A Computational Pipeline for Crowdsourced Transcriptions of Ancient Greek Papyrus Fragments

Alex C. Williams*†, John F. Wallin*, Haoyu Yu§, Marco Perale¶†, Hyrum D. Carroll*,
Anne-Francoise Lamblin§, Lucy Fortson§, Dirk Obbink†, Chris J. Lintott†, and James H. Brusuelas†
*_Middle Tennessee State University_, Murfreesboro, Tennessee, USA
{Alex.Williams,John.Wallin,Hyrum.Carroll}@mtsu.edu
†_University of Oxford_, Oxford, Oxfordshire, UK
cjl@astro.ox.ac.uk,{Dirk.Obbink,James.Bruseulas}@classics.ox.ac.uk
¶_University of Liverpool_, Liverpool, Merseyshire, UK
Marco.Perale@liverpool.ac.uk
§_University of Minnesota_, Minneapolis, Minnesota, USA
{yuxxx084,lambl001}@umn.edu,fortson@physics.umn.edu

***Abstract*—In the late nineteenth century, two excavators from the University of Oxford uncovered a vast trove of naturally deteriorated papyri, numbering over 500,000 fragments, from the city of Oxyrhynchus. With varying levels and forms of deterioration, the identification of a papyrus fragment can become a repetitive, long, and exhausting process for a professional papyrologist. The University of Oxford's Ancient Lives project aims to accelerate the identification process through citizen science (or crowdsourcing). In the Ancient Lives interface, volunteer users identify letters by clicking on a location in the image to designate the presence of a letter. To date, over 7 million letter identifications from users across the world have been recorded in the Ancient Lives database. In this paper, we present a computational pipeline for converting crowdsourced letter identifications made through the Ancient Lives interface into digital consensus transcriptions of papyrus fragments. We conclude by explaining the usefulness of the pipeline output in the context of additional computational projects that aim to further accelerate the identification process.**

Keywords*—*crowdsourcing; human computation; big data; papyrus transcription

I. Introduction

Over a century ago, two excavators, B.P. Grenfell and A.S. Hunt, of the University of Oxford, uncovered a vast trove of papryi, numbering over 500,000 fragments, from the city of Oxyrhynchus [1]. After transporting the collection back to the university, the field of papyrology emerged. Grenfell and Hunt began transcribing and editing the papyrus fragments, and to this day only a fraction of this vast trove have been published. Transcribing the collection has not been a simple task as each fragment suffers a unique level of deterioration with varying sections of missing papyrus or illegible handwritten text. Due to the meticulous process of transcribing fragments with limited information, the rate of manual transcription for fragments is extremely slow. In order to quicken the process of transcription, the University of Oxford enlisted volunteers by establishing Ancient Lives[1], a web-based interface for identifying letters on digital images of papyrus. Users can log onto the Ancient Lives site and help transcribe ancient papyrus fragments by clicking on a location in the image and designating the presence of a specific letter. Each letter identification and its associated characteristics (*i.e.*, x,y coordinates) are stored in a database of user identifications. To date, over 7 million letter identifications have been been recorded internationally via the Ancient Lives interface.

In other projects that confront the task of transcribing historical documents, such as Transcribe Bentham [2], user transcriptions are given in plain-text with supplemental XML tags. However, for Ancient Lives, user transcriptions cannot be given in plain-text due to the absence of certain Ancient Greek characters and accents on the modern keyboard. In substitution of the physical keyboard, users use an on-screen keyboard that has the characters and accents necessary to transcribe any ancient Greek papyrus fragment. By capturing letter identifications through click data and the on-screen keyboard instead of plain-text, the interpretation of the letter identification data has become a nontrivial task. In order to more easily interpret the large amount of letter identification data, we present a new computational pipeline for automating the process of converting the crowdsourced letter identifications into digital consensus transcriptions of papyrus fragments. The Methods section details the design of each pipeline component in depth. The Evaluation section presents an assessment of each pipeline component performed by comparing the consensus letter identifications and consensus line sequences for a set of fragments in Ancient Lives to the fragment transcriptions and sequences for the same fragments as they appear in published Oxyrhynchus (P. Oxy.) volumes. The Results section provides an interpretation of the results found in each evaluation. The Conclusion section reiterates on the value of the pipeline and ends with discussion on future work.

Using the digital consensus transcriptions from the pipeline, both professional papyrologists and papyrology students will be able to more quickly and easily begin the papyrological process. Despite being designed specifically for the domain of papyrus, additional classification projects that are tasked with forming consensus letter identifications or consensus line sequences from data-click coordinates can make use of the pipeline architecture.

[1] https://ancientlives.org

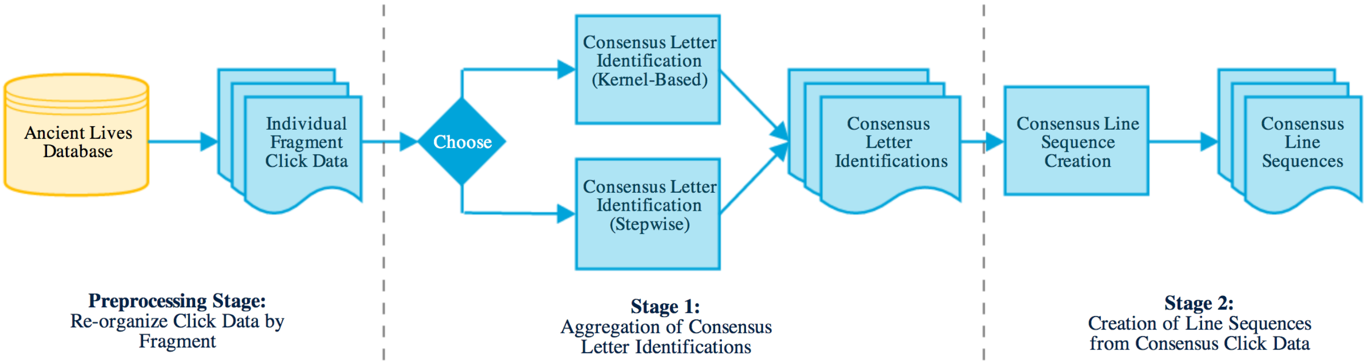


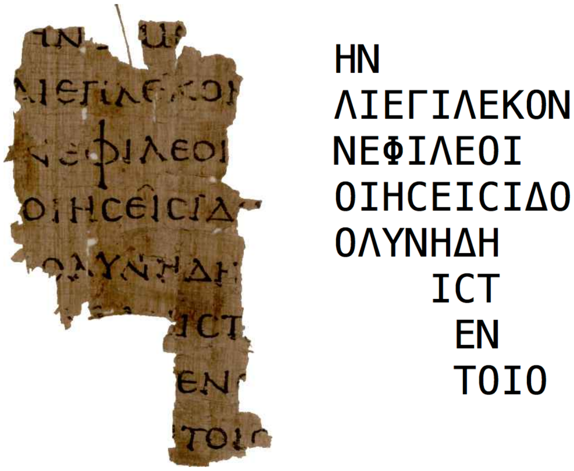
Fig. 1: The architecture of the computational pipeline used to create consensus transcriptions from user-clicks made through the Ancient Lives interface.

II. Methods

The computational pipeline can be separated into two stages (see Figure 1):

Stage 1: Aggregating User Clicks into Consensus Clicks.
Stage 2: Creating Line Sequences from Consensus Clicks.

The pipeline begins with a collection of plain-text files where each file contains click-data information for a specific fragment. The pipeline processing requires no human intervention after the input has been given. Once processing has finished, the pipeline will yield two files for each papyrus fragment. The first file contains the relative fragment's consensus letter identifications with consensus x,y coordinates. The second file, which is the final output of the pipeline, contains the relative fragment's consensus line sequence that closely resembles the original papyrus fragment (see Figure 2). The consensus letter identification components are written in both Matlab and Python and the line sequence creation component is written only in Python. All supplemental visualization in Python is performed with version 2.4.9 of the OpenCV image processing and computer vision package[2].



```
ΗΝ
ΛΙΕΓΙΛΕΚΟΝ
ΝΕΦΙΛΕΟΙ
ΟΙΗCΕΙCΙΔΟ
ΟΛΥΝΗΔΗ
    ΙCΤ
     ΕΝ
     ΤΟΙΟ
```

Fig. 2: An Ancient Lives fragment image (left). A digital consensus transcription for the same Ancient Lives fragment (right).

A. *Preprocessing Stage*

The Ancient Lives interface is directly linked to a MySQL relational database that houses all transcription information. For every click a user makes on a digital papyrus image, the database will store the unique user-id of the ”citizen”, the user's relative click location for the letter (*i.e.*, x,y coordinates), and the citizen's letter choice in unicode. From the database, we retrieve all click information and categorize user clicks into separate files for each fragment. Separating and organizing the click data by fragment allows us to more easily analyze click information on the basis of individual fragments. More specifically, the procedure encourages the detection of strong dissimilarities between individual user clicks and the consensus clicks for a given fragment (*i.e.*, an accidental click on the image).

In some cases, a digital papyrus image is incorrectly oriented in the Ancient Lives database or a user might transcribe a fragment sideways while maintaining an accurate transcription. Subsequently, the click-data coordinates that are relative to the image are also incorrectly oriented. Where applicable, processing rotations in click-data is a necessary step in order to ensure line sequences are correctly formed. From the Ancient Lives MySQL database, we also query relevant rotation information (*i.e.*, the rotation degree used by users to transcribe) for each fragment. Using the rotation information retrieved from the database, we determine the correct rotation degree of each fragment by identifying the most frequently used orientation among all users during the transcription process. Afterwards, all click-data undergoes a rotation filter that adjusts x,y coordinates based on the identified orientation degree.

B. *Aggregation of User Clicks into Consensus Clicks*

Two unique approaches were developed for the task of aggregating all user clicks for a given fragment to form consensus identifications for letters. We refer to the first approach as the kernel-based approach. This approach was written in Matlab and leverages kernel density estimation, a mathematical approach for inferring the likelihood that a variable will take

[2] https://opencv.org

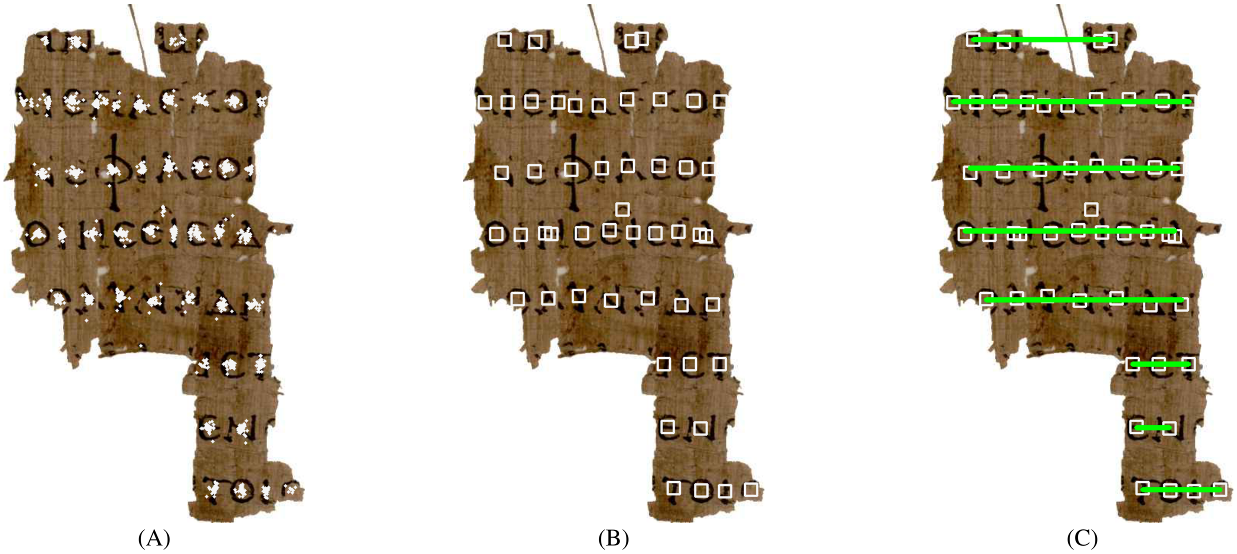


Fig. 3: A visualization of the processing performed by the pipeline using the same Ancient Lives fragment from Figure 2. A) White dots represent the 1,591 clicks of all users. B) White squares represent calculated consensus clicks from the aggregated click data of all users for the fragment using the stepwise aggregation approach . C) Green lines represent the regression line for the nearby consensus clicks.

on a given value, to identify consensus clicks and letters [3]. The algorithm begins by distributing all user click data into a number of bins based on the click's x,y coordinates. The number of bins is determined by multiplying a user-specified kernel width by 2. If no kernel width is specified by the user, the kernel width is assigned a default value of 8. Within each unique bin, the algorithm will identify the highest kernel density peaks, which represent the presence of a consensus letter. Once peaks have been identified within each bin, a filtering function is imposed to prevent duplicates and eliminate suspected false consensus letters. The x,y coordinates of the remaining kernel density peaks are clustered and used to determine the location of consensus letters.

Due to the nature of calculating the kernel density estimation for millions of user clicks, the kernel-based approach requires a large amount of computational overhead and takes multiple days to process user click data to yield consensus letter identifications. In order to hasten the processing time, a second approach, referred to in this paper as the stepwise aggregation approach, was developed in Python. This approach relies on a recently established concept that citizen scientists who complete more classification tasks have an elevated level of knowledge and reason in classifying data correctly than those who complete fewer classification tasks [4]. Based on the concept that expertise can be represented by experience or frequency of activity, the algorithm will first identify the user that has made the highest number of clicks on the fragment and use their clicks as seed locations for potential consensus letter identifications. Depending on an unprocessed click's proximity to pre-existing seed locations, the remaining user clicks are either merged with a preexisting seed location or used to establish a new consensus letter location and added as a seed location. Once all user clicks have been processed, a centroid, or center point, of each agglomeration of clicks is identified and recorded as a consensus letter (see Figure 3B).

C. Creation of Line Sequences from Consensus Clicks

Using the consensus letter identifications from the previous stage, the line sequence creation component will attempt to form line sequences that closely resemble the text presented in the digital image of the papyrus fragment. The input of line sequence creation component is a text file containing a list of x,y coordinates with associated Greek characters. The output of the line sequence creation component is a text file containing a line sequence, or string, composed of all of the characters that appear in the input file.

The algorithm begins by sorting all clicks into a list based on the y coordinate. Beginning at a y-coordinate of 0 and ending at a y-coordinate equal to the height of the relative fragment's digital image, the algorithm searches the sorted y-coordinates and identifies the presence of lines based on gaps of vertical space between neighboring y-coordinates (see Figure 3C). When a line is identified, the y-coordinate is added to a list of line regions. After each click has been grouped into a line region based on its relative x,y coordinate, the best fit line, or regression line, for each line region is calculated. In addition to the equation, the average space between neighboring line regions is calculated. Using the equation of the best fit line as a reference, a second pass of all y-coordinates is made in order to ensure that each letter was categorized in the correct line region. If an x,y coordinate is not within half of the calculated average space between neighboring line regions from the relative line's median y-coordinate, the coordinate is categorized into another line region. After the second pass of y-coordinates has finished, each line region is sorted by the x-coordinate in order to ensure characters appear in the same order they appear on the papyrus. After the regions have been sorted by x-coordinate, the regions are concatenated into a single string, which represents the line sequence for the fragment.

There are two types of styles of line that are presented in papyrus. The first style is parallel where lines are written in straight, distinct, and predictable lines and are equidistant from neighboring lines. The second style is curvilinear where lines are written in the shape of an arc and are unpredictable in direction. For most papyrus fragments with curvilinear lines, identifying a consistent amount of vertical space between line regions is nontrivial. As a result, the described approach could produce duplicate lines by identifying multiple line regions from a single curvilinear line due to incorrect measurements of vertical space between line regions. In order to filter duplicate line regions, a final post-processing stage will remove a line region if it shares 70% or more identity with its neighboring line region.

III. Pipeline Evaluation

In this evaluation, we examine the efficacy of Stage 1 and Stage 2 separately. In addition to fragments that have yet to be transcribed, a group of published fragments with known transcriptions were deposited into Ancient Lives in order to make assessing the effectiveness of each pipeline component possible. Given that each is supplied with the same set of click data, both the kernel-based approach and the stepwise aggregation approach are scrutinized on the ability to correctly classify a letter by comparing consensus click data to the click data of published fragments given by a professional papyrologist. Similarly, the performance of the line sequence creation component is evaluated based on the similarity between line sequences produced by the component and digital fragment transcriptions as they appear in a published P. Oxy. volume.

A. Evaluation of Consensus Letter Identifications from Users

We randomly selected 54 published fragments from the Oxyrhynchus collection to be used as evaluation criteria for measuring the accuracy of consensus letter identifications from both the kernel-based approach and stepwise aggregation approach in comparison to known letter identifications that appear in P. Oxy. volumes. In this assessment, the default kernel width value of 8 is used in the kernel-based approach. In order to evaluate on the basis of user clicks made by citizen scientists, clicks made by professional papyrologists have been removed from the click data set used to make consensus letter identifications. Each fragment was categorized based on handwriting style and legibility into one of the following groups: non-cursive, semi-cursive, and cursive. We utilize three established metrics, precision, recall, and F_1 score [5], for determining the classification performance of each approach. Precision is calculated by dividing the number of correct letter identifications by the number of total letter identifications in the consensus transcription. Recall is calculated by dividing the number of correct letter identifications by the total number of letter identifications in the relative fragment's P. Oxy. transcription. Both precision and recall are combined into a composite metric, the F_1 score. The equation to calculate the F_1 score for an individual fragment is:

$$F_1 = 2 \times \frac{P \times R}{P\text{+}R} \tag{1}$$

where P and R are respectively the precision and recall of the fragment's click data. A F_1 score of 0.0 can be interpreted as an approach correctly classifying next to none of the letters while a F_1 score of 1.0 can be interpreted as an approach correctly classifying most or all of the letters.

B. Evaluation of Line Sequence Creation Component

In this evaluation, the accuracy of the line sequence creation component is examined by supplying the component with professionally curated click-data, where each click represents a correct letter at the correct relative x,y coordinate. From the same set of fragments used in the previous evaluation, a subset of 41 fragments were selected to be used as evaluation criteria to examine the sequence similarity of consensus line sequences produced through the line sequence creation component and the published digital transcription of the same fragment. All 41 fragments were chosen on the basis that a digital transcription exists for the fragment. For the remaining 13 fragments in the set of fragments used in the previous evaluation, a digital transcription does not exist. Each fragment with a digital transcription was categorized as having either parallel lines or curvilinear lines. Of the 41 fragments, 30 were categorized as having parallel lines and 11 were categorized as having curvilinear lines. The edit distance, or Levenshtein distance [6], is employed as a metric to measure the similarity between the transcription produced through the line sequence component and the transcription that appears in a P. Oxy. volume. An edit distance of 0 represents complete identity between two strings (*i.e.*, an exact match). For an edit distance that is greater than 0, at least one insertion, deletion, or substitution was required for one sequence to resolve to the other.

IV. Results

A. Consensus Letter Identification Evaluation Results

The results of the evaluation for Stage 1 suggest the stepwise aggregation approach produces a higher level of accuracy for correctly determining consensus letter identifications than the kernel-based approach, especially for fragments with cursive handwriting (see Table I). The small difference in performance between the kernel-based approach and the stepwise aggregation approach can be explained by how each

Handwriting Style	**Average Kernel-Based F_1 Score**	**Average Stepwise Aggr. F_1 Score**
Non-cursive (34)	0.65	0.67
Semi-cursive (12)	0.64	0.68
Cursive (8)	0.61	0.69
Aggregated (54)	0.64	0.67

TABLE I: Average F_1 scores of consensus letter identifications in each handwriting style from both the kernel-based approach and stepwise aggregation approach.

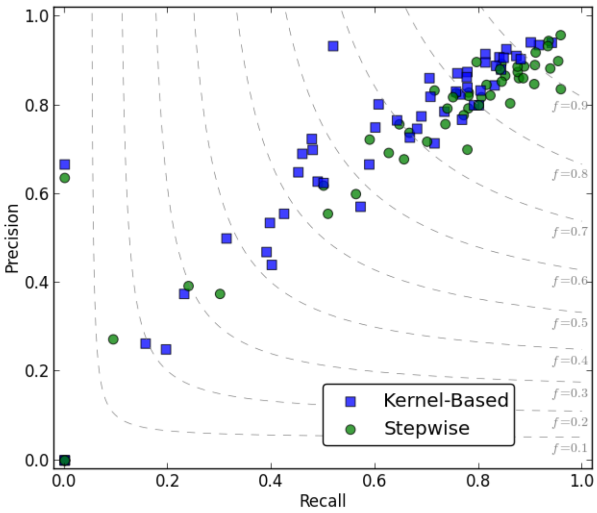


Fig. 4: A Precision-Recall graph that visualizes how the precision and recall of each consensus letter identification from both approaches align with respect to calculated F_1 scores in Table I.

method extrapolates on user clicks. In the stepwise aggregation approach, every user click is used to create the consensus click data set for a fragment. In the kernel-based approach, every user click is also used to create the consensus click data set, but in order to prevent duplicate letter identifications, the list of suspected consensus clicks undergoes a filtering process, which has the potential to remove true-positive letter identifications and decrease identification performance. The difference in accuracy of correctly classifying true-positives and true-negatives can be visualized with the precision and recall (see Figure 4). In addition to producing a higher level of accuracy, the stepwise aggregation approach has an accelerated execution time in comparison to the kernel-based approach. The total execution time for the stepwise aggregation approach is currently about fifteen minutes while the execution time or the kernel-based approach spans a few days.

B. Line Sequence Creation Component Evaluation Results

The results of the evaluation for Stage 2 suggest that the line sequence creation component is effective at creating sequences correctly for most fragments with parallel lines (see Table II). Of the 30 fragments categorized as having parallel lines, eleven sequences created through the line sequence component were exact matches to their digital transcription counterpart. Of the 11 fragments categorized as having curvilinear lines, only one sequence created through the line sequence was exactly matched to its digital transcription counterpart. There is a clear distinction between the component's effectiveness for fragments containing parallel lines and the component's effectiveness for fragments containing curvilinear lines. The difference in performance can be explained by the current approach's inability to consistently determine which letters belong to which line in fragments with curvilinear lines.

V. Conclusion

In order to more easily interpret the large amount of letter classification data from over a million users, we presented a new computational pipeline for translating millions of user-clicks on digital images of Ancient Greek papyri to digital, consensus transcriptions that closely resemble the format of the original papyrus. Both professional papyrologists and student papyrologists can utilize the digital consensus transcriptions produced through the pipeline to more quickly examine, edit, and publish fragments with confidence. Despite being designed specifically for Ancient Greek papyrus fragments, classification projects that share the task of forming either consensus letter identifications or consensus lines of text from coordinate click-data can take advantage of the computational pipeline.

By engineering a pipeline for interpreting the millions of user-clicks from Ancient Lives, we are dramatically re-defining how professional papyrologists and scholars interact with ancient papyri. Typically, a papyrologist could spend days, weeks, or months manually transcribing multiple papyrus fragments. Both the pipeline and the Ancient Lives project leverage the work of citizens in order to help the papyrologist more quickly transcribe and evaluate fragments. There are a number of computational systems that have already made use of the pipeline's output and share the goal of bringing ease to the transcription process. Greek-BLAST [7], for example, is a variant of BLAST, a popular genetic sequence alignment tool, designed specifically for suggesting identifications for literary papyrus fragments. Consensus transcriptions of literary fragments made through Ancient Lives can be supplied to Greek-BLAST directly as input and quickly aligned with matches in Ancient Greek literary manuscript databases (*i.e.*, The Perseus Digital Library [8]). Additionally, collaborators at the University of Minnesota have developed a web-based tool for quickly curating the digital consensus transcriptions produced from the computational pipeline. Curated consensus transcriptions are stored in a database that will later be used for data mining purposes. Lastly, the consensus transcriptions

Line Style	**Average Fragment Length**	**Average Edit Distance**	**Error Ratio**
Parallel (30)	43.7	6.8	15.6%
Curvilinear (11)	234.0	83.3	35.6%
Aggregated (41)	94.7	26.1	27.6%

TABLE II: Average fragment lengths, average edit distances, and error ratios for the 41 fragments used in the line sequence creation component evaluation. Error ratios are calculated by dividing the average edit distance by the average fragment length.

made through Ancient Lives will be the basis for many fragments that will be further studied, edited, and published in *The Oxyrhynchus Papyri* volume series and Proteus, a new interactive, web-based platform that leverages advanced computational methods and techniques to both the study and analysis of ancient texts and the creation of next-generation digital editions.

A. Future Work

A key component of future work is improving the line sequencing stage of the pipeline. For fragments written in a curvilinear manner, forming lines is a nontrivial task. We will investigate measures to help identify the presence of curvilinear lines and how the accuracy of the existing approach for developing line sequences can be improved. A final re-design of the pipeline will take place after additional classification information (*i.e.*, methods for identifying line information or missing papyrus) is added to the Ancient Lives framework.

Acknowledgments

We would like to acknowledge both the Adler Planetarium and the University of Minnesota Supercomputing Institute for access to their computing resources. Additionally, we would like to acknowledge the Arts and Humanities Research Council for funding ”Proteus” and Jisc and the National Endowment for Humanities’ Digging into Data Challenge program for funding ”Resurrecting Early Christian Lives”. We would also like to acknowledge Nita Krevans, Trevor Wennblom, and Philip Sellew for their contributions to this study. Last, but certainly not least, we would like to thank the Egypt Exploration Society and the Ancient Lives project for access to the data set used by the pipeline.

References

[1] A. K. Bowman, R. A. Coles, N. Gonis, D. Obbink, and P. J. Parsons, *Oxyrhynchus: a City and its Texts*. Egypt Exploration Society, 2007, vol. 93.

[2] M. Moyle, J. Tonra, and V. Wallace, “Manuscript transcription by crowdsourcing: Transcribe Bentham,” *Liber Quarterly*, vol. 20, no. 3/4, pp. 347–356, 2011.

[3] K. Fukunaga and L. Hostetler, “The estimation of the gradient of a density function, with applications in pattern recognition,” *Information Theory, IEEE Transactions on*, vol. 21, no. 1, pp. 32–40, 1975.

[4] E. E. Prather, S. Cormier, C. S. Wallace, C. Lintott, M. J. Raddick, and A. Smith, “Measuring the Conceptual Understandings of Citizen Scientists Participating in Zooniverse Projects: A First Approach,” *Astronomy Education Review*, vol. 12, no. 1, p. 010109, 2013.

[5] C. J. Van Rijsbergen, “Foundation of evaluation,” *Journal of Documentation*, vol. 30, no. 4, pp. 365–373, 1974.

[6] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions and reversals,” in *Soviet physics doklady*, vol. 10, 1966, p. 707.

[7] A. C. Williams, H. D. Carroll, J. F. Wallin, J. Brusuelas, L. Fortson, A.-F. Lamblin, and H. Yu, “Identification of Ancient Greek Papyrus Fragments Using Genetic Sequence Alignment Algorithms,” to appear in Proceedings of the 1st Workshop on Digital Humanities and e-Science, 2014.

[8] D. A. Smith, J. A. Rydberg-Cox, and G. R. Crane, “The Perseus Project: A digital library for the humanities,” *Literary and Linguistic Computing*, vol. 15, no. 1, pp. 15–25, 2000.