NATIONAL ENDOWMENT FOR THE HUMANITIES

OFFICE OF **DIGITAL HUMANITIES**

Narrative Section of a Successful Application

The attached document contains the grant narrative and selected portions of a previously funded grant application. It is not intended to serve as a model, but to give you a sense of how a successful application may be crafted. Every successful application is different, and each applicant is urged to prepare a proposal that reflects its unique project and aspirations. Program guidelines also change and the samples may not match exactly what is now required. Please use the current set of application instructions to prepare your application.

Prospective applicants should consult the current Office of Digital Humanities program application guidelines at https://www.neh.gov/grants/odh/digital-humanities-advancement-grants for instructions.

Applicants are also strongly encouraged to consult with the NEH Office of Digital Humanities staff well before a grant deadline.

Note: The attachment only contains the grant narrative and selected portions, not the entire funded application. In addition, certain portions may have been redacted to protect the privacy interests of an individual and/or to protect confidential commercial and financial information and/or to protect copyrighted materials.

Project Title: *Transkribus and the Georgian Papers Programme Tabular-Formatted Manuscripts*

Institution: College of William and Mary

Project Directors: Deborah Cornell and Zhenming Liu

Grant Program: Digital Humanities Advancement Grants, Level II

1. List of Participants

Principal Investigators:
Cornell, Deborah, Head of Digital Services, Librarian, William and Mary
Liu, Zhenming, Assistant Professor, Computer Science, William and Mary

Researchers/Transcribers:
Zhu, Xiaodan, PhD candidate, Computer Science, William and Mary
[TBN], Graduate student, Humanities, William and Mary

Undergraduate students:
[TBN]

Evaluation Board:
Hervé Déjean, Computer Science Researcher, Naver Labs Europe, READ Project
Florian Kleber, Senior Scientist, Computer Vision Lab, Vienna University of Technology, READ Project
Arthur Burns, Professor of Modern British History, and Academic Director of the Georgian Papers Programme, Kings College London.
Cynthia A Kierner, Professor of United States History, George Mason University, Virginia
[TBN] 5 self-selected evaluators that will include Transkribus users, historians, and archival professionals.

Administrative Support:
Beth Brown, Library Accountant, William and Mary

Letters of Support:
Jennifer Stertzer, Director, Center for Digital Editing; Senior Editor, Washington Papers
Anne Helmreich, Associate Director, Digital Initiatives, The Getty Research Institute

1. Narrative

Enhancing the Humanities through Innovation

In the past ten years, researchers have made remarkable advances to handwritten text recognition (HTR) software. Through optical character recognition, optical layout recognition, and deep learning for handwritten text recognition, many applications can now analyze handwritten documents and produce transcriptions that are usable for computer and text-processing applications. Text-based transcriptions are then available for complex analysis by scholars and students. Transcription applications are not perfect, but, depending on legibility of handwriting and the size of the sample available, they can approach 95% accuracy in capturing text.

However, machine-reading and recognition often fails in its approach to tabular data, or information presented in columns and rows. Programmers are developing applications that can detect a well-populated and symmetrical table consisting of columns and rows, but when the handwritten document contains a ledger or chart that scatters data across the page in increasingly sporadic cells, the applications do not create the correct number of blank spaces and accurate relationships in a tabular fashion.

Over the past two years, William and Mary Libraries staff have built a strong program of transcription, centered both around our own special collections and transcriptions of the papers of King George III in the Georgian Papers Programme (https://www.rct.uk/collection/georgian-papers-programme), a project managed in collaboration with the Royal Archives at Windsor Castle in the United Kingdom, and King's College London. During our work with these collections, we have encountered many such ledgers and tabular records where no machine reading can be utilized in their transcription. Though we have worked with an open-source handwritten text recognition tool, Transkribus (https://transkribus.eu/) to develop a HTR model to read the script of George III, the coding is in its infancy in handling complex tabular data, where text is not formatted in consistent grid-like formats. Moreover, we, like other Transkribus users, would benefit from improvement in functionality in export of text into spreadsheets.

William and Mary Libraries possess enormous collections of tabular-formatted manuscripts from account books and inventories, to business and household records. These materials contains a wealth of data that provide insight into the lives and happenings of multiple communities. For example, the bursars' records of our 325 year old school include ledgers and rosters with details about students and faculty over a history that spans multiple wars and historical events. The stories that can be told through the tables and charts and ledgers throughout our handwritten texts are captivating and important.

As we have worked on the Georgian Papers, we have encountered similar examples: the Georgian mensil books detail the supply of food and household goods, including the names of the suppliers and costs for the royal residences. Scholars use this information for analysis and research to understand the development of a wide range of humanities questions in social and political structures, economics, foodways, and culture. A remarkable example is a volume in Princess Charlotte's records of the food and household items such as candles, lamp oil, china, and glass for her residences. From this type of document, we know more about the quotidian life in these residences, especially about the service help that was employed and their daily work.

Proposal Goal

This proposal seeks a Level II Digital Humanities Advancement Grant from the NEH-ODH to experiment with Transkribus on a subset of tabular formatted materials in the Georgian Papers Programme (GPP).

NEH funding would support: a) development of new machine learning techniques will further enhance Transkribus' capabilities in layout analysis, document comprehension, and automatic text recognition of tabular manuscript materials;; b) algorithmic processing of approximately 50,000 pages of images from GPP manuscripts; c) writing documentation, code, and user guides; and d) presentation of the new developments and standards at Transkribus user conferences, and host a workshop for historians to evaluate the tool.

Through Transkribus' API, we will advance the functionality to read tabular data, notably asymmetrical data, and translate it into a standard tabular format. As part of this development, we will build upon the capability to translate and output the text into spreadsheets. Transcribing text into a spreadsheet will allow the ability to download this data in delimited files. In addition to the creation of the application, we will share this functionality and transcription standard with others engaging in the transcription of handwritten text in order to solve a common challenge in computer science, the recognition and transcription of handwritten text when it does not correspond to a narrative style and structure. Our goal is not only to solve a problem which we encounter regularly as part of our transcription projects, but to address a thorny issue that digital humanities scholars will continue to grapple with as machine recognition tools become more widely used and trusted.

Problem Statement and Motivations

As described, the exact design of ledgers and account books is problematic for HTR transcription especially when the complex layout of data and the relationships between data are difficult for the HTR processors to detect. A frequently cited example is the idioms of ledgers, commonly referred to as the 'curly bracket' problem, where the relationship of **one to many** is frequently expressed in multiple ways (See Appendix A for examples).

As open source HTR has matured and HTR tools have become available to manage simple manuscripts such as correspondence and diaries with high success rates, development for HTR on tabular manuscript documents is just beginning. In Europe, the large e-infrastructure project called READ (Recognition and Enrichment of Archival Documents https://read.transkribus.eu), funded by the European Commission, combines research, services, and network building. One aspect of READ's service portfolio is Transkribus (https://transkribus.eu), an HTR tool and platform maintained by researchers at the University of Innsbruck in Austria. In 2017, we presented our progress and use of the tool on the large corpus of the Georgian Papers at the Transkribus users' conference; the corpus includes 425,000 pages of documents in approximately 170 separate collections. In 2018 we continued to work on a limited basis with Transkribus, as we awaited more content from our GPP partners. Our relationship with Transkribus is notable especially because we are a U.S. partner, where many of their current users and projects have come from European universities and grant projects.

Approaches

Taking an intuitive, bottom-up approach, we will validate our algorithm by using historical documents from the Georgian Papers, although we anticipate our technologies will be more widely applicable.

Our technologies will feature three major components (steps):

1. Building tabular classifiers that decide whether a specific page contains a table.
2. Building a cell detector that distinguishes all cells in a table and a layout analyzer that interprets the relationships among cells in the tabular data.
3. Translating the relationship representation of cells into knowledge graphs that enables easy searching and interpretation by end-users.

Component 1: Building Tabular Classifiers

Existing table-detection software and techniques are optimized for modern typography and thus are unlikely to be directly applicable to historical archives. Our idea is to leverage both supervised and unsupervised techniques simultaneously to construct new classifiers optimized for historical documents. In supervised learning, we ask humans to label the location and patterns of tables across a small collection of scanned pages, and then we design computer algorithms that can recognize patterns from the labelled data and generalize the patterns to increase the capability of the program to determine whether a previously unlabeled page contains a table. In unsupervised learning, a computer algorithm can automatically discover the hidden structure of the data. It will know there are two types of pages (tabular and non-tabular), but it will not understand the semantics of the tabular/non-tabular data.

Supervised learning more effectively extracts hidden patterns from labeled data but labeling data is costly. Unsupervised learning does not need to rely on labeled data but it does need a substantially larger volume of data (i.e., many pages of documents) to identify patterns, and it performs poorly for low signal-to-noise data (i.e., when the pattern is not clear).

Our supervised classifier will be constructed via the so-called "convolutional neural net" technology inspired by the biological structure of neurons. Here, the input is the scanned documents and the output/response is the labels. Our unsupervised classifier will use the observation that tabular data usually consist of small blocks of text as well as horizontal and vertical lines, and it will execute a k-means algorithm (or PCA k-means) to extract the hidden structure. Finally, we will build an in-house boosting algorithm (similar to adaboost) that consolidates the signals from both supervised and unsupervised learning procedures.

Component 2: Cell Detector and Relationship Graphs

After we identify a table in a page by using the building block above, the first challenge is to detect all cells in a table. We will focus on interpreting the geometric structure of the text blocks and the lines. For example, when the blocks align horizontally or vertically or lie inside rectangular line boxes, they are likely to be the cells we need to detect. To represent the relationship among the cells, we will build graph-based data-mining methods, where a cell corresponds to a node and two cells/nodes connect if and only if they are adjacent.

Component 3: Knowledge Graphs.

While graph representation of tables makes it possible to understand the local interactions between cells, they do not directly help the user to search and interpret these tables. We will create a specialized natural language processing technique for interpreting the meaning of the cells (represented by relational graphs) by clustering tables with similar semantics together and building a knowledge graph on top of each "type" of the tables. Users will be able to answer sophisticated questions from the data (i.e., "how much does the King spend on dining?"). Our algorithm/technology will have two components: an efficient heuristics that clusters similar relational graphs together, and neural nets that extract knowledge graphs from relational graphs. When the relational graphs for two tables are homomorphic to each other, they are more likely to have the same semantic meaning. Testing whether two graphs are homomorphic is usually computationally intractable, but the relational graphs we construct will have unique characteristics (e.g., connecting all the planar graphs) which can be leveraged in the design of our heuristics.

Regarding the neural nets, our idea is to embed the various subgraphs of the relational graphs and their associated texts into high-dimensional space, and use the neural nets to optimize the embedding so that similar subgraphs are embedded to the points close to each other in the high dimensional space.

Environmental Scan

This project draws on and advances recent research on text digitization tools and their applications for humanities scholarship. Much of the current body of research has sought to advance Optical Character Recognition (OCR) technologies, addressing techniques such as optical modeling, text enhancement, and computer learning. The Early Modern OCR Project at Texas A&M University (http://emop.tamu.edu/), merges book history, textual analysis, and machine learning to produce a corpus of keyed texts with a higher level of accuracy than previously possible. The University of Salford's PRImA Research Lab (http://www.primaresearch.org/) creates advanced programs using document image analysis, character recognition, and pattern analysis to enable humanities scholars to run text recognition and annotation software on scanned early modern printed texts. Developing out of this work, Handwriting Text Recognition (HTR) is an emerging technique in computer science, and there is a growing body of research addressing its application to archival documents and humanities research. The READ Project (https://www.read-project.eu/) has begun to conduct research into HTR technology, producing technologies like Transkribus. Enhancing this innovative technology further, developers at the University of Innsbruck, Computer Vision Lab (https://cvl.tuwien.ac.at/) and Naver Labs Europe (http://www.europe.naverlabs.com/) collaborated to develop the Table Editor (http://truben.no/table/), a segmentation tool that allows the program to read and transcribe simple tables.

This project will build on the work undertaken by these three labs, developing a tool to improve upon reading of elaborate tables and address the problem of complex relational issues within tabular data. Currently, projects that work with difficult handwritten data rely on crowd-source citizen science or Text Encoding Initiative (TEI) mark-up to transcribe challenging text. The most recent Zooniverse initiative, Weather Rescue (https://www.zooniverse.org/projects/edh/weather-rescue), is an example that relies on crowd-sourcing to uncover lost meteorological observations. Similarly, the Smithsonian Transcription Center (https://transcription.si.edu/) mobilizes a crowd-source platform to transcribe a number of handwritten texts in their collection. The NEH funded project, 'Encoding Financial Records for Historical Research' by Kathryn Tomasek, explored models for TEI mark-up of financial records to capture the semantic relationships between data. These processes are labor-intensive and time-consuming. In transcription workshops help for the Georgian Papers, historians acknowledged the difficulty of acquiring reusable data from tabular manuscript materials. The development of an HTR technology to address these complicated tabular and data relational issues will enhance Transkribus user capabilities, providing accurate transcriptions that will decrease labor-intensive processes, and allow historians and researchers easier access to the data.

History of Project

This proposal is an offshoot of the Georgian Papers Programme (GPP), a partnership between the Royal Collection Trust and King's College London and is joined by primary United States partners: the Omohundro Institute of Early American History & Culture, William and Mary Libraries, and a few other participating U.S. institutions, including Mount Vernon and the Library of Congress. The Programme is a ten-year interdisciplinary project capitalizing on the mutual work of scholars, librarians, technologists, and digital specialists. In a long-range initiative, the GPP will digitize, disseminate, and interpret in overlapping stages of discovery, access, and interpretation the archive of the Georgian monarchs held in the Royal Archives at Windsor Castle. The project aims to provide these digitized materials documenting the Hanoverian Dynasty, dating from 1714–1837, online by 2027, creating an online archive and library

available to all, academics and the public.

William and Mary Libraries are working closely with King's College and the Omohundro Institute on the technical infrastructure and process for transcription and metadata enrichment. This process was supported in part by a 2017 NEH HCRR planning grant to the Omohundro, which enabled William and Mary Libraries to explore transcription tools. From the outset, the GPP partners realized multiple transcription methods would be necessary for a project with 425,000 pages. In the planning stage, two methods were selected for development, Omeka + Scripto for public crowd-source transcription (http://transcribegeorgianpapers.wm.edu/) and the Handwritten Text Recognition platform, Transkribus, developed under the READ Project, a European Union Horizon 2020 programme.

William and Mary Libraries' evaluation and testing of Transkribus through a small pilot was encouraging in terms of usefulness of product and indicative workflows. The pilot produced 3,500 manually transcribed pages that enabled the software to learn the handwriting and create a Geo III HTR model. The Geo III model originally returned an average error rate of 13% when run over new (non-transcribed) images. In the last year, the READ project made enhancements to the HTR + engine, and the model was improved to a 7% error rate.

In the evaluation period, an area of challenge was in HTR for tabular format materials, specifically account and ledger books. In discussions with the historians doing researcher with these materials, we hit upon their need to have the information contained within these materials to be in a data format, e.g. spreadsheet. As approximately 20% of the GPP collections are account type bound manuscripts, largely written in secretary hand, this challenge presented itself as the perfect use case to experiment on tabular formatted manuscripts and contribute to the Transkribus client. In 2018, Deborah Cornell approached computer science faculty at W&M, Dr. Zhenming Liu to enlist his expertise in machine learning and programming and to collaborate on the table project.

William and Mary, on behalf of the GPP, has a Memorandum of Understanding with the READ Project. In addition to, standing Memos of Understanding among the GPP partners, with William and Mary Libraries committed to the transcription for the length of the GPP.

Work Plan
W&M Libraries request Level II funds to support work for 18 months from September 2019 to February 2020. The work will be an iterative process – design, test, evaluate, refine, and deliver – which will continue through the entire work plan. Each plan task serves as a building block for the subsequent task. Evaluation will take place in two areas: 1) technical accuracy rate of output during each task of the project, and 2) complexity of process required to define and output a table by users (transcribers) of Transkribus. Appendix B provides a detailed work plan. The five main tasks are outlined here:

1. Task 1: Classifiers | September 2019 – January 2020
   Development of table detection with student transcribers working to manually label cells within Transkribus and use of their work to build supervised and unsupervised classifiers to teach pattern recognition in the program and test its performance.

2. Task 2: Cell Detection | February 2020 – May 2021
   Interpretation of the geometric structures of tabular data in images, interpreting the manuscript content and output requirements.

3. Task 3: Algorithms for Relationship Graphs | June 2020 – September 2020
   Development of the algorithms to represent the relationships among the table cells in graph form. Presentation of results at the Transkribus User and Digital Humanities conferences.

4. Task 4: Knowledge Graphs and Integration | September 2020 – December 2020
   Creation of a natural language processing algorithm for interpreting relationships between cells. Analysis by Evaluation Board that includes humanities researchers, and Transkribus users for tabular transcription process and project goals.

5. Task 5: Write Project Documentation and Create Tutorials | January 2021 – February 2021
   Create Transkribus user guides and tutorials for new functionality. Host a workshop for historians and archival professionals to introduce them to Transkribus, and evaluate the process and user guides developed in this project. Write the NEH White Paper.

Risks

The nature of risks is different for different tasks. The major risk for task 1 and 2 is that existing machine learning algorithms are not directly optimized for detecting tables' structure so customized machine learning solutions may be needed. In general, these are relatively simple tasks. Co-PI Liu has completed projects in similar complexity and is confident that task 1 and 2 can be delivered successfully.

The major risk for task 3 and 4 is that generic solutions exist, but these solutions may not be optimized. Developing specialized solutions for historical archives could substantially improve the performance, but it is also riskier. We can mitigate our risk by developing two solutions simultaneously: polishing existing open source packages and conducting original research to build specialized algorithm for our datasets.

Final Product and Dissemination

This project will produce products in four areas:

1. Transkribus page layout templates and output of transcription in spreadsheet format. The source code and documentation will be freely and publicly available via GitHub for download, distribution and, for continued contributions and modifications.

2. Documentation outlining the Transkribus workflow and procedures undertaken with the GPP tables project. Documents will include user guides, tutorials, and transcription protocols. The http://transcribegeorgianpapers.wm.edu website includes behind-the-scenes news of GPP transcription, and reports on the evolving transcription work will be posted. The Transkribus documents will be made freely and publicly available there, in William and Mary's institutional repository, and contributed to the READ project.

3. A white paper for the NEH assessing Transkribus for HTR of tabular formatted manuscript materials.

4. Project participants will produce articles, conference papers and presentations on the project and research undertaken to share with the digital libraries and archives, humanities and computer science communities

1. Biographies

Deborah Cornell, Head of Digital Services and Metadata Librarian at William and Mary Libraries provides leadership for the libraries' emerging digital collections and initiatives. She holds a M.I.S. in digital collections from University of North Texas. Her experience in metadata, digitization, transcription and project management was obtained from positions as Digital Archives Cataloger at UCLA, digital asset coordinator at JibJab.com, and through several positions, including her most recent role, at William and Mary. Cornell leads the Georgian Papers Programme transcription project and serves as the Virginia Hub representative on the DPLA National Council.

Cornell will manage the humanities perspective on this project and oversee the Transkribus transcription.

Zhenming Liu is an Assistant Professor in the Computer Science Department at William and Mary. He received his Ph.D. in the theory of computation from Harvard University in 2012. He was a Postdoctoral Research Associate at Princeton University in 2012 and 2014, and a machine learning researcher at Two Sigma Investments from 2014 through 2016. Currently, Dr. Liu is using tools from applied probability, theoretical computer science, and optimization to build scalable systems for analyzing massive datasets.

Dr. Liu will manage the research effort on design of machine learning systems in this project.

Xiaodan Zhu is currently a PhD candidate at William and Mary. His research interests center around computer vision and computer graphics, and he works with Dr. Zhenming Liu. His current research is handwriting ancient archive content recognition. Zhu has an ME in Computer Science from University of Electronic Science and Technology of China and a Bachelor of Computer Science from Jiangnan University in China.

Zhu will lead the effort of research, implementation, and evaluation of machine learning algorithms described in the proposal.

Graduate Assistants [TBN] at William and Mary will provide Transkribus transcription instruction, quality control, and supervision to undergraduate transcribers, in addition to updating transcription protocol documentation. The assistants serve as liaisons to the computer science Ph.D. candidate, Xiaodan Zhu.

Undergraduate Students [TBN] at William and Mary will carry out Transkribus transcription, while computer science undergraduates will assist Zhu and Liu.

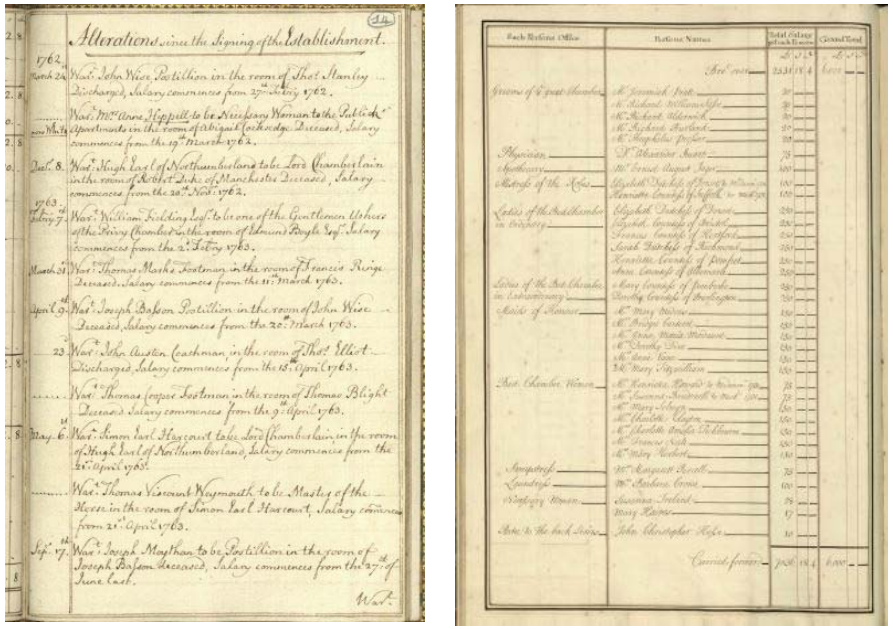
An **Evaluation Board** composed of Transkribus developers and users, historians and archival professionals will evaluate, the developed transcription process and technological improvement to the functionality for analyzing tabular formatted documents, and the data (csv/spreadsheet)output. Two Transkribus developers, **Florian Kleber** and **Hervé Déjean**, are confirmed as technical evaluators, as well as, historians, **Arthur Burns** and **Cynthia Kierner**, and additional members that will include historians and archival professionals will be recruited from the United States and the United Kingdom.

Beth Brown, William and Mary Libraries' accountant, will provide administrative support for the financial aspects of the project. She has been working for the University for fourteen years in positions of increasing responsibility and oversight of fiscal matters.

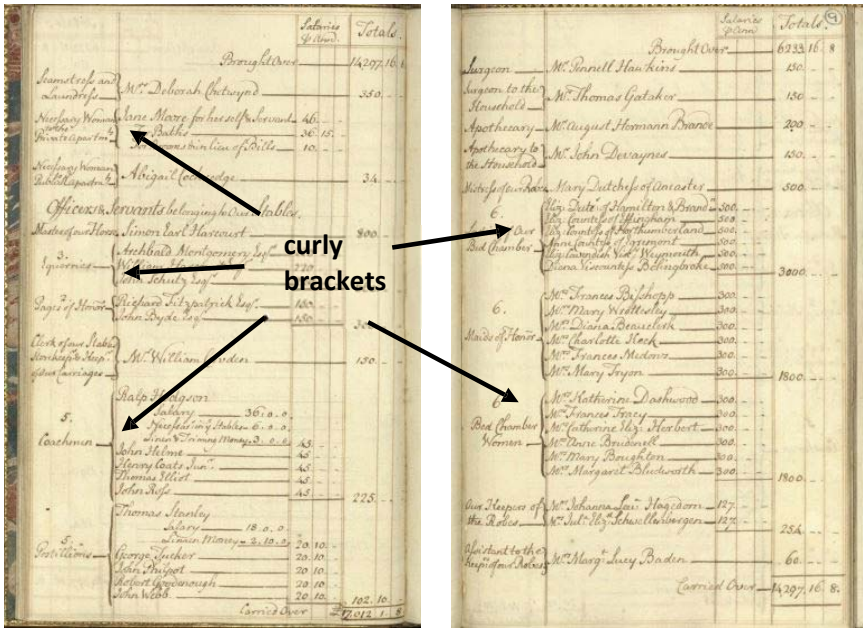
Appendix A: Appearance of Complex Tabular Data in the Georgian Papers Collection

A selection of pages from the Georgian Papers Project illustrating the variety of tabular data found in this collection.

One of the HTR data issues present in the collection is the one to many relationship. This complex relational problem is evident in the Establishment Lists in two forms: bracketed (the 'curly bracket' problem) and not bracketed. Each relationship provides a unique issue for the HTR program that will be addressed in this project.

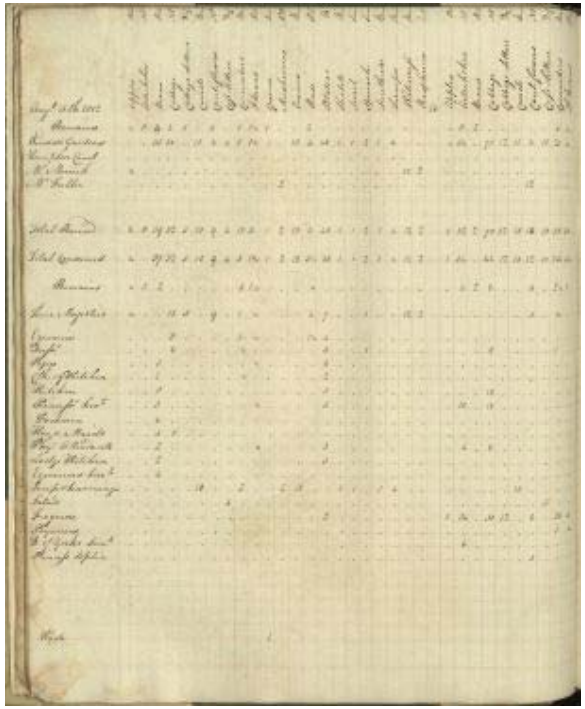


Examples of non-bracketed "one to many" relationships

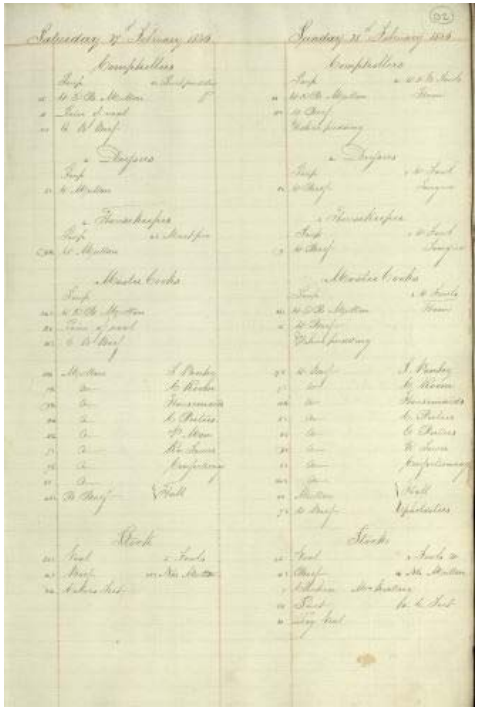


Example of bracketed "one to many" relationships

A second HTR issue present in the collection is the issue of tables; both in grid (hand drawn in) and non-grid formats. Examples these tabular issues can be found both in menu books and mensils. Similarly to the “one to many” relationships, these tabular issues each present a unique issue for the HTR program which will be addressed in this project.

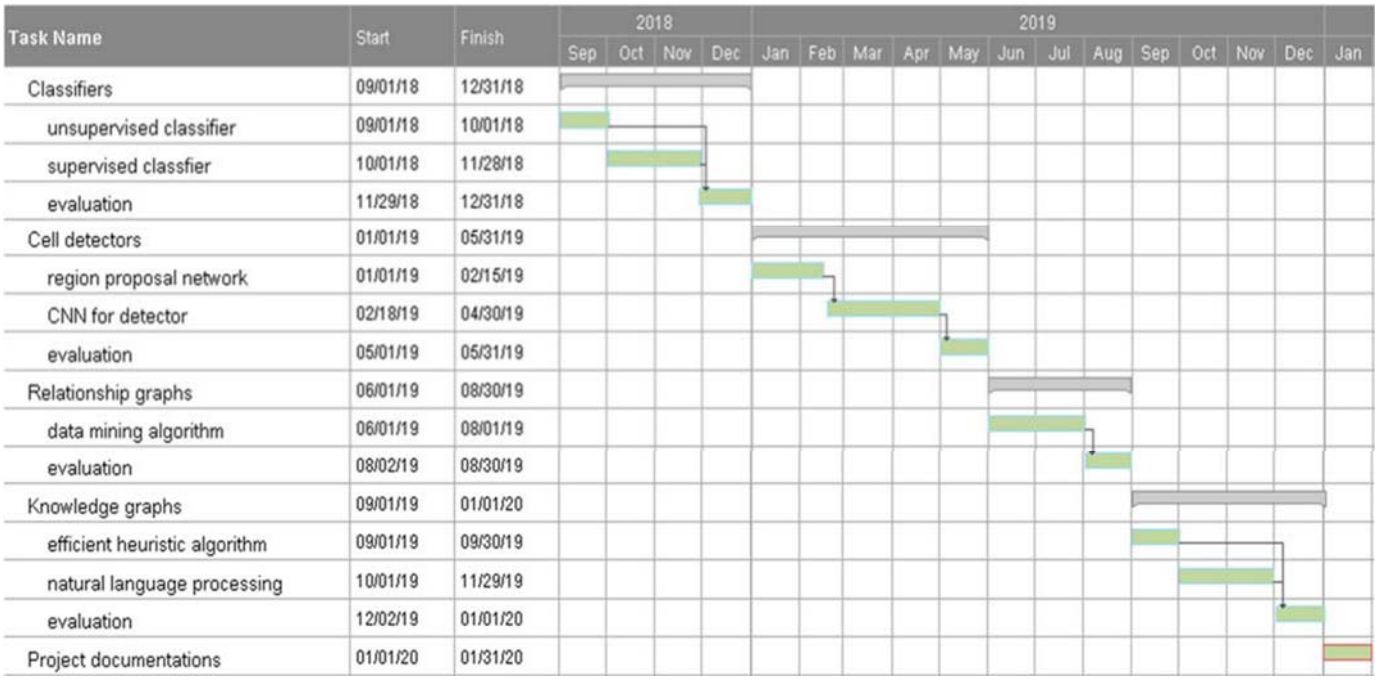


Example of a grid table



Example of a non-grid table

Appendix B. Detailed work plan



Task Name	Start	Finish
Classifiers	09/01/18	12/31/18
unsupervised classifier	09/01/18	10/01/18
supervised classfier	10/01/18	11/28/18
evaluation	11/29/18	12/31/18
Cell detectors	01/01/19	05/31/19
region proposal network	01/01/19	02/15/19
CNN for detector	02/18/19	04/30/19
evaluation	05/01/19	05/31/19
Relationship graphs	06/01/19	08/30/19
data mining algorithm	06/01/19	08/01/19
evaluation	08/02/19	08/30/19
Knowledge graphs	09/01/19	01/01/20
efficient heuristic algorithm	09/01/19	09/30/19
natural language processing	10/01/19	11/29/19
evaluation	12/02/19	01/01/20
Project documentations	01/01/20	01/31/20

Work Plan Outline

Task 1: Classifiers – Table detection (September 2019 - January 2020)

- Technology lead: Zhenming Liu and Xiaodan Zhu
- Data management and platform lead: Deborah Cornell and TBN Humanities graduate students.

Development of table detection begins with Cornell, a graduate student (TBN) and transcribers manually labeling tables, cells, and data in the manuscript images with the Transkribus Client. Using labeled data, Liu and Zhu will develop pattern recognition algorithms that automatically detect tabular structures within manuscript images. We will build both supervised and unsupervised classifiers for detecting tables, allowing one to two months for each classifier, and develop a rigorous procedure to evaluate their performance.

- Unsupervised classifiers: we will examine k-means, kernel methods, and kernel-PCA on the image datasets.
- Supervised classifiers: we will focus on support vector machines and convolutional neural nets.
- Evaluation: We will primarily use labeled data to evaluate the performance of algorithms we examine.

Task 2: Cell Detectors – Cell detection (February – May 2020)

- Program management and evaluation: Zhenming Liu
- Technology lead: Zhenming Liu and Xiaodan Zhu
- Data management and platform lead: Xidaodan Zhu and undergraduate researchers hired through computer science department.

Liu and Zhu will analyze and interpret the geometric structures of the tabular data in images. Cornell and the graduate student will examine two deep-learning based technologies, including region proposal network, and convolutional neural net (Goodfellow et al. 2016), for interpreting the manuscript content and output requirements. Task 2 will require labeled data. We will focus on examining region proposal networks and convolutional neural nets.

Task 3: Algorithms for relationship graphs – Construction of algorithm (June – September 2020)

- Program management and evaluation: Deborah Cornell
- Technology and research lead: Zhenming Liu and Xiaodan Zhu.
- Domain expert consultant: Deborah Cornell and TBN Humanities graduate assistants.

Liu and Zhu will develop the algorithms to represent the relationships among the table cells in graph form. Their technologies will rely on prior graph-based data mining algorithms developed by Dr. Liu's lab and other open source software (Deepayan et al. 2012, Li et al. 2017). Techniques we will examine include dynamic programming (for cell alignment), spectral graph techniques for data mining, and frequent-items detection algorithms.

Cornell and the graduate student will provide guidance and context for interpretation of manuscript content and output requirements.

We will present the project's status at Transkribus Users and/or Digital Humanities conferences in 2020.

Task 4: Knowledge graphs and integration – Creation of simple language queries (September 2020 – December 2021)

- Program management and evaluation: Deborah Cornell
- Technology and research lead: Zhenming Liu and Xiaodan Zhu.
- Domain expert consultant: Deborah Cornell and TBN Humanities graduate assistants.

Liu and Zhu will create a natural language processing algorithm for interpreting relationships between cells. This consists of two main components: an efficient heuristic that detects similarities between relationship graphs, and natural language processing techniques. Algorithms will enable users to make simple string searches and writing code to output data in spreadsheet form. Cornell and the graduate student will lead a final evaluation and assessment of the complete tabular transcription processes with Transkribus.

Cornell will create Transkribus user guides and tutorials for new functionality. Technical members of the Evaluation Board will critique and provide feedback on the HTR functionality. Where historians, humanities researchers and archival professionals Board members will evaluate the use of Transkribus, user guides, and results of product output.

Task 5: Write project documentation and create tutorials (January 2021 – February 2021)
Cornell and Liu will write the NEH white paper and finalize Transkribus user guides/tutorials.

Appendix C. Transkribus and The READ project.

The Transkribus platform development came forth under the tranScriptorium project (http://transcriptorium.eu/), a Specific Targeted Research Project (STREP) funded by the European Commission Seventh Framework from 2013-2015. tranScriptorium's aims were to "mature" Handwritten Text Recognition (HTR) technology to create an user-friendly and efficient computer-assisted transcription application which would enable non-technologically sophisticated researchers and cultural heritage institutions to transcribe historical handwritten documents. A consortium was formed among six research institutions with expertise in HTR and Document Image Analysis (DIA) and historical documents transcription to provide integration experience and content. Their research evolved methods in interactive-predictive HTR, namely in indexing, searching on handwritten text images, and word spotting. These methods are the technological foundation that were integrated into a user client, the Transkribus platform (http://transcriptorium.eu/demonstrations/).

In 2016, tranScriptorium project members joined with ten additional partner institutions for the READ (Recognition and Enrichment of Archival Documents) project (https://read.transkribus.eu/about/). Funded by the European Union's Horizon 2020 research and innovation programme grant, the objective of READ is further development of Transkribus as a service platform through research, services, and network building. Collaboration among the domains of computer science, archives & libraries, and humanities research remains a core aspect of READ. Fundamental research in HTR, DIA, computer vision, natural language processing and web technology integration continues to improve the platform. New services and tools that automate processes required in the production of training data were made available in the last year (https://read.transkribus.eu/services/). The advanced document layout analysis can detect regions and lines of text (a process called segmentation), thereby eliminating the need for transcribers to manually identify these on document images. In cases where transcript copy exist, the Text2Image matching tool, automatically syncs the transcript txt files and image files of documents then preforms segmentation on the content. In 2018, READ made strides in tabular format reading for documents of a consistent and graphically visible grid lines, such as the Passau Diocesan Archives (https://read.transkribus.eu/news/page/2/). A guide on 'How to Process Tables in Transkribus" was produced in November 2018.

A major objective of the READ programme is the creation of a network of institutions and researchers involved in Transkribus, ensuring sustainability and continued development of the application. Institutions and projects may join the project consortium through a READ Memorandum of Understanding (MoU). To date, 81 institutions have signed MOUs with READ (https://read.transkribus.eu/network/). Notable parties include the British Library, National Library of Spain, and The Hessian State Archives. William and Mary signed a MOU with READ in 2018.

Appendix D: References

Early Modern OCR Project. Texas A&M University. http://emop.tamu.edu/ [accessed January 10, 2018]

University of Salford. PRImA Research Lab. http://www.primaresearch.org/ [accessed January 10, 2018]

Encoding Historical Financial Records. Kathryn Tomasek, Wheaton College, Massachusetts. http://www.customization.encodinghfrs.org/ [accessed January 10, 2018]

Weather Rescue: European Daily Weather Reports. Zooniverse. https://www.zooniverse.org/projects/edh/weather-rescue [accessed January 10, 2018]

Smithsonian Transcription Center. Smithson Digital Volunteers. https://transcription.si.edu/ [accessed January 10, 2018]

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016. Harvard

Chakrabarti, Deepayan, and Christos Faloutsos. "Graph mining: laws, tools, and case studies." Synthesis Lectures on Data Mining and Knowledge Discovery 7, no. 1 (2012): 1-207.
Harvard

Li, Cheng, Felix MF Wong, Zhenming Liu, and Varun Kanade. "From which world is your graph." In Advances in Neural Information Processing Systems, pp. 1468-1478. 2017.

9. Data Management Plan

Final Project Data

Data Type	Point of Dissemination	Condition of Availability
Open source computer code developed for templates, algorithms, and export tool, including documentation	Conclusion of project	Code will be freely available on GitHub.
Documentation detailing Transkribus workflows, user guides, and transcription protocols	Conclusion of project	Documentation will be freely available from the GPP transcription website*, shared with the READ project and W&M's institutional repository.
Interim reports	At the time of being written over the duration of the project	Reports will be freely available on the GPP transcription website and W&M's institutional repository.
White paper	After the completion of the project	Reports will be freely available on the GPP transcription website and W&M's institutional repository.
NEH final report	At the conclusion of the project	The final report will be submitted to the NEH for dissemination. A copy will be freely available on the GPP transcription website and W&M's institutional repository.

*The Transcribe Georgian Papers website at http://transcribegeorgianpapers.wm.edu/

Period of Data Retention

Data and formal reports will be made publically available within 6 months of project completion. Data will be retained for a minimum of 5 years beyond the completion of the project.

Data Formats and Dissemination

Computer code and related documentation will be freely available on GitHub, a publicly accessible and widely used code repository. Documentation and reports will be available in PDF format and deposited in William and Mary's institutional repository, W&M Scholar Works (https://scholarworks.wm.edu/). The GPP transcription website will serve as a portal and communications arm for the project. Documentation and code will also be shared with the READ project, which has permission to disseminate as they see fit.

Data Storage and Preservation of Access

Over the duration of the project, working data and documentation will be stored and maintained on W&M's secured campus network. Final project computer code will be made available on GitHub and be maintained by the project team. Final documentation and reports will be deposited to William and Mary's institutional repository for long-term storage and public access. The institutional repository is a Digital Commons, bepress platform that follows LOCKSS (Lots of Copies Keep Stuff Safe), an archival compliant preservation protocol by providing 'backbone connections and backup generators or in redundant Amazon Web Services availability zones" with offsite backups on Amazon Glacier. W&M Libraries administer and manage the repository and archives quarterly back-up copies of the content. The Libraries are responsible for the long-term preservation of the content.