The opened trays of carbonized papyrus from the Villa of the Papyri at Herculaneum form an enormously important collection of literary writing from the ancient world.
1. **Project Summary**

In this project we teamed up with a library (the National Library of Naples) and a museum (the Morgan Library and Museum) to gather and process data from iconic objects in those collections. The goal was to acquire representations of damaged manuscripts non-invasively using photography and X-rays, and then process that data in such a way as to restore the ability to read the writing. The core approach is an extension of “virtual unwrapping,” (Seales et al., 2016) which is based on X-ray micro-computed tomography and sits at the intersection of emerging techniques in artificial intelligence (and machine learning) and heritage science. In this study we were able to develop and analyze a new method for recovering and enhancing ink signals from within scrolls and manuscripts, and we then engineered a new AI-inspired way to render those signals and to experiment with the end-goal of readability and enhancement techniques. At the conclusion of the project we launched a global contest with partners from Silicon Valley based on the experimental findings of this work (Fig. 1). The kaggle-based contest\(^1\) involved more than 1,200 teams of experts competing to improve the methods and beat our benchmarks. We released our code as open-source and our datasets to the global research community.

![](image)

**Figure 1**: Landing page for the partner-funded competition based on many of the findings and technical advances in this report.

\(^1\) https://www.kaggle.com/competitions/vesuvius-challenge-ink-detection
2. Project Origins and Goals: Virtually Unwrapping the Invisible Library
This project emerged from the observation that new techniques in artificial intelligence and machine learning are beginning to play a significant role in finding solutions to long-standing problems in Heritage Science (Seales & Learner, forthcoming). As this report will show, machine learning played a central role in the breakthrough experiments we completed in this project. The particular challenge of this work revolved around the interplay of three of the core technologies required for virtual unwrapping: X-ray-based micro-computed tomography, the method for non-invasively imaging the interior of closed/wrapped/folded objects; machine learning methods, used to enhance and amplify the desired signals in the tomography; and the digital provenance chain, designed to engender trust in the final claims of virtual unwrapping.

The objects we studied in this project were authentic and cherished. From Herculaneum, we collected data non-invasively from still-rolled-up scrolls (Fig. 2). These enigmatic objects have captured the imagination of scholars and scientists worldwide because not only are they verifiably from the only library from antiquity ever discovered in situ (Lapatin, 2019), but also because many scrolls in the collection (perhaps as many as 500) are still wrapped up and their texts are completely unknown. Many have speculated as to what these hundreds of scrolls could bear witness to if we could read them (Sider, 2005; CBS, 2018). We were also able to collect data from four fragments of already-opened scrolls, each of which contained visible writing. Although fragmentary, the collection of opened Herculaneum material is vast and offers evidence that inspires a deep desire to read the still-closed scrolls. For example, some open fragments bear witness to lost Greek and Latin literary texts, with references to events and philosophies of the ancient world (Sider, 2005). The assigned titles such as “Rhetoric,” “On Poems,” “On Anger,” and “On Freedom of Speech” reveal the classical period’s version of contemporary discourse around the very things that make us human (THS, 2023).

In addition to the data from the venerated Herculaneum collection (which is scattered among several institutions, including the National Library in Naples, the Institut de France, the British Library, and Oxford’s Bodleian Library), we also used data from a badly damaged early medieval Egyptian Coptic codex from the Morgan Library and Museum in New York City (Fig. 3). The data had been acquired in 2017 (Wade, 2018), but this project enabled us to develop and apply machine learning techniques in support of virtual unwrapping across the entire invisible library of authentic world-class objects.

Figure 2: (Left) Research team member Christy Chapman viewing intact Herculaneum scrolls in Naples, Italy. (Center) Research team member Seth Parker comparing ID numbers on the intact Neapolitan scrolls. (Right) One of the scrolls being imaged using photogrammetry to prepare for “fixturing,” which is the construction of custom cases that facilitate safe imaging with micro-computed tomography.
In this white paper we summarize our findings from the virtual unwrapping of these objects, and we report on the newly developed machine learning tools that we have applied in the domain of micro-computed tomography (micro-CT) and associated spectral photography. Despite the challenge of doing this work during the global Covid-19 pandemic (Covid, 2023), with all the restrictions and unpredictable maneuvers that entailed, we have accomplished four primary goals, which we set out in our original proposal:

1) We broadened our virtual unwrapping pipeline (a series of geometric algorithms implemented computationally) to include a machine learning component that enables the revelation of carbon-based ink on severely compressed papyrus. In addition, we made discoveries that took us beyond our originally proposed idea of detecting the ink and developed a method to render the ink evidence, from micro-CT, in an enhanced photorealistic manner.

2) We employed this enhanced technique to create a “system and a method” for making a born-digital version of unseen text buried beneath layers of opened Herculaneum papyri as well as a photorealistic digital version of the text within the leaves of the M.910 codex.

3) We disseminated our results and made the software available as open source under a generous non-commercial license agreement² so that our tools can be applied to any and all types of damaged cultural objects containing hidden text.

4) We developed a metadata protocol for our complex digital images to address the unique provenance and scholarly usage issues that these types of born-digital images create.

2.1 The Invisible Library
We set out to capitalize on the emerging power of machine learning methods at scale for the purpose of virtual unwrapping and ink enhancement. We achieved that, and while these findings may seem tailored and hyper-focused on a specific (albeit important) collection of objects, through this work we believe we have forged a more general pathway for revealing any type of ink on any type of substrate in any type of damaged cultural artifact.

In fact, our work originated because we recognized that a multitude of text-based heritage objects in libraries and museums worldwide are essentially lost to scholars and the public, due to the massive damage that prevents their use, study, understanding, and enjoyment (Baumann et al., 2008; Brown et al., 2007; Lin and Seales, 2005). This “invisible library” is estimated to contain hundreds of thousands of critical texts that could significantly enhance or even alter our understanding of world history and culture, if they could only be read. Instead, their words are locked away behind the destruction and decay wrought by time. Figure 4 shows some examples.

![Figure 4: Objects from the “invisible library.”](image)

(A) Coptic papyrus codex containing Mani’s The Living Gospel at the Chester Beatty Library in Dublin; unknown medieval text (B) burned when the library from the Cathedral School of Chartres (one the most important schools during the Middle Ages, second only to Paris institutions) was bombed in World War II; scrolls from the Dead Sea Scrolls Collection (C), and phylacteries (D), tiny leather boxes that were worn like a pendant and contain Hebrew texts on vellum; fused and waterlogged papers (E) found on a frozen cadaver from the infamous doomed Franklin expedition to the Arctic; autographed version of Jack London’s personal copy of The Sea Wolf (F) tragically burned in the San Francisco earthquake conflagration of 1906 after London and his wife decided to safeguard it in a “flameproof” bank vault.

Subjecting these valuable objects to the potentially injurious effects of physical manipulation is obviously fraught with risk. Yet their benefits to humankind are severely limited when they are closeted away on dark shelves, well-protected but prevented from proffering knowledge and encouraging inquiry.

2.1.1 The Library at the Villa of the Papyri: Herculaneum

The Herculaneum Papyri are among the most iconic and inaccessible of this “invisible library” of irreparably damaged manuscripts. Buried and carbonized in the eruption of Mount Vesuvius in 79 CE, a library of between 800 and 1,100 papyrus rolls (the exact number is unknown and considered unknowable; see Sider, 2005; and Janko, 2016) was excavated in the 18th century

---

from the Italian city of Herculaneum. These highly prized artifacts form a unique window into the ancient world. The collection was found in a luxurious home believed to belong to the family of Julius Caesar's father-in-law, Lucius Calpurnius Piso Caesoninus. Now known as the Villa dei Papiri, the site produced the only large-scale library to have survived from Greco-Roman antiquity and the only classical one to have been found in situ (Lapatin, 2019). Horace and Virgil were known to have spent time in this villa. Experts therefore postulate that restored texts from Herculaneum might include lost or hitherto unknown Greek and Latin masterpieces, such as a Greek epic cycle of poetry on the Trojan War, or lost portions of Livy's History of Rome, or undiscovered writings by Aristotle (Sider, 2005; “The Herculaneum Scrolls,” 2018). According to Francesco Mercurio, former Director of the Biblioteca Nazionale in Naples, “If we can read all the papyri, we will be able to give humanity the opportunity to know Greek and Latin works which have become irreparably lost or forgotten; and we will be able to allow scholars to correct any errors and omissions made by medieval copyists for known works.”

In addition, due to the remarkable imbalance between Latin and Greek texts found in the excavated library (only ten percent of the opened and studied scrolls are in Latin, yet the villa belonged to a Roman aristocrat), many experts believe that a significant portion of the library remains buried, possibly containing hundreds, maybe thousands, more intact carbonized papyrus rolls (Sider, 2005; CBS, 2018). In providing a proven method for reading unopened carbonized texts, it is hoped that this project will spur momentum for further archaeological exploration of the Villa, an exciting initiative that would offer up a wealth of new literature for scholarly analysis (Sider, 2005; La Villa dei Papiri, 2013).

In short, this work promises a potential “renaissance of classical antiquity” as Professor Heyworth describes it (Marchant, 2018, p. 118). “We’d change the canon,” Heyworth claims. “I think the next generation is going to have a very different picture of antiquity.”

2.1.2 The Morgan M.910
As we were putting the project together, another highly prized yet severely damaged manuscript was brought to our attention by Professor Paul Dilley (Dilley, 2022): The manuscript M.910 at the Morgan Library and Museum in New York. One of the oldest near-complete copies of the Acts of the Apostles, this Sahidic Coptic Christian manuscript was copied in Egypt during the fifth or sixth century. According to Maria Fredericks, the Sherman Fairchild Head of the Thaw Conservation Center at the Morgan, both fire and water had damaged the codex sometime in the years prior to its acquisition by the Morgan in 1962. The thinner-than-average parchment is thus charred, tattered, gelatinized, and brittle in many areas. In addition, some of the book leaves are so distorted that they are actually interlocked, making it impossible to open without causing more damage and loss. Finally, due to its exposure to moisture and the long-term corrosive effects, the iron-gall ink used to write the manuscript has deteriorated in certain places, resulting in localized perforation of the parchment.

2.2 Virtual Unwrapping
This project built on our prior work on the scroll from En-Gedi, Israel (Seales et al., 2016), where we first demonstrated our “virtual unwrapping” software pipeline. The technique is designed specifically for objects that are in a distressed and precarious physical state, making collections such as the Herculaneum scrolls a perfect fit for its application. We spent some years studying and prototyping a step-by-step computational approach that works in a constrained digitization environment (such as in situ X-ray-based volumetric scanning via micro-CT), yet still manages to produce a digital restoration rich enough to enable true textual scholarship (Seales, forthcoming, 2023).

4 Spelling of “papiri” in the Villa’s title follows the Italian. In English it is normal to write “papyri.”
5 https://www.trismegistos.org/text/108698
The virtual unwrapping pipeline is organized into a series of geometric algorithms that transform the unstructured, volumetric (3D) micro-CT scan data into localized 2D images of pages. These geometric transformations capitalize on shape attributes and the spatial localization of the substrate (e.g., parchment, papyrus) to effect the necessary transformations of data. The first stage, “segmentation,” is concerned with isolating the surfaces of the document in the scan and producing a 3D model for each page. Next, the “texturing” stage searches the dataset for an ink signal and embeds that signal onto the surface model. The final stage of our pipeline, “flattening,” transforms the 3D surface - which is almost never flat and often is wrapped up because it comes from a scroll - into 2D images that scholars can easily read. The choice of algorithm to be used at each stage in this pipeline is extremely important and must be tailored to the object being studied.

In 2015, we applied the virtual unwrapping software, which we developed within the research team, to a CT-scan of the unopened, burned Hebrew scroll from En-Gedi and were able to visualize the never-before and never-to-be-seen writing trapped inside. We restored and rendered the text of five complete wraps of the scroll. For the first time, a complete text from an object so severely damaged that it could never be opened physically was digitally retrieved and recreated, representing a true technical breakthrough (Seales, 2016).\(^6\)

But the accomplishment drew accolades not just for its scientific leap forward. Of even greater significance was the biblical and textual scholarship that the project enabled. Thanks to the clarity of the writing – the text was rendered at a resolution comparable to archival quality images in digital libraries – Hebrew and biblical scholars were able to identify the scroll from En-Gedi as a 2,000-year-old copy of the Bible (Seales et al., 2017). Two distinct columns of Hebrew writing were clearly visible and revealed the scroll to be the book of Leviticus. In addition, scholars were able to perform a complete paleographical analysis of the written contents. “We have never found something as striking as this,” Dr. Emanuel Tov, Professor Emeritus at Hebrew University in Jerusalem and one of the world’s top experts on the Dead Sea Scrolls, told The New York Times (Wade, 2016).

Such interdisciplinary collaboration and discovery is, in fact, the great promise of this work. Without the scientific approach to digitally restore the writing, the scholarly value of the En-Gedi scroll would have remained lost. Restoring the full scholarly value of these prized artifacts is thus the ultimate goal of this project and those that follow in its wake.

2.2.1 New Questions

In prototyping the virtual unwrapping pipeline on the scroll from En-Gedi, we demonstrated the indisputable power of the computational process. But we also revealed its limitations and the need for further work, which we sought to do in this project. In particular, we noted that the combination of materials used in Herculaneum texts — carbon black ink on a highly compressed papyrus substrate — creates an additional challenge for segmentation and texturing, because the ink signature that is captured is very subtle and the layers of papyrus are so compressed and distorted that it becomes extremely difficult to isolate them from one another.

Another segmentation challenge arose when we began working with the M.910 manuscript: how to differentiate the writing on each side of a two-sided text. X-ray images contain details from both sides of an object in one image, conflating the writing from recto and verso sides of the

---

\(^6\) See [The Scroll from En-Gedi - Digital Restoration Initiative](https://www.wwtc.org)
Thus, even when ink signals readily appear in X-ray, as in the case with iron gall, it is difficult if not impossible to distinguish which signals are on one side of the page and which are on the other. This was not an issue with the scroll from En-Gedi, since the Hebrew scribes wrote only on one side of the parchment roll. But the book form of the M.910, plus the abnormally thin parchment used for its pages (in some cases thinner than a sheet of modern paper), made it necessary to find a way to address this conflation.

We also noted that scholars conduct their best analysis with image alternatives: some that appear more like photographs, and some that are processed and enhanced. It can be problematic to produce and carefully align these various alternatives when those images are coming from volumetric data that has been generated by X-ray and not visible light on a 2D surface.

Finally, we realized that we owed scholars a provenance chain that substantiates claims of invisible text made visible. Because the born-digital text cannot be physically verified, the individual steps of the computational pipeline must be represented in a way that can be inspected and understood, from raw data to final result. This chain of transformations, when made explicit and visualizable, is incredibly valuable not only for the scholar performing the initial reading, but also to support third party confirmation through peer review of the entire chain of data comprising the final text.

Although we believe the technical approaches we developed for amplifying ink sensitivity, improving the segmentation and texturing processes, enhancing final images, and supporting provenance chaining will be a breakthrough, our goal is not just another technological advance. Rather, we intend to apply – and encourage others to apply – these technical advances to support novel text discovery and its scholarly analysis.
3. Project Activities, Team, and Participants

We set out in this project to further develop and apply non-invasive technological approaches so that the writing hidden in the most iconic collection of lost and damaged classical manuscripts – the scrolls from Herculaneum – could be visualized, read, and studied. The first step was to acquire quality data. As noted earlier, Herculaneum scrolls represent the perfect storm of important content, massive damage, extreme fragility, and difficult-to-detect ink. However, gaining access to this material and acquiring the necessary high-resolution micro-CT data proved extremely difficult. Prior to the award, we worked with proxy material we created ourselves using modern day papyrus and carbon black ink that we scanned in our own laboratory (Parker et al., 2019), as well as with a miniscule authentic fragment containing one single character, a lunate sigma, which had been scanned at a synchrotron (Parker et al., 2019; Seales and Chapman, 2018). But with the support of the NEH Advancement Grant, in 2018 we were able to procure another authentic fragment from the Herculaneum collection of the Institut de France. This fragment contained significantly more visible text, and through vital partnerships with the Institut de France, Microphotonics, and the school of dentistry at UCLA, we acquired micro-CT data of the fragment at a 3-micron resolution over four different incident energies. Using this data, we were able to confirm spatial resolution constraints as well as information about the invariance of the method to incident energies in the ranges we collected.

Data acquisition of the M.910 proved challenging as well. Because of the M.910’s fragility and the Morgan Library & Museum’s reluctance to see it travel, we arranged for two scanning sessions: a primary one at the Morgan in Manhattan, for which a micro-CT system was set up on site to conduct the imaging of the entire M.910 manuscript (Wade, 2018), and a secondary one some months later for portions of the manuscript that were already separated from the overall codex and therefore able to travel safely. This secondary scan also provided access to additional equipment as well as test data using different scan parameters, such as resolution and incident energy distributions. The scans of fragments containing visible ink also provided the training data necessary for creating the “ink-ID” machine learning network tuned to this manuscript.

3.1 Collaboration

One of the most important aspects of any project, but especially one where the stakes of access to and handling of material is so high, is that of relationship development and management. As much time was spent thinking about the project from the perspective of the conservators and curators as it was focused on technical parameters. Our first step in any project is always to spend time on site viewing the object with its caretakers. We dedicated time and travel from other funding sources to fully explain our approach to our partners and, most importantly, to listen to their concerns and advice.

3.2 Fixturing

Collaboration is especially key when it comes to placing strangely-shaped, fragile, and highly valuable objects into a high-powered X-ray machine that requires precise positioning to get the desired image data. Each unique object that we imaged for this project required careful “fixturing” to secure it for safe handling and positioning within the computed tomography equipment. In each case, a custom solution was necessary. While we had developed protocols for mounting the Herculaneum scrolls when we performed our first scans of the papyrus rolls (Seales, 2009; Seales, 2013), lessons learned and advances in materials during the intervening years informed new approaches. In addition, the fixturing of the M.910 book form factor had to

---

7 We named this convolutional neural network as a portmanteau of “ink” and “identification”.

be completely different from that of the scrolls. We developed a new fixturing process for each object organically in conjunction with the conservation staff of the institutions where we worked (The Institut de France; The Morgan Library; The National Library of Naples, Italy). And while the establishment of best practice fixturing was not a specified outcome of this project, the activity is nonetheless key to its success.

![Figure 5: PhD student Stephen Parsons working with the photogrammetry setup to capture images of an intact Herculaneum scroll.](image)

3.2.1 Scroll Cases

Generally, the process starts with measurements by hand – dimensions, weights – and a study of safe orientations led by conservators. In the case of the scrolls, this progressed to a careful and complete reconstruction of the 3D shape of each scroll using a non-invasive, camera-based system known as photogrammetry (Fig. 5). The conservation staff placed each scroll on a motorized rotating stage against a white backdrop. To ensure proper sizing, known scale markers were placed on the stage next to the scroll. The stage slowly rotates a few (2-4) degrees at a time while a photographer takes pictures of the scroll from each angle. This process continues for a full 360 degrees. After a complete rotation, the conservation staff flips the scroll onto a new resting edge, and the process is repeated until the full external shape of the scroll has been photographed.

This process generates hundreds of 2D photographs, all of which are fed into a photogrammetry reconstruction software and compiled to create a full 3D digital model of the scroll. Using the scale markers in the photographs, the model is scaled to match the physical dimensions of the scroll. This 3D proxy is then imported into CAD software and used to design a cylindrical, form-fitting scan case that follows the exact curvature of the model. Once an initial design is completed, a physical 3D model of the scroll and the case that will hold the actual scroll are printed using selective laser sintering (SLS) 3D printing of Nylon 12 PA material. This material has an ASTM D638 tensile strength measurement of 6,815 psi and an ASTM D256 impact strength measurement of 43 J/m. The printed model is used to conduct an initial fit test of the case to ensure no manufacturing defect occurred during its printing.
The shape reconstruction, shown in Fig. 6, creates the modeling necessary to fabricate a complete, form-fitting fixture designed to hold the object snugly and prevent movement.

3.2.2 A Case for a Codex
Because each object is unique, the fixturing process can vary and require out-of-the-box thinking along with mechanical engineering and 3D modeling expertise. For the Morgan M.910 manuscript, for example, a form-fitted case proved not to be the best solution (Fig. 7). After several iterations and careful measurements with the lead conservator, our team landed on a design that solved a few problems. First, the safest orientation for a book or codex is horizontal, resting on its face or back as if placed on a table to be opened and read. However, this orientation is exactly the wrong one for micro-CT capture. For the X-ray beam to attenuate evenly through a book’s pages so that a quality and complete 3D image can be generated, a book must stand vertically on one of its four edges, with the X-ray beam perpendicular to the page surface. Due to the thin, brittle edges of its pages, which were much too flimsy and weak to support any downward pressure, the only side on which M.910 could rest was its spine. Yet, this position was also problematic due to the codex’s rare, exquisite, and quite fragile book binding.
Figure 7: (Top left) Conservator Maria Fredericks carefully holding the Morgan M.910 manuscript as we prepare it for imaging in the micro-computed tomography machine. One can clearly see the fragile, brittle edges of the manuscript pages. (Top right) The spine of the M.910, one of the few surviving examples from this period of codex book production. (bottom) Micro-CT images of the binding.

In addition, evenly distributed attenuation requires that a scanned object be of relatively homogeneous shape and density, such as a solid sphere, cube, or cylinder. However, since the large flat surface of the book page is significantly wider than the total thickness of the manuscript, this uniformity is missing, a characteristic of almost all oddly shaped damaged heritage objects. The form factor of the codex (thin in most directions, but very thick in the edge-on direction) required engineering a mount that would even out the density in all directions. Without some sort of mitigation, X-rays projected through the thickness of the manuscript as it rotated would be significantly brighter than those which passed through the thinner, edge-on orientation, resulting in a disparity that can greatly affect the quality of a scan. We solved this issue by fabricating a hollow, cylindrical mount and packing the rounded cavities with dense material to create an approximately uniform density for the x-rays in all directions around the mount.

Finally, acquiring a full set of images from all 360 degrees required the object to rotate as the scan progressed increasing the risk of additional damage due to movement. The mount thus needed to apply just the right pressure to the manuscript to keep pages from moving or separating during scanning.
The conservator also wanted a very gentle way to place the manuscript into the mounting apparatus. Our research team worked with the conservation team to design a mount that would hold the manuscript in the scanner with its binding down, safely wedging the codex between two vertical uprights (Fredericks, 2020). This resulting case (Fig. 8) consisted of two half-cylinders made from 3D printed plastic that slid together and left a rectangular gap in the middle for manuscript placement. It also included a quick-connect mounting plate for easily securing the loaded fixture into the SkyScan 1173 micro-CT scanner. Figure 9 shows the manuscript placed into the holder and then, once positioned, binding down, placed and mounted into the micro-CT machine. The black circular base remained in the scanner, allowing the manuscript case to be easily, gently, and securely installed using nylon set screws.

![Figure 8: Fixturing for the Morgan M.910 manuscript.](image)

![Figure 9: The M.910 manuscript is prepared for placement in the scanner. Foam and support boards ensure a firm but reasonable fit against the manuscript. The fabricated, 3D printed fixture satisfied all the design constraints and mounted directly to the rotational plate in the SkyScan 1173 micro-CT scanner.](image)

3.2.3 Fixturing Fragments
Because our machine learning ink-ID network is trained using corresponding, registered X-ray images of visible text, it was necessary to also create holders for Herculaneum papyrus and M.910 parchment fragments. Once again, this required bespoke approaches. For the M.910 fragments, we used 3D printing to create small cylinders with a sloped hollow in the middle that served as a tray for the fragment. These cylinders were then directly mounted onto the scanner's sample holders. The conservation team pre-mounted each fragment on conservation paper board and wrapped it in conservation grade plastic to make tiny envelopes that fit into the cylinder's trays (Fig. 10).
The fixturing process is extremely important and is a crucial first step for non-invasive work. Without proper fixturing, all the many benefits of non-invasive analysis are put at risk through the danger of physically handing. This crucial first step truly sets the stage for a successful project, both in terms of conservator trust and support and in the acquisition of high-quality data, two necessary elements that are often in conflict. Capturing data at high enough resolution for the development of the proposed machine learning based process for virtual unwrapping requires the positioning of fragile material in ways that automatically generate considerable angst in any well-trained conservator. And since these customized cases are also used in combination with Pelican cases for transport, the objects’ custodians are more willing to allow the objects to leave their home.

3.3 Experiments: Orchestration, Automation, Leaderboards
We spent significant activity orchestrating and then managing the many machine learning experiments we performed. The outcomes from these activities, summarized in Section 4, were only possible due to the apparatus we constructed to facilitate the organization and analysis of the machine learning experiments. The high-quality data, which we had in hand after fixturing and scanning, was the basis for extensive experimentation.

We developed a surprising amount of software scripts and tools to orchestrate the machine learning experiments once the data was structured and organized. While we did not initially set out in this project to produce an orchestration process that we would disseminate as an outcome, it was a substantial project activity and has led to outcomes that will have impact for our future projects.

The input to the orchestration suite is the primary data and metadata describing the many images, tomography volumes, label sets, and other components necessary for the machine learning algorithms to run. The output is an organized set of metrics and results, and an organized storage protocol for keeping track of the results, as well as computed visualizations and associated metadata that summarize the collective results of simultaneous experiments. The snapshot below (Fig. 11) of a personal lab notebook, not really intended as a broad dissemination platform, but incredibly useful in day-to-day research, shows how this infrastructure facilitates ongoing evaluation and development by producing metrics and summary images for the machine learning experiments.
The orchestration software helps to organize complex jobs that run on a cloud-based architecture (CPUs, GPUs, shared memory and data storage systems) and then manages the results to support archiving performance numbers and visualizing key indicators.

The scale of the experimentation that we were able to perform thanks to structured, automated orchestration was enormous. In preparing this report we surveyed our metadata logs, which indicated records of 2,093 GPU-based machine learning training jobs. Estimating an average run time of 12 hours for each job (an educated guess from the normal run times we have come to expect) we spent more than 25,116 GPU hours of experimental time training machine learning models on this data. This translates to almost 3 “GPU-years” of run-time. Compared to current large-scale machine learning training - Large Language Models - this is about 2.4% of the GPU time it takes to train a large language model with 65 billion parameters, like LLaMA (Meta-AI, 2023). While this comparison clearly shows that our experiments were much smaller scale, we ran many different jobs rather than one monolithic one, and we were able to do that thanks to structured, automated processes.

A complete review of the experimental development, with successive results over the project period, is available as part of a recent PhD thesis (Parsons, 2023). The best outcomes from the machine learning experiments, focused around ink-ID, are summarized in the next section. Fig. 12 shows a schematic of the storage and computing environment, with the numbered, color-coded arrows representing the steps for running an experiment:

1. Data stored in a cloud-based repository is moved to a storage system connected to a GPU cluster.
2. From the cluster, a GPU-based training session is launched and runs, often for many hours; GPUs may not have all data memory-resident, but they can access the storage system if necessary.
3. Results and metadata describing the training session are moved from the GPU cluster to another storage repository that is accessible to a visualization platform.
4. Results and metadata support construction of visualizations on another compute platform, which allows calculation of various metrics and a regression analysis.
5. Results, metadata, visualizations, and metrics are all returned to the original repository and associated with the original data, structured with a metadata schema.

Fig. 12 (right) shows a snippet of the metadata that helps to organize this multi-step processing pathway. We used a few helpful tools to automate most of this process, since each arrow in Fig. 12 represents time lost from other activities if those movements are not automated and require human intervention. We used Singularity containers to run experiments in a reliable software environment. For parallel jobs we used SLURM, and Rclone for data transfers. Most of these actions were coded with python scripts with email notifications for events like experiment completion.

**Figure 12:** Conducting the machine learning experiments required a cloud-based computing network to support data storage, GPU-based training of convolutional neural networks, visualizations of results, structured storage of results, and a metadata schema to manage everything.
4. Project Outcomes
This project’s focus was to build a machine learning apparatus that would produce results and advance our major goal of ink identification and enhancement of ink visibility in Herculaneum material. The following sections illustrate the various outcomes.

**Collected Data:** Micro-CT data from both the Herculaneum fragments imaged at Diamond Light Source (fragments from PHerc.Paris 1 and 2 and scans of the intact scrolls PHerc.Paris 3 and 4) and the Morgan M.910 codex and fragments.

We collected full volumetric data at high resolution of multiple artifacts (two complete papyrus scrolls with carbon ink; four papyrus fragments with visible text; one medieval parchment codex and associated fragments with visible iron gall ink). This data formed the basis of the machine learning work to support ink-ID and “trans-modal rendering.” Figure 13 shows example slices from the interior structure of one of the complete scrolls and makes it clear that the high spatial resolution from the Diamond synchrotron data is much sharper than the data collected on the same scroll in 2009 with a conventional x-ray source.

![Figure 13](image1.png)

*Figure 13:* (Top) The slice views of the tomography of a papyrus scroll shows the clear difference between the extremely sharp 8-micron resolution (left) and the much blurrier 25-micron resolution (right) collected in 2009. (Bottom) The tomography of the Herculaneum fragment does not clearly show evidence of ink, although it is clear in infrared photography. Madame Francois Berard, Director of the Library of the Institute de France, displays a fragment that has been fixtured and is ready to be placed within the synchrotron beam. Beam line scientist Robert Atwood (center in photo) and Laurent Chapon, former director of physical sciences at Diamond Light Source, look on as the object is positioned in the beam.
**Virtual Unwrapping Code:** Improved prototype code for virtual unwrapping, written in a combination of C++, Python, and leveraging OpenCV, ITK, VTK and other open-source image toolkits.

Code development continued throughout the project period and has been released as open source (GNU General Public License 3.0: [https://github.com/educelab/volume-cartographer/blob/develop/LICENSE](https://github.com/educelab/volume-cartographer/blob/develop/LICENSE)). The screenshot in Fig. 14 shows the various software components developed by our research group available at Github under the GNU software license. As this work continues even beyond the reach of this particular NEH-funded portion, the software tools can continue to evolve, improve, and create future impact. Through the Vesuvius Challenge, this has already materialized, as people around the world have downloaded, used, and contributed to these tools.

**Figure 14:** Our Github repository allows for the wide distribution and dissemination of the software developed in this project: [https://github.com/educelab](https://github.com/educelab)

**Text from M.910:** A “virtually unwrapped” digital edition of the Morgan M.910.

As shown in Figs. 15 and 17 we scanned the Morgan M.910 manuscript and then virtually unwrapped its pages using the Volume Cartographer software suite. Our result revealed never before seen text on approximately 80 percent of the codex’s 94 pages (written on both recto and verso). Collaborators were able to read portions of the text, aligning it with the known text of the Acts of the Apostles (Dilley, 2022). In Fig. 15 the orientation of the page is laid out for the readers to understand the text more easily. The virtually-unwrapped page shown on the left comes from the center of the manuscript, shown in cross-section in blue on the right.

Fig. 17 shows a very clear page, with the text identified as coming from The Acts of the Apostles chapter 10 verse 4, where Cornelius encounters an angel:

> Cornelius stared at him in fear. “What is it, Lord?” he asked. The angel answered, “Your prayers and gifts to the poor have come up as a memorial offering before God.”

This was very satisfying outcome, proving again – as with the Scroll from En-Gedi – that textual scholarship is possible based on virtually-unwrapped results. Analysis of this data and the text it contains is ongoing.
**Figure 15:** The virtually unwrapped page from the center of the still-closed Morgan M.910 (identified in blue in the cross section on the right) shows clear writing. The annotations indicate the page orientation. The language is Coptic. The M.910 is one of the earliest known versions of the Acts of the Apostles.

**Ink-ID:** A new machine learning component of our virtual unwrapping (VU) pipeline, called “ink-ID”, that is capable of detecting subtle patterns in micro-CT using machine learning.

The ink-ID framework we developed successfully demonstrates that the carbon-based ink of Herculaneum can indeed be imaged and detected from micro-CT. This breakthrough result is based on the “morphological hypothesis,” which postulates that carbon ink is not “invisible” in micro-CT but rather is present as a complex set of subtle cues: textural patterns, changes to the papyrus fibers of the substrate material, and so on.

The empirical results we have collected support the morphological hypothesis, even though these patterns are typically unobservable in the micro-CT data with the naked eye. Using a supervised machine learning framework and opened fragments with visible ink, we have confirmed that the differences between inked surfaces and un-inked surfaces can be successfully characterized and then reliably detected by the network.

Figure 16 shows final results of having tuned our ink-ID experiments to train a classifier to detect the localized presence of ink on the papyrus substrate. Successfully training this classifier and demonstrating it working on ground-truth data means that it is possible to use such an approach on the still-closed Herculaneum scrolls. The systematic study as part of this project explores many variants to the method and evaluates numerous parameters involved (Parsons, 2023). On the left (Fig. 16) is a color depiction of the 8-fold cross-validation structure of the machine learning experiment. Eight separate ink detection models are trained. Each time, one of the indicated regions (half of one fragment surface) is left out, to be used as the validation dataset. The ink-ID network is trained with all the other data and once trained, attempts to classify the ink portions in the left-out region. After repeating for all eight fragment halves, the composite (center column in Fig. 16, labeled “Generated image”) shows the ink-ID performance across all trained models. We expect the results to approach the readability of photographs of the visible layers once more training data is collected.
Figure 16: This 8-fold experiment, taken from Stephen Parsons’s PhD thesis (Parsons, 2023) shows the performance of “ink-ID” in making the ink of Herculaneum appear from micro-CT. Shown here are the label images (for training, since the ink in the micro-CT cannot be seen with the naked eye) and the result of an 8-fold training and classification experiment.

**Trans-modal Rendering:** A powerful generalization to ink-ID that can render photorealistic color images, with only micro-CT images as inputs.

Our novel trans-modal rendering method demonstrates that the ink-ID network can learn appropriate color attributions from registered 2D label images. We postulated in the proposal that we could make progress on the question of rendering micro-CT images into a color image.
format, making the results look much more like photography than X-ray. Indeed, by training ink-ID with labels directly from color images, we demonstrate that the interior pages (which have never been photographed, but only imaged using volumetric micro-CT and then virtually unwrapped using Volume Cartographer) can be rendered photorealistically. This type of “digital restoration” learns the restorative mapping, or the “style transfer,” between what characteristics micro-CT captures about the material and that which visible light captures. Fig. 17 shows two examples, one from the Morgan M.910 and one from a small fragment from the Dead Sea Scroll collection, courtesy of the Israel Antiquities Authority (IAA). While ink detection models learn only to classify the presence or absence of ink, trans-modal rendering models are more powerful, learning how to visually represent subtle cues in CT such as papyrus fibers, surface cracks, sand grains, and more.

**Figure 17:** The manuscript page from the interior of the Morgan M.910 shows clear text (upper left). The text is Acts 10:4 where Cornelius encounters an angel. The multimodal rendering, which is trained to mimic photography using training examples from other fragments, for the same page is shown on the right. On the second row is an example from a fragment from the Dead Sea Scroll collection.

**METS Schema and the Digital Provenance Chain:** an associated metadata schema to record parameters and hyperparameters throughout the stages of image processing.

We call this approach a “digital provenance chain” to facilitate a reproducible record of processing for the purpose of recording experimental results and allowing peer review. While we recognize the incredible importance of the reproducibility of something as complex as virtual unwrapping and ink-ID, it turns out to be non-trivial to capture and record all the processing steps involved in the highly computation-based images like those in Figs. 16 and 17. As part of this project, we explored a metadata schema designed to capture the details of these complex processing paths, and store them in a structured way. Doing this supports visualization, peer review, and the durability of final claims that are made from non-invasive imaging of important artifacts. Figure 18 depicts the data pathways (input data, processing steps, and output texture claims) that are captured in the METS schema we designed as part of this project.

We designed and implemented a preliminary version of the schema and reported on it at the annual conference for Metadata and Semantic Research (MTSR 2020), where it won the Best Paper award (Chapman, 2020). The structure supports an organization of all the data playing a
role in the production of the complex digital object that makes up a final unwrapped texture claim, and allows associated software "actions" to be developed. The METS behavior section "provides a means to link digital content with applications or computer programming code that can be used in conjunction with the other information in the METS document to render or display the digital object, or to transform one or more of its component content files." For our digital objects, the METS "behaviorSec" includes a set of behavior mechanism "xlink:href" pointers to code that, when invoked, will assemble all of the files and interface information necessary to execute the behavior. The various behavior mechanisms will launch visualizations that (1) explain how input files were manipulated at various iterations, (2) enable the replication of such manipulations, and (3) depict all the processes and associated metadata.

For example, we employ various working pieces of software internally to create ad hoc visualizations of our processes and their outputs. We refer to these behaviors as EduceData Visualizations and are in the process of transforming these tools into the METS behaviors format. Applying the METS structure to our tools will afford them greater portability, scalability, and searchability so that they and the digital objects we create can achieve widespread distribution and use. The development of these tools around the central METS schema is work that is ongoing.

---

9 To view an example using example data from Herculaneum, see http://infoforest.cs.uky.edu/pherc118/
The METS schema is a structured way to capture the complex processing paths that make up a final texture claim from these volumetric digital objects. The machine learning enhancement (ink-ID) adds new complexity because there are labels, registration parameters, and many training parameters that must be captured in order to meaningfully record what has been done in making a final claim of a virtually-unwrapped, readable text.

Data Management Ecosystem: an experimental ecosystem for managing the many different data sources and metadata items produced in the machine learning phase, and a “leaderboard” mechanism for evaluating and visualizing the results of the experiments.

Although less glamorous than ink-ID, and perhaps less challenging than METS-based metadata schemas, the management ecosystem is crucially important, as shown in Section 3.3. We engineered a number of scripts, metadata structures, and file organizations for managing all the digital assets in this project work. The image in Fig. 19 illustrates the kind of data that the orchestration packages help us manage and then visualize to understand results and make decisions about parameters and improvements. Comparing new experiments to baselines, understanding the performance of different kinds of loss functions, and visualizing output images to assess readability are all facilitated with the orchestration infrastructure.
Figure 19: The orchestration infrastructure we created helps to organize all the parameters around ink-ID and the many experiments we ran and tracked.

**Fixturing Protocols:** Fixturing protocols for transporting and imaging highly damaged material in a safe way.

As described in Section 3.2, we built and executed a process for fixturing that included software, interactions with conservation staff, modeling and fitting and 3D printing, etc. As mundane as it may seem, appropriate fixturing is the oil needed for the race car to run. Although we have not yet released a primer on what we learned, Maria Fredericks, the Sherman Fairchild Head of the Thaw Conservation Center at the Morgan Library and Museum, wrote an excellent piece that detailed the overall project and aspects of the fixturing problems for the M.910 (Fredericks, 2020).
5. Project Evaluation and Impact
Evaluation for this project involved our internal review (code review, results from experiments, shared laboratory notebooks with incremental progress reports) followed by external reviews from colleagues such as machine learning experts, papyrologists, and experienced practitioners with micro-CT. The purpose of the work was to develop the machine learning framework and apply it to the specific data sets we have available. At the various decision points of the project, we were able to evaluate results based on controls such as

- Proxy objects for which we had known results.
- Fragments from which we could see visible layers and compare results.
- Machine learning metrics indicating precision and recall relative to known labels.
- Obvious readability confirmed by papyrologists who could validate how well the machine learning algorithms were producing ink signals.

We believe this project will have a high impact, although we have not yet completed the dissemination of all our results. Parsons’s PhD thesis will be released in September 2023 after which we will prepare major publications. Datasets are now available, and practitioners are already exploring the data we have structured and released.

The most tangible and immediate impact we have seen, which we mentioned in the introduction of this white paper, is the launch of the “Vesuvius Challenge.” The work we did for this project directly prepared us to launch a Kaggle contest, which required careful tutorials and an organized webpage (scrollprize.org) with examples and references and experimental results. Our partners are tremendously talented; we were able to bridge the gap between what was needed to run that competition and our day-to-day project work thanks to those partners and NEH’s support for this project.

The Vesuvius Challenge (VC) was launched on the Ides of March 2023 to scale up the experimental effort that we established in this project. We formulated baseline results, and, with the help of the VC production and implementation team, we created an automated metric for the kaggle competition. This allowed competitors to submit the results from working solutions and then have them evaluated automatically on held-back data. The kaggle framework created a leaderboard, showing competitors where they sat in the solution hierarchy relative to other competitors.

The VC organizing committee decided to offer a series of prizes designed to entice participants with low-hanging fruit but also involve as many as possible in the grand prize: reading the interior wraps of intact scrolls. Example early prizes: first letters prize; open-source tooling prize; and the kaggle competition prizes, which are awarded based on the final leaderboard positions of the competitors (ranked according to the automatic metric, which measures a mixture of precision and recall of the open and held-back fragments). Thanks to the concerted fundraising capacity of the VC organizing committee (led by Nat Friedman), the prize pool grew to over $1M, capturing the attention of over 1000 competitors in the kaggle phase of the competition.

---

10 https://arxiv.org/abs/2304.02084
11 https://www.kaggle.com/competitions/vesuvius-challenge-ink-detection
12 Nat Friedman, Daniel Gross, JP Posma, Daniel Havir
The grand prize of $700,000 will be awarded based on the following qualifying criteria:

- Read at least 4 separate passages from the two Herculaneum scrolls, each containing at least 140 characters of continuous text.
- Verify that each passage contains no more than 15% of characters which are missing or illegible.
- Confirm that submissions contain legitimate and linguistically plausible text.
- Independently reproduce and verify your results using your code and documented techniques.

This is a substantial challenge, which is ongoing, with submissions evaluated at any point they are received. We expect the competition to run through the end of the calendar year 2023. More information is available at scrollprize.org and from other media coverage (Sample, 2023).

From the competition, we gained a threefold boost in productivity:

1. Over 1514 competitors (grouped into 1249 teams) entered the kaggle contest and competed for the first set of prizes. Such significant engagement had a truly remarkable impact and produced a number of very interesting contributions from all around the globe.
2. Peer review is never more thorough than when 1249 independent competitive teams can verify that techniques work and that data is carefully and clearly curated. The peer review and confirmation of our work by talented and motivated third parties was important and gratifying.
3. The human ecosystem around the competition has increased enormously - in the form of conversations, live streams, videos and news coverage, and a community of more than 1300 members on Discord. These individuals are having daily discussions around the various technical and scholarly questions, methods, and materials generated by this project. The expanding community and conversations are another tremendous boost in awareness and interest.

In terms of “Lessons Learned” from the overall project funded by the NEH Digital Humanities Advancement Grant, we offer the following:

- Big ideas remain small steps without a fair amount of pedestrian work. We took the time to build orchestration and management tools, and it has paid off.
- Covid had a bigger impact than we thought it would. Although most of our work was software development and data processing, the national/global Covid-19 panic had a strong effect.
- Open software, open data, and open scholarship is tremendously valuable for dissemination and broader impact.
- The tires of conventional wisdom must be kicked periodically. With changing tools and techniques, what was formerly impossible or improbable may suddenly become achievable.

---

13 https://discord.com/invite/6FgWYNjb4N
6. Project Continuation and Long-Term Impact

In the end, this project settled once and for all the crucial question of whether carbon inks in Herculaneum material are detectable using non-invasive micro-CT plus virtual unwrapping. They are. Ink-ID works and was confirmed by many contestants working on the kaggle framework. The task facing us now is to scale up the effort by creating a very large scale training set for the machine learning network and automating the following technical approaches that are barriers to a full-scale approach to the Herculaneum material:

1. Automatic segmentation: we need methods that can virtually unwrap a papyrus scroll or the pages of a book with high accuracy in a completely automatic way.
2. Generation of labels for ink-ID training: we need either to move away from supervised learning techniques or find a way to generate large scale labels that will improve ink-ID to the point of uncontested readability.
3. Dial in the imaging protocols: because we can image the material so infrequently, we have not settled on the best imaging protocols. Questions remain, such as what is the best spatial resolution? What is the required incident energy? Is a synchrotron required?

This project was a huge step forward in our long-term plan of reading all Herculaneum material using non-invasive techniques. We will continue the work through EduceLab, the heritage science midscale infrastructure being funded through an NEH Infrastructure Challenge Grant and the NSF’s midscale infrastructure program (NSF, 2021; Digital Restoration Initiative, 2021). While both programs primarily fund equipment, we are actively seeking a path to sustainability that will allow us to continue developing human expertise in this area and keep this progress moving forward. The goals we achieved through this NEH Advancement Grant funding firmly set us on that path.
References and Additional Reading


Mid-scale RI-1 (M1:IP): EduceLab: Infrastructure for Next-Generation Heritage Science, A 2021 National Science Foundation Mid-Scale Research Infrastructure Project. NSF Award Number 2131940. https://educelab.engr.uky.edu/


Seales, W. B., & Learner, T. (Forthcoming). Fostering Collaborative Breakthroughs in Heritage Science through Machine Learning and Data Science. Final Report to the National Science Foundation, Award #2035533.


