prove otherwise. The median estimated completion ratio for all retired subjects is only 58.5%, meaning on average more than 40% of lines in each retired subject could not reach consensus when aggregated using Zooniverse's algorithm. This is already a more optimistic estimation, considering we excluded the subjects that were not able to result in any aggregation lines when calculating the median transcribe ratio. Another key finding from the summary statistics for the estimated transcribed ratio is that subjects retired by `consensus` don't have the best transcribed ratio. In fact, it has one of the worst transcribed ratios among the four retirement reasons generated by Zooniverse, only slightly better than `nothing_here`, which only were recorded for five subjects. Although subjects retired by `other` have significantly more subjects that were unable to produce any line aggregations, subjects retired by `consensus` have significantly more subjects with only a 50% transcribed ratio. Subjects retired by `consensus` only have a slight advantage in the number of 100% transcribed subjects when compared to `other` (22 versus 20). This suggests that significant errors occurred either during the live aggregation and subject retirement built

into *Shakespeare's World's* Zooniverse workflow, or issues occurred in the later aggregation of individual transcriptions after the project went on hiatus.

Interestingly, subjects that were forced to retire produced the majority of 100% transcribed subjects according to Zooniverse aggregation (48 out of 90). It also has a higher average estimated transcribed ratio – second highest in median agreement ratio and the highest in the mean agreement ratio. The initial decision to force retire subjects was made due to concerns about subjects not retiring over a prolonged period. This forced retirement seems partly justified on the basis that many of those subjects actually contained a significant number of high-quality transcriptions but did not retire properly according to the predetermined threshold.

The macro subject retirement pattern shows some weak correlation between the retirement timespan and the estimated transcribed ratio. The forced retired subjects have high variations in transcribed ratio and clustered between 280 and 300 days which matches the previous analysis. The number of classifications positively correlates with the retirement timespan but has a varied impact on the transcribed ratio. Classification counts for subjects forced to retire or retired by consensus tend to have a stronger positive correlation with the estimated transcribed ratio. However, when examining only subjects retired with an estimated transcribed ratio higher than 95%, no significant correlation could be spotted between the number of classifications and transcribed ratio for those two retirement reasons. This relationship points to the potential influence of other factors on high transcribed ratio, and more classifications might have diminishing effects, caused by the nuances of materials. For example, some subjects

5.3 Consensus and Estimated Transcribed Ratio

Based on our analysis, it is evident that the `retirement_reason` term consensus does not truly speak to the quality of aggregated transcriptions based on volunteer classifications. To better understand the best metrics for measuring the quality of aggregated transcriptions produced by *Shakespeare's World* and similar online crowdsourced transcription projects, additional studies are required, which may require a thorough inspection of all pages or inspection of representative samples from all subjects containing text.

The estimated transcribed ratio alone also does not constitute a robust measure of either the quality of the aggregated transcriptions or the quality of the work by the volunteers. Subjects receiving zero aggregated lines could be caused by blank or image-only subjects. In this situation, while the estimated transcribed ratio would be zero or null, it represents an accurate transcription by the volunteers that are being successfully aggregated.

It is also important to note that OCR at the time of the project was also performing inadequately with Early Modern English manuscripts. Folger's own experiments with GoogleOCR in 2018 showed high error rates for using computer vision to process Early Modern English manuscripts as well [48].

6. Discussion

6.1 Implications

6.1.1 Reuse Potential

The cleaned *Shakespeare's World* outcome dataset is carefully documented and restructured to improve its usability for both volunteers and the public. These improvements help support the potential future reuse of both the metadata and volunteer transcriptions recorded.

The submission metadata could not only provide insights into the contribution patterns of *Shakespeare's World* volunteers but could also be used along with other crowdsourced projects' metadata to better understand crowdsourced volunteers' contribution patterns. Scholars like Ruth Ahnert and Sebastian E. Ahnert, for example, have used metadata from historical letter networks to perform network analysis and revealed social interactions among Protestant correspondences during the reign of Mary I [49].

With the recent rapid development in machine learning algorithms and computational technology, Large Language Model (LLM) has emerged as a significant field of study. Researchers in the humanities domain have also started to utilize this method to examine language patterns, such as the scaling properties of natural languages [50]. However, models trained on contemporary language data are often less effective when handling historic text. Previous studies have tried addressing this challenge through various methods like fine-tuning

language models on historic languages [51], modernizing historic text [52] and even pre-training language models on historic text (MacBERTh) [53]. Although pre-trained models, like MacBERTh, have been shown to perform more reliably [54], the volume of text and diversity of genres within the training corpus is still limited, especially when compared to modern text.

Online crowdsourced transcription projects hosted on Zooniverse and similar crowdsourcing platforms could potentially be a valuable addition for researchers using language models to understand historic English. These crowdsourced projects, which often involve the transcription of historic documents, contain unique potentials for those interested in the linguistic nuances and semantic evolution present in historical primary. The manuscripts used across these projects include genres beyond official records and published works, and some projects, like *Voices Through Time: The Story of Care* [55] and *Shakespeare's World*, focused specifically on underrepresented authors, such as women writers. The transcriptions generated by the Zooniverse projects provide rich linguistic and social diversity which could provide invaluable materials for training robust language models to accurately capture the depth and complexity of historic language use and its evolution over time.

The processed dataset could also be used to support education in classrooms for students with different levels of data manipulation and analysis skills as it contains structured and semi-structured metadata and unstructured volunteer transcription data.

In short, our work in *Revisiting Shakespeare's World* marked the first step towards making this dataset accessible to the public and more sustainable in the long run.

6.1.2 Volunteers in *Shakespeare's World* and Similar Crowdsourced Projects

Through our analysis, we found that *Shakespeare's World* has helped to curate a dynamic community with highly engaged registered volunteers with many sustained contributors and continues to attract new persistent volunteers even towards the end of the project's life cycle. While we could not conclude the individual contribution patterns of anonymous volunteers, they have contributed to more than 46.5% of all classifications towards the project and more than 40% of the most challenging and involved classification tasks (T2). Our analysis also reveals the difficulty in retaining volunteers' interest. This could be seen in the steady decline in contribution levels over time. While significant events like Shakespeare's anniversary and the inclusion of new subject genres would help temporarily bump up active volunteers, a rapid fall off often followed soon.

Besides top contributors' tendency to have a persistent engagement profile (see Figure 20), similar to the pattern identified by Ponciano and Brasileiro [32], we have noticed a slightly different contribution pattern among the top 10 contributors in *Shakespeare's World*. Many of the top 10 contributing volunteers have contributed frequently throughout the project's lifespan (see Figure 21), though at a lower intensity compared to the hardworking profiles. This contribution pattern does not seem to match any of the five engagement profiles, while these 10 volunteers contributed to more than 16.5% of the total classifications submitted. Since Ponciano and Brasileiro's study was primarily conducted on scientific-oriented projects, we

suspected that there might be a difference in volunteer engagement profile across different types of crowdsourced projects.

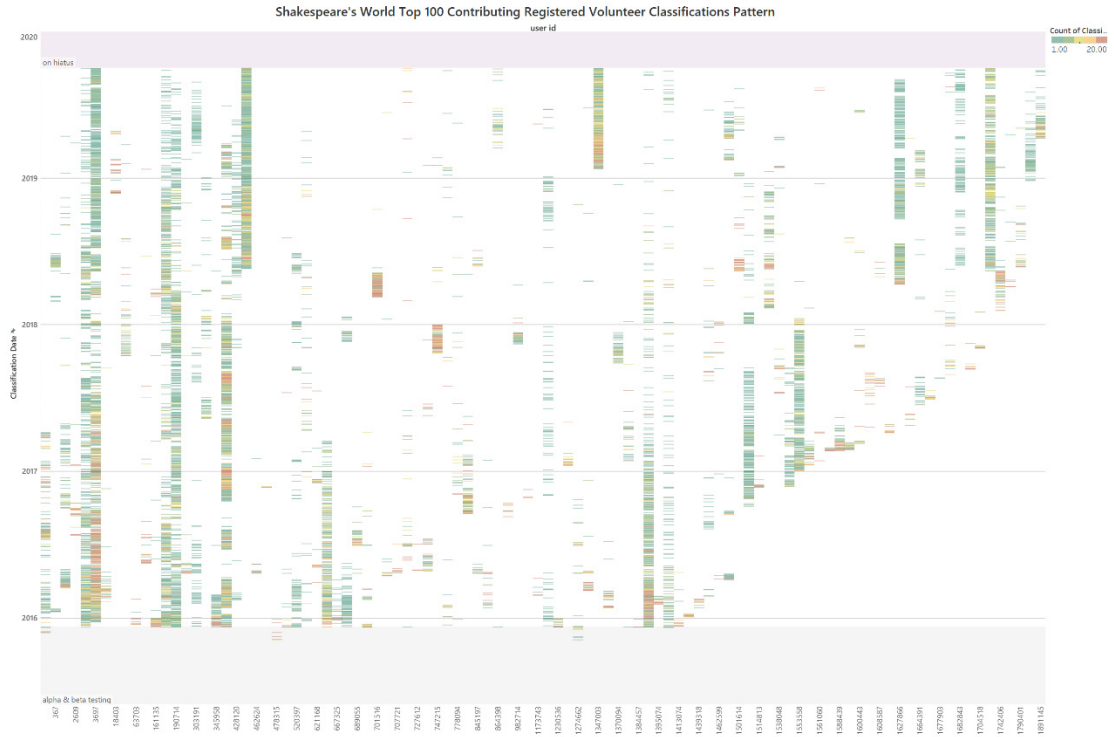


Figure 20 – Shakespeare's World Top 100 Contributing Registered Volunteer Classifications Pattern, Sorted by Their user_ids

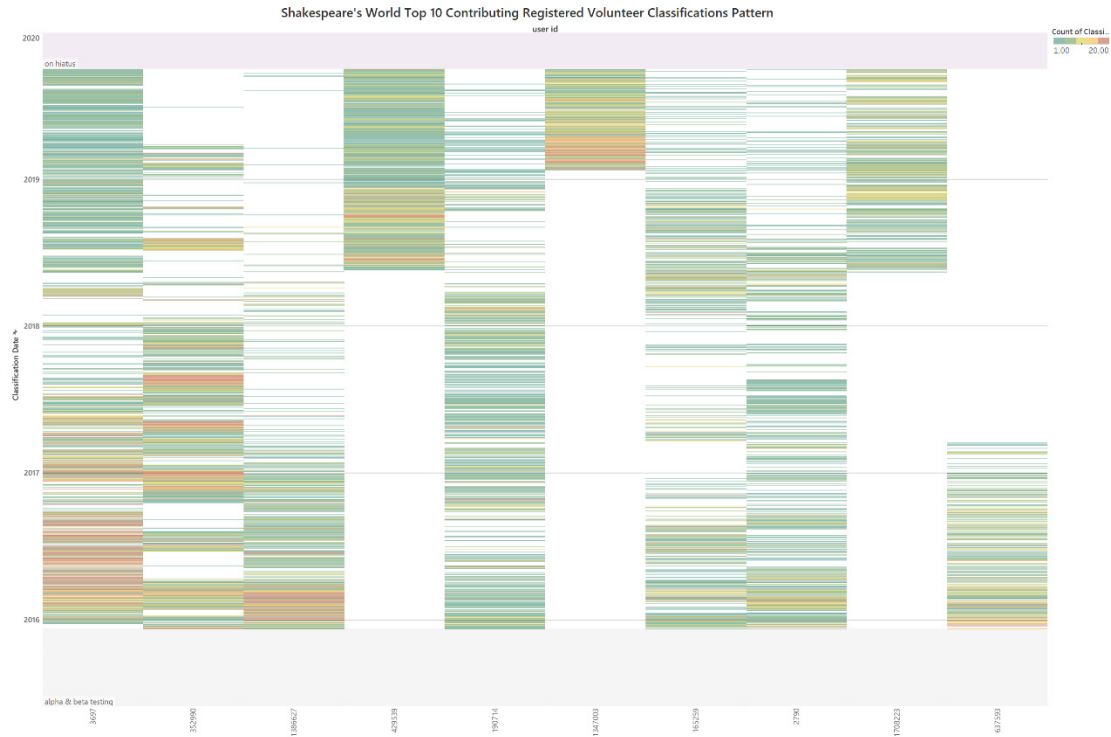


Figure 21 – Shakespeare's World Top 10 Contributing Registered Volunteer Classifications Pattern, Sorted by Total Classification Counts

These findings about consistent contributions from a small subset of registered volunteers and difficulties in recruiting new frequent contributors during *Shakespeare's World* echo the contribution patterns found in other crowd-powered initiatives. Previous research on Wikipedia editors discovered that editors who contributed most of the work and produced the most value contributed differently from casual participants on their the very first days, and that it is rare for such contribution patterns to evolve over time [56]. A similar pattern could be identified for *Shakespeare's World* volunteers as well. Few volunteers, like `user_id` 428120, started

with very little contributions and started to contribute more consistently during the later phases of the project. Panciera et. al discussed the importance for Wikipedia to identify potential power users based on their contribution patterns early on in order to make sure they remain in the community and rise a suitable position in the volunteer hierarchy. Crowdsourced transcriptions project managers could take the same approach. These findings and patterns in the data may diverge from some of the fundamental design choices behind *Shakespeare's World* and other similar Zooniverse projects, which aimed to engage “dabblers” and harness small one-day or even one-off contributions [57].

6.1.3 Online Crowdsourced Projects Management

Another crucial takeaway we gained is the vital role of comprehensive documentation during long-term online crowdsourced projects. This includes thoroughly documenting the meaning of different attributes used in the project and recording every key decision made during its lifespan. When cleaning the data for *Shakespeare's World*, we invested a significant amount of time and effort speculating, examining, discussing, and verifying the significance of different variables. Despite our efforts, the precise meanings behind some attributes (like the `id` used in the original retirement metadata and `user groups id`) remain elusive.

If possible, online crowdsourced project teams should also try to evaluate the crowdsourced outcomes during the project, or at the end of each project phase, and identify potential errors early in the data collection or aggregation pipelines. This proactive approach ensures maximum availability of information on potential factors causing errors while enabling timely

interventions and adjustments. These measures could not only enhance the quality and integrity of data collected but also help avoid potential confusion for participating volunteers.

6.3 Limitations

One key limitation of the project is the potential inaccuracies in attribute description. Since the project went on hiatus more than four years ago and many original documentation were missing, some inaccuracies might occur when reannotating the attributes based on the known data and drawing from memory. Due to the scope of our study, we also did not conduct interviews or workshops with potential users from other domains. This means we might have missed some attributes other scholars and researchers deemed essential when flattening the dataset. Thus, additional cleaning and processing might still be needed to better fit different research goals. We also understand that while the dataset has been cleaned and flattened, there are still technical barriers towards effective public reuse. For a wider audience to understand the outcome of *Shakespeare's World*, other forms of project outputs like visualizations or cleaned transcriptions could be used as better supplementary materials to facilitate public understanding of the project. Furthermore, our preliminary analysis did not assess the quality of aggregated transcriptions at the subject level. We also did not have enough time to fully explore the reasons behind all patterns spotted in the retirement duration, like the “little hump” between 750 seconds and 800 seconds in session durations (see Figure 8), the bimodal distributions of retirement timespan for subjects retired by consensus (see Figure 16) and the low estimated transcribed ratio for the retired subjects.

6.4 Future Work

Looking ahead, we envisioned several potential directions to further build on this project and address its current limitations.

First, we hope to conduct comprehensive interviews and co-design workshops with professionals from other domains to better understand what information is commonly used across different related academic domains. We would also like to conduct interviews with other Zooniverse transcription teams to better understand their data collection practice, project workflows and retirement threshold. Using insights from these qualitative studies, we aim to construct a data primer for related Zooniverse transcription projects. This primer could not only help make other projects' data outcomes more accessible but also help enable further study of volunteer engagement patterns across different projects and genres.

Future studies should also systematically explore the discrepancies between the designed retirement agreement ratio and the occurrences of low consensus lines and low transcription ratios seen in the aggregation results post-hiatus. Professionally cleaned transcription in Folger's EMMO collection could be used as the gold standard (for example, subject id 1276972's aggregated transcription could be evaluated against L.a.457 in EMMO) to help assess the accuracy of aggregated volunteer transcriptions. Doing so might help us understand the underlying causes of low transcribed ratio and could help improve the quality of aggregated transcriptions and determine proper retirement thresholds for future stages of *Shakespeare's World* and other crowdsourced transcription projects hosted on Zooniverse.

A more detailed quantitative analysis of volunteer contribution patterns could also be performed using the existing dataset. This could help reveal whether volunteer contribution patterns in *Shakespeare's World* significantly differ from the engagement profiles described in Ponciano and Brasileiro's paper [30]. Other researchers within the Zooniverse community have explored using early-stage performance metrics to predict later contributions [58]. A similar methodology could also be applied to *Shakespeare's World*, considering the new genre introduced in May 2018 naturally dividing the project into two stages. Doing such could help uncover how volunteer performance evolves throughout the project. This could lead to intriguing research questions and workflow design implications, such as whether performance on simpler tasks, like binary classification or identifying images and lines, can be a reliable indicator of a volunteer's overall contribution quality.

6.4 Conclusion

In this study, we conducted a case study for cleaning the *Shakespeare's World* output dataset and a preliminary analysis using the dataset to answer the following questions: 1) What were *Shakespeare's World's* volunteer contribution patterns like? How did its contribution pattern change over time and across different subjects? and 2) Could we spot any previously unknown anomalies in volunteer contribution patterns? What can these anomalies tell us about the structure of the *Shakespeare's World* project and perhaps about transcription approaches on the Zooniverse platform?

We thoroughly documented our data cleaning process and created a robust dataset that captures all essential information while protecting volunteer privacy and being easily reusable through low-programming tools like OpenRefine and Tableau. Through the preliminary analysis, we found that registered and logged-in volunteers were significantly more involved across different task categories compared to anonymous volunteers. However, anonymous volunteers also contributed consistently and even occasionally surpassed the efforts of logged-in volunteers. We found that the overall contribution level steadily declined over time, and while significant events could help significantly drive up volunteer contribution, the effect usually does not persist. When examining the subjects' retirement patterns, we spot a forced retirement during the early phase of the project. Using the Zooniverse aggregation, we estimated the transcribed ratio of every retired subject. We found that most estimated transcribed ratios for retired subjects were lower than predesignated retirement thresholds and, interestingly, forced retired subjects showed a similar or even higher transcribed ratio compared to subjects retired by "consensus".

Through this project, we helped update the project's outcome dataset and rectify the missing documentation, preparing *Shakespeare's World* outcome dataset for future use and reuse across different academic domains. We also shed some light on the project's volunteer contribution pattern and made the first step in quantitatively studying the project and its community. By sharing this project, we hope to encourage more scholars to engage with the outcome of *Shakespeare's World*, and other similar online crowdsourced transcription projects, and keep the collective wisdom and efforts of the crowd alive and relevant for years to come.

Appendices

The screenshot shows a Google Form titled "Zooniverse Username Update". At the top, there are tabs for "Questions", "Responses" (with a count of 1), and "Settings". The form is set against a light blue background. The title "Zooniverse Username Update" is in a large, bold, black font. Below the title is a rich text editor with icons for bold, italic, underline, link, and unlink. The text of the form is as follows:

Dear former Shakespeare's World Project Volunteers,

You are receiving this form because you participated in the Shakespeare's World project, which ran between 2015 and 2019. Shakespeare's World was an online crowdsourced transcription project hosted on the Zooniverse platform, inviting volunteers to transcribe handwritten documents from William Shakespeare's time.

We are currently preparing data outcomes of the Shakespeare's World project for publication in the Journal of Open Humanities Data (JOHD) and are committed to properly recognizing all volunteers while ensuring their privacy. We have noticed that you used an email address as your Zooniverse username during the project. To protect your privacy in the soon to be public dataset, we will change your username to either your existing Zooniverse ID in the Zooniverse data output or a username of your choice. This will ensure that your individual contributions are recognized without running the risk that your email is published online.

Please fill out this form to indicate if you wish to use your Zooniverse ID* or submit an alternative username for credit. If you do not respond to this form by February 20, 2024 (EST), we will, by default, anonymize your contribution by replacing your username with your Zooniverse ID*, to ensure that your privacy is maintained.

*The Zooniverse ID is an arbitrary number automatically generated by the Zooniverse platform when you first registered.

The form contains three main sections:

- Question 1:** "What email address did you use as your username for the Zooniverse project, Shakespeare's World (sometime between 2015 and 2019)?" This is a short answer text field.
- Question 2:** "Please specify if you wish to be credited in the public dataset by your Zooniverse generated ID number, or enter a new username of your choice in the 'Other...' field provided below". This is a radio button question with two options: "Zooniverse ID" and "Other...".
- Question 3:** "Do you have any questions about this project? If yes, please describe them below." This is a long answer text field.
- Question 4:** "If you have a question and want to hear back from Victoria Van Hynning or other members of the research team preparing the data, provide your email below. This is optional, and only necessary if you wish to have a reply." This is a short answer text field.

On the right side of the form, there is a vertical toolbar with icons for adding, deleting, and duplicating questions, as well as a "Print" icon.

Figure 22 – Screenshot of the Google Form Distributed to Collect Alternative Usernames

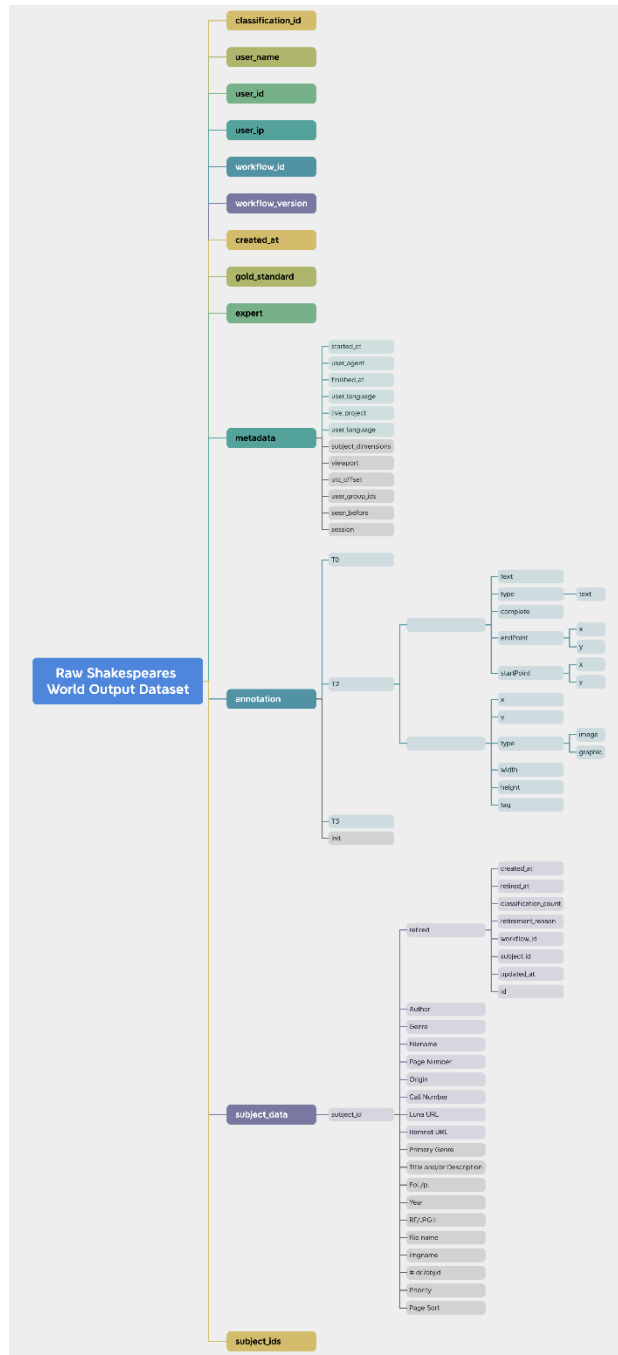


Figure 23 – Data Structure of the Raw Output Dataset

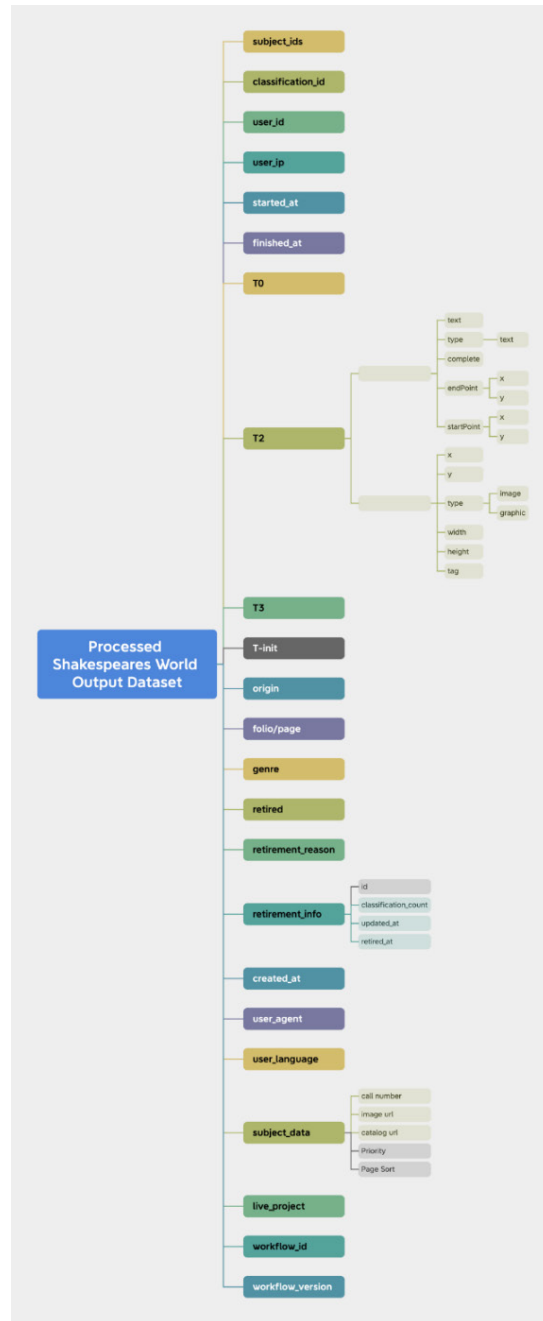


Figure 24 - Data Structure of the Cleaned Dataset

Bibliography

- [1] J. Howe, “The Rise of Crowdsourcing,” *Wired*. Accessed: Mar. 28, 2024. [Online]. Available: <https://www.wired.com/2006/06/crowds/>
- [2] “Crisis Mapping Haiti: Some Final Reflections,” Ushahidi. Accessed: Mar. 28, 2024. [Online]. Available: <https://www.ushahidi.com//about/blog/crisis-mapping-haiti-some-final-reflections/>
- [3] K. Wazny, “Crowdsourcing’s ten years in: A review,” *Journal of Global Health*, vol. 7, no. 2, p. 020601, Dec. 2017, doi: [10.7189/jogh.07.020601](https://doi.org/10.7189/jogh.07.020601).
- [4] E. Estellés-Arolas and F. González-Ladrón-de-Guevara, “Towards an integrated crowdsourcing definition,” *Journal of Information Science*, vol. 38, no. 2, pp. 189–200, Apr. 2012, doi: [10.1177/0165551512437638](https://doi.org/10.1177/0165551512437638).
- [5] C. J. Lintott *et al.*, “Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey★,” 2008.
- [6] “About — Zooniverse” Accessed: Apr. 04, 2024. [Online]. Available: <https://www.zooniverse.org/about>
- [7] V. Van Hyning, “Harnessing crowdsourcing for scholarly and GLAM purposes,” *Literature Compass*, vol. 16, no. 3–4, p. e12507, Apr. 2019, doi: [10.1111/lic3.12507](https://doi.org/10.1111/lic3.12507).
- [8] S. Blickhan, C. Krawczyk, D. Hanson, A. Boyer, A. Simenstad, and V. Van Hyning, “Individual vs. Collaborative Methods of Crowdsourced Transcription,” *Journal of*

- Data Mining & Digital Humanities*, vol. Special Issue on Collecting,..., p. 5759, Dec. 2019, doi: [10.46298/jdmdh.5759](https://doi.org/10.46298/jdmdh.5759).
- [9] “Publications | Zooniverse - People-powered research” Accessed: Mar. 28, 2024. [Online]. Available: <https://www.zooniverse.org/about/publications>
 - [10] V. Van Hyning *et al.*, “Shakespeare’s World,” Shakespeare’s World. Accessed: Jul. 11, 2022. [Online]. Available: <https://www.zooniverse.org/projects/zooniverse/shakespeares-world>
 - [11] “Home - Early Modern Manuscripts Online - EMMO - Folger Shakespeare Library,” Early Modern Manuscripts Online - EMMO - Folger Shakespeare Library. Accessed: Apr. 09, 2024. [Online]. Available: <https://emmo.folger.edu/>
 - [12] “EMMO announces the launch of Shakespeare’s World | Folger Shakespeare Library.” Accessed: Mar. 13, 2024. [Online]. Available: <https://www.folger.edu/blogs/collation/emmo-announces-shakespeares-world/>
 - [13] spowellzoon, “Shakespeare’s World launches Newdigate newsletters!,” Shakespeare’s World. Accessed: Mar. 14, 2024. [Online]. Available: <https://blog.shakespearesworld.org/2018/05/15/shakespeares-world-launches-newdigate-newsletters/>
 - [14] Hines, G. (2017). *Zooniverse Aggregation Engine Documentation, Release 0.9*. Zooniverse. http://developer.zooniverse.org/_/downloads/aggregation/en/latest/pdf/

- [15] V. A. Van Hyning and M. A. Jones, “Data’s Destinations: Three Case Studies in Crowdsourced Transcription Data Management and Dissemination,” Dec. 2021, doi: [10.5281/ZENODO.5750691](https://doi.org/10.5281/ZENODO.5750691).
- [16] Van Hyning, V., & Wolfe, H. (2023). More Content, Less Context: Rethinking Access. In A. Prescott & A. Wiggins (Eds.), *Archives* (1st ed., pp. 174–191). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780198829324.013.0013>
- [17] The Zooniverse, “ShaxWorldFullSiteArchive,” YouTube, [Online]. Available: <https://www.youtube.com/watch?v=OKiv-9c5-28>. [Accessed: Mar. 28, 2024]. Screenshot at 00:54.
- [18] philipdurkin, “Shakespeare’s World and updating the OED: a splendid antedating of ‘white lie,’” Shakespeare’s World. Accessed: Mar. 28, 2024. [Online]. Available: <https://blog.shakespearesworld.org/2017/04/05/shakespeares-world-and-updating-the-oed-a-splendid-antedating-of-white-lie/>
- [19] “white lie, n. meanings, etymology and more | Oxford English Dictionary.” Accessed: Apr. 10, 2024. [Online]. Available: https://www.oed.com/dictionary/white-lie_n
- [20] M. Ridge, “Crowdsourcing in Cultural Heritage,” in *Routledge International Handbook of Research Methods in Digital Humanities*, 1st ed., K. Schuster and S. Dunn, Eds., Routledge, 2020, pp. 363–383. doi: [10.4324/9780429777028-26](https://doi.org/10.4324/9780429777028-26).

- [21] F. Rohden, C. Kullenberg, N. Hagen, and D. Kasperowski, “Tagging, Pinging and Linking – User Roles in Virtual Citizen Science Forums,” *Citizen Science: Theory and Practice*, vol. 4, no. 1, p. 19, Jun. 2019, doi: [10.5334/cstp.181](https://doi.org/10.5334/cstp.181).
- [22] P. Organisciak, M. Efron, K. Fenlon, and M. Senseney, “Evaluating rater quality and rating difficulty in online annotation activities,” *Proc of Assoc for Info*, vol. 49, no. 1, pp. 1–10, Jan. 2012, doi: [10.1002/meet.14504901166](https://doi.org/10.1002/meet.14504901166).
- [23] H. Zhou, D. Baskov, and M. Lease, “Crowdsourcing Transcription Beyond Mechanical Turk,” *HCOMP*, vol. 1, pp. 9–16, Nov. 2013, doi: [10.1609/hcomp.v1i1.13093](https://doi.org/10.1609/hcomp.v1i1.13093).
- [24] N. Putri, A. Carroll, J. Chen, J. Hershberger, and M. Khan, “The-Shakespeares-UMD/Final Datasets,” GitHub. Accessed: Jun. 29, 2023. [Online]. Available: <https://github.com/nisaputri/The-Shakespeares-UMD>
- [25] “Newdigate family collection of newsletters | FromThePage.” Accessed: Apr. 29, 2024. [Online]. Available: <https://fromthepage.com/folger/newdigate-family-collection-of-newsletters>
- [26] L. Seaward *et al.*, “Engaging Crowds: citizen research and cultural heritage data at scale,” [object Object], Oct. 2022. doi: [10.5281/ZENODO.7152030](https://doi.org/10.5281/ZENODO.7152030).
- [27] “HMS NHS: The Nautical Health Service | Zooniverse - People-powered research.” Accessed: Mar. 31, 2024. [Online]. Available: <https://www.zooniverse.org/projects/msalmon/hms-nhs-the-nautical-health-service>

- [28] “Scarlets and Blues | Zooniverse - People-powered research.” Accessed: Mar. 31, 2024. [Online]. Available: <https://www.zooniverse.org/projects/bogden/scarlets-and-blues>
- [29] “The RBGE Herbarium: Exploring Gesneriaceae, the African violet family | Zooniverse - People-powered research.” Accessed: Mar. 31, 2024. [Online]. Available: <https://www.zooniverse.org/projects/emhaston/the-rbge-herbarium-exploring-gesneriaceae-the-african-violet-family>
- [30] “TANC Engaging Crowds.” Accessed: Mar. 31, 2024. [Online]. Available: <https://tanc-ahrc.github.io/EngagingCrowds/Data.html>
- [31] B. Ogden, “User engagement analysis report.” [object Object], Oct. 06, 2022. doi: [10.5281/ZENODO.7151973](https://doi.org/10.5281/ZENODO.7151973).
- [32] L. Ponciano and F. Brasileiro, “Finding Volunteers’ Engagement Profiles in Human Computation for Citizen Science Projects,” *HCj*, vol. 1, no. 2, Dec. 2014, doi: [10.15346/hc.v1i2.12](https://doi.org/10.15346/hc.v1i2.12).
- [33] “Part III: In For the Long Haul - Zooniverse Help.” Accessed: Apr. 26, 2024. [Online]. Available: <https://help.zooniverse.org/best-practices/3-long-haul/>
- [34] “OpenRefine.” Accessed: Mar. 28, 2024. [Online]. Available: <https://openrefine.org/>
- [35] “Extensions | OpenRefine.” Accessed: Mar. 28, 2024. [Online]. Available: <https://openrefine.org/extensions>
- [36] A. Ahmi, “OpenRefine: An approachable tool for cleaning and harmonizing bibliographical data,” presented at the 27TH INTERNATIONAL MEETING OF

THERMOPHYSICS 2022, Dalešice, Czech Republic, 2023, p. 030006. doi:

[10.1063/5.0164724](https://doi.org/10.1063/5.0164724).

- [37] K. M. Hill, “In Search of Useful Collection Metadata: Using OpenRefine to Create Accurate, Complete, and Clean Title-level Collection Information,” *Serials Review*, vol. 42, no. 3, pp. 222–228, Jul. 2016, doi: [10.1080/00987913.2016.1214529](https://doi.org/10.1080/00987913.2016.1214529).
- [38] S. V. Hooland, R. Verborgh, and M. D. Wilde, “Cleaning Data with OpenRefine,” *Programming Historian*, no. 2, Aug. 2013, doi: [10.46430/phen0023](https://doi.org/10.46430/phen0023).
- [39] E. P. Williamson, “Fetching and Parsing Data from the Web with OpenRefine,” *Programming Historian*, Aug. 2017, Accessed: Mar. 28, 2024. [Online]. Available: <https://programminghistorian.org/en/lessons/fetch-and-parse-data-with-openrefine>
- [40] A. D. Chapman, *Principles and methods of data cleaning*. GBIF, 2005. [Online]. Available: <https://books.google.com/books?id=44gDJTEJoVIC>
- [41] A. D. Chapman and G. B. I. Facility, *Principles of Data Quality*. in Report for the Global Biodiversity Information Facility Copenhagen. GBIF, 2005. [Online]. Available: <https://books.google.com/books?id=LL5YQ5GdhzoC>
- [42] K. Rawson and T. Muñoz, “Against Cleaning,” Jul. 2016, Accessed: Feb. 16, 2024. [Online]. Available: <http://www.curatingmenus.org/articles/against-cleaning/>
- [43] L. Aroyo and C. Welty, “Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation,” *AI Magazine*, vol. 36, no. 1, pp. 15–24, Mar. 2015, doi: [10.1609/aimag.v36i1.2564](https://doi.org/10.1609/aimag.v36i1.2564).

- [44] “Our new catalog is here! | Folger Shakespeare Library.” Accessed: Mar. 27, 2024. [Online]. Available: <https://www.folger.edu/blogs/collation/our-new-catalog-is-here/>
- [45] “Our new digital collections site is here! | Folger Shakespeare Library.” Accessed: Mar. 27, 2024. [Online]. Available: <https://www.folger.edu/blogs/collation/our-new-digital-collections-site-is-here/>
- [46] R. Kwok, “Crowdsourcing for Shakespeare,” *The New Yorker*, Jan. 16, 2017. Accessed: Apr. 08, 2024. [Online]. Available: <https://www.newyorker.com/tech/annals-of-technology/crowdsourcing-for-shakespeare>
- [47] “20 verso | Digital Collections.” Accessed: Apr. 10, 2024. [Online]. Available: <https://digitalcollections.folger.edu/img44982>
- [48] M. J. Brown, M. Dai, C. Yang, and R. Ingle, “Experiments with early modern manuscripts and computer-aided transcription | Folger Shakespeare Library.” Accessed: Mar. 30, 2024. [Online]. Available: <https://www.folger.edu/blogs/collation/computer-aided-transcription/>
- [49] R. Ahnert and S. E. Ahnert, “Protestant Letter Networks in the Reign of Mary I: A Quantitative Approach,” *elh*, vol. 82, no. 1, pp. 1–1, Mar. 2015, doi: [10.1353/elh.2015.0000](https://doi.org/10.1353/elh.2015.0000).
- [50] S. Takahashi and K. Tanaka-Ishii, “Evaluating Computational Language Models with Scaling Properties of Natural Language,” *Computational Linguistics*, vol. 45, no. 3, pp. 481–513, Sep. 2019, doi: [10.1162/coli_a_00355](https://doi.org/10.1162/coli_a_00355).

- [51] X. Han and J. Eisenstein, “Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 4237–4247. doi: [10.18653/v1/D19-1433](https://doi.org/10.18653/v1/D19-1433).
- [52] M. Bollmann, “A Large-Scale Comparison of Historical Text Normalization Systems,” in *Proceedings of the 2019 Conference of the North*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 3885–3898. doi: [10.18653/v1/N19-1389](https://doi.org/10.18653/v1/N19-1389).
- [53] E. Manjavacas Arevalo and L. Fonteyn, “MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950),” in *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, M. Hämmäläinen, K. Alnajjar, N. Partanen, and J. Rueter, Eds., NIT Silchar, India: NLP Association of India (NLP AI), Dec. 2021, pp. 23–36. [Online]. Available: <https://aclanthology.org/2021.nlp4dh-1.4>
- [54] E. Manjavacas and L. Fonteyn, “Adapting vs. Pre-training Language Models for Historical Languages,” *Journal of Data Mining & Digital Humanities*, vol. NLP4DH, no. Digital humanities in..., p. 9152, Jun. 2022, doi: [10.46298/jdmdh.9152](https://doi.org/10.46298/jdmdh.9152).

- [55] “Voices Through Time: The Story of Care — Zooniverse” Accessed: Mar. 28, 2024. [Online]. Available: <https://www.zooniverse.org/projects/jojo38/voices-through-time-the-story-of-care/>
- [56] K. Panciera, A. Halfaker, and L. Terveen, “Wikipedians are born, not made: a study of power editors on Wikipedia,” in *Proceedings of the 2009 ACM International Conference on Supporting Group Work*, Sanibel Island Florida USA: ACM, May 2009, pp. 51–60. doi: [10.1145/1531674.1531682](https://doi.org/10.1145/1531674.1531682).
- [57] A. Eveleigh, C. Jennett, A. Blandford, P. Brohan, and A. L. Cox, “Designing for dabblers and deterring drop-outs in citizen science,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Toronto Ontario Canada: ACM, Apr. 2014, pp. 2985–2994. doi: [10.1145/2556288.2557262](https://doi.org/10.1145/2556288.2557262).
- [58] P. J. Marshall *et al.*, “Space Warps – I. Crowdsourcing the discovery of gravitational lenses,” *Monthly Notices of the Royal Astronomical Society*, vol. 455, no. 2, pp. 1171–1190, Jan. 2016, doi: [10.1093/mnras/stv2009](https://doi.org/10.1093/mnras/stv2009).

