Scholarly Editing and AI: Machine Predicted Text and Herculaneum Papyri

James H. Brusuelas
University of Kentucky, USA

Abstract   In 2016 the Digital Restoration Initiative (DRI) at the University of Kentucky, under the direction of Professor Brent Seales, virtually unrolled a carbonized parchment scroll from Ein Gedi, revealing a copy of Leviticus written in iron gall ink. In 2019 the DRI applied a new machine learning method to reveal a Greek character written in carbon ink from an actual Herculaneum papyrus fragment. Virtual unwrapping of cultural heritage objects is a reality. The application of machine and deep learning methods to enhance difficult-to-detect ink signals in tomography will continue to evolve. This raises an important question. How will the process of editing texts that are 'true-born virtual' (the object can never be opened to verify the results) change to reflect the presence and dependency on AI? This paper produces a theoretical model for how a critical edition of a virtually unwrapped papyrus text must document the role of the machine. It also engages the possible requirements, in terms of Data Science, that this new type of text compels in order to ensure transparency at the level of its 'birth'. Put simply, a new virtual edition model that is a fusion of humanities and science is needed.

Keywords   AI. Tomography. Born-virtual text. Scholarly editions. Textual criticism. Herculaneum. Papyri.

Summary   1 Editing with AI: Essential Methods First, Not Fantasy. – 2 Herculaneum Papyri and AI. – 3 Challenges in Visualizing and Working with Born-Virtual Text. – 4 Conclusion.

1 Editing with AI: Essential Methods First, Not Fantasy[1]

Editing an ancient language with artificial intelligence. It sounds like a scene in a science fiction movie. But before we get distracted with any fanciful images of sitting at a computer terminal, or perhaps just a screen in a post-keyboard world, and talking with an Alexa, Cortana, or Siri type of entity with expertise in ancient human languages, let us step back and remember that we have been assisted by computational resources for some time now. Regardless of the language of a given manuscript, advanced imaging techniques and image processing have been critical in the scholarly editing process. In the fields of Classics and Greek and Latin Papyrology, which will be the focus of this paper, that process produces the first edition of a papyrus manuscript (the *editio princeps*), the subsequent versions as papyri are constantly re-edited over time, and the collation of manuscripts in the critical edition of one work, such as Aeschylus' *Agamemnon*. As a papyrologist, a basic example would be the use of multispectral imaging and software like Adobe Photoshop (or even just Mac Preview) to enhance the contrast between the ink and the substrate surface. Reconstruction of a text is thus often made possible because the editor is viewing a spectral image of a manuscript produced by imaging it under incident light that has a wavelength of 940 nanometers, for example, and at various contrast settings within an image viewing application. Yet when we read the *editio princeps* or the critical edition, whether it is found in the *Oxford Classical Texts* series, the *Teubner* series, or in a papyrus edition series like *The Oxyrhynchus Papyri*, documentation of the spectral bands used in the editing process is not always added. The software applications used to manipulate the images, let alone the contrast settings used, are most definitely omitted. Put simply, documentation to reproduce the conditions under which the text has been reconstructed, or seen, is seldom, if at all, provided. No metadata. No scientific reproducibility. And while we have survived relatively well without providing such metadata in our editions, the current situation requires change. The introduction of machine learning and its 'black box' of prediction requires adjustments in the methodology of constructing an edition. Even more so, the process of virtually unwrapping and extracting text from cultural heritage artefacts that cannot be opened – due to their fragile state – requires a general re-assessment of how that extracted text should be edited. After all, one cannot verify the text with the human eye. This new kind of text, which I will refer to as born-

virtual text, will only exist virtually and is the product of an artificial intelligence; although born-digital might be the expected term, virtual seems more nuanced for a digital text that is not the product of direct observation.

The purpose of this paper is both to start the discussion about the editing of born-virtual text and to put forth some possible ways of presenting such text in our editions. First things first, virtual unwrapping is real. A carbonized parchment scroll from En-Gedi was virtually unwrapped by the Digital Restoration Initiative, which has now evolved into EduceLab, at the University of Kentucky in 2016, revealing an early copy of Leviticus (Seales et al. 2016). Moreover, the technique is no longer considered a unique methodology, or a concept that still must be developed. It is being applied by many research groups.[2] Second, although iron gall ink so far tends to be fairly visible in micro-CT scans, as in the case of En-Gedi, carbon ink is not. Enter machine and/or deep learning and the prediction of the presence of ink in tomography. To even see the text, the human eye requires the aid of artificial intelligence; for the purposes of this paper, we will use artificial intelligence (AI) as a generic term inclusive of both machine and deep learning methods. Humanities scholars must now embrace further a concept that their colleagues in the sciences have been aware of for ages: scientific reproducibility. To interrogate a scholar's reconstruction of the text, one must be able at any time to reproduce the initial findings or, at the very least, be aware of what produced the output. To do so, means not simply knowing where to access the data, but, more importantly, being aware of key aspects of metadata associated with the output of both the AI and the algorithmic process involved. What was responsible for detecting and enhancing the ink? Where is it located within the physical object that cannot be opened? Accordingly, in current print and digital edition models we will need supplementary conventions to account for this metadata. A reader would thus have the essential data that is a traceback to the 'what', 'where', and 'how' regarding the born-virtual text before them. That said, extracting text in 3D space – from voxels rather than pixels – should also make us consider augmenting existing digital edition models. For example, we will likely need to move beyond the level of 'behind the scenes' metadata markup, JSON or XML files stored somewhere on a server or downloadable via a website, and one image as the 'canonical' representation of the object. To fully grasp the data which we are looking at – and subsequently making scholarly arguments based upon – one needs the full context of this virtual birth, i.e. structured data. We will need access to multiple image datasets and visualizations that facilitate the

---

2 E.g. Ziesche et al. 2020; Stromer et al. 2019; Liu et al. 2018; Bukreeva et al. 2016.

comprehension of the digital provenance. Only then can we achieve transparency. To explore these ideas and to put forth some possible methods, I will now offer a few hypothetical scenarios based on current research at EduceLab on the virtual unwrapping of carbonized papyri from Herculaneum and the detection of carbon ink therein.

2 Herculaneum Papyri and AI

The problem of seeing carbon ink in tomography is well documented, especially in the context of the carbonized papyri from Herculaneum (Parker et al. 2019). The ink and the papyrus substrate have different densities; the chemistry is different. And so, one will often hear how ink is, or should be, brighter in micro-CT, i.e. the density of the ink should attenuate x-rays more than the density of the papyrus substrate. Great for iron gall ink, as it is generally visible to the human eye in tomography. But the density of carbon ink just seems to resist being 'bright' enough to appear. At one point the idea that carbon ink is actually invisible in tomography even emerged (Gibson et al. 2018). That idea, however, has been proven to be inaccurate (Parker et al. 2019). Still, the problem persists. How does one make the carbon ink from an actual Herculaneum papyrus appear in tomography? Well, this has also been done using AI. In 2019 Brent Seales' presentation at the Getty Museum included a video showing how a Herculaneum fragment was used in an AI experiment to accurately reveal a Greek character in a micro-CT scan. Using a 3D Convolutional Neural Network (3DCNN), our AI was trained on one half of the visible layer of P.Herc.Paris Objet 59, while the other side was reserved for evaluation and prediction. To the human eye, a carbon ink Greek omega was made visible. It is not just the fact that virtual unwrapping is real. The visualization of carbon ink in x-ray scans is also becoming a reality, though much work remains to be done. And even though iron gall ink is generally visible, we are also applying AI to further enhance its signal in micro-CT for greater legibility (Gessel et al. 2021). AI is indeed poised to become a persistent entity or assistant in reading damaged manuscripts **[figs 1a-b]**.
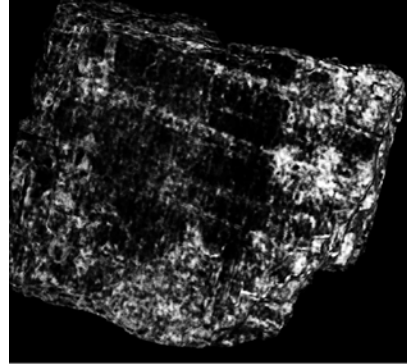
The initial AI method created at EduceLab has been published (Parker et al. 2019). Without diving too deep into the science, to be repetitive, a summary of the approach is warranted here, especially to convey the process to the general papyrological and digital humanities audiences. Morphology is the key term. Now, although this is not the kind of morphology of which papyrologists might immediately think – inflection/conjugation of verbs, nouns etc. – there is a fundamental similarity: change in structure, albeit at the micro-level this constitutes papyrus fibers *vs* papyrus fibers with ink. If the density of carbon ink will persist in not attenuating x-rays to be bright enough for the human eye, then perhaps the morphological pattern of ink on papyrus substrate is, or should be, a feature detectable and thus learnable by the machine. After all, there is a physical change, and thus a difference in structure between papyrus with no ink and papyrus with ink; this is rather visible using an electron microscope (Parker et al. 2019, 5). Thus far, this has been the basic logic upon which we continue to refine and train our AI. It thinks in terms of ink and no ink features, not alphabets nor languages. Based on what it has learned, it predicts the presence of ink and amplifies its signal to be visible to the human eye **[figs 2a-b]**.
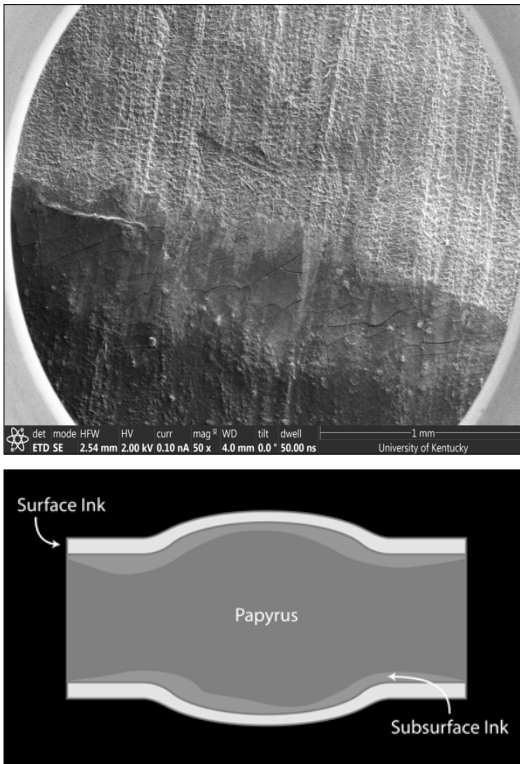
Figure 2a Morphology. A: papyrus with ink *vs* no ink is clear with an electronic microscope

Figure 2b Morphology. B: the feature, the signal, we are detecting is the ink both penetrating and resting on top of a papyrus fiber

The process starts with Volume Cartographer (VC), a custom software application developed at EduceLab for virtual unwrapping. Raw micro-CT data (sinograms) undergo reconstruction and that data is then rendered into a volume package that can be passed through VC. Put very simply, the VC pipeline allows for the efficient segmentation of volumetric image data (the slicing of the volume to isolate writing surfaces) and the subsequent texturing, flattening, and generating of 3D and even 2D images of those segments. Now, it is the texturing process that is critical for our AI. As the 3D mesh of a given segment is textured (the process of applying the visual details to a 3D model – the point clouds that represent the structure – to give it definition in terms of surface shape), a per-pixel map that stores all 3D positions is generated. For any segment, areas or points from this map are then selected and used to create sub-volumes that constitute the input for our 3D CNN. These sub-volumes, oriented toward the surface of the writing substrate, is where prediction will occur. This is where the so-called 'black box' of AI exists, the point at which something is purportedly seen or predicted based on prior training.

Understanding how learning takes place determines how transparent this black box will be.

The greatest challenge in applying AI to visualize carbon ink in Herculaneum papyri is a lack of training data. The most effective AI is the one with extensive reference libraries. The more data to reference and from which to learn, the greater confidence in its ability to predict. To prove the concept, we used a carbon phantom (a fabricated facsimile) scanned at 12 microns. Training labels were made by aligning and registering images containing the ground truth of ink/no ink to the x-ray images in which it is not visible; with both x-ray and conventional images, we thus know where the ink is, even if we cannot see it in the former. Multiple sub-volumes, each with their own label, were then used to successfully train a neural network. For actual fragments, creating training labels is essentially the same process. For evaluation, however, we have used a form of k-fold cross validation in early experiments to validate the concept, especially since we have limited training data. The writing surface is partitioned spatially into k-regions of interest. These regions are used for training, with one reserved for evaluation, i.e. one region is the input upon which the network applies what it has learned and predicts ink/no ink. Training runs on P.Herc.Paris Objet 59 demonstrate that this morphological approach is working. Nevertheless, one caveat must be pointed out: resolution. To detect this ink signal, that morphological pattern of ink covering and penetrating the substrate surface, a high resolution is required. Our current projection so far is that a resolution of 3-5 microns is needed; yet this could change over time as we learn more about the ink signal that we are detecting. The end results are not only images documenting the blank x-ray scan and the prediction of ink (thus a character), but also a photo-realistic rendition that offers a virtual facsimile of the manuscript as it would appear under natural light conditions to the human eye **[figs 3a-b]**.
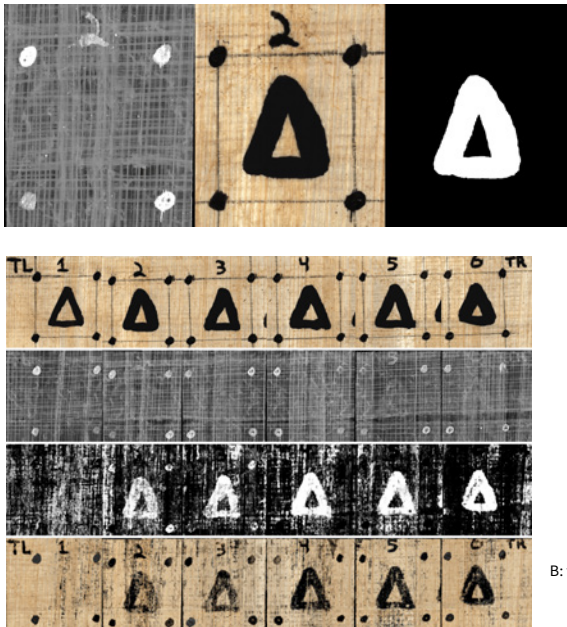
Figure 3a   Carbon phantom.
A: from left to right: the x-ray scan,
the natural light view, the training label

Figure 3b   Carbon phantom.
B: from top to bottom: natural light, x-ray,
prediction (i.e. reconstruction),
the photo-realistic rendition

These ongoing experiments raise issues rarely, if at all, discussed. How do we edit this text? In recent scanning of Herculaneum papyri using X-ray phase-contrast tomography (XPCT), attempts have been made. In Mocella et al. (2015) cropped images of XPCT data (P.Herc. Paris. 1 fr. 101 and P.Herc.Paris. 4) were paired with basic transcriptions of the Greek text purportedly seen, as well as individually cropped images constituting an entire Greek alphabet (2015). In Bukreeva et al. (2016; 2017) we find the most extensive attempt to pair cropped images of XPCT data (P.Herc. 375 and 495) with both diplomatic and articulated transcriptions; the use of the typical editorial conventions in these transcriptions, such as the underdot and square brackets, indicate the application of papyrological method. Now, I have no interest here in debating the reliability of the ink allegedly seen in these publications. There are lingering issues, especially regarding resolution, and the reality that it may not be ink at all persists. Rather, I am interested in data that is missing in the presentation of this text. Let us hypothetically say that all the text published is, in fact, indicative of carbon ink. First, from where does this text come? For example, P.Herc.Paris 1 fr. 101 is actually a multi-layer fragment removed from an intact scroll. Where was its original location? Second, two lines are revealed from the hidden layer (Mocella et al. 2015). Which layer? Moreover, which line is first in

succession? There is no indication. In Bukreeva et al. (2017) an image of a layer virtually removed from P.Herc. 375 is provided. However, only a cropped, magnified image is later provided with annotation indicating possible lines of Greek text. As for the text edited and presented according to papyrological method, exiguous as it may be, it is actually from P.Herc. 495. Again, where is this text coming from? Finally, in Bukreeva et al. (2016) we find more cropped, close-up images from P.Herc. 375 and 495, albeit with better papyrological transcriptions. Whether P.Herc. 375 or 495, where is the text located in relation to the overall structure of the intact scrolls? The only indication is that the text comes from the inner part. Obviously, these are first attempts in the process to reveal the hidden ink. Be that as it may, for confidence and trust, we must be more precise.

With virtual unwrapping and AI prediction and enhancement, we cannot just pretend that we are looking at the usual 2D image, or even the actual fragment, and whatever text is or is not visible under natural light conditions. No, we cannot verify the text with our eyes at all. This is a moment in which metadata associated with the virtual unwrapping process and AI prediction becomes important. In the examples of published text mentioned above, virtual unwrapping and/or segmentation metadata is ignored. And while no AI was used in those experiments, in our work at EduceLab we plan on incorporating metadata in a JSON file during any application of AI, which will notably include a Git Hash that references the specific code used and thus responsible for ink prediction; this is a part of the on-going development of our AI work. This metadata is critical for scientific reproducibility. Normally this is just metadata stored on a server somewhere and (hopefully) accessible in some way. Yet due to the increasing role of virtual unwrapping in digital restoration and the on-going developments in the use of AI to virtually enhance text, some of this data should be moved into the workflow of the humanities scholars who will edit this born-virtual text. We are looking at a near future in which both multiple versions of AI (multiple versions of code) and multiple scans might be used to predict and to enhance text from one cultural heritage object over time. Furthermore, the segmentation process in virtual unwrapping must be tracked for understanding the location of the text and the virtual reconstruction of the physical object.

So, AI, segments, and sub-volumes. How does this affect the editing process? Intact scrolls from Herculaneum offer a good sandbox in which to approach that question. For any intact scroll, the degree of damage varies over the internal structure, potentially resulting in random rather than consistent areas that are initially strong candidates for virtual unwrapping. The odds are highly unlikely that we can start at the beginning of the work and slowly unroll to reveal the text. Moreover, while some characters may be successfully amplified and made visible on the first try, ambiguity due to damage and, of course, the

random noise of ink smudges, ink drops, crossed out letters, severely faded ink, tears, holes etc. will likely persist and drive further improvement of the AI over time. The more noise, the more training and improvement are needed. The AI will thus require an ever-evolving reference library as more training data is gathered to reduce noise that appears as we start to see accurately within in a Herculaneum scroll. Nevertheless, when our AI begins to detect, predict, and enhance ink, the scholarly community will want the publication of that text to start immediately, just as we saw above with the published XPCT data. Waiting until an entire scroll is virtually unrolled is not a welcome strategy at this point. Unfortunately, there is currently no indication of how long the full process of segmentation to completion, i.e. an entire intact scroll, will take; this is a massive optimization problem **[figs 4ab]**.

Figure 4a P.Herc.Paris 3. A: vertical Y axis
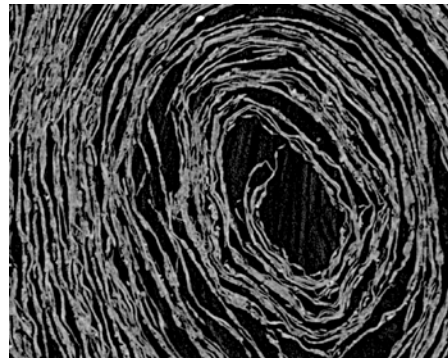

Figure 4b P.Herc.Paris 3. B: circumference at 7.91 microns

Based on the current state of virtual unwrapping and ink enhancement using AI, let us now explore a few theoretical scenarios in which we first present born-virtual text from an intact Herculaneum scroll in our edition models. A papyrologist that sits down to produce the *editio princeps* will need to document, at the very least, three attributes along with the expected metadata (e.g. publication/inventory number, measurements, date etc.). We need a segment ID, a volume ID, and even perhaps an AI ID **[fig. 5]**.
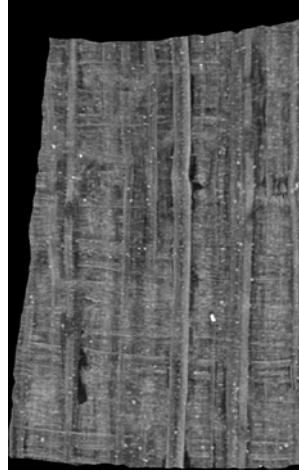
Figure 5 Segment of P.Herc.Paris 3

In figure 5 we see a segment of P.Herc.Paris 3 scanned at 7.91 microns. This data stems from scanning sessions at Diamond Light Facility, UK, in 2019. If we recall the general description of our AI process, multiple points across this segment will be selected to create sub-volumes that will be the input for the 3D CNN. Within these sub-volumes lies the ink signal that will be subsequently enhanced for both visibility and legibility. Moreover, these sub-volumes are so small (approximately 90 um × 90 um) that multiple sub-volumes are used in the reconstruction of just one Greek character. As we begin to see clearly successive lines of text based on multiple sub-volumes, perhaps it is best to follow the standards implemented by Obbink (1996, 99-103) and Janko (2000, 194-200) as we begin to edit that text. This method is characterized by the utilization of facing pages: 1) the diplomatic/articulated text, according to column structure, and a critical/testimonial apparatus on the left; 2) a modern layout of the text with a translation and notes on the right. Focusing on just the left facing portion, if we could now see text in this segment of P.Herc.Paris 3, we could possibly add our three attributes as follows:

P.Herc.Paris 3 V# SG#   Col.# AI#
                             Greek Text

Testimonia

Apparatus
(palaeographical/critical)

V# is the identifier for the volume scan which contains the ink signal. SG# is the identifier for the segment produced during virtual unwrapping. Lastly, the AI# is the unique identifier of the AI model

or code used in prediction. Now, it is important to understand that this is just a theoretical approach at the moment. As work progresses, we might find ways to be more efficient and reduce the number of identifiers required in an edition. The AI#, for example, is a part of the metadata associated with the images (3D or 2D) produced. The exact identification of the AI model could just remain there. However, giving credit to the AI, or at least contemplating how we should do this at this very early stage, is worth considering. Overall, it is a mapping between text, segment, scan, AI, and the physical object that should be borne in mind **[fig. 6]**.
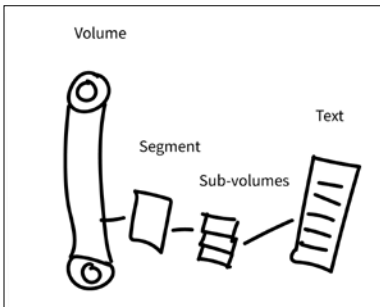

Figure 6 Mapping the volume, segment, and sub-volumes to the text

The general logic of the mapping in figure 6 should seem straightforward. However, it could (or will) get more complicated. In the above example we have a V# from P.Herc.Paris 3, a SG#, and text reconstructed from the sub-volumes by an AI#. The SG# from P.Herc.Paris 3 is a partial layer that is near the core of the intact scroll. The surface area is approximately 1 cm × 0.6 cm, and the scroll itself is approximately 6 cm in diameter and 18 cm in height. If we think about the dimensions of the papyrus sheets glued together to make the scroll, particularly the height, the segment is approximately 6 cm in height and thus a small portion of the scroll's, or a given papyrus sheet's, height. Clearly, we do not know the exact column height nor size of the upper and lower margins. Nevertheless, it is accurate to say that we only have a small portion of the text from one sheet. If for one column of Greek text we have multiple segments, as a hypothetical example, in the reconstruction of the born-virtual text, we will need to annotate accordingly. We would have text that points to more than one segment in our edited reconstruction.

P.Herc.Paris3 V#	Col.#, AI#
	(SG#) Text (lines 1-10)
Testimonia	
Apparatus	
(palaeographical/critical)	(SG#) Text (lines 11-20)

In order to know the correspondence between text and segments, marginal annotation can be employed; this kind of annotation is not unfamiliar in critical editions of works with a complex reconstruction based on both a mediaeval manuscript tradition and other witnesses (e.g. texts that provide quotations of text missing in the manuscript tradition). In the above example, we see that two different segments, containing twenty lines in sum, reconstruct one column. Now, while that is not too complicated, what happens when we have multiple scans, multiple segments unique to those scans, and multiple versions of AI used in the reconstruction of one column over time? Scans at different resolutions and further training to improve and to change the AI are also very likely to occur. Marginal annotation might indeed be necessary to convey how the text is reconstructed from these critical elements, e.g. (V#, SG#, AI#).

Too many segments? Too many volumes? Too many AIs? Exactly how many segments and volumes? Will every drop of ink appear perfectly clear? We do not know yet. Obviously, keeping the number of segments and volumes to a minimum would be ideal. Moreover, we want our AI to completely reveal every drop of ink with ease. Yet in the context of segmentation to completion some uncertainty remains. However, while a large number of volumes, segments, and AI versions seems cumbersome, this might not be a bad thing. That would indicate the possible existence of areas of persistent ambiguity. These areas might constitute points in the scroll where internal damage does not permit a clear virtual reconstruction by the AI, whether that is because noise persists (further training is required), or the ink signal itself has been irreparably damaged in some fashion. Even with AI, we could still have the ink traces with which all papyrologists are very familiar. Humanities scholars would thus continue to apply their skillsets to conjecture and to debate the reconstruction of a born-virtual text. Virtually extracting text embedded in cultural heritage artefacts is indeed exciting, especially in the case of Herculaneum papyri. Yet papyrologists might see the possibility of being replaced by an AI that is essentially recognizing and reconstructing Greek characters, even though it does not think in terms of the Greek language

or alphabet – yet.[3] Be that as it may, I envision a process where the human papyrologist is still very much in the loop. In terribly damaged cultural heritage objects, the human and the AI will work together to elucidate the text.

Now, one might still ask: why do we need to keep track of this metadata, let alone include it in our editions? We need to keep in mind that a sequence of segment IDs will not indicate the logical order of the work. If we virtually unrolled a well-known work, like a copy of Homer's *Iliad*, we would know our exact location based on the text itself. But for unknown works and works only known by title or random quotation this creates a slight problem in visualization. Without any visible data, such as stichometric counting (line counting) or a numbering of columns, the segment IDs are basically 'puzzle pieces' that we need to move around to reconstruct the proper order of the work as it is slowly revealed. One way to mitigate this issue is to expand or 'grow' a segment over time to extract large areas of continuous text. Still, this would not change the fact that we are likely to 'grow' multiple segments from different areas within a scroll. Note also how this even makes assigning columns an alphabetic or numerical sequence problematic. In the examples above, the Col.# is unique to a segment ID. Whether we call it Col. I or Col. A, that ordinal sequence pertains only to that segment ID, not to an alphabetic or numerical sequence of columns from the start to the end of a work. It is perhaps ironic that we are, in a way, creating virtual fragments (the segments) of a physically intact scroll in order to get to the text; invasively or non-invasively, we cannot seem to stop fragmenting Herculaneum papyri. At any rate, visually keeping track of the location of every segment within the physical object is critical [fig. 7].

---

3 Along with colleagues at Middle Tennessee State University, the University of Tennessee at Knoxville, and the University of Minnesota Twin Cities, we have successfully trained a few machine learning models to classify Greek characters in images of papyrus fragments. This was uniquely done using crowdsourced transcription data from the Ancient Lives project (Williams et. al. 2014) as training data. Results to be published soon.
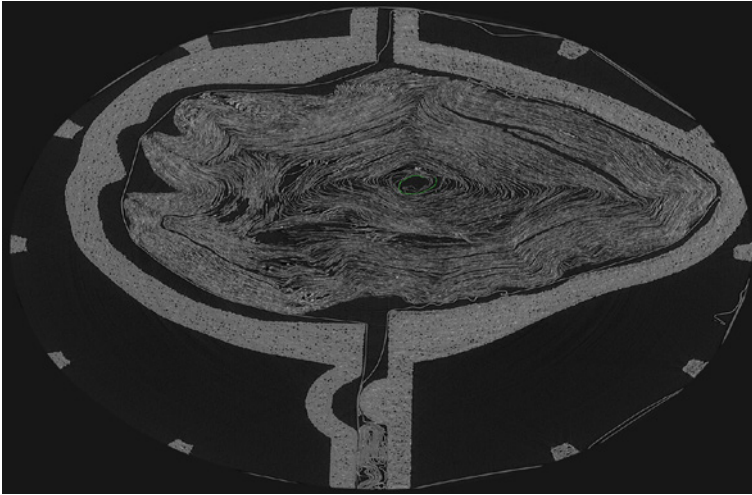
Although we have focused on the virtual extraction of text from intact scrolls thus far, the concepts discussed also apply to the opened Herculaneum fragments. P.Herc.Paris 1 fr. 101, which was a part of the study of Mocella et al. (2015), is a multi-layered fragment physically removed from an intact scroll. With the text purportedly seen in Mocella et al. (2015), again, we have two questions. Which layer? Which line of text comes first? Just as in our hypothetical examples above, we can be more precise. Let us take P.Herc.Paris Objet 59 **[figs 1a-b]**, an important subject in Seales' 2019 talk at the Getty, as another hypothetical example. This small fragment has a few layers with clear ink on the top and even some visible ink on the second layer.

Let us hypothetically peel that first layer off to reveal the second layer. Now, we can still use identifiers for the volume, segment (or in this case layer), and the AI. On the top layer, just below the ε, we can see what might be a ν in the second layer. Although two columns are distinct on the top layer, we cannot assume another intercolumnium in the second. So, in the example below, we simply present areas likely containing text and the possible ν, along with the volume ID, segment ID (or in this case we could call it a layer ID) and AI ID **[fig. 8]**.
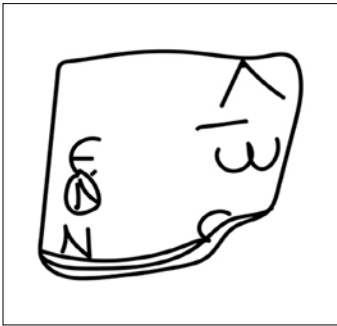
P.Herc.Paris Objet 59
Layer 1   V# SG# (Layer 2) AI#

Col. #	Col. #	
]	__λ	Greek Text
]. ε	ω	ν
] ν	.	Greek Text

Even for Herculaneum fragments that are the result of human, physical intervention, nothing changes. These fragments too can benefit from virtually unwrapping and AI ink prediction and enhancement. Significant logistic challenges, nevertheless, remain for these fragments. For all those stored in *cornici* in Naples (the trays in which they are preserved), we cannot bring the synchrotron to them to acquire the desired resolution. And even if we could bring them to the synchrotron, is scanning even possible or safe due to the state of their conservation (in the *cornici*)? While that logistic issue remains a problem, we still have fragments like P.Herc.Paris Objet 59 and P.Herc.Paris 1 fr. 101 that can benefit from virtual unwrapping and AI ink prediction. In editing the born-virtual text from the hidden layers of opened Herculaneum fragments, we still need to account for the volumes, segments, and the AI involved.

Before moving to the final section of this paper, we should also briefly address the use of born-virtual text in the critical editions of Greek and Latin works, such as those found in the *Teubner* and *Oxford Classical Texts* series. Even now papyri published in *The Oxyrhynchus Papyri* series continue to confirm or reject emendations in critical editions, as well as offer new readings that are eventually printed in either the text itself or the critical apparatus. In collating manuscripts, the standard practice is to assign a papyrus a value in the *sigla*, which will then represent the papyrus in the critical apparatus. In Diggle's *OCT* edition of Euripides' *Medea*, for example, we find the following:

Π²   P.Oxy. 1370 fr. 1: uu. 20-6, 57-63   v p.C.

We see the expected publication series, publication number, location data within the papyrus (fr. 1), location data within the work (verse numbers), and the date of the papyrus (5th century CE). Do we need to augment this? For specificity and clarity, yes. If hypothetically our segment of P.Herc.Paris 3 contained quotations of Euripides' *Medea* with variant or new readings, we should, at the very least, see the following:

Π#   P.Herc.Paris 3 V#, SG#, verse numbers, papyrus date

To have a simple and clear traceback, the volume and the segment IDs are required. Remember, we do not know how long the process of segmentation to completion will take. Accordingly, for one intact scroll, we could see a progression of their segments published over time; perhaps even their volumes too, if the object is scanned multiple times. And for the specific text incorporated into the critical apparatus or into the text of the critical edition itself, that volume and segment ID constitute precise location data for the born-virtual text. For papyrologists and philologists, this issue is not unlike the re-assigning and re-ordering of fragments in different editions over time, in which a system of mapping illustrates the change in fragment identifiers/publication numbers.

Now, if the above example satisfies the necessary requirements for the *sigla* of a critical edition, what about the critical apparatus? The current standard is to place the Greek lemma (text) followed by Π#, so that the reader knows that the text stems from a papyrus. But does that constitute transparency? After all, whether a fragment with multiple layers or an intact scroll, we cannot see the text in the physical object, nor can we see it in the x-ray image. The reading is there because of the AI. One could perhaps argue that the AI ID should also be included, since an artificial intelligence is responsible for the text. Perhaps we should give it credit, e.g. Greek lemma Π# AI#. Furthermore, in the case an area of persistent ambiguity, if the AI only reveals ink traces, an editor will reconstruct according to established practices. We could, therefore, even find in a critical apparatus a listing of Greek lemma Π# AI# editor's name. Perhaps. Yet I will assume the continued use of a lemma followed only by Π# will suffice for now.

3 Challenges in Visualizing and Working with Born-Virtual Text

So far, we have reviewed the process of virtual unwrapping, AI ink prediction and enhancement, and how the resulting born-virtual text might be presented in editions that conform to the methodologies of Papyrology and Classics. The 2D space of print publication has been the tacit focus. What about digital editions? A careful reader will have noticed one issue percolating in the discussion above: the amount of image data inextricably tied to this born-virtual text. Furthermore, while we have suggested simple ways to introduce essential metadata into the critical editing workflow, there is so much more metadata associated with the generation of born-virtual text. The digital edition model might seem better suited in that context. But does this new kind of text deserve its own unique environment for editing and publication?

Digital papyrology has been around for some time now, and its history and current trajectory has been well documented by Reggiani (2017; 2018). For our purposes here, we will get straight to the point. Any text extracted from an intact scroll or from the hidden layers of Herculaneum fragments can be presented in a digital edition. The fundamental model is EpiDoc (TEI/XML) and the most critical resource is Papyri.info, which implements the EpiDoc standard for documentary papyri and allows for a robust search of the Greek text. For literary papyri, Papyri.info's recent Digital Corpus of Literary Papyri (DCLP) is now advancing digital editions for the kinds of text associated with Herculaneum (e.g. not documentary). Editions of Herculaneum fragments, in fact, already appear in the DCLP. And these texts and editions are indicative of the scholarly work that is dependent on a combination of the autopsy of the original fragment, the Oxford/Neapolitan Disegni (hand drawn facsimiles made mostly at the time of unrolling, or in some cases later), multispectral imaging conducted by Brigham Young University, and the user interface of Papyri.info, which allows for the creation of a digital edition for its platform. For bibliography and images, links are provided to Chartes (chartes.it), an online catalogue of Herculaneum papyri. Without a doubt, the virtually extracted text from Herculaneum papyri will appear in the expected Epidoc standard, or a modified version of it. However, as we have seen above, simply showing the born-virtual text within the parameters of standard papyrological conventions is not enough. Whether in print or in digital form, there are further data outputs and algorithmic metadata that should, if not must, accompany the text.

Let us start with the image data. At the end of the virtual unwrapping and ink prediction and enhancement process, we have multiple images. In 2D and 3D there are images of the whole object (intact scroll or fragment) and of the segments from which the sub-volumes are extracted. For the segments, the 2D images document the x-ray

scan (no visible ink), the ink prediction (visible characters), and the photo-realistic reconstruction (what the text would look like under natural light conditions). As noted in the second section, in the publication of any segment/s, a visualization documenting the location within the object is required for the 'bigger picture' view. So, for every line of text revealed, a reader of the edited text should be able to access the x-ray, the prediction, and the photo-realistic rendition images for that area of papyrus, as well as a visualization of its location within the scroll/fragment. Again, here is the refrain. We do not yet know how long segmentation to completion of an intact Herculaneum scroll will take. In addition, although keeping the number of volumes and segments to a minimum would be ideal, uncertainty also remains in that context. Building an edition of the born-virtual text from an entire, intact scroll will both take time and include an ongoing increase in associated image data [fig. 9].
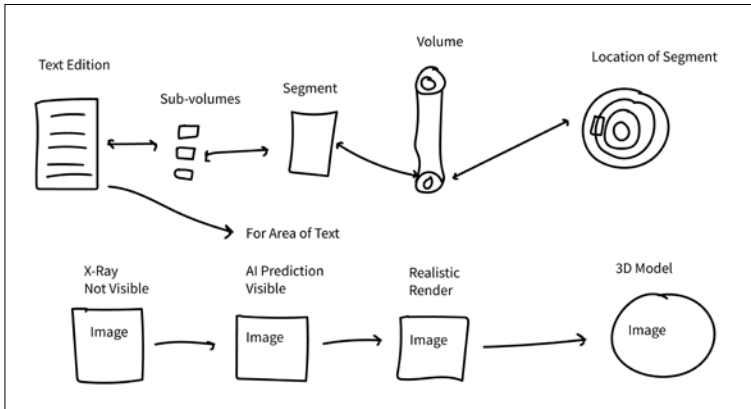
Next, digital provenance is the only way to comprehend the full process involved in generating born-virtual text. For virtual unwrapping, this is a process of segmentation, texturing, flattening, and finally merging and visualization (2D and 3D images); as already noted, within this process, ink prediction and enhancement take place in the texturing phase. In Chapman, Parker, Parsons, Seales (2021), EduceLab presents its plan to use a METS container[4] to systematically document the digital provenance of any digital surrogates or digital twins generated in our lab; an important point to remember is that, for damaged cultural heritage objects, our digital versions are surrogates to

---

[4]   Metadata Encoding Transmission Standard, http://www.loc.gov/standards/mets/.

be used in lieu of an object that can no longer be physically handled (with the exception of conservation). In the first instance, the 3D compilation of P.Herc. 118, housed in the Bodleian Library of the University of Oxford, is presented as a test case (Chapman et al. 2021). The 3D compilations of all 12 *pezzi* (fragments) are created using 3D photogrammetry, digitized versions of analog photos, multispectral and hyperspectral images, and digitized versions of the Disegni. Through complex processes such as segmentation, image stitching, and image registration, the resulting 3D model is an unprecedented digital entity in which multiple, and formally separate, datasets are now accessible in one 'place'. Chapman et al. demonstrate how a METS container, which can incorporate multiple schemata like Dublin Core and MIX to track administrative, technical, and descriptive metadata, offers an efficient means to document an entire digital provenance chain, from image acquisition, whether starting with x-ray scanning or photogrammetry, to the final 3D and 2D images that can include AI predicted and enhanced text. Essentially an XML wrapper, METS provides a familiar and easy to use human and machine-readable format. More importantly, EduceLab intends to repurpose the behavior section (behaviorSec) of a METS container to document, to describe, and ultimately to visualize complex algorithmic processes (7-10). Stop for a moment and think about viewing a 3D model of an opened Herculaneum fragment, like P.Herc. 118, to which archival analog photos, multispectral images, and hyperspectral images have been registered, so that you can easily view any of those images at will in one 3D space. But at the same time, there is also a way for you not only to visualize the image registration pipeline used, and thus ensure comprehension of all the fixed and moving parts (the 3D mesh and the unique 2D images registered to it), but also to recreate that digital object. Utilizing the METS behavior section, Chapman et al. lay out how EduceLab envisions implementing graph visualizations and hands-on scientific reproducibility to reveal how algorithmic processes like registration, segmentation, and stitching work 'under the hood'. As an initial example, Seth Parker has released the Structured Metadata Engine and Graph Objects Library,[5] a C++ library for visualizing dataflow pipelines. For one digital surrogate, whether it is a digital object in 3D and/or 2D created through the process of photogrammetry, scanning sessions at a synchrotron facility, or any available x-ray imaging equipment, the combination of a canonical METS document plus the visualizations of dataflow pipelines represents not only an unprecedented account of the metadata involved, but also a significant advance in accessibility and transparency in digital provenance.

---

5   SMEAGOL, https://gitlab.com/educelab/smeagol; https://zenodo.org/record/4298710#.YADk9MVKimE.

To ensure trust in the generation of born-virtual text, much more is needed than the typical editorial conventions applied in both current print and digital formats. Put simply, the text itself is not enough. To present it alone is nearly useless, in fact, since it is text that cannot be verified by the human eye. In the case of Herculaneum, presentation of the text with the usual papyrological editorial conventions and new conventions (like those suggested in the second section above) must be inextricably linked to the variety of image and metadata generated in the process of revealing that text. Ideally, we should be able to see and to work with these relationships in one place. And thus, one might see how the 2D space of print publication might not be the best medium. A digital edition format, which allows 3D functionality, is far better suited to see the relationships between edited text, the 3D and 2D image data revealing its location in the object and the reality of ink prediction, and the data visualizations representing the dataflow pipelines involved. This is scientific reproducibility and transparency at the level of working with the published text and at the level of accessing the image data and metadata itself – preferably in live time.

Is there a current application, web or desktop based, to facilitate these kinds of user interactions between born-virtual text, the image data, and both the metadata and the graph visualizations of it, let alone a digital environment for creating a digital edition with that data and the additional editorial conventions required? No. In his account of digital papyrology, Reggiani goes through many in great detail, and he highlights a very important agenda in the creation of digital edition platforms for papyri and similarly fragmented texts: integrating the existing print/bibliographical resources (2017, 146-59). In Vindolanda Tablets Online, Codex Sinaiticus, Derveni Papyrus Online, and, but certainly not least, Papyri.info, we see digital texts accompanied by digital translations, essential descriptive metadata of the fragment, bibliographical sources, addenda and corrigenda, and reference sections and/or commentary. In the context of visual integration between digital text and image, in Anagnosis and Codex Sinaiticus, in particular, visual alignment between a word in the digital text and the location of its sequence of characters in the image of the manuscript is provided (151-9; Reggiani 2018, 63-74). Devoted entirely to Herculaneum papyri, the Anagnosis UI can facilitate the ongoing process of editing the opened carbonized scrolls, as it allows one immediately to anchor a transcription to the image of the carbonized surface, in which contrast between black carbon ink and black papyrus substrate is often wanting. Indeed, these platforms have been a great success and are indispensable to the academic community. But now we are facing something new. We are entering a new phase in which text embedded or hidden in cultural heritage objects can be extracted through complex algorithmic processes and

artificial intelligence. Furthermore, we have also noted that the AIs used to predict and enhance ink will also likely be used to enhance the legibility of visible but damaged ink in the typical papyrus fragments we find in collections around the world. Where do we go to work with this data? How do we work with this data?

To even begin to visualize these connections, one would need multiple and disconnected applications, such as Papyri.info, Adobe Photoshop, Mac Preview, and Meshlab for 3D object files. Yet even with them, none of these applications were designed to facilitate these connections in a manageable space designed with the workflows of editors and researchers in mind. Now, while the creation of new tools and platforms can often receive pushback – why not concentrate efforts on improving existing tools? – the process of virtual unwrapping and the subsequent production of born-virtual text is a massive step forward in the digital restoration of damaged cultural heritage objects. And even though we often omit cultural heritage objects that are not damaged in this discussion, these methods, for example, can be used for the digital preservation of centuries old codices that just should not be opened anymore; we are literally breaking their spines. With this type of leap forward, a combination of improving existing tools and building new ones, which can be integrated, is required.

What is the answer? What is the path forward? Without providing one definitive answer at this time, let us say that we fundamentally know how to approach the question. There are a few existing tools that point in the right direction. For visualizing, annotating, and interacting with 3D models, IIIF, the 3D Heritage Online Presenter (3DHOP), and the Smithsonian Voyager are notable. 3DHOP allows one to create interactive 3D models, at high resolutions, that can be embedded in a standard web page.[6] The Smithsonian Voyager[7] is a unique authoring tool that allows users to create visual presentations or 'stories' using 3D models. One can position them freely, export 2D versions, inspect the 3D mesh, and annotate the 3D object with 'articles' or content to add critical context to the visualization. For the cultural heritage community, these tools are very effective in creating content or virtual exhibitions for both academic and general audiences. The standards set forth by IIIF[8] allow one to set up a data server that accepts IIIF API calls and to utilize predefined viewers that let you access data from multiple, disparate servers. For the editing and presentation of the born-virtual text, in the context of Greek and Latin Papyrology, Papyri.info and especially its DCLP are foundational tools with proven methodologies for presenting digi-

---

[6] https://www.3dhop.net/index.php.

[7] https://smithsonian.github.io/dpo-voyager/introduction/.

[8] https://iiif.io/.

tal editions. However, the requirements for producing born-digital, critical editions of literary and sub-literary papyri are still wanting. From the ability to visualize the variety of marginalia (notably symbols) present in literary and sub-literary papyri and the more custom methods of annotation, such as the combining asterisk and bold font unique to Herculaneum papyri editions (to mark editorial correction of the Disegni and the placement of sovrapposti/sottoposti respectively), to the visualization needs of the critical, testimonial, and palaeographical apparatuses, much more development is needed. Proteus is a project designed to fill in these gaps (Williams et al. 2015). The project began at the University of Oxford in 2015 and successfully built a stable editor for creating born-digital, critical editions that included all the desired attributes: diplomatic and articulated text, palaeographical, critical, and testimonial apparatuses, a translation, and critical notes. Moreover, Proteus' editor allows a user to create these components without any hardcoding of the XML. As one types, the XML and HTML are generated in live time. And, of course, this editor is a part of a larger web application for presenting these editions online in a similar fashion to Papyri.info. Unfortunately, due to funding the project's development timeline has slowed down considerably. Be that as it may, as of January 2021 Proteus has been upgraded from Python 2.7 to 3 and is now undergoing preparation for a small beta test. In the end, we can thus see the components we need across existing applications.

In thinking about the virtual unwrapping of Herculaneum papyri and the virtual extraction of the text hidden inside, the ideal tool for editing and interacting with the born-virtual text and the amount of image data and metadata that accompanies its creation is likely to be combination of what we see in 3DHOP, Smithsonian Voyager, Papyri.info, and Proteus. The papyrologist needs an editor that allows them to create a critical edition that meets the requirements of literary and sub-literary papyri. Yet they also need, within that same editorial interface, the ability to pair annotated (especially for text location purposes) 2D and 3D image data, AI ink prediction metadata, and digital provenance metadata with the edited born-virtual text itself. And for the readers of published born-virtual text, they too will need a UI that allows them to interact with and understand in a meaningful way the relationship between the text, image data, and metadata. Without this functionality, at both the level of editing and publishing editions for subsequent research, we will only pretend to understand the text before us.

4   Conclusion

Scholarly editing with AI. It probably does not seem very exciting at the moment. As we have seen in this paper, the topics of concern have been metadata, images, and the editorial conventions used in Greek and Latin Papyrology and Classics. Be that as it may, at this fundamental stage in the application of AI to reveal and to enhance text in manuscripts, it is the perfect time to initiate conversations about how we edit text produced by AI before the process becomes, perhaps, even more complex. The basic conventions proposed here, for carbon ink text embedded in Herculaneum papyri, are by no means definitive solutions. They are simply ideas meant to encourage thinking and further discussion on the topic. And while digital editions seem to be more appropriate, the 2D space of print could also still be used. What is clear is that we cannot treat text extracted from Herculaneum papyri by AI and through virtual unwrapping in the same fashion as the legible, and even illegible, ink in manuscripts and papyrus fragments that preserve a natural contrast between text and substrate surface – even if spectral bands are required to see that contrast. To do so, we will only pretend to understand the nature of the text before us, and upon which we are making scholarly arguments in research. To ensure trust in the born-virtual text before us, we need to understand its virtual birth. We need to understand the data, i.e. the structured data describing and visualizing the entire process from start to finish. As for AI, identifying it and bringing it into our editions may or may not be necessary at this time. It is probably not the type of AI about which our science fiction induced imaginations think and dream. It is an intelligence that makes predictions. But we cannot talk to it. We cannot interact with it. Still, it represents the very beginning of the kinds of digital minds of which Bostrom speaks (Bostrom, Shulman 2020). We must, then, begin to think about how we represent AI, as a critical assistant, if not full partner, in our work.

Bibliography

Bostrom, N.k; Shulman, C. (2020). "Sharing the World with Digital Minds". Working Paper. https://www.nickbostrom.com/papers/digital-minds.pdf.

Bukreeva, I. et al. (2016). "Virtual Unrolling and Deciphering of Herculaneum Papyri by X-ray Phase-contrast Tomography". *Scientific reports*, 6, 27227. https://doi.org/10.1038/srep27227 PMID: 27265417.

Bukreeva, I.; Alessandrelli, M.; Formoso, V.; Ranocchia, G.; Cedola, A. (2017). "Investigating Herculaneum Papyri: An Innovative 3D Approach for the Virtual Unfolding of the Rolls". https://arxiv.org/abs/1706.09883.

Chang, L. et al. (2018). "Robust Virtual Unrolling of Historical Parchment XMT Images". *IEEE Transactions on Image Processing*, 27(4), 1914-26.

Chapman, C.Y.; Parker, C.S.; Bertelsman, A.; Gessel, K.; Hatch, H.; Seevers, K., Brusuelas, J.H.; Parsons, S.; Seales, W.B. (2021). "The Digital Compilation and Restoration of Herculaneum Fragment P.Herc.118". *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies*, 6(1), 1-32. https://doi.org/10.1353/mns.2021.0000.

Chapman, C.Y.; Parker, C.S.; Parsons, S.; Seales, W.B. (2021). "Using METS to Express Digital Provenance for Complex Digital Objects". *Metadata and Semantic Research 1355*, 143-54. https://doi.org/10.1007/978–3–030–71903–6_15.

Gessel, K.; Parsons, S.; Parker, C.S.; Seales, W.B. (forthcoming). "Towards Automating Volumetric Segmentation for Virtual Unwrapping". *Open Archaeology*.

Gibson, A.; Piquette, K.E.; Bergmann, U.; Christens-Barry, W.; Davis, G.; Endrizzi, M. et al. (2018). "An Assessment of Multimodal Imaging of Subsurface Text in Mummy Cartonnage Using Surrogate Papyrus Phantoms". *Heritage Science*, 6(1), 7. https://doi.org/10.1186/s40494–018–0175–4.

Janko, R. (2000). *Philodemus: On Piety Book 1: Edited with Introduction, Translation, and Commentary*. Oxford: Oxford University Press.

Mocella, V.; Brun, E.; Ferrero, C.; Delattre, D. (2015). "Revealing Letters in Rolled Herculaneum Papyri by X-ray Phase-contrast Imaging". *Nature communications*, 6. https://doi.org/10.1038/ncomms6895.

Obbink, D. (1996). *Philodemus: On Piety Part 1: Critical Text with Commentary*. Oxford: Oxford University Press.

Parker, C.S.; Parsons, S.; Bandy, J.; Chapman, C.; Coppens, F.; Seales, W.B. (2019). "From Invisibility to Readability: Recovering the Ink of Herculaneum". *Public Library of Science*, PLOS ONE, 14(5). https://doi.org/10.1371/journal.pone.0215775.

Reggiani, N. (2017). D*igital Papyrology*. Vol. 1, *Methods, Tools, and Trends*. Berlin; Boston: de Gruyter.

Reggiani, N. (ed.) (2018). *Digital Papyrology*. Vol. 2, *Case Studies on the Digital Edition of Ancient Greek Papyri*. Berlin; Boston: de Gruyter.

Seales, W.B.; Parker, C.S.; Segal, M.; Tov, E.; Shor, P.; Porath, Y (2016). "From Damage to Discovery Via Virtual Unwrapping: Reading the Scroll from En-Gedi". *Science Advances*, 2(9). https://doi.org/10.1126/sci-adv.1601247.

Stromer, D.; Vincent, C.; Xiaolin, H. et al. (2019). "Virtual Cleaning and Unwrapping of Non-invasively Digitized Soiled Bamboo Scrolls". *Scientific reports*, 9(1), 2311. https://doi.org/10.1038/s41598–019–39447–0.

Williams, A.C. et al. (2014). "A Computational Pipeline for Crowdsourced Transcriptions of Ancient Greek Papyrus Fragments". *IEEE International Conference on Big Data (Big Data)*. Washington, DC: IEEE, 100-5. https://doi.org/10.1109/BigData.2014.7004460.

Williams, A.C. et al. (2015). "Proteus: A Platform for Born Digital Critical Editions of Literary and Subliterary Papyri". *Digital Heritage*, 453-6. https://doi.org/10.1109/DigitalHeritage.2015.7419546.

Ziesche, R.F.; Arlt, T.; Finegan, D.P. et al. (20120). "4D Imaging of Lithium-batteries Using Correlative Neutron and X-ray Tomography with a Virtual Unrolling Technique". *Nature Communications*, 11(1), 1-11. https://doi.org/10.1038/s41467–019–13943–3.