Modern Technology / Ancient Manuscripts

Michael Penn, Department of Religion, Mount Holyoke College [Project Leader]
mpenn@mtholyoke.edu

Nicholas Howe, Department of Computer Science, Smith College
nhowe@smith.edu
The *Modern Technology / Ancient Manuscripts* project is committed to the idea that the best way for students to learn about the digital humanities is through their direct involvement in all aspects of the design, implementation, analysis, and use of a cutting-edge, digital humanities tool. The proposed, two year project seeks to build on an already successful, student-centered, digital humanities collaboration of two five-college professors and nine five-college students that began in 2009. The end-product would be a digital tool useful to students and scholars alike in their investigations of ancient manuscripts and the communities that produced them.

**Project Background and Status**

In the Spring of 2009, professors Michael Penn (Mount Holyoke College, Department of Religion) and Nicholas Howe (Smith College, Department of Computer Science) began using recent advances in digital handwriting recognition to identify the scribal hands of ancient manuscripts and to uncover relationships between historical documents. The goal was a digital system that would embody and extend the expertise of human scholars. A student could scan in manuscript pages of interest, and the system would then compare the test document to those in its database, list the degrees of similarity between this document’s handwriting and those of other manuscripts, note any close enough to suggest that they were written by the same scribe, and compare the test document’s handwriting to the handwriting of documents with dated colophons to help estimate when the manuscript was originally written.

Without technological intervention, this level of analysis would not be possible. Less than a third of extant manuscripts contain a colophon providing the name of the copyist or the date of its composition. We simply do not know how many different scribes were responsible for the other 70 percent of documents, which documents were written by the same scribes, or—in most cases—when these documents were written. Even the few scholars who have the necessary
paleographic expertise cannot possibly compare the handwriting of a given manuscript with thousands of others to properly trace the relationships between them.

In addition to providing information about the provenance of individual manuscripts, the use of digital analysis would enable students to discover new connections between these texts. The ability to identify multiple documents from the same scribe will help correlate data from one manuscript to another, discern manuscripts that have been divided between modern libraries, trace the development of scribal schools, approximate the contents of ancient monastic collections, and track the circulation of ancient codices. The digital analysis of ancient handwriting thus provides a key for understanding not only the composition of individual documents, but also the larger context and communities in which they were written.

In addition to providing new information about individual manuscripts, when combined with traditional tools of historical and literary analysis, this project would lead to new questions and new fields of inquiry. Because most manuscripts currently lack verifiable dates of composition and their relationships with each other are indeterminate, one most often uses manuscripts solely as a means to reconstruct the earliest form of the individual texts they contain. In contrast, the data this project generates through the clustering of related manuscripts would allow students to explore manuscripts as collections. For example, what if a student discovered that many anti-Jewish writings were copied in the decades following the Islamic conquests? How might this affect our understanding of the connections between Christian anti-Judaism and the rise of Islam? How might this strengthen the relationship between seventh-century anti-Jewish policies in Byzantium and anti-Judaism further east? What if one of the scribes who copied Aristotle was also responsible for transmitting a number of anti-Muslim treatises? What if it were discovered that he was writing from Baghdad? How could this affect our notions of why eastern Christians were so interested in classical logic? How would this affect their role in the Abbasid caliphate translating and preserving Greek science and philosophy? The ability to ask (and in some cases, to begin answering) such questions
transforms scribes from neutral, after-the-fact transmitters of ancient texts into active historical agents whom students would be able to effectively study.

In order to test the feasibility of such a system, Dr. Howe initially enlisted the help of one of his undergraduate engineering students, Ms. Emma Dalton. The resulting senior thesis included a working prototype developed in MATLAB that quantified how closely the handwriting of a test document matched the handwriting of other manuscripts in the computer’s database. It did this by first measuring two elements of handwriting style: (1) contour hinge—how rounded is the text and (2) run length—capturing qualities such as how squinted or expansive the handwriting is [Figures 1–2]. The program then used a technique recently developed at UMass, called letter congealing, to compare individual letters to each other [Figure 3]. The computer produced a mean of all its samples of a given letter to create a standard against which the individual letters from a given document can be measured. The system then compared each letter in a document against this standardized letter form in three ways: (1) whole letter (2) letter parts (3) the blank space between letter strokes [Figure 4].

Ms. Dalton next assembled a database of handwriting samples from twenty ancient Aramaic codices to help test the system’s accuracy. The results of Ms. Dalton’s thesis project were quite impressive. If the computer never looked at a single letter but simply the overarching writing style, it would correctly match the documents’ handwriting three times out of four. If the computer compared the exterior shape of individual letters, the accuracy jumped to 92 percent. If the computer also compared individual parts of the letter and the spaces within the letters, the accuracy was 98.7 percent. When it combined all five handwriting algorithms together, its accuracy was 100 percent [figure 4]. Every time the system would properly match a test page with the manuscript from which it came.

A 100 percent success rate for twenty manuscripts does not guarantee a perfect success rate for a database containing hundreds of manuscripts. But it was a very encouraging result for a prototype and motivated us to substantially expand student involvement in the project. Internal grants from Mount Holyoke College supported a computer science undergraduate who worked
for two summers with Dr. Howe designing the system’s Java interface. A grant from the American Academy of Religion allowed Dr. Penn to spend a month at the British Library obtaining additional images for the project. Mount Holyoke later funded six religion and one other computer science student to enlarge the computer’s manuscript database. By the end of last semester, the system was fully networked and, by the end of this spring, it will have a database of seventy-five manuscripts.

We seek five college digital humanities funding to build upon this successful model of student involvement in the digital humanities. In addition to Dr. Penn and Dr. Howe, the Modern Technology / Ancient Manuscripts project would include (depending on student continuity between the two years of the project) from four to eight five-college students, half with background in computer science, half in the humanities. Over a two year period, this digital humanities team would 1) Expand the current system; 2) Construct and implement new algorithms for manuscript analysis; 3) Illustrate the system’s utility through student initiated projects that use the digital tool they helped developed; 4) Refine the interface so that the system could be used by future students, both in independent research projects and as part of existing course curriculum; 5) Help disseminate project results through academic publication and the mentoring of future student assistants.

Project Schedule

**Academic year 2013 – 2014.** Half the students will focus on building digital tools and the other half on manuscript analysis. Nevertheless, they will work closely with each other so that experience with manuscripts influences programming, and software development affects the design of individual research projects. The goal is for students to quickly become active participants in the digital humanities and to provide them with an in-depth pedagogical experience. First, they will be directly involved in the expansion and refinement of a newly developed digital humanities tool as they double the size of the current manuscript database, test the system’s efficacy with this larger sample set, and further enhance the Java interface. Second,
the students will begin developing their own projects for how to use this digital tool to address larger computational and historical questions. For example, computational projects might examine the effects of different types of learning algorithms on the system’s efficacy, or test new style comparison techniques that have recently appeared in the scientific literature. Historical projects might focus on how the system’s new dating of manuscripts changes our understanding of early Christian polemics or how a set of documents that the system newly identified as coming from the same scribe affects modern interpretation of their content.

Phase 2. Summer, 2014. Two computer science students will work full-time with Dr. Howe to further refine the system’s structure in MATLAB, to integrate an optical character recognition component into the Java interface, and to begin implementing their computational projects.

For the humanists, summer research would concentrate on acquiring manuscript images appropriate for the project. Because the system has been developed to work with Aramaic manuscripts and almost none of the most ancient Aramaic manuscripts have been digitized, it is necessary to see these manuscripts in person. Only through a direct examination can one determine the specific script used, note changes of hand, marginalia, and colophons, and determine which pages are sufficiently preserved for the computer to read. Without direct observation, the majority of the images ordered would be useless for the project, a particularly costly mistake given that each set of manuscript images costs $40. In truth, a trained scholar could analyze these manuscripts more quickly than students. Nevertheless, with the pedagogical goals of providing students with direct research experience and of having them fully participate in the project’s implementation and design, our proposal also includes a request for student travel. Two humanities students would travel with Dr. Penn to the British Library to work directly with the world’s largest collection of ancient Aramaic manuscripts. In London, Dr. Penn will train them in manuscript handling, cataloging, script identification, and image acquisition. While in London, the students would choose one hundred and fifty manuscripts that are most important for the system’s database and for their individual research projects. For the remainder of the summer, they will add these manuscripts to the database and implement their individual projects.
Phase 3. Academic Year 2014-2015. The final project phase focuses on three modes of dissemination. The first is to finalize the individual student research projects and ready them for publication. This Smith-Mount Holyoke collaboration has already resulted in two scholarly articles co-authored by five-college undergraduates (Proceedings of the Workshop on Historical Document Imaging in Processing, 2011 and Journal of Syriac Studies, 2012). As the project is the first of its kind in digital paleography and continues to generate new historical data, it is very likely that student research projects will also produce publishable results. Dr. Howe and Dr. Penn will work with students in order to submit their work for publication. The second mode of dissemination will be to transform the system into a form more suitable for future students. On the computer science side, this involves identifying portions of the system that could form appropriate case studies or lab activities for computer and engineering classes. On the humanities side, this involves identifying specific projects that could be integrated into a humanities class syllabus and providing appropriate documentation that would allow professors to use this newly developed digital tool in their existing courses. The third mode of dissemination consists of peer mentorship. At the end of the Mellon grant period, it is anticipated that the project would continue on a smaller scale, modestly supported—as it has been in the past—by MHC and Smith research assistant grants, as well as students doing independent thesis work. Toward the end of the Mellon grant period, students from the digital humanities team would thus mentor a new set of students. They would fully documenting the project, as well as directly train incoming research assistants and any students planning to use the system as part of their thesis work.

Pedagogical Goals

Building on an already successful, student-centered collaboration, the project’s pedagogical impact would be four fold:

First, the project would provide a core group of students with the opportunity to lead all aspects of the development and implementation of a highly significant digital humanities tool. They will direct all elements of the tool’s production from project design, to implementation, to assessment, to
dissemination. At the end of two years, these students will have obtained a wide range of in-depth experiences ranging from software coding to the handling of ancient manuscripts; they will also have become full fledged digital humanities practitioners.

Second, the project will have created a significant pedagogical tool that could be integrated into the current curriculum. The project would provide a platform for student research projects—short assignments, research papers, senior theses—that would allow a wide range of students to directly engage in meaningful computational and manuscript research. The system does not require the student to have mastered an entire ancient language, simply to recognize the letters in its alphabet. As a result, it can be easily incorporated into five college classes in computer science, classics, history, and religion, as well as Smith’s recently developed concentration in book studies.

Third, the project will offer important opportunities for collaborative learning and peer mentorship. Students will be part of a truly interdisciplinary team in which students in the humanities and students in computer science will co-design a digital humanities tool. As part of this process, they also will coordinate their efforts with related digital humanities endeavors such as the Five College Digital Manuscripts Project, work on multi-author publications, and help mentor their successors.

Perhaps most importantly, the project’s overall development—as well as the students’ individual research projects—can serve as an important model for future endeavors. Modern Technology / Ancient Manuscripts will provide a core group of students with a set of deep, meaningful, interdisciplinary research experiences. Through publication, peer mentorship, and curricular design they, in turn, will reach a much wider audience of fellow students and scholars. Modern Technology / Ancient Manuscripts will thus demonstrate how undergraduates as producers, not simply consumers, of the digital humanities can substantially advance both learning and scholarship.

Resource Sharing and Project Sustainability

Mount Holyoke LITS has already provided server space and support for this project and both Mount Holyoke and Smith have site licenses for the one piece of proprietary software that the project
employs. As a result, there are only two equipment expenses. The first is for the manuscript images necessary for the system’s database. The aim is to increase the project database by 150 more manuscripts in order to have a sufficiently large sample size to accurately assess and refine the system’s algorithms and to support student research projects. These are provided by the British Library at a fixed fee of $40 for a manuscript section. The other material expense is the acquisition of a computer workstation (approximate cost of $2700). Given the high amount of student use, only through a dedicated machine can one keep all files in place, keep configurations consistent, and effectively back-up data. Although much of the data entry can be done on student laptops, the actual analysis of the data requires access to modern graphic processing hardware not available on the older workstations in Dr. Howe’s lab. The securing of a new machine will not only enable students to process the data when they need to, it also will ensure that they have the best possible platform for the proposed research. To help off-set such expenses, the proposed budget includes $6500 of cost-sharing.

For the last two years this project has been partially supported through Mount Holyoke’s research assistant grants and faculty grants. Through these programs, Mount Holyoke has contributed approximately $1250 / year to help pay for research assistants and $1000 to help pay for faculty travel. Mount Holyoke cannot guarantee project support ahead of time. Nevertheless, the project budget has been calculated in the expectation of a similar level of support over the Mellon grant period. Smith College has not previously supported the project, but we suspect that the project would be competitive for Smith’s faculty grants. Smith also cannot guarantee support in advanced. Nevertheless, the project budget has been calculated with the assumption that Smith would be able to pay $1500 / year ($3000 total) toward the employment of student assistants.

There is another set of area resources that the Modern Technology / Ancient Manuscripts project aims to both draw upon and contribute to. Our project is particularly complementary to the Five College Digital Manuscripts Project currently funded by the Mellon Foundation. The tools that project has developed for digital codicology, such as the automated measurement of mise-en-page and ink density would be of great help to our project. So too, our work on digital handwriting analysis could
help advance the *Digital Manuscript Project’s* work on the identification of scribal houses. Students on each project will be working on manuscripts written in different languages and each project has its own analytical focus. Nevertheless, these two projects’ share similar enough goals to allow students from each project to work with those of the other, providing additional opportunities for meaningful research collaboration and peer mentorship.

A Five College Mellon Grant would substantially advance the *Modern Technology / Ancient Manuscripts* project. It would provide a core set of students with in-depth research experiences in the digital humanities, help them create a curricular and analytical tool applicable to a wide range of class room and research settings, and illustrate the project’s long term feasibility and import. At the end of the two year period, we anticipate that Mount Holyoke and Smith would continue to provide modest support for future research assistants and that the project would continue to attract students interested in doing independent work and senior theses in the humanities and in computer science. Internal funding opportunities such as research assistant, Reese, and UAF grants would suffice for additional, incremental changes to the system. At the end of the Mellon grant, the core group of students would be involved in training a new group of research assistants who will continue to move the project forward.

At this point, the project would also be extremely competitive for external funding opportunities such as the ACLS Digital Innovation Grant, the NEH Level Two Digital Humanities Grant, the NEH Collaborative Grant, ACM-W Scholarships, and the NSF’s Creative Research Awards for Transformative Interdisciplinary Ventures. Building on the strong foundation obtained by student led digital humanities research, these external funding opportunities would allow for a further expansion of the project into other language groups such as Arabic, Greek, and Latin. This would enable students and scholars to analyze in fundamentally new ways thousands of pre-modern manuscripts and the societies that produced them.
Proposed Budget:

Physical Equipment and Supplies:
- dedicated work station at Smith College: $2700
- manuscript images from the British Library (150 manuscripts): $6000

Student Hours:
- during academic years: $4840 [7340 minus $2500 cost share from MHC research assistantship grants]
- summer hours: $11,400 [14,400 minus $3000 cost share from Smith faculty grants]

Travel to British Library:
- airfare (two students, one professor): $3600
- lodging (3 weeks for professor): $2000 [3000 minus $1000 cost share from MHC faculty grant]
- lodging (5 weeks for two students): $5000

Total: $35,540

Explanation:

Mount Holyoke LITS has already provided server space dedicated to this project. Both MHC and Smith already have site licenses for MATLAB, the one piece of proprietary software required for the project. In terms of infrastructure, the only additional requirement is a dedicated PC workstation based in Dr. Howe’s lab so that students can use advanced graphic hardware to analyze the project data.

The British Library has the world’s most comprehensive collection of early Aramaic manuscripts. They provide imaging services as a fixed rate of approximately $40 / manuscript. The estimated cost for 150 manuscripts is thus $6000.

Academic year based on four students, each working five hours a week, twelve weeks a semester for four semesters, which is a total of 960 hours. Compensation figured at $9 / hour. In previous years, Mount Holyoke has provided modest research assistant support. This budget assumes a similar level of future support, that is $2500 over the duration of the project.

Summer hours based on four students working forty hours a week for ten weeks.
- Compensation figured at $9 / hour. This budget assumes a level of research support similar to what Dr. Howe has received in the past through faculty grants, that is $3000 over the duration of the project to help off-set the cost of summer student hours.
- Airfare based on average round trip summer fare from Hartford-London.
- Lodging approximated for professor’s three week stay single room while training the students and helping them begin their research. Mount Holyoke has traditionally provided modest faculty grant support to off-set travel for this project and the budget assumes a similar level of support, that is $1000. Student lodging was calculated for two students staying five weeks in a shared room in London.
Figure 1. Contour hinge. The system measures the interior angles of each letter to determine the “roundness” of the contour. 3D probability distributions of handwriting from two different scribes illustrate the measurable difference between these two scribes’ handwriting. Because the method is text-independent, the user needs to provide the computer with no further input than the facsimile manuscript page.

Figure 2. Run length. The system calculates the vertical and horizontal lengths for each point along each contour indicating how “squinty” or “expansive” the scribe’s handwriting is. Horizontal and vertical run length histograms from handwriting samples of two Aramaic scribes illustrate the measurable difference between their handwriting. Because the method is text-independent, the user needs to provide the computer with no further input than the facsimile manuscript page.
Figure 3. Letter congealing. To compare individual letters, the system first makes a “weighted average” from the database’s samples of that letter. This provides the software with a standard against which it can quantify a scribe’s individual variance.

Figure 4. Letter parts and concave hulls. In addition to comparing the outline of individual letters, the system also compares the shape of individual elements of each letter. For example, with the Aramaic letter alaph it compares three different sections of the whole letter as well as the three different concavities created by the letter strokes.

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text-Independent</td>
<td>Contour Hinge</td>
<td>70%</td>
</tr>
<tr>
<td></td>
<td>Vertical Run Length</td>
<td>68%</td>
</tr>
<tr>
<td></td>
<td>Contour Hinge + Vertical Run Length</td>
<td>76%</td>
</tr>
<tr>
<td>Text-Dependent</td>
<td>Whole Letters</td>
<td>92%</td>
</tr>
<tr>
<td></td>
<td>Letter Parts</td>
<td>79%</td>
</tr>
<tr>
<td></td>
<td>Concave Hulls</td>
<td>88%</td>
</tr>
<tr>
<td>Combined</td>
<td>Whole Letters + Parts + Hulls</td>
<td>98.7%</td>
</tr>
<tr>
<td></td>
<td>Text-Dependent + Text-Independent</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 5. Assessment of current prototype. Results indicate the rate for the prototype to successfully identify manuscript pages written by the same scribe. Given a dataset of twenty manuscripts, the chance of the system randomly guessing correctly was just over 5 percent. When the system used only text-independent methods (overall writing style), the computer correctly identified the scribe 76 percent of the time. When it used only text-dependent methods (comparing individual letters) the success rate was over 98 percent. When it used all five methods the computer was 100 percent accurate.