CHAPTER 10

Engaging Greek: Ancient Lives

James Brusuelas
University of Oxford

Abstract

Since July 2011, Ancient Lives has recorded well over 1.5 million transcriptions of ancient Greek papyri (over 9 million characters), the work of over 105,000 unique online collaborators. The result was not simply the creation of big data, but the inception of an entirely different way of conceiving and interfacing ancient digital texts. Put simply, Ancient Lives has created something that has never existed before: a database of unedited Greek. We have strings of Greek characters without word division or any modern editorial convention. The purpose of this chapter is to discuss, first, the Ancient Lives' methodology of public engagement, the inclusive process by which the public participates in the fundamental tasks of papyrology (this includes both untrained and the unique users targeted by Almas and Beaulieu). Next, the success of any crowdsourcing project depends not only on data input but also how that data is subsequently processed and utilized. An overview of current development then follows, which particularly addresses Ancient Lives' interest and continual use of machine intelligence and genetic sequence alignment algorithms (examples of successfully repurposed field-specific algorithms, an often challenging process as discussed by Tarte), to process multiple transcriptions of a single fragment (version control), query, data mine, and edit these crowdsourced transcriptions within an innovative digital environment. More importantly, in providing public access to data that was for a century viewed only by a handful of scholars, Ancient Lives continues to engage in changing models of traditional scholarship.

How to cite this book chapter:
Brusuelas, J. 2016. Engaging Greek: Ancient Lives. In: Bodard, G & Romanello, M (eds.) *Digital Classics Outside the Echo-Chamber: Teaching, Knowledge Exchange & Public Engagement*, Pp. 187–204. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bat.k. License: CC-BY 4.0.

1 Introduction

It's madness. Within the field of Classics and its subset Papyrology, that phrase, in one form or another, was often the response to even the slightest mention of Ancient Lives (hereafter AL) in 2010, when I arrived at Oxford to begin work on the project.[1] A collaboration between Oxford Classics and Astrophysics,[2] AL was to join the many other crowdsourcing projects hosted by the Zooniverse.[3] As one might guess, upon hearing the word crowdsourcing, such a reaction usually came from senior academics. But perhaps the most colorful comment I can recall was the description of AL as the 'bastardization of the papyri,' uttered by a young postgraduate student at an academic gathering in Leiden in 2012, months after the project had launched—I happened to be sitting on the other side of the table and, with arguably too much delight, responded, 'Oh yes, my project.' What a thought, an experiment indeed. Let anyone, trained or untrained, transcribe a papyrus fragment of ancient Greek online. Let the world assist in transcribing the seemingly countless papyrus fragments from the ancient city of Oxyrhynchus, housed in the Sackler Library of the Ashmolean Museum.

Since their discovery this body of well-known fragments has reintroduced to the world texts that have not been seen since antiquity, such as the Gospel of Thomas and the poetry of Sappho,[4] and although Oxford has held them for over a century, the opportunity for discovery still lingers seductively; due to the sheer volume of fragments, more texts and authors are still waiting to be found. From its very inception, then, AL touched a distinct nerve: access. Looking back, it was not so much about crowdsourcing but access to viewing unpublished material. What happens if someone with no formal training accurately transcribes a fragment? Worse still, what happens when a self-taught individual, using the same tools available to scholars, contextualizes or even identifies a fragment? The cardinal rule, after all, of working with ancient manuscripts is that their text looks nothing like the modern printed editions through which students and the vast majority of scholars engage their content. A Greek papyrus fragment is a perfect example. It is just a string of characters without word division and little to no punctuation, not to mention issues such as scribal errors, variant readings, new words, and cursive handwriting reminiscent of a doctor's prescription. It is not a simple reading experience. Accordingly, a distinct scholarly identity has been constructed around them; one that, as noted above, cuts across generations. For the laymen to walk in off the street and successfully perform certain academic and papyrological tasks, even if only at a rudimentary level… That idea was not just brushed off, but seemed threatening to some. It seemed that any success achieved by AL would be at the expense of Papyrology and even Classics, or at least demystification of the academic process to a certain degree.

Be that as it may, I found myself in a peculiar position. I was tasked with creating a dialogue between academics, our beloved primary source material, and … everyone else. Better (or worse) still, I had to create this dialogue even if certain segments of my field simply were not interested. Now, it is not that AL was devoid of supporters in the beginning. There were, and still are, many colleagues interested in engaging the general public about Classics, classicists, and the Greek and Latin languages in a living dialogue rather than from any position or notion of 'gatekeeper;' these languages are not oracles nor are we the Delphic priestesses and priests uniquely capable of interpreting them. And so many colleagues and friends have long suggested I write something about my involvement with AL and my ongoing role as its leading voice. For the invitation to contribute to this book, I am thus very grateful for the opportunity (or, better put, the motivation of a deadline) to write a simple essay about my involvement in the development of Ancient Lives and what the project has thus far achieved.

As I write, AL is in the process of being rebuilt for re-launch. This is both to improve its functionality, its overall frontend and backend design (a Rails app about to become a Backbone.js app), and to upgrade the application to conform with current Zooniverse standards. AL is changing, morphing into something else.[5] What follows now is nothing more than a simple reflection on how AL initially produced millions of transcriptions of useable data, engaged in machine learning for processing this data, and dabbled in Bioinformatics for the purposes of automated text identification.

2 Patterns and Users

With the lure of discovering new texts of Greek literature or even a new gospel, we always expected classicists of all skill levels to play with the interface. And they did. Nevertheless, although AL embraces the volunteer community as a whole, including trained classicists, development of the interface was always focused on the individual with no knowledge of ancient Greek. So, how does one produce an environment that facilitates participation and contribution from those outside of academia? Pattern recognition.

The fundamental premise upon which AL operates is pattern recognition. It is a task at which the human brain excels. And so one does not need to know the dynamics of ancient Greek grammar and syntax to recognize the triangular shape of alpha and delta (A, Δ) or especially a familiar shape like nu (N). A simple image and a keyboard of Greek characters are all one needs (Figure 1).

Be that as it may, like any evidence of human generated script, character shape is not consistent and the degree of cursive can be slight to severe. Moreover, the alphabet present in the papyri is devoid of any cognitive notion of modern upper or lower case; there really is only an 'upper' case, even though it appears to mimic those distinctions, such as the case of alpha being triangular (A) or round (α). From a development standpoint, that caused a bit of a

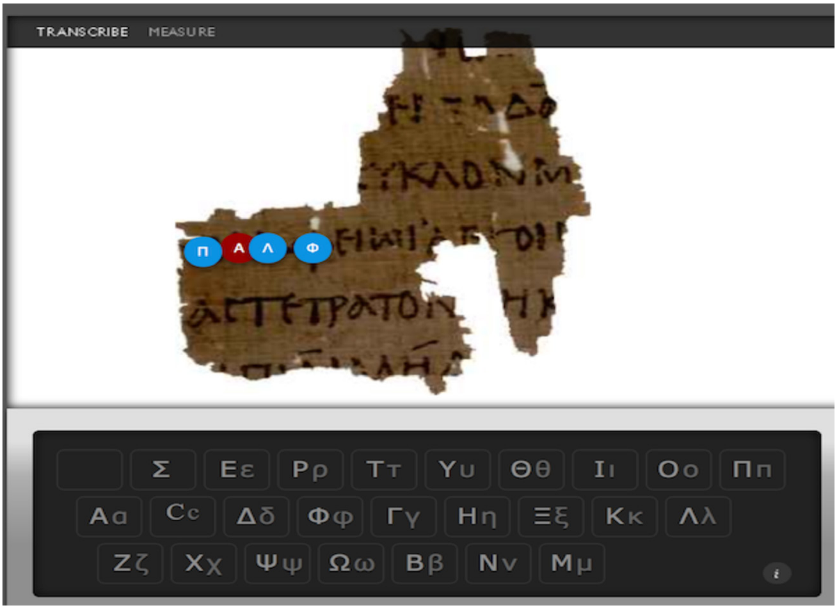


Figure 1: Current AL keyboard.

dilemma. There is no 'paleographical' keyboard devoted to the character shapes found in ancient papyri. For a virtual keyboard we only have Unicode character shapes, of which many directly correlate to the later Byzantine Greek minuscule preserved in parchment codices. Moreover, to produce useful data, it is indeed those Unicode characters that are the required input. In a papyrological context, while a trained user could immediately recognize that they are looking for lunate sigma (c) and not medial and final sigma (σ, ς), or that delta (Δ) does not look like the minuscule form (δ) in papyri, the question over what visual data to provide the crowd lingered. Furthermore, from the crowdsourcing standpoint, users not only needed to focus on finding patterns and matching character shapes, but also the intuitive freedom to provide data without being bogged down in a host of variables that would cause hesitation. In other words, that first moment of interface is not a moment for actual training or official indoctrination in Greek manuscripts. Motivation to engage in the task of classification must arise from the simple notion of pattern matching, not necessarily knowledge of ancient Greek or Greek paleography. The user must recognize that the digital tools before them facilitate their contribution. If there were even the slightest hint that formal training was needed, participation by the general public would have most likely been stifled to a large degree. Accordingly, the solution at the time was to provide users with a standard Greek Unicode keyboard, including both shapes that appear in papyri and even those that do not. In addition, to assist with the difficulty that arises as letter shapes become cursive, images of cursive forms were made available by hovering over a given character.

In psychological parlance, AL would most likely straddle both the 'template matching' and 'feature analysis' theories of human pattern recognition. Some literary hands are nearly font-like, and users are explicitly pairing character shapes in an image with a character 'template' in the keyboard. But as documentary hands become more and more cursive, the general feature of epsilon (Figure 2), for example, must be recalled when classifying cursive character shapes.

In late July of 2011, when AL went live, the Zooniverse had a community of roughly 400,000 users – it is now over 1 million. To a large extent, and considering the positive reception during beta testing, we knew the Zooniverse community would provide data, at least enough data to evaluate the experiment. But would the general public engage in transcribing ancient Greek papyrus fragments? The question still remained. Fortunately, the answer was not only a resounding 'yes,' but AL, due to the media attention we received, even brought in new users into the Zooniverse. The general public was and is indeed interested in what papyrologists do. Moreover, the characters shapes themselves, both clear and cursive, and the random bits of ancient art visible on some papyrus fragments inspired the imagination of the volunteer community.[6] And as the world outside academia became more informed about this vast number of papyrus fragments from Oxyrhynchus, the idea of contributing to the discovery of a lost work was a profound source of motivation. By the end of the first year of the project, AL recorded 1.5 million transcriptions, roughly 7 million Greek character classifications – currently over 9 million have been recorded. What became immediately apparent, and not unexpected in the Zooniverse, was the appearance of so-called 'power' or 'super' users, individuals who contribute hundreds of transcriptions as opposed to the majority of users that were only producing a few.[7] And so there was this segment of the crowd that wanted to talk with papyrologists and classicists about what they love, discuss ancient literature and history, and simply help. This nodal point of interaction and outreach is unique. This is Classics in culture, happening in real time and not defined by the parameters of a classroom or even a university campus.

Despite such interesting variables, however, no in-depth study of AL users, both in the context of the Zooniverse community as a whole and in relation to other crowdsourcing projects, has been conducted. To get a feel for the AL community one must visit Talk. Every Zooniverse project is equipped with a Talk section, a place where the members of a specific community of a given project talk to one another, as well as project members. This is a place to isolate interesting images, flag them, ask questions, and essentially acquire further knowledge about a project's data. For the AL community we should note their engagement in self-learning. As an expert, one of the unique aspects of working with users in Talk is not being the never-ending voice of 'no' or 'wrong.' They may be there to help you, but they did not sign up for your class. This is an

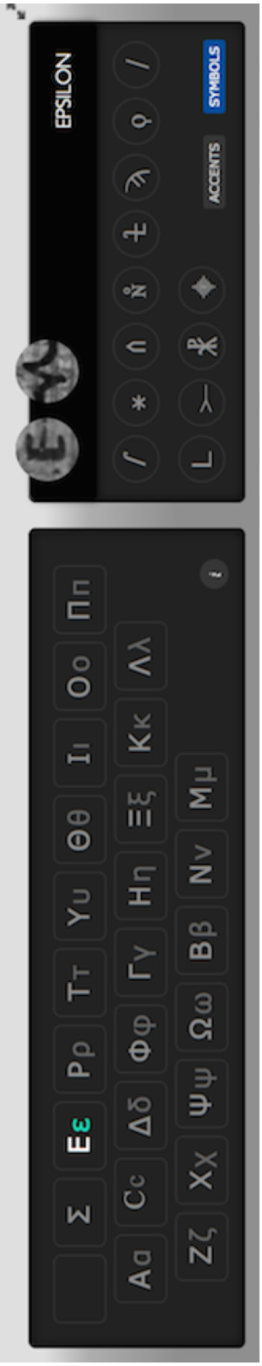


Figure 2: Cursive shapes.

exploration in real time, in their private time. The biggest mistake would be to drive volunteers away by constantly hammering home what they do not know. The user who actively engages in Talk is someone that wants more information, wants to improve the accuracy of their classifications. As an expert, your voice is simply one of information, not evaluation. After providing basic guidance it is often better to step back and let the users help each other. After AL launched and users began to get acquainted with the various kinds of handwriting present, it was not long before individuals began posting online links for Greek paleography, especially those showing examples of the more difficult cursive forms of certain Greek characters. Soon users were helping each other classify the more difficult forms of cursive epsilon, for example. And discussion pertaining to the AL keyboard and characters shapes not present in the papyri become commonplace, especially the topic of lunate sigma vs. the medial and final sigma forms of the later Greek minuscule.

How did AL generate so many crowdsourced transcriptions? We simply gave the crowd images, a virtual Greek keyboard, and an intuitive task.[8] With so many users in the Zooniverse, AL then generated what can be described as Big Data, a term not necessarily devoid of ambiguity. But in our case, since AL creates multiple transcriptions of a given fragment, processing the data posed a great challenge.

3 Enter the Machine: Consensus, Line Sequencing, and Greek BLAST

Having over a million transcriptions tucked away in a MySQL database allowed for easy interaction with AL data, if one knew how to write a MySQL query. The number of papyrologists and classicists that can, however, never seems to be very large. Consequently it became rapidly clear that AL required serious computational support if its data was going to be made useful to those without any knowledge of coding.[9]

One of the principal tenets of papyrology is that more than one pair of eyes is always better. Whether a student or an experienced scholar, establishing a transcription and eventually a final edition is not produced in isolation. We often see different shapes, and in reconstructing a fragmented ancient text the most accurate product is never the result of just one pair of eyes. The size of ALs papyrological database may have been unprecedented, but the required methodology for processing was no different. For each fragment we needed a consensus transcription. How to extract a consensus from the database then emerged as a machine learning challenge. We needed an algorithm that could be trained to batch process millions of transcriptions. Accordingly, it also offered the opportunity to bring the transcription data face to face with the experts.

To tackle this problem AL collaborated with the Minnesota Supercomputing Institute (MSI) and the departments of Classics and Near Eastern Studies and Physics and Astronomy at the University of Minnesota. Dr. Haoyu Yu from MSI was tasked with writing a consensus algorithm. In doing so, one must remember that AL is very different from a transcription project like Transcribe Bentham, whose input is plain text and supplemental XML tags.[10] Again, if you want the world to help transcribe ancient Greek, one cannot assume their varied devices are equipped with the necessary Greek keyboard. Instead of recording plain text, a virtual mapping is employed. For each click on a papyrus image, the database not only stores the Unicode character selected, but also the relative click location as x,y coordinates. For the aggregation of user clicks, the initial approach was written in Matlab, employing kernel density estimation—that is, mathematically inferring the likelihood that a variable will take on a given value—to isolate consensus clicks and letters. Besides giving different transcriptions, users will also not click the same exact location on an image, resulting in both multiple characters and multiple sets of x,y coordinates for one character position. Looking at the multiple transcription data of one fragment, the algorithm essentially takes the x,y coordinates for each click position and distributes them into a number of bins according to the search radius, a number determined by multiplying a user-specified kernel width by 2 (the default value is 8 if no kernel width is specified). Within each bin the algorithm finds a consensus letter by identifying the highest kernel density peaks. The x,y coordinates of those peaks are then clustered to create consensus characters and their locations (pixel locations), whereby a virtual image of the fragment can be visualized (Figure 3).

Training the algorithm to successfully render consensus also meant evaluating the resulting user consensus. This was accomplished by performing kernel density estimation against a select group of fragments transcribed by volunteers and then compared with the transcription of expert papyrologists. On this select group of fragments, which included examples of clear literary hands, semi-cursive, and cursive documents (marriage certificates, land leases, personal accounts, private letters, etc), we created a correspondence between the expert's characters and locations and that of the consensus. For clear literary book hands, as seen in Table 1, comparison yielded the following.

Number of expert locations	Number of consensus locations	Number of overlapping locations	Percent coverage of expert locations by consensus locations	Number of matches in the overlapping locations	Percent match in the overlapping locations
100	150	95	95.00	85	89.47

Table 1: Consensus evaluation.

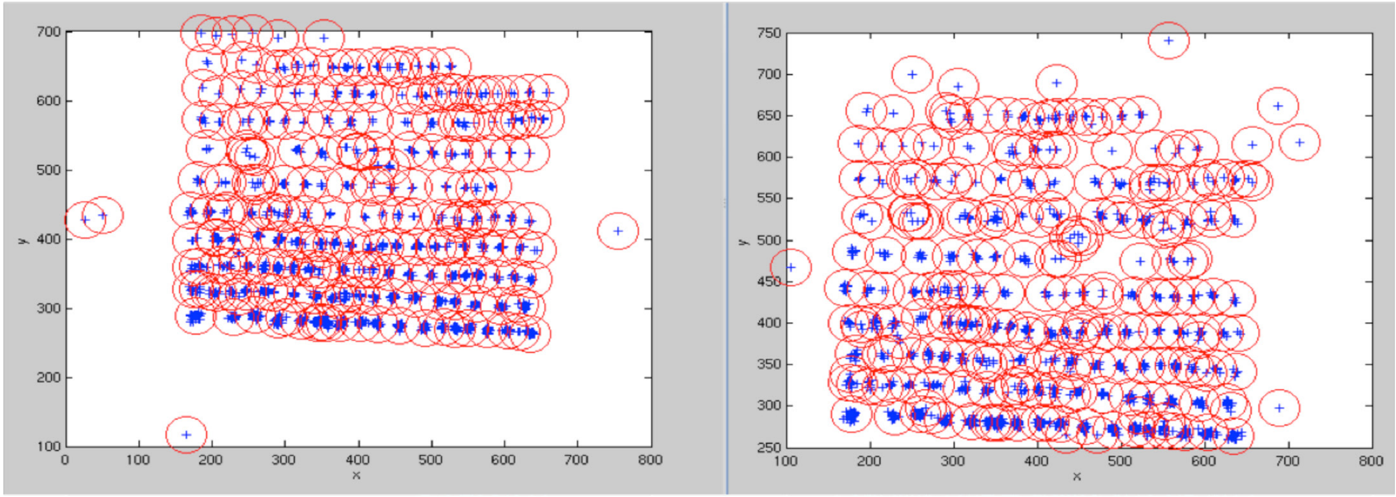


Figure 3: Kernel Density Clustering. This image shows how clustering the highest kernel density peaks (user clicks storing character and x,y coordinate data) in Matlab also reveals an abstract image of the papyrus itself.

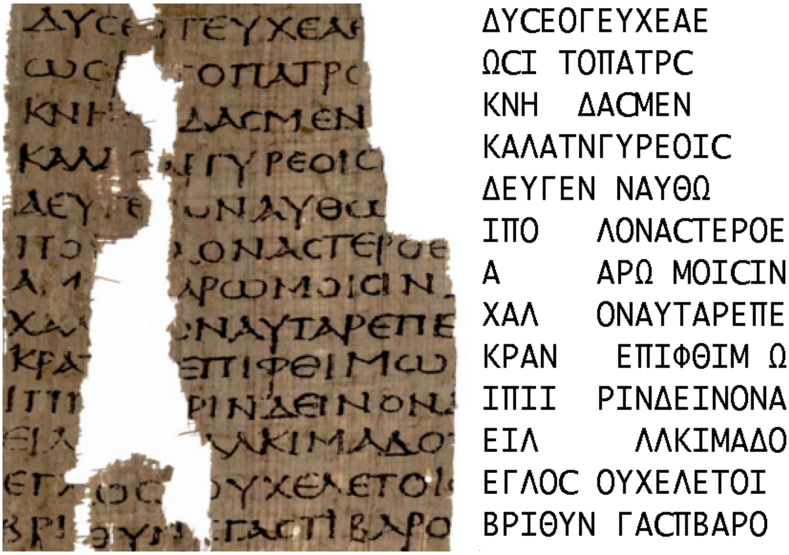
AL volunteers, as a community, could thus provide excellent data. Users were capable of providing transcriptions that were nearly identical to those given by experts. That said, gathering good data from a clear literary hand was never really in question. It was the idiosyncratic hand styles and cursive writing that posed the largest potential problem. Conducting further isolated investigations we began to see what was expected. Looking at 31 examples of semi-cursive scripts, the percentage of agreement dropped to 65%, while 14 examples of very difficult cursive hands yielded a percentage agreement of only 51%.

Although consensus characters could be successfully gathered, kernel density estimation in Matlab proved to be a computationally cumbersome task. It took multiple days, hundreds of hours, to process the data. More recently we addressed this issue at Oxford. Alex Williams, research programmer on the Proteus project,and Dr. John Wallin of Middle Tennessee State University developed the stepwise approach using the high level programming language Python.[11] This new method utilizes the recently established concept that members of the Zooniverse community who complete more classifications, the so-called 'super' user, demonstrate a higher ability to correctly classify data than those who complete fewer classifications. This new algorithm thus identifies the user that has made the most number of clicks first and isolates their character positions as potential nodes of consensus. The remaining clicks are then either merged according to pre-existing locations or, depending on their frequency, established as another possible node of consensus. Once finished, a centroid of each agglomeration of clicks is isolated, yielding a consensus letter. Unlike Matlab, the Python script processes the data in minutes. In a Big Data context, this was quite an achievement. However, we did not intend stepwise to supersede kernel density estimation entirely, as both have their merits. Speed is obviously the benefit that comes with stepwise. But the Matlab approach, though slow in processing time, records and allows visualization of all the user data for a given fragment. In comparing user transcriptions of cursive documents with expert transcriptions, we noticed instances where the correct character was essentially hidden under the incorrect consensus character. More evaluation needs to be conducted in order to fully grasp how the AL pipeline might processes the more difficult cursive manuscripts.

The purpose of AL was to explore new methods that could potentially increase the pace at which scholars study and organize this massive body of fragmentary ancient texts, and the fundamental way to do that is through transcriptions, which are important for identifying and contextualizing fragments. This is how we determine what is Homer, Demonsthenes, Simonides, Pindar, etc. Although both the stepwise and kernel density methods could extract a consensus transcription, the output consisted only of characters and x,y coordinates, not an actual text file of Unicode characters in lines corresponding to the papyrus image. To create lines, Dr. Wallin created another Python script to identify the presence of lines based on gaps of vertical space between neighboring y-coordinates. As shown below in Figure 4, the code smartly deduces lines.

The creation of usable strings of Greek Unicode was perhaps the most important achievement for AL; it also completed the initial AL pipeline (Figure 5). From the moment of launch, we proved that the crowd was interested in transcribing, but the onus was always on us to turn their volunteer efforts into useful data. And with these strings there was always one target in the distance: algorithmic identification of fragments.

When I first arrived at Oxford, I was given a few boxes of black and white images and asked to identify whatever fragments I could. After a few weeks of compiling a long list of identifications, it became clear not only how time consuming the process was – and this was just a tiny fraction of the total number of fragments – but authors we expect to find due to the canon in ancient education, like Homer and Plato, were indeed in great abundance. There was so much Homer! For every high priority discovery, such as a new text or the first papyrus evidence for a known author, one had to slowly make their way through multiple copies of works like the *Iliad*. But with the creation of Unicode transcriptions, AL had the opportunity to leverage them against a database of known Greek texts for rapid algorithmic matching. This would not only result in discovering important texts, specifically works only known through select quotation by other ancient sources, but also allow us to quickly isolate and batch known material, and thus turn our attention to the literary texts that could not be matched.



```
ΔΥϹΕΟΓΕΥΧΕΑΕ
ΩϹΙ ΤΟΠΑΤΡϹ
ΚΝΗ  ΔΑϹΜΕΝ
ΚΑΛΑΤΝΓΥΡΕΟΙϹ
ΔΕΥΓΕΝ ΝΑΥΘΩ
ΙΠΟ   ΛΟΝΑϹΤΕΡΟΕ
Α     ΑΡΩ ΜΟΙϹΙΝ
ΧΑΛ   ΟΝΑΥΤΑΡΕΠΕ
ΚΡΑΝ   ΕΠΙΦΘΙΜ Ω
ΙΠΙΙ  ΡΙΝΔΕΙΝΟΝΑ
ΕΙΛ      ΛΛΚΙΜΑΔΟ
ΕΓΛΟϹ ΟΥΧΕΛΕΤΟΙ
ΒΡΙΘΥΝ ΓΑϹΠΒΑΡΟ
```

Figure 4: Line sequencing.

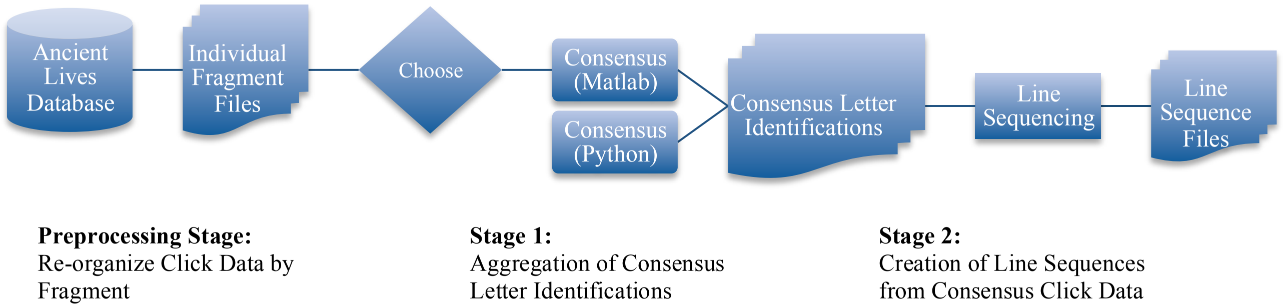

Ancient Lives Database
Individual Fragment Files
Choose
Consensus (Matlab)
Consensus (Python)
Consensus Letter Identifications
Line Sequencing
Line Sequence Files
Preprocessing Stage: Re-organize Click Data by Fragment
Stage 1: Aggregation of Consensus Letter Identifications
Stage 2: Creation of Line Sequences from Consensus Click Data

Figure 5: AL pipeline.

Instead of creating another algorithm for this task, we instead decided to repurpose one from Bioinformatics. The Basic Local Alignment Search Tool (BLAST) is the standard tool for matching amino acid sequences in proteins or nucleotide sequences in DNA.[12] Genes are digitally represented by a sequence of continuous letters, in which each letter represents a specific nucleotide or amino acid. Figure 6 below shows the typical BLAST output.

The serendipitous realization that occurred was that BLAST was essentially thinking in terms of an alphabet, especially in the case of proteins in which twenty amino acids are found. Better still, when comparing genetic sequences an exact match is not necessarily the goal. Thus BLAST was already equipped to account for gaps between aligned sequences. The Greek alphabet not only consisted of 25 characters, but a papyrus text is nothing but a string of Greek characters often separated by gaps, the literal holes in a papyrus, let alone the appearance of variant material such as changes in spelling and scribal errors. All we had to do was simply substitute the characters of the Greek alphabet for those representing amino acids, supply a database of known strings for comparison, and alter how the algorithm scores the identified relationships. In a short period of time BLAST was beginning to think in Greek, as shown in Figure 7.

And so we have Greek-BLAST, which, instead of using the BLOSUM (BLOcks Substitution Matrix)[13] substitution matrix for scoring alignments between protein sequences, now has the Greek Letter Oriented Substitution Matrix (GLOSUM). Put simply, scoring is critical for evaluating instances of match and mismatch resulting from alignment. Greek-BLAST, in particular,

```
Score = 98.6 bits [244),   Expect = 1e-23, Method: Compositional matrix adjust.
Identities = 42/66 (64%), Positives = 57/66 (86%), Gaps = 0/66 (0%)
Frame = -2

Query  199  VAPSITNTPLAQRLLSSSDKEEASAKRHPLHRVGKAKDIGSMAAFLLSDQSGWMTGAILG   20
            +APS+TNTPLA++LLS+   +K++         +RHPL   RVG+AKDI    +M     FLLS++S    WMTGQ+LG
Sbjct   162  IAPSLTNTPLAEKLLSNDEKKKKMDERHPLKRVGEAKDIANMVVFLLSEKSSWMTGQVLG  221

Query  19    VDGGLS    2
             +DGGLS
Sbjct   222   MDGGLS   227
```

Figure 6: BLAST.

```
Score = 68.4 bits (154), Expect = 8e- 13

Ancient Lives fragment: 131383

FRAGMENT  ΠΑ?ΑΔΟ?Ι?ΑΓΑΘΗΚΑΙΠΑΝΔΩΡΗΜΑΤΕΛΕΙΟΝΑΝΩΘΕΝΕ?ΤΙΝΚΑΤΑΒΑ
TEXT          ΠΑΣΑΔΟΣΙΣΑΓΑΘΗΚΑΙΠΑΝΔΩΡΗΜΑΤΕΛΕΙΟΝΑΝΩΘΕΝΕΣΤΙ-ΚΑΤΑΒΑ
SIMILAR       ΠΑ  ΑΔΟ  Ι   ΑΓΑΘΗΚΑΙΠΑΝΔΩΡΗΜΑΤΕΛΕΙΟΝΑΝΩΘΕΝΕ  ΤΙ ΚΑΤΑΒΑ

FRAGMENT  ΙΝΟΝΑΠΟΤΟΥΠΑΤΡΟ?ΤΩΝΦΩΤΩΝ
TEXT          ΙΝΟΝΑΠΟΤΟΥΠΑΤΡΟΣΤΩΝΦΩΤΩΝ
SIMILAR       ΙΝΟΝΑΠΟΤΟΥΠΑΤΡΟ  ΤΩΝΦΩΤΩΝ
```

Figure 7: Greek-BLAST.

evaluates letter-pair matches uniquely, taking into account the frequency of the letter in the known database and confusion likelihood from AL volunteers. Recalling again the reality that literary papyrus fragments are not always a verbatim match with the tradition of medieval manuscripts, which was the principal source of transmission from antiquity, a positive match will not always be an exact one. Consequently, Greek-BLAST needs to bring to our attention output that shows both exact matches and those of potentially interesting similarity. In early 2016 Greek-BLAST will begin interrogating the AL database of papyri from Oxyrhynchus.

4 The Papyrologist in the Shell

As I said in the beginning, Ancient Lives is transforming, becoming something else. In its next iteration new projects based on other collections and even new languages will be incorporated. So this essay is somewhat timely. In the end, what have I learned from AL and crowdsourcing? More importantly, what has AL done? How does AL fit within the community of Digital Classics? What is its significance?

Crowdsourcing, in the context of moving beyond one's niche academic community, works. But this should not come as a surprise. The Zooniverse model has been in operation since 2007. Using that model, AL was also launched successfully without even conducting prior workshops, case studies, or surveys. That may sound cavalier. But AL was and still is more about directly engaging the world outside academia; not employing academic methods was crucial in this respect. Moreover, if untrained users can, as they have shown, produce good transcriptions, that certainly does not come at the expense of Papyrology. AL simply gathers transcriptions in order to re-think how the vast collection of papyrus fragments from Oxyrhynchus might be studied. It is certainly not about the mass publishing of papyri; oversight of the study and publication of these fragments is maintained by the Egypt Exploration Society and the Oxyrhynchus Papyri Project. In aggregating these transcriptions AL has produced something that has never existed before: a database of largely unedited and unpublished ancient texts. Looking at the raw data, there are just strings of information, including that data ultimately removed in the editing process. It is also devoid of XML or any markup convention or standard. Regardless of the kind of manuscript processed through AL in the future, this is what it does. Consequently, this is predominately why AL has been slow to collaborate with other Digital Humanities or Digital Classics projects. Its data is something else. The methods needed for data analysis thus did not exist and required new thinking. We had to invent as we went along, designing the consensus, line sequencing, and Greek-Blast methods.

With this unique database our initial focus has been on fragments of Greek literature. This is primarily due to the research interests of project staff and the quality of data generated from literary fragments. Although Greek-BLAST has yet to be fully deployed, its ability to advance and expedite the identification process has great promise. Bringing to light new texts, whether that means more Sophocles, Aristotle, or even unknown uncanonical gospels, is one way AL can impact the production of new knowledge. But users have also engaged the more cursive documents, the texts of everyday life. How ALs database can impact the study of ancient documentary evidence is very much a topic of future research. And as more data is gathered over time, there are possibly more ways to analyze and visualize this vast dataset that spans roughly from the first century BCE/CE to the Muslim conquest of Egypt in the 8th-century, such as studying and modeling scribal errors and habits, the development and spread of Koine, and the rate and characteristics of bilingualism. I did not even mention the fact that we take measurements of margins, and that AL is potentially housing data that can statistically either prove or modify the way we think about the aesthetics of ancient bookrolls.[14] Machine learning and automated algorithmic mining is the way forward. And it is perhaps time to start thinking not so much about so-called Omega, in the textual criticism sense of trying to reconstruct what an ancient author actually wrote, but the reality of ancient reading and cognition. In the end, one can continue to invent methods for exploiting AL data.

In creating this database I have also been asked numerous questions about digital editing and digital editions – rightfully so if so many transcriptions have been generated. Naturally, this was the next step Dirk Obbink and I took. The year 2016 will not only see the re-launch of AL but also the launch of Proteus, a new ecosystem for digital philology and the creation of born digital critical editions and the textual criticism that underwrites them. Our initial focus is on Greek literary (primarily those constituting direct evidence for an author and/or text) and subliterary papyri (i.e. commentaries, lexica, glossaries, anthologies, etc.). Proteus is a virtual space for parallel critical editing, a process whereby multiple scholars and students can produce digital editions, suggest conjectures, and submit critical notes and translations. As the data from these fragments evolves over time through the re-editing process, Proteus provides a way to interface and examine this change through its search platform; it is designed to not just house multiple editions of a given text, but to spawn multiple editions while simultaneously applying version control. The architecture consists of two components: the Proteus Search Interface and the Digital Editor for Classical Philology (DELPHI). The project is implemented using Python, HTML5, CSS, JavaScript, PostgreSQL database management system, and Apache Solr for search. Its new Digital Editor for Classical Philology (DELPHI) allows for the creation of all the attributes that make an edition critical and citable: critical apparatus, testimonia, paleographical apparatus, diplomatic transcriptions, even the ability to edit marginalia. Along with updating the TEI/EpiDoc/XML standards for Greek literary and subliterary fragments by creating the necessary tags required for creating digital critical editions, DELPHI also employs a markdown concept similar to the Leiden+ system in the SoSOL editor used by Papyri.info and the Perseids project. DELPHI, however, provides automated translation of markdown into full XML and HTML5 in live time; XSLT stylesheets are not used. Moreover, for accents and diacritics, the editor employs a built-in on-screen menu inspired by the Apple OS X Character Accent menu. Proteus' ecosystem is also not Oxyrhynchus-centric. Born digital critical editions of fragments from any collection can be produced. But in order to integrate AL data into the scholarly process of editing, DELPHI will have a user workflow for those working on unpublished fragments from Oxyrhynchus. For those fragments the consensus transcription produced by AL will be provided to their editors. This is an important step for unpublished material. The capture of the digital edition at the inception of the *editio princeps* will remove the need for another party to encode the text at a later stage for use in other projects and digital research.

To conclude, I should say that I am very fond of coding and promote coding literacy whenever possible. As a Classicist, Papyrologist, or any other humanities scholar, coding may not be your job, but whether you are managing or just participating in a Digital Humanities or Digital Classics project, coding literacy ensures that you actually understand the nature of your data. This also facilitates communication with the developers and computer scientists involved. As of now, if your data is going to be useful to your colleagues, new digital tools will most likely be required. When your development team asks what you want to do with your data, the correct answer needs to reflect an actual knowledge of the data. There is a saying in the entertainment industry that it takes just as much time, effort, and money to make a bad movie as it does a great one. Development, especially academic development, is not immune. We can build as many digital tools and algorithms as we like, but if these tools and their output are not being used and cited by the field of Classics at large, then there is a disconnect that needs to be addressed. In that context AL still has more work to do.

Notes

[1] Ancient Lives: <http://ancientlives.org/>.

[2] The project is led by Dirk Obbink (Classics) and Chris Lintott (Astrophysics).

[3] Zooniverse: <http://zooniverse.org/>.

[4] For an introduction to the city of Oxyrhynchus and the importance of the Oxyrhynchus papyri collection, see Bowman et al. 2007 and Parsons 2007.

[5] Ancient Lives will no longer focus on transcribing Greek papyrus fragments from Oxyrhynchus. But other collections and even Coptic manuscripts will be included. Along with this transformation Ancient Lives is now a full partnership between the University of Oxford and the University of Minnesota.

[6] User comments and discussions within AL Talk document a wide range of reactions to the content and images found in Oxyrhynchus papyri. Their opinions and expressions, however, are their own. Since Talk is not open to the public, but a forum for registered users, I encourage exploration of the Ancient Lives site to get a feel for its community.

[7] See Prather et al. 2013.

[8] For a recent study on crowdsourcing in the humanities, see Dunn & Hedges 2013: 147–169.

[9] For further reading on the algorithms involved, see our computational papers: Williams et al. 2014a: 100–105 and Williams et al. 2014b: 5–10. For support in creating these algorithms, I would like to thank the following funding bodies: The John Fell Fund, Minnesota Futures, The Arts and Humanities Research Council, and the National Endowment for the Humanities. Images provided by Alex Williams.

[10] Moyle et al. 2011: 347–356.

[11] Proteus, available: <http://www.proteusproject.uk>; see also <http://www.papyrology.ox.ac.uk/ProteusProject/>. Python: <https://www.python.org>.

[12] Altschul et al. 1990: 403–410.

[13] Henikoff & Henikoff 1992: 10919.

[14] To date Johnson 2004 remains the only comprehensive study of the aesthetics of the ancient papyrus bookroll; the dataset notably comprises of only 413 papyri fragments.

References

Altschul, S. F. et al. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3): 403–410.

Bowman, A. K., Coles, R. A., Gonis, N., Obbink, D. & Parsons, P. J. (2007). Oxyrhynchus: a City and its Texts. *Egypt Exploration Society,* 93.

Dunn S. & Hedges. M. (2013). Crowd-sourcing as a Component of Humanities Research Infrastructures. *International Journal of Humanities and Arts Computing*, 7(1–2): 147–169.

Henikoff S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22): 10915–10919.

Johnson, W. A. (2004). *Bookrolls and Scribes in Oxyrhynchus*. Toronto, University of Toronto Press.

Moyle, M., Trona, J. & Wallace, V. (2011). Manuscript transcription by crowdsourcing: Transcribe Bentham. *Liber. Quarterly*, 20(3/4): 347–356.

Parsons, P. J. (2007). *City of the Sharp-Nosed Fish: Greek Lives in Roman Egypt*. London, Weidenfeld & Nicolson.

Prather, E. E., Cormier, S., Wallace, C. S., Lintott, Raddick, M. J. & Smith, A. (2013). Measuring the Conceptual Understandings of Citizen Scientists

Participating in Zooniverse Projects: A First Approach. *Astronomy Education Review*, 12(1).

Williams, A. C., Wallin, J. F., Yu, H., Perale, M., Carroll, H. D., Lamblin, A-F., Fortson, L., Obbink, D., Lintott, C. J. & Brusuelas J. H. (2014a). A Computational Pipeline for Crowdsourced Transcriptions of Ancient Greek Papyrus Fragments. *Big Data*, *IEEE International Conference on*, 100–105.

Williams, A. C., Carroll, H. D., Wallin, J. F., Brusuelas, J., Fortson, L., Lamblin, A-F. & Yu, H. (2014b). Identification of Ancient Greek Papyrus Fragments using Genetic Sequence Alignment Algorithms. *e-Science*, *2014 IEEE 10th International Conference on*, 2: 5–10.